A Practical Guide To Quantitative Portfolio Trading

Daniel Bloch

30th of December 2014

A Practical Guide To Quantitative Portfolio Trading

Daniel BLOCH [1]
QUANT FINANCE LTD
eBook

30th of December 2014
Version 1.3.1

[1] db@quantfin.eu

Abstract

We discuss risk, preference and valuation in classical economics, which led academics to develop a theory of market prices, resulting in the general equilibrium theories. However, in practice, the decision process does not follow that theory since the qualitative aspect coming from human decision making process is missing. Further, a large number of studies in empirical finance showed that financial assets exhibit trends or cycles, resulting in persistent inefficiencies in the market, that can be exploited. The uneven assimilation of information emphasised the multifractal nature of the capital markets, recognising complexity. New theories to explain financial markets developed, among which is a multitude of interacting agents forming a complex system characterised by a high level of uncertainty. Recently, with the increased availability of data, econophysics emerged as a mix of physical sciences and economics to get the best of both world, in view of analysing more deeply assets' predictability. For instance, data mining and machine learning methodologies provide a range of general techniques for classification, prediction, and optimisation of structured and unstructured data. Using these techniques, one can describe financial markets through degrees of freedom which may be both qualitative and quantitative in nature. In this book we detail how the growing use of quantitative methods changed finance and investment theory. The most significant benefit being the power of automation, enforcing a systematic investment approach and a structured and unified framework. We present in a chronological order the necessary steps to identify trading signals, build quantitative strategies, assess expected returns, measure and score strategies, and allocate portfolios.

I would like to thank my wife and children for their patience and support during this adventure.

I would like to thank Antoine Haddad and Philippe Ankaoua for giving me the opportunity, and the means, of completing this book. I would also like to thank Sebastien Gurrieri for writing a section on CUDA programming in finance.

Contents

0.1 Introduction

0.1.1 Preamble

There is a vast literature on the investment decision making process and associated assessment of expected returns on investments. Traditionally, historical performances, economic theories, and forward looking indicators were usually put forward for investors to judge expected returns. However, modern finance theory, including quantitative models and econometric techniques, provided the foundation that has revolutionised the investment management industry over the last 20 years. Technical analysis have initiated a broad current of literature in economics and statistical physics refining and expanding the underlying concepts and models. It is remarkable to note that some of the features of financial data were general enough to have spawned the interest of several fields in sciences, from economics and econometrics, to mathematics and physics, to further explore the behaviour of this data and develop models explaining these characteristics. As a result, some theories found by a group of scientists were rediscovered at a later stage by another group, or simply observed and mentioned in studies but not formalised. Financial text books presenting academic and practitioners findings tend to be too vague and too restrictive, while published articles tend to be too technical and too specialised. This guide tries to bridge the gap by presenting the necessary tools for performing quantitative portfolio selection and allocation in a simple, yet robust way. We present in a chronological order the necessary steps to identify trading signals, build quantitative strategies, assess expected returns, measure and score strategies, and allocate portfolios. This is done with the help of various published articles referenced along this guide, as well as financial and economical text books. In the spirit of Alfred North Whitehead, we aim to seek the simplest explanations of complex facts, which is achieved by structuring this book from the simple to the complex. This pedagogic approach, inevitably, leads to some necessary repetitions of materials. We first introduce some simple ideas and concepts used to describe financial data, and then show how empirical evidences led to the introduction of complexity which modified the existing market consensus. This book is divided into in five parts. We first present and describe quantitative trading in classical economics, and provide the paramount statistical tools. We then discuss quantitative trading in inefficient markets before detailing quantitative trading in multifractal markets. At last, we we present a few numerical tools to perform the necessary computation when performing quantitative trading strategies. The decision making process and portfolio allocation being a vast subject, this is not an exhaustive guide, and some fields and techniques have not been covered. However, we intend to fill the gap over time by reviewing and updating this book.

0.1.2 An overview of quantitative trading

Following the spirit of Focardi et al. [2004], who detailed how the growing use of quantitative methods changed finance and investment theory, we are going to present an overview of quantitative portfolio trading. Just as automation and mechanisation were the cornerstones of the Industrial Revolution at the turn of the 19th century, modern finance theory, quantitative models, and econometric techniques provide the foundation that has revolutionised the investment management industry over the last 20 years. Quantitative models and scientific techniques are playing an increasingly important role in the financial industry affecting all steps in the investment management process, such as

- defining the policy statement
- setting the investment objectives
- selecting investment strategies
- implementing the investment plan
- constructing the portfolio
- monitoring, measuring, and evaluating investment performance

The most significant benefit being the power of automation, enforcing a systematic investment approach and a structured and unified framework. Not only completely automated risk models and marking-to-market processes provide a powerful tool for analysing and tracking portfolio performance in real time, but it also provides the foundation for complete process and system backtests. Quantifying the chain of decision allows a portfolio manager to more fully understand, compare, and calibrate investment strategies, underlying investment objectives and policies.

Since the pioneering work of Pareto [1896] at the end of the 19th century and the work of Von Neumann et al. [1944], decision making has been modelled using both

1. utility function to order choices, and,
2. some probabilities to identify choices.

As a result, in order to complete the investment management process, market participants, or agents, can rely either on subjective information, in a forecasting model, or a combination of both. This heavy dependence of financial asset management on the ability to forecast risk and returns led academics to develop a theory of market prices, resulting in the general equilibrium theories (GET). In the classical approach, the Efficient Market Hypothesis (EMH) states that current prices reflect all available or public information, so that future price changes can be determined only by new information. That is, the markets follow a random walk (see Bachelier [1900] and Fama [1970]). Hence, agents are coordinated by a central price signal, and as such, do not interact so that they can be aggregated to form a representative agent whose optimising behaviour sets the optimal price process. Classical economics is based on the principles that

1. the agent decision making process can be represented as the maximisation of expected utility, and,
2. that agents have a perfect knowledge of the future (the stochastic processes on which they optimise are exactly the true stochastic processes).

The essence of general equilibrium theories (GET) states that the instantaneous and continuous interaction among agents, taking advantage of arbitrage opportunities (AO) in the market is the process that will force asset prices toward equilibrium. Markowitz [1952] first introduced portfolio selection using a quantitative optimisation technique that balances the trade-off between risk and return. His work laid the ground for the capital asset pricing model (CAPM), the most fundamental general equilibrium theory in modern finance. The CAPM states that the expected value of the excess return of any asset is proportional to the excess return of the total investible market, where the constant of proportionality is the covariance between the asset return and the market return. Many critics of the mean-variance optimisation framework were formulated, such as, oversimplification and unrealistic assumption of the distribution of asset returns, high sensitivity of the optimisation to inputs (the expected returns of each asset and their covariance matrix). Extensions to classical mean-variance optimisation were proposed to make the portfolio allocation process more robust to different source of risk, such as, Bayesian approaches, and Robust Portfolio Allocation. In addition, higher moments were introduced in the optimisation process. Nonetheless, the question of whether general equilibrium theories are appropriate representations of economic systems can not be answered empirically.

Classical economics is founded on the concept of equilibrium. On one hand, econometric analysis assumes that, if there are no outside, or exogenous, influences, then a system is at rest. The system reacts to external perturbation by reverting to equilibrium in a linear fashion. On the other hand, it ignores time, or treats time as a simple variable by assuming the market has no memory, or only limited memory of the past. These two points might explain why classical economists had trouble forecasting our economic future. Clearly, the qualitative aspect coming from human decision making process is missing. Over the last 30 years, econometric analysis has shown that asset prices present some level of predictability contradicting models such as the CAPM or the APT, which are based on constant trends. As a result, a different view on financial markets emerged postulating that markets are populated by interacting agent, that is, agents making only imperfect forecasts and directly influencing each other, leading to feedback in financial markets and potential asset prices predictability. In consequence, factor models and other econometric techniques developed

to forecast price processes in view of capturing these financial paterns at some level. However, until recently, asset price predictability seemed to be greater at the portfolio level than at the individual asset level. Since in most cases it is not possible to measure the agent's utility function and its ability to forecast returns, GET are considered as abstract mathematical constructs which are either not easy or impossible to validate empirically. On the other hand, econometrics has a strong data-mining component since it attempts at fitting generalised models to the market with free parameters. As such, it has a strong empirical basis but a relatively simple theoretical foundation. Recently, with the increased availability of data, econophysics emerged as a mix of physical sciences and economics to get the best of both world in view of analysing more deeply asset predictability.

Since the EMH implicitly assumes that all investors immediately react to new information, so that the future is unrelated to the past or the present, the Central Limit Theorem (CLT) could therefore be applied to capital market analysis. The CLT was necessary to justify the use of probability calculus and linear models. However, in practice, the decision process do not follow the general equilibrium theories (GET), as some agents may react to information as it is received, while most agents wait for confirming information and do not react until a trend is established. The uneven assimilation of information may cause a biased random walk (called fractional Brownian motion) which were extensively studied by Hurst in the 1940s, and by Mandelbrot in the 1960s and 1970s. A large number of studies showed that market returns were persistent time series with an underlying fractal probability distribution, following a biased random walk. Stocks having Hurst exponents, H, greater than $\frac{1}{2}$ are fractal, and application of standard statistical analysis becomes of questionable value. In that case, variances are undefined, or infinite, making volatility a useless and misleading estimate of risk. High H values, meaning less noise, more persistence, and clearer trends than lower values of H, we can assume that higher values of H mean less risk. However, stocks with high H values do have a higher risk of abrupt changes. The fractal nature of the capital markets contradicts the EMH and all the quantitative models derived from it, such as the Capital Asset Pricing Model (CAPM), the Arbitrage Pricing Theory (APT), and the Black-Scholes option pricing model, and other models depending on the normal distribution and/or finite variance. This is because they simplify reality by assuming random behaviour, and they ignore the influence of time on decision making. By assuming randomness, the models can be optimised for a single optimal solution. That is, we can find optimal portfolios, intrinsic value, and fair price. On the other hand, fractal structure recognises complexity and provides cycles, trends, and a range of fair values.

New theories to explain financial markets are gaining ground, among which is a multitude of interacting agents forming a complex system characterised by a high level of uncertainty. Complexity theory deals with processes where a large number of seemingly independent agents act coherently. Multiple interacting agent systems are subject to contagion and propagation phenomena generating feedbacks and producing fat tails. Real feedback systems involve long-term correlations and trends since memories of long-past events can still affect the decisions made in the present. Most complex, natural systems, can be modelled by nonlinear differential, or difference, equations. These systems are characterised by a high level of uncertainty which is embedded in the probabilistic structure of models. As a result, econometrics can now supply the empirical foundation of economics. For instance, science being highly stratified, one can build complex theories on the foundation of simpler theories. That is, starting with a collection of econometric data, we model it and analyse it, obtaining statistical facts of an empirical nature that provide us with the building blocks of future theoretical development. For instance, assuming that economic agents are heterogeneous, make mistakes, and mutually interact leads to more freedom to devise economic theory (see Aoki [2004]).

With the growing quantity of data available, machine-learning methods that have been successfully applied in science are now applied to mining the markets. Data mining and more recent machine-learning methodologies provide a range of general techniques for the classification, prediction, and optimisation of structured and unstructured data. Neural networks, classification and decision trees, k-nearest neighbour methods, and support vector machines (SVM) are some of the more common classification and prediction techniques used in machine learning. Further, combinatorial optimisation, genetic algorithms and reinforced learning are now widespread. Using these techniques, one can describe financial markets through degrees of freedom which may be both qualitative and quantitative in nature, each node being the siege of complicated mathematical entity. One could use a matrix form to represent interactions

between the various degrees of freedom of the different nodes, each link having a weight and a direction. Further, time delays should be taken into account, leading to non-symmetric matrix (see Ausloos [2010]).

Future success for portfolio managers will not only depend on their ability to provide excess returns in a risk-controlled fashion to investors, but also on their ability to incorporate financial innovation and process automation into their frameworks. However, the quantitative approach is not without risk, introducing new sources of risk such as model risk, operational risk, and an inescapable dependence on historical data as its raw material. One must therefore be cautious on how the models are used, understand their weaknesses and limitations, and prevent applications beyond what they were originally designed for. With more model parameters and more sophisticated econometric techniques, we run the risk of over-fitting models, and distinguishing spurious phenomena as a result of data mining becomes a difficult task.

In the rest of this guide we will present an overview of asset valuation in presence of risk and we will review the evolution of quantitative methods. We will then present the necessary tools and techniques to design the main steps of an automated investment management system, and we will address some of the challenges that need to be met.

Part I

Quantitative trading in classical economics

Chapter 1

Risk, preference, and valuation

1.1 A brief history of ideas

Pacioli [1494] as well as Pascal and Fermat (1654) considered the problem of the points, where a game of dice has to be abandoned before it can be concluded, and how is the pot (the total wager) distributed among the players in a fair manner, introducing the concept of fairness (see Devlin [2008] for historical details). Pascal and Fermat agreed that the fair solution is to give to each player the expectation value of his winnings. The expectation value they computed is an ensemble average, where all possible outcomes of the game are enumerated, and the products of winnings and probabilities associated with each outcome for each player are added up. Instead of considering only the state of the universe as it is, or will be, an infinity of additional equally probable universes is imagined. The proportion of those universes where some event occurs is the probability of that event. Following Pascal's and Fermat's work, others recognised the potential of their investigation for making predictions. For instance, Halley [1693] devised a method for pricing life annuities. Huygens [1657] is credited with making the concept of expectation values explicit and with first proposing an axiomatic form of probability theory. A proven result in probability theory follows from the axioms of probability theory, now usually those of Kolmogorov [1933].

Once the concept of probability and expectation values was introduced by Pascal and Fermat, the St Petersburg paradox was the first well-documented example of a situation where the use of ensembles leads to absurd conclusions. The St Petersburg paradox rests on the apparent contradiction between a positively infinite expectation value of winnings in a game and real people's unwillingness to pay much to be allowed to participate in the game. Bernoulli [1738-1954] (G. Cramer 1728, personal communication with N. Bernoulli) pointed out that because of this incongruence, the expectation value of net winnings has to be discarded as a descriptive or prescriptive behavioural rule. As pointed out by Peters [2011a], one can decide what to change about the expectation value of net winnings, either the expectation value or the net winnings. Bernoulli (and Cramer) chose to replace the net winnings by introducing utility, and computing the expectation value of the gain in utility. They argued that the desirability or utility associated with a financial gain depends not only on the gain itself but also on the wealth of the person who is making this gain. The expected utility theory (EUT) deals with the analysis of choices among risky projects with multidimensional outcomes. The classical resolution is to apply a utility function to the wealth, which reflects the notion that the usefulness of an amount of money depends on how much of it one already has, and then to maximise the expectation of this. The choice of utility function is often framed in terms of the individual's risk preferences and may vary between individuals. The first important use of the EUT was that of Von Neumann and Morgenstern (VNM) [1944] who used the assumption of expected utility maximisation in their formulation of game theory. When comparing objects one needs to rank utilities but also compare the sizes of utilities. VNM method of comparison involves considering probabilities. If a person can choose between various randomised events (lotteries), then it is possible to additively compare for example a shirt and a sandwich. Later, Kelly [1956], who contributed to the debate on time averages, computed time-average exponential growth rates in games of chance (optimise wager sizes in a hypothetical horse race using private information) and

argued that utility was not necessary and too general to shed any light on the specific problems he considered. In the same spirit, Peters [2011a] considered an alternative to Bernoulli's approach by replacing the expectation value (or ensemble average) with a time average, without introducing utility.

It is argued that Kelly [1956] is at the origin of the growth optimal portfolio (GOP), when he studied gambling and information theory and stated that there is an optimal gambling strategy that will accumulate more wealth than any other different strategy. This strategy is the growth optimal strategy. We refer the reader to Mosegaard Christensen [2011] who presented a comprehensive review of the different connections in which the GOP has been applied. Since one aspect of the GOP is the maximisation of the geometric mean, one can go back to Williams [1936] who considered speculators in a multi-period setting and reached the conclusion that due to compounding, speculators should worry about the geometric mean and not the arithmetic one. One can further go back in time by recognising that the GOP is the choice of a log-utility investor, which was first discussed by Bernoulli and Cramer in the St. Petersurg paradox. However, it was argued (leading to debate among economists) that the choice of the logarithm appears to have nothing to do with the growth properties of the strategy (Cramer solved the paradox with a square-root function). Nonetheless, the St. Petersurg paradox inspired Latane [1959], who independently from Kelly, suggested that investors should maximise the geometric mean of their portfolios, as this would maximise the probability that the portfolio would be more valuable than any other portfolio. It was recently proved that when denominated in terms of the GOP, asset prices become supermartingales, leading Long [1990] to consider change of numeraire and suggest a method for measuring abnormal returns. The change of numeraire technique was then used for derivative pricing. No matter the approach chosen, the perspective described by Bernoulli and Cramer has consequences far beyond the St Petersburg paradox, including predictions and investment decisions, as in this case, the conceptual context change from moral to predictive. In the latter, one can assume that the expected gain (or growth factor or exponential growth rate) is the relevant quantity for an individual deciding whether to take part in the lottery. However, considering the growth of an investment over time can make this assumption somersault into being trivially false.

In order to explain the prices of economical goods, Walras [1874-7] started the theory of general equilibrium by considering demand and supply and equating them, which was formalised later by Arrow-Debreu [1954] and McKenzie [1959]. In parallel Arrow [1953] and then Debreu [1953] generalised the theory, which was static and deterministic, to the case of uncertain future by introducing contingent prices (Arrow-Debreu state-prices). Arrow [1953] proposed to create financial markets, and was at the origin of the modern theory of financial markets equilibrium. This theory was developed to value asset prices and define market equilibium. Radner [1976] improved Arrow's model by considering more general assets and introducing the concept of rational anticipation. Radner is also at the origin of the incomplete market theory. Defining an arbitrage as a transaction involving no negative cash flow at any probabilistic or temporal state, and a positive cash flow in at least one state (that is, the possibility of a risk-free profit after transaction costs), the prices are said to constitute an arbitrage equilibrium if they do not allow for profitable arbitrage (see Ross [1976]). An arbitrage equilibrium is a precondition for a general economic equilibrium (see Harrison et al. [1979]). In complete markets, no arbitrage implies the existence of positive Arrow-Debreu state-prices, a risk-neutral measure under which the expected return on any asset is the risk-free rate, and equivalently, the existence of a strictly positive pricing kernel that can be used to price all assets by taking the expectation of their payoffs weighted by the kernel (see Ross [2005]).

More recently, the concepts of EUT have been adapted for derivative security (contingent claim) pricing (see Hodges et al. [1989]). In a financial market, given an investor receiving a particular contingent claim offering payoff C_T at future time $T > 0$ and assuming market completeness, then the price the investor would pay can be found uniquely. Option pricing in complete markets uses the idea of replication whereby a portfolio in stocks and bonds recreates the terminal payoff of the option, thus removing all risk and uncertainty. However, in reality, most situations are incomplete as market frictions, transactions costs, non-traded assets and portfolio constraints make perfect replication impossible. The price is no-longer unique and several potential approaches exist, including utility indifference pricing (UIP), superreplication, the selection of one particular measure according to a minimal distance criteria (for example the minimal martingale measure or the minimal entropy measure) and convex risk measures. The UIP will be of

particular interest to us in the rest of this book (see Henderson et al. [2004] for an overview). In that setting, the investor can maximise expected utility of wealth and may be able to reduce the risk due to the uncertain payoff through dynamic trading. As explained by Hodges et al. [1989], the investor is willing to pay a certain amount today for the right to receive the claim such that he is no worse off in expected utility terms than he would have been without the claim. Some of the advantages of UIP include its economic justification, incorporation of wealth dependence, and incorporation of risk aversion, leading to a non-linear price in the number of units of claim, which is in contrast to prices in complete markets.

1.2 Solving the St. Petersburg paradox

1.2.1 The simple St. Petersburg game

The expected utility theory (EUT) deals with the analysis of choices among risky projects with (possibly multidimensional) outcomes. The expected utility model was first proposed by Nicholas Bernoulli in 1713 and solved by Daniel Bernoulli [1738-1954] in 1738 as the St. Petersburg paradox. A casino offers a game of chance for a single player in which a fair coin is tossed at each stage. The pot starts at \$1 and is doubled every time a head appears. The first time a tail appears, the game ends and the player wins whatever is in the pot. The player wins the payout $D_k = \$2^{k-1}$ where k heads are tossed before the first tail appears. That is, the random number of coin tosses, k, follows a geometric distribution with parameter $\frac{1}{2}$, and the payouts increase exponentially with k. The question being on the fair price to pay the casino for entering the game. Following Pascal and Fermat, one answer is to consider the average payout (expected value)

$$E[D_k] =< D_k >= \sum_{k=1}^{\infty}(\frac{1}{2})^k 2^{k-1} = \frac{1}{2}1 + \frac{1}{4}2 + ... = \sum_{k=1}^{\infty}\frac{1}{2} = \infty$$

A gamble is worth taking if the expectation value of the net change of wealth, $< D_k > -c$ where c is the cost charged to enter the game, is positive. Assuming infinite time and unlimited resources, this sum grows without bound and so the expected win for repeated play is an infinite amount of money. Hence, considering nothing but the expectation value of the net change in one's monetary wealth, one should therefore play the game at any price if offered the opportunity, but people are ready to pay only a few dollars. The paradox is the discrepancy between what people seem willing to pay to enter the game and the infinite expected value. Instead of computing the expectation value of the monetary winnings, Bernoulli [1738-1954] proposed to compute the expectation value of the gain in utility. He argued that the paradox could be resolved if decision-makers displayed risk aversion and argued for a logarithmic cardinal utility function $u(w) = \ln w$ where w is the gambler's total initial wealth. It was based on the intuition that the increase in wealth should correspond to an increase in utility which is inversely proportional to the wealth a person already has, that is, $\frac{du}{dx} = \frac{1}{x}$, whose solution is the logarithm. The expected utility hypothesis posits that a utility function exists whose expected net change is a good criterion for real people's behaviour. For each possible event, the change in utility will be weighted by the probability of that event occurring. Letting c be the cost charged to enter the game, the expected net change in logarithmic utility is

$$E[\Delta u] =< \Delta u >= \sum_{k=1}^{\infty}\frac{1}{2^k}\left(\ln\left(w + 2^{k-1} - c\right) - \ln w\right) < \infty \tag{1.2.1}$$

where $w + 2^{k-1} - c$ is the wealth after the event, converges to a finite value. This formula gives an implicit relationship between the gambler's wealth and how much he should be willing to pay to play (specifically, any c that gives a positive expected utility). However, this solution by Cramer and Bernoulli is not completely satisfying, since the lottery can easily be changed in a way such that the paradox reappears. For instance, we just need to change the game so that it gives the (even larger) payoff e^{2^k}. More generally, it is argued that one can find a lottery that allows for a variant

of the St. Petersburg paradox for every unbounded utility function (see Menger [1934]). But, this conclusion was shown by Peters [2011b] to be incorrect. Nicolas Bernoulli himself proposed an alternative idea for solving the paradox, conjecturing that people will neglect unlikely events, since only unlikely events yield the high prizes leading to an infinite expected value. The idea of probability weighting resurfaced much later in the work of Kahneman et al. [1979], but their experiments indicated that, very much to the contrary, people tend to overweight small probability events. Alternatively, relaxing the unrealistic assumption of infinite resources for the casino, and assuming that the expected value of the lottery only grows logarithmically with the resources of the casino, one can show that the expected value of the lottery is quite modest.

1.2.2 The sequential St. Petersburg game

As a way of illustrating the GOP (presented in Section (2.3.2)) for constantly rebalanced portfolio, Gyorfi et al. [2009] [2011] introduced the sequential St. Petersburg game which is a multi-period game having exponential growth. Before presenting that game we first discuss an alternative version (called iterated St. Petersburg game) where in each round the player invest $C_A = \$1$, and let X_n denotes the payoff for the n-th simple game. Assuming the sequence $\{X_n\}_{n=1}^{\infty}$ to be independent and identically distributed, after n rounds the player's wealth in the repeated game becomes

$$C_A(n) = \sum_{i=1}^{n} X_i$$

so that in the limit we get

$$\lim_{n\to\infty} \frac{C_A(n)}{n \log_2 n} = 1$$

in probability, where $\log_2$ denotes the logarithm with base 2. We can now introduce the sequential St. Petersburg game. Starting with initial capital $C_A(0) = \$1$ and assuming an independent sequence of simple St. Petersburg games, for each simple game the player reinvest his capital. If $C_A^c(n-1)$ is the capital after the $(n-1)$-th simple game, then the invested capital is $C_A^c(n-1)(1-f_c)$, while $C_A^c(n-1)f_c$ is the proportional cost of the simple game with commission factor $0 < f_c < 1$. Hence, after the n-th round the capital is

$$C_A^c(n) = C_A(n-1)^{fc}(1-f_c)X_n = C_A(0)(1-f_c)^n \prod_{i=1}^{n} X_i = (1-f_c)^n \prod_{i=1}^{n} X_i$$

Given its multiplicative definition, $C_A^c(n)$ has exponential trend

$$C_A^c(n) = e^{nW_n^c} \approx e^{nW^c} \tag{1.2.2}$$

with average growth rate

$$W_n^c = \frac{1}{n} \ln C_A^c(n)$$

and with asymptotic average growth rate

$$W^c = \lim_{n\to\infty} \frac{1}{n} \ln C_A^c(n)$$

From the definition of the average growth rate, we get

$$W_n^c = \frac{1}{n}\left(n \ln(1-f_c) + \sum_{i=1}^{n} \ln X_i\right)$$

and applying the strong law of large numbers, we obtain the asymptotic average growth rate

$$W^c = \ln(1 - f_c) + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln X_i = \ln(1 - f_c) + E[\ln X_1] \ a.s.$$

so that W^c can be calculated via expected log-utility. The commission factor f_c is called fair if

$$W^c = 0$$

so that the growth rate of the sequential game is 0. We can calculate the fair factor f_c as

$$\ln(1 - f_c) = -E[\ln X_1] = -\sum_{k=1}^{\infty} k \ln 2 \frac{1}{2^k} = -2 \ln 2$$

and we get

$$f_c = \frac{3}{4}$$

Note, Gyorfi et al. [2009] studied the portfolio game, where a fraction of the capital is invested in the simple fair St. Petersburg game and the rest is kept in cash.

1.2.3 Using time averages

Peters [2011a] used the notion of ergodicity in stochastic systems, where it is meaningless to assign a probability to a single event, as the event has to be embedded within other similar events. While Fermat and Pascal chose to embed events within parallel universes, alternatively we can embed them within time, as the consequences of the decision will unfold over time (the dynamics of a single system are averaged along a time trajectory). However, the system under investigation, a mathematical representation of the dynamics of wealth of an individual, is not ergodic, and that this manifests itself as a difference between the ensemble average and the time average of the growth rate of wealth. The origins of ergodic theory lie in the mechanics of gases (large-scale effects of the molecular dynamics) where the key rationale is that the systems considered are in equilibrium. It is permissible under strict conditions of stationarity (see Grimmet et al. [1992]). While the literature on ergodic systems is concerned with deterministic dynamics, the basic question whether time averages may be replaced by ensemble averages is equally applicable to stochastic systems, such as Langevin equations or lotteries. The essence of ergodicity is the question whether the system when observed for a sufficiently long time t samples all states in its sample space in such a way that the relative frequencies $f(x,t)dx$ with which they are observed approach a unique (independent of initial conditions) probability $P(x)dx$

$$\lim_{t\to\infty} f(x,t) = P(x)$$

If this distribution does not exist or is not unique, then the time average $\overline{A} = \lim_{T\to\infty} \frac{1}{T} \int_0^T A(x(t))dt$ of an observable A can not be computed as an ensemble average in Huygens' sense, $< A >= \int_x A(x)P(x)dx$. Peters [2011a] pointed out that computing the naive expected payout is mathematically equivalent to considering multiple outcomes of the same lottery in parallel universes. It is therefore unclear why expected wealth should be a quantity whose maximization should lead to a sound decision theory. Indeed, the St. Petersburg paradox is only a paradox if one accepts the premise that rational actors seek to maximize their expected wealth. The choice of utility function is often framed in terms of the individual's risk preferences and may vary between individuals. An alternative premise, which is less arbitrary and makes fewer assumptions, is that the performance over time of an investment better characterises an investor's prospects and, therefore, better informs his investment decision. To compute ensemble averages, only a probability distribution is required, whereas time averages require a dynamic, implying an additional assumption. This assumption corresponds to the multiplicative nature of wealth accumulation. That is, any wealth gained can itself be employed to generate further wealth, which leads to exponential growth (banks and governments offer exponentially growing interest payments on savings). The accumulation of wealth over time is well characterized by an exponential

growth rate, see Equation (1.2.2). To compute this, we consider the factor r_k by which a player's wealth changes in one round of the lottery (one sequence of coin tosses until a tails-event occurs)

$$r_k = \frac{w - c + D_k}{w}$$

where D_k is the kth (positive finite) payout. Note, this factor corresponds to the payoff X_k for the k-th simple game described in Section (1.2.2). To convert this factor into an exponential growth rate g (so that e^{gt} is the factor by which wealth changes in t rounds of the lottery), we take the logarithm $g_k = \ln r_k$. The ensemble-average growth factor is

$$< r >= \sum_{k=1}^{\infty} p_k r_k$$

where p_k is the (non-zero) probability. The logarithm of $< r >$ expresses this as the ensemble-average exponential growth rate, that is $< g >= \ln < r >$, and we get

$$< g >= \ln \sum_{k=1}^{\infty} p_k r_k$$

Note, the exponential growth rate $< g >$ should be rescaled by $\frac{1}{t}$ to be consistent with the factor e^{gt}. Further, this rate corresponds to the rate of an M-market. The main idea in the time averages is to consider the rate of an ergotic process and to average over time. In this case, the passage of time is incorporated by identifying as the quantity of interest the Average Rate of Exponential Growth of the player's wealth in a single round of the lottery. Repeating the simple game in sequences, the time-average growth factor $\overline{r}$ is

$$\overline{r} = \prod_{k=1}^{\infty} r_k^{p_k}$$

corresponding to the player's wealth $C_A^c(\infty)$. The logarithm of $\overline{r}$ expresses this as the time-average exponential growth rate, that is, $\overline{g} = \ln \overline{r}$. Hence, the time-average exponential growth rate is

$$\overline{g}(w, c) = \sum_{k=1}^{\infty} p_k \ln \left(\frac{w - c + D_k}{w}\right)$$

where p_k is the (non-zero) probability of receiving it. In the standard St. Petersburg lottery, $D_k = 2^{k-1}$ and $p_k = \frac{1}{2^k}$. Note, given Jensen's inequality (see Appendix (A.1.1)), when f is concave (here, the logarithm function), we get

$$< g >\geq \overline{g}(w, c)$$

Although the rate $\overline{g}(w, c)$ is an expectation value of a Growth Rate r_k (the time unit being one lottery game), and may therefore be thought of in one sense as an average over parallel universes, it is in fact equivalent to the time average Growth Rate that would be obtained if repeated lotteries were played over time. This is because the logarithm function, taken as a utility function, has the special property of encoding the multiplicative nature common to gambling and investing in a linear additive object. The expectation value is

$$\sum_{k=1}^{\infty} p_k \ln r_k = \ln \left(\lim_{T \to \infty} \left(\prod_{i=1}^{T} r_i\right)^{\frac{1}{T}}\right)$$

which is the geometric average return (for details see Section (2.3.2)). It is reasonable to assume that the intuition behind the human behaviour is a result of making repeated decisions and considering repeated games. While $\overline{g}(.,.)$ is identical to the rate of change of the expected logarithmic utility in Equation (1.2.1), it has been obtained without making any assumptions about the player's risk preferences or behaviour, other than that he is interested in the Rate of Growth of his wealth. Under this paradigm, an individual with wealth w should buy a ticket at a price c provided $\overline{g}(w, c) > 0$. Note, this equation can also be considered a criterion for how much risk a person should take.

1.2.4 Using option pricing theory

Bachelier [1900] asserted that every price follows a martingale stochastic process, leading to the notion of perfect market. One of the fundamental concept in the mathematical theory of financial markets is the no-arbitrage condition. The fundamental theorem of asset pricing states that in an arbitrage free market model there exists a probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ such that every discounted price process $\overline{X}$ is a martingale under $\mathbb{Q}$, and $\mathbb{Q}$ is equivalent to $\mathbb{P}$. Using the notion of Arrow-Debreu state-price density from economics, Harrison et al. [1979] showed that the absence of arbitrage implies the existence of a density or pricing kernel, also called stochastic discount factor, that prices all asset. We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}_t$ is a right continuous filtration including all $\mathbb{P}$ negligible sets in $\mathcal{F}$. Given the payoff C_T at maturity T, the prices π_t seen at time t can be calculated as the expectation under the physical measure

$$\pi_t = E[\xi_T C_T | \mathcal{F}_t] \tag{1.2.3}$$

where ξ_T is the state-price density at time T which depend on the market price of risk λ. The pricing kernel measures the degree of risk aversion in the market, and serves as a benchmark for preferences. In the special case where the interest rates and the market price of risk are null, one can easily compute the price of a contingent claim as the expected value of the terminal flux. These conditions are satisfied in the M-market (see Remark (F.1.2) in Appendix (F.1)) also called the market numeraire and introduced by Long [1990]. A common approach to option pricing in complete markets, in the mathematical financial literature, is to fix a measure $\mathbb{Q}$ under which the discounted traded assets are martingales and to calculate option prices via expectation under this measure (see Harrison et al. [1981]). The option pricing theory (OPT) states that when the market is complete, and market price of risk and rates are bounded, then there exists a unique risk-neutral probability $\mathbb{Q}$, and the risk-neutral rule of valuation can be applied to all contingent claims which are square integrable (see Theorem (F.1.5)). Risk-neutral probabilities are the product of an unknown kernel (risk aversion) and natural probabilities. While in a complete market there is just one martingale measure or state-price density, there are an infinity of state-price densities in an incomplete market (see Cont et al. [2003] for a description of incomplete market theory). In the utility indifference pricing (UIP) theory (see an introduction to IUP in Appendix (F.5)), assuming the investor initially has wealth w, the value function (see Equation (F.5.16)) is given by

$$V(w,k) = \sup_{W_T \in \mathcal{A}(w)} E[u(W_T + kC_T)]$$

with $k > 0$ units of the claim, and where the supremum is taken over all wealth W_T which can be generated from initial fortune w. The utility indifference buy price $\pi_b(k)$ (see Equation (F.5.18)) is the solution to

$$V(w - \pi_b(k), k) = V(w, 0)$$

which involve solving two stochastic control problems (see Merton [1969] [1971]). An alternative solution is to convert this primal problem into the dual problem which involves minimising over state-price densities or martingale measures (see Equation (F.5.19)) (a simple example is given in Appendix (E.5)). A consequence of the dual problem is that the market price of risk plays a fundamental role in the characterisation of the solution to the utility indifference pricing problem (see Remark (F.5.1)).

Clearly the St Petersburg paradox is neither an example of a complete market situation nor one of an incomplete market situation, since the payout grows without bound, making the payoff not square integrable. Further, the expectation value of net winnings proposed by Pascal and Fermat implicitly assume the situation of an M-market, corresponding to a market with null rates and null market price of risk. As pointed out by Huygens [1657], this concept of expectation is agnostic regarding fluctuations, which is harmless only if the consequences of the fluctuations, such as associated risks, are negligible. The ability to bear risk depends not only on the risk but also on the risk-bearer's resources. Similarly, Bernoulli [1738-1954] noted "if I am not wrong then it seems clear that all men can not use the

same rule to evaluate the gamble". That is, in the M-market investors are risk-neutral which does not correspond to a real market situation, and one must incorporate rates and market price of risk in the pricing of a claim.

Rather than explicitly introducing the market price of risk, Bernoulli and Cramer proposed to compute the expectation value of the gain in some function of wealth (utility). It leads to solving Equation (1.2.1) which is to be compared with Equation (F.5.16) in the IUP discussed above. If the utility function is properly defined, it has the advantage of bounding the claim so that a solution can be found. Clearly, the notion of time is ignored, and there is no supremum taken over all wealth generated from initial fortune w. Note, the lottery is played only once with wealth w and one round of the game is assumed to be very fast (instantaneous). Further, there is no mention of repetitions of the game in N. Bernoulli's letter, only human behaviour. It is assumed that the gambler's wealth is only modified by the outcome of the game, so that his wealth after the event becomes $w + 2^{k-1} - c$ where c is the ticket price. Hence, the absence of supremum in the ensemble average. As a result, the ensemble average on gain by Pascal and Fermat has been replaced by an ensemble average on a function of wealth (bounding the claim), but no notion of time and market price of risk has been considered. As discussed by Peters [2011a], utility functions (in Bernoulli's framework) are externally provided to represent risk preferences, but are unable by construction to recommend appropriate levels of risk. A quantity that is more directly relevant to the financial well-being of an individual is the growth of an investment over time. In UIP and time averages, any wealth gained can itself be employed to generate further wealth, leading to exponential growth. By proposing a time average, Peters introduced wealth optimisation over time, but he had to assume something about wealth W in the future. In the present situation, similarly to the sequential St. Petersburg game discussed in Section (1.2.2), he based his results on the assumption that equivalent lotteries can be played in sequence as often as desired, implying that irrespective of how close a player gets to bankruptcy, losses will be recovered over time. To summarise, UIP and time averages are meaningless in the absence of time (here sequences of equivalent rounds).

1.3 Modelling future cashflows in presence of risk

The rationally oriented academic literature still considered the pricing equation (1.3.4), but in a world of uncertainty and time-varying expected returns (see Arrow [1953] and Debreu [1953]). That is, the discount rate $(\mu_t)_{t\geq 0}$ is now a stochastic process, leading to the notion of stochastic discount factor (SDF). As examined in Section (1.1), a vast literature discussed the various ways of valuing asset prices and defining market equilibrium. In view of introducing the main ideas and concepts, we present in Appendix (E) some simple models in discrete time with one or two time periods and with a finite number of states of the world. We then consider in Appendix (F) more complex models in continuous time and discuss the valuation of portfolios on multi-underlyings where we express the dynamics of a self-financing portfolio in terms of the rate of return and volatility of each asset. As a consequence of the concept of absence of arbitrage opportunity (AAO) (see Ross [1976]), the idea that there exists a constraint on the rate of return of financial assets developed, leading to the presence of a market price of risk λ_t which is characterised in Equation (F.1.2). Further, in a complete market with no-arbitrage opportunity, the price of a contingent claim is equal to the expected value of the terminal flux expressed in the cash numeraire, under the risk-neutral probability $\mathbb{Q}$ (see details in Appendix (E) and Appendix (F) and especially Theorem (F.1.5)).

1.3.1 Introducing the discount rate

We saw in Section (1.2.4) that at the core of finance is the present value relation stating that the market price of an asset should equal its expected discounted cash flows under the right probability measure (see Harrison et al. [1979], Dana et al. [1994]). The question being: How to define the pricing Kernel? or equivalently, how to define the martingale measures? We let π_t be the price at time t of an asset with random cash flows $F_k = F(T_k)$ at the time T_k for $k = 1, .., N$ such that $0 < T_1 < T_2 < ... < T_N$. Note, N can possibly go to infinity. Given the price of an asset in Equation (1.2.3) and assuming $\xi_{T_k} = e^{-\mu(T_k - t)}$, the present value of the asset at time t is given by

$$\pi_t = E[\sum_{k=1}^{N} e^{-\mu(T_k - t)} F_k | \mathcal{F}_t] \tag{1.3.4}$$

where μ is the discount rate, and such that the most common cash flows F_k can be coupons and principal payments of bonds, or dividends of equities. The main question becomes: What discount rate to use? As started by Walras [1874-7], equilibrium market prices are set by supply and demand among investors applying the pricing Equation (1.3.4), so that the discount rates they require for holding assets are the expected returns they can rationally anticipate. Hence, the discount rate or the expected return contains both compensation for time and compensation for risk bearing. Williams [1938] discussed the effects of risk on valuation and argued that "the customary way to find the value of a risky security has been to add a premium for risk to the pure rate of interest, and then use the sum as the interest rate for discounting future receipts". The expected excess return of a given asset over the risk-free rate is called the expected risk premium of that asset. In equity, the risk premium is the growth rate of earnings, plus the dividend yield, minus the riskless rate. Since all these variables are dynamic, so must be the risk premium. Further, the introduction of stock and bond option markets confirmed that implied volatilities vary over time, reinforcing the view that expected returns vary over time. Fama et al. [1989] documented counter-cyclical patterns in expected returns for both stocks and bonds, in line with required risk premia being high in bad times, such as cyclical troughs or financial crises. Understanding the various risk premia is at the heart of finance, and the capital asset pricing model (CAPM) as well as the option pricing theory (OPT) are two possible answers among others. To proceed, one must find a way to observe risk premia in view of analysing them and possibly forecasting them. Through out this guide we are going to describe the tools and techniques used by institutional asset holders to estimate the risk premia via historical data as well as the approach used by option's traders to implicitly infer the risk premia from option prices.

1.3.2 Valuing payoffs in continuous time

A consequence of the fundamental theorem of asset pricing introduced in Section (1.2.4) is that the price process S need to be a semimartingale under the original measure $\mathbb{P}$. In this section we give a bief introduction to some fundamental finance concepts. While some more general models might include discontinuous semimartingales and even stochastic processes which are not semimartingales (for example fractional Brownian motion), we consider the continuous decomposable semimartingales models for equity securities

$$S(t) = A(t) + M(t)$$

where the drift A is a finite variation process and the volatility M is a local martingale. For simplicity of exposition we focus on an $\mathcal{F}_t$-adapted market consisting of the $(N+1)$ multi-underlying diffusion model described in Appendix (F.1) with $0 \leq t \leq T$ and dynamics given by

$$\frac{dS_t}{S_t} = b_t dt + < \sigma_t, \hat{W}_t > \tag{1.3.5}$$

where the instantaneous rate of return b_t is an adapted vector in $\mathbb{R}^N$, $\hat{W}_t$ is a k-dimensional Brownian motion with components $\hat{W}_t^j$, and σ_t is a $N \times k$ adapted volatility matrix with elements $\sigma_j^i(t)$. The market $\{S_t\}_{t\in[0,T]}$ is called normalised if $S_t^0 = 1$. We can always normalise the market by defining

$$\overline{S}_t^i = \frac{S_t^i}{S_t^0}, 1 \leq i \leq N$$

so that

$$S_t = \left(1, \overline{S}_t^1, ..., \overline{S}_t^N\right)$$

is the normalisation of S_t. Hence, it corresponds to regarding the price S_t^t of the safe investment as the unit of price (the numeraire) and computing the other prices in terms of this unit. We define the riskless asset or accumulation

factor $B(t)$ as the value at time t of a fund created by investing \$1 at time 0 on the money market and continuously reinvested at the instantaneous interest rate $r(t)$. Assuming that for almost all ω, $t \to r_t(\omega)$ is strictly positive and continuous, and r_t is an $\mathcal{F}_t$ measurable process, then the riskless asset is given by

$$B(t) = S_t^0 = e^{\int_0^t r(s)ds} \tag{1.3.6}$$

We can now introduce the notion of arbitrage in the market.

Lemma 1.3.1 *Suppose there exists a measure Q on $\mathcal{F}_T$ such that $Q \sim P$ and such that the normalised price process $\{\overline{S}_t\}_{t\in[0,T]}$ is a martingale with respect to Q. Then the market $\{S_t\}_{t\in[0,T]}$ has no arbitrage.*

Definition 1.3.1 *A measure $Q \sim P$ such that the normalised process $\{\overline{S}_t\}_{t\in[0,T]}$ is a martingale with respect to Q is called an equivalent martingale measure.*

That is, if there exists an equivalent martingale measure then the market has no arbitrage. In that setting the market also satisfies the stronger condition called no free lunch with vanishing risk (NFLVR) (see Delbaen et al. [1994]). We can consider a weaker result. We let Q be an equivalent martingale measure with $dQ = \xi dP$, such that ξ is strictly positive and square integrable. Further, the process $\{\xi_t\}_{0\leq t\leq T}$ defined by $\xi_t = E[\xi|\mathcal{F}_t]$ is a strictly positive martingale over the Brownian fields $\{\mathcal{F}_t\}$ with $\xi_T = \xi$ and $E[\xi_t] = E[\xi] = 1$ for all t. Given ϕ an arbitrary bounded, real-valued function on the real line, Harrison et al. [1979] showed that the Radon-Nikodym derivative is given by

$$\xi_t = \frac{dQ}{dP}\Big|_{\mathcal{F}_t} = e^{-\frac{1}{2}\int_0^t \phi^2(W_s)ds + \int_0^t \phi(W_s)d\tilde{W}(s)}, 0 \leq t \leq T$$

where $\tilde{W}$ is a Brownian motion on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Note, $\lambda_t = \phi(\tilde{W}_t)$ is the market price of risk such that $(\lambda_t)_{0\leq t\leq T}$ is a bounded adapted process. Then, ξ is a positive martingale under $\mathbb{P}$ and one can define the new probability measure $\mathbb{Q}$ for arbitrary $t > 0$ by

$$Q(A) = E[\xi_t I_A]$$

where I_A is the indicator function for the event $A \in \mathcal{F}_t$. Moreover, using the theorem of Girsanov [1960], the Brownian motion

$$W(t) = \tilde{W}(t) + \int_0^t \lambda_s ds$$

is a Brownian motion on the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ (see Appendix (F.2)). Using the Girsanov transformation (see Girsanov [1960]), we choose λ_t in a clever way, and verify that each discounted price process is a martingale under the probability measure $\mathbb{Q}$. For instance, in the special case where we assume that the stock prices are lognormally distributed with drift μ and volatility σ, the market price of risk is given by Equation ((F.2.4)) (see details in Appendix (F.2.1)). Hence, we see that modelling the rate of return of risky asset is about modelling the market price of risk λ_t.

Assuming a viable Ito market, Theorem (F.1.1) applies, that is, there exists an adapted random vector λ_t and an equivalent martingale measure $\mathbb{Q}$ such that the instantaneous rate of returns b_t of the risky assets satisfies Equation (F.1.2), that is

$$b_t = r_t I + \sigma_t \lambda_t \, , d\mathbb{P} \times dt \text{ a.s.}$$

***Remark* 1.3.1** *As a result of the absence of arbitrage opportunity, there is a constraint on the rate of return of financial assets. The riskier the asset, the higher the return, to justify its presence in the portfolio.*

Hence, in absence of arbitrage opportunity, the multi-underlyings model has dynamics

$$\frac{dS_t^i}{S_t^i} = r_t dt + \sum_{j=1}^{k} \sigma_j^i(t)\lambda_t^i dt + \sum_{j=1}^{k} \sigma_j^i(t)\hat{W}_t^j$$

and we see that the normalised market $\overline{S}_t$ is a Q-martingale, and the conclusion of no-arbitrage follows from Lemma (1.3.1). Geman et al. [1995] proved that many other probability measures can be defined in a similar way, which reveal themselves to be very useful in complex option pricing. We can now price claims with future cash flows. Given X_T an $\mathcal{F}_T$-measurable random variable, and assuming $\pi_T(H) = X$ for some self-financing H, then by the Law of One Price, $\pi_t(H)$ is the price at time t of X defined as

$$\pi_t(H) = B_t E^Q[\frac{\pi_T}{B_T}|\mathcal{F}_t] = B_t E^Q[\frac{X}{B_T}|\mathcal{F}_t]$$

where $B_t = e^{rt}$ is the riskless asset, and $\frac{\pi(H)}{B}$ is a martingale under $\mathbb{Q}$ (see Harrison et al. [1981] [1983]). As a result, the market price π_t in Equation (1.3.4) becomes

$$\pi_t = E^Q[\sum_{k=1}^{N} e^{-\int_t^{T_k} r_s ds} F_k] = \sum_{k=1}^{N} \hat{F}_k P(t, T_k)$$

where $(r_t)_{0\leq t\leq T}$ is the risk-free rate (possibly stochastic), $P(t, T_k)$ is the discount factor, and

$$\hat{F}_k(t) = \frac{E_t^Q[e^{-\int_t^{T_k} r_s ds} F_k]}{P(t, T_k)}$$

is the expected cash flows at time t for the maturities T_k for $k = 1, .., N$.

1.3.3 Modelling the discount factor

We let the zero-coupon bond price $P(t, T)$ be the price at time t of \$1 paid at maturity. Given the pair $(r_t, \lambda_t)_{t\geq 0}$ of bounded adapted stochastic processes, the price of a zero-coupon bond in the period $[t, T]$ and under the risk-neutral probability measure $\mathbb{Q}$ is given by

$$P(t, T) = E^Q[e^{-\int_t^T r_s ds}|\mathcal{F}_t]$$

reflecting the uncertainty in time-varying discount rates. As a result, to model the bond price we can characterise a dynamic of the short rate r_t. Alternatively, the AAO allows us to describe the dynamic of the bond price from its initial value and the knowledge of its volatility function. Therefore, assuming further hypothesis, the shape taken by the volatility function fully characterise the dynamic of the bond price and some specific functions gave their names to popular models commonly used in practice. Hence, the dynamics of the zero-coupon bond price are

$$\frac{dP(t, T)}{P(t, T)} = r_t dt \pm \Gamma_P(t, T) dW_P(t) \text{ with } P(T, T) = 1 \tag{1.3.7}$$

where $(W_P(t))_{t\geq 0}$ is valued in $\mathbb{R}^n$ and $\Gamma_P(t, T)$ [1] is a family of local volatilities parameterised by their maturities T. However, practitioners would rather work with the forward instantaneous rate which is related to the bond price by

$$f_P(t, T) = -\partial_T \ln P(t, T)$$

[1] $\Gamma_P(t, T)dW(t) = \sum_{j=1}^{n} \Gamma_{P,j}(t, T)dW_j(t)$

The relationship between the bond price and the rates in general was found by Heath et al. [1992] and following their approach the forward instantaneous rate is

$$f_P(t,T) = f_P(0,T) \mp \int_0^t \gamma_P(s,T) dW_P(s) + \int_0^t \gamma_P(s,T)\Gamma_P(s,T)^T ds$$

where $\gamma_P(s,T) = \partial_T \Gamma_P(t,T)$. The spot rate $r_t = f_P(t,t)$ is therefore

$$r_t = f_P(0,t) \mp \int_0^t \gamma_P(s,t) dW_P(s) + \int_0^t \gamma_P(s,t)\Gamma_P(s,t)^T ds$$

Similarly to the bond price, the short rate is characterised by the initial yield curve and a family of bond price volatility functions. However, either the bond price or the short rate above are too general and additional constraints must be made on the volatility function. A large literature flourished to model the discount-rate process and risk premia in continuous time with stochastic processes. The literature on term structure of interest rates is currently dominated by two different frameworks. The first one is originated by Vasicek [1977] and extended among others by Cox, Ingersoll, and Ross [1985]. It assumes that a finite number of latent factors drive the whole dynamics of term structure, among which are the Affine models. The other framework comprises curve models which are calibrated to the relevant forward curve. Among them are forward rate models generalised by Heath, Jarrow and Morton (HJM) [1992], the libor market models (LMM) initiated by Brace, Gatarek and Musiela (BGM) [1997] and the random field models introduced by Kennedy [1994].

As an example of the HJM models, Frachot [1995] and Duffie et al. [1993] considered respectively the special case of the quadratic and linear factor model for the yield price. In that setting, we assume that at a given time t all the zero-coupon bond prices are function of some state variables. We can further restrain the model by assuming that market price of claims is a function of some Markov process.

***Assumption* 1** *We assume there exists a Markov process X_t valued in some open subset $D \subset \mathbb{R}^n \times [0,\infty)$ such that the market value at time t of an asset maturing at $t+\tau$ is of the form*

$$f(\tau, X_t)$$

where $f \in C^{1,2}(D \times [0,\infty[)$ and $\tau \in [0,\mathcal{T}]$ with some fixed and finite $\mathcal{T}$.

For tractability, we assume that the drift term and the market price of risk of the Markov process X are nontrivially affine under the historical probability measure $\mathbb{P}$. Only technical regularity is required for equivalence between absence of arbitrage and the existence of an equivalent martingale measure. That is, the price process of any security is a $\mathbb{Q}$ martingale after normalisation at each time t by the riskless asset $e^{\int_0^t R(X_s)ds}$ (see Lemma (1.3.1)). Therefore, there is a standard Brownian motion W in $\mathbb{R}^n$ under the probability measure $\mathbb{Q}$ such that

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t \tag{1.3.8}$$

where the drift $\mu : D \to \mathbb{R}^n$ and the diffusion $\sigma : D \to \mathbb{R}^{n\times n}$ are regular enough to have a unique strong solution valued in D. To be more precise, the domain D is a subset of $\mathbb{R}^n \times [0,\infty)$ and we treat the state process X defined so that (X_t, t) is in D for all t. We assume that for each t, $\{x : (t,x) \in D\}$ contains an open subset of $\mathbb{R}^n$. We are now going to be interested in the choice for (f,μ,σ) that are compatible in the sense that f characterises a price process. El Karoui et al. [1992] explained interest rates as regular functions of an n-dimensional state variable process X

$$f_P(t,T) = F(t,T,X_t)\,, t \le T$$

where F is at most quadratic in X. By constraining X to be linear, it became the Quadratic Gaussian model (QG). Similarly, introducing n state variables Frachot in [1995] described the linear factor model

$$f_P(t,T) = b(t,T) + a(t,T).X_t \ , \forall t \leq T \tag{1.3.9}$$

where $f_P(t,T)$ is the instantaneous forward rate and the functions $b(t,T)$ and $a(t,T) = [a_1(t,T), ..., a_n(t,T)]^\top$ are deterministic. He showed that by discarding the variables $X_i(t)$ one can identify the state variables to some particular rates such that

$$f_P(t,T) = b(t,T) + a_1(t,T)f_P(t,t+\theta_1) + ... + a_n(t,T)f_P(t,t+\theta_n) \ , \forall t \leq T$$

where θ_i for $i = 1, .., n$ are distinct maturities and the functions $a(.,.)$ and $b(.,.)$ must satisfy extra compatibility conditions

$$b(t,t+\theta_i) = 0 \ , a_i(t,t+\theta_i) = 1 \ , a_j(t,t+\theta_i) = 0 (j \neq i)$$

that is, the rate with maturity $(t+\theta_i)$ can only be identified with a single state variable. Frachot also showed that the QG model could be seen as a linear factor model with constraints, that is by considering a model linear in X and $XX^\top$. In that case, the entire model can not be identified in term of some particular rates but only $XX^\top$ as in the case of X it would leads to $a_i(t,T) = 0$ for all i belonging to X. Frachot [1995] and Filipovic [2001] proved that the quadratic class represents the highest order of polynomial functions that one can apply to consistent time-separable term structure models. Another example is the Yield factor model defined by Duffie et al. [1993] which is a special case of the linear factor model where the asset $f(\tau, X_t)$ in assumption (1) is the price of a zero-coupon bond of maturity $t+\tau$

$$f(T-t, X_t) = E^Q[e^{-\int_t^T R(X_s)ds} | \mathcal{F}_t] \tag{1.3.10}$$

so that given the definition of the risk-free zero coupon bond, $P(t,T)$ is its price at time t. The short rate is assumed to be such that there is a measurable function $R : D \to \mathbb{R}$ defined as the limit of yields as the maturity goes to zero, that is

$$R(x) = \lim_{\tau \to 0} -\frac{1}{\tau} \log f(\tau, x) \text{ for } x \in D$$

Depending on the asset prices and the markets under consideration, many different authors have constructed such compatible set (f, ν, σ). For instance, the Factors models were extended by Duffie, Pan and Singleton [2000] in the case of jump-diffusion models, and a unified presentation was developed by Duffie, Filipovic and Schachermayer [2003]. Affine models is a class of time-homogeneous Markov processes that has arisen from a large and growing range of useful applications in finance. They imply risk premium on untraded risks which are linear functions of the underlying state variables, creating a method for going between the objective and risk-neutral probability measures while retaining convenient affine properties for both measures. Duffie et al. [2003] provided a definition and complete characterization of regular affine processes. Given a state space of the form $D = \mathbb{R}_+^m \times \mathbb{R}^n$ for integers $m \geq 0$ and $n \geq 0$ the key affine property, is roughly that the logarithm of the characteristic function of the transition distribution $p_t(x, .)$ of such a process is affine with respect to the initial state $x \in D$. Given a regular affine process X, and a discount-rate process $\{R(X_t) : t \geq 0\}$ defined by an affine map $x \to R(x)$ on D into $\mathbb{R}$, the discount factor

$$P(t,T) = E[e^{-\int_t^T R(X_u)du} | X_t]$$

is well defined under certain conditions, and is of the anticipated exponential-affine form in X_t.

1.4 The pricing kernel

We saw in Section (1.3.2) that risk neutral returns are risk-adjusted natural returns. That is, the return under the risk neutral measure is the return under the natural measure with the risk premium subtracted out. Hence, to use risk neutral prices to estimate natural probabilities, we must know the risk adjustment to add it back in. This is equivalent

to knowing both the agent's risk aversion and his subjective probability which are non-observable. Various authors tried to infer these risks from the market with the help a model, but with more or less success. Further, the natural expected return of a strategy depends on the risk premium for that strategy, so that knowledge on the kernel can help in estimating the variability of the risk premium. At last, we are unable to obtain the current market forecast of the expected returns on equities directly from their prices, and we are left to using historical returns. Even though the risk premium is not directly observable from option prices various authors tried to infer it. This is because there is a rich market in equity option prices and a well developed theory to extract the martingale or risk neutral probabilities from these prices. As a result, one can use these probabilities to forecast the probability distribution of future returns.

1.4.1 Defining the pricing kernel

The asset pricing kernel summarises investor preferences for payoffs over different states of the world. In absence of arbitrage, all asset prices can be expressed as the expected value of the product of the pricing kernel and the asset payoff (see Equation (1.2.3)). Discounting payoffs using time and risk preferences, it is also called the stochastic discount factor. Hence, combined with a probability model for the states, the pricing kernel gives a complete description of asset prices, expected returns, and risk premia. In a discrete time world with asset payoffs $h(X)$ at time T, contingent on the realisation of a state of nature $X \in \Omega$, absence of arbitrage opportunity (AAO) (see Dybvig et al. [2003]) implies the existence of positive state space prices, that is, the Arrow-Debreu contingent claims prices $p(X)$ paying \$1 in state X and nothing in any other states (see Theorem (E.1.1)). If the market is complete, then these state prices are unique. The current value π_h of an asset paying $h(X)$ in one period is given by

$$\pi_h = \int h(X)dP(X)$$

where $P(X)$ is a price distribution function. Letting $r(X^0)$ be the riskless rate as a function of the current state X^0, such that $\int p(X)dX = e^{-r(X^0)T}$, we can rewrite the price as

$$\begin{aligned} \pi_h &= \int h(X)dP(X) = (\int dP(X)) \int h(X)\frac{dP(X)}{\int dP(X)} = e^{-r(X^0)T}\int h(X)dq^*(X) \\ &= e^{-r(X^0)T}E^*[h(x)] = E[h(X)\xi(X)] \end{aligned} \tag{1.4.11}$$

where the asterisk denotes the expectation in the martingale measure and where the pricing kernel, that is, the state-price density $\xi(X)$ is the Radon-Nikodym derivative of $P(X)$ with respect to the natural measure denoted $F(X)$. With continuous distribution, we get $\xi(X) = \frac{p(X)}{f(X)}$ where $f(X)$ is the natural probability, that is, the actual or relevant objective probability distribution, and the risk-neutral probabilities are given by

$$q^*(X) = \frac{p(X)}{\int p(X)dX} = e^{r(X^0)T}p(X)$$

so that

$$\xi(X) = e^{-r(X^0)T}\frac{q^*(X)}{f(X)} \tag{1.4.12}$$

We let X_t denote the current state and X_{t+1} be a state after one period and assume that it fully describe the state of nature. Then, $\xi_{t,t+1}$ is the empirical pricing kernel associated with returns between date t and $t+1$, conditional on the information available at date $t \leq t+1$. It is estimated as

$$\xi_{t,t+1} = \xi(X_t, X_{t+1}) = \frac{p(X_{t+1}|X_t)}{f(X_{t+1}|X_t)} = e^{-r\Delta t}\frac{q(X_{t+1}|X_t)}{f(X_{t+1}|X_t)}$$

where q is the risk-neutral density, and f is the objective (historical) density. Hence, the kernel is defined as the price per unit of probability in continuous state spaces (see Equation (E.6.13)). Note, we can always rewrite the risk-neutral probability as

$$q(X_{t+1}|X_t) = e^{r\Delta t}\xi_{t,t+1}f(X_{t+1}|X_t)$$

where the natural probability transition function $f(X_{t+1}|x_t)$, the kernel $\xi_{t,t+1}$, and the discount factor $e^{-r\Delta t}$ are unknowns. One can therefore spend his time either modelling each separate element and try to recombine them or directly infer them from the risk-neutral probability. However, one can not disentangle them, and restrictions on the kernel, or the natural distribution must be imposed to identify them separately from the knowledge of the risk-neutral probability.

One approach is to use the historical distribution of returns to estimate the unknown kernel and then link the historical estimate of the natural distribution to the risk neutral distribution. For instance, Jackwerth et al. [1996] used implied binomial trees to represent the stochastic process, Ait Sahalia et al. [1998] [2000] combined state prices derived from option prices with estimates of the natural distribution to determine the kernel, Bollerslev et al. [2011] used high frequency data to estimate the premium for jump risk in a jump diffusion model. Assuming a transition independent kernel, Ross [2013] showed that the equilibrium system above could be solved without the need of historical data.

1.4.2 The empirical pricing kernel

The discount factor can be seen as an index of bad times such that the required risk premium for any asset reflects its covariation with bad times. Investors require higher risk premia for assets suffering more in bad times, where bad times are periods when the marginal utility (MU) of investors is high.

Lucas [1978] expressed the pricing kernel as the intertemporal marginal rate of substitution

$$\xi_{t,t+1} = \frac{u^{'}(c_{t+1})}{u^{'}(c_t)}$$

where $u(\bullet)$ is a utility function and c_t is the consumption at time t. Under the assumption of power utility, the pricing kernel becomes

$$\xi_{t,t+1} = e^{-\rho}(\frac{c_{t+1}}{c_t})^{-\gamma}$$

where ρ is the rate of time preference and γ is a level of relative risk aversion. The standard risk-aversion measures are usually functions of the pricing kernel slope, and Pratt showed that it is the negative of the ratio of the derivative of the pricing kernel to the pricing kernel, that is

$$\gamma_t = -\frac{c_{t+1}\xi^{'}_{t,t+1}(c_{t+1})}{\xi_{t,t+1}(c_{t+1})}$$

Generally, the pricing kernel depends not just on current and future consumption, but also on all variables affecting the marginal utility. When the pricing kernel is a function of multiple state variables, the level of risk aversion can also fluctuate as these variables change. Several authors used Equation (Equation (1.2.3)) to investigate the characteristics of investor preferences in relation to equity securities, that is

$$S_t = E[\xi_{t,t+1}X_{t+1}|\mathcal{F}_t]$$

where S_t is the current stock price and X_{t+1} is the asset payoff in one period. Hansen et al. [1982] identified the pricing kernel equation above with the unconditional version

$$E[\frac{S_{t+1}}{S_t}\xi_{t,t+1}] = 1 \tag{1.4.13}$$

They specified the aggregate consumption growth rate as a pricing kernel state variable, and measured consumption using data from the National Income and Products Accounts. Campbell et al. [1997] and Cochrane [2001] provided comprehensive treatments of the role of the pricing kernel in asset pricing. As it is not clear among researchers about what state variables should enter the pricing kernel, Rosenberg et al. [2001] [2002] considered a pricing kernel projection estimated without specifying the state variables. Writing the original pricing kernel as $\xi_{t,t+1} = \xi_{t,t+1}(Z_t, Z_{t+1})$ where Z_t is a vector of pricing kernel state variables they re-wrote the pricing equation by factoring the joint density $f_t(X_{t+1}, Z_{t+1})$ into the product of the conditional density $f_t(Z_{t+1}|X_{t+1})$ and the marginal density $f_t(X_{t+1})$. The expectation is then evaluated in two step, first the pricing kernel is integrated using the conditional density, giving the projected pricing kernel $\xi^*_{t,t+1}(X_{t+1})$. second, the product of the projected pricing kernel and the payoff variable is integrated using the marginal density, giving the asset price

$$S_t = E_t[\xi^*_{t,t+1}(X_{t+1})X_{t+1}] \, , \xi^*_{t,t+1}(X_{t+1}) = E_t[\xi_{t,t+1}(Z_t, Z_{t+1})|X_{t+1}]$$

Thus, for the valuation of an asset with payoffs depending only on X_{t+1}, the pricing kernel is summarised as a function of the asset payoff which can vary over time, reflecting time-variation in the state variables. It is called the empirical pricing kernel (EPK) and was estimated on monthly basis from 1991 to 1995 on the $S\&P$ 500 index option data. Barone-Adesi et al. [2008] relaxed the restriction that the variances of the objective return distribution and risk neutral distribution are equal, along with higher moments such as skewness and kurtosis. Further, Barone-Adesi et al. combined an empirical and a theoretical approach relaxing the restriction that the objective return distribution and risk neutral distribution share the same volatility and higher order moments.

1.4.3 Analysing the expected risk premium

One of the consequence of the absence of arbitrage opportunities (AAO) is that the expected value of the product of the pricing kernel and the gross asset return R^g_t defined in Equation (3.3.4) must equal unity (see Equation (1.4.13)). That is, assuming dividend adjusted prices and setting the one period gross return as $R^g_{t+1} = 1 + R_{t,t+1} = \frac{S_{t+1}}{S_t}$, we get

$$\pi_t = E[\xi_{t+1}R^g_{t+1}|\mathcal{F}_t] = 1$$

Hence, the one-period risk-free rate r_{ft} can be written as the inverse of the expectation of the pricing kernel

$$r_{ft} = E[\xi_{t+1}|\mathcal{F}_t]^{-1}$$

Further, Equation (1.4.13) implies that the expected risk premium is proportional to the conditional covariance of its return with the pricing kernel

$$E_t[R^g_{t+1} - r_{ft}] = -r_{ft}Cov_t(\xi_{t+1}, R^g_{t+1})$$

where $Cov_t(\bullet, \bullet)$ is the covariance conditional on information available at time t (see Whitelaw [1997]). As a result, the conditional Sharpe ratio of any asset, defined as the ratio of the conditional mean excess return to the conditional standard deviation of this return, can be written in terms of the volatility of the pricing kernel and the correlation between the kernel and the return

$$\begin{aligned}\frac{E_t[R^g_{t+1} - r_{ft}]}{\sigma_t(R^g_{t+1} - r_{ft})} &= -r_{ft}\frac{Cov_t(\xi_{t+1}, R^g_{t+1})}{\sigma_t(R^g_{t+1})} \\ &= -r_{ft}\sigma_t(\xi_{t+1})Corr_t(\xi_{t+1}, R^g_{t+1})\end{aligned}$$

where $\sigma_t(\bullet)$ and $Corr_t(\bullet,\bullet)$ are respectively the standard deviation and correlation, conditional on information at time t. Hence, this equation shows that the conditional Sharpe ratio $M_{SR,t}$ is proportional to the correlation between the pricing kernel and the return on the market $R_{m,t}$

$$M_{SR,t} = -r_{ft}\sigma_t(\xi_{t+1})Corr_t(\xi_{t+1}, R^g_{m,t+1})$$

Hence, if the Sharpe ratio varies substantially over time, then the variation is mostly attributable to the variation in the conditional correlation and depend critically on the modelling of the pricing kernel.

One approach is to specify the kernel ξ_t as a function of asset returns. For instance, modelling the pricing kernel as a linear function of the market return produces the conditional CAPM. Since risk aversion implies a negative coefficient on the market return, modelling the pricing kernel as a linear function of the market return leads the correlation to be around -1 so that the market Sharpe ratio is approximately constant over time. Alternatively, modelling the discount factor as a quadratic function of the market return gives the conditional three moment CAPM proposed by Kraus et al. [1976] allowing for some time variation in the market SRs due to to the pricing of skewness risk, but the correlation is still pushed towards -1. Other modelling of the pricing kernel exists, such as nonlinear function of the market return or a linear function of multiple asset returns, but with limited time variation in the correlation (see Bansal et al. [1993]). Based on explanatory and predictive power, a number of additional factors, including small firm returns and return spreads between long-term and short-term bonds, have been proposed and tested but the correlations between the discount factor and the market return tend to be relatively stable. Another approach uses results from a representative agent and exchange economy, and models the pricing kernel as the marginal rate of substitution (MRS) over consumption leading to the consumption CAPM (see Breeden et al. [1989]). When the MRS depends on consumption growth and the stock market is modelled as a claim on aggregate consumption, one might expect the correlation and the Sharpe ratio to be relatively stable. Assuming consumption growth follows a two-regime autoregressive process, Whitelaw [1997a] obtained regime shifts (phases of business cycle) with mean and volatility being negatively correlated, implying significant time variation in the Sharpe ratio.

1.4.4 Infering risk premium from option prices

In a recent article, assuming that the state-price transition function $p(X_i, X_j)$ is observable, Ross [2013] used the recovery theorem to uniquely determine the kernel, the discount rate, future values of the kernel, and the underlying natural probability distribution of return from the transition state prices alone. Note, the notion of transition independence was necessary to separately determine the kernel and the natural probability distribution. Other approaches used the historical distribution of returns to estimate the unknown kernel and thereby link the historical estimate of the natural distribution to the risk-neutral distribution. Alternatively, one could assume a functional form for the kernel. Ross [2013] showed that the equilibrium system in Equation (E.6.15) could be solved without the need to use either the historical distribution of returns or independent parametric assumptions on preferences to find the market's subjective distribution of future returns. The approach relies on the knowledge of the state transition matrix whose elements give the price of one dollar in a future state, conditional on any other state. One way forward is to find these transaction prices from the state-prices for different maturities derived from the market prices of simple options by using a version of the forward equation for Markov processes. As pointed out by Ross, there are a lot of possible applications if we know the kernel (market's risk aversion) and the market's subjective assessment of the distribution of returns. We can use the market's future distribution of returns much as we use forward rates as forecasts of future spot rates. Rather than using historical estimates of the risk premium on the market as an input into asset allocation models (see Section (1.4.3)), we should use the market's current subjective forecast. The idea can extend to all project valuation using historical estimates of the risk premium.

1.5 Modelling asset returns

1.5.1 Defining the return process

Setting $\alpha_t = \log(B_t)$ for $0 \leq t \leq T$ where $B(t)$ is the riskless asset defined in Equation (1.3.6), we call α the return process for B, such that $dB = Bd\alpha$ with $\alpha_0 = 0$. Since the riskless asset, B, is absolutely continuous, then

$$\alpha_t = \int_0^t r_s ds$$

and, r_s, is the time-s interest rate with continuous compounding. Similarly to the riskless asset, assuming any semi-martingale price process, S, possibly with jumps, we want to consider its return process for a stock to satisfy the equation

$$dS = S_- dR$$

which is equivalent to

$$S_t = S_0 + \int_0^t S_u dR_u \tag{1.5.14}$$

as well as

$$R_t = \int_0^t \frac{1}{S_{u-}} dS_u \tag{1.5.15}$$

where the return, R_t, is defined over the range $[0, t]$. Note, $\frac{dS_t}{S_{t-}}$, is the infinitesimal rate of return of the price process, S, while, R_t, is its integrated version. Given a (reasonable) price process, S, the above equation defines the corresponding return process R. Similarly, given, S_0, and a semimartingale, R, Equation (1.5.14) always has a semimartingale solution S. It is unique and given by

$$S_t = S_0 \psi_t(R) \, , 0 \leq t \leq t$$

where

$$\psi_t(R) = e^{R_t - R_0 - \frac{1}{2}[R,R]_t} \prod_{s \leq t} (1 + \Delta R_s) e^{-\Delta R_s + \frac{1}{2}(\Delta R_s)^2}$$

with $\Delta R_s = R_s - R_{s-}$ and quadratic variation

$$[R, R]_t = R_t R_t - 2 \int_0^t R_{s-} dR_s$$

Note, R, is such that $1 + \Delta R > 0$ for any and all jumps if and only if $\psi_t(R) > 0$ for all t. Similarly with weak inequalities. Further, we have $\psi_0(R) = 1$. In that setting, the discounted price process is given by

$$\overline{S} = \frac{S}{B} = S_0 \psi(R) e^{-\alpha}$$

Since $-\alpha$ is continuous, $\alpha_0 = 0$, and $[-\alpha, -\alpha] = 0$ we get the semimartingale exponential expression

$$\overline{S} = S_0 \psi(R) \psi(-\alpha)$$

Further, since $[R, -\alpha] = 0$, using a probability property, we get

$$\overline{S} = S_0 \psi(R - \alpha) = \overline{S}_0 \psi(R - \alpha)$$

so that $Y = R - \alpha$ can be interpreted as the return process for the discounted price process $\overline{S}$. For example, in the Black-Scholes [1973] model the return process is given by

$$R_t = \int_0^t \frac{1}{S_u} dS_u = \mu t + \sigma \hat{W}_t$$

where $\hat{W}_t$ is a standard Brownian motion in the historical probability measure $\mathbb{P}$. With a constant interest rate, r, we get $\alpha_t = rt$, and the return process for the discounted price process becomes

$$Y_t = (\mu - r)t + \sigma \hat{W}_t$$

1.5.2 Valuing potfolios

In order to describe general portfolio valuation and its optimisation, we consider a general continuous time framework where the stock price, S, is the decomposable semimartingale. The example of a multi-underlying diffusion model is detailed in Appendix (F.1). We let the world be defined as in Assumption (F.1.1), and we assume $(N+1)$ assets $S = (S^0, S^1, .., S^N)$ traded between 0 and T_H where the risky-assets $S^i; 1 \leq i \leq N$ are Ito's random functions given in Equation (1.3.5). The risk-free asset S^0 satisfies the dynamics

$$dS_t^0 = r_t S_t^0 dt$$

The data format of interest is returns, that is, relative price change. From the definition of the percent return given in Equation (3.3.4), we consider the one period net returns $R_{t,t+1}^i = \frac{S_{t+1}^i}{S_t^i} - 1$, which for high-frequency data are almost identical to log-price changes $r_L^i(t, t+1) = \ln S_{t+1}^i - \ln S_t^i$ (at daily or higher frequency). The portfolio strategy is given by the process $(\delta_i(t))_{0 \leq i \leq N}$ corresponding to the quantity invested in each asset. The financial value of the portfolio δ is given by $V(\delta)$, and its value at time t satisfies

$$V_t(\delta) =< \delta(t), S_t >= \sum_{i=0}^N \delta^i(t) S_t^i \tag{1.5.16}$$

Assuming a self-financing portfolio (see Definition (F.1.2)), its dynamics satisfy

$$dV_t(\delta) = \sum_{i=0}^N \delta^i(t) dS_t^i = \delta^0(t) dS_t^0 + \sum_{i=1}^N \delta^i(t) dS_t^i$$

Assuming a simple strategy, for any dates $t < t^{'}$, the self-financing condition can be written as

$$V_{t'} - V_t = \int_t^{t'} \sum_{i=0}^N \delta^i(u) dS_u^i$$

From the definition of the portfolio we get $\delta^0(t) S_t^0 = V_t(\delta) - \sum_{i=1}^N \delta^i(t) S_t^i$, so that, plugging back in the SDE and factorising, the dynamics of the portfolio become

$$dV_t(\delta) = V_t(\delta) r_t dt + \sum_{i=1}^N \delta^i(t) S_t^i \big(\frac{dS_t^i}{S_t^i} - r_t dt \big)$$

Further, we let $(h_i(t))_{1 \leq i \leq N} = (\delta^i(t) S_t^i)_{1 \leq i \leq N}$ be the fraction of wealth invested in the ith risky security, or the dollars invested in the ith stock, so that the dynamics of the portfolio rewrite

$$dV_t(h) = \big(V_t(h) - \sum_{i=1}^N h_i(t) \big) r_t dt + \sum_{i=1}^N h_i(t) R_{t,t+1}^i \tag{1.5.17}$$

which can be rewritten as

$$dV_t(h) = V_t(h)r_t dt + \sum_{i=1}^{N} h_i(t)(R^i_{t,t+1} - r_t dt)$$

corresponding to Equation (F.1.1), that is

$$dV_t(h) = r_t V_t(h)dt + < h_t, R_t - r_t I dt >$$

where $h_t = (\delta S)_t$ corresponds to the vector with component $(\delta^i(t)S^i_t)_{1\leq i\leq N}$ describing the amount to be invested in each stock, and where R_t is a vector with component $(R^i_{t,t+1})_{1\leq i\leq N}$ describing the return of each risky security. It is often convenient to consider the portfolio fractions

$$\pi_v = \left\{\pi_v(t) = (\pi^0_v(t), .., \pi^N_v(t))^\top, t \in [0, \infty)\right\}$$

with coordinates defined by

$$\pi^i_v(t) = \frac{\delta^i(t)S^i_t}{V_t(h)} = \frac{h_i(t)}{V_t(h)} \tag{1.5.18}$$

denoting the proportion of the investor's capital invested in the ith asset at time t. It leads to the dynamics of the portfolio defined as

$$\frac{dV_t(h)}{V_t(h)} = r_t dt + < \pi_v(t), R_t - r_t I dt >$$

The linear Equation having a unique solution, knowing the initial investment and the weights of the portfolio is enough to characterise the value of the portfolio. In a viable market, one of the consequences of the absence of arbitrage opportunity is that the instantaneous rate of returns of the risky assets satisfies Equation (F.1.2). Hence, in that setting, the dynamics of the portfolio become

$$\frac{dV_t(h)}{V_t(h)} = r_t dt + < \pi_v(t), \sigma_t \lambda_t dt > + < \pi_v(t), \sigma_t d\tilde{W}_t >$$

where

$$< \pi_v(t), \sigma_t \lambda_t dt >$$

represents the systematic component of returns of the portfolio, and

$$< \pi_v(t), \sigma_t d\tilde{W}_t >$$

represents the idiosyncratic component of the portfolio.

***Remark* 1.5.1** *To value the portfolio, one must therefore assume a model for the process* $R^i_{t,t+1}$.

One way forward is to consider Equation (1.5.15) and follow the interest rate modelling described in Section (1.3.3) by assuming that the equity return is a function of some latent factors or state variables.

1.5.3 Presenting the factor models

1.5.3.1 The presence of common factors

Following Ilmanen [2011], we state that the expected returns on a zero-coupon government bond are known, but for all other assets the expected returns are uncertain ex-ante and unknowable ex-post, while the realised returns are knowable ex-post but do not reveal what investors expected. As a result, apart from ZC-bond, institutional asset holders, such as pension funds, must infer expected returns from empirical data, past returns, and investor surveys, and from statistical models. The aim being to forecast expected future returns in time and across assets in view of valuing portfolios (see Section (1.5.2)) and performing the asset allocation process described in Section (2.1.2). Given constant expected returns, the best way for estimating expected future return is to calculate the sample average realised return over a long sample period. However, when the risk premia λ_t is time-varying or stochastic, this average becomes biased. It is understood that time-varying expected returns (risk-premia) can make historical average returns highly misleading since it can reflect historical average unexpected returns. As a result, practitioners became interested in ex-ante return measures such as valuation ratios or dividend discount models. Even though market's expected returns are unobservable, both academics and practitioners focused on forward-looking indicators such as simple value and carry measures as proxies of long-run expected returns. The carry measures include any income return and other returns that are earned when capital market conditions are unchanged. Among these measures, dividend yield was the early leader, but broader payout yields including share buybacks have replaced it as the preferred carry measure, while earnings yield and the Gordon model (DDM) equity premium became the preferred valuation measures (see Campbell et al. [1988]). This is because they give better signals than extrapolation of past decade realised returns. For instance, in order to estimate prospective returns, one can relate current market prices to some sort of value anchor (historical average price, yield, spread etc.) assuming that prices will revert to fair values which are proxied by the anchors. Note, value or carry indicators are inherently narrow and one should consider broader statistical forecasting models such as momentum, volatility, macroeconomic environment etc. The presence of portfolio forecastability, even if single assets are sometime unforecastable, is one of the key features of cointegration. Two or more processes are said to be cointegrated if there are long-term stable regression relationships between them, even if the processes themselves are individually integrated. This means that there are linear combinations of the processes that are autocorrelated and thus forecastable. Hence, the presence of cointegrating relationships is associated with common trends, or common factors, so that cointegration and the presence of common factors are equivalent concepts. This equivalence can be expressed more formally in terms of state-space models which are dynamic models representing a set of processes as regressions over a set of possibly hidden factors or state variables. One approach is to express asset returns in terms of factor models.

1.5.3.2 Defining factor models

The economical literature considered Factors models of the risk premia in discrete time, and concentrated on the special case of Affine models. In the equity setting, given the one period percentage return $R_t = R_{t-1,t}$, the econometric linear factor model in Equation (1.3.9) can be expressed as

$$R_t = \alpha_t + \sum_{k=1}^{m} \beta_k F_{kt} + \epsilon_t \tag{1.5.19}$$

where the terms F_{kt} for $k = 1, .., m$ represent returns of the risk factors associated with the market under condition, m is the number of factors explaining the stock returns, and α_t is the drift of the idiosyncratic component. We can then view the residuals as increments of a process that will be estimated

$$X_t = X_0 + \sum_{s=1}^{t} \epsilon_{i,s}$$

As a result, using continuous-time notation, the continuous-time model for the evolution of stock prices becomes

$$\frac{dS(t)}{S(t)} = \alpha_t dt + \sum_{k=1}^{m} \beta_k F_k(t) + dX(t) \tag{1.5.20}$$

where the term $\sum_{k=1}^{m} \beta_k F_k(t)$ represents the systematic component of returns. The coefficients β_k are the corresponding loadings. The idiosyncratic component of the stock returns is given by

$$d\tilde{X}(t) = \alpha_t dt + dX(t)$$

where α_t represents the drift of the idiosyncratic component, which implies that $\alpha_t dt$ represents the excess rate of return of the stock with respect to the stock market, or some industry sector, over a particular period of time. The term $dX(t)$ is assumed to be the increment of a stationary stochastic process which models price fluctuations corresponding to over-reactions or other idiosyncratic fluctuations in the stock price which are not reflected the industry sector.

1.5.3.3 CAPM: a one factor model

The capital asset pricing model (CAPM) introduced by Sharpe [1964], which is an approach to understanding expected asset returns or discount rates beyond the riskless rate, is a simplification of Factors models defined above, based on some restrictive assumptions

- one-period world (constant investment opportunity set and constant risk premia over time)
- access to unlimited riskless borrowing and lending as well as tradable risky assets
- no taxes or transaction costs
- investors are rational mean variance optimisers (normally distributed asset returns, or quadratic utility function)
- investors have homogeneous expectations (all agree about asset means and covariances)

These assumptions ensure that every investor holds the same portfolio of risky assets, combining it with some amount of the riskless asset (based on the investor's risk aversion) in an optimal way. Even though these assumptions are too restrictive, the main insight is that only systematic risk is priced in the sense that it influences expected asset returns. In the CAPM, the ith asset's return is given by

$$R_{it} = \alpha_i(t) + \beta_{iM}(t) R_{Mt} + \epsilon_{it}$$

where R_{Mt} is the market return, and the residual ϵ_{it} is normally distributed with zero mean and variance $\sigma_i^2(t)$. It represent a straight line in the (R_i, R_M) plane where α_i is the intercept and $\beta_{iM} = \frac{Cov(R_i, R_M)}{\sigma_M^2}$ is the slope called the beta value. According to the CAPM, in equilibrium, the ith asset's expected excess return $E_t[R_{i,(t+1)}]$ is a product of its market beta $\beta_{iM}(t)$ and the market risk premium $\lambda_M(t)$, given by

$$E_t[R_{i,(t+1)}] = \beta_{iM}(t)\lambda_M(t)$$

Note, λ_M is the market risk premium common to all assets, but it also reflects the price of risk (investor risk aversion). Defining the portfolio as $P(t) = \sum_{i=1}^{N} w_i R_{it}$ where w_i is the ith weight of the portfolio, the portfolio beta is $\beta_P(t) = \sum_{i=1}^{N} w_i \beta_{iM}(t)$, and the variance of the portfolio is given by

$$Var(P(t)) = \beta_P^2(t)\sigma_M^2 + \sum_{i=1}^{N} w_i^2 Var(\epsilon_{it})$$

Note, in the case where the weights are uniform $w_1 = w_2 = .. = w_N = \frac{1}{N}$, we get the limit

$$\sum_{i=1}^{N}(\frac{1}{N})^2 Var(\epsilon_{it}) \to 0 \, , N \to \infty$$

and the idiosyncratic risk vanishes. Realised returns reflect both common (systematic) risk factors and asset specific (idiosyncratic) risk. While idiosyncratic risk can be diversified away as N increases, risk premia compensate investors for systematic risk that can not be diversified away. However, the CAPM does not specify how large the market risk premium should be. While the CAPM is a static model, a more realistic approach should reflect the view that market risk aversion varies with recent market moves and economic conditions, and that the amount of risk in the market varies with stock market volatility and asset correlations.

1.5.3.4 APT: a multi-factor model

Relaxing some of the simplifying assumptions of the CAPM, new approaches flourished to explain expected returns such as multiple systematic factors, time-varying risk premia, skewness and liquidity preferences, market frictions, investor irrationalities etc. The Arbitrage Pricing Theory (APT) developed by Ross [1976] is one of the theories that relate stock returns to macroeconomic state variables. However, identifying risk factors in the multi-factor models in Equation (1.5.20) that are effective at explaining realised return variations over time is an open problem, since theory gives limited guidance. One can either consider theoretical factor models or empirical factor models where factors are chosen to fit empirical asset return behaviour (see Ilmanen [2011]). Factors with a strong theoretical basis include aggregate consumption growth, investment growth, as well as overall labour income growth and idiosyncratic job risk. Equity factors that are primarily empirical include value, size, and often momentum. The list can be extended to indicators like return reversals, volatility, liquidity, distress, earnings momentum, quality factors such as accruals, and corporate actions such as asset growth and net issuance. Beyond equities, sensitivities to major asset classes are the most obvious factors to consider. For instance, the inflation factor is especially important for bonds, and liquidities for other assets. In the case where several common factors generate undiversifiable risk, then a multi-factor relation holds. For instance, with K sistematic factors, the ith asset's expected excess return reflects its factor sensitivities $(\beta_{i_1}, .., \beta_{i_K})$ and the factor risk premia $(\lambda_1, .., \lambda_k)$

$$E_t[R_{i,(t+1)}] = \beta_{i_1}(t)\lambda_1(t) + ... + \beta_{i_K}(t)\lambda_k(t)$$

More generally, if we assume that the stochastic discount factor (SDF) is linearly related to a set of common risk factors, then asset returns can be described by a linear factor model. Moreover, the idea that the risk premium depends on covariation with the SDF (bad times, high MU periods) also applies to the risk factors, not just to individual assets.

1.6 Introducing behavioural finance

We saw in Section (1.3) that understanding the risk premia in view of predicting asset returns was at the heart of finance, and that several answers existed. One of them, called behavioural finance, implies that market prices do not only reflect the rationally expected discounted cash flows from an asset (see Section (1.4.12)), but also incorporate noise from irrational investors. In complete markets, rational investors make forecast that correctly incorporate all available information, leaving no room for systematic forecasting errors. Behavioural economics and behavioural finance have challenged this paradigm (see Barberis et al. [2002]). As fully rational behaviour requires quite complex calculations, an alternative idea developed suggesting that investors exhibit bounded rationality, rationality that is limited by their cognitive resources and observational powers. Psychological biases predict specific systematic deviations from rationality causing mispricings. The main biases are heuristic simplifications such as rules of thumb, mental shortcuts, attention and memory biases, representativeness, conservatism, and/or self-deception such as overconfidence, overoptimism, biased self-attribution, confirmation bias, hindsight bias. The best known market-level mispricings are speculative bubbles, and the best known relative mispricings are value and momentum effects. Even though it is argued that rational traders will view any such mispricings as trading opportunities, which when realised,

will undo any price impact irrational traders might have, it did not happen in practice since these strategies are risky and costly (see Shleifer et al. [1997]). In fact, any observed predictability pattern can be interpreted to reflect either irrational mispricings or rational time-varying risk premia (rational learning about structural changes). It is likely that both irrational and rational forces drive asset prices and expected returns. That is, available information is not fully reflected in current market prices, leading to incomplete market situations. Hence, by challenging the complete market paradigm, behavioural finance attempted to propose an alternative theory, which is to be compared with the incomplete market theory (IMT).

1.6.1 The Von Neumann and Morgenstern model

The first important use of the EUT was that of Von Neumann and Morgenstern (VNM) [1944] who used the assumption of expected utility maximisation in their formulation of game theory. When comparing objects one needs to rank utilities but also compare the sizes of utilities. VNM method of comparison involves considering probabilities. If a person can choose between various randomised events (lotteries), then it is possible to additively compare for example a shirt and a sandwich. It is possible to compare a sandwich with probability 1, to a shirt with probability p or nothing with probability $1-p$. By adjusting p, the point at which the sandwich becomes preferable defines the ratio of the utilities of the two options. If options A and B have probability p and $1-p$ in the lottery, we can write the lottery L as a linear combination

$$L = pA + (1-p)B$$

and for a lottery with n possible options, we get

$$L = \sum_{i=1}^{n} p_i A_i$$

with $\sum_{i=1}^{n} p_i = 1$. VNM showed that, under some assumptions, if an agent can choose between the lotteries, then this agent has a utility function which can be added and multiplied by real numbers, which means the utility of an arbitrary lottery can be calculated as a linear combination of the utility of its parts. This is called the expected utility theorem. The required assumptions are made of four axioms about the properties of the agent's preference relation over simple lotteries, which are lotteries with just two options. Writing $B \preceq A$ for A is weakly preferred to B, the axioms are

1. Completeness: for any two simple lotteries L and M, either $L \preceq M$ or $M \preceq L$ (or both).
2. Transitivity: for any three lotteries L, M, N, if $L \preceq M$ and $M \preceq N$, then $L \preceq N$.
3. Convexity: if $L \preceq M \preceq N$, then there is $p \in [0,1]$ such that the lottery $pL + (1-p)N$ is equally preferable to M.
4. Independence: for any three lotteries L, M, N, $L \preceq M$ if and only if $pL + (1-p)N \preceq pM + (1-p)N$.

A VNM utility function is a function from choices to the real numbers $u : X \to \mathbb{R}$ which assigns a real number to every outcome in a way that captures the agent's preferences over simple lotteries. Under the four assumptions mentioned above, the agent will prefer a lottery L_2 to a lottery L_1, if and only if the expected utility of L_2 is greater than the expected utility of L_1

$$L_1 \preceq L_2 \text{ if and only if } u(L_1) \leq u(L_2)$$

More formally, we assume a finite number of states of the world such that the jth state happens with the probability p_j, and we let the consumption C be a random variable taking the values c^j for $j = 1, .., k$ (see Appendix (E.2) for details). To get a Von-Neumann Morgenstern (VNM) utility there must exist $v : \mathbb{R}_+ \to \mathbb{R}$ such that

$$u(P) = \int_0^\infty v(x)dP(x)$$

In the special case where P_C is a discrete sum, the VNM utility simplifies to

$$u(P) = \sum_{j=1}^{k} p_j v(c^j)$$

Hence, the criterion becomes that of maximising the expected value of the utility of consumption where $u(P) = E[v(C)]$, with

$$E[v(C)] =< v(C) >= \sum_{j=1}^{k} p_j v(c^j)$$

A variety of generalized expected utility theories have arisen, most of which drop or relax the independence axiom. One of the most common uses of a utility function, especially in economics, is the utility of money. The utility function for money is a nonlinear function that is bounded and asymmetric about the origin. The utility function is concave in the positive region, reflecting the phenomenon of diminishing marginal utility. The boundedness reflects the fact that beyond a certain point money ceases being useful at all, as the size of any economy at any point in time is itself bounded. The asymmetry about the origin reflects the fact that gaining and losing money can have radically different implications both for individuals and businesses. The nonlinearity of the utility function for money has profound implications in decision making processes: in situations where outcomes of choices influence utility through gains or losses of money, which are the norm in most business settings, the optimal choice for a given decision depends on the possible outcomes of all other decisions in the same time-period.

1.6.2 Preferences

In traditional decision theory, people form beliefs, or make judgments about some probabilities by assigning values, or utilities, to outcomes (see Arrow [1953] [1971]). Attitudes toward risk (risk aversion) determine what choices people make among the various opportunities that exist, given their beliefs. We saw in Section (1.6.1) that people calculate the expected utility of each gamble or lottery as the probability weighted sum of utility outcomes, then choose the gamble with the highest expected utility (see Von Neumann et al. [1944] and Friedman et al. [1948]). More formaly, one can view decision making under risk as a choice between prospects or gambles (lotteries). A prospect $(x_1, p_1; ...; x_n, p_n)$ is a contract yielding outcome x_i with probability p_i where $\sum_{i=1}^n p_i = 1$. To simplify notation, omitting null outcomes, we use (x, p) to denote the prospect $(x, p; 0, (1-p))$ that yields x with probability p and 0 with probability $(1-p)$. It is equivalent to a simple lottery. The riskless prospect that yields x with certainty is denoted (x). The application of expected utility theory is based on

1. Expectation: $u(x_1, p_1; ...; x_n, p_n) =< u(X) >= p_1 u(x_1) + ... + p_n u(x_n)$ where $u(.)$ is the utility function of a prospect. The utilities of outcomes are weighted by their probabilities.

2. Asset integration: $(x_1, p_1; ...; x_n, p_n)$ is acceptable at asset position w if and only if $u(w + x_1, p_1; ...; w + x_n, p_n) \geq u(w)$. A prospect is acceptable if the utility resulting from integrating the prospect with one's assets exceeds the utility of those assets alone. The domain of the utility function is final states rather than gains or losses.

3. Risk aversion: u is concave, that is, $u'' < 0$.

However, experimental work found substantial violations of this rational model (see Allais [1953]), leading to the development of various alternative theories, such as the prospect theory (PT). Kahneman et al. [1979] described several classes of choice problems in which preferences systematically violate the axioms of expected utility theory,

and developed an intentionally positive, or descriptive, model of preferences that would best characterise the deviations of actual choices from the normative expected utility model. Their experimental studies revealed that people

- overweight outcomes that are considered certain, relative to outcomes which are merely probable (certainty effect).
- care more about gains and losses (changes in wealth) than about overall wealth
- exhibit loss aversion and can be risk seeking when facing the possibility of loss (reflection effect).
- overweight low probability events.

The reflection effect implies that risk aversion in the positive domain is accompanied by risk seeking in the negative domain. Further, outcomes which are obtained with certainty are overweighted relative to uncertain outcomes. In the positive domain, the certainty effect contributes to a risk averse preference for a sure gain over a larger gain that is merely probable. In the negative domain, the same effect leads to a risk seeking preference for a loss that is merely probable over a smaller loss that is certain. The following example (Problem 11, 12) illustrates these findings.

1. suppose that you are paid $\$1000$ to participate in a gamble that presents you with the following further choices
   (a) a sure $\$500$ gain
   (b) a 50% chance of winning $\$1000$, a 50% chance of $\$0$
2. suppose that you are paid $\$2000$ to participate in a gamble that presents you with the following further choices
   (a) a sure $\$500$ loss
   (b) a 50% chance of winning $\$1000$, a 50% chance of $\$0$

The typical strategies chosen are $(1a)$ (certainty) and $(2b)$ (risk seeking), even though the total wealth outcomes in $(1a)$ and $(2a)$ are identical, and that in $(1b)$ and $(2b)$ are likewise identical. People act as if they are risk averse when only gains are involved but become risk seeking when facing the possibility of loss and they view the initial bonus and the gamble separately (isolation effect). The experimental studies showed that the intuitive notion of risk is not adequately captured by the assumed concavity of the utility function for wealth. Prospect theory (PT) distinguishes two phases in the choice process, an early phase of editing and a subsequent phase of evaluation. The major operations of the editing phase are coding, combination, segregation, and cancellation. The edited prospects are then evaluated and the prospect of highest value is chosen. As the editing operations facilitate the task of decision, it can explain many anomalies of preference. For example, the inconsistencies associated with the isolation effect result from the cancellation of common components. In PT, people maximise the weighted sum of values (utilities) where weights are not probability themselves (probabilities are rational) but their nonlinear transformation.

- whether an outcome is seen as a gain or a loss (relative to some neutral reference point) determines the value (utility) of a dollar. Its value depends on the context (coding). The reference point can be affected by the formulation of the offered prospects, and by the expectations of the decision maker.
- the value function is kinked at the origin (reference point) where the steeper slope below zero implies loss aversion. Many studies show that losses hurt twice to two and a half times as much as same-sized gains satisfy. In general, the value function has concave (convex) shape to the right (left) of the reference point, implying risk aversion among gambles involving only gains but risk seeking among gambles involving only losses.
- overweighting low-probability events is the main feature of the probability weighting function, explaining the simultaneous demand for both lotteries and insurance. Such overweighting of small probabilities can be strong enough to reverse the sign of risk appetite in the value function. Lotteries offering a small chance of very large gains can induce risk seeking, despite a general tendency to risk aversion when gambles only involve gains.

To summarise, the main idea was to assume that values are attached to changes rather than to final states, and that decision weights do not coincide with stated probabilities, leading to inconsistencies, intransitivities, and violations of dominance. More formally, the overall value of an edited prospect, denoted V, is expressed in terms of two scales π and v, where the former associates with each probability p a decision weight $\pi(p)$ reflecting the impact of p on the over-all value of the prospect, and the latter assigns to each outcome x a number $v(x)$ reflecting the subjective value of that outcome. Note, v measures the value of deviations from that reference point (gains and losses). Kahneman et al. [1979] considered simple prospects $(x, p; y, q)$ with at most two non-zero outcomes, where one receives x with probability p, y with probability q, and nothing with probability $1 - p - q$, with $p + q \leq 1$. An offered prospect is strictly positive if its outcomes are all positive, if $x, y > 0$ and $p + q = 1$, and it is strictly negative if its outcomes are all negative. A prospect is regular if it is neither strictly positive nor strictly negative (either $p + q < 0$, or $x \geq 0 \geq y$, or $x \leq 0 \leq y$). If $(x, p; y, q)$ is a regular prospect, then the value is given by

$$V(x, p; y, q) = \pi(p)v(x) + \pi(q)v(y)$$

where $v(0) = 0$, $\pi(0) = 0$, and $\pi(1) = 1$. The two scales coincide for sure prospects, where $V(x, 1) = V(x) = v(x)$. In that setting, the expectation principle (1) of the expected utility theory is relaxed since π is not a probability measure. The evaluation of strictly positive and strictly negative prospects is slightly modified. If $p+q = 1$, and either $x > y > 0$, or $x < y < 0$, then

$$V(x, p; y, q) = v(y) + \pi(p)(v(x) - v(y))$$

since $\pi(p) + \pi(1 - p) = 1$. Markowitz [1952b] is at the origin of the idea that utility be defined on gains and losses rather than on final asset positions. He proposed a utility function which has convex and concave regions in both the positive and the negative domains. However, in PT, the carriers of value are changes in wealth or welfare, rather than final states, which is consistent with basic principles of perception and judgment. Value should be treated as a function in two arguments, the asset position that serves as reference point, and the magnitude of the change (positive or negative) from that reference point. Kahneman et al. [1979] assumed that the value function for changes of wealth is normally concave above the reference point ($v''(x) < 0$, for $x > 0$) and often convex below it ($v''(x) > 0$, for $x < 0$). Put another way, PT postulates the leaning S-shape of the value function. However, the actual scaling is considerably more complicated than in utility theory, because of the introduction of decision weights. They measure the impact of events on the desirability of prospects, and should not be interpreted as measures of degree or belief. Note, $\pi(p) = p$ only if the expectation principle holds, but not otherwise. In general, π is an increasing function of p, with $\pi(0) = 0$ and $\pi(1) = 1$. In the case of small values of p, π is a subadditive function of p, that is, $\pi(rp) > r\pi(p)$ for $0 < r < 1$, and that very low probabilities are generally overweighted, that is, $\pi(p) > p$ for small p. In general, for $0 < p < 1$ we get $\pi(p) + \pi(1 - p) < 1$ called subcertainty. Thus, for a fixed ratio of probabilities, the ratio of the corresponding decision weights is closer to unity when the probabilities are low than when they are high. This holds if and only if $\log \pi$ is a convex function of $\log p$. The slope of π in the interval $(0, 1)$ can be viewed as a measure of the sensitivity of preferences to changes in probability. These properties entail that π is relatively shallow in the open interval and changes abruptly near the end-points where $\pi(0) = 0$ and $\pi(1) = 1$. Because people are limited in their ability to comprehend and evaluate extreme probabilities, highly unlikely events are either ignored or overweighted, and the difference between high probability and certainty is either neglected or exaggerated. Consequently, π is not well-behaved near the end-points (sharp drops, discontinuities).

1.6.3 Discussion

Note, most applications of the PT theory have been concerned with monetary outcomes. We saw that one consequence of the reflection effect is that risk aversion in the positive domain is accompanied by risk seeking in the negative domain. That is, the preference between negative prospects is the mirror image of the preference between positive prospects. Denoting $>$ the prevalent preference, in Problem 3 we get $(3000) > (4000, 0.8)$ with 80% against 20%, and in Problem $3'$ we get $(-4000, 0.8) > (-3000)$ with 92% against 8%. The majority of subjects were willing to accept a risk of 0.8 to lose 4000, in preference to a sure loss of 3000, although the gamble has a lower expected value.

These Problems demonstrate that outcomes which are obtained with certainty are overweighted relative to uncertain outcomes. The same psychological principle, the overweighting of certainty, favours risk aversion in the domain of gains and risk seeking in the domain of losses. Referencing Markowitz [1959] and Tobin [1958], the authors postulate that the aversion for uncertainty or variability is the explanation of the certainty effect in the rational theory. That is, people prefer prospects with high expected value and small variance (high Sharpe ratio). For instance, (3000) is chosen over $(4000, 0.8)$ despite its lower expected value because it has no variance. They argued that the difference in variance between $(3000, 0.25)$ and $(4000, 0.20)$ may be insufficient to overcome the difference in expected value. They further postulate that since (-3000) has both higher expected value and lower variance than $(-4000, .80)$, the sure loss should be preferred, contrary to the data. They concluded that their data was incompatible with the notion that certainty is generally desirable and that certainty increases the aversiveness of losses as well as the desirability of gains. In all these Problems, there is no notion of volatility so that one can not define preferences in terms of measures such as the Sharpe ratio, except in the case of certainty events. In problem 3, the prospect (3000) has an infinite Sharpe ratio $(M_{SR} = \infty)$, while in Problem $3^{'}$ the prospect (-3000) has a negative infinite Sharpe ratio $(M_{SR} = -\infty)$ contradicting the founding that (-3000) should be preferred over $(-4000, .80)$.

Other research proposed an alternative foundation based on salience while using standard risk preferences (see Bordalo et al. [2010]). Decision makers overweight the likelihood of salient states (where lotteries have extreme, contrasting payoffs), explaining both the reflecting shape of the value function and the overweighting of low probability events. Note, one feature of PT preferences is that people derive decision utility from the gains and losses of a single trade. To obtain testable predictions, PT must be combined with other assumptions such as narrow framing or the house money effect. The former stipulates that ignoring the rest of wealth implies narrow framing (analysing problems in a too isolated framework). For instance, focusing too much on asset specific risks (volatility, default) and ignoring correlation effects. One important aspect of framing is the selection of a reference point as the benchmark for comparisons (doing nothing, one's current asset position). The latter stipulates that gamblers tend to become less loss averse and more willing to take risks when they are ahead (playing with house money). Risk preferences in a sequence of gambles depend on how prior gains and losses influence loss aversion over time (more aggressive risk taking following successful trading, and cautiousness following losses). Further, people dislike vague uncertainty (ambiguity) more than objective uncertainty. While risky choices always involve the possibility of adverse outcomes, some outcomes are more likely to trigger regret than others. Hence, we might want to minimise regret (when hanging on to losers) which is also a motivation for diversification. At last, moods and feelings may be the most plausible explanations for empirical observations that average stock returns tend to be higher or lower near major events (holidays, weather, lunar cycles).

1.6.4 Some critics

Kahneman et al. [1979] considered a very simplistic approach that can be interpreted (in the positive domain) as a special case of the model of Von Neumann Morgenstern with a single agent, one period of time, a single good of consumption and a finite number of states of the world, described in Appendix (E.2). However, in their settings, there is no notion of good of consumption, and as a result, no objective of maximising wealth over time via optimal consumption. The idea being to compute the expected value of the utility function rather than to maximise that expected value over consumption. Consequently, their class of choice problems is to be related to the St Petersburg paradox described in Section (1.2.1). Hence, all the critics addressed in Section (1.2.3) to Bernoulli's expected utility theory apply. That is, it is meaningless to assign a probability to a single event, as the event has to be embedded within other similar events. The main difference with the EUT being that the decision weights $\pi(p)$ do not coincide with stated probabilities, relaxing the expectation principle of the EUT, since π is no-longer a probability measure. Recognising that EUT can represent risk preferences, but is unable to recommend appropriate levels of risk (see Section (1.2.4)), PT, via the modification of the historical probability, is an artifact to include the missing notion of risk premia in the EUT. It is a non-mathematical approach where the choices of the decision weights have to be related to the minimisation over martingale measures introduced in the utility indifference pricing theory discussed in Section (1.2.4). Obviously the previous example (Problem 11, 12) is a free lunch and one can not use the results from

this experimental work to elaborate a pricing theory. However, the main idea of behavioural finance is to recognise that rationality requires complex calculations, and that when facing a situation where the decision process must be instantaneous investors exhibit bounded rationality. Put another way, investors can not by themselves value the fair price of uncertain future outcomes, and requires shortcuts and heuristic simplifications with an early phase of editing and a subsequent phase of evaluation.

1.7 Predictability of financial markets

1.7.1 The martingale theory of asset prices

As told by Mandelbrot [1982], the question of the predictability of financial markets is an old one, as financial newspapers have always presented analysis of charts claiming that they could predict the future from the geometry of those charts. However, as early as 1900, Bachelier [1900] asserted that successive price changes were statistically independent, implying that charting was useless. Weakening that statement, he stated that every price follows a martingale stochastic process, leading to the concept of perfect market (for the definition of independent processes and martingales see Appendix (B.3)). That is, everything in its past has been discounted fully for the definition of independent processes and martingales. Bachelier introduced an even weaker statement, the notion of efficient market where imperfections remain only as long as they are smaller than transaction costs. A more specific assertion by Bachelier is that any competitive price follows, in the first approximation, a one-dimensional Brownian motion. Since the thesis of Bachelier, the option pricing theory (see Sections (1.2.4) and (1.3.2)) developed around the martingale theory of dynamic price processes, stating that discounted traded assets are martingales under the appropriate probability measure. In a martingale, the expected value of a process at future dates is its current value. If after the appropriate discounting (by taking into account the time value of money and risk), all price processes behave as martingales, the best forecast of future prices is present prices. That is, prices might have different distributions, but the conditional mean, after appropriate discounting, is the present price. Hence, under the appropriate probability measure, discounted prices are not exponentially diverging. In principle, it is satisfied by any price process that precludes arbitrage opportunities. Delbaen et al. [1994] proved that an arbitrage opportunity exists if a price process, P, is not a semimartingale. Hence, an important question in the quantitative analysis of financial data is therefore to check whether an observed process is a semimartingale. However, the discount factors taking risk into account, also called kernels, are in general stochastic making the theory difficult to validate.

Estimating the Hurst exponent (see Hurst [1951]) for a data set provides a measure of whether the data is a pure white noise random process or has underlying trends (see details in Section (10.1)). For instance, processes that we might assume are purely white noise sometimes turn out to exhibit Hurst exponent statistics for long memory processes. In practice, asset prices have dependence (autocorrelation) where the change at time t has some dependence on the change at time $t-1$, so that the Brownian motion is a poor representation of financial data. Actual stock returns, especially daily returns, do not have a normal distribution as the curve of the distribution exhibits fatter tails (called the stylised facts of asset distribution). One approach to validate the theory is to introduce a test about the Hurst coefficient H of a fractional Brownian motion (fBm). A fBm is an example for a stochastic process that is not a semimartingale except in the case of a Hurst coefficient H equal to $\frac{1}{2}$. Hence, a financial market model with a price process, P, that is assumed to be a fBm with $H \neq \frac{1}{2}$ implies an arbitrage opportunity. Rogers [1997] provided a direct construction of a trading strategy producing arbitrage in this situation.

Many applied areas of financial economics such as option pricing theory (see Black et al. [1973]) and portfolio theory (see Markowitz [1952] and [1959]) followed Bachelier's assumption of normally distributed returns. The justification for this assumption is provided by the law of large numbers stating that if price changes at the smallest unit of time are independently and identically distributed (i.i.d.) random numbers, returns over longer intervals can be seen as the sum of a large number of such i.i.d. observations, and, irrespective of the distribution of their summands, should under some weak additional assumptions converge to the normal distribution. While this seemed plausible

and the resulting Gaussian distribution would also come very handy for many applied purposes, Mandelbrot [1963] was the first to demonstrate that empirical data are distinctly non-Gaussian, exhibiting excess kurtosis and higher probability mass in the center and in their tails than the normal distribution. Given sufficiently long record of stock market, foreign exchange or other financial data, the Gaussian distribution can always be rejected with statistical significance beyond all usual boundaries, and the observed largest historical price changes would be so unlikely under the normal law that one would have to wait for horizons beyond at least the history of stock markets to observe them occur with non-negligible probability.

1.7.2 The efficient market hypothesis

From a macroeconomic perspective, it is often assumed that the economy is the superposition of different economic cycles ranging from short to long periods. They have first been studied by Clément Juglar in the 19th-Century to prevent France from being hit by repetitive crises. Then, many economists have been interested in these phenomena such as Mitchell [1927], Kondratiev [1925] or Schumpeter [1927], to name but a few. Nowadays, many new classical economists such as Kydland (Nobel Prize in 1995), Prescott (Nobel Prize in 2004), Sargent (Nobel Prize in 2011) are still working on this area. As a consequence, it is widely believed that equity prices react sensitively to the developments of these macroeconomic fundamentals. That is, the changes in the current fundamental value of the firm will depend upon the present value of the future earnings, explaining the behaviour of stock markets in the long-run. However, some economists such as Fama [1965a] [1970] demonstrated that stock prices were extremely difficult to predict in the short run, with new information quickly incorporated into prices. Even though Bachelier is at the origin of using statistical methods to analyse returns, his work was largely ignored and forgotten until the late 1940s where the basis of the efficient market hypothesis (EMH) was collected in a book by Cootner [1964]. The book presents the rationale for what was to be formalised as the EMH by Fama in the 1960s. Originally, during the 1920s through the 1940s, on one hand the Fundamentalists assumed investors to be rational, in order for value to reassert itself, and on the other hand the Technicians assumed markets were driven by emotions. In the 1950s the Quants made an appeal for widespread use of statistical analysis (see Roberts [1964]). At the same time, Osborn [1964] formalised the claim that stock prices follow a random walk. Similarly, Samuelson [1965] postulated that properly anticipated prices fluctuate randomly. Later, Malkiel [1973] stated that the past movement or direction of the price of a stock or overall market could not be used to predict its future movements, leading to the Random Walk Theory (RWT). Note, in his conclusion (Assumption 7), Osborn states that since price changes are independent (random walk), we expect the distribution of changes to be normal, with a stable mean and finite variance. This is a result of the Central Limit Theorem stating that a sample of i.i.d. random variables will be normally distributed as the sample gets larger. This postulate relies on the fact that capital markets are large systems with a large number of degrees of freedom (investors or agents), so that current prices must reflect the information everyone already has. Hence, investors value stocks based on their expected value (expected return), which is the probability weighted average of possible returns. It is assumed that investors set their subjective probabilities in a rational and unbiased manner. Consequently, if we can not beat the market, the best investment strategy we can apply is buy-and-hold where an investor buys stocks and hold them for a long period of time, regardless of market fluctuations.

Fama [1970] formalised the concept of EMH by presenting three basic models which states that the market is a martingale model and a random walk model, or a fair game model. The main implications of the EMH are

- Homogeneity of investors based on their rationality: if all investors are rational and have the access to the same information, they necessarily arrive at the same expectations and are therefore homogeneous.
- Normal distribution of returns: random walk can be represented by $AR(1)$ process in the form of $P_t = P_{t-1}+\epsilon_t$ implying that $r_t = P_t - P_{t-1} = \epsilon_t$ where $\epsilon_t \sim N(0, \sigma)$ is independent normally distributed variable.
- Standard deviation as a measure of volatility and thus risk : since returns are normally distributed, it implies that standard deviation is stable and finite and thus is a good measure of volatility.

- Tradeoff between risk and return: since standard deviation is stable and finite, there is a relationship between risk and return based on non-satiation and risk-aversion of the investors.
- Unpredictability of future returns: since returns follow random walk, the known information is already incorporated in the prices and thus their prediction is not possible.

While the EMH does not require independence through time or accept only i.i.d. observations, the random walk does. That is, if returns are random, the markets are efficient, but the converse may not be true. Over time, a semistrong version of the EMH was accepted by the investment community which states that markets are efficient because prices reflect all public information. In a weak form efficient market, the price changes are independent and may be a random walk. It can be restated in terms of information sets such that, in the weak form, only historical prices of stocks are available for current price formation, while the semi-strong form broadens the information set by all publicly available information, and the strong form includes insider information into the information set. Market is then said to be weakly efficient when investors can not reach above-average risk-adjusted returns based on historical prices and similarly for the other forms. While the efficient market hypothesis (EMH) implies that all available information is reflected in current market prices, leading to future returns to be unpredictable, this assumption has been rejected both by practitioners and academics. To test the EMH requires understanding the restrictions it imposes on probabilistic models of returns. In theory, the EMH is embodied in the martingale theory of dynamic price process (see Section 1.7.1) which can be hardly, if at all, tested in practice since the full reflection of information in prices is hard to define [2]. LeRoy [1976] criticised all presented definitions of EMH and argued that they were tautologies and therefore impossible to test. Further, some authors raised the problem of joint-hypothesis which state that even when the potential inefficiency of the market is uncovered, it can be due to wrongly chosen asset-pricing model. It is therefore impossible to reject EMH in general, and one must state under which conditions to test EMH (see Lo [2008]). For instance, we can either follow Fama [1965a] (Fa65) and assume that the market is weakly efficient if it follows a random walk, or we can follow Samuelson [1965] (Sa65) and assume that the market is efficient if it follows a martingale process. Note, the assumption of martingale process is more general than the random walk one and allows for dependence in the process.

1.7.3 Some major critics

For the technical community, this idea of purely random movements of prices was totally rejected. A large number of studies showed that stock prices are too volatile to be explained entirely by fundamentals (see Shiller [1981] and LeRoy et al. [1981]). Even Fama [1965] found that returns were negatively skewed, the tails were fatter, and the peak around the mean was higher than predicted by the normal distribution (leptokurtosis). This was also noted by Sharpe [1970] on annual returns. Similarly, using daily $S\&P$ returns from 1928 till 1990, Turner et al. [1990] found the distributions to be negatively skewed and having leptokurtosis. While any frequency distribution including October 1987 will be negatively skewed with a fat negative tail, earlier studies showed the same phenomenon. Considering quaterly $S\&P$ 500 returns from 1946 till 1999, Friedman et al. [1989] noted that in addition to being leptokurtotic, large movements have more often been crashes than rallies, and significant leptokurtosis appears regardless of the period chosen. Analysing financial futures prices of Treasury Bond, Treasury Note, and Eurodollar contracts, Sterge [1989] found that very large (three or more standard deviations from the norm) price changes could be expected to occur two or three times as often as predicted by normality. Evidence suggests that in the short-run equity prices deviate from their fundamental values, and are also driven by non-fundamental forces (see Chung et al. [1998] and Lee [1998]). This is due to noise and can be driven by irrational expectations such as irrational waves of optimism or pessimism, feedback trading, or other inefficiencies. The facts that stock market returns are not normally distributed weaken statistical analysis such as correlation coefficients and t-statistics as well as the concept of random walk. Nonetheless, over a longer period of time, the deviations from the fundamentals diminish, and stock prices can be compatible with economic theories such as the Present Value Model.

[2] The problem was mentioned by Fama in Fama [1970].

Mandelbrot [1964] postulated that capital market returns follow the family of Stable Paretian distribution with high peaks at the mean, and fat tails. These distributions are characterised by a tendency to have trends and cycles as well as abrupt and discontinuous changes, and can be adjusted for skewness. However, since variance can be infinite (or undefined), if financial returns fall into the Stable Paretian family of distributions, then variance is only stable and finite for the normal distribution. Hence, given that market returns are not normally distributed, volatility was found to be disturbingly unstable. For instance, Turner et al. [1990] found that monthly and quarterly volatility were higher than they should be compared to annual volatility, but daily volatility was lower. Engle [1982] proposed to model volatility as conditional upon its previous level, that is, high volatility levels are followed by more high volatility, while low volatility is followed by more low volatility. This is consistent with the observation by Mandelbrot [1964] that the size of price changes (ignoring the sign) seems to be correlated. Engle [1982] and LeBaron [1990], among others, found supportive evidence of the autoregressive conditional heteroskedastic (ARCH) model family, such that standard deviation is not a standard measure (at least over the short term).

Defining rationality as the ability to value securities on the basis of all available information, and to price them accordingly, we see that the EMH is heavily dependent on rational investors. Even though investors are risk-averse, Kahneman et al. [1979] and Tversky [1990] suggested that when losses are involved, people tend to be risk-seeking. That is, they are more likely to gamble if gambling can minimise their loss (see Section (1.6)). In addition, in practice, markets are not complete, and investors are not always logical. Shiller [1981] showed that investors could be irrational and that assets from stocks to housing could develop into bubbles. He concluded that rational models of the stock market, in which stock prices reflect rational expectations of future payouts, were in error. He argued that a clever combination of logic, statistics, and data implied that stock markets were, instead, prone to irrational exuberance. Following from these studies and from the paper by DeBondt et al. [1985], behavioural finance developed where people do not recognise, or react to, trends until they are well established. This behaviour is quite different from that of the rational investor, who would immediately adjust to new information, leading to market inefficiencies as all information has not yet been reflected in prices. As explained by Peters [1991-96], if the reaction to information occurs in clumps, and investors ignore information until trends are well in place, and then react in a cumulative fashion to all the information previously ignored, then people react to information in a nonlinear way. This sequence implies that the present is influenced by the past, which is a clear violation of the EMH.

This old debate was partly put to an end by The Royal Swedish Academy of Sciences which attributed the Nobel prize 2013 for economics to both Fama and Shiller. Nonetheless, there is a consensus in empirical finance around the idea that financial assets may exhibit trends or cycles, resulting in persistent inefficiencies in the market that can be exploited (see Keim et al. [1986], Fama et al. [1989]). For instance, Kandel et al. [1996] showed that even a low level of statistical predictability can generate economic significance with abnormal returns attained even if the market is successfully timed only one out of hundred times. One argument put forward is that risk premiums are time varying and depend on business cycle so that returns are related to some slow moving economic variables exhibiting cyclical patterns in accordance with business cycles (see Cochrane [2001]). Another argument states that some agents are not fully rational (for theory of behavioural finance see Barberis et al. [1998], Barberis et al. [2002]), leading prices to underreact in the short run and to overreact at longer horizons.

1.7.4 Contrarian and momentum strategies

Since the seminal article of Fama [1970], a large number of articles have provided substantial evidence that the stock returns and portfolio returns can be predicted from historical data. For instance, Campbell et al. [1997] showed that while the returns of individual stocks do not seem to be autocorrelated, portfolio returns are significantly autocorrelated. The presence of significant cross-autocorrelations lead to more evident predictability at the portfolio level than at the level of individual stocks. The profit generated by trading strategies based on momentum and reversal effect are further evidence of cross-autocorrelations. Lo et al. [1990b] and Lewellen [2002] showed that momentum and reversal effects are not due to significant positive or negative auto-correlation of individual stocks but to cross-autocorrelation effects and other cross-sectional phenomena. The empirical evidences challenged the paradigm of the

weak form efficient market hypothesis (EMH) putting into questions the well accepted capital asset pricing model (CAPM).

The two most popular strategies that emerged from the literature are the contrarian and the momentum strategies. A contrarian strategy takes advantage of the negative autocorrelation of asset returns and is constructed by taking a long position in stocks performing badly in the past and shorting stocks performing well in the past. In contrast, a momentum strategy is based on short selling past losers and buying past winners. Empirical evidence suggested that these two strategies mutually co-exist and that their profitability are international (see Griffin et al. [2003], Chui et al. [2005]). However, although there are sufficient supportive evidence for both strategies, the source and interpretations of the profits is a subject of much debate. Three alternative explanations for such an outcome were proposed:

1. the size effect: with the losers tending to be those stocks with small market value and overreaction being most significant for small firms.
2. time-varying risk: the coefficients of the risk premia of the losers are larger than those of the winners in the period after the formation of the portfolios.
3. market microstructure related effect: part of the return reversal is due to bid-ask biases, illiquidity, etc.

Although contrarian strategies (buying past losers and selling past winners) have received a lot of attention in the early literature on market efficiency, recent literature focused on relative strength strategies that buy past winners and sell past losers. It seems that the authors favoring contrarian strategies focuses on trading strategies based on either very short-term return reversals (1 week or 1 month) or very long-term return reversals (3 to 5 years), while evidence suggests that practitioners using relative strength rules base their selections on price movements over the past 3 to 12 months. As individual tend to overreact to information, De Bondt et al. [1985] [1987] assumed that stock prices also overreact to information, and showed that over 3 to 5-year holding periods stocks performing poorly over the previous 3 to 5 years achieved higher returns than stocks performing well over the same period. When ranked by the previous cumulated returns in the past 3 to 5 years, the losers outperformed the previous winners by nearly 25% in the subsequent 3 to 5 years. Jegadeesh [1990] and Lehmann [1990] provided evidence of shorter-term return reversal. That is, contrarian strategies selecting stocks based on their returns in the previous week or month generated significant abnormal returns. In the case of momentum strategies, Levy [1967] claimed that a trading rule that buys stocks with current prices being substantially higher than their average prices over the past 27 weeks realised significant abnormal returns. Jegadeesh et al. [1993] provided analysis of relative strength trading strategies over 3 to 12-month horizon on the NYSE and AMEX stock, and showed significant profits in the 1965 to 1989 sample period for each of the relative strength strategies examined. For example, a strategy selecting stocks based on their past 6-month returns and holds them for 6 months realised a compounded excess return of 12.01% per year on average. Additional evidence indicates that the profitability of the relative strength strategies are not due to their systematic risk. These results also indicated that the relative strength profits can not be attributed to lead-lag effects resulting from delayed stock price reactions to common factors. However, the evidence is consistent with delayed price reactions to firm specific information. That is, part of the abnormal returns generated in the first year after portfolio formation dissipates in the following two years. Using out-of-sample tests, Jegadeesh et al. [2001] found that momentum profits continued after 1990, indicating that their original findings were not due to a data snooping bias. They suggested that the robustness of momentum returns could be driven by investors' cognitive biases or under reaction to information, such as earning announcements.

Time series momentum or trend is an asset pricing anomaly with effect persisting for about a year and then partially reversing over longer horizons. Hurst et al. [2010] noted that the main driver of many managed futures strategies pursued by CTAs is trend-following or momentum investing. That is, buying assets whose price is rising and selling assets whose price is falling. Rather than focus on the relative returns of securities in the cross-section, time series momentum focuses purely on a security's own past return. These findings are robust across a number of subsamples, look-back periods, and holding periods (see Moskowitz et al [2012]). They argued that time series momentum

directly matches the predictions of many prominent behavioral and rational asset pricing theories. They found that the correlations of time series momentum strategies across asset classes are larger than the correlations of the asset classes themselves, suggesting a stronger common component to time series momentum across different assets than is present among the assets themselves. They decomposed the returns to a time series and cross-sectional momentum strategy to identify the properties of returns that contribute to these patterns and found that positive auto-covariance in futures contracts' returns drives most of the time series and cross-sectional momentum effects. One explanation is that speculators trade in the same direction as a return shock and reduce their positions as the shock dissipates, whereas hedgers take the opposite side of these trades. In general, spot price changes are mostly driven by information shocks, and they are associated with long-term reversals, consistent with the idea that investors may be over-reacting to information in the spot market. This finding of time series momentum in virtually every instrument challenges the random walk hypothesis on the stock prices. Further, they showed that a diversified portfolio of time series momentum across all assets is remarkably stable and robust, yielding a Sharpe ratio greater than one on an annual basis, or roughly 2.5 times the Sharpe ratio for the equity market portfolio, with little correlation to passive benchmarks in each asset class or a host of standard asset pricing factors. At last, the abnormal returns to time series momentum also do not appear to be compensation for crash risk or tail events. Rather, the return to time series momentum tends to be largest when the stock market's returns are most extreme, performing best when the market experiences large up and down moves. Note, the studies of autocorrelation examine, by definition, return predictability where the length of the look-back period is the same as the holding period over which returns are predicted. This restriction masks significant predictability that is uncovered once look-back periods are allowed to differ from predicted or holding periods. Note, return continuation can be detected implicitly from variance ratios. Also, a significant component of the higher frequency findings in equities is contaminated by market microstructure effects such as stale prices. Focusing on liquid futures instead of individual stocks and looking at lower frequency data mitigates many of these issues (see Ahn et al. [2003]).

Recently, Baltas et al. [2012a] extended existing studies of futures time-series momentum strategies in three dimensions (time-series, cross-section and trading frequency) and documented strong return continuation patterns across different portfolio rebalancing frequencies with the Sharpe ratio of the momentum portfolios exceeding 1.20. These strategies are typically applied to exchange traded futures contracts which are considered relatively liquid compared to cash equity or bond markets. However, capacity constraints have limited these funds in the past. The larger they get, the more difficult it is to maintain the diversity of their trading books. Baltas et al. rigorously establish a link between CTAs and momentum strategies by showing that time-series momentum strategies have high explanatory power in the time series of CTA returns. They do not find evidence of capacity constraints when looking at momentum strategies in commodities markets only, suggesting that the futures markets are relatively deep and liquid enough to accommodate the trading activity of the CTA industry.

1.7.5 Beyond the EMH

In financial markets, volatility is a measure of price fluctuations of risky assets over time which can not be directly observed and must be estimated via appropriate measures. These measures of volatility show volatility clustering, asymmetry and mean reversion, comovements of volatilities across assets and financial markets, stronger correlation of volatility compared to that of raw returns, (semi-) heavy-tails of the distribution of returns, anomalous scaling behaviour, changes in shape of the return distribution over time horizons, leverage effects, asymmetric lead-lag correlation of volatilities, strong seasonality, and some dependence of scaling exponents on market structure. Mandelbrot [1963] showed that the standard Geometric Brownian motion (gBm) proposed by Bachelier was unable to reproduce these stylised facts. In particular, the fat tails and the strong correlation observed in volatility are in sharp contrast to the mild, uncorrelated fluctuations implied by models with Brownian random terms. He presented an alternative description of asset prices constructed on the basis of a scaling assumption. From simple observation, a continuous process can not account for a phenomenon characterised by very sharp discontinuities such as asset prices. When $P(t)$ is a price at time t, then $\log(P(t))$ has the property that its increment over an arbitrary time lag d, that is

$\Delta(d) = \log(P(t+d)) - \log(P(t))$, has a distribution independent of d, except for a scale factor. Hence, in a competitive market no time lag is more special than any other. Under this assumption, typical statistics to summarise data such as sample average to measure location and sample root mean square to measure dispersion have poor descriptive properties. This lead Mandelbrot to assume that the increment $\Delta(d)$ has infinite variance and to conclude that price change is ruled by Levy stable distribution. It was motivated by the fact that in a generalised version of the central limit law dispensing with the assumption of a finite second moment, sums of i.i.d. random variables converge to these more general distributions (where the normal law is a special case of the Levy stable law obtained in the borderline case of a finite second moment). Therefore, the desirable stability property indicates the choice of the Levy stable law which has a shape that, in the standard case of infinite variance, is characterised by fat tails. It is interesting to note that Fama [1963] discussed the Levy stable law applied to market returns and Fama et al. [1971] proposed statistical techniques for estimating the parameters of the Levy distributions. While the investment community accepted variance and standard deviation as the measures of risk, the early founders of the capital market theory (Samuelson, Sharpe, Fama, and others) were well aware of these assumptions and their limitations as they all published work modifying the MPT for non-normal distributions. Through the 1960s and 1970s, empirical evidence continued to accumulate proving the non-normallity of market returns (see Section (1.7.3)). Sharpe [1970] and Fama et al. [1972] published books including sections on needed modifications to standard portfolio theory accounting for the Stable Paretian Hypothesis of Mandelbrot [1964]. As the weak-form EMH became widely accepted more complex applications developed such as the option pricing of Black et al. [1973] and the Arbitrage Pricing Theory (APT) of Ross [1976]. The APT postulates that price changes come from unexpected changes in factors, allowing the structure to handle nonlinear relationships.

In the statistical extreme value theory the extremes and the tail regions of a sample of i.i.d. random variables converge in distribution to one of only three types of limiting laws (see Reiss et al. [1997])

1. exponential decay
2. power-law decay
3. the behaviour of distributions with finite endpoint of their support

Fat tails are often used as a synonym for power-law tails, so that the highest realisations of returns would obey a law like $P(x_t < x) \sim 1 - x^{-\alpha}$ after appropriate normalisation with transformation $x_t = ar_t + b$. Hence, the universe of fat-tailed distributions can be indexed by their tail index α with $\alpha \in (0, \infty)$. Levy stable distributions are characterised by tail indices $\alpha < 2$ (2 characterising the case of the normal distribution). All other distributions with a tail index smaller than 2 converge under summation to the Levy stable law with the same index, while all distributions with an asymptotic tail behaviour with $\alpha > 2$ converge under aggregation to the Gaussian law. Various authors such as Jansen et al. [1991] and Lux [1996] used semi-parametric methods of inference to estimate the tail index without assuming a particular shape of the entire distribution. The outcome of these studies on daily records is a tail index α in the range of 3 to 4 counting as a stylised fact. Using intra-daily data records Dacorogna et al. [2001] confirmed the previous results on daily data giving more weight to the stability of the tail behaviour under time aggregation as predicted by extreme-value theory. As a result, it was then assumed that the unconditional distribution of returns converged toward the Gaussian distribution, but was distinctly different from it at the daily (and higher) frequencies. Hence, the non-normal shape of the distribution motivated the quest for the best non-stable characterisation at intermediate levels of aggregation. A large literature developed on mixtures of normal distributions (see Kon [1984]) as well as on a broad range of generalised distributions (see Fergussen et al. [2006]) leading to the distribution of daily returns close to a Student-t distribution with three degrees of freedom. Note, even though a tail index between 3 and 4 was typically found for stock and foreign exchange markets, some other markets were found to have fatter tails (see Koedijk et al. [1992]).

Even though the limiting laws of extreme value theory apply to samples of i.i.d. random variables it may still be valid for certain deviations from i.i.d. behaviour, but dependency in the time series of return will dramatically slow down convergence leading to a long regime of pre-asymptotic behaviour. While this dependency of long lasting

autocorrelation is subject to debate for raw (signed) returns, it is plainly visible in absolute returns, squared returns, or any other measure of the extent of fluctuations (volatility). For instance, Ausloos et al. [1999] identified such effects on raw returns. With sufficiently long time series, significant autocorrelation can be found for time lags (of daily data) up to a few years. This positive feedback effect, called volatility clustering or turbulent (tranquil) periods, are more likely followed by still turbulent (tranquil) periods than vice versa. Lo [1991] proposed rigorous statistical test for long term dependence with more or less success on finding deviations from the null hypothesis of short memory for raw asset returns, but strongly significant evidence of long memory in squared or absolute returns. In general, short memory comes along with exponential decay of the autocorrelation, while one speaks of long memory if the decay follows a power-law. Evidence of the latter type of behaviour both on rate of returns and volatility accumulated over time. Lobato et al. [1998] claimed that such long-range memory in volatility measures was a universal stylised fact of financial markets. Note, this long-range memory effect applies differently on the foreign exchange markets and the stock markets (see Genacy et al. [2001a]). Further, LeBaron [1992] showed that, due to the leverage effect, stock markets exhibited correlation between volatility and raw returns.

The hyperbolic decay of the unconditional pdf together with the hyperbolic decay of the autocorrelations of many measures of volatility (squared, absolute returns) fall into the category of scaling laws in the natural sciences. The identification of such universal scaling laws in financial markets spawned the interest of natural scientists to further explore the behaviour of financial data and to develop models explaining these characteristics. From this line of research, multifractality, multi-scaling or anomalous scaling emerged gradually over the 90s as a more subtle characteristic of financial data, motivating the adaptation of known generating mechanisms for multifractal processes from the natural sciences in empirical finance. The background of these models is the theory of multifractal measures originally developed by Mandelbrot [1974] in order to model turbulent flows. The formal analysis of such measures and processes, called multifractal formalism, was developed by Frisch et al. [1985], Mandelbrot [1989], and Evertz et al. [1992], among others. Mandelbrot et al. [1997] introduced the concept of multifractality in finance by adapting an earlier asset pricing framework of Mandelbrot [1974]. Subsequent literature moved from the more combinatorial style of the multifractal model of asset returns (MMAR) to iterative, causal models of similar design principles such as the Markov-switching multifractal (MSM) model proposed by Calvet et al. [2004] and the multifractal random walk (MRW) by Bacry et al. [2001] constituting the second generation of multifractal models.

Mantegna et al. [2000] and Bouchaud et al. [2000] considered econophysics to study the herd behaviour of financial markets via return fluctuations, leading to a better understanding of the scaling properties based on methods and approaches in scientific fields. To measure the multifractals of dynamical dissipative systems, the generalised dimension and the spectrum have effectively been used to calculate the trajectory of chaotic attractors that may be classified by the type and number of the unstable periodic orbits. Even though a time series can be tested for correlation in many different ways (see Taqqu et al. [1995]), some attempts at computing these statistical quantities emerged from the box-counting method, while others extended the R/S analysis. Using detrended fluctuation analysis (DFA) or detrended moving average (DMA) to analyse asset returns on different markets, various authors observed that the Hurst exponent would change over time indicating multifractal process (see Costa et al. [2003], Kim et al. [2004]). Then methods for the multifractal characterisation of nonstationary time series were developed based on the generalisation of DFA, such as the MFDFA by Kantelhardt et al. [2002]. Consequently, the multifractal properties as a measure of efficiency (or inefficiency) of financial markets were extensively studied in stock market indices, foreign exchange, commodities, traded volume and interest rates (see Matia et al. [2003], Ho et al. [2004], Moyano et al. [2006], Zunino et al. [2008], Stosic et al. [2014]). It was also shown that observable in the dynamics of financial markets have a richer multifractality for emerging markets than mature one. As a rule, the presence of multifractality signalises time series exhibiting a complex behaviour with long-range time correlations manifested on different intrinsic time scales. Considering an artificial multifractal process and daily records of the $S\&P$ 500 index gathered over a period of 50 years, and using multifractal detrended fluctuation analysis (MFDFA) and multifractal diffusion entropy analysis (MFDEA), Jizba et al. [2012] showed that the latter posses highly nonlinear, and long-ranged, interactions which is the manifestation of a number of interlocked driving dynamics operating at different time scales each with its own scaling function. Such a behaviour typically points to the presence of recurrent economic

cycles, crises, large fluctuations (spikes or sudden jumps), and other non-linear phenomena that are out of reach of more conventional multivariate methods (see Mantegna et al. [2000]).

1.7.6 Risk premia and excess returns

1.7.6.1 Risk premia in option prices

The Black-Scholes model [1973] for pricing European options assumes a continuous-time economy where trading can take place continuously with no differences between lending and borrowing rates, no taxes and short-sale constraints. Investors require no compensation for taking risk, and can construct a self-financing riskless hedge which must be continuously adjusted as the asset price changes over time. In that model, the volatility is a parameter quantifying the risk associated to the returns of the underlying asset, and it is the only unknown variable. However, since the market crash of October 1987, options with different strikes and expirations exhibit different Black-Scholes implied volatilities (IV). The out-of-the-money (OTM) put prices have been viewed as an insurance product against substantial downward movements of the stock price and have been overpriced relative to OTM calls that will pay off only if the market rises substantially. As a result, the implicit distribution inferred from option prices is substantially negatively skewed compared to the lognormal distribution inferred from the Black-Scholes model. That is, given the Black-Scholes assumptions of lognormally distributed returns, the market assumes a higher return than the risk-free rate in the tails of the distributions.

Market efficiency states that in a free market all available information about an asset is already included in its price so that there is no good buy. However, in financial markets, perfect hedges do not exist and option prices induce market risks called gamma risk and vega risk whose order of magnitude is much larger than market corrections such as transaction costs and other imperfections. In general, these risks can not be hedged away even in continuous time trading, and hedging becomes approximating a target payoff with a trading strategy. The value of the option price is thus the cost of the hedging strategy plus a risk premium required by the seller to cover his residual risk which is unhedgeable. The no-arbitrage pricing theory tells us about the first component of the option value while the second component depends on the preferences of investors. Thus, the unhedgeable portion is a risky asset and one must decide how much he is willing to pay for taking the risk. The no-arbitrage argument implies a unique price for that extra risk called the market price of risk. Hence, when pricing in incomplete market, the market price of risk enter explicitly the pricing equation leading to a distribution of prices rather than a single price such that one consider bounds. Therefore, one can either simply ignore the risk premium associated to a discontinuity in the underlying, or one can choose any equivalent martingale measure as a self-consistent pricing rule but in that case the option price does not correspond to the cost of a specific hedging strategy. Hence, one should first discuss a hedging strategy and then derive a valuation for the options in terms of the cost of hedging plus a risk premium.

Incorporating market incompleteness, alternative explanations for the divergence between the risk-neutral distributions and observed returns include peso problems, risk premia and option mispricing but no consensus has yet been reached. For instance, in the analysis performed by Britten-Jones et al. [2000] the bias may not be due to model misspecification or measurement errors, but to the way the market prices volatility risk. Similarly Duarte et al. [2007] documented strong evidence of conditional risk premium that varies positively with the overall level of market volatility. Their results indicate that the bias induced by censoring options that do not satisfy arbitrage bounds can be large, possibly resulting in biases in expected returns as large as several percentage points per day. Option investors are willing to pay more to purchase options as hedges under adverse market conditions, which is indicative of a negative volatility risk premium. These results are consistent with the existence of time-varying risk premiums and volatility feedback, but there may be other factors driving the results. Nonetheless, negative market price of volatility risk is the key premium in explaining the noticeable differences between implied volatility and realised volatility in the equity market. Thus, research now proposes the volatility risk premium as a possible explanation (Lin et al. [2009] , Bakshi et al. [2003] found supportive evidence of a negative market volatility risk premium).

1.7.6.2 The existence of excess returns

To capture the extra returns embedded in the tails of the market distributions, the literature focused on adding stochastic processes to the diffusion coefficient of the asset prices or even jumps to the asset prices as the drift was forced to match the risk-free rate. The notion that equity returns exhibit stochastic volatility is well documented in the literature, and evidence indicates the existence of a negative volatility risk premium in the options market (see Bakshi et al. [2003]). CAPM suggests that the only common risk factor relevant to the pricing of any asset is its covariance with the market portfolio, making beta the right measure of risk. However, excess returns on the traded index options and on the market portfolio explain this variation, implying that options are non-redundant securities. As a result, Detemple et al. [1991] argued that there is a general interaction between the returns of risky assets and the returns of options, implying that option returns should help explain stock returns. That is, option returns should appear as factors in explaining the cross section of asset returns. For example, Bekaert et al. [2000] investigated the leverage effect and the time-varying risk premium explanations of the asymmetric volatility phenomenon at both the market and firm level. They found covariance asymmetry to be the main mechanism behind the asymmetry for the high and medium leverage portfolios. Negative shocks increase conditional covariances substantially, whereas positive shocks have a mixed impact on conditional covariances. While the above evidence indicates that volatility risk is priced in options market, Arisoy et al. [2006] used straddle returns (volatility trade) on the $S\&P$ 500 index and showed that it is also priced in securities markets.

Chapter 2

Introduction to asset management

2.1 Portfolio management

2.1.1 Defining portfolio management

A financial portfolio consists in a group of financial assets, also called securities or investments, such as stocks, bonds, futures, or groups of these investment vehicles referred as exchange-traded-funds (ETFs). The building of financial portfolio constitutes a well known problem in financial markets requiring a rigorous analysis in order to select the most profitable assets. Portfolio construction consists of two interrelated tasks

1. an asset allocation task for choosing how to allocate the investor's wealth between a risk-free security and a set of N risky securities, and
2. a risky portfolio construction task for choosing how to distribute wealth among the N risky securities.

Therefore, in order to construct a portfolio, we must define investments objectives by focusing on accepted degree of risk for a given return. Portfolio management is the act of deciding which assets need to be included in the portfolio, how much capital should be allocated to each kind of security, and when to remove a specific investment from the holding portfolio while taking the investor's preferences into account. We can apply two forms of management (see Maginn et al. [2007])

1. Passive management in which the investor concentrates his objective on tracking a market index. This is related to the idea that it is not possible to beat the market index, as stated by the Random Walk Theory (see Section (1.7.2)). A passive strategy aims only at establishing a well diversified portfolio without trying to find under or overvalued stocks.
2. Active management where the main goal of the investor consists in outperforming an investment benchmark index, buying undervalued stocks and selling overvalued ones.

As explained by Jacobs et al. [2006], a typical equity portfolio is constructed and managed relative to an underlying benchmark. Designed to track a benchmark, an indexed equity portfolio is a passive management style with no active returns and residual risk constrained to be close to zero (see Equation (8.2.1)). While the indexed equity portfolio may underperform the benchmark after costs are considered, enhanced indexed portfolios are designed to provide an index-like performance plus some excess return after costs. The latter are allowed to relax the constraint on residual risk by slightly overweighting securities expected to perform well and slightly underweighting securities expected to perform poorly. This active portfolio incurs controlled anticipated residual risk at a level generally not exceeding 2%. Rather than placing hard constraint on the portfolio's residual risk, active equity management seek portfolios

with a natural level of residual risk based on the return opportunities available and consistent with the investor's level of risk tolerance. The aim of most active equity portfolios is to generate attractive risk-adjusted returns (or alpha). While both the portfolio and the benchmark are defined in terms of constituent securities and their percentage weights, active equity portfolios have active weights (differ from their weights in the benchmark) giving rise to active returns measured as the difference between the returns of the actively managed equity portfolio and the returns of its benchmark. In general, an actively managed portfolio overweights the securities expected to perform the benchmark and underweights the securities expected to perform below the benchmark. In a long-only portfolio, while any security can be overweighted to achieve a significant positive active weight, the maximum attainable underweight is equal to the security's weight in the underlying benchmark index which is achieved by not holding any of the security in the portfolio. As the weights of most securities in most benchmarks are very small, there is extremely limited opportunity to profit from underweighting unattractive securities in long-only portfolios. Allowing short-selling by relaxing the long-only constraint gives the investor more flexibility to underweight overvalued stocks and enhance the actively managed portfolio's ability to produce attractive active equity returns. It also reduces the portfolio's market exposure. Greater diversification across underweighted and overweighted opportunities should result in greater consistency of performance relative to the benchmark (see details in Section (7.2)).

An active portfolio management tries to find under or overvalued stocks in order to achieve a significant profit when prices are rising or falling. Both trend measurement and portfolio allocation are part of momentum trading strategies. The former requires the selection of trend filtering techniques which can involve a pool of methods and the need for an aggregation procedure. This can be done through averaging or dynamic model selection. The resulting trend indicator can be used to analyse past data or to forecast future asset returns for a given horizon. The latter requires quantifying the size of each long or short position given a clear investment process. This process should account for the risk entailed by each position given the expected return. In general, individual risks are calculated in relation to asset volatility, while a correlation matrix aggregate those individual risks into a global portfolio risk. Note, rather than considering the correlation of assets one can also consider the correlation of each individual strategy. In any case, the distribution of these risks between assets or strategies remains an open problem. One would like the distribution to account for the individual risks, their correlations, and the expected return of each asset. Wagman [1999] provided a simple framework based on Genetic Programming (GP), which tries to find an optimal portfolio with recurrence to a simple technical analysis indicator, the moving average (MA). The approach starts by generating a set of random portfolios and the GP algorithm tries to converge in an optimal portfolio by using an evaluation function which considers the weight of each asset within the portfolio and the respective degree of satisfaction against the MA indicator, using different period parameters. Similarly, Yan [2003] and Yan et al. [2005] used a GP approach to find an optimal model to classify the stocks within the market. The top stocks adopt long positions while the bottom ones follows short positions. Their model is based on the employment of Fundamental Analysis which consists on studying the underlying forces of the economy to forecast the market development.

In order to outperform a benchmark index by buying and selling properly selected assets, a portfolio manager must detect profitable opportunities. For instance, in the capital asset pricing model (CAPM), or extension to a multi-factor model (APT), the skill of a fund manager relies on accurate forecasts of the expected returns and systematic risk on all risky assets, and on the market portfolio. That is, conditional on the returns on the market portfolio and the risk free asset, and given forecasts of the systematic risks of risky assets, the fund manager must identify those assets presenting promising investment opportunities. As a result, markets need to be predictable in some way in order to apply successfully active management. Fortunately, we saw in Section (1.7.4) that there was substantial evidence showing that market returns and portfolio returns could be predicted from historical data. We also saw that the two most popular strategies in financial markets are the contrarian and the momentum strategies. The literature, which tries to explain the reasons behind why momentum exists, seems to be split into two categories:

1. behavioural explanations (irrational),
2. and market-based explanations (rational).

The literature on behavioural explanations usually focuses around investors under-reacting to information such as news. This under-reaction can be manifested by either not reacting early enough or the actions that they take are insufficiently drastic in order to protect themselves from the volatility of the market. As a result, prices rise or fall for longer than would normally be expected by the market players. Market-based explanations regarding momentum are based around the fact that poor market performance can establish diminishing illiquidity, putting downward pressure on performance (see Smith [2012]). Another market-based explanation for momentum is that an investor's appetite for risk changes over time. When the values of an investor's assets are forced towards their base level of wealth, the investor begins to worry about further losses. This leads the investor to sell, putting downward pressures on prices, further lowering risk appetites and prices. Markets can therefore generate their own momentum when risk appetites fall or liquidity is low.

In a study on the inequality of capital returns, Piketty [2013] stressed the importance of active portfolio management by showing that skilled portfolio managers could generate substantial additional profits. Analysing capital returns from the world's richest persons ranked in Forbes magazine as well as the returns generated from donations to the US universities, he showed that the rate of return was proportional to the size of the initial capital invested. The net annualised rate of return of the wealthiest persons and universities is around $6-7\%$ against $3-4\%$ for the rest. The main reason being that a larger fraction of the capital could be invested in riskier assets, necessitating the services of skilled portfolio managers to identify and select in an optimum way the best portfolio. For instance, with about 30 billion dollars of donations invested in 2010, Harvard university obtained a rate of return of about 10.2% from 1980-2010. On the other hand, for roughly 500 universities out of 850 having a capital smaller or equal to 100 million dollars invested, they obtained a rate of return of about 6.2% from 1980-2010 (5.1% from 1990-2010). Universities investing over 1 billion dollars have 60% or more of their capital invested in risky assets, while for universities investing between 50 and 100 million dollars 25% of the capital is invested in risky assets, and for universities investing less that 50 million dollars only 10% is invested in risky assets. While Harvard university spent about 100 million dollars of management fees, which is 0.3% of 30 billion of dollars, it represents 10% for a university investing 1 billion dollars. Considering that a university pays between 0.5% and 1% of management fees, it would spend 5 million dollars to manage 1 billion dollars. A university such as North Iowa Community College which is investing 11.5 million dollars, would only spend $150,000$ dollars in management fees.

2.1.2 Asset allocation

2.1.2.1 Objectives and methods

While the objective of investing is to increase the purchasing power of capital, the main goal of asset allocation is to improve the risk-reward trade-off in an investment portfolio. As explained by Darst [2003], investors pursue this objective by selecting an appropriate mix of asset classes and underlying investments based on

- the investor's needs and temperament
- the characteristics of risk, return, and correlation coefficients of the assets under consideration in the portfolio
- the financial market outlook

The objective being

1. to increase the overall return from a portfolio for a given degree of risk, or,
2. to reduce the overall risk from the portfolio for a targeted level of return.

For most investors asset allocation often means

1. calculating the rates of return from, standard deviations on, and correlations between, various asset classes

2. running these variables through a mean-variance optimisation program to select asset mixes with different risk-reward profiles
3. analysing and implementing some version of the desired asset allocation in light of the institution's goals, history, preferences, constraints, and other factors

A disciplined asset allocation process tends to proceed in a series of sequential steps

1. investor examines and proposes some assumptions on future expected returns, risk, and the correlation of future returns between asset classes
2. investor selects asset classes that best match his profile and objectives with the maximum expected return for a given level of risk
3. investor establishes a long-term asset allocation policy (Strategic Asset Allocation (SAA)) reflecting the optimal long-term standard around which future asset mixes might be expected to vary
4. investor may decide to implement Tactical Asset Allocation (TAA) decisions against the broad guidelines of the Strategic Asset Allocation
5. investor will periodically rebalance the portfolio of assets, with sensitivity to the tax and transaction cost consequences of such rebalancing, taking account of the SAA framework
6. from time to time, the investor may carefully review the SAA itself to ensure overall appropriateness given current circumstances, frame of mind, the outlook for each of the respective asset classes, and overall expectations for the financial markets

Asset allocation seeks, through diversification, to provide higher returns with lower risk over a sufficiently long time frame and to appropriately compensate the investor for bearing non-diversifiable volatility. Some of the foundations of asset allocation are related to

- the asset - such as the selection of asset classes, the assessment of asset characteristics, the evaluation of the outlook for each asset class
- the market - such as gauging divergence, scenario analysis, risk estimation
- the investor - such as investor circumstances review, models efficacy analysis, application of judgment

While the scope of asset allocation for any investor defines his universe of investment activity, the types of asset allocation are classified according their style, orientation, and inputs and can be combined accordingly

- The style of an asset allocation can be described as conservative, moderate, or aggressive (cash, bonds, equities, derivatives). A conservative style should exhibit lower price volatility (measured by the standard deviation of returns from the portfolio), and generate a greater proportion of its returns in the form of dividend and interest income. An aggressive style may exhibit higher price volatility and generate a greater proportion of its returns in the form of capital gains.
- The orientation type can be described as strategic, tactical, or a blend of the two. A strategic asset allocation (SAA) attempts to establish the best long-term mix of assets for the investor, with relatively less focus on short-term market fluctuations. It helps determine which asset classes to include in the long-term asset mix. Some investors may adopt a primarily tactical approach to asset allocation by viewing the long term as an ongoing series of short term time frames. Others can use TAA to either reinforce or counteract the portfolio's strategic allocation policies. Owing to the price-aware, opportunistic nature of TAA, special forms of tactical risk management can include price alerts, limit and stop-loss orders, simultaneous transaction techniques, and value-at-risk (VaR) models. While SAA allows investors to map out a long-term plan, TAA helps investors to anticipate and respond to significant shifts in asset prices.

- Investors can use different inputs to formulate the percentages of the overall portfolio that they will invest in each asset class. These percentages can be determined with the help of quantitative models, qualitative judgements, or a combination of both. The quantitative approach consists in selecting the asset classes and subclasses for the portfolio, propose assumptions on future expected returns, risk of the asset classes, correlations of future expected returns between each pair of asset classes. Then, portfolio optimisation program can generate a set of possible asset allocations, each with its own level of expected risk and return. As a result, investors can select a series of Efficient Frontier asset allocation showing portfolios with the minimum risk for a given level of expected return. Investors may then decide to set upper and lower percentage limits on the maximum and minimum amounts allowed in the portfolio by imposing constraints on the optimisation. Qualitative asset allocation assesses fundamental measures, valuation measures, psychology and liquidity measures. These assessments, carried out on an absolute basis and relative to long-term historical averages, are often expressed in terms of the number of standard deviations above or below their long-term mean.

2.1.2.2 Active portfolio strategies

The use of predetermined variables to predict asset returns in view of constructing optimum portfolios, has produced new insights into asset pricing models which have been applied on improving existing policies based upon unconditional estimates. Several strategies exist in taking advantage of market predictability in view of generating excess return. Extending Sharpe's CAPM [1964] to account for the presence of pervasive risk, Fama et al. [1992] decomposed portfolio returns into:

- systematic market risk,
- systematic style risk, and,
- specific risk.

As a result, a new classification of active portfolio strategies appears where

- Market Timing or Tactical Asset Allocation (TAA) strategies aim at exploiting evidence of predictability in market factors, while,
- Style Timing or Tactical Style Allocation (TSA) strategies aim at exploiting evidence of predictability in style factors, and,
- Stock Picking (SP) strategies are based on stock specific risk.

As early as the 1970s, Tactical Asset Allocation (TAA) was considered as a way of allocating wealth between two asset classes, typically shifting between stocks and bonds. It is a style timing strategy, that is, a dynamic investment strategy actively adjusting a portfolio's asset allocation in view of improving the risk-adjusted returns of passive management investing. The objectives being to maximise total return on investment, limit risk, and maintain an appropriate degree of portfolio diversification. Systematic TAA use quantitative investment models such as trend following or relative strength techniques, capitalising on momentum, to exploit inefficiencies and produce excess returns. Market timing is another form of asset allocation where investors attempt to time the market by adding funds to or withdrawing funds from the asset class in question according to a periodic schedule, seeking to take advantage of downward or upward price fluctuations. Momentum strategies are examples of market timing. For instance, momentum strategies try to benefit from either market trends or market cycles. Being an investment style based only on the history of past prices, one can identify two types of momentum strategies:

1. On one hand the trend following strategy consisting in buying (or selling) an asset when the estimated price trend is positive (or negative).

2. On the other hand the contrarian (or mean-reverting) strategy consisting in selling (or buying) an asset when the estimated price trend is positive (or negative).

For example, earning momentum strategies involve buying the shares of companies exhibiting strong growth in reported earnings, and selling shares experiencing a slowdown in the rate of growth in earnings. Similarly, price momentum strategies are based on buying shares with increasing prices, and selling shares with declining prices. Note, such momentum based methods involves high rate of portfolio turnover and trading activity, and can be quite risky. Stock Selection criteria or Stock Picking strategies aim at exploiting evidence of predictability in individual stock specific risk. Perhaps one of the most popular stock picking strategies is that of the long/short with the majority of equity managers still favouring this strategy to generate returns. The stock investment or position can be long to benefit from a stock price increase or short to benefit from a stock price decrease, depending on the investor's expectation of how the stock price is going to move. The stock selection criteria may include systematic stock picking methods utilising computer software and/or data. Note, most mutual fund managers actually make discretionary, and sometimes unintended, bets on styles as much as they make bets on stocks. In other words, they perform tactical asset allocation (TAA), tactical style allocation (TSA) and stock picking (SP) at the same time.

2.1.2.3 A review of asset allocation techniques

We present a few allocation techniques among the numerous methodologies developed in the financial literature. For more details see text book by Meucci [2005].

- Equally-weighted is the simplest allocation algorithm, where the same weight is attributed to each strategy. It is used as a benchmark for other allocation methods.
- Inverse volatility consists of weighting the strategies in proportion to the inverse of their volatility, that is, it takes a large exposure to assets with low volatility.
- Minimum variance seeks at building a portfolio such that the overall variance is minimal. If the correlation between all assets in the basket is null, the minimum variance will allocate everything to the lowest volatility asset, resulting in poor diversification.
- Improved minimum variance improves the portfolio's covariance by using a correlation matrix based on the Speaman correlation which is calculated on the ranks of the variables and tends to be a more reliable estimate of correlation.
- Minimum value-at-risk (VaR) seeks to build a portfolio such that the overall VaR is minimal. The marginal distribution of each strategy is measured empirically and the relationship between the strategies is modelled by the Gaussian copula which consider a single correlation coefficient.
- Minimum expected shortfall seeks to build a portfolio such that the overall expected shortfall (average risk above the VaR) is minimal.
- Equity-weighted risk contribution (ERC) seeks to equalise the risk contribution of each strategy. The risk contribution of a strategy is the share of the total portfolio risk due to the strategy represented by the product of the standard deviation and correlation with the portfolio.
- Factor based minimum variance applies the minimum variance method on a covariance matrix reduced to the first three factors of a rolling principal component analysis (PCA).
- Factor based ERC applies the ERC method on a covariance matrix reduced to the first three factors of a rolling PCA.

Optimal portfolios are designed to offer best risk metrics by computing estimates of future covariances and risk metrics measured by looking at past data over a certain rolling window. However, future returns may vary widely from the past and strategies may subsequently fail. Investors have preferences in terms of risk, return, and diversification. One can classify the allocation strategies into two groups:

1. the low volatility allocation techniques (inverse volatility, minimum variance),
2. and the strong performance allocation techniques (equally-weighted allocation, minimum VaR, minimum expected shortfall).

While low volatility techniques deliver the best risk metrics once adjusted for volatility, strong performance techniques, such as minimum VaR and shortfall, lead to higher extreme risks. Techniques such as inverse volatility and minimum variance reduce risk and improve Sharpe ratios, mostly by steering away from the most volatile strategies at the right time. But in some circumstances this is done at the cost of lower diversification. Note, getting away from too volatile strategies by reducing the scope of the portfolio may be preferable. While the equally weighted allocation is the most diversified allocation strategy, equal risk contribution offers an attractive combination of low risk and high returns.

2.1.3 Presenting some trading strategies

2.1.3.1 Some examples of behavioural strategies

For applications of behavioural finance such as bubbles and other anomalies see Shiller [2000]. To summarise, the bubble story states that shifting investor sentiment over time creates periods of overvaluation and undervaluation in the aggregate equity market level that a contrarian market timer can exploit. In addition, varying investor sentiment across individual stocks creates cross-sectional opportunities that a contrarian stock-picker can exploit. However, short-term market timers may also join the bandwagon while the bubble keeps growing (there is a cross-sectional counterpart for this behaviour). Specifically, momentum-based stock selection strategies involving buying recent winners appear profitable. Cross-sectional trading strategies may be relatively value oriented (buy low, sell high) or momentum oriented (buy rising stocks, sell falling ones), and they may be applied within one market or across many asset markets. Micro-inefficiency refers to either the rare extreme case of riskless arbitrage opportunities or the more plausible case of risky trades and strategies with attractive reward-to-risk ratios. Note, cross-sectional opportunities are safer to exploit than market directional opportunities since one can hedge away directional risk and diversify specific risk much more effectively. Also, the value effect refers to the pattern that value stocks. For instance, those with low valuation ratios (low price/earnings, price/cash flow, price/sales, price/dividend, price/book value ratios) tend to offer higher long-run average return than growth stocks or glamour stocks with high valuation ratios. Some of the most important biases of behavioural finance are

1. momentum,
2. and reversal effects.

DeBondt et al. [1985] found stocks that had underperformed in the previous 12 to 36 months tended to subsequently outperform the market. Jegadeesh et al. [1993] found a short to medium term momentum effect where stocks that had outperformed in recent months tended to keep outperforming up to 12 months ahead. In addition, time series evidence suggests that many financial time series exhibit positive autocorrelation over short horizons and negative autocorrelation over multi-year horizons. As a result, trend following strategies are profitable for many risky assets in the short run, while value strategies, which may in part rely on long-term (relative) mean reversion, are profitable in the long run. Momentum and value strategies also appear to be profitable when applied across countries, within other asset classes, and across asset classes (global tactical asset allocation) but with different time horizons. It seems that behavioural finance was better equipped than rational finance to explain the combination of momentum patterns up to 12 months followed by reversal patterns beyond 12 months. Other models relying on different behavioural errors were developed to explain observed momentum and reversal patterns. Assuming that noise traders follow positive

feedback strategies (buying recent winners and selling recent losers) which could reflect extrapolative expectations, stop-loss orders, margin calls, portfolio insurance, wealth dependent risk aversion or sentiment, De Long et al. [1990] developed a formal model to predict both short-term momentum and long-term reversal. Positive feedback trading creates short-term momentum and price over-reaction with eventual return toward fundamental values creating long-term price reversal. Hong et al. [1999] developed a model relying on the interaction between two not-fully rational investor groups, news-watchers and momentum traders, under condition of heterogeneous information. Slow diffusion of private information across news-watchers creates underreaction and momentum effects. That is, momentum traders jumping on the bandwagon when observing trends in the hope to profit from the continued diffusion of information, generating further momentum and causing prices to over-react beyond fundamental values. All these models use behavioural finance to explain long-term reversal as a return toward fundamental values as a correction to over-reation.

2.1.3.2 Some examples of market neutral strategies

As described by Guthrie [2006], equity market neutral hedge funds buy and sell stocks with the goal of neutralising exposure to the market, while capturing a positive return, regardless of the market's direction. It includes different equity strategies with varying degrees of volatility seeking to exploit the market inefficiencies. This is in direct contradiction with the efficient market hypothesis (EMH) (see Section (1.7.2)). The main strategy involves simultaneously holding matched long and short stock positions, taking advantage of relatively under-priced and over-priced stocks. The spread between the performance of the longs and the shorts, and the interest earned from the short rebate, provides the primary return for this strategy. An equity market neutral strategy can be established in terms of dollar amount, beta, country, currency, industry or sector, market capitalisation, style, and other factors or a combination thereof. The three basic steps to build a market neutral strategy are

1. Select the universe: The universe consists of all equity securities that are candidates for the portfolio in one or more industry sectors, spanning one or more stock exchanges. The stock in the universe should have sufficient liquidity so that entering and exiting positions can be done quickly, and it should be feasible to sell stocks short. with reasonable borrowing cost.
2. Generate a forecast: Fund managers use proprietary trading models to generate potential trades. The algorithms should indicate each trade's expected return and risk, and implementation costs should be included when determining the net risk-return profile.
3. Construct the portfolio: In the portfolio construction process, the manager assigns weights (both positive and negative) to each security in the universe. There are different portfolio construction techniques, but in any case risk management issues must be considered. For instance, the maximum exposure to any single security or sector, and the appropriate amount of leverage to be employed.

One can distinguish two main approaches to equity market neutral:

1. the statistical arbitrage,
2. and the fundamental arbitrage which can be combined.

Statistical arbitrage involves model-based, short-term trading using quantitative and technical analysis to detect profit opportunities. A particular type of arbitrage opportunity is hypothesised, formalised into a set of trading rules and back-tested with historical data. This way, the manager hopes to discover a persistent and statistically significant method to detect profit opportunities. Three typical statistical arbitrage techniques are

1. Pairs or peer group involves simultaneously buying and selling short stocks of companies in the same economic sector or peer group. Typical correlations are measured and positions are established when current prices fall outside of a normal band. Position sizes can be weighted to achieve dollar, beta, or volatility neutrality. Positions are closed when prices revert to the normal range or when stop losses are breached. Portfolios of multiple pair trades are blended to reduce stock specific risk.

2. Stub trading involves simultaneously buying and selling short stocks of a parent company and its subsidiaries, depending on short-term discrepancies in market valuation versus actual stock ownership. Position sizes are typically weighted by percentage ownership.

3. Multi-class trading involves simultaneously buying and selling short different classes of stocks of the same company, typically voting and non-voting or multi-volting and single-volting share classes. Much like pairs trading, typical correlations are measured and positions are established when current prices fall outside of a normal band.

Fundamental arbitrage consists mainly of building portfolios in certain industries by buying the strongest companies and selling short companies showing signs of weakness. Even though the analysis is mainly fundamental and less quantitative than statistical arbitrage, some managers use technical and price momentum indicators (moving averages, relative strength and trading volume) to help them in their decision making. Fundamental factors used in the analysis include valuation ratios (price/earnings, price/cash flow, price/earnings before interest and tax, price/book), discounted cash flows, return on equity, operating margins and other indicators. Portfolio turnover is generally lower than in statistical arbitrage as the signals are stronger but change less frequently.

Among the factors contributing to the different sources of return are

- No index constraint: Equity market neutral removes the index constraints limiting the buy-and-hold market participants. Selling a stock short is different from not owning a stock in the index, since the weight of the short position is limited only by the manager's forecast accuracy, confidence and ability to offset market risk with long positions.

- Inefficiencies in short selling: Significant inefficiencies are available in selling stocks short.

- Time arbitrage: Equity market neutral involves a time arbitrage for short-term traders at the expense of long-term investors. With higher turnover and more frequent signals, the equity market neutral manager can often profit at the expense of the long-term equity investor.

- Additional active return potential: Equity market neutral involves double the market exposure by being both long and short stocks. At a minimum, two dollars are at work for every one dollar of invested capital. Hence, a market neutral manager has the potential to generate more returns than the active return of a long-only equity manager.

- Managing volatility: Through an integrated optimisation, the co-relationship between all stocks in an index can be exploited. Depending on the dispersion of stock returns, risk can be significantly reduced by systematically reweighting positions to profit from offsetting volatility. Reducing volatility allows for leverage to be used, which is an additional source of return.

- Profit potential in all market conditions: By managing a relatively fixed volatility portfolio, an equity market neutral manager may have an advantage over a long-only equity manager allowing him to remain fully invested in all market conditions.

The key risk factors of an equity market neutral strategy are

- Unintended beta mismatch: Long and short equity portfolios can easily be dollar neutral, but not beta neutral. Reaction to large market movements is therefore unpredictable, as one side of the portfolio will behave differently than the other.

- Unintended factor mismatch: Long and Short equity portfolios can be both dollar neutral and beta neutral, but severely mismatched on other important factors (liquidity, turnover, value/growth, market capitalisation). Again, large market moves will affect one side of the portfolio differently from the other.

- Model risk: All risk exposures of the model should be assessed to prevent bad forecast generation, and practical implementation issues should be considered. For instance, even if the model indicates that a certain stock should be shorted at a particular instant in time, this may not be feasible due to the uptick rule. Finally, the effectiveness of the model may diminish as the market environment changes.
- Changes in volatility: The total volatility of a market neutral position depends on the volatility of each position, so that the manager must carefully assess the volatility of each long and short position as well as the relationship between them.
- Low interest rates: As part of the return from an equity market neutral strategy is due to the interest earned on the proceeds from a short sale (rebate), a lower interest rate environment places more pressure on the other return sources of this strategy.
- Higher borrowing costs for stock lending: Higher borrowing costs cause friction on the short stock side, and decreases the number of market neutral opportunities available.
- Short squeeze: A sudden increase in the price of a stock that is heavily shorted will cause short sellers to scramble to cover their positions, resulting in a further increase in price.
- Currency risk: Buying and selling stocks in multiple countries may create currency risk for an equity market neutral fund. The cost of hedging, or not hedging, can significantly affect the fund's return.
- Lack of rebalancing risk: The success of a market neutral fund is contingent on constantly rebalancing the portfolio to reflect current market conditions. Failure to rebalance the portfolio is a primary risk of the strategy.

2.1.3.3 Predicting changes in business cycles

We saw in Section (2.1.3) that there are evidence in the market for different equity styles to perform better at different points in time. For instance, the stock market can be divided into two types of stocks, value and growth where value stocks are bargain or out-of-favour stocks that are inexpensive relative to company earnings or assets, and growth stocks represent companies with rapidly expanding earnings growth. Hence, an investment style which is based around growth searches for investments whose returns are expected to grow at a faster pace than the rest of the market, while the value style of investment seeks to find investments that are thought to be under-priced. Investors have an intuitive understanding that equity indexes have contrasted performance at different points of the business cycle. Historically, value investing tends to be more prominent in periods when the economy is experiencing a recession, while growth investing is performing better during times of economic booms (BlackRock, 2013). As a result, excess returns are produced by value and growth styles at different points within the business cycle since growth assets and sectors are affected in a different way than their value equivalents. Therefore, predicting the changes in the business cycle is very important as it has a direct impact on the tactical style allocation decisions. There are actually two approaches which can be considered when predicting these changes:

- One approach consists in forecasting returns by first forecasting the values of various economic variables (scenarios on the contemporaneous variables).
- The other approach to forecasting returns is based on anticipating market reactions to known economic variables (econometric model with lagged variables).

A number of academic studies (see Bernard et al. [1989]) suggested that the reaction of market participants to known variables was easier to predict than financial and economic factors. The performance of timing decisions based on an econometric model with lagged variables results from a better ability to process available information, as opposed to privileged access to private information. This makes a strong case towards using time series modelling in order to gain insights into the momentum that a market exhibits. Therefore, the objective of a Systematic Tactical Allocator is to set up an econometric model capable of predicting the time when a given Style is going to outperform other Styles.

For instance, using a robust multi-factor recursive modelling approach, Amenc et al. [2003] found strong evidence of predictability in value and size style differentials. Since forecasting returns based on anticipating market reactions to known economic variables is more favourable than trying to forecast financial or economic factors, econometric models which include lagged variables are usually used. This type of modelling is usually associated with univariate time series models but can be extended to account for cross-sections. These types of models attempt at predicting variables using only the information contained in their own past values. The class of time series models one should first consider is the ARMA/ARIMA family of univariate time series models. These types of models are usually a-theoretical therefore their construction is not based on any underlying theoretical model describing the behaviour of a particular variable. ARMA/ARIMA models try to capture any empirically relevant features of the data, and can be used to forecast past stock returns as well as to improve the signals associated with time series momentum strategies. More sophisticated models, such as Exponential Smoothing models and more generally State-Space models models can also be used.

2.1.4 Risk premia investing

As discussed in Section (1.5.3), a large range of effective multi-factor models exist to explain realised return variation over time. A different approach is to use multi-factor models directly on strategies. Risk premia investing is a way of improving asset allocation decisions with the goal of delivering a more dependable and less volatile investment return. This new approach is about allocating investment to strategies rather than to assets (see Ilmanen [2011]). Traditional portfolio allocation such as a $60/40$ allocation between equities and bonds remain volatile and dominated by equity risk. Risk premia investing introduce a different approach to portfolio diversification by constructing portfolios using available risk premia within the traditional asset classes or risk premia from systematic trading strategies rather than focusing on classic risk premia, such as equities and bonds. Correlations between many risk premia have historically been low, offering significant diversification potential, particularly during periods of distress (see Bender et al [2010]). There is a large selection of risk premia strategies across assets covering equities, bonds, credit, currency and derivative markets, and using risk-return characteristics, most of them can be classified as either income, momentum, or relative value.

1. Income strategies aim at receiving a certain steady flow of money, typically in the form of interest rates or dividend payments. These strategies are often exposed to large losses during confidence crises, when the expected income no longer offsets the risk of holding the instruments. Credit carry, VIX contango, variance swap strategies, and volatility premium strategy, equity value, dividend and size, FX carry, the rates roll-down strategy, and volatility tail-event strategies belongs to that grouping.

2. Momentum strategies are designed to bring significant gains in market downturns, whilst maintaining a decent performance in other circumstances. For example, CTA-type momentum strategies perform well when markets rise or fall significantly. An equity overwriting strategy performs best when stock prices fall, and still benefits from the option premia in other circumstances. Equity/rates/FX momentum, quality equity, overwriting, and interest rates carry belongs to that grouping. Further, a momentum system has a lot in common with a strategy that buys options and can be used as a hedge during crisis (see Ungari [2013]).

3. Relative value outright systematic alpha strategies based on market anomalies and inefficiencies, for example convertible arbitrage and dividend swaps. Such discrepancies are expected to correct over a time frame varying from a few days for technical strategies to a few months or even years for more fundamental strategies.

These categories represent distinct investment styles which is a key component in understanding risk premia payoffs and can be compared to risk factors in traditional asset classes. Portfolio managers have always tried to reap a reward for bearing extra risk, and risk premia investing is one way forward since risk premium has

- demonstrated an attractive positive historical return profile
- fundamental value allowing for a judgement on future expected returns

- some diversification benefits when combined with multi-asset portfolio

The basic concept recognises that assets are driven by a set of common risk factors for which the investor typically gets paid and, by controlling and timing exposures to these risk factors the investor can deliver a superior and more robust outcome than through more traditional forms of asset allocation (see Clarke et al. [2005]). Principal Components Analysis (PCA), which is an efficient way of summing up the information conveyed by a correlation matrix, is generally used to determine the main drivers of the strategy. For instance, in order to assess the performances of risk premia investing, Turc et al. [2013] compiled two equity portfolios based on the same five factors (value, momentum, size, quality and yield). The former is an equal-weighted combination of the five factors in the form of risk premia indices, and the latter is a quant specific model with a quantitative stock selection process scoring each stock on a combination of the five factors and creating a long-short portfolio. Using PCA, they identified three risk factors to which each strategy is exposed, market crisis, equity direction, and volatility premium. In their study, the traditional equity quant portfolio outperformed by far the combined risk premia approach. Nonetheless, when comparing the two approaches, most risk premia are transparent, easy of access, and obtained through taking a longer term view exploiting the short term consideration of other active market participants. An important issue with risk premia strategies is the way in which they are combined, as their performance and behaviour are linked to the common factors, making them difficult to implement. In the equity market, returns are expressed on a long-short basis, adding a significant amount of cost and complexity, as portfolios are often rebalanced to such a degree that annual turnover rates not only eat into returns, but also involve a considerable amount of portfolio management. Capacity constraints are another concern faced by many portfolio managers.

In order to compare risk premia strategies across different asset classes, one need to use a variety of risk metrics based on past returns such as volatility, skew and kurtosis. The skewness is a measure of the symmetry of a distribution, such that a negative skew means that the most extreme movements are on the downside. On the other hand, a strategy with a positive skew is more likely to make large gains than suffer large loss. Kurtosis is a measure of extreme risk, and a high kurtosis indicates potential fat-tails, that is, a tendency to post unusually large returns, whether either on the upside or the downside. Practitioners compares some standard performance ratios and statistics commonly used in asset management including the Sharpe ratio (returns divided by volatility), the Sortino ratio (returns divided by downside volatility), maximum drawdown and time to recovery, or measures designed to evaluate extreme risks such as value-at-risk and expected shortfall.

2.1.5 Introducing technical analysis

2.1.5.1 Defining technical analysis

We saw in Section (1.7.2) that the accumulating evidence against the efficiency of the market has caused a resurgence of interest in the claims of technical analysis as the belief that the distribution of price dynamics is totally random is now being questioned. There exists a large body of work on the mathematical analysis of the behaviour of stock prices, stock markets and successful strategies for trading in these environments. While investing involves the study of the basic market fundamentals which may take several years to be reflected in the market, trading involves the study of technical factors governing short-term market movements together with the behaviour of the market. As a result, trading is riskier than long-term investing, but it offers opportunities for greater profits (see Hill et al. [2000]).

Technical analysis is about market traders studying market price history with the view of predicting future price changes in order to enhance trading profitability. Technical trading rules involve the use of technical analysis to design indicators that help a trader determine whether current behaviour is indicative of a particular trend, together with the timing of a potential future trade. As a result, in order to apply Technical Analysis, which tries to analyse the securities past performance in view of evaluating possible future investments, we must assume that the historical data in the markets forms appropriate indications about the market future performance. Technical Analysis relies on three principles (see Murphy [1999])

1. market action discounts everything
2. prices move in trends or are contrarian
3. history tends to repeat itself

Hence, by analysing financial data and studying charts, we can anticipate which way the market is most likely to go. That is, even though we do not know when we pick a specific stock if its price is going to rise or fall, we can use technical indicators to give us a future perspective on its behaviour in order to determine the best choice when building a portfolio. Technical indicators try to capture the behaviour and investment psychology in order to determine if a stock is under or overvalued. For instance, in order to classify each stock within the market, we can employ a set of rules based on technical indicators applied to the asset's prices, their volumes, and/or other financial factors. Based on entry/exit signals and other plot characteristics, we can define different rules allowing us to score the distinct stocks within the market and subsequently pick the best securities according to the indicator employed. However, there are several problems occurring when using technical indicators. There is no better indicator, so that the indicators should be combined in order to provide different perspectives. Further, a technical indicator always need to be applied to a time window, and determining the best time window is a complex task. For instance, the problem of determining the best time window can be the solution to an optimisation problem (see Fernandez-Blanco et al. [2008]).

New techniques combining elements of learning, evolution and adaptation from the field of Computational Intelligence developed, aiming at generating profitable portfolios by using technical analysis indicators in an automated way. In particular, subjects such as Neural Networks (28), Swarm Intelligence, Fuzzy Systems and Evolutionary Computation can be applied to financial markets in a variety of ways such as predicting the future movement of stock's price or optimising a collection of investment assets (funds and portfolios). These techniques assume that there exist patterns in stock returns and that they can be exploited by analysis of the history of stock prices, returns, and other key indicators (see Schwager [1996]). With the fast increase of technology in computer science, new techniques can be applied to financial markets in view of developing applications capable of automatically manage a portfolio. Consequently, there is substantial interest and possible incentive in developing automated programs that would trade in the market much like a technical trader would, and have it relatively autonomous. A mechanical trading systems (MTS), founded on technical analysis, is a mathematically defined algorithm designed to help the user make objective trading decisions based on historically reoccurring events. Some of the reasons why a trader should use a trading systems are

- continuous and simultaneous multimarket analysis
- elimination of human emotions
- back test and verification capabilities

Mechanical system traders assume that the trending nature of the markets can be understood through the use of mathematical formulas. For instance, properly filtering time series by removing the noise (congestion), they recover the trend which is analysed in view of inferring trading signals. Assuming that assets are in a continuous state of flux, a single system can profitably trade many markets allowing a trader to be exposed to different markets without fully understanding the nuances of all the individual markets. Since MTS can be verified and analysed with accuracy through back testing, they are very popular. Commodity Trading Advisors (CTA) use systems due to their ease of use, non-emotional factor, and their ability to be used as a foundation for an entire trading platform. Everything being mathematically defined, a CTA can demonstrate a hypothetical track record based on the different needs of his client and customise a specific trading plan. However, one has to make sure the the system is not over-fitting the data in the back test. A system must have robust parameters, that is, parameters not curve fitted to historical data. Hence, one must understand the logic and mathematics behind a system. Unlike actual performance records, simulated results do not represent actual trading. As a rule of thumb, one should expect only one half of the total profit and twice the maximum draw down of a hypothetical track record.

2.1.5.2 Presenting a few trading indicators

Any mechanical trading system (MTS) must have some consistent method or trigger for entering and exiting the market based on some type of indicator or mathematical statistics with price forecasting capability. Anything that indicates what the future may hold is an indicator. Most system traders and developers spend 90% of their time developing entry and exit technique, and the rest of their time is dedicated to the decision process determining profitability. Some of the most popular indicators include moving average, rate of change, momentum, stochastic (Lane 1950), Relative Strength Iindex (RSI) (see Wilder [1978]), moving average convergence divergence (MACD) (see Appel [1999]), Donchian breakout, Bollinger bands, Keltner bands (see Keltner [1960]). However, indicators can not stand alone, and should be used in concert with other ideas and logic. There is a large list of indicators with price forecasting capability and we will describe a few of them. For more details we refer the reader to Hill et al. [2000].

- In the 50s Lane introduced an oscillatoring type of indicator called Stochastics that compares the current market close to the range of prices over a specific time period, indicating when a market is overbought or oversold (see Lane [1984]). This indicator is based on the assumption that when an uptrend/downtrend approaches a turning point, the closing prices start moving away from the high/low price of a specific range. The number generated by this indicator is a percentage in the range $[0, 100]$ such that for a reading of 70 or more it indicates that the close is near the high of the range. A reading of 30 or below indicates the close is near the low of the range. Note, in general the values of the indicator are smoothed with a moving average (MA).

- The Donchian breakout is an envelope indicator involving two lines that are plotted above and below the market. The top line represents the highest high of n days back (or weeks) and conversely the bottom line represents the lowest low of n days back. The idea being buying when the day's high penetrates the highest high of four weeks back and selling when the day's low penetrates the lowest low of four weeks back.

- The Moving Average Crossover involves two or more moving averages usually consisting of a longer-term and shorter-term average. When the short-term MA crosses from below the long-term MA, it usually indicates a buying opportunity, and selling opportunities occur when the shorter-term MA crosses from above the longer-term MA. Moving averages can be calculated as simple, exponential, and weighted average. Exponential and weighted MAs tend to skew the moving averages toward the most recent prices, increasing volatility.

- In the 70s Appel developed another price oscillator called the moving average convergence divergence (MACD) which is derived from three different exponentially smoothed moving averages (see Appel et al. [2008]). It is plotted as two different lines, the first line (MACD line) being the difference between the two MAs (long-term and short-term MAs), and the second line (signal or trigger line) being an exponentially smoothed MA of the MACD line. Note, the difference (or divergence) between the MACD line and the signal line is shown as a bar graph. The purpose being to try to eliminate the lag associated with MA type systems. This is done by anticipating the MA crossover and taking action before the actual crossover. The system buys when the MACD line crosses the signal line from below and sells when the MACD line crosses the signal line from above.

- As an example of channel trading, Keltner [1960] proposed a system called the 10-day moving average rule using a constant width channel to time buy-sell signals with the following rules

  1. compute the daily average price $\frac{high+low+close}{3}$.
  2. compute a 10-day average of the daily average price.
  3. compute a 10-day average of the daily range.
  4. add and subtract the daily average range from the 10-day moving average to form a band or channel.
  5. buy when the market penetrates the upper band and sell when it breaks the lower band.

  While this system is buying on the strength and selling on weakness, some practitioners have modified the rules as follow

1. instead of buying at the upper band, you sell and vice versa.
2. the number of days are changed to a three-day average with bands around that average.

- The Donchian channel is an indicator formed by taking the highest high and the lowest low of the last n periods. The area between the high and the low is the channel for the period chosen. While it is an indicator of volatility of a market price, it is used for providing signals for long and short positions. If a security trades above its highest n periods high, then a long is established, otherwise if it trades below the lowest n periods, a short is established.
- The Bollinger bands, also called alpha-beta bands, usually uses 20 or more days in its calculations and does not oscillate around a fixed point. It consist of three lines, where the middle line is a simple moving average and the outside lines are plus or minus two (that number can vary) standard deviations above and below the MA. A typical BB type system buys when price reaches the bottom and liquidates as the price moves up past the MA. The sell side is simply the opposite. It is assumed that when a price goes beyond two standard deviations it should revert to the MA. Note, some practitioners revert the logic and sell rather than buy when prices reach the lower band, and vice versa with the upper band. Alternatively, one can use a 20-day MA with one and two standard deviations above and below the MA. Looking at the chart we can deduce trend, volatility, and overbought/oversold conditions. A market above one standard deviation is overbought, and it becomes extremely overbought above two standard deviation. That is, most underlyings will pullback to the average even in strongly trending markets. Further, with narrow bands we should buy volatility (calls and puts) and when the bands are widening we should sell volatility.

Considering projected charts to map future market activity, Drummond et al. [1999] introduced the Drummond Geometry (DG) which is both a trend-following and a congestion-action methodology, leading rather than lagging the market. It tries to foretell the most likely scenario that shows the highest probability of occurring in the immediate future and can be custom fitted to one's personality and trading style. The key elements of DG include a combination of the following three basic categories of trading tools and techniques

1. a series of short-term moving averages
2. short-term trend lines
3. multiple time-period overlays

The concept of Point and Line (PL) reflecting how all of life moves from one extreme to another, flowing back and forth in a cyclical or wave-like manner, is applied to the market. The PLdot (average of the high, low, and close of the last three bars) is a short-term MA based on three bars (or time periods) of data capturing the trend/nontrend activity of the time frame that is being charted. It represent the center of all market activity. The PLdot is very sensitive to trending markets, it is also very quick at registering the change of a market out of congestion (noise) into trend, and it is sensitive to ending trend.

Since pattern is defined as a predictable route or movement, all trading systems are pattern recognition systems. For instance, a long-term moving average cross over system uses pattern recognition, the crossover, in its decision to buy or sell. Similarly, an open range breakout is pattern recognition given by the movement from the open to the breakout point. All systems look for some type of reoccurring event and try to capitalise on it. Hill demonstrated the success of pattern recognition when used as a filter. The system was developed around the idea of a pattern consisting of the last four days' closing prices. A buy or sell signal is not generated unless the range of the past four days' closes is less than the 30-day average true range, indicating that the market has reached a state of rest and any movement, up or down, from this state will most likely result in a significant move.

Eventually, we want to develop an approach into a comprehensive, effective, trading methodology that combines analytical sophistication with tradable rules and principles. One way forward is to consider a multiple time frame approach. A time frame is any regular sampling of prices in a time series, from the smallest such as one minute up to the longest capped out at ten year. The multiple time frame approach has proven to be a fundamental advance in the field of TA allowing for significant improvement in trading results. For instance, if market analysis is coordinated to show the interaction of these time frames, then the trader can monitor what happens when the support and resistance lines of the different time frames coincide. Assuming we are interested in analysing the potential of the integration of time frames, then we need to look at both a higher time frame and a lower time frame.

2.1.5.3 The limitation of indicators

Since each indicator has a significant failure rate, the random nature of price change being one reason why indicators fail, Chande et al. [1994] explained how most traders developed several indicators to analyse prices. Traders use multiple indicators to confirm the signal of one indicator with respect to another one, believing that the consensus is more likely to be correct. However, this is not a viable approach due to the strong similarities existing between price based momentum indicators. In general, momentum based indicators fail most of the time because

- none of them is a pure momentum oscillator that measures momentum directly.
- the time period of the calculations is fixed, giving a different picture of market action for different time periods.
- they all mirror the price pattern, so that it may be better trading prices themselves.
- they do not consistently show extremes in prices because they use a constant time period.
- the smoothing mechanism introduces lags and obscures short-term price extremes that are valuable for trading.

2.1.5.4 The risk of overfitting

While a simplified representation of reality can either be descriptive or predictive in nature, or both, financial models are predictive to forecast unknown or future values based on current or known values using mathematical equations or set of rules. However, the forecasting power of a model is limited by the appropriateness of the inputs and assumptions so that one must identify the sources of model risk to understand these limitations. Model risk generally occurs as a result of incorrect assumptions, model identification or specification errors, inappropriate estimation procedures, or in models used without satisfactory out-of-sample testing. For instance, some models can be very sensitive to small changes in inputs, resulting in big changes in outputs. Further, a model may be overfitted, meaning that it captures the underlying structure or the dynamics in the data as well as random noise. This generally occurs when too many model parameters are used, restricting the degrees of freedom relative to the size of the sample data. It often results in good in-sample fit but poor out-of-sample behaviour. Hence, while an incorrect or misspecified model can be made to fit the available data by systematically searching the parameter space, it does not have a descriptive or predictive power. Familiar examples of such problems include the spurious correlations popularised in the media, where over the past 30 years when the winner of the Super Bowl championship in American football is from a particular league, a leading stock market index historically goes up in the following months. Similar examples are plentiful in economics and the social sciences, where data are often relatively sparse but models and theories to fit the data are relatively prolific. In economic time series prediction, there may be a relatively short time-span of historical data available in conjunction with a large number of economic indicators. One particularly humorous example of this type of prediction was provided by Leinweber who achieved almost perfect prediction of annual values of the $S\&P$ 500 financial index as a function of annual values from previous years for butter production, cheese production, and sheep populations in Bangladesh and the United States. There are no easy technical solutions to this problem, even though various strategies have been developed. In order to avoid model overfitting and data snooping, one should decide upon the

framework by defining how the model should be specified before beginning to analyse the actual data. First, by properly formulating model hypothesis making financial or economic sense, and then carefully determining the number of dependent variables in a regression model, or the number of factors and components in a stochastic model one can expect avoiding or reducing storytelling and data mining. To increase confidence in a model, true out-of-sample studies of model performance should be conducted after the model has been fully developed. We should also be more comfortable with a model working cross-sectionally and producing similar results in different countries. Note, in a general setting, sampling bootstraping, and randomisation techniques can be used to evaluate whether a given model has predictive power over a benchmark model (see White [2000]). All forecasting models should be monitored and compared on a regular basis, and deteriorating results from a model or variable should be investigated and understood.

2.1.5.5 Evaluating trading system performance

We define a successful strategy to be one that maximise the number of profitable days, as well as positive average profits over a substantial period of time, coupled with reasonably consistent behaviour. As a result, while we must look at profit when evaluating trading system performance, we must also look at other statistics such as

- maximum drawdown: the highest point in equity to the subsequent lowest point in equity. It is the largest amount of money the system lost before it recovered.
- longest flat time: the amount of time the system went without making money.
- average drawdown: maximum drawdown is one time occurrence, but the average drawdown takes all of the yearly drawdowns into consideration.
- profit to loss ratio: it represents the magnitude of winning trade dollars to the magnitude of losing trade dollars. As it tells us the ratio of wins to losses, the higher the ratio the better.
- average trade: the amount of profit or loss we can expect on any given trade.
- profit to drawdown ratio: risk in this statistic comes in the form of drawdown, whereas reward is in the form of profit.
- outlier adjusted profit: the probability of monstrous wins and/or losses reoccurring being extremely slim, it should not be included in an overall track record.
- most consecutive losses: it is the total number of losses that occurred consecutively. It gives the user an idea of how many losing trades one may have to go through before a winner occurs.
- Sharpe ratio: it indicates the smoothness of the equity curve. The ratio is calculated by dividing the average monthly or yearly return by the standard deviation of those returns.
- long and short net profit: as a robust system would split the profits between the long trades and the short trades, we need to make sure that the money made is well balanced between both sides.
- percent winning months: it checks the number of winning month out of one year.

2.2 Portfolio construction

We assume that the assets have already been selected, but we do not know the allocations, and we try to make the best choice for the portfolio weights. In his article about the St. Petersburg paradox, Bernoulli [1738-1954] argued that risk-averse investors will want to diversify: " ... it is advisable to divide goods which are exposed to some small danger into several portions rather than to risk them all together ". Later, Fisher [1906] suggested variance as a measure of economic risk, which lead investors to allocate their portfolio weights by minimising the variance

of the portfolio subject to several constraints. For instance, the theory of mean-variance based portfolio selection, proposed by Markowitz [1952], assumes that rational investors choose among risky assets purely on the basis of expected return and risk, where risk is measured as variance. Markowwitz concluded that rational investors should diversify their investments in order to reduce the respective risk and increase the expected returns. The author's assumption focus on the basis that for a well diversified portfolio, the risk which is assumed as the average deviation from the mean, has a minor contribution to the overall portfolio risk. Instead, it is the difference (covariance) between individual investment's levels of risk that determines the global risk. Based on this assumption, Markowitz provided a mathematical model which can easily be solved by meta-heuristics such as Simulated Annealing (SA) or Genetic Algorithm (GA). Solutions based on this model focus their goals on optimising either a single objective, the risk inherent to the portfolio, or two conflicting objectives, the global risk and the expected returns of the securities within the portfolio. That is, a portfolio is considered mean-variance efficient

- if it minimises the variance for a given expected mean return, or,
- if it maximise the expected mean return for a given variance.

On a theoretical ground, mean-variance efficiency assumes that

- investors exhibit quadratic utility, ignoring non-normality in the data, or
- returns are multivariate normal, such that all higher moments are irrelevant for utility function.

2.2.1 The problem of portfolio selection

The expected return on a linear portfolio being a weighted sum of the returns on its constituents, we denote the expected return by $\mu_p = w^\top E[r]$, where

$$E[r] = (E[r_1], .., E[r_N])^\top \text{ and } w = (w_1, ..., w_N)^\top$$

are the vectors of expected returns on N risky assets and portfolio weights. The variance of a linear portfolio has the quadratic form $\sigma^2 = w^\top Q w$ where Q is the covariance matrix of the asset returns. In practice, we can either

- minimise portfolio variance for all portfolios ranging from minimum return to maximum return to trace out an efficient frontier, or
- construct optimal potfolios for different risk-tolerance parameters, and by varying the parameters, find the efficient frontier.

2.2.1.1 Minimising portfolio variance

We assume that the elements of the $N \times 1$ vector w are all non-negative and sum to 1. We can write the $N \times N$ covariance matrix Q as $Q = DCD$ where D is the $N \times N$ diagonal matrix of standard deviations and C is the correlation matrix of the asset returns (see details in Appendix (A.6)). One can show that whenever asset returns are less than perfectly correlated, the risk from holding a long-only portfolio will be less than the weighted sum of the component risks. We can write the variance of the portfolio return R as

$$Q(R) = w^\top Q w = w^\top DCDw = x^\top Cx$$

where

$$x = Dw = (w_1\sigma_1, .., w_N\sigma_N)^\top$$

is a vector where each portfolio weight is multiplied by the standard deviation of the corresponding asset return. If all asset returns are perfectly correlated, then $C = I_N$, and the volatility of the portfolio becomes

$$(w^\top Qw)^{\frac{1}{2}} = w_1\sigma_1 + ... + w_N\sigma_N$$

where the standard deviation of the portfolio return is the weighted sum of the asset return standard deviation. However, when some asset returns have less than perfect correlation, then C has elements less than 1. As the portfolio is long-only, the vector x has non-negative elements, and we get

$$Q(R) = w^\top Qw = x^\top Cx \leq I^\top CI$$

which is an upper bound for the portfolio variance. It correspond to the Principle of Portfolio Diversification (PPD). Maximum risk reduction for a long-only portfolio occurs when correlations are highly negative. However, if the portfolio contains short positions, we want the short positions to have a high positive correlation with the long positions for the maximum diversification benefit. The PPD implies that investors can make their net specific risk very small by holding a large portfolio with many assets. However, they are still exposed to irreducible risk since the exposure to a general market risk factor is common to all assets.

We obtain the minimum variance portfolio (MVP) when the portfolio weights are chosen so that the portfolio variance is as small as possible. That is

$$\min_w w^\top Qw \tag{2.2.1}$$

with the constraint

$$\sum_{i=1}^N w_i = 1$$

in the case of a long-only portfolio. Any constraint on the portfolio weights restricts the feasible set of solutions to the minimum variance problem. In the case of the single constraint above, the solution to the MVP is given by

$$\tilde{w}_i = \psi_i \big(\sum_{i=1}^N \psi_i\big)^{-1}$$

where ψ_i is the sum of all the elements in the ith column of Q^{-1}. The portfolio with these weights is called the global minimum variance portfolio with variance

$$V^* = \big(\sum_{i=1}^N \psi_i\big)^{-1}$$

In general, there is no analytic solution to the MVP when more constraints are added. While the MVP ignores the return characteristics of portfolio, more risk may be perfectly acceptable for higher returns. As a result, Markowitz [1952] [1959] considered adding another constraint to the MVP by allowing the portfolio to meet or exceed a target level of return $\overline{R}$ leading to the optimisation problem of solving Equation (2.2.1) subject to the constraints

$$\sum_{i=1}^N w_i = 1 \text{ and } w^\top E[r] = \overline{R}$$

where $E[r] = \overline{R}$ is a target level for the portfolio return. Using the Lagrange multipliers we can obtain the solution analytically.

2.2.1.2 Maximising portfolio return

An alternative approach is to maximise portfolio return by defining the utility function

$$U = \mu_p - \frac{1}{2\lambda}\sigma^2 = w^\top \mu - \frac{1}{2\lambda} w^\top Q w \tag{2.2.2}$$

where $\mu = E[r]$. We let λ be a risk-tolerance parameter, and compute the optimal solution by taking the first derivative with respect to portfolio weights, setting the term to zero

$$\frac{dU}{dw} = \mu - \frac{1}{2\lambda} 2Qw = \mu - \frac{1}{\lambda} CQ = 0$$

and solving for the optimal vector w^*, getting

$$w^* = \lambda Q^{-1} \mu$$

To be more realistic, we introduce general linear constraints of the form $Aw = b$, where A is a $N \times M$ matrix where N is the number of assets and M is the number of equality constraints and b is a $N \times 1$ vector of limits. We now maximise

$$U = w^\top \mu - \frac{1}{2\lambda} w^\top Q w \text{ subject to } Aw = b$$

We can write the Lagrangian

$$L = w^\top \mu - \frac{1}{2\lambda} w^\top Q w - \delta^\top (Aw - b)$$

where δ is the $M \times 1$ vector of Lagrangian multipliers (one for each constraint). Taking the first derivatives with respect to the optimal weight vector and the vector of multipliers yields

$$\begin{aligned} \frac{dL}{dw} &= \mu - \frac{1}{\lambda} Qw - \delta^\top A = 0 \, , \, w^* = \lambda Q^{-1}(\mu - \delta^\top A) \\ \frac{dL}{d\delta} &= Aw - b \, , \, Aw = b \end{aligned}$$

From the above equations, we obtain

$$\begin{aligned} \lambda A Q^{-1} \mu - b &= \lambda A Q^{-1} A^\top \delta \\ \delta &= \frac{AC^{-1}\mu}{AC^{-1}A^\top} - \frac{1}{\lambda} b A C^{-1} A^\top \end{aligned}$$

Replacing in the derivative of the Lagrangian, we get the optimal solution under linear equality constraints

$$w^* = Q^{-1} A^\top (AQ^{-1}A^\top) b + \lambda Q^{-1} \big(\mu - A^\top (AQ^{-1}A^\top)^{-1} AQ^{-1}\mu\big)$$

The optimal solution is split into a (constrained) minimum variance portfolio and a speculative portfolio. It is called a two-fund separation because the first term does not depend on expected returns or on risk tolerance, and the second term is sensitive to both inputs. To test for the significance between the constrained and unconstrained optimisation we can use the Shape ratio R_S. Assuming an unconstrained optimisation with $N^{'}$ assets and a constrained optimisation with only N assets $(N^{'} > N)$ we an use the measure

$$\frac{(T - N^{'})(N^{'} - N)(R_S^2(N^{'}) - R_S^2(N))}{(1 + R_S^2(N))} \sim F_{N^{'}, T-(N^{'}+N+1)}$$

where T is the number of observations. This statistic is F-distributed.

2.2.1.3 Accounting for portfolio risk

While there are many competing allocation procedures such as Markowwitz portfolio theory (PT), or risk budgeting methods to name a few, in all cases risk must be decided. Scherer [2007] argued that portfolio construction, using various portfolio optimisation tools to assess expected return versus expected risk, is equivalent to risk budgeting. In both cases, investors have to trade off risk and return in an optimal way. Even though the former is an allocation either in nominal dollar terms or in percentage weights, and the latter arrives at risk exposures expressed in terms of value at risk (VaR) or percentage contributions to risk, this is juste a presentational difference. While the average target volatility of the portfolio is closely related to the risk aversion of the investors, this amount is not constant over time. In general, any consistent investment process should measure and control the global risk of a portfolio. Nonetheless, it seems that full risk-return optimisation at the portfolio level is only done in the most quantitative firms, and that portfolio management remains a pure judgemental process based on qualitative, not quantitative, assessments. Portfolio managers developed risk measures to represent the level of risk in a particular portfolio, where risk is defined as underperformance relative to a mandate. In the financial industry, there is a large variety of risk indicators, and portfolio managers must decide which ones to consider. For instance, one may consider the maximum drawdown of the cumulative profit or total open equity of a financial trading strategy. One can also consider performance measures such as the Sharpe ratio, the Burke ratio, the Calmar ratio and many more at their disposal. In practice, portfolio managers must consider other issues such as execution policies and transaction cost management on a regular basis.

The main reason for using qualitative measures being the difficulty to apply practically optimisation technology. For instance, the classical mean-variance optimisation is very sensitive to inputs such as expected returns of each asset and their covariance matrix. Chopra et al. [1993] have done elementary research to the sensitivity of the classic Mean-Variance to errors in the input parameters. According to their research, errors in expected returns have a much bigger influence on the performance than errors in variances or covariances. This lead to optimal portfolios having extreme or non-intuitive weights for some of the individual assets. Consequently, practitioners added constraints to the original problem to limit or reduce these drawbacks, resulting in an optimum portfolio dominated by the constraints. Additional problems to portfolio optimisation exist, such as

- poor model ex-post performance, coupled with the risk of maximising error rather than minimising it.
- difficulty in estimating a stable covariance matrix for a large number of assets.
- sensitivity of portfolio weights to small changes in forecasts.

Different methods exist to make the portfolio allocation process more robust to different sources of risk (estimation risk, model risk etc.) among which are

- Bayesian approches
- Robust Portfolio Allocation

In the classical approach future expected returns are estimated by assuming that the true expected returns and covariances of returns are unknown and fixed. Hence, a point estimate of expected returns and (co)variances is obtained using forecasting models of observed market data, influencing the mean-variance portfolio allocation decision by the estimation error of the forecasts. Once the expected returns and the covariance matrix of returns have been estimated, the portfolio optimisation problem is typically treated and solved as a deterministic problem with no uncertainty.

A more realistic model would consider the uncertainty of expected returns and risk into the optimisation problem. One way forward is to choose an optimum portfolio under different scenarios that is robust in some worst case model misspecification. The goal of the Robust Portfolio Allocation (RPA) framework is to get a portfolio, which will perform well under a number of different scenarios instead of one scenario. However, to obtain such a portfolio the investor has to give up some performance under the most likely scenarios to have some insurance for the less

likely scenarios. In order to construct such portfolios an expected returns distribution is necessary instead of a point-estimate. One method to obtain such distributions is the Bayesian method which assumes that the true expected returns are unknown and random. A prior distribution is used, reflecting the investor's knowledge about the probability before any data are observed. The posterior distribution, computed with Bayes' formula, is based on the knowledge of the prior probability distribution plus the new data. For instance, Black et al. [1990] estimated future expected returns by combining market equilibrium (CAPM equilibrium) with an investor's views. The Bayesian framework allows forecasting systems to use external information sources and subjective interventions in addition to traditional information sources. The only restriction being that additional information is combined with the model following the law of probability (see Carter et al. [1994]).

One alternative approach, discussed by Focardi et al. [2004], is to use Monte Carlo technique by sampling from the return distribution and averaging the resulting portfolios. In this method a set of returns is drawn iteratively from the expected return distribution. In each iteration, a mean-variance optimisation is run on the set of expected returns. The robust portfolio is then the average of all the portfolios created in the different iterations. Although this method will create portfolios that are more or less robust, it is computationally very expensive because an optimisation must be run for each iteration step. Furthermore there is no guarantee that the resulting average portfolio will satisfy the constraints on which the original portfolios are created. Note, in the Robust Portfolio Allocation approach, the portfolio is not created with an iterative process but the distribution of the expected returns is directly taken into account, resulting in a single optimisation process. Therefore this approach is computationally more effective than the Monte Carlo process.

2.3 A market equilibrium theory of asset prices

In the problem of portfolio selection, the mean-variance approach introduced by Markowitz [1952] is a simple trade-off between return and uncertainty, where one is left with the choice of one free parameter, the amount of variance acceptable to the individual investor. For proofs and rigorous introduction to the mean-variance portfolio technique see Huang et al. [1988]. For a retrospective on Markowitz's portfolio selection see Rubinstein [2002]. Investment theory based on growth is an alternative to utility theory with simple goal. Following this approach, Kelly [1956] used the role of time in multiplicative processes to solve the problem of portfolio selection.

2.3.1 The capital asset pricing model

2.3.1.1 Markowitz solution to the portfolio allocation problem

We showed in Section (2.2.1) how a rational investor should allocate his funds between the different risky assets in his universe, leading to the portfolio allocation problem. To solve this problem Markowitz [1952] introduced the concept of utility functions (see Section (1.6.1) and Appendix (A.7)) to express investor's risk preferences. Markowitz first considered the rule that the investor should maximise discounted expected, or anticipated returns (which is linked to the St. Petersurg paradox). However, he showed that the law of large numbers (see Bernoulli [1713]) can not apply to a portfolio of securities since the returns from securities are too intercorrelated. That is, diversification can not eliminate all variance. Hence, rejecting the first hypothesis, he then considered the rule that the investor should consider expected return a desirable thing and variance of return an undesirable thing. Note, Marschak [1938] suggested using the means and covariance matrix of consumption of commodities as a first order approximation to utility.

We saw in Section (2.2.1) that the mean-variance efficient portfolios are obtained as the solution to a quadratic optimisation program. Its theoretical justification requires either a quadratic utility function or some fairly restrictive assumptions on the class of return distribution, such as the assumption of normally distributed returns. For instance, we assume zero transaction costs and portfolios with prices V_t taking values in $\mathbb{R}$ and following the geometric Brownian motion with dynamics under the historical probability measure $\mathbb{P}$ given by

$$\frac{dV_t}{V_t} = \mu dt + \sigma_V dW_t \tag{2.3.3}$$

where μ is the drift, σ_V is the volatility and W_t is a standard Brownian motion. Markowitz first considered the problem of maximising the expected rate of return

$$g = \frac{1}{dt} < \frac{dV_t}{V_t} >= \mu$$

(also called ensemble average growth rate) and rejected such strategy because the portfolio with maximum expected rate of return is likely to be under-diversified, and as a result, to have an unacceptable high volatility. As a result, he postulated that while diversification would reduce risk, it would not eliminate it, so that an investor should maximise the expected portfolio return μ while minimising portfolio variance of return σ_V^2. It follows from the relation between the variance of the return of the portfolio σ_V^2 and the variance of return of its constituent securities σ_j^2 for $j = 1, 2, ..., N$ given by

$$\sigma_V^2 = \sum_j w_j^2 \sigma_j^2 + \sum_j \sum_{k \neq j} w_j w_k \rho_{jk} \sigma_j \sigma_k$$

where the w_j are the portfolio weights such that $\sum_j w_j = 1$, and ρ_{jk} is the correlation of the returns of securities j and k. Therefore, $\rho_{jk}\sigma_j\sigma_k$ is the covariance of their returns. So, the decision to hold any security would depend on what other securities the investor wants to hold. That is, securities can not be properly evaluated in isolation, but only as a group. Consequently, Markowitz suggested calling portfolio i efficient if

1. there exists no other portfolio j in the market with equal or smaller volatility, $\sigma_j \leq \sigma_i$, whose drift term μ_j exceeds that of portfolio i. That is, for all j such that $\sigma_j \leq \sigma_i$, we have $\mu_j \leq \mu_i$.
2. there exists no other portfolio j in the market with equal or greater drift term, $\mu_j \geq \mu_i$, whose volatility σ_j is smaller than that of portfolio i. That is, for all j such that $\mu_j \geq \mu_i$, we have $\sigma_j \geq \sigma_i$.

In the presence of a riskless asset (with $\sigma_i = 0$), all efficient portfolios lie along a straight line, the efficient frontier, intersecting in the space of volatility and drift terms, the riskless asset r_f and the so-called market portfolio M. Since any point along the efficient frontier represents an efficient portfolio, additional information is needed in order to select the optimal portfolio. For instance, one can specify the usefulness or desirability of a particular investment outcome to a particular investor, namely his risk preference, and represent it with a utility function $u = u(V_t)$. Following the work of von Neumann et al. [1944] and Savage [1954], Markowitz [1959] found a way to reconcile his mean-variance criterion with the maximisation of the expected utility of wealth after many reinvestment periods. He advised using the strategy of maximising the expected logarithmic utility of return each period for investors with a long-term horizon, and developed a quadratic approximation to this strategy allowing the investor to choose portfolios based on mean and variance.

As an alternative to the problem of portfolio selection, Kelly [1956] proposed to maximise the expected growth rate

$$g^b = \frac{1}{dt} < d\ln V_t >= \mu - \frac{1}{2}\sigma^2$$

obtained by using Ito's formula (see Peters [2011c]). This rate is called the expected growth rate, or the logarithmic geometric mean rate of return (also called the time average growth rate). In that setting, we observe that large returns and small volatilities are desirable. Note, Ito's formula changes the behaviour in time without changing the noise term. That is, Ito's formula encode the multiplicative effect of time (for noise terms) in the ensemble average (see Oksendal [1998]). Hence, it can be seen as a mean of accounting for the effects of time. For self-financing portfolios, where eventual outcomes are the product over intermediate returns, maximising g^b yields meaningful results. This

is because it is equivalent to using logarithmic utility function $u(V_t) = \ln V_t$. In that setting, the rate of change of the ensemble average utility happens to be the time average of the growth rate in a multiplicative process. Note, the problem that additional information is needed to select the right portfolio disappears when using the expected growth rate g^b. That is, there is no need to use utility function to express risk preferences as one can use solely the role of time in multiplicative processes.

2.3.1.2 The Sharp-Lintner CAPM

As discusses in Section (2.2.1), in equilibrium and under proper diversification, market prices are made of the risk-free rate and the price of risk in such a way that an investor can attain any desired point along a capital market line (CML). Higher expected rate of return can be obtained only by incurring additional risk. In view of properly describing the price of risk, Sharpe [1964] extended the model of investor behaviour (see Arrow [1952] and Markowitz [1952]) to construct a market equilibrium theory of asset prices under conditions of risk. That is, the purpose of the capital asset pricing model (CAPM) is to deduce how to price risky assets when the market is in equilibrium. The conditions under which a risky asset may be added to an already well diversified portfolio depend on the Systematic Risk of the asset, also called the undiversifiable risk of the asset. Assuming that an investor views the possible result of an investment in terms of some probability distribution, he might consider the first two moments of the distribution represented by the total utility function

$$u = f(E_w, \sigma_w)$$

where E_w is the expected future wealth and σ_w the predicted standard deviation, and such that $\frac{du}{dE_w} > 0$ and $\frac{du}{d\sigma_w} < 0$. Letting W_i be the quantity of the investor's present wealth and W_t be his terminal wealth, we get

$$W_t = W_i(1 + R)$$

where R is the rate of return on the investment. One can therefore express the utility function in terms of R, getting

$$u = g(E_R, \sigma_R)$$

so that the investor can choose from a set of investment opportunities, represented by a point in the (E_R, σ_R) plane (with E_R on the x-axis and σ_R on the y-axis), the one maximising his utility. Both Markowitz [1952] [1959] and Tobin [1958] derived the indifference curves by maximising the expected utility with total utility represented by a quadratic function of R with decreasing marginal utility. The investor will choose the plan placing him on the indifference curve representing the highest level of utility. A plan is said to be efficient if and only if there is no alternative with either

1. the same E_R and a lower σ_R
2. the same σ_R and a hiher E_R
3. a higher E_R and a lower σ_R

For example, in the case of two investment plans A and B, each with one or more assets, such that α is the proportion of the individual's wealth placed in plan A and $(1 - \alpha)$ in plan B, the expected rate of return is

$$E_{R_c} = \alpha E_{R_a} + (1 - \alpha)E_{R_b}$$

and the predicted standard deviation of return is

$$\sigma_{R_c} = \sqrt{\alpha^2\sigma_{R_a}^2 + (1 - \alpha)^2\sigma_{R_b}^2 + 2\rho_{ab}\alpha(1 - \alpha)\sigma_{R_a}\sigma_{R_b}} \tag{2.3.4}$$

where ρ_{ab} is the correlation between R_a and R_b. In case of perfect correlation between the two plans $(\rho_{ab} = 1)$, both E_{R_c} and σ_{R_c} are linearly related to the proportions invested in the two plans and the standard deviation simplifies to

$$\sigma_{R_c} = \sigma_{R_b} + \alpha(\sigma_{R_a} - \sigma_{R_b})$$

Considering the riskless asset P with $\sigma_{R_p} = 0$, an investor placing α of his wealth in P and the remainder in the risky asset A, we obtain the expected rate of return

$$E_{R_c} = \alpha E_{R_p} + (1-\alpha)E_{R_a}$$

and the standard deviation reduces to

$$\sigma_{R_c} = (1-\alpha)\sigma_{R_a}$$

such that all combinations involving any risky asset or combination of assets with the riskless asset must have the values (E_{R_c}, σ_{R_c}) lying along a straight line between the points representing the two components. To prove it, we set $(1-\alpha) = \frac{\sigma_{R_c}}{\sigma_{R_a}}$ and replace in the expected rate of return, getting

$$E_{R_c} = E_{R_p} + \frac{E_{R_a} - E_{R_p}}{\sigma_{R_a}}\sigma_{R_c}$$

***Remark* 2.3.1** *The investment plan lying at the point of the original investment opportunity curve where a ray from point P is tangent to the curve will dominate.*

Since borrowing is equivalent to disinvesting, assuming that the rate at which funds can be borrowed equals the lending rate, we obtain the same dominant curve.

To reach equilibrium conditions, Sharpe [1964] showed that by assuming a common risk-free rate with all investors borrowing or lending funds on equal terms, and homogeneity of investor expectations, capital asset prices must keep changing until a set of prices is attained for which every assets enters at least one combination lying on the capital market line (CML). While many alternative combinations of risky assets are efficient, they must be perfectly positively correlated as they lie along a linear border of the (E_R, σ_R) region, even though the contained individual securities are not perfectly correlated. For individual assets, the pair (E_{R_i}, σ_i) for the ith asset (with E_{R_i} on the x-axis and σ_i on the y-axis) will lie above the capital market line (due to inefficiency of undiversified holdings) and be scattered throughout the feasible region. Given a single capital asset (point i) and an efficient combination of assets (point g) of which it is part, we can combine them in a linear way such that the expected return of the combination is

$$E = \alpha E_{R_i} + (1-\alpha)E_{R_g}$$

In equilibrium, Sharpe obtained a tangent curve to the CML at point g, leading to a simple formula relating E_{R_i} to some risk in combination g. The standard deviation of a combination of i and g is given by Equation (2.3.4) with a and b replaced with i and g, respectively. Further, at $\alpha = 0$ we get

$$\frac{d\sigma}{dE} = \frac{\sigma_{R_g} - \rho_{ig}\sigma_{R_i}}{E_{R_g} - E_{R_i}}$$

Letting the equation of the capital market line (CML) be

$$\sigma_R = s(E_R - P) \text{ or } E_R = P + b\sigma_R \tag{2.3.5}$$

with $b = \frac{1}{s}$, where P is the risk-free rate, and since we have a tangent line at point g with the pair (E_{R_g}, σ_{R_g}) lying on that line, we get

$$\frac{\sigma_{R_g} - \rho_{ig}\sigma_{R_i}}{E_{R_g} - E_{R_i}} = \frac{\sigma_{R_g}}{E_{R_g} - P}$$

Given a number of ex-post observations of the return of the two investments with E_{R_f} approximated with $\overline{R}_f$ for $f = i, g$ and total risk σ_{R_f} approximated with $\overline{\sigma}$, we call B_{ig} the slope of the regression line between the two returns, and observe that the response of $\overline{R}_i$ to changes in $\overline{R}_g$ account for much of the variation in $\overline{R}_i$. This component B_{ig} of the asset's total risk is called the Systematic Risk, and the remainder which is uncorrelated with $\overline{R}_g$ is the unsystematic component. This relationship between $\overline{R}_i$ and $\overline{R}_g$ can be employed ex-ante as a predictive model where B_{ig} is the predicted response of $\overline{R}_i$ to changes in $\overline{R}_g$. Hence, all assets entering efficient combination g have B_{ig} and E_{R_i} values lying on a straight line (minimum variance condition)

$$E_{R_i} = B_{ig}(E_{R_g} - P) + P \tag{2.3.6}$$

where P is the risk-free rate and

$$B_{ig} = \frac{\rho_{ig}\sigma_{R_i}}{\sigma_{R_g}} = \frac{Cov(R_i, R_g)}{\sigma^2_{R_g}} \tag{2.3.7}$$

The slope B_{ig}, also called the CAPM beta, represents the part of an asset's risk which is due to its correlation with the return on a combination and can not be diversified away when the asset is added to the combination. Consequently, it should be directly related to the expected return E_{R_i}. This result is true for any efficient combinations because the rates of return from all efficient combinations are perfectly correlated. Risk resulting from swings in economic activity being set aside, the theory states that after diversification, only the responsiveness of an asset's rate of return to the level of economic activity is relevant in assessing its risk. Therefore, prices will adjust until there is a linear relationship between the magnitude of such responsiveness and expected return.

2.3.1.3 Some critics and improvements of the CAPM

Fama et al. [2004] discussed the CAPM and argued that whether the model's problems reflect weakness in the theory or in its empirical implementation, the failure of the CAPM in empirical test implies that most applications of the model are invalid. The CAPM is based on the model of portfolio choice developed by Markowitz [1952] where an investor selects a portfolio at time $t-1$ that produces a stochastic return at time t. Investors are risk averse in choosing among portfolios, and care only about the mean and variance of their one-period investment return. It results in algebraic condition on asset weights in mean-variance efficient portfolios (see Section (2.3.1.2)). In view of making prediction about the relation between risk and expected return Sharpe [1964] and Lintner added two assumptions, complete agreement of all investors on the joint distribution of asset returns from $t-1$ to t supposed to be the true one, and that there is borrowing and lending at a risk-free rate. As a result, the market portfolio M (tangency portfolio) must be on the minimum variance frontier if the asset market is to clear and satisfy Equation (2.3.6) with g replaced by M. The slope B_{iM} measures the sensitivity of the asset's return to variation in the market return, or put another way, it is proportional to the risk each dollar invested in asset i contributes to the market portfolio. Consequently, it can be seen as a sensitivity risk measure relative to the market risk factor. To stress the proportionality of the normalised excess return of the risky asset with that of the market portfolio, we can rewrite Equation (2.3.6) as

$$\frac{E_{R_i} - P}{\sigma_{R_i}} = \rho_{iM}\frac{E_{R_M} - P}{\sigma_{R_g}}$$

Black [1972] developed a version of the CAPM without risk-free borrowing or lending by allowing unrestricted short sales of risky assets. These unrealistic simplifications were tested by Fama et al. [2004]. They were faced with numerical errors when estimating the beta of individual assets to explain average returns, and they obtained positive correlation in the residuals producing bias in ordinary least squares (OLS) estimates. Since the CAPM explains security returns, it also explains portfolio returns so that one can work with portfolios rather than securities to estimate betas. Letting w_{ip} for $i = 1, .., N$ be the weights for the assets in some portfolio p, the expected return and market beta for the portfolio are given by

$$E_{R_p} = \sum_{i=1}^{N} w_{ip} E_{R_i} \text{ and } \beta_{pM} = \sum_{i=1}^{N} w_{ip} \beta_{ip}$$

so that the CAPM relation in Equation (2.3.6) also holds when the ith asset is a portfolio. However, grouping stocks shrinks the range of betas and reduces statistical power, so that one should sort securities on beta when forming portfolios where the first portfolio contains securities with the lowest betas, and so on, up to the last portfolio with the highest beta assets (see Black et al. [1972]). Jensen [1968] argued that the CAPM relation in Equation (2.3.6) was also a time-series regression test

$$E_{R_{it}} = \alpha_i + B_{iM}(E_{R_{Mt}} - R_{ft}) + R_{ft} + \epsilon_{it} \tag{2.3.8}$$

where R_{ft} is the risk-free rate at time t, ϵ_{it} is assumed to be a white noise, and such that the intercept term in the regression, also called the Jensen's alpha, is zero for each asset. In the CAPM equilibrium, no single asset may have abnormal return where it earns a rate of return of alpha above (or below) the risk free rate without taking any market risk. In the case where $\alpha_i \neq 0$ for any risky asset i, the market is not in equilibrium, and pairs (E_{R_i}, B_{iM}) will lie above or below the CML according to the sign of α_i. If the market is not in equilibrium with an asset having a positive alpha, it should have an expected return in excess of its equilibrium return and should be bought. Similarly, an asset with a negative alpha has expected return below its equilibrium return, and it should be sold. In the CAPM, abnormal returns should not continue indefinitely, and price should rise as a result of buying pressure so that abnormal profits will vanish. Forecasting the alpha of an asset, using a regression model based on the CAPM, one can decide whether to add it or not in a portfolio. While the CAPM is a cross-sectional model, it is common to cast the model into time series context and to test the hypothesis

$$H_0 : \alpha_1 = \alpha_2 = ... = 0$$

using historical data on the excess returns on the assets and the excess return on the market portfolio. A large number of tests rejected the Sharpe-Lintner version of the CAPM by showing that the regressions consistently got intercept greater than the average risk-free rate as well as a beta less than the average excess market return (see Black et al. [1972], Fama et al. [1973], Fama et al. [1992]). More recently, Fama et al. [2004] considered market return for 1928-2003 to estimate the predicted line, and confirmed that the relation between beta and average return is much flatter than the Sharpe-Lintner CAPM predicts. They also tested the prediction of mean-variance efficiency of the portfolio (portfolios are entirely explained by differences in market beta) by considering additional variables with cross-section regressions and time-series regressions. They found that standard market proxies seemed to be on the minimum variance frontier, that is, market betas suffice to explain expected returns and the risk premium for beta is positive, but the idea of SL-CAPM that the premium per unit of beta is the expected market return minus the risk-free rate was consistently rejected. Nonetheless, further research on the scaling of asset prices (or ratios), such as earning-price, debt-equity and book-to-market ratios (B/M), showed that much of the variation in expected return was unrelated to market beta (see Fama et al. [1992]). Among the possible explanations for the empirical failures of the CAPM, some refers to the behaviour of investors over-extrapolating past performance, while other point to the need of a more complex asset pricing model. For example, in the intertemporal capital asset pricing model (ICAPM) presented by Merton [1973], investors prefer high expected return and low return variance, but they are also concerned with the covariances of portfolio returns with state variables, so that portfolios are multifactor efficient. The ICAPM is a generalisation of the CAPM requireing additional betas, along with a market beta, to explain expected returns, and necessitate the specification of state variables affecting expected returns (see Fama [1996]). One approach is to derive an extension to the CAPM equilibrium where the systematic risk of a risky asset is related to the higher moments of the joint distribution between the return on an asset and the return on the market portfolio. Kraus et al. [1976] considered the coskewness to capture the asymmetric nature of returns on risky asset, and Fang et al. [1997] used the cokurtosis to capture the returns leptokurtosis. In both cases, the derivation of higher moment CAPM models is based on the higher moment extension of the investor's utility function derived in Appendix (A.7.4). Alternatively, to avoid specifying state variables, Fama et al. [1993] followed the arbitrage pricing theory by Ross [1976] and considered

unidentified state variables producing undiversifiable risks (covariances) in returns not captured by the market return and priced separately from market betas. For instance, the returns on the stocks of small firms covary more with one another than with returns on the stocks of large firms, and returns on high B/M stocks covary more with one another than with returns on low B/M stocks. As a result, Fama et al. [1993] [1996] proposed a three-factor model for expected returns given by

$$E[R_{it}] - R_{ft} = \beta_{iM}(E[R_{Mt}] - R_{ft}) + \beta_{is}E[SMB_t] + \beta_{ih}E[HML_t]$$

where SMB_t (small minus big) is the difference between the returns on diversified portfolios of small and big stocks, HML_t (high minus low) is the difference between the returns on diversified portfolios of high and low B/M stocks, and the betas are the slopes in the multiple regression of $R_{it} - R_{ft}$ on $R_{Mt} - R_{ft}$, SMB_t and HML_t. Given the time-series regression

$$E[R_{it}] - R_{ft} = \alpha_i + \beta_{iM}(E[R_{Mt}] - R_{ft}) + \beta_{is}E[SMB_t] + \beta_{ih}E[HML_t] + \epsilon_{it}$$

they found that the intercept α_i is zero for all assets i. Estimates of α_i from the time-series are used to calibrate the speed to which stock prices respond to new information as well as to measure the special information of portfolio managers such as performance. The momentum effect of Jegadeesh et al. [1993] states that stocks doing well relative to the market over the last three to twelve months tend to continue doing well for the next few months, and stocks doing poorly continue to do poorly. Even though this momentum effect is not explained by the CAPM or the three-factor model, one can add to these model a momentum factor consisting of the difference between the return on diversified portfolios of short-term winners and losers. For instance, Carhart [1997] proposed the four factor model

$$E[R_{it}] - R_{ft} = \alpha_i + \beta_{iM}(E[R_{Mt}] - R_{ft}) + \beta_{is}E[SMB_t] + \beta_{ih}E[HML_t] + \beta_{im}E[UMD_t] + \epsilon_{it}$$

where UMD_t is the monthly return of the style-attribution Carhart momentum factor.

2.3.2 The growth optimal portfolio

The growth optimal portfolio (GOP) is a portfolio having maximal expected growth rate over any time horizon, and as such, it is sure to outperform any other significantly different strategy as the time horizon increases. As a result, it is an investment tool for long horizon investors. Calculating the growth optimal strategy is in general very difficult in discrete time (in incomplete market), but it is much easier in the continuous time continuous diffusion case and was solved by Merton [1969]. Solutions to the problem exists in a semi-explicit form and in the general case, the GOP can be characterised in terms of the semimartinale characteristic triplet. Following Mosegaard Christensen [2011], we briefly review the discrete time case, providing the main properties of the GOP and extend the results to the continuous case. Details can be found in Algoet et al. [1988], Goll et al. [2000], Becherer [2001], Christensen et al. [2005].

2.3.2.1 Discrete time

Consider a market consisting of a finite number of non-dividend paying assets. The market consists of $N+1$ assets, represented by a $N+1$ dimensional vector process S where

$$S = \{S(t) = (S^0(t), .., S^N(t)), t \in [0, 1, .., T]\}$$

and T is assumed to be a finite number. The first asset S^0 is sometimes assumed to be risk-free from one period to the next, that is, it is a predictable process. The price of each asset is known at time t, given the information $\mathcal{F}_t$. Define the return process

$$R = \{R(t) = (R^0(t), .., R^N(t)), t \in [1, .., T]\}$$

by

$$R^i(t) = \frac{S^i(t)}{S^i(t-1)} - 1$$

Often it is assumed that returns are independent over time, and for simplicity this assumption is made in this section. Investors in such a market consider the choice of a strategy

$$b = \{b(t) = (b^0(t), .., b^N(t)), t \in [0, 1, .., T]\}$$

where $b^i(t)$ denotes the number of units of asset i that is being held during the period $(t, t+1]$.

Definition 2.3.1 *A trading strategy b generates the portfolio value process $S^b(t) = b(t).S(t)$. The strategy is called admissible if it satisfies the three conditions*

1. *Non-anticipative: the process b is adapted to the filtration $\mathcal{F}$.*
2. *Limited liability: the strategy generates a portfolio process $S^b(t)$ which is non-negative.*
3. *Self-financing: $b(t-1).S(t) = b(t).S(t)$ for $t \in [1, .., T]$ or equivalently $\Delta S^b(t) = b(t-1).\Delta S(t)$.*

where $x.y$ denotes the standard Euclidean inner product. The set of admissible portfolios in the market is denoted $\Theta(S)$, and $\underline{\Theta}(S)$ denotes the strictly positive portfolios. It is assumed that $\underline{\Theta}(S) \neq 0$. The third part requires that the investor re-invests all money in each time step. No wealth is withdrawn or added to the portfolio. This means that intermediate consumption is not possible. Consider an investor who invests a dollar of wealth in some portfolio. At the end of period T his wealth becomes

$$S^b(T) = S^b(0) \prod_{j=1}^{T} (1 + R^b(j))$$

where $R^b(t)$ is the return in period t. The ratio is given by

$$\frac{S^b(T)}{S^b(T-1)} = (1 + R^b(T))$$

If the portfolio fractions are fixed during the period, the right-hand-side is the product of T independent and identically distributed (i.i.d.) random variables. The geometric average return over the period is then

$$\Big(\prod_{j=1}^{T} (1 + R^b(j))\Big)^{\frac{1}{T}}$$

Because the returns of each period are i.i.d., this average is a sample of the geometric mean value of the one-period return distribution. For discrete random variables, the geometric mean of a random variable X taking (not necessarily distinct) values $x_1, ..., x_S$ with equal probabilities is defined as

$$G(X) = \Big(\prod_{s=1}^{S} x_s\Big)^{\frac{1}{S}} = \Big(\prod_{k=1}^{K} \tilde{x}_k^{f_k}\Big) = e^{E[\log X]}$$

where $\tilde{x}_k$ is the distinct values of X and f_k is the frequency of which $X = x_k$, that is, $f_k = P(X = x_k)$. In other words, the geometric mean is the exponential function of the growth rate

$$g^b(t) = E[\log(1 + R^b)(t)]$$

of some portfolio. Hence if Ω is discrete or more precisely if the σ-algebra $\mathcal{F}$ on Ω is countable, maximising the geometric mean is equivalent to maximising the expected growth rate. Generally, one defines the geometric mean of an arbitrary random variable by

$$G(X) = e^{E[\log X]}$$

assuming the mean value $E[\log X]$ is well defined. Over long stretches intuition dictates that each realised value of the return distribution should appear on average the number of times dictated by its frequency, and hence as the number of periods increase, it would hold that

$$\Big(\prod_{j=1}^{T}(1+R^b(j))\Big)^{\frac{1}{T}} = e^{\frac{1}{T}\sum_{j=1}^{T}\log S^b(j)} \to G(1+R^b(1))$$

as $T \to \infty$. This states that the average growth rate converges to the expected growth rate. In fact this heuristic argument can be made precise by an application of the law of large numbers. In multi-period models, the geometric mean was suggested by Williams [1936] as a natural performance measure, because it took into account the effects from compounding. Instead of worrying about the average expected return, an investor who invests repeatedly should worry about the geometric mean return. It explains why one might consider the problem

$$\sup_{S^b(T)\in\Theta} E[\log(\frac{S^b(T)}{S^b(0)})] \tag{2.3.9}$$

Definition 2.3.2 *A soulution $S^{\underline{b}}$ to Equation (2.3.9) is called a GOP.*

Hence the objective given by Equation (2.3.9) is often referred to as the geometric mean criteria. Economists may view this as the maximisation of expected terminal wealth for an individual with logarithmic utility. However, the GOP was introduced because of the properties of the geometric mean, when the investment horizon stretches over several periods. For simplicity it is always assumed that $S^b(0) = 1$, i.e. the investors start with one unit of wealth.

Definition 2.3.3 *An admissible strategy b is called an arbitrage strategy if*

$$S^b(0) = 0\,,\; P(S^b(T) \geq 0) = 1\,,\; P(S^b(T) > 0) > 0$$

It is closely related to the existence of a solution to problem in Equation (2.3.9), because the existence of a strategy that creates something out of nothing would provide an infinitely high growth rate.

Theorem 2.3.1 *There exists a GOP $S^{\underline{b}}$ if and only if there is no arbitrage. If the GOP exists its value process is unique.*

The necessity of no arbitrage is straightforward as indicated above. The sufficiency will follow directly once the numeraire property of the GOP has been established. It is possible to infer some simple properties of the GOP strategy, without further specifications of the model:

Theorem 2.3.2 *The GOP strategy has the following properties:*

1. *The fractions of wealth invested in each asset are independent of the level of total wealth.*
2. *The invested fraction of wealth in asset i is proportional to the return on asset i.*
3. *The strategy is myopic*

To see why the GOP strategy depends only on the distribution of asset returns one period ahead note that

$$E[\log S^b(T)] = \log S^b(0) + \sum_{j=1}^{T} E[\log(1+R^b(j))]$$

In general, obtaining the strategy in an explicit closed form is not possible, as it involves solving a non-linear optimisation problem. To see this, we derive the first order conditions of Equation (2.3.9). Since the GOP strategy is myopic and the invested fractions are independent of wealth, one needs to solve the problem

$$\sup_{b(t)} E_t[\log(\frac{S^b(t+1)}{S^b(t)})]$$

for each $t \in [0, 1, .., T-1]$. This is equivalent to solving the problem

$$\sup_{b(t)} E_t[\log(1 + R^b(t+1))]$$

Using the fractions $\pi_b^i(t) = \frac{b^i(t)S^i(t)}{S^b(t)}$ the problem can be written

$$\sup_{\pi_b(t)\in\mathbb{R}^N} E[\log\big(1 + (1 - \sum_{k=1}^N \pi_b^k(t))R^0(t+1) + \sum_{k=1}^N \pi_b^k(t)R^k(t+1)\big)]$$

since

$$1 + (1 - \sum_{i=1}^N \pi_b^i(t))R^0(t+1) + \sum_{i=1}^N \pi_b^i(t)R^i(t+1) =$$
$$\frac{1}{S^b(t)}\big((1 + R^0(t+1))S^b(t) - \sum_{i=1}^N b^i(t)S^i(t)R^0(t+1) + \sum_{i=1}^N b^i(t)S^i(t)R^i(t+1)\big)$$

which gives

$$1 + (1 - \sum_{i=1}^N \pi_b^i(t))R^0(t+1) + \sum_{i=1}^N \pi_b^i(t)R^i(t+1) =$$
$$\frac{1}{S^b(t)}\big((1 + R^0(t+1))S^b(t) - (1 + R^0(t+1))\sum_{i=1}^N b^i(t)S^i(t) + \sum_{i=1}^N b^i(t)S^i(t+1)\big)$$

Since $S^b(t) - b^0(t)S^0(t) = \sum_{i=1}^N b^i(t)S^i(t)$, we get

$$1 + (1 - \sum_{i=1}^N \pi_b^i(t))R^0(t+1) + \sum_{i=1}^N \pi_b^i(t)R^i(t+1) =$$
$$\frac{1}{S^b(t)}\big((1 + R^0(t+1))b^0(t)S^0(t) + \sum_{i=1}^N b^i(t)S^i(t+1)\big)$$

and since the portfolio is self-financing, we get

$$1 + (1 - \sum_{i=1}^N \pi_b^i(t))R^0(t+1) + \sum_{i=1}^N \pi_b^i(t)R^i(t+1) = \frac{S^b(t+1)}{S^b(t)}$$

The properties of the logarithm ensures that the portfolio will automatically become admissible. By differentiation, the first order conditions become

$$E_{t-1}[\frac{1+R^k(t)}{1+R^{\underline{b}}(t)}] = 1 \ , \ k = 0, 1, .., N$$

This constitutes a set of $N+1$ non-linear equation to be solved simultaneously such that one of which is a consequence of the others, due to the constraint that $\sum_{i=0}^{N} \pi_b^i = 1$. Although these equations do not generally posses an explicit closed-form solution, there are some special cases which can be handled.

2.3.2.2 Continuous time

Being a $(N+1)$-dimensional semimartingale and satisfying the usual conditions, S can be decomposed as

$$S(t) = A(t) + M(t)$$

where A is a finite variation process and M is a local martingale. The reader is encouraged to think of these as drift and volatility respectively, but should beware that the decomposition above is not always unique. If A can be chosen to be predictable, then the decomposition is unique. This is exactly the case when S is a special semimartingale (see Protter [2004]). Following standard conventions, the first security is assumed to be the numeraire, and hence it is assumed that $S^0(t) = 1$ almost surely for all $t \in [0, T]$. The investor needs to choose a strategy, represented by the $N+1$ dimensional process

$$b = \{b(t) = (b^0(t), .., b^N(t)), t \in [0, T]\}$$

Definition 2.3.4 *An admissible trading strategy b satisfies the three conditions:*

1. *b is an S-integrable, predictable process.*
2. *The resulting portfolio value $S^b(t) = \sum_{i=0}^{N} b^i(t)S^i(t)$ is nonnegative.*
3. *The portfolio is self-financing, that is $S^b(t) = \int_0^t b^s dS(s)$.*

The last requirement states that the investor does not withdraw or add any funds. It is often convenient to consider portfolio fractions, i.e

$$\pi_b = \{\pi_b(t) = (\pi_b^0(t), .., \pi_b^N(t))^\top, t \in [0, \infty)\}$$

with coordinates defined by

$$\pi_b^i(t) = \frac{b^i(t)S^i(t)}{S^b(t)}$$

One may define the GOP $S^{\underline{b}}$ as the solution to the problem

$$S^{\underline{b}} = arg \sup_{S^b \in \underline{\Theta}(S)} E[\log(\frac{S^b(T)}{S^b(0)})] \tag{2.3.10}$$

Definition 2.3.5 *A portfolio is called a GOP if it satisfies Equation (2.3.10).*

The essential feature of No Free Lunch with Vanishing Risk (NFLVR) is the fact that it implies the existence of an equivalent martingale measure. More precisely, if asset prices are locally bounded, the measure is an equivalent local martingale measure and if they are unbounded, the measure becomes an equivalent sigma martingale measure. Here, these measures will all be referred to collectively as equivalent martingale measures (EMM).

Theorem 2.3.3 *Assume that*

$$\sup_{S^b} E[\log(\frac{S^b(T)}{S^b(0)})] < \infty$$

and that NFLVR holds. Then there is a GOP.

A less stringent and numeraire invariant condition is the requirement that the market should have a martingale density. A martingale density is a strictly positive process Z, such that $\int SdZ$ is a local martingale. In other words, a Radon-Nikodym derivative of some EMM is a martingale density, but a martingale density is only the Radon-Nikodym derivative of an EMM if it is a true martingale. Modifying the definition of the GOP slightly, one may show that:

***Corollary* 1** *There is a GOP if and only if there is a martingale density.*

We present a simple example to get a feel of how to finf the growth optimal strategy in the continuous setting.

Example : two assets

Let the market consist of two assets, a stock and a bond. Specifically the SDEs describing these assets are given by

$$\begin{aligned} dS^0(t) &= S^0(t)rdt \\ dS^1(t) &= S^1(t)\big(adt + \sigma dW(t)\big) \end{aligned}$$

where W is a Wiener process and r, a, σ are constants. Since $S^b(t) = b^0(t)S^0(t) + b^1(t)S^1(t)$, applying Ito's lemma we get

$$dS^b(t) = b^0(t)S^0(t)rdt + b^1(t)S^1(t)adt + b^1(t)S^1(t)\sigma dW(t)$$

Using fractions $\pi^1(t) = \frac{b^1(t)S^1(t)}{S^b(t)}$, any admissible strategy can be written

$$dS^b(t) = S^b(t)\big((r + \pi^1(t)(a-r))dt + \pi^1(t)\sigma dW(t)\big)$$

since

$$dS^b(t) = S^b(t)\big(rdt + \frac{b^1(t)S^1(t)}{S^b(t)}(a-r)dt + \frac{b^1(t)S^1(t)}{S^b(t)}\sigma dW(t)\big)$$

which gives

$$dS^b(t) = (S^b(t) - b^1(t)S^1(t))rdt + b^1(t)S^1(t)adt + b^1(t)S^1(t)\sigma dW(t)\big)$$

and since $S^b(t) - b^1(t)S^1(t) = b^0(t)S^0(t)$, we recover the SDE

$$dS^b(t) = b^0(t)S^0(t)rdt + b^1(t)S^1(t)adt + b^1(t)S^1(t)\sigma dW(t)\big)$$

Applying Ito's lemma to $Y(t) = \log S^b(t)$ we get

$$dY(t) = \big(\big[r + \pi^1(t)(a-r) - \frac{1}{2}(\pi^1(t))^2\sigma^2\big]dt + \pi^1(t)\sigma dW(t)\big)$$

Hence, assuming the local martingale with differential $\pi^1(t)\sigma dW(t)$ to be a true martingale, it follows that

$$E[\log S^b(T)] = E[\int_0^T \big[r + \pi^1(t)(a-r) - \frac{1}{2}(\pi^1(t))^2\sigma^2\big]dt]$$

so by maximizing the expression for each (t, ω) the optimal fraction is obtained as

$$\pi^1_{\underline{b}}(t) = \frac{a-r}{\sigma^2}$$

Hence, inserting the optimal fractions into the wealth process, the GOP is described by the SDE

$$dS^b(t) = S^b(t)\big((r + (\frac{a-r}{\sigma})^2)dt + \frac{a-r}{\sigma}dW(t)\big)$$

which we rewrite as

$$dS^b(t) = S^b(t)\big((r+\theta^2)dt + \theta dW(t)\big)$$

where $\theta = \frac{a-r}{\sigma}$ is the market price of risk process.

Fix a truncation function h i.e. a bounded function with compact support $h : \mathbb{R}^N \to \mathbb{R}^N$ such that $h(x) = x$ in a neighbourhood around zero. For instance, a common choice would be

$$h(x) = xI_{\{|x|\leq 1\}}$$

For such truncation function, there is a triplet (A, B, ν) describing the behaviour of the semimartingale. There exists a locally integrable, increasing, predictable process $\hat{A}$ such that (A, B, ν) can be written as

$$A = \int a d\hat{A}\ ,\ B = \int b d\hat{A} \text{ and } \nu(dt, dv) = d\hat{A}_t F(t, dv)$$

The process A is related to the finite variation part of the semimartingale, and it can be thought of as a generalised drift. The process B is similarly interpreted as the quadratic variation of the continuous part of S, or in other words it is the square volatility where volatility is measured in absolute terms. The process ν is the compensated jump measure, interpreted as the expected number of jumps with a given size over a small interval and F essentially characterises the jump size.

Example
Let S^1 be geometric Brownian Motion. Then $\hat{A} = t$ and

$$dA(t) = S^1(t)adt\ ,\ dB(t) = (S^1(t)\sigma)^2 dt$$

Theorem 2.3.4 *(Goll and Kallsen [2000])*
Let S have a characteristic triplet (A, B, ν) as described above. Suppose there is an admissible strategy $\underline{b}$ with corresponding fractions $\pi_{\underline{b}}$ such that

$$a^k(t) - \sum_{i=1}^N \frac{\pi^i_{\underline{b}}}{S^i(t)} b^{i,k}(t) + \int_{\mathbb{R}^N} \big(\frac{x^k}{1 + \sum_{i=1}^N \frac{\pi^i_{\underline{b}}}{S^i(t)} x^i} - h(x)\big) F(t, dx) = 0$$

for $\mathbb{P} \times d\hat{A}$ almost all $(\omega, t) \in \Omega \times [0, T]$ where $k \in [0, .., N]$ and $\times$ denotes the standard product measure. Then $\underline{b}$ is the GOP strategy.

This Equation represents the first order conditions for optimality and they would be obtained easily if one tried to solve the problem in a pathwise sense.

Example
Assume that discounted asset prices are driven by an m-dimensional Wiener process. The locally risk free asset is used as numeraire, whereas the remaining risky assets evolve according to

$$dS^i(t) = S^i(t)a^i(t)dt + \sum_{k=1}^{m} S^i(t)b^{i,k}(t)dW^k(t)$$

for $i \in [1, .., N]$. Here $a^i(t)$ is the excess return above the risk free rate. From this equation, the decomposition of the semimartingale S follows directly. Choosing $\hat{A} = t$, a good version of the characteristic triplet becomes

$$(A, B, \nu) = \big(\int a(t)S(t)dt, \int S(t)b(t)(S(t)b(t))^\top dt, 0\big)$$

Consequently, in vector form and after division by $S^i(t)$, the above Equation yields that

$$a(t) - (b(t)b(t)^\top)\pi_{\underline{b}}(t) = 0$$

In the particular case where $m = N$ and the matrix b is invertible, we get the well-known result that

$$\pi(t) = b^{-1}(t)\theta(t)$$

where $\theta(t) = b^{-1}(t)a(t)$ is the market price of risk. Generally, whenever the asset prices can be represented by a continuous semimartingale, a closed form solution to the GOP strategy may be found. The cases where jumps are included are less trivial. In general when jumps are present, there is no explicit solution in an incomplete market. In such cases, it is necessary to use numerical methods.

As it was done in discrete time, the GOP can be characterised in terms of its growth properties.

Theorem 2.3.5 *The GOP has the following properties:*

1. *The GOP maximises the instantaneous growth rate of investments*
2. *In the long term, the GOP will have a higher realised growth rate than any other strategy, i.e.*

$$\lim_{T\to\infty} \sup \frac{1}{T} \log S^b(T) \leq \lim_{T\to\infty} \sup \frac{1}{T} \log S^{\underline{b}}(T)$$

for any other admissible strategy S^b.

The instantaneous growth rate is the drift of $\log S^b(t)$.

Example
Given the previous example, the instantaneous growth rate $g^b(t)$ of a portfolio S^b was found by applying the Ito's formula to get

$$dY(t) = \big(\big[r + \pi(t)(a - r) - \frac{1}{2}\pi^2(t)\sigma^2\big]dt + \pi(t)\sigma dW(t)\big)$$

Hence, the instantaneous growth rate is

$$g^b(t) = r + \pi(t)(a - r) - \frac{1}{2}\pi^2(t)\sigma^2 \tag{2.3.11}$$

Differentiating the instantaneous growth rate $g^b(t)$ with respect to the fraction $\pi(t)$ and setting the results to zero, we recover $\pi^1_b(t)$ in the Example with two assets. Hence, by construction, the GOP maximise the instantaneous growth rate. As in the discrete setting, the GOP enjoys the numeraire property. However, there are some subtle differences.

Theorem 2.3.6 *Let S^b denote any admissible portfolio process and define $\hat{S}^b(t) = \frac{S^b(t)}{S^{\underline{b}}(t)}$. Then*

1. *$\hat{S}^b(t)$ is a supermartingale if and only if $S^{\underline{b}}(t)$ is the GOP.*
2. *The process $\frac{1}{\hat{S}^b(t)}$ is a submartingale.*
3. *If asset prices are continuous, then $\hat{S}^b(t)$ is a local martingale.*

2.3.2.3 Discussion

Mosegaard Christensen [2011] reviewed the discussion around the attractiveness of the GOP against the CAPM and concluded that there is an agreement on the fact that the GOP can neither proxi for, nor dominate other strategies in terms of expected utility, and no matter how long (finite) horizon the investor has, utility based preferences can make other portfolios more attractive because they have a more appropriate risk profile. Authors favouring the GOP believe growth optimality to be a reasonable investment goal, with attractive properties being relevant to long horizon investors (see Kelly [1956], Latane [1959]). On the other hand, authors disagreeing do so because they do not believe that every investor could be described as log-utility maximising investors (see Markowitz [1976]). To summarise, the disagreement had its roots in two very fundamental issues, namely whether or not utility theory is a reasonable way of approaching investment decisions in practice, and whether utility functions, different from the logarithm, is a realistic description of individual long-term investors. The fact that investors must be aware of their own utility functions is a very abstract statement which is not a fundamental law of nature. Once a portfolio has been build using the CAPM, it is impossible to verify ex-post that it was the right choice. On the other hand, maximising growth over time is formulated in dollars, so that one has a good idea of the final wealth he will get.

2.3.2.4 Comparing the GOP with the MV approach

The main results on the comparison between mean variance (MV) and growth optimality can be found in Hakansson [1971]. Mosegaard Christensen [2011] presented a review of whether or not the GOP and MV approach could be united or if they were fundamentally different. He concluded by stating that they were in general two different things. The mean-variance efficient portfolios are obtained as the solution to a quadratic optimisation program. Its theoretical justification requires either a quadratic utility function or some fairly restrictive assumptions on the class of return distribution, such as the assumption of normally distributed returns. As a result, the GOP is in general not mean-variance efficient. Mean-variance efficient portfolios have the possibility of ruin, and they are not consistent with first order stochastic dominance. We saw earlier that the mean-variance approach was further developed into the CAPM, where the market is assumed mean-variance efficient. Similarly, it was assumed that if all agents were to maximise the expected logarithm of wealth, then the GOP becomes the market portfolio and from this an equilibrium asset pricing model appears. As with the CAPM, the conclusion of the analysis provide empirically testable predictions. As a result of assuming log-utility, the martingale or numeraire condition becomes a key element of the equilibrium model. Recall, $R^i(t)$ is the return on the ith asset between time $t-1$ and t, and $R^{\underline{b}}$ is the return process for the GOP. Then the equilibrium is

$$1 = E_{t-1}[\frac{1+R^i(t)}{1+R^{\underline{b}}}]$$

which is the first order condition for a logarithmic investor. We assume a world with a finite number of states, $\Omega = \{\omega_1, .., \omega_n\}$ and define $p_i = \mathbb{P}(\{\omega_i\})$. Then if $S^i(t)$ is an Arrow-Debreu price, paying off one unit of wealth at time $t+1$, we get

$$S^i(t) = E_t[\frac{I_{\{\omega=\omega_i\}}}{1+R^{\underline{b}}(t+1)}]$$

and consequently summing over all states provides an equilibrium condition for the risk-free rate

$$1 + r(t, t+1) = E_t[\frac{1}{1+R^{\underline{b}}(t+1)}]$$

Combining with the previous equations, defining $\overline{R}^i = R^i - r$, and performing some basic calculations, we get

$$E_t[\overline{R}^i(t+1)] = \beta_t^i E_t[\overline{R}^b(t+1)]$$

where

$$\beta_t^i = \frac{Cov(\overline{R}^i(t+1), \frac{\overline{R}^b(t+1)}{R^b(t+1)})}{Cov(\overline{R}^b(t+1), \frac{\overline{R}^b(t+1)}{R^b(t+1)})}$$

This is to be compared with the CAPM, where the β is given by

$$\beta_{CAPM} = \frac{Cov(R^i, R^*)}{Var(R^*)}$$

In some cases only is the CAPM and the CAPM based on the GOP similar. Note, the mean-variance portfolio provides a simple trade-off between expected return and variance which can be parametrised in a closed-form, requiring only the estimation of a variance-covariance matrix of returns and the ability to invert that matrix. Further, choosing a portfolio being either a fractional Kelly strategy or logarithmic mean-variance efficient provides the same trade-off, but it is computationally more involved.

In the continuous case, for simplicity of exposition, we use the notation given in Appendix (F.1) and rewrite the dynamics of the ith risky asset as

$$\frac{dS^i(t)}{S^i(t)} = a_t^i dt + \sigma_t^i dW_t$$

where dW_t is a column vector of dimension $(M, 1)$ of independent Brownian motions with elements $(dW_t^j)_{j=1}^M$, and σ_t^i is a volatility matrix of dimension $(1, N)$ with elements $(\sigma_j^i(t))_{j=1}^N$ such that

$$< \sigma_t^i, dW_t >= \sum_{j=1}^M \sigma_j^i(t) dW_t^j$$

with Euclidean norm

$$|\sigma_t^i|^2 = \sum_{j=1}^M (\sigma_j^i(t))^2$$

In that setting, the portfolio in Equation (F.1.1) becomes

$$dV_t^b = r_t V_t^b dt + < (bS)_t, a_t - r_t I > dt + < (bS)_t, \sigma_t dW_t >$$

where we let σ_t be an adapted matrix of dimension $N \times M$, $(bS)_t$ corresponds to the vector with component $(b^i(t)S_t^i)_{1 \leq i \leq N}$ describing the amount to be invested in each stock. Also, $\frac{1}{S^0(t)}\left(V_t^b - (bS)_t\right)$ is invested in the riskless asset. Writting $\pi(t) = \frac{(bS)_t}{V_t^b}$ as a $(N, 1)$ vector with elements $(\pi^i(t))_{i=1}^N = \frac{(b^i(t)S_t^i)_{1 \leq i \leq N}}{V_t^b}$, the dynamics of the portfolio become

$$\frac{dV_t^b}{V_t^b} = r_t dt + < \pi(t), a_t - r_t I > dt + < \pi(t), \sigma_t dW_t >$$

In that setting, the instantaneous mean-variance efficient portfolio is the solution to the probelem

$$\sup_{b\in\underline{\Theta}(S)} a^b(t)$$
$$\text{s.t. } \sigma_t^b \leq \kappa(t)$$

where $\kappa(t)$ is some non-negative adapted process. Defining the process $Y_t^b = \ln V_t^b$ and applying Ito's lemma, we get

$$dY_t^b = r_t dt + < \pi(t), a_t - r_t I > dt - \frac{1}{2}|\pi(t)\sigma_t^i|^2 dt + < \pi(t), \sigma_t dW_t >$$

with instantaneous growth rate being

$$g^b(t) = r_t + < \pi(t), a_t - r_t I > -\frac{1}{2}|\pi(t)\sigma_t^i|^2$$

which is a generalisation of Equation (2.3.11). By construction, the GOP maximise the instantaneous growth rate. Taking the expectation, we get

$$E[Y_T^b] = E[\int_0^T \big(r_s + < \pi(s), a_s - r_s I > -\frac{1}{2}|\pi(s)\sigma_s^i|^2\big)ds]$$

Note, we can define the minimal market price of risk as

$$\theta(t) = \sigma_t \frac{a_t - r_t I}{\sigma_t \sigma_t^\top}$$

Any efficient portfolio along the straight efficient frontier can be specified by its fractional holdings of the market portfolio, called the leverage and denoted by α. The instantaneously mean-variance efficient portfolios have fractions solving the equation

$$\pi_{\underline{b}}(t)\sigma_t = \alpha(t)\theta(t) = \alpha(t)\sigma_t \frac{a_t - r_t I}{\sigma_t \sigma_t^\top}$$

for some non-negative process α. Hence, the optimum fractions become

$$\pi_{\underline{b}}(t) = \alpha(t) \frac{a_t - r_t I}{\sigma_t \sigma_t^\top} \tag{2.3.12}$$

In the special case where we assume the volatility matrix to be of dimension (N, N) and invertible, the market price of risk become $\theta(t) = \frac{a_t - r_t I}{\sigma_t}$ and the optimum fractions simplify to

$$\pi_{\underline{b}}(t) = \alpha(t) \frac{\theta(t)}{\sigma_t} \tag{2.3.13}$$

Using the optimum fractions, the SDE for such leveraged portfolios become

$$\frac{dV_t^{\underline{b}}}{V_t^{\underline{b}}} = r_t dt + \alpha(t)|\theta(t)|^2 dt + \alpha(t) < \theta(t), dW_t >$$

where the volatility of the portfolio is now θ_t. The GOP is instantaneously mean-variance efficient,corresponding to the choice of $\alpha = 1$. The GOP belongs to the class of instantaneous Sharpe ratio maximising strategies, where for some strategy b, it is defined as

$$M_{SR}^b = \frac{b_t r_t + < b_t, a_t - r_t I > -r_t}{|(b\sigma_j)_t|^2} = \frac{< b_t, a_t > + r_t(b_t^0 - 1)}{|(b\sigma_j)_t|^2} \tag{2.3.14}$$

where $(b\sigma_j)_t$ is a weighted volatility vector with elements $(b_t^i \sigma_j^i(t))_{i=1}^N$. Recall, $b_t r_t - < b_t, r_t I >= b_t^0 r_t$, so that the mean-variance portfolios consist of a position in the GOP and the rest in the riskless asset, that is, a fractional Kelly strategy.

2.3.2.5 Time taken by the GOP to outperfom other portfolios

As the GOP was advocated, not as a particular utility function, but as an alternative to utility theory relying on its ability to outperform other portfolios over time, it is important to document this ability over horizons relevant to actual investors. We present a simple example illustrating the time it takes for the GOP to dominate other assets.

Example Assume a two asset Black-Scholes model with constant parameters with risk-free asset $S^0(t) = e^{rt}$ and and solving the SDE, the stock price is given as

$$S^1(t) = e^{(a-\frac{1}{2}\sigma^2)t+\sigma W(t)}$$

The GOP is given by the process

$$S^{\underline{b}}(t) = e^{(r-\frac{1}{2}\theta^2)t+\theta W(t)}$$

where $\theta = \frac{a-r}{\sigma}$. Some simple calculations imply that the probability

$$P_0(t) = P(S^{\underline{b}}(t) \geq S^0(t))$$

of the GOP outperforming the savings account over a period of length t and the probability

$$P_1(t) = P(S^{\underline{b}}(t) \geq S^1(t))$$

of the GOP outperforming the stock over a period of length t are given by

$$P_0(t) = N(\frac{1}{2}\theta\sqrt{t})$$

and

$$P_1(t) = N(\frac{1}{2}|\theta - \sigma|\sqrt{t})$$

where the cumulative distribution function of the standard Gaussian distribution $N(.)$ are independent of the short rate. Moreover, the probabilities are increasing in the market price of risk and time horizon. They converge to one as the time horizon increases to infinity, which is a manifestation of the growth properties of the GOP. The time needed for the GOP to outperform the risk free asset for a 99% confidence level is 8659 year for a market price of risk $\theta = 0.05$ and 87 year for $\theta = 0.5$. Similarly, for a 95% confidence level the time is 4329 year for $\theta = 0.05$ and 43 year for $\theta = 0.5$. One can conclude that the long run may be very long. Hence, the argument that one should choose the GOP to maximise the probability of doing better than other portfolios is somewhat weakened.

2.3.3 Measuring and predicting performances

We saw in Section (2.3.1.1) that since any point along the efficient frontier represents an efficient portfolio, the investor needs additional information in order to select the optimal portfolio. That is, the key element in mean-variance portfolio analysis being one's view on expected return and risk, the selection of a preferred combination of risk and expected return depends on the investor. However, we saw in Section (2.3.1.2) that one can attempt at finding efficient portfolios promising the greatest expected return for a given degree of risk (see Sharpe [1966]). Hence, one must translate predictions about security performance into predictions of portfolio performance, and select one efficient portfolio based on some utility function. The process for mutual funds becomes that of security analysis and portfolio analysis given some degree of risk. As a result, there is room for major and persisting differences in the performance of different funds. Over time, security analysis moved towards evaluating the interrelationships between securities, while portfolio analysis focused more on diversification as any diversified portfolio should be efficient in a perfect market. For example, one may only require the spreading of holdings among standard industrial classes.

In the CAPM presented in Section (2.3.1.2), one assumes that the predicted performance of the ith portfolio is described with two measures, namely the expected rate of return E_{R_i} and the predicted variability or risk expressed as the standard deviation of return σ_i. Further, assuming that all investors can invest and borrow at the risk-free rate, all efficient portfolios satisfy Equation (2.3.5) and follow the linear representation

$$E_{R_i} = a + b\sigma_i$$

where a is the risk-free rate and b is the risk premium. Hence, by allocating his funds between the ith portfolio and borrowing or lending, the investor can attain any point on the line

$$E_R = a + \frac{E_{R_i} - a}{\sigma_i}\sigma_R$$

for a given pair (E_R, σ_R), such that the best portfolio is the one for which the slope $\frac{E_{R_i}-a}{\sigma_i}$ is the greatest (see Tobin [1958]). The predictions of future performance being difficult to obtain, ex-post values must be used in the model. That is, the average rate of return of a portfolio $\overline{R}_i$ must be substituted for its expected rate of return, and the actual standard deviation $\overline{\sigma}_i$ of its rate of return for its predicted risk. In the ex-post settings, funds with properly diversified portfolios should provide returns giving $\overline{R}_i$ and $\overline{\sigma}_i$ lying along a straight line, but if they fail to diversify the returns will yield inferior values for $\overline{R}_i$ and $\overline{\sigma}_i$. In order to analyse the performances of different funds, Sharpe [1966] proposed a single measure by substituting the ex-post measures $\overline{R}$ and $\overline{\sigma}_R$ for the ex-ante measures E_R and σ_R obtaining the formula

$$\overline{R} = a + \frac{\overline{R}_i - a}{\overline{\sigma}_i}\overline{\sigma}_R$$

which is a reward-to-variability ratio (RV) or a reward per unit of variability. The numerator is the reward provided the investor for bearing risk, and the denominator measures the standard deviation of the annual rate of return. The results of his analysis based on 34 funds on two periods, from 1944 till 1953 and from 1954 till 1963, showed that differences in performance can be predicted, although imperfectly, but one can not identifies the sources of the differences. Further, there is no assurance that past performance is the best predictor of future performance. During the period 1954-63, almost 90% of the variance of the return a typical fund of the sample was due to its comovement with the return of the other securities used to compute the Dow-Jones Industrial Average, with a similar percentage for most of the 34 funds. Taking advantage of this relationship, Treynor [1965] used the volatility of a fund as a measure of its risk instead of the total variability used in the RV ratio. Letting B_i be the volatility of the ith fund defined as the change in the rate of return of the fund associated with a 1% change in the rate of return of a benchmark or index, the Treynor index can be written as

$$M_{TI} = \frac{\overline{R}_i - a}{B_i}$$

According to Sharpe [1966], Treynor intended that his index be used both for measuring a fund's performance, and for predicting its performance in the future. Treynor [1965] argued that a good historical performance pattern is one which, if continued in the future, would cause investors to prefer it to others. Given the level of contribution of volatility to the over-all variability, one can expect the ranking of funds on the basis of the Treynor index to be very close to that based on the RV ratio, especially when funds hold highly diversified portfolios. Differences appear in the case of undiversified funds since the TI index do not capture the portion of variability due to the lack of diversification. For this reason Sharpe concluded that the TI ratio was an inferior measure of past performance but a possibly superior measure for predicting future performance.

Note, independently from Markowitz, Roy [1952] set down the same equation relating portfolio variance of return to the variances of return of the constituent securities, developing a similar mean-variance efficient set. However, while Markowitz left it up to the investor to choose where along the efficient set he would invest, Roy advised choosing the single portfolio in the mean-variance efficient set maximising

$$\frac{\mu - d}{\sigma_M^2}$$

where d is a disaster level return the investor places a high priority on not falling below. It is very similar to the reward-to-variability ratio (RV) proposed by Sharpe. In its measure of quality and performance of a portfolio, Sharpe [1966] did not distinguish between time and ensemble averages. We saw in Section (2.3.2.4) that the measure is also meaningful in the context of time averages in geometric Brownian motion and derived in Equation (2.3.14) the GOP Sharpe ratio. Assuming a portfolio following the simple geometric Brownian motion in Equation (2.3.3), Peters [2011c] derived the dynamics of the leveraged portfolio, and, applying Ito's lemma to obtain the dynamics of the log-portfolio, computed the time-average leveraged exponential growth rate as

$$g_\alpha^b = \frac{1}{dt} < d \ln V_t^\alpha >= r + \alpha\mu - \frac{1}{2}\alpha^2\sigma_M^2$$

where σ_M is the volatility of the market portfolio. Differentiating with respect to α and setting the result to zero, the optimum leverage becomes

$$\alpha^* = \frac{\mu}{\sigma_M^2}$$

corresponding to the optimum fraction in Equation (2.3.13) with $\alpha = 1$. Note, Peters chose to optimise the leverage rather than optimising the fraction $\pi(t)$. Differing from the Sharpe ratio for the market portfolio only by a square in the volatility, the optimum leverage or GOP Sharpe ratio is also a fundamental measure of the quality and performance of a portfolio. Further, unless the Sharpe ratio, the optimum leverage is a dimensionless quantity, and as such can distinguish between fundamentally different dynamical regimes.

2.3.4 Predictable variation in the Sharpe ratio

The Sharpe ratio (SR) is the most common measure of risk-adjusted return used by private investors to assess the performance of mutual funds (see Modigliani et al. [1997]). Given evidence on predictable variation in the mean and volatility of equity returns (see Fama et al. [1989]), various authors studied the predictable variation in equity market SRs. However, due to the independence of the sample mean and sample variance of independently normally distributed variables (see Theorem (B.8.3)), predictable variation in the individual moments does not imply predictable variation in the Sharpe ratio. One must therefore ask whether these moments move together, leading to SRs which are more stable and potentially less predictable than the two components individually. The intuition being that volatility in the SR is not a good proxy for priced risk. Using regression analysis, some studies suggested a negative relation between the conditional mean and volatility of returns, indicating the likelihood of substantial predictable variation in market SRs. Using linear functions of four predetermined financial variables to estimate conditional moments, Whitelaw [1997] showed that estimated conditional SRs exhibit substantial time-variation that coincides with the variation in ex-post SRs and with the phases of the business cycle. For instance, the conditional SRs had monthly values ranging from less than -0.3 to more than 1.0 relative to an unconditional SR of 0.14 over the full sample period. This variation in estimated SRs closely matches variation in ex-post SRs measured over short horizons. Subsamples chosen on the basis of in-sample regression have SRs more than three times larger than SRs over the full sample. On an out-sample basis, using 10-year rolling regressions, subsample SRs exhibited similar magnitudes. As a result, Whitelaw showed that relatively naive market-timing strategies exploiting this predictability could generate SRs more than 70% larger than a buy-and-hold strategy. These active trading strategies involve switching between the market and the risk-free asset depending on the level of the estimated SR relative to a specific threshold. This result is critical in asset allocation decisions, and it has implications for the use of SRs in investment performance evaluation.

While the Sharpe ratio is regarded as a reliable measure during periods of increasing stock prices, it leads to erroneous conclusions during periods of declining share prices. However, there are still contradictions in the literature with respect to the interpretation of the SR in bear market periods. Scholz et al. [2006] showed that ex-post Sharpe

ratios do not allow for meaningful performance assessment of funds during non-normal periods. Using a single factor model, they showed the resulting SRs to be subject to random market climates (random mean and standard deviation of market excess returns). Considering a sample of 532 US equity mutual funds, funds exhibiting relatively high proportions of fund-specific risk showed on average superior ranking according to the SR in bear markets, and vice versa. Using regression analysis, they ascertained that the SRs of funds significantly depend on the mean excess returns of the market.

2.4 Risk and return analysis

Asset managers employ risk metrics to provide their investors with an accurate report of the return of the fund as well as its risk. Risk measures allow investors to choose the best strategies per rebalancing frequency in a more robust way. Performance evaluation of any asset, strategy, or fund tends to be done on returns that are adjusted for the average risk taken. We call active return and active risk the return and risk measured relative to a benchmark. Since all investors in funds have some degree of risk aversion and require limits on the active risk of the funds, they consider the ratio of active return to active risk in a risk adjusted performance measure (RAPM) to rank different investment opportunities. In general, RAPMs are used to rank portfolios in order of preference, implying that preferences are already embodied in the measure. However, we saw in Section (1.6.2) that to make a decision we need a utility function. While some RAPMs have a direct link to a utility function, others are still used to rank investments but we can not deduce anything about preferences from their ranking so that no decision can be based on their ranks (see Alexander [2008]).

The three measures by which the risk/return framework describes the universe of assets are the mean (taken as the arithmetic mean), the standard deviation, and the correlation of an asset to other assets' returns. Concretely, historical time series of assets are used to calculate the statistics from it, then these statistics are interpreted as true estimators of the future behaviour of the assets. In addition, following the central limit theorem, returns of individual assets are jointly normally distributed. Thus, given the assumption of a Gaussian (normal) distribution, the first two moments suffice to completely describe the distribution of a multi-asset portfolio. As a result, adjustment for volatility is the most common risk adjustment leading to Sharpe type metrics. Implicit in the use of the Sharpe ratio is the assumption that the preferences of investors can be represented by the exponential utility function. This is because the tractability of an exponential utility function allows an investor to form optimal portfolios by maximising a mean-variance criterion. However, some RAPMs are based on downside risk metrics which are only concerned with returns falling short of a benchmark or threshold returns, and are not linked to a utility function. Nonetheless these metrics are used by practitioners irrespectively of their theoretical foundation.

2.4.1 Some financial meaning to alpha and beta

2.4.1.1 The financial beta

In finance, the beta of a stock or portfolio is a number describing the correlated volatility of an asset in relation to the volatility of the benchmark that this asset is being compared to. We saw in Section (2.3.1.2) that the beta coefficient was born out of linear regression analysis of the returns of a portfolio (such as a stock index) (x-axis) in a specific period versus the returns of an individual asset (y-axis) in a specific year. The regression line is then called the Security characteristic Line (SCL)

$$SCL : R_a(t) = \alpha_a + \beta_a R_m(t) + \epsilon_t$$

where α_a is called the asset's alpha and β_a is called the asset's beta coefficient. Note, if we let R_f be a constant rate, we can rewrite the SCL as

$$SCL : R_a(t) - R_f = \alpha_a + \beta_a(R_m(t) - R_f) + \epsilon_t \tag{2.4.15}$$

Both coefficients have an important role in Modern portfolio theory. For

- $\beta < 0$ the asset generally moves in the opposite direction as compared to the index.
- $\beta = 0$ movement of the asset is uncorrelated with the movement of the benchmark
- $0 < \beta < 1$ movement of the asset is generally in the same direction as, but less than the movement of the benchmark.
- $\beta = 1$ movement of the asset is generally in the same direction as, and about the same amount as the movement of the benchmark
- $\beta > 1$ movement of the asset is generally in the same direction as, but more than the movement of the benchmark

We consider that a stock with $\beta = 1$ is a representative stock, or a stock that is a strong contributor to the index itself. For $\beta > 1$ we get a volatile stock, or stocks which are very strongly influenced by day-to-day market news. Higher-beta stocks tend to be more volatile and therefore riskier, but provide the potential for higher returns. Lower-beta stocks pose less risk but generally offer lower returns. For instance, a stock with a beta of 2 has returns that change, on average, by twice the magnitude of the overall market's returns: when the market's return falls or rises by 3%, the stock's return will fall or rise (respectively) by 6% on average.

The Beta measures the part of the asset's statistical variance that cannot be removed by the diversification provided by the portfolio of many risky assets, because of the correlation of its returns with the returns of the other assets that are in the portfolio. Beta can be estimated for individual companies by using regression analysis against a stock market index. The formula for the beta of an asset within a portfolio is

$$\beta_a = \frac{Cov(R_a, R_b)}{Var(R_b)}$$

where R_a measures the rate of return of the asset, R_b measures the rate of return of the portfolio benchmark, and $Cov(R_a, R_b)$ is the covariance between the rates of return. The portfolio of interest in the Capital Asset Pricing Model (CAPM) formulation is the market portfolio that contains all risky assets, and so the R_b terms in the formula are replaced by R_m, the rate of return of the market. Beta is also referred to as financial elasticity or correlated relative volatility, and can be referred to as a measure of the sensitivity of the asset's returns to market returns, its non-diversifiable risk, its systematic risk, or market risk. On an individual asset level, measuring beta can give clues to volatility and liquidity in the market place. As beta also depends on the correlation of returns, there can be considerable variance about that average: the higher the correlation, the less variance; the lower the correlation, the higher the variance.

In order to estimate beta, one needs a list of returns for the asset and returns for the index which can be daily, weekly or any period. Then one uses standard formulas from linear regression. The slope of the fitted line from the linear least-squares calculation is the estimated beta. The y-intercept is the estimated alpha. Beta is a statistical variable and should be considered with its statistical significance (R square value of the regression line). Higher R square value implies higher correlation and a stronger relationship between returns of the asset and benchmark index.

Using beta as a measure of relative risk has its own limitations. Beta views risk solely from the perspective of market prices, failing to take into consideration specific business fundamentals or economic developments. The price level is also ignored. Beta also assumes that the upside potential and downside risk of any investment are essentially equal, being simply a function of that investment's volatility compared with that of the market as a whole. This too is inconsistent with the world as we know it. The reality is that past security price volatility does not reliably predict future investment performance (or even future volatility) and therefore is a poor measure of risk.

2.4.1.2 The financial alpha

Alpha is a risk-adjusted measure of the so-called active return on an investment. It is the part of the asset's excess return not explained by the market excess return. Put another way, it is the return in excess of the compensation for the risk borne, and thus commonly used to assess active managers' performances. Often, the return of a benchmark is subtracted in order to consider relative performance, which yields Jensen's [1968] alpha. It is the intercept of the security characteristic line (SCL), that is, the coefficient of the constant in a market model regression in Equation (2.4.15). Therefore the alpha coefficient indicates how an investment has performed after accounting for the risk it involved:

- $\alpha < 0$ the investment has earned too little for its risk (or, was too risky for the return)
- $\alpha = 0$ the investment has earned a return adequate for the risk taken
- $\alpha > 0$ the investment has a return in excess of the reward for the assumed risk

For instance, although a return of 20% may appear good, the investment can still have a negative alpha if it is involved in an excessively risky position.

A simple observation: during the middle of the twentieth century, around 75% of stock investment managers did not make as much money picking investments as someone who simply invested in every stock in proportion to the weight it occupied in the overall market in terms of market capitalisation, or indexing. A belief in efficient markets spawned the creation of market capitalisation weighted index funds that seek to replicate the performance of investing in an entire market in the weights that each of the equity securities comprises in the overall market. This phenomenon created a new standard of performance that must be matched: an investment manager should not only avoid losing money for the client and should make a certain amount of money, but in fact he should make more money than the passive strategy of investing in everything equally.

Although the strategy of investing in every stock appeared to perform better than 75% of investment managers, the price of the stock market as a whole fluctuates up and down. The passive strategy appeared to generate the market-beating return over periods of 10 years or more. This strategy may be risky for those who feel they might need to withdraw their money before a 10-year holding period. Investors can use both Alpha and Beta to judge a manager's performance. If the manager has had a high alpha, but also a high beta, investors might not find that acceptable, because of the chance they might have to withdraw their money when the investment is doing poorly.

2.4.2 Performance measures

When considering the performance evaluation of mutual funds, one need to assess whether these funds are earning higher returns than the benchmark returns (portfolio or index returns) in terms of risk. Three measures developed in the framework of the Capital Asset Pricing Model (CAPM) proposed by Treynor [1965], Sharpe [1964] and Lintner [1965] directly relate to the beta of the portfolio through the security market line (SML). Jensen's [1968] alpha is defined as the portfolio excess return earned in addition to the required average return, while the Treynor ratio and the Information ratio are defined as the alpha divided by the portfolio beta and by the standard deviation of the portfolio residual returns. More recent performance measures developed along hedge funds, such as the Sortino ratio, the M2 and the Omega, focus on a measure of total risk, in the continuation of the Sharpe ratio applied to the capital market line (CML). In the context of the extension of the CAPM to linear multi-factor asset pricing models, the development of measures has not been so prolific (see Hubner [2007]).

The Sharpe ratio or Reward to Variability and Sterling ratio have been widely used to measure commodity trading advisor (CTA) performance. One can group investment statistics as Sharpe type combining risk and return in a ratio, or descriptive statistics (neither good nor bad) providing information about the pattern of returns. Examples of the latter are regression statistics (systematic risk), covariance and R^2. Additional risk measures exist to accommodate the risk concerns of different types of investors. Some of these measures have been categorised in Table (2.1).

Table 2.1: List of measure

Type	Combined Return and Risk Ratio
Normal	Sharpe, Information, Modified Information
Regression	Apraisal, Treynor
Partial Moments	Sortino, Omega, Upside Potential, Omega-Sharpe, Prospect
Drawdown	Calmar, Sterling, Burke, Sterling-Calmar, Pain, Martin
Value at Risk	Reward to VaR, Conditional Sharpe, Modified Sharpe

2.4.2.1 The Sharpe ratio

The Sharpe ratio measures the excess return per unit of deviation in an investment asset or a trading strategy defined as

$$M_{SR} = \frac{E[R_a - R_b]}{\sigma} \tag{2.4.16}$$

where R_a is the asset return and R_b is the return of a benchmark asset such as the risk free rate or an index. Hence, $E[R_a - R_b]$ is the expected value of the excess of the asset return over the benchmark return, and σ is the standard deviation of this expected excess return. It characterise how well the return of an asset compensates the investor for the risk taken. If we graph the risk measure with a the measure of return in the vertical axis and the measure of risk in the horizontal axis, then the Sharpe ratio simply measures the gradient of the line from the risk-free rate to the combined return and risk of each asset (or portfolio). Thus, the steeper the gradient, the higher the Sharpe ratio, and the better the combined performance of risk and return.

***Remark* 2.4.1** *The ex-post Sharpe ratio uses the above equation with the realised returns of the asset and benchmark rather than expected returns.*

$$M_{SR} = \frac{r_P - r_F}{\sigma_P}$$

where r_p is the asset/portfolio return (annualised), r_F is the annualised risk-free rate, and σ_P is the portfolio risk or standard deviation of return.

This measure can be compared with the Information ratio in finance defined in general as mean over standard deviation of a series of measurements. The Sharp ratio is directly computable from any observed series of returns without the need for additional information surrounding the source of profitability. While the Treynor ratio only works with systemic risk of a portfolio, the Sharp ratio observes both systemic and idiosyncratic risks. The SR has some shortcomings because all volatility is not equal, and the volatility taken in the measure ignores the distinction between systematic and diversifiable risks. Further, volatility does not distinguish between losses occurring in good or bad time or even between upside and downside surprises.

***Remark* 2.4.2** *The returns measured can be any frequency (daily, weekly, monthly or annually) as long as they are normally distributed, as the returns can always be annualised. However, not all asset returns are normally distributed.*

The SR assumes that assets are normally distributed or equivalently that the investors' preferences can be represented by the quadratic (exponential) utility function. That is, the portfolio is completely characterised by its mean and volatility. As soon as the portfolio is invested in technology stocks, distressed companies, hedge funds or high yield bonds, this ratio is no-longer valid. In that case, the risk comes not only from volatility but also from higher moments like skewness and kurtosis. Abnormalities like kurtosis, fatter tails and higher peaks or skewness on the distribution can be problematic for the computation of the ratio as standard deviation does not have the same effectiveness when these problems exist. As a result, we can get very misleading measure of risk-return. In addition, the Sharp ratio being a dimensionless ratio it may be difficult to interpret the measure of different investments. This weakness was

well addressed by the development of the Modigliani risk-adjusted performance measure, which is in units of percent returns. One need to consider a proper risk-adjusted return measure to get a better feel of risk-adjusted out-performance such as M^2 defined as

$$M^2 = (r_p - r_F)\frac{\sigma_M}{\sigma_P} + r_F$$

where σ_M is the market risk or standard deviation of a benchmark return (see Modigliani et al. [1997]). It can also be rewritten as

$$M^2 = r_P + M_{SR}(\sigma_M - \sigma_P)$$

where the variability can be replaced by any measure of risk and M^2 can be calculated for different types of risk measures. This statistic introduces a return penalty for asset or portfolio risk greater than benchmark risk and a reward if it is lower.

2.4.2.2 More measures of risk

Treynor proposed a risk adjusted performance measure associated with abnormal returns in the CAPM. The Treynor ratio or reward to volatility is a Sharpe type ratio where the numerator (or vertical axis graphically speaking) is identical but the denominator (horizontal axis) replace total risk with systematic risk as calculated by beta

$$M_{TR} = \frac{r_P - r_F}{\beta_P}$$

where β_P is the market beta (see Treynor [1965]). Although well known, the Treynor ratio is less useful precisely because it ignores specific risk. It will converge to the Sharpe ratio in a fully diversified portfolio with no specific risk. The Appraisal ratio suggested by Treynor & Black [1973] is a Sharpe ratio type with excess return adjusted for systematic risk in the numerator, and specific risk and not total risk in the denominator

$$M_{AR} = \frac{\alpha}{\sigma_\epsilon}$$

where α is the Jensen's alpha. It measures the systematic risk adjusted reward for each unit of specific risk taken. While the Sharpe ratio compares absolute return and absolute risk, the Information ratio compares the excess return and tracking error (the standard deviation of excess return). That is, the Information ratio is a Sharpe ratio type with excess return on the vertical axis and tracking error or relative risk on the horizontal axis. As we are not using the risk-free rate, the information ratio lines radiate from the origin and can be negative indicating underperformance.

2.4.2.3 Alpha as a measure of risk

The Jensen's alpha, which is the excess return adjusted for systematic risk, was argued by Jensen to be a more appropriate measure than TR or IR for ranking the potential performance of different portfolios, implying that asset managers should view the best investment as the one with the largest alpha, irrespective of its risk. In view of keeping track of the portfolio's returns Ross extended the CAPM to multiple risk factors, leading to the multi-factor return models commonly used to identify alpha opportunities. Some multi-factor models include size and value factors besides the market beta factor, and others include various industry and style factors for equities. In contrast to the CAPM that has only one risk factor, namely the overall market, APT has multiple risk factors. Each risk factor has a corresponding beta indicating the responsiveness of the asset being priced to that risk factor. Whatever risk factors are used, significant average loadings on any risk factor are viewed as evidence of a systematic risk tilt. Therefore, while the Jensen's alpha is the intercept when regressing asset returns on equity market returns, it is also the intercept of any risk factor model and can be used as a metric. Howerver, it is extremely difficult to obtain a consistent ranking of portfolios using Jensen's alpha because the estimated alpha is too dependent on the multi-factor model used. For example, as the number of factors increase, there is ever less scope for alpha.

2.4.2.4 Empirical measures of risk

In the CAPM, the empirical derivation of the security market line (SML) corresponds to the market model

$$R_i^e = \alpha_i + \beta_i R^m + \epsilon_i$$

where $R_i^e = R_i - R_f$ denotes the excess return on the ith security. The risk and return are forecast ex-ante using a model for the risk and the expected return. They may also be estimated ex-post using some historical data on returns, assuming that investors believe that historical information are relevant to infer the future behaviour of financial assets. The ex-post (or realised) version of the SML is

$$\overline{R}_i^e = \hat{\alpha}_i + \hat{\beta}_i \overline{R}^m$$

where $\overline{R}_i^e$ is the average excess return for the ith security, $\hat{\beta}_i = \hat{C}ov(R_i^e, R^m)$ and $\hat{\alpha}_i = \overline{R}_i^e - \hat{\beta}_i \overline{R}^m$ are the estimators of β_i and α_i, respectively. The Jensen's alpha is measured by the $\hat{\alpha}_i$ in the ex post SML, while the Treynor ratio is defined as the ratio of Jensen's alpha over the stock beta, that is, $M_{TR}(i) = \frac{\hat{\alpha}_i}{\hat{\beta}_i}$. Finally, the Information Ratio is a measure of Jensen's alpha per unit of portfolio specific risk, measured as the standard deviation of the market model residuals $M_{IR} = \frac{\hat{\alpha}_i}{\hat{\sigma}(\epsilon_i)}$.

The sample estimates being based on monthly, weekly or daily data, the moments in the risk measures must be annualised. However,the formula to convert returns or volatility measures from one time period to another assume a particular underlying model or process. When portfolio returns are autocorrelated, the standard deviation does not obey the square-root-of-time rule and one must use higher moments leading to Equation (3.3.11). When returns are perfectly correlated with 100% autocorrelation then a positive return is followed by a positive return and we get a trending market. On the other hand, when the autocorrelation is -100% then a positive return is followed by a negative return and we get a mean reverting or contrarian market. Therefore, assuming a 100% daily correlated market with 1% daily return (5% weekly return), then the daily volatility is 16% ($1\% \times \sqrt{252}$) but the weekly volatility is 35% ($5\% \times \sqrt{252}$) which is more than twice as large (see Bennett et al. [2012]).

When the use of a single index or benchmark in a market model is not sufficient to keep track of the systematic sources of portfolio returns in excess of the risk free rate, one can consider the families of linear multi-index unconditional asset pricing models among which is the ex-post multidimensional equation

$$\overline{R}_i^e = \hat{\alpha}_i + \sum_{j=1}^{k} \hat{\beta}_{ij} \overline{R}_j = \hat{\alpha}_i + \hat{B}_i \overline{R}^e$$

where $j = 1, ..., k$ is the number of distinct risk factors, the line vector $\hat{B}_i = (\hat{\beta}_{i1}, .., \hat{\beta}_{ik})$ and the column vector $\overline{R}^e = (\overline{R}_1^e, .., \overline{R}_k^e)^\top$ represent risk loadings and average returns for the factors, respectively. In this setting, the alpha remains a scalar, and the standard deviation of the regression residuals is also a positive number, so that the multi-index counterparts of the Jensen's alpha and the information ratio are similar to the performance measures applied to the single index model. To conserve the same interpretation as the original Treynor ratio, Hubner [2005] proposed the following generalisation

$$M_{GTR}(i) = \hat{\alpha}_i \frac{\hat{B}_l \overline{R}^e}{\hat{B}_i \overline{R}^e}$$

where l denotes the benchmark portfolio against which the ith portfolio is compared.

2.4.2.5 Incorporating tail risk

While the problem of fat tails is everywhere in financial risk analysis, there is no solution, and one should only consider partial solutions. Hence, one practical strategy for dealing with the messy but essential issues related to measuring and managing risk starts with the iron rule of never relying on one risk metric. Even though all of the standard metrics have well-known flaws, that does not make them worthless, but it is a reminder that we must understand where any one risk measure stumbles, where it can provide insight, and what are the possible fixes, if any.

The main flaw of the Sharpe ratio (SR) is that it uses standard deviation as a proxy for risk. This is a problem because standard deviation, which is the second moment of the distribution, works best with normal distributions. Therefore, investors must have a minimal type of risk aversion to variance alone, as if their utility function was exponential. Even though normality has some validity over long periods of time, in the short run it is very unlikely (see short maturity smile on options in Section (1.7.6.1)). Note, when moving away from the normality assumption for the stock returns, only the denominator in the Sharpe ratio is modified. Hence, all the partial solutions are attempts at expressing one way or another the proper noise of the stock returns. While extensions of the SR to normality assumption have been successful, extension to different types of utility function have been more problematic.

The problem is that extreme losses occur more frequently than one would expect when assuming that price changes are always and forever random (normally distributed). Put another way, statistics calculated using normal assumption might underestimate risk. One must therefore account for higher moments to get a better understanding of the shape of the distribution of returns in view of assessing the relative qualities of portfolios. Investors should prefer high average returns, lower variance or standard deviation, positive skewness, and lower kurtosis. The adjusted Sharpe ratio suggested by Pezier et al. [2006] explicitly rewards positive skewness and low kurtosis (below 3, the kurtosis of a normal distribution) in its calculation

$$M_{ASR} = M_{SR}\big[1 + \frac{S}{6}M_{SR} - \frac{K-3}{24}M_{SR}^2\big]$$

where S is the skew and K is the kurtosis. This adjustment will tend to lower the SR if there is negative skewness and positive excess kurtosis in the returns. Hence, it potentially removes one of the possible criticisms of the Sharpe ratio. Hodges [1997] introduced another extension of the SR accounting for non-normality and incorporating utility function. Assuming that investors are able to find the expected maximum utility $E[u^*]$ associated with any portfolio, the generalised Sharpe ratio (GSR) of the portfolio is

$$M_{GSR} = \big(-2\ln\left(-E[u^*]\right)\big)^{\frac{1}{2}}$$

One can avoid the difficulty of computing the maximum expected utility by assuming the investor has an exponential utility function. Using the fourth order Taylor approximation of the certain equivalent, and approximating the multiplicative factor Pezier et al. [2006] obtained the maximum expected utility function in that setting and showed that the GSR simplifies to the ASR. Thus, when the utility function is exponential and the returns are normally distributed, the GSR is identical to the SR. Otherwise a negative skewness and high positive kurtosis will reduce the GSR relative to the SR.

2.4.3 Some downside risk measures

Downside risk measures the variability of underperformance below a minimum target rate which could be the risk free rate, the benchmark or any other fixed threshold required by the client. All positive returns are included as zero in the calculation of downside risk or semi-standard deviation. Investors being less concerned with variability on the upside, and extremely concerned about the variability on the downside, an extended family of risk-adjusted measures flourished, reflecting the downside risk tolerances of investors seeking absolute and not relative returns. Lower partial moments (LPMs) measure risk by negative deviations of the returns realised in relation to a minimal acceptable return r_T. The LPM of kth-order is computed as

$$LPM(k) = E[\max(r_T - r, 0)^k] = \sum_{i=1}^{n} \frac{1}{n} \max[r_T - r_i, 0]^k$$

Kaplan et al. [2004] introduced a Sharpe type denominator with lower partial moments in the denominator given by $\sqrt[k]{LPM(k)}$. The Kappa index of order k is

$$M_K = \frac{r_P - r_T}{\sqrt[k]{LPM(k)}}$$

Kappa indices can be tailord to the degree of risk aversion of the investor, but can not be used to rank portfolio's performance according to investor's preference. One possible calculation of semi-standard deviation or downside risk in the period $[0, T]$ is

$$\sigma_D = \sqrt{\sum_{i=1}^{n} \frac{1}{n} \min[r_i - r_T, 0]^2}$$

where r_T is the minimum target return. Downside potential is simply the average of returns below target

$$\sum_{i=1}^{n} \frac{1}{n} \min[r_i - r_T, 0] = \sum_{i=1}^{n} \frac{1}{n} I_{\{r_i < r_T\}}(r_i - r_T)$$

Alternatively, one can measure excess return by using a higher potential moment (HPM), which measures positive deviations from the minimal acceptable return r_T. The LPM of kth-order is computed as

$$HPM(k) = \sum_{i=1}^{n} \frac{1}{n} \max[r_i - r_T, 0]^k$$

The upside statistics are

$$\sigma_U = \sqrt{\sum_{i=1}^{n} \frac{1}{n} \max[r_i - r_T, 0]^2}$$

with upside potential being the average of returns above target

$$\sum_{i=1}^{n} \frac{1}{n} \max[r_i - r_T, 0] = \sum_{i=1}^{n} \frac{1}{n} I_{\{r_i > r_T\}}(r_i - r_T)$$

Shadwick et al. [2002] proposed a gain-loss ratio, called Omega, that captutres the information in the higher moments of return distribution

$$M_\Omega = \frac{E[\max(r - r_T, 0)]}{E[\max(r_T - r, 0)]} = \frac{\frac{1}{n}\sum_{i=1}^{n} \max(r_i - r_T, 0)}{\frac{1}{n}\sum_{i=1}^{n} \max(r_T - r_i, 0)}$$

This ratio implicitly adjusts for both skewness and kurtosis. It can also be used as a ranking statistics (the higher, the better). Note, the ratio is equal to 1 when r_T is the mean return. Kaplan et al. [2004] showed that the Omega ratio can be rewritten as a Sharpe type ratio called Omega-Sharpe ratio

$$M_{OSR} = \frac{r_P - r_T}{\frac{1}{n}\sum_{i=1}^{n} \max(r_T - r_i, 0)}$$

which is simply $\Omega - 1$, thus generating identical ranking than the Omega ratio. Setting $r_T = 0$ in the Omega ratio, Bernardo et al. [1996] obtained the Bernardo Ledoit ratio (or Gain-Loss ratio)

$$M_{BLR} = \frac{\frac{1}{n}\sum_{i=1}^{n} \max(r_i, 0)}{\frac{1}{n}\sum_{i=1}^{n} \max(-r_i, 0)}$$

Sortino et al. [1991] proposed an extension of the Omega-Sharpe ratio by using downside risk in the denominator

$$M_{SoR} = \frac{r_p - r_T}{\sigma_D}$$

In that measure, portfolio managers will only be penalised for variability below the minimum target return, but will not be penalised for upside variability. In order to rank portfolio performance while combining upside potential with downside risk, Sortino et al. [1999] proposed the Upside Potential ratio

$$M_{UPR} = \frac{\frac{1}{n}\sum_{i=1}^{n} \max(r_i - r_T, 0)}{\sigma_D}$$

This measure is similar to the Omega ratio except that performance below target is penalised further by using downside risk rather than downside potential. Going further, we can replace the upside potential in the numerator with the upside risk, getting the Variability Skewness

$$M_{VSR} = \frac{\sigma_U}{\sigma_D}$$

2.4.4 Considering the value at risk

2.4.4.1 Introducing the value at risk

The Value at Risk (VaR) is a widely used risk measure of the risk of loss on a specific portfolio of financial assets. VaR is defined as a threshold value such that the probability that the mark-to-market loss on the portfolio over the given time horizon exceeds this value (assuming normal markets and no trading in the portfolio) is the given probability level. For example, if a portfolio of stocks has a one-day 5% VaR of $\$1$ million, there is a 0.05 probability that the portfolio will fall in value by more than $\$1$ million over a one day period if there is no trading. Informally, a loss of $\$1$ million or more on this portfolio is expected on 1 day out of 20 days (because of 5% probability). VaR represents a percentile of the predictive probability distribution for the size of a future financial loss. That is, if you have a record of portfolio value over time then the VaR is simply the negative quantile function of those values.

Given a confidence level $\alpha \in (0, 1)$, the VaR of the portfolio at the confidence level α is given by the smallest number z_p such that the probability that the loss L exceeds z_p is at most $(1 - \alpha)$. Assuming normally distributed returns, the Value-at-Risk (daily or monthly) is

$$VaR(p) = W_0(\mu - z_p\sigma)$$

where W_0 is the initial portfolio wealth, μ is the expected asset return (daily or monthly), σ is the standard deviation (daily or monthly), and z_p is the number of standard deviation at $(1 - \alpha)$ (distance between μ and the VaR in number of standard deviation). It ensures that

$$P(dW \leq -VaR(p)) = 1 - \alpha$$

Note, $VaR(p)$ represents the lower bound of the confidence interval given in Appendix (B.9.6). For example, setting $\alpha = 5\%$ then $z_p = 1.96$ with $p = 97.5$ which is a 95% probability.

If returns do not display a normal distribution pattern, the Cornish-Fisher expansion can be used to include skewness and kurtosis in computing value at risk (see Favre et al. [2002]). It adjusts the z-value of a standard VaR for skewness and kurtosis as follows

$$z_{cf} = z_p + \frac{1}{6}(z_p^2 - 1)S + \frac{1}{24}(z_p^3 - 3z_p)K - \frac{1}{36}(2z_p^3 - 5z_p)S^2$$

where z_p is the critical value according to the chosen α-confidence level in a standard normal distribution, S is the skewness, K is the excess kurtosis. Integrating them into the VaR measure by means of the Cornish-Fisher expansion z_{cf} , we end up with a modified formulation for the VaR, called MVaR

$$MVaR(p) = W_0(\mu - z_{cf}\sigma)$$

2.4.4.2 The reward to VaR

We saw earlier that when the risk is only measured with the volatility it is often underestimated, because the assets returns are negatively skewed and have fat tails. One solution is to use the value-at-risk as a measure of risk, and consider Sharpe type measures using VaR. For instance, replacing the standard deviation in the denominator with the VaR ratio (Var expressed as a percentage of portfolio value rather than an amount) Dowd [2000] got the Reward to VaR

$$M_{RVaR} = \frac{r_P - r_F}{\text{VaR ratio}}$$

Note, the VaR measure does not provide any information about the shape of the tail or the expected size of loss beyond the confidence level, making it an unsatisfactory risk measure.

2.4.4.3 The conditional Sharpe ratio

Tail risk is the possibility that investment losses will exceed expectations implied by a normal distribution. One attempt at trying to anticipate non-normality is the modified Sharpe ratio, which incorporates skewness and kurtosis into the calculation. Another possibility is the so-called conditional Sharpe ratio (CSR) or expected shortfall, which attempts to quantify the risk that an asset or portfolio will experience extreme losses. VaR tries to tell us what the possibility of loss is up to some confidence level, usually 95%. So, for instance, one might say that a certain portfolio is at risk of losing $X\%$ for 95% of the time. What about the remaining 5%? Conditional VaR, or CVaR, dares to tread into this black hole of fat taildom (by accounting for the shape of the tail). For the conditional Sharpe ratio, CVaR replaces standard deviation in the metric's denominator

$$M_{CVaR} = \frac{r_P - r_F}{CVaR(p)}$$

The basic message in conditional Sharpe ratio, like that of its modified counterpart, is that investors underestimate risk by roughly a third (or more?) when looking only at standard deviation and related metrics (see Agarwal et al. [2004]).

2.4.4.4 The modified Sharpe ratio

The modified Sharpe ratio (MSR) is one of several attempts at improving the limitations of standard deviation. MSR is far from a complete solution, still, it factors in two aspects of non-normal distributions, skewness and kurtosis. It does so through the use of what is known as a modified Value at Risk measure (MVaR) as the denominator. The MVaR follows the Cornish-Fisher expansion, which can adjust the VaR in terms of asymmetric distribution (skewness) and above-average frequency of earnings at both ends of the distribution (kurtosis). The modified Sharpe ratio is

$$M_{MSR} = \frac{r_P - r_F}{MVaR(p)}$$

Similarly to the Adjusted Sharpe ratio, the Modified Sharpe ratio uses modified VaR adjusted for skewness and kurtosis (see Gregoriou et al. [2003]). Given a 10 years example, in all cases the modified Sharpe ratio was lower than its traditional Sharpe ratio counterpart. Hence, for the past decade, risk-adjusted returns were lower than expected after

adjusting for skewness and kurtosis. Note, depending on the rolling period, MSR has higher sensitivity to changes in non-normal distributions whereas the standard SR is immune to those influences.

2.4.4.5 The constant adjusted Sharpe ratio

Eling et al. [2006] showed that even though hedge fund returns are not normally distributed, the first two moments describe the return distribution sufficiently well. Furthermore, on a theoretical basis, the Sharpe ratio is consistent with expected utility maximisation under the assumption of elliptically distributed returns. Taking all the previous remarks into consideration, we propose a new very simple Sharpe ratio called the Constant Adjusted Sharpe ratio and defined as

$$M_{CASR} = \frac{r_P - r_F}{\sigma(1+\epsilon_S)}$$

where $\epsilon_S > 0$ ($\epsilon_S = \frac{1}{3}$ to recover the conditional Sharpe ratio) is the adjusted volatility defined in Section (**??**). In that measure, the volatility is simply modified by a constant.

2.4.5 Considering drawdown measures

For an investor wishing to avoid losses, any continuous losing return period or drawdown constitutes a simple measure of risk. The drawdown measures the decline from a historical peak in some variable (see Magdon-Ismail et al. [2004]). It is the pain period experienced by an investor between peak (new highs) and subsequent valley (a low point before moving higher). If $(X_t)_{t\geq 0}$ is a random process with $X_0 = 0$, the drawdown $D(T)$ at time T is defined as

$$D(T) = \max\left(0, \max_{t\in(0,T)}(X_t - X_T)\right)$$

One can count the total number of drawdowns n_d in the entire period $[0, T]$ and compute the average drawdown as

$$\overline{D}(T) = \frac{1}{n_d}\sum_{i=1}^{n_d} D_i$$

where D_i is the ith drawdown over the entire period. The maximum drawdown (MDD) up to time T is the maximum of the drawdown over the history of the variable (typically the Net Asset Value of an investment)

$$MDD(\tau) = \max_{\tau\in(0,T)} D(\tau)$$

In a long-short portfolio, the maximum drawdown is the maximum loss an investor can suffer in the fund buying at the highest point and selling at the lowest. We can also define the drawdown duration as the length of any peak to peak period, or the time between new equity highs. Hence, the maximum drawdown duration is the worst (maximum/longest) amount of time an investment has seen between peaks. Martin [1989] developed the Ulcer index where the impact of the duration of drawdowns is incorporated by selecting the negative return for each period below the previous peak or high water mark

$$\text{Ulcer Index } = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i')^2}$$

where D_i' is the drawdown since the previous peak in ith period. This way, deep, long drawdowns will have a significant impact as the underperformance since the last peak is squared. Being sensitive to the frequency of time period, this index penalises managers taking time to recovery from previous high. If the drawdowns are not squared, we get the Pain index

$$\text{Pain Index } = \frac{1}{n}\sum_{i=1}^{n}|D_i^{'}|$$

which is similar to the Zephyr Pain index in discrete form proposed by Becker.

We are considering measures which are modification of the Sharpe ratio in the sense that the numerator is always the excess of mean returns over risk-free rate, but the standard deviation of returns in the denominator is replaced by some function of the drawdown.

The Calmar ratio (or drawdown ratio) is a performance measurement used to evaluate hedge funds which was created by T.W. Young [1991]. Originally, it is a modification of the Sterling ratio where the average annual rate of return for the last 36 months is divided by the maximum drawdown for the same period. It is computed on a monthly basis as opposed to other measures computed on a yearly basis. Note, the MAR ratio, discussed in Managed Account Reports, is equal to the compound annual return from inception divided by the maximum drawdown over the same period of time. As discussed by Bacon [2008], later version of the Calmar ratio introduce the risk-free rate into the numerator to create a Sharpe type ratio

$$M_{CR} = \frac{r_P - r_F}{MDD(\tau)}$$

The Sterling ratio replaces the maximum drawdowns in the Calmar ratio with the average drawdown. According to Bacon, the original definition of the Sterling ratio is

$$M_{SterR} = \frac{r_P}{\overline{D}_{lar} + 10\%}$$

where $\overline{D}_{lar}$ is the average largest drawdown, and the 10% is an arbitrary compensation for the fact that $\overline{D}_{lar}$ is inevitably smaller than the maximum drawdown. In vew of generalising the measure, Bacon rewrote it in as a Sharpe type ratio given by

$$M_{SterR} = \frac{r_P - r_F}{\overline{D}(T)}$$

where the number of observations n_d is fixed by the investor's preference. Other variation of the Sterling ratio uses the average annual maximum drawdown $\overline{MDD}(\tau)$ in the denominator over three years. Combining the Sterling and Calmar ratio, Bacon proposed the Sterling-Calmar ratio as

$$M_{SCR} = \frac{r_P - r_F}{\overline{MDD}(\tau)}$$

In order to penalise major drawdowns as opposed to many mild ones, Burke [1994] used the concept of the square root of the sum of the squares of each drawdown, getting

$$M_{BR} = \frac{r_P - r_F}{\sqrt{\sum_{i=1}^{n_d} D_i^2}}$$

where the number of drawdowns n_d used can be restricted to a set number of the largest drawdowns. In the case where the investor is more concerned by the duration of the drawdowns, the Martin ratio or Ulcer performance index is similar to the Burke ratio but with the Ulcer index in the denominator

$$M_{MR} = \frac{r_P - r_F}{\sqrt{\sum_{i=1}^{d} \frac{1}{n}(D_i')^2}}$$

and the equivalent to the Martin ratio but using the Pain index is the Pain ratio

$$M_{PR} = \frac{r_P - r_F}{\sum_{i=1}^{d} \frac{1}{n} D_i'}$$

In view of assessing the best measure to use, Eling et al. [2006] concluded that most of these measures are all highly correlated and do not lead to significant different rankings. For Bacon, the investor must decide ex-ante which measures of return and risk best describe his preference, and choose accordingly.

2.4.6 Some limitation

2.4.6.1 Dividing by zero

Statistical inference with measures based on ratios, such as the Treynor performance measure, is delicate when the denominator tends to zero as the ratio goes to infinity. Hence, this measure provides unstable performance measures for non-directional portfolios such as market neutral hedge funds. When the denominator is not bounded away from zero, the expectation of the ratio is infinite. Further, when the denominator is negative, the ratio would assign positive performance to portfolios with negative abnormal returns. As suggested by Hubner [2007], one way arround when assessing the quality of performance measures is to consider only directional managed portfolios. However, hedge funds favour market neutral portfolios. We present two artifacts capable of handeling the beta in the denominator of a ratio. We let, $\beta_i^a(j\delta)$ taking values in $\mathbb{R}$, be the statistical Beta for the stock $S_i(j\delta)$ at time $t = j\delta$. We want to define a mapping $\beta_i(j\delta)$ such that the ratio $\frac{1}{\beta_i(j\delta)}$ allocates maximum weight to stocks with $\beta \approx 0$, and decreasing weight as the β moves away from zero. One possibility is to set

$$\beta_i(j\delta) = a + b\beta_i^a(j\delta) \, , i = 1, .., N$$

with $a = \frac{1}{3}$ and $b = \frac{2}{3}$, but it does not stop the ratio from being negative. An alternative approach is to consider the inverse bell shape for the distribution of the Beta

$$\beta_i(j\delta) = a\big(1 - e^{-b(\beta_i^a(j\delta))^2}\big) + c$$

such that for $\beta_i^a(j\delta) = 0$ we get $\beta_i(j\delta) = c$. In that setting $\beta_i(j\delta) \in [c, a + c]$ and a good calibration gives $a = 1.7$, $b = 0.58$, and $c = 0.25$. Modifying the bell shape, we can directly define the ratio as

$$\frac{1}{\beta_i(j\delta)} = ae^{-b(\beta_i^a(j\delta))^2}$$

with $a = 3$ and $b = 0.25$. In that setting $\frac{1}{\beta_i(j\delta)} \in [0, a]$ with the property that when $\beta = 0$ we get the maximum value a.

2.4.6.2 Anomaly in the Sharpe ratio

The (ex post) Sharpe ratio of a sequence of returns $x_1, ..., x_N \in [-1, \infty)$ is $M(N) = \frac{\mu_N}{\sigma_N}$ where μ_N is the sample mean and σ_N^2 is the sample variance. Note, the returns are bounded from below by -1. Intuitively, the Sharpe ratio is the return per unit of risk. Another way of measuring the performance of a portfolio with the above sequence of returns is to see how this sequence of returns would have affected an initial investment of $C_A = 1$ assuming no capital inflows and outflows after the initial investment. The final capital resulting from this sequence of returns is

$$P_N = C_A \prod_{i=1}^{N} (1 + x_i)$$

We are interested in conditions under which the following anomaly is possible: the Sharpe ratio $M(N)$ is large while $P_N < 1$. We could also consider the condition that in the absence of capital inflows and outflows the returns $x_1, ..., x_N$ underperform the benchmark portfolio. Vovk [2011] showed that if the return is 5% over $k - 1$ periods, and then it is

-100% in the kth period then as $k \to \infty$ we get $\mu_k \to 0.05$ and $\sigma_k \to 0$. Therefore, making k large enough, we can make the Sharpe ratio $M(k)$ as large as we want, despite losing all the money over the k periods. In this example the returns are far from being Gaussian (strictly speaking, returns cannot be Gaussian unless they are constant, since they are bounded from below by -1). Note, this example leads to the same conclusions when the Sharpe ratio is replaced by the Sortino ratio. However, this example is somewhat unrealistic in that there is a period in which the portfolio loses almost all its money. Fortunately,Vovk [2011] showed that it is the only way a high Sharpe ratio can become compatible with losing money. That is, in the case of the Sharpe ratio, such an abnormal behaviour can happen only when some one-period returns are very close to -1. In the case of the Sortino ratio, such an abnormal behaviour can happen only when some one-period returns are very close to -1 or when some one-period returns are huge.

2.4.6.3 The weak stochastic dominance

The stochastic dominance axiom of utility implies that if exactly the same returns can be obtained with two different investments A and B, but the probability of a return exceeding any threshold τ is always greater with investment A, then A should be preferred to B. That is, investment A strictly dominates investment B if and only if

$$P_A(R > \tau) > P_B(R > \tau) \forall \tau$$

and A weakly dominates B if and only if

$$P_A(R > \tau) \geq P_B(R > \tau) \forall \tau$$

Hence, no rational investor should choose an investment which is weakly dominated by another one. With the help of an example, Alexandrer showed that the SR can fail to rank investment according to the weak stochastic dominance. We consider two portfolios A and B with the distribution of their returns in excess of the risk-free rate given in Table (2.2).

Table 2.2: Distribution of returns

Probability	Excess return A	Excess return B
0.1	20%	40%
0.8	10%	10%
0.1	-20%	-20%

The highest excess return from portfolio A is only 20%, whereas the highest excess return from portfolio B is 40%. We show the result of the SR of the two investments in Table (2.3). The mean is given by $E[R] = \sum_i P_i R_i$ and the variance satisfies $Var(R) = \sum_i P_i R_i^2 - (E[R])^2$.

Table 2.3: Sharpe ratios

Portfolio	A	B
Expected excess return	8.0%	10.0%
Standard deviation	9.79%	13.416%
Sharpe ratio	0.8165	0.7453

Following the SR, investor would choose portfolio A, whereas the weak stochastic dominance indicates that any rational investor should prefer B to A. As a result, one can conclude that the SRs are not good metrics to use in the decision process on uncertain investments.

Chapter 3

Introduction to financial time series analysis

For details see text books by Makridakis et al. [1989], Brockwell et al. [1991] and Tsay [2002].

3.1 Prologue

A time series is a set of measurements recorded on a single unit over multiple time periods. More generally, a time seies is a set of statistics, usually collected at regular intervals, and occurring naturally in many application areas such as economics, finance, environmental, medecine, etc. In order to analyse and model price series to develop efficient quantitative trading, we define returns as the differences of the logarithms of the closing price series, and we fit models to these returns. Further, to construct efficient security portfolios matching the risk profile and needs of individual investors we need to estimate the various properties of the securities consitituting such a portfolio. Hence, modelling and forecasting price return and volatility is the main task of financial research. Focusing on closing prices recorded at the end of each trading day, we argue that it is the trading day rather than the chronological day which is relevant so that constructing the series from available data, we obtain a process equally spaced in the relevant time unit. We saw in Section (2.1.5) that a first step towards forecasting financial time series was to consider some type of technical indicators or mathematical statistics with price forecasting capability, hoping that history trends would repeat itself. However, following this approach we can not assess the uncertainty inherent in the forecast, and as such, we can not measure the error of forecast. An alternative is to consider financial time series analysis which is concerned with theory and practice of asset valuation over time. While the methods of time series analysis pre-date those for general stochastic processes and Markov Chains, their aims are to describe and summarise time series data, fit low-dimensional models, and make forecasts. Even though it is a highly empirical discipline, theory forms the foundation for making inference. However, both financial theory and its empirical time series contain some elements of uncertainty. For instance, there are various definitions of asset volatility, and in addition, volatility is not directly observable. Consequently, statistical theory and methods play an important role in financial time series analysis. One must therefore use his knowledge of financial time series in order to use the appropriate statistical tools to analyse the series. In the rest of this section we are going to describe financial time series analysis, and we will introduce statistical theory and methods in the following sections.

3.2 An overview of data analysis

3.2.1 Presenting the data

3.2.1.1 Data description

The data may consists in equity stocks, equity indices, futures, FX rates, commodities, and interest rates (Eurodollar and 10-year US Treasury Note) spanning a period from years to decade with frequency of intraday quotes, close-to-close, weeks, or months. As the contracts are traded in various exchanges, each with different trading hours and holidays, the data series should be appropriately aligned to avoid potential lead-lag effects by filling forward any missing asset prices (see Pesaran et al. [2009]). Daily, weekly or montly return series are constructed for each contract by computing the percentage change in the closing end of day, week or month asset price level. The mechanics of opening and maintaining a position on a futures contract involves features like initial margins, potential margin call, interest accured on the margin account, and no initial cash payment at the initiation of the contract (see Miffre et al. [2007]). As a result, the construction of a return data series for a futures contract does not have an objective nature and various methodologies have been used in the literature. Pesaran et al. [2009], Fuertes et al. [2010] compute returns similarly as the percentage change in the price level, wheras Pirrong [2005] and Gorton et al. [2006] also take into account interest rate accruals on a fully collateralised basis, and Miffre et al. [2007] use the change in the logarithms of the price level. Lastly, Moskowitz et al. [2012] use the percentage change in the price level in excess of the risk-free rate. Knowing the percentage returns of the time series, we can compute the annualised mean return, volatility, and Sharpe ratios.

3.2.1.2 Analysing the data

We apply standard econometric theory described in Section (5.1) to test for the presence of heteroskedasticity and autocorrelation, and we adjust the models accordingly when needed. In general, there exists a great amount of cross-sectional variation in mean returns and volatilities with the commodities being historically the most volatile contracts (see Pesaran et al. [2009]). Further, the distribution of buy-and-hold or buy-and-sell return series exhibits fat tails as deduced by the kurtosis and the maximum likelihood estimated degrees of freedom for a Student t-distribution. A normal distribution is almost universally rejected by the Jarque and Bera [1987] and the Lilliefors [1967] tests of normality (see Section (3.3.4.2)). It is more difficult to conclude about potential first-order time-series autocorrelation using tools such as the Ljung and Box [1978] test. However, very strong evidence of heteroscedasticity is apparent across all frequencies deduced by the ARCH test of Engle [1982]. Baltas et al. [2012b] found that this latter effect of time variation in the second moment of the return series was also apparent in the volatility. We also perform a regression analysis with ARMA (autoregressive moving average) modelling of the serial correlation in the disturbance. In addition we can perform several robustness checks. First, we check the robustness of the model through time by using a Chow [1960] test to test for stability of regression coefficients between two periods. When we find significant evidence of parameter instability, we use a Kalman filter analysis described in Section (3.2.5.2), which is a general form of a linear model with dynamic parameters, where priors on model parameters are recursively updated in reaction to new information (see Hamilton [1994]).

3.2.1.3 Removing outliers

We follow an approach described by Zhu [2005] consisting in finding the general trend curve for the time series, and then calculating the spread which is the distance between each point and the trend curve. The idea is to replace each data point by some kind of local average of surrounding data points such that averaging reduce the level of noise without biasing too much the value obtained. To find the trend, we consider the Savitzky-Golay low-pass smoothing filter described in Section (4.3.3). After some experiments, Zhu [2005] found that the filter should be by degree 1 and span size 3. Given the corresponding smoothed data representing the trend of the market data we get the spread for each market data point from the trend. The search for outliers uses the histogram of $(f_i - \overline{f}_i)$ with $M = 10$ bins of equal width. We label a threshold T and define all f_i with $|f_i - \overline{f}_i| > T$ to be outliers. The next question is how to

select the value M, and the threshold T. Suppose we are given a set of market data which contain previously known errors. Adjust M and T until we find proper pairs of M and T which can successfully find all the errors. We then can tune the parameters with more historical data from the same market. We can have an over-determined solution for the value of M and T by enough training data provided. Outliers are replaced by interpolation. On the market, one common way to deal with error data is to replace it with the previous data, that is, zeroth order interpolation. This method neglects the trend, while we usually expect movements on a liquid market. Instead of utilising much training data, an alternative to search for T is to iteratively smooth the data points. Step one, we choose a start T, say T_0, and smooth the data according to M_0 and T_0. Second step we stop the iteration if the histogram has a short tail, since we believe all the outliers are removed. Else we replace the outliers by interpolations, and repeat step one.

3.2.2 Basic tools for summarising and forecasting data

We assume that we have available a database from which to filter data and build numerical forecasts, that is, a table with multiple dimensions. Cross sectional data refer to measurements on multiple units, recorded in a single time period. Although forecasting practice involves multiple series, the methods we are going to examine use data from the past and present to predict future outcomes. Hence, we will first focus on the use of time series data.

3.2.2.1 Presenting forecasting methods

Forecasting is about making statements on events whose actual outcomes have not yet been observed. We distinguish two main types of forecasting methods:

1. Qualitative forecasting techniques are subjective, based on opinions and judgements, and are appropriate when past data are not available. For example, one tries to verify whether there is some causal relationship between some variables and the demand. If this is the case, and if the variable is known in advance, it can be used to make a forecast.

2. On the contrary, quantitative forecasting models are used to forecast future data as a function of past data, and as such, are appropriate when past data are available. The main idea being that the evolution in the past will continue into the future. If we observe some correlations between some variables, then we can use these correlations to make some forecast. A dynamic model incorporating all the important internal and external variables is implemented and used to test different alternatives. For instance, to estimate the future demand accurately, we need to take into account facts influencing the demand.

Subjective forecasts are often time-consuming to generate and may be subject to a variety of conscious or unconscious biases. In general, simple analysis of available data can perform as well as judgement procedures, and are much quicker and less expensive to produce. The effective possible choices are judgement only, quantitative method only and quantitative method with results adjusted by user judgement. All three options have their place in the forecasting lexicon, depending upon costs, available data and the importance of the task in hand. Careful subjective adjustment of quantitative forecasts may often be the best combination, but we first need to develop an effective arsenal of quantitative methods. To do so, we need to distinguish between methods and models.

- A forecasting method is a (numerical) procedure for generating a forecast. When such methods are not based upon an underlying statistical model, they are termed heuristic.

- A statistical (forecasting) model is a statistical description of the data generating process from which a forecasting method may be derived. Forecasts are made by using a forecast function that is derived from the model.

For example, we can specify a forecasting method as

$$F_t = b_0 + b_1 t$$

where F_t is the forecast for time period t, b_0 is the intercept representing the value at time zero, and b_1 is the slope representing the increase in forecast values from one period to the next. All we need to do to obtain a forecast is to calibrate the model. However, we lack a basis for choosing values for the parameters, and we can not assess the uncertainty inherent in the forecasts. Alternatively, we may formulate a forecasting model as

$$Y_t = \beta_0 + \beta_1 t + \epsilon$$

where Y denotes the time series being studied, β_0 and β_1 are the level and slope parameters, and ϵ denotes a random error term corresponding to that part of the series that cannot be fitted by the trend line. Once we make appropriate assumptions about the nature of the error term, we can estimate the unknown parameters, β_0 and β_1. These estimates are typically written as b_0 and b_1. Thus the forecasting model gives rise to the forecast function

$$F_t = b_0 + b_1 t$$

where the underlying model enables us to make statements about the uncertainty in the forecast, something that the heuristic method do not provide. As a result, risk and uncertainty are central to forecasting, as one must indicate the degree of uncertainty attaching to forecasts. Hence, some idea about its probability distribution is necessary. For example, assuming a forecast for some demand has the distribution of Gauss (normal) with average μ and standard deviation σ, the coefficient of variation of the prediction is $\frac{\sigma}{\mu}$.

3.2.2.2 Summarising the data

Following Brockwell et al. [1991], we write the real-valued series of observations as $..., Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, ...$ a doubly infinite sequence of real-valued random variables indexed by $\mathbb{Z}$. Given a set of n values $Y_1, Y_2, .., Y_n$, we place these values in ascending order written as $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$. The median is the middle observation. When n is odd it can be written $n = 2m + 1$ and the median is $Y_{(m+1)}$, and when n is even we get $n = 2m$ and the median is $\frac{1}{2}(Y_{(m)} + Y_{(m)})$. It is possible to have two very different datasets with the same means and medians. For that reason, measures of the middle are useful but limited. Another important attribute of a dataset is its dispersion or variability about its middle. The most useful measures of dispersion (or variability) are the range, the percentiles, the mean absolute deviation, and the standard deviation. The range denotes the difference between the largest and smallest values in the sample

$$Range = Y_{(n)} - Y_{(1)}$$

Therefore, the more spread out the data values are, the larger the range will be. However, if a few observations are relatively far from the middle but the rest are relatively close to the middle, the range can give a distorted measure of dispersion. Percentiles are positional measures for a dataset that enable one to determine the relative standing of a single measurement within the dataset. In particular, the pth percentile is defined to be a number such that $p\%$ of the observations are less than or equal to that number and $(100 - p)\%$ are greater than that number. So, for example, an observation that is at the 75th percentile is less than only 25% of the data. In practice, we often can not satisfy the definition exactly. However, the steps outlined below at least satisfies the spirit of the definition:

1. Order the data values from smallest to largest, including ties.
2. Determine the position
$$k.ddd = 1 + \frac{p(n-1)}{100}$$
3. The pth percentile is located between the kth and the $(k + 1)$th ordered value. Use the fractional part of the position, .ddd as an interpolation factor between these values. If $k = 0$, then take the smallest observation as the percentile and if $k = n$, then take the largest observation as the percentile.

The 50th percentile is the median and partitions the data into a lower half (below median) and upper half (above median). The 25th, 50th, 75th percentiles are referred to as quartiles. They partition the data into 4 groups with 25% of the values below the 25th percentile (lower quartile), 25% between the lower quartile and the median, $25th$ between the median and the 75th percentile (upper quartile), and 25% bove the upper quartile. The difference between the upper and lower quartiles is referred to as the inter-quartile range. This is the range of the middle 50% of the data. Given $d_i = Y_i - \overline{Y}$ where $\overline{Y}$ is the arithmetic mean, the Mean Absolute Deviation (MAD) is the average of the deviations about the mean, ignoring the sign

$$MAD = \frac{1}{n}\sum_i |d_i|$$

The sample variance is an average of the squared deviations about the mean

$$S^2 = \frac{1}{n-1}\sum_i d_i^2$$

The population variance is given by

$$\sigma_p^2 = \frac{1}{n}\sum_i (Y_i - \mu_p)^2$$

where μ_p is the population mean. Note that the unit of measure for the variance is the square of the unit of measure for the data. For that reason (and others), the square root of the variance, called the standard deviation, is more commonly used as a measure of dispersion. Note that datasets in which the values tend to be far away from the middle have a large variance (and hence large standard deviation), and datasets in which the values cluster closely around the middle have small variance. Unfortunately, it is also the case that a dataset with one value very far from the middle and the rest very close to the middle also will have a large variance. Comparing the variance with the MAD, S gives greater weight to the more extreme observations by squaring them and it may be shown that $S > MAD$ whenever MAD is greater than zero. A rough relationship between the two is

$$S = 1.25MAD$$

The standard deviation of a dataset can be interpreted by Chebychev's Theorem (see Tchebichef [1867]):

Theorem 3.2.1 *For any $k > 1$, the proportion of observations within the interval $\mu_p \pm k\sigma_p$ is at least $(1 - \frac{1}{k^2})$.*

Hence, knowing just the mean and standard deviation of a dataset allows us to obtain a rough picture of the distribution of the data values. Note that the smaller the standard deviation, the smaller is the interval that is guaranteed to contain at least 75% of the observations. Conversely, the larger the standard deviation, the more likely it is that an observation will not be close to the mean. Note that Chebychev's Theorem applies to all data and therefore must be conservative. In many situations the actual percentages contained within these intervals are much higher than the minimums specified by this theorem. If the shape of the data histogram is known, then better results can be given. In particular, if it is known that the data histogram is approximately bell-shaped, then we can say

- $\mu_p \pm \sigma_p$ contains approximately 68%,
- $\mu_p \pm 2\sigma_p$ contains approximately 95%,
- $\mu_p \pm 3\sigma_p$ contains essentially all

of the data values. This set of results is called the empirical rule. Several extensions of Chebyshev's inequality have been developed, among which is the asymmetric two-sided version given by

$$P(k_1 < Y < k_2) \geq \frac{4\big((\mu_p - k_1)(k_2 - \mu_p) - \sigma_p^2\big)}{(k_2 - k_1)^2}$$

In mathematical statistics, a random variable Y is standardized by subtracting its expected value $E[Y]$ and dividing the difference by its standard deviation $\sigma(Y)$

$$Z = \frac{Y - E[Y]}{\sigma(Y)}$$

The Z-score is a dimensionless quantity obtained by subtracting the population mean μ_p from an individual raw score Y_i and then dividing the difference by the population standard deviation σ_p. That is,

$$Z = \frac{Y_i - \mu_p}{\sigma_p}$$

From Chebychev's theorem, at least 75% of observations in any dataset will have Z-scores in the range $[-2, 2]$. The standard score is the (signed) number of standard deviations an observation or datum is above the mean, and it provides an assessment of how off-target a process is operating. The use of the term Z is due to the fact that the Normal distribution is also known as the Z distribution. They are most frequently used to compare a sample to a standard normal deviate, though they can be defined without assumptions of normality. Note, considering the Z-score, Cantelli obtained sharpened bounds given by

$$P(Z \geq k) \leq \frac{1}{1 + k^2}$$

The Z-score is only defined if one knows the population parameters, but knowing the true standard deviation of a population is often unrealistic except in cases such as standardized testing, where the entire population is measured. If one only has a sample set, then the analogous computation with sample mean and sample standard deviation yields the Student's t-statistic. Given a sample mean $\overline{Y}$ and sample standard deviation S, we define the standardised scores for the observations, also known as Z-scores as

$$Z = \frac{Y_i - \overline{Y}}{S}$$

The Z-score is to examine forecast errors and one proceed in three steps:

- Check that the observed distribution of the errors is approximately normal
- If the assumption is satisfied, relate the Z-score to the normal tables
  - The probability that $|Z| > 1$ is about 0.32
  - The probability that $|Z| > 2$ is about 0.046
  - The probability that $|Z| > 3$ is about 0.0027
- Create a time series plot of the residuals (and/or Z-scores) when appropriate to determine which observations appear to be extreme

Hence, whenever you see a Z-score greater than 3 in absolute value, the observation is very atypical, and we refer to such observations as outliers.

The change in the absolute level of the series from one period to the next is called the first difference of the serries, given by

$$DY_t = Y_t - Y_{t-1}$$

where Y_{t-1} is known at time t. Letting $\hat{D}_t$ be the forecast for the difference, the forecast for Y_t becomes

$$F_t = \hat{Y}_t = Y_{t-1} + \hat{D}_t$$

The growth rate for Y_t is

$$G_t = GY_t = 100\frac{D_t}{Y_{t-1}}$$

so that the forecast for Y_t can be written as

$$F_t = \hat{Y}_t = Y_{t-1}\big(1 + \frac{\hat{G}_t}{100}\big)$$

If we think of changes in the time series in absolute terms we should use DY and if we think of it in relative terms we should use GY. Note, reducing a chocolate ration by 50% and then increasing it by 50% does not gives you as much chocolate as before since

$$(1 - \frac{50}{100})(1 + \frac{50}{100}) = 0.75$$

To avoid this asymmetry we can use the logarithm transform $L_t = \ln Y_t$ with first difference in logarithm being

$$DL_t = \ln Y_t - \ln Y_{t-1}$$

converting the exponential (or proportional) growth into linear growth. If we generate a forecast of the log-difference, the forecast for the original series, given the previous value Y_{t-1} becomes

$$\hat{Y}_t = Y_{t-1}e^{D\hat{L}_t}$$

3.2.2.3 Measuring the forecasting accuracy

When selecting a forecasting procedure, a key question is how to measure performance. A natural approach would be to look at the differences between the observed values and the forecasts, and to use their average as a performance measure. Suppose that we start from forecast origin t so that the forecasts are made successively (one-step-ahead) at times $t+1, t+2, ..., t+h$, there being h such forecasts in all. The one-step-ahead forecast error at time $t+i$ may be denoted by

$$e_{t+i} = Y_{t+i} - F_{t+i}$$

The Mean Error (ME) is given by

$$ME = \frac{1}{h}\sum_{i=1}^{h}(Y_{t+i} - F_{t+i}) = \frac{1}{h}\sum_{i=1}^{h} e_{t+i}$$

ME will be large and positive (negative) when the actual value is consistently greater (less) than the forecast. However, this measure do not reflect variability, as positive and negative errors could virtually cancel each other out, yet substantial forecasting errors could remain. Hence, we need measures that take account of the magnitude of an error regardless of the sign. The simplest way to gauge the variability in forecasting performance is to examine the absolute errors, defined as the value of the error ignoring its sign and expressed as

$$|e_i| = |Y_i - F_i|$$

We now present various averages, based upon the errors or the absolute errors.

- the Mean Absolute Error

$$MAE = \frac{1}{h}\sum_{i=1}^{h}|Y_{t+i} - F_{t+i}| = \frac{1}{h}\sum_{i=1}^{h}|e_{t+i}|$$

- the Mean Absolute Percentage Error

$$MAPE = \frac{100}{h}\sum_{i=1}^{h}\frac{|Y_{t+i} - F_{t+i}|}{Y_{t+i}} = \frac{100}{h}\sum_{i=1}^{h}\frac{|e_{t+i}|}{Y_{t+i}}$$

- the Mean Square Error

$$MSE = \frac{1}{h}\sum_{i=1}^{h}(Y_{t+i} - F_{t+i})^2 = \frac{1}{h}\sum_{i=1}^{h}e_{t+i}^2$$

- the Normalised Mean Square Error

$$NMSE = \frac{1}{\sigma^2 h}\sum_{i=1}^{h}(Y_{t+i} - F_{t+i})^2 = \frac{1}{\sigma^2 h}\sum_{i=1}^{h}e_{t+i}^2$$

  where σ^2 is the variance of the true sequence over the prediction period (validation set).

- the Root Mean Square Error

$$RMSE = \sqrt{MSE}$$

- the Mean Absolute Scaled Error

$$MASE = \frac{\sum_{i=1}^{h}|Y_{t+i} - F_{t+i}|}{\sum_{i=1}^{h}|Y_{t+i} - Y_{t+i-1}|}$$

- the Directional Symmetry

$$DS = \frac{1}{h-1}\sum_{i=1}^{h}\mathcal{H}(Y_{t+i}.F_{t+i})$$

  where $\mathcal{H}(x) = 1$ if $x > 0$ and $\mathcal{H}(x) = 0$ otherwise (Heaviside function).

- the Direction Variation Symmetry

$$DVS = \frac{1}{h-1}\sum_{i=2}^{h}\mathcal{H}((Y_{t+i} - Y_{t+i-1}).(F_{t+i} - F_{t+i-1}))$$

Note, the Mean Absolute Scaled Error was introduced by Hyndman et al. [2006]. It is the ratio of the MAE for the current set of forecasts relative to the MAE for forecasts made using the random walk. Hence, for $MASE > 1$ the random walk forecasts are superior, otherwise the method under consideration is superior to the random walk. We now give some general comments on these measures.

- MAPE should only be used when $Y > 0$, MASE is not so restricted.
- MAPE is the most commonly used error measure in practice, but it is sensitive to values of Y close to zero.
- MSE is measured in terms of (dollars), and taking the square root to obtain the RMSE restores the original units.
- A value of the $NMSE = 1$ corresponds to predicting the unconditional mean.
- The RMSE gives greater weight to large (absolute) errors. It is therefore sensitive to extreme errors.
- The measure using absolute values always equals or exceeds the absolute value of the measure based on the errors, so that $MAE \geq |ME|$ and $MAPE \geq |MPE|$. If the values are close in magnitude that suggests a systematic bias in the forecasts.
- Both MAPE and MASE are scale-free and so can be used to make comparisons across multiple series. The other measures need additional scaling.
- DS is the percentage of correctly predicted directions with respect to the target variable. It provides a measure of the number of times the sign of the target was correctly forecast.
- DVS is the percentage of correctly predicted direction variations with respect to the target variable.

3.2.2.4 Prediction intervals

So far we have considered point forecasts which are future observations for which we report a single forecast value. However, confidence in a single number is often misplaced. We assume that the predictive distribution for the series Y follows the normal law (although such an assumption is at best an approximation and needs to be checked). If we assume that the standard deviation (SD) of the distribution is known, we may use the upper 95% point of the standard normal distribution (this value is 1.645) so that the one-sided prediction interval is

$$\hat{Y} + 1.645 \times SD$$

where $\hat{Y}$ is the point forecast. The normal distribution being the most widely used in the construction of prediction intervals, it is critical to check that the forecast errors are approximately normally distributed. Typically the SD is unknown and must be estimated from the sample that was used to generate the point forecast, meaning that we use the RMSE to estimate the SD. We can also use the two-sided $100(1-\alpha)$ prediction intervals given by

$$\hat{Y} \pm z_{1-\frac{\alpha}{2}} \times RMSE$$

where $z_{1-\frac{\alpha}{2}}$ denotes the upper $100(1-\frac{\alpha}{2})$ percentage point of the normal distribution. In the case of a 95% one-step-ahead prediction intervals we set $\alpha = 5\%$ and get $z_{1-\frac{0.05}{2}} = 1.96$. The general purpose of such intervals is to provide an indication of the reliability of the point forecasts.

An alternative approach to using theoretical formulae when calculating prediction intervals is to use the observed errors to show the range of variation expected in the forecasts. For instance, we can calculate the 1-step ahead errors made using the random walk forecasts which form a histogram. We can also fit a theoretical probability density to the observed errors. Other distributions (than normal) are possible since in many applications more extreme errors are observed than those suggested by a normal distribution. Fitting a distribution gives us more precise estimates of the prediction intervals which are called empirical prediction intervals. To be useful, these empirical prediction intervals need to be based on a large sample of errors.

3.2.2.5 Estimating model parameters

Usually we partition the original series into two parts and refer to the first part (containing $75-80\%$ of the observations) as the estimation sample, which is used to estimate the starting values and the smoothing parameters. The parameters are commonly estimated by minimizing the mean squared error (MSE), although the mean absolute error (MAE) or mean absolute percentage error (MAPE) are also used. The second part called hold-out sample represents the remaining $20-25\%$ of the observations and is used to check forecasting performance. Some programs allow repeated estimation and forecast error evaluation by advancing the estimation sample one observation at a time and repeating the error calculations.

When forecating data, we should not rely on an arbitrary pre-set smoothing parameter. Most computer programs nowdays provide efficient estimates of the smoothing constant, based upon minimizing some measure of risk such as the mean squared error (MSE) for the one-step-ahead forecasts

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - F_i)^2$$

where $Y_i = Y_{t_i}$ and $F_i = F_{t_i|t_{i-1}}$. More formally, we let $\hat{Y}_{T+\tau}(T)$ (or $F_{T+\tau|T}$) denote the forecast of a given time series $\{Y_t\}_{t\in\mathbb{Z}^+}$ at time $T+\tau$, where T is a specified origin and $\tau \in \mathbb{Z}^+$. In that setting, the MSE becomes

$$MSE(T) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_{(t-1)+1}(t-1))^2$$

Similarly, we define the MAD as

$$MAD(T) = \frac{1}{T}\sum_{t=1}^{T}|Y_t - \hat{Y}_{(t-1)+1}(t-1)|$$

which is the forecast $\hat{Y}_{T+\tau}(T)$ with $T = t-1$ and $\tau = 1$. An approximate 95% prediction interval for $\hat{Y}_{T+\tau}(T)$ is given by

$$\hat{Y}_{T+\tau}(T) \pm z_{0.25}1.25MAD(T)$$

where $z_{0.25} \approx 1.96$ (see Appendix (5.7.5)).

3.2.3 Modelling time series

Given the time series $(X_t)_{t\in\mathbb{Z}}$ at time t, we can either decompose it into elements and estimate each components separately, or we can directly model the series with a model such as an autoregressive integrated moving average (ARIMA).

3.2.3.1 The structural time series

When analysing a time series, the classical approach is to decompose it into components: the trend, the seasonal component, and the irregular term. More generally, the structural time series model proposed by Majani [1987] decomposes the time series $(X_t)_{t\in\mathbb{Z}}$ at time t into four elements

1. the trend (T_t) : long term movements in the mean
2. the seasonal effects (I_t) : cyclical fluctuations related to the calendar
3. the cycles C_t : other cyclycal fluctuations (such as business cycles)

4. the residuals E_t : other random or systematic fluctuations

The idea being to create separate models for these four elements and to combine them either additively

$$X_t = T_t + I_t + C_t + E_t$$

or multiplicatively

$$X_t = T_t \cdot I_t \cdot C_t \cdot E_t$$

which can be obtained by applying the logarithm. Forecasting is done by extrapolating T_t, I_t and C_t, and expecting $E[E_t] = c \in \mathbb{R}$. One can therefore spend his time either modelling each seperate element and try to recombine them, or, directly modelling the process X_t. Howver, the decomposition is not unique, and the components are interrelated, making identification difficult. Several methods have been proposed to extract the components in a time series, ranging from simple weighted averages to more sophisticated methods, such as Kalman filter or exponential smoothing, Fourier transform, spectral analysis, and more recently wavelet analysis (see Kendall [1976b], Brockwell et al. [1991], Arino et al. [1995]). In economic time series, the seasonal component has usually a constant period of 12 months, and to assess it one uses some underlying assumptions or theory about the nature of the series. Longer-term trends, defined as fluctuations of a series on time scales of more than one year, are more difficult to estimate. These business cycles are found by elimination of the seasonal component and the irregular term. Further, forecasting is another reason for decomposing a series as it is generally easier to forecast components of a time series than the whole series itself. One approach for decomposing a continuous, or discrete, time series into components is through spectral analysis. Fourier analysis uses sum of sine and cosine at different wavelengths to express almost any given periodic function, and therefore any function with a compact support. However, the non-local characteristic of sine and cosine implies that we can only consider stationary signals along the time axis. Even though various methods for time-localising a Fourier transform have been proposed to avoid this problem such as windowed Fourier transform, the real improvement comes with the development of wavelet theory. In the rest of this guide, we are going to discribe various technique to model the residual components (E_t), the trend T_t, and the business cycles C_t, and we will also consider different models to forecast the process X_t directly.

3.2.3.2 Some simple statistical models

Rather than modelling the elements of the time series $(X_t)_{t\in\mathbb{Z}}$ we can directly model the series. For illustration purpose we present a few basic statistical models describing the data which will be used and detailed in Chapter (5). Note, each of these models has a number of variants, which are refinements of the basic models.

AR process An autoregressive (AR) process is one in which the change in the variable at a point in time is linearly correlated with the previous change. In general, the correlation declines exponentially with time and desapear in a relatively short period of time. Letting Y_n be the change in Y at time n, with $0 \leq Y \leq 1$, then we get

$$Y_n = c_1 Y_{n-1} + ... + c_p Y_{n-p} + e_n$$

where $|c_l| \leq 1$ for $l = 1, .., p$, and e is a white noise series with mean 0 and variance σ_e^2. The restrictions on the coefficients c_l ensure that the process is stationary, that is, there is no long-term trend, up or down, in the mean or variance. This is an $AR(p)$ process where the change in Y at time n is dependent on the previous p periods. To test for the possibility of an AR process, a regression is run where the change at time n is the dependent variable, and the changes in the previous q periods (the lags) are used as independent variables. Evaluating the t-statistic for each lag, if any of them are significant at the 5% level, we can form the hypothesis that an AR process is at work.

MA process In a moving average (MA) process, the time series is the result of the moving average of an unobserved time series

$$Y_n = d_1 e_{n-1} + ... + d_p e_{n-q} + e_n$$

where $|d_l| < 1$ for $l = 1, .., q$. The restriction on the coefficients d_l ensure that the process is invertible. In the case where $d_l > 1$, future events would affect the present, and the process is statioary. Because of the moving average process, there is a linear dependence on the past and a short-term memory effect.

ARMA process In an autoregressive moving average ($ARMA$) model, we have both some autoregressive terms and some moving average terms which are unobserved random series. We get the general $ARMA(p, q)$ form

$$Y_n = c_0 + c_1 Y_{n-1} + ... + c_p Y_{n-p} - d_1 e_{n-1} - ... - d_q Y_{n-q} + e_n$$

where p is the number of autoregressive terms, q is the number of moving average terms, and e_n is a random variable with a given distribution F and $c_0 \in \mathbb{R}$ is the drift.

ARIMA process Both AR and $ARMA$ models can be absorbed into a more general class of processes called autoregressive integrated moving average (ARIMA) models which are specifically applied to nonstationary time series. While they have an underlying trend in their mean and variance, by taking successive differences of the data, these processes become stationary. For instance, a price series is not stationary merly because it has a long-term growth component. That is, the price will not tend towards an average value as it can grow without bound. Fortunately, in the efficient market hypothesis (EMH), it is assumed that the changes in price (or returns) are stationary. Typically, price changes are specified as percentage changes, or, log differences, which is the first difference. However, in some series, higher order differences may be needed to make the data stationary. Hence, the difference of the differences is a second-order $ARIMA$ process. In general, we say that Y_t is a homogeneous nonstationary process of order d if

$$Z_t = \Delta^d Y_t$$

is stationary, where Δ represents differencing, and d represents the level of differencing. If Z_t is an $ARMA(p, q)$ process, then Y_t is considered an $ARIMA(p, d, q)$ process. The process does not have to be mixed as if Y_t is an $ARIMA(p, d, 0)$ process, then Z_t is an $AR(p)$ process.

ARCH process We now introduce popular models to describe the conditional variance of market returns. The basic autoregressive conditional heteroskedasticity $(ARCH)$ model developed by Engle [1982] became famous because

- they are a family of nonlinear stochastic processes (as opposed to $ARMA$ models)
- their frequency distribution is a high-peaked, fat-tailed one
- empirical studies showed that financial time series exhibit statistically significant $ARCH$.

In the ARCH model, time series are defined by normal probability distributions but time-dependent variances. That is, the expected variance of a process is conditional on its previous value. The process is also autoregressive in that it has a time dependence. A sample frequency distribution is the average of these expanding and contracting normal distributions, leading to fat-tailed, high-peaked distribution at any point in time. The basic model follows

$$\begin{aligned} Y_n &= S_n e_n \\ S_n^2 &= \alpha_0 + \alpha_1 e_{n-1}^2 \end{aligned}$$

where e is a standard normal random variable, and α_1 is a constant. Typical values are $\alpha_0 = 1$ and $\alpha_1 = \frac{1}{2}$. Once again, the observed value Y is the result of an unobserved series, e, depending on past realisations of itself. The

nonlinearity of the model implies that small changes will likely be followed by other small changes, and large changes by other large changes, but the sign will be unpredictable. Further, large changes will amplify, and small changes will contrac, resulting in fat-tailed high-peaked distribution.

GARCH process Bollerslev [1986] formalised the generalised $ARCH$ (or $GARCH$) by making the S variable dependent on the past as well,

$$\begin{aligned} Y_n &= S_n e_n \\ S_n^2 &= \alpha_0 + \alpha_1 e_{n-1}^2 + \beta_1 S_{n-1}^2 \end{aligned}$$

where the three values range from 0 to 1, but $\alpha_0 = 1$, $\alpha_1 = 0.1$, and $\beta_1 = 0.8$ are typical values. $GARCH$ also creates a fat-tailed high-peaked distribution.

Example of a financial model The main idea behind (G)ARCH models is that the conditional standard deviations of a data series are a function of their past values. A very common model in financial econometric is the $AR(1) - GARCH(1,1)$ process given by

$$\begin{aligned} r_n &= c_0 + c_1 r_{n-1} + a_n \\ v_n &= \alpha_0 + \alpha_1 e_{n-1}^2 + \beta_1 v_{n-1} \end{aligned}$$

where r_n is the log-returns of the data series for each n, v_n is the conditional variance of the residuals for the mean equation [1]for each n, and c_0, c_1, α_0 and α_1 are known parameters that need to be estimated. The $GARCH$ process is well defined as long as the condition $\alpha_1 + \beta_1 < 1$ is satisfied. If this is not the case, the variance process is non-stationary and we have to fit other processes for conditional variance such as Integrated GARCH (IGARCH) models.

3.2.4 Introducing parametric regression

Given a set of observations, we want to summarise the data by fitting it to a model that depends on adjustable parameters. To do so we design a merit function measuring the agreement between the data and the model with a particular choice of parameters. We can design the merit function such that either small values represent close agreement (frequentist), or, by considering probabilities, larger values represent closer agreement (bayesians). In either case, the parameters of the model are adjusted to find the corresponding extremum in the merit function, providing best-fit parameters. The adjustment process is an optimisation problem which we will treat in Chapter (14). However, in some special cases, specific modelling exist, providing an alternative solution. In any case, a fitting procedure should provide

1. some parameters
2. error estimates on the parameters, or a way to sample from their probability distribution
3. a statistical measure of goodness of fit

In the event where the third item suggests that the model is an unlikely match to the data, then the first two items are probably worthless.

[1] $a_n = r_n - c_0 - c_1 r_{n-1}$, $a_n = \sqrt{v_n} e_n$.

3.2.4.1 Some rules for conducting inference

The central frequentist idea postulates that given the details of a null hypothesis, there is an implied population (probability distribution) of possible data sets. If the assumed null hypothesis is correct, the actual, measured, data set is drawn from that population. When the measured data occurs very infrequently in the population, then the hypothesis is rejected. Focusing on the distribution of the data sets, they neglect the concept of a probability distribution of the hypothesis. That is, for frequentists, there is no statistical universe of models from which the parameters are drawn. Instead, they identify the probability of the data given the parameters, as the likelihood of the parameters given data. Parameters derived in this way are called maximum likelihood estimators (MLE). An alternative approach is to consider Bayes's theorem relating the conditional probabilities of two events, A and B,

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{3.2.1}$$

where $P(A|B)$ is the probability of A given that B has occured. A and B need not to be repeatable events, and can be propositions or hypotheses, obtaining a set of consistent rules for conducting inference. All Bayesian probabilities are viewed as conditional on some collective background information I. Assuming some hypothesis H, even before any explicit data exist, we can assign some degree of plausibility $P(H|I)$ called the Bayesian prior. When some data D_1 comes along, using Equation (3.2.1), we reassess the plausibility of H as

$$P(H|D_1I) = P(H|I)\frac{P(D_1|HI)}{P(D_1|I)}$$

where the numerator is calculable as the probability of a data set given the hypothesis, and the denominator is the prior predictive probability of the data. The latter is a normalisation constant ensuring that the probability of all hypotheses sums to unity. When some additional data D_2 come along, we can further refine the estimate of the probability of H

$$P(H|D_2D_1I) = P(H|D_1I)\frac{P(D_2|HD_1I)}{P(D_2|D_1I)}$$

and so on. From the product rule for probabilities $P(AB|C) = P(A|C)P(B|AC)$, we get

$$P(H|D_2D_1I) = P(H|I)\frac{P(D_2D_1|HI)}{P(D_2D_1|I)}$$

obtaining the same answer as if all the data D_1D_2 had been taken together.

3.2.4.2 The least squares estimator

Maximum likelihood estimator Given N data points (X_i, Y_i) for $i = 0, .., N-1$, we want to fit a model having M adjustable parameters a_j, $j = 0, .., M-1$, predicting a functional relationship between the measured independent and dependent varibles, defined as

$$Y(X) = Y(X|a_0, .., a_{M-1})$$

Following the frequentist, given a set of parameters, if the probability of obtaining the data set is too amall, then we can conclude that the parameters are unlikely to be right. Assuming that each data point Y_i has a measurement error that is independently random and distributed as a Gaussian distribution around the true model $Y(X)$, and assuming that the standard deviations σ of these normal distributions are the same for all points, then the probability of obtaining the data set is the product of the probabilities of each point

$$P(\text{ data } | \text{ model }) \propto \prod_{i=0}^{N-1} e^{-\frac{1}{2}(\frac{Y_i - Y(X_i)}{\sigma})^2} \Delta Y$$

Alternatively, calling Bayes' theorem in Equation (3.2.1), we get

$$P(\text{ model } | \text{ data }) \propto P(\text{ data } | \text{ model }) P(\text{ model })$$

where $P(\text{ model }) = P(a_0, .., a_{M-1})$ is the prior probability distribution on all models. The most probable model is to maximise the probability of obtaining the data set above, or equivalently, minimise the negative of its logarithm

$$\sum_{i=0}^{N-1} \frac{1}{2} \Big(\frac{Y_i - Y(X_i)}{\sigma} \Big)^2 - N \log \Delta Y$$

which is equivalent to minimising the probability above since N, σ, and ΔY are all contants. In that setting, we recover the least squares fit

$$\text{minimise over } a_0, .., a_{M-1} : \sum_{i=0}^{N-1} (Y_i - Y(X_i | a_0, .., a_{M-1}))^2$$

Under specific assumptions on measurement errors (see above), the least-squares fitting is the most probable parameter set in the Bayesian sense (assuming flat prior), and it is the maximum likelihood estimate of the fitted parameters. Relaxing the assumption of constant standard deviations, by assuming a known standard deviation σ_i for each data point (X_i, Y_i), then the MLE of the model parameters, and the Bayesian most probable parameter set, is given by minimising the quantity

$$\chi^2 = \sum_{i=0}^{N-1} \Big(\frac{Y_i - Y(X_i)}{\sigma_i} \Big)^2 \tag{3.2.2}$$

called the chi-square, which is a sum of N squares of normally distributed quantities, each normalised to unit variance. Note, in practice measurement errors are far from Gaussian, and the central limit theorem does not apply, leading to fat tail events skewing the least-squares fit. In some cases, the effect of nonnormal errors is to create an abundance of outlier points decreasing the probability Q that the chi-square should exceed a particular value χ^2 by chance.

Linear models So far we have made no assumption about the linearity or nonlinearity of the model $Y(X|a_0, .., a_{M-1})$ in its parameters $a_0, .., a_{M-1}$. The simplest model is a straight line

$$Y(X) = Y(X|a, b) = a + bX$$

called linear regression. Assuming that the uncertainty σ_i associated with each measurement Y_i is known, and that the dependent variables X_i are known excatly, we can minimise Equation (3.2.2) to determine a and b. At its minumum, derivatives of $\xi^2(a, b)$ with respect to a and b vanish. See Press et al. [1992] for explicit solution of a and b, covariance of a and b characterising the uncertainty of the parameter estimation, and an estimate of the goodness of fit of the data. We can also consider the general linear combination

$$Y(X) = a_0 + a_1 X + a_2 X^2 + ... + a_{M-1} X^{M-1}$$

which is a polynomial of degree $M - 1$. Further, linear combination of sines and cosines is a Fourier series. More generally, we have models of the form

$$Y(X) = \sum_{k=0}^{M-1} a_k \phi_k(X)$$

where the quantities $\phi_0(X), .., \phi_{M-1}(X)$ are arbitrary fixed functions of X called basis functions which can be non-linear (linear refers only to the model's dependence on its parameters a_k). In that setting, the chi-square merit function becomes

$$\chi^2 = \sum_{i=0}^{N-1} \Big(\frac{Y_i - \sum_{k=0}^{M-1} a_k \phi_k(X_i)}{\sigma_i} \Big)^2$$

where σ_i is the measurement error of the ith data point. We can use optimisation to minimise χ^2, or in special cases we can use specific techniques. We let A be an $N \times M$ matrix constructed from the M basis functions evaluated at the N abscissas X_i, and from the N measurement errors σ_i with element

$$A_{ij} = \frac{\phi_j(X_i)}{\sigma_i}$$

This matrix is called the design matrix, and in genral $N \geq M$. We also define the vector b of length N with element $b_i = \frac{Y_i}{\sigma_i}$, and denote the M vector whose components are the parameters to be fitted $a_0, .., a_{M-1}$ by a. The minimum of the merit function occurs where the derivative of χ^2 with respect to all M parameters a_k vanishes. It yields M equations

$$\sum_{i-0}^{N-1} \frac{1}{\sigma_i} \big(Y_i - \sum_{j=0}^{M-1} a_j \phi_j(X_i) \big) \phi_k(X_i) = 0 \, , k = 0, .., M-1$$

Interchanging the order of summations, we get the normal equations of the least-squares problem

$$\sum_{j=0}^{M-1} \alpha_{kj} a_j = \beta_k$$

where $\alpha = A^\top . A$ is an $M \times M$ matrix, and $\beta = A^\top . b$ is a vector of length M. In matrix form, the normal equations become

$$(A^\top . A) a = A^\top . b$$

which can be solved for the vector a by LU decomposition, Cholesky decomposition, or Gauss-Jordan elimination. The inverse matrix $C = \alpha^{-1}$ is called the covariance matrix, and is closely related to the uncertainties of the estimated parameters a. These uncertainties are estimated as

$$\sigma^2(a_j) = C_{jj}$$

the diagonal elements of C, being the variances of the fitted parameters a. The off-diagonal elements C_{jk} are the covariances between a_j and a_k.

Nonlinear models In the case where the model depends nonlinearly on the set of M unknown parameters a_k for $k + 0, .., M - 1$, we use the same method as above where we define a χ^2 merit function and determine best-fit parameters by its minimisation. This is similar to the general nonlinear function minimisation problem. If we are sufficiently close to the minimum, we expect the χ^2 function to be well approximated by a quadratic form

$$\chi^2(a) \approx \gamma - d.a + \frac{1}{2} a.D.a$$

where d is an M-vector and D is an $M \times M$ matrix. If the approximation is a good one, we can jump from the current trial parameters a_{cur} to the minimising ones a_{min} in a single leap

$$a_{min} = a_{cur} + D^{-1} . [-\nabla \chi^2(a_{cur})]$$

However, in the where the approximation is a poor one, we can take a step down the gradient, as in the steepest descent method, getting

$$a_{next} = a_{cur} - cst \times \nabla\chi^2(a_{cur})$$

for small constant cst. In both cases we need to compute the gradient of the χ^2 function at any set of parameters a. For a_{min}, we also need the matrix D, which is the second derivative matrix (Hesian matrix) of the χ^2 merit function, at any a. In this particular case, we know exactly the form of χ^2, since it is based on a model function that we specified, so that the Hessian matrix is known to us.

3.2.5 Introducing state-space models

3.2.5.1 The state-space form

The state-space form of time series models represent the actual dynamics of a data generation process. We let Y_t denote the observation from a time series at time t, related to a vector α_t, called the state vector, which is possibly unobserved and whose dimension m is independent of the dimension n of Y_t. The general form of a linear state-space model, is given by the following two equations

$$\begin{aligned} Y_t &= Z_t\alpha_t + d_t + G_t\epsilon_t \ , t = 1, .., T \\ \alpha_{t+1} &= T_t\alpha_t + c_t + H_t\epsilon_t \end{aligned} \tag{3.2.3}$$

where Z_t is an $(n \times m)$ matrix, d_t is an $(n \times 1)$ vector, G_t is an $(n \times (n+m))$ matrix, T_t is an $(m \times m)$ matrix, and H_t is an $(m \times (n+m))$ matrix. The process ϵ_t is an $((n+m) \times 1)$ vector of serially independent, identically distributed disturbances with $E[\epsilon_t] = 0$ and $Var(\epsilon_t) = I$ the identity matrix. We let the initial state vector α_1 be independent of ϵ_t at all time t. The first equation is the observation or measurement equation, and the second equation is transition equation. The general (first-order Markov) state equation takes the form

$$\alpha_t = f(\alpha_{t-1}, \theta_{t-1}) + \eta_{t-1}$$

and the general observation equation takes the form

$$Y_t = h(\alpha_t, \theta_t) + \epsilon_t$$

with independent error processes $\{\eta_t\}$ and $\{\epsilon_t\}$. If the system matrices do not evolve with time, the state-space model is called time-invariant or time-homogeneous. If the disturbances ϵ_t and initial state vector α_1 are assumed to have a normal distribution, then the model is termed Gaussian. Further, if $G_t H_t^\top = 0$ for all t then the measurement and transition equations are uncorrelated. The fundamental inference mechanism is Bayesian and consists in computing the posterior quantities of interest sequentially in the following recursive calculation:

1. Letting $\psi_t = \{Y_1, .., Y_t\}$ be the information set up to time t, we get the prior distribution

$$p(\alpha_t|\psi_{t-1}) = \int p(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|\psi_{t-1})d\alpha_{t-1}$$

   corresponding to the distribution of the parameters before any data is observed.

2. Then, the updating equation becomes

$$p(\alpha_t|\psi_t) = \frac{p(Y_t|\alpha_t)p(\alpha_t|\psi_{t-1})}{p(Y_t|\psi_{t-1})}$$

   where the sampling distribution $p(Y_t|\alpha_t)$ is the distribution of the observed data conditional on its parameters.

The updates provides an analytical solution if all densities in the state and observation equation are Gaussian, and both the state and the observation equation are linear. If these conditions are met, the Kalman filter (see Section (3.2.5.2)) provides the optimal Bayesian solution to the tracking problem. Otherwise we require approximations such as the Extended Kalman filter (EKF), or Particle filter (PF) which approximates non-Gaussian densities and non-linear equations. The particle filter uses Monte Carlo methods, in particular Importance sampling, to construct the approximations.

3.2.5.2 The Kalman filter

The Kalman filter is used for prediction, filtering and smoothing. If we let $\psi_t = \{Y_1, .., Y_t\}$ denote the information set up to time t, then the problem of prediction is to compute $E[\alpha_{t+1}|\psi_t]$. Filtering is concerned with calculating $E[\alpha_t|\psi_t]$, while smoothing is concerned with estimating $E[\alpha_t|\psi_T]$ for all $t < T$. In the Linear Gaussian State-Space Model we assume $G_t H_t^\top = 0$ and drop the terms d_t and c_t from the observation and transition equations (3.2.3). Further, we let

$$G_t G_t^\top = \Sigma_t \ , \ H_t H_t^\top = \Omega_t$$

and the Kalman filter recursively computes the quantities

$$\begin{aligned}
&a_{t|t} = E[\alpha_t|\psi_t] \text{ filtering} \\
&a_{t+1|t} = E[\alpha_{t+1}|\psi_t] \text{ prediction} \\
&P_{t|t} = MSE(\alpha_t|\psi_{t-1}) \\
&P_{t+1|t} = MSE(\alpha_{t+1}|\psi_t)
\end{aligned}$$

where MSE is the mean-square error or one-step ahead prediction variance. Then, starting with $a_{1|0}$, $P_{1|0}$, then $a_{t|t}$ and $a_{t+1|t}$ are obtained by running for $t = 1, .., t$ the recursions

$$\begin{aligned}
&V_t = Y_t - Z_t a_{t|t-1} \ , \ F_t = Z_t P_{t|t-1} Z_t^\top + \Sigma_t \\
&a_{t|t} = a_{t|t-1} + P_{t|t-1} Z_t^\top F_t^{-1} V_t \\
&P_{t|t} = P_{t|t-1} - P_{t|t-1} Z_t^\top F_t^{-1} Z_t P_{t|t-1} \\
&a_{t+1|t} = T_t a_{t|t} \\
&P_{t+1|t} = T_t P_{t|t} T_t^\top + \Omega_t
\end{aligned}$$

where V_t denotes the one-step-ahead error in forecasting Y_t conditional on the information set at time $(t-1)$ and F_t is its MSE. The quantities $a_{t|t}$ and $a_{t|t-1}$ are optimal estimators of α_t conditional on the available information. The resulting recursions for $t = 1, .., T-1$ follows

$$\begin{aligned}
&a_{t+1|t} = T_t a_{t|t-1} + K_t V_t \\
&K_t = T_t P_{t|t-1} Z_t^\top F_t^{-1} \\
&P_{t+1|t} = T_t P_{t+1|t} L_t^\top + \Omega_t \\
&L_t = T_t - K_t Z_t
\end{aligned}$$

Parameter estimation Another application of the Kalman filter is the estimation of any unknown parameters θ that appear in the system matrices. The likelihood for data $Y = (Y_1, .., Y_Y)$ can be constructed as

$$p(Y_1,..,Y_Y) = p(Y_T|\psi_{T-1})...p(Y_2|\psi_1)p(Y_1) = \prod_{t=1}^{T} p(Y_t|\psi_{t-1})$$

Assuming that the state-space model is Gaussian, by taking conditional expectations on both sides of the observation equation, with $d_t = 0$ we deduce that for $t = 1,..,T$

$$\begin{aligned} E[Y_t|\psi_{t-1}] &= Z_t a_{t|t-1} \\ Var(Y_t|\psi_{t-1}) &= F_t \end{aligned}$$

the one-step-ahead prediction density $p(Y_t|\psi_{t-1})$ is the density of a multivariate normal random variable with mean $Z_t a_{t|t-1}$ and and covariance matrix F_t. Thus, the log-likelihood function is given by

$$\log L = -\frac{nT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log detF_t - \frac{1}{2}\sum_{t=1}^{T} V_t^\top F_t^{-1} V_t$$

where $V_t = Y_t - Z_t a_{t|t-1}$. Numerical procedures are used in order to maximise the log-likelihood to obtain the ML estimates of the parameters θ which are consistent and asymptotically normal. If the state-space model is not Gaussian, the likelihood can still be constructed in the same way using the minimum mean square linear estimators of the state vector. However, the estimators $\hat{\theta}$ maximising the likelihood are the quasi-maximum likelihood (QML) estimators of the parameters. They are also consistent and asymptotically normal.

Smoothing Smoothing is another application of Kalman filter where, given a fixed set of data, estimates of the state vector are computed at each time t in the sample period taking into account the full information set available. The algorithm computes $a_{t|T} = E[\alpha_t|\psi_T]$ along with its MSE, and $P_{t|T}$ computed via a set of backward recursions for all $t = 1,..,T-1$. To obtain $a_{t|T}$ and $P_{t|T}$ we start with $a_{T|T}$ and $P_{T|T}$ and run backwards for $t = T-1,..,0$

$$\begin{aligned} a_{t|T} &= a_{t|t} + P_t^*(a_{t+1|T} - a_{t+1|t}) \\ P_{t|T} &= P_{t|t} + P_t^*(P_{t+1|T} - P_{t+1|t})P_t^* \text{ , } P_t^* = P_t T_t^\top P_{t+1|t} \end{aligned}$$

the extensive use of the Markov chain Monte Carlo (MCMC), in particular the Gibbs sampler, has given rise to another smoothing algorithm called the simulation smoother and is also closely related to the Kalman filter. In contrast, to the fixed interval smoother, which computes the conditional mean and variance of the state vector at each time t in the sample, a simulation smoother is used for drawing samples from the density $p(\alpha_0,...,\alpha_T|Y_T)$. The first simulation smoother is based on the identity

$$p(\alpha_0,...,\alpha_T|Y_T) = p(\alpha_T|Y_T)\prod_{t=0}^{T-1} p(\alpha_t|\psi_t,\alpha_{t+1})$$

and a draw from $p(\alpha_0,...,\alpha_T|Y_T)$ s recursively constructed in terms of α_t. Starting with a draw $\hat{\alpha}_T \sim N(\alpha_{T|T}, P_{T|T})$, the main idea is that for a Gaussian state space model $p(\alpha_t|\psi_t,\alpha_{t+1})$ is a multivariate normal density and hence it is completely characterized by its first and second moments. The usual Kalman filter recursions are run, so that $\alpha_{t|t}$ is initially obtained. Then, the draw $\hat{\alpha}_{t+1} \sim p(\alpha_{t+1}|\psi_t,\alpha_{t+2})$ is treated as m additional observations and a second set of m Kalman filter recursions is run for each element of the state vector $\hat{\alpha}_{t+1}$. However, the latter procedure involves the inversion of system matrices, which are not necessarily non-singular.

3.2.5.3 Model specification

While state space models are widely used in time series analysis to deal with processes gradually changing over time, model specification is a challenge for these models as one has to specify which components to include and to decide whether they are fixed or time-varying. It leads to testing problems which are non-regular from the view-point of classical statistics. Thus, a classical approach toward model selection which is based on hypothesis testing such as a likelihood ratio test or information criteria such as AIC or BIC cannot be easily applied, because it relies on asymptotic arguments based on regularity conditions that are violated in this context. For example, we consider the time series $Y = (Y_1, ..., Y_T)$ for $t = 1, .., T$ modelled with the linear trend model

$$Y_t = \mu_t + \epsilon_t \, , \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

where μ_t is a random walk with a random drift starting from unknown initial values μ_0 and a_0

$$\begin{aligned} \mu_t &= \mu_{t-1} + a_{t-1} + \omega_{1t} \, , \omega_{1t} \sim N(0, \theta_1) \\ a_t &= a_{t-1} + \omega_{2t} \, , \omega_{2t} \sim N(0, \theta_2) \end{aligned}$$

In order to decide whether the drift a_t is time-varying or fixed we could test $\theta_2 = 0$ versus $\theta_2 > 0$. However, it is a nonregular testing problem since the null hypothesis lies on the boundary of the parameter space. Testing the null hypothesis $a_0 = a_1 = ... = a_T$ versus the alternative a_t follows a random walk is, again, non-regular because the size of the hypothesis increases with the number of observations. One possibility is to consider the Bayesian approach when dealing with such non-regular testing problems. We assume that there are K different candidates models $M_1, .., M_K$ for generating the time series Y. In a Bayesian setting each of these models is assigned a prior probability $p(M_k)$ with the goal of deriving the posterior model probability $p(M_k|Y)$ (the probability of a hypothesis M_k given the observed evidence Y) for each model M_k for $k = 1, .., K$. One strategie for computing the posterior model probabilities is to determine the posterior model probabilities of each model separately by using Bayes' rule $p(M_k|Y) \propto p(Y|M_k)p(M_k)$ where $p(Y|M_k)$ is the marginal likelihood for model M_k (it is the probability of observing Y given M_k). An explicit expression for the marginal likelihood exists only for conjugate problems like linear regression models with normally distributed errors, whereas for more complex models numerical techniques are required. For Gaussian state space models, marginal likelihoods have been estimated using methods such as importance sampling, Chib's estimator, numerical integration and bridge sampling. The modern approach to Bayesian model selection is to apply model space MCMC methods by sampling jointly model indicators and parameters, using for instance the reversible jump MCMC algorithm (see Green [1995]) or the stochastic variable selection approach (see George and McCulloch [1993] [1997]). The stochastic variable selection approach applied to model selection for regression models aims at identifying non-zero regression effects and allows parsimonious covariance modelling for longitudinal data. Fruhwirth-Schnatter et al. [2010a] considered the variable selection approach for many model selection problems occurring in state space modelling. In the above example, they used binary stochastic indicators in such a way that the unconstrained model corresponds to setting all indicators equal to 1. Reduced model specifications result by setting certain indicators equal to 0.

3.3 Asset returns and their characteristics

3.3.1 Defining financial returns

We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}_t$ is a right continuous filtration including all $\mathbb{P}$ negligible sets in $\mathcal{F}$. For simplicity, we let the market be complete and assume that there exists an equivalent martingale measure $\mathbb{Q}$ as defined in a mixed diffusion model by Bellamy and Jeanblanc [2000]. In the presence of continuous dividend yield, that unique probability measure equivalent to $\mathbb{P}$ is such that the discounted stock price plus the cumulated dividends are martingale when the riskless asset is the numeraire. In a general setting, we let the underlying process $(S_t)_{t\geq 0}$ be a one-dimensional Ito process valued in the open subset D.

3.3.1.1 Asset returns

Return series are easier to handle than price series due to their more attractive statistical properties and to the fact that they represent a complete and scale-free summary of the investment opportunity (see Campbell et al. [1997]). Expected returns need to be viewed over some time horizon, in some base currency, and using one of many possible averaging and compounded methods. Holding the asset for one period from date t to date $(t+1)$ would result in a simple gross return

$$1 + R_{t,t+1} = \frac{S_{t+1}}{S_t} \tag{3.3.4}$$

where the corresponding one-period simple net return, or simple return, $R_{t,t+1}$ is given by

$$R_{t,t+1} = \frac{S_{t+1}}{S_t} - 1 = \frac{S_{t+1} - S_t}{S_t}$$

More generally, we let

$$R_{t-d,t} = \frac{\nabla_d S_t + D_{t-d,t}}{S_{t-d}}$$

be the discrete return of the underlying process where $\nabla_d S_t = S_t - S_{t-d}$ with period d and where $D_{t-d,t}$ is the dividend over the period $[t-d, t]$. For simplicity we will only consider dividend-adjusted prices with discrete dividend-adjusted returns $R_{t-d,t} = \frac{\nabla_d S_t}{S_{t-d}}$. Hence, holding the asset for d periods between dates $t-d$ and t gives a d-period simple gross return

$$\begin{aligned} 1 + R_{t-d,t} &= \frac{S_t}{S_{t-d}} = \frac{S_t}{S_{t-1}} \times \frac{S_{t-1}}{S_{t-2}} \times ... \times \frac{S_{t-d+1}}{S_{t-d}} \\ &= (1 + R_{t-1,t})(1 + R_{t-2,t-1})...(1 + R_{t-d,t-d+1}) = \prod_{j=1}^{d}(1 + R_{t-j,t-j+1}) \end{aligned} \tag{3.3.5}$$

so that the d-period simple gross return is just the product of the d one-period simple gross returns which is called a compound return. Holding the asset for d years, then the annualised (average) return is defined as

$$R^A_{t-d,t} = \Big(\prod_{j=1}^{d}(1 + R_{t-j,t-j+1})\Big)^{\frac{1}{d}} - 1$$

which is the geometric mean of the d one-period simple gross returns involved and can be computed (see Appendix (B.10.4)) by

$$R^A_{t-d,t} = e^{\frac{1}{d}\sum_{j=1}^{d}\ln(1+R_{t-j,t-j+1})} - 1$$

It is simply the arithmetic mean of the logarithm returns $(1 + R_{t-j,t-j+1})$ for $j = 1, .., d$ which is then exponentiated to return the computation to the original scale. As it is easier to compute arithmetic average than geometric mean, and since the one-period returns tend to be small, one can use a first-order Taylor expansion [2] to approximate the annualised (average) return

$$R^A_{t-d,t} \approx \frac{1}{d}\sum_{j=1}^{d} R_{t-j,t-j+1}$$

[2] since $\log(1+x) \approx x$ for $|x| \leq 1$

Note, the arithmetic mean of two successive returns of $+50\%$ and -50% is 0%, but the geometric mean is -13% since $[(1+0.5)(1-0.5)]^{\frac{1}{d}} = 0.87$ with $d = 2$ periods. While some financial theory requires arithmetic mean as inputs (single-period Markowitz or mean-variance optimisation, single-period CAPM), most investors are interested in wealth compounding which is better captured by geometric means.

In general, the net asset value A of continuous compounding is

$$A = Ce^{r \times n}$$

where r is the interest rate per annum, C is the initial capital, and n is the number of years. Similarly,

$$C = Ae^{-r \times n}$$

is referred to as the present value of an asset that is worth A dollars n years from now, assuming that the continuously compounded interest rate is r per annum. If the gross return on a security is just $1 + R_{t-d,t}$, then the continuously compounded return or logarithmic return is

$$r_L(t-d,t) = \ln(1 + R_{t-d,t}) = L_t - L_{t-d} \tag{3.3.6}$$

where $L_t = \ln S_t$. Note, on a daily basis we get $R_t = R_{t-1,t}$ and $r_L(t) = \ln(1 + R_t)$. The change in log price is the yield or return, with continuous compounding, from holding the security from trading day $t-1$ to trading day t. As a result, the price becomes

$$S_t = S_{t-1}e^{r_L(t)}$$

Further, the return $r_L(t)$ has the property that the log return between the price at time t_1 and at time t_n is given by the sum of the $r_L(t)$ between t_1 and t_n

$$\log \frac{S_n}{S_1} = \sum_{i=1}^{n} r_L(t_i)$$

which implies that

$$S_n = S_1 e^{\sum_{i=1}^{n} r_L(t_i)}$$

so that if the $r_L(t)$ are independent random variables with finite mean and variance, the central limit theorem implies that for very large n, the summand in the above equation is normally distributed. Hence, we would get a log-normal distribution for S_n given S_1. In addition, the variability of simple price changes for a given security is an increasing function of the price level of the security, whereas this is not necessarily the case with the change in log price. Given Equation (3.3.5), then Equation (3.3.6) becomes

$$r_L(t-d,t) = r_L(t) + r_L(t-1) + r_L(t-2) + ... + r_L(t-d+1) \tag{3.3.7}$$

so that the continuously compounded multiperiod return is simply the sum of continuously compounded one-period returns. Note, statistical properties of log returns are more tractable. Moreover, in the cross section approach aggregation is done across the individual returns.

***Remark* 3.3.1** *That is, simple returns $R_{t-d,t}$ are additive across assets but not over time (see Equation (3.3.5)), wheras continuously compounded returns $r_L(t-d,t)$ are additive over time but not across assets (see Equation (3.3.7)).*

3.3.1.2 The percent returns versus the logarithm returns

In the financial industry, most measures of returns and indices use change of returns $R_t = R_{t-1,t}$ defined as $\frac{S_t - S_{t-1}}{S_{t-1}}$ where S_t is the price of a series at time t. However, some investors and researchers prefer to use returns based on logarithms of prices $r_t = \log \frac{S_t}{S_{t-1}}$ or compound returns. As discussed in Section (3.3.1.1), continuous time generalisations of discrete time results are easier, and returns over more than one day are simple functions of a single day return. In order to compare change and compound returns, Longerstaey et al. [1995a] compared kernel estimates of the probability density function for both returns. As opposed to the histogram of the data, this approach spreads the frequency represented by each observation along the horizontal axis according to a chosen distribution function, called the kernel, and chosen to be the normal distribution. They also compared daily volatility estimates for both types of returns based on an exponential weighting scheme. They concluded that the volatility forecasts were very similar. They used that methodology to compute the volatility for change returns and then replaced the change returns with logarithm returns. The same analysis was repeated on correlation by changing the inputs from change returns to logarithm returns. They also used monthly time series and found little difference between the two volatility and correlation series. Note, while the one month volatility and correlation estimators based on change and logarithm returns do not coincide, the difference between their point estimates is negligible.

3.3.1.3 Portfolio returns

We consider a portfolio consisting of N instruments, and let $r_L^i(t)$ and $R_i(t)$ for $i = 1, 2, .., N$ be respectively the continuously compounded and percent returns. We assign weights ω_i to the ith instrument in the portfolio together with a condition of no short sales $\sum_{i=1}^N \omega_i = 1$ (it is the percentage of the portfolio's value invested in that asset). We let P_0 be the initial value of the portfolio, and P_1 be the price after one period, then by using discrete compounding, we derive the usual expression for a portfolio return as

$$P_1 = \omega_1 P_0(1 + R_1) + \omega_2 P_0(1 + R_2) + ... + \omega_N P_0(1 + R_N) = \sum_{i=1}^N \omega_i P_0(1 + R_i)$$

We let $R_p(1) = \frac{P_1 - P_0}{P_0}$ be the return of the portfolio for the first period and replace P_1 with its value. We repeat the process at periods $t = 2, 3, ...$ to get the portfolio at time t as

$$R_p(t) = \omega_1 R_1(t) + \omega_2 R_2(t) + ... + \omega_N R_N(t) = \sum_{i=1}^N \omega_i R_i(t)$$

Hence, the simple net return of a portfolio consisting of N assets is a weighted average of the simple net returns of the assets involved. However, the continuously compounded returns of a portfolio do not have the above convenient property. The portfolio of returns satisfies

$$P_1 = \omega_1 P_0 e^{r_1} + \omega_2 P_0 e^{r_2} + ... + \omega_N P_0 e^{r_N}$$

and setting $r_p = \log \frac{P_1}{P_0}$, we get

$$r_p = \log \left(\omega_1 P_0 e^{r_1} + \omega_2 P_0 e^{r_2} + ... + \omega_N P_0 e^{r_N} \right)$$

Nonetheless, RiskMetrics (see Longerstaey et al. [1996]) uses logarithmic returns as the basis in all computations and the assumption that simple returns R_i are all small in magnitude. As a result, the portfolio return becomes a weighted average of logarithmic returns

$$r_p(t) \approx \sum_{i=1}^N \omega_i r_L^i(t)$$

since $\log(1 + x) \approx x$ for $|x| \leq 1$.

3.3.1.4 Modelling returns: The random walk

We are interested in characterising the future changes in the portfolio of returns described in Section (3.3.1.3), by forecasting each component of the portfolio using only past changes of market prices. To do so, we need to model

1. the temporal dynamics of returns
2. the distribution of returns at any point in time

Traditionally, to get tractable statistical properties of asset returns, financial markets assume that simple returns $\{R_{it}|t = 1, .., T\}$ are independently and identically distributed as normal with fixed mean and variance. However, while the lower bound of a simple return is -1, normal distribution may assume any value in the real line (no lower bound). Further, assuming that R_{it} is normally distributed, then the multi-period simple return $R_{it}(k)$ is not normally distributed. At last, the normality assumption is not supported by many empirical asset returns which tend to have positive excess kurtosis. Still, the random walk is one of the widely used class of models to characterise the development of price returns. In order to guarantee non-negativity of prices, we model the log price L_t as a random walk with independent and identically distributed (iid) normally distributed changes with mean μ and variance σ^2. It is given by

$$L_t = \mu + L_{t-1} + \sigma\epsilon_t \ , \epsilon_t \sim iidN(0, 1)$$

The use of log prices, implies that the model has continuously compounded returns, that is, $r_t = L_t - L_{t-1} = \mu + \sigma\epsilon_t$ with mean and variance

$$E[r_t] = \mu \ , Var(r_t) = \sigma^2$$

Hence, an expression for prices can be derived as

$$S_t = S_{t-1}e^{\mu+\sigma\epsilon_t}$$

and S_t follows the lognormal distribution. Hence, the mean and variance of simple returns become

$$E[R_t] = e^{\mu+\frac{1}{2}\sigma^2} - 1 \ , Var(R_t) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

which can be used in forecasting asset returns. There is no lower bound for r_t, and the lower bound for R_t is satisfied using $1 + R_t = e^{r_t}$. Assuming that logarithmic price changes are i.i.d. implies that

- at each point in time t the log price changes are identically distributed with mean μ and variance σ^2 implying homoskedasticity (unchanging prices over time).
- log price changes are statistically independent of each other over time (the values of returns sampled at different points are completely unrelated).

However, the lognormal assumption is not consistent with all the properties of historical stock returns. The above models assume a constant variance in price changes, which in practice is flawed in most financial time series data. We can relax this assumption to let the variance vary with time in the modified model

$$L_t = \mu + L_{t-1} + \sigma_t\epsilon_t \ , \epsilon_t \sim N(0, 1)$$

These random walk models imply certain movement of financial prices over time.

3.3.2 The properties of returns

3.3.2.1 The distribution of returns

As explained by Tsay [2002] when studying the distributional properties of asset returns, the objective is to understand the behaviour of the returns across assets and over time, in order to characterise the portfolio of returns described in Section (3.3.1.3). We consider a collection of N assets held for T time periods $t = 1, .., T$. The most general model for the log returns $\{r_{it}; i = 1, .., N; t = 1, .., T\}$ is its joint distribution function

$$F_r\big(r_{11}, ..., r_{N1}; r_{12}, ..., r_{N2}; ..; r_{1T}, ..., r_{NT}; Y; \theta\big) \tag{3.3.8}$$

where Y is a state vector consisting of variables that summarise the environment in which asset returns are determined and θ is a vector of parameters that uniquely determine the distribution function $F_r(.)$. The probability distribution $F_r(.)$ governs the stochastic behavior of the returns r_{it} and the state vector Y. In general Y is treated as given and the main concern is defining the conditional distribution of $\{r_{it}\}$ given Y. Empirical analysis of asset returns is then to estimate the unknown parameter θ and to draw statistical inference about behavior of $\{r_{it}\}$ given some past log returns. Consequently, Equation (3.3.8) provides a general framework with respect to which an econometric model for asset returns r_{it} can be put in a proper perspective. For instance, financial theories such as the Capital Asset Pricing Model (CAPM) of Sharpe focus on the joint distribution of N returns at a single time index t, that is, $\{r_{1t}, ..., r_{Nt}\}$, while theories emphasise the dynamic structure of individual asset returns, that is, $\{r_{i1}, ..., r_{iT}\}$ for a given asset i. When dealing with the joint distribution of $\{r_{it}\}_{t=1}^T$ for asset i, it is useful to partition the joint distribution as

$$\begin{aligned} F_r(r_{i1}, ..., r_{iT}; \theta) &= F(r_{i1})F(r_{i2}|r_{i1})...F(r_{iT}|r_{i,T-1}, ..., r_{i,1}) \\ &= F(r_{i1}) \prod_{t=2}^{T} F(r_{it}|r_{i,t-1}, ..., r_{i,1}) \end{aligned} \tag{3.3.9}$$

highlighting the temporal dependencies of the log return. As a result, one is left to specify the conditional distribution $F(r_{it}|r_{i,t-1}, ..., r_{i,1})$ and the way it evolves over time. Different distributional specifications leads to different theories. In one version of the random-walk, the hypothesis is that the conditional distribution is equal to the marginal distribution $F(r_{it})$ so that returns are temporally independent and, hence, not predictable. In general, asset returns are assumed to be continuous random variables so that one need to know their probability density functions to perform some analysis. Using the relation among joint, marginal, and conditional distributions we can write the partition as

$$f_r(r_{i1}, ..., r_{iT}; \theta) = f(r_{i1}; \theta) \prod_{t=2}^{T} f(r_{it}|r_{i,t-1}, ..., r_{i,1}; \theta)$$

In general, it is easier to estimate marginal distributions than conditional distributions using past returns. Several statistical distributions have been proposed in the literature for the marginal distributions of asset returns (see Tsay [2002]). A traditional assumption made in financial study is that the simple returns $\{R_{it}\}_{t=1}^T$ are independently and identically distributed as normal with fixed mean and variance. However, the lower bound of a simple return is -1 but normal distribution may assume any value in the real line having no lower bound. Further, the normality assumption is not supported by many empirical asset returns, which tend to have a positive excess kurtosis. To overcome the first problem, assumption is that the log returns r_t of an asset is independent and identically distributed (iid) as normal with mean μ and variance σ^2.

The multivariate analyses are concerned with the joint distribution of $\{r_t\}_{t=1}^T$ where $r_t = (r_{1t}, .., r_{Nt})^\top$ is the log returns of N assets at time t. This joint distribution can be partitioned in the same way as above so that the analysis focusses on the specification of the conditional distribution function

$$F(r_t|r_{t-1}, .., r_1; \theta)$$

in particular, how the conditional expectation and conditional covariance matrix of r_t evolve over time. The mean vector and covariance matrix of a random vector $X = (X_1, .., X_p)$ are defined as

$$\begin{aligned} E[X] &= \mu_X = (E[X_1], .., E[X_p])^\top \\ Cov(X) &= \Sigma_X = E[(X - \mu_X)(X - \mu_X)^\top] \end{aligned}$$

provided that the expectations involved exist. When the data $\{x_1, .., x_T\}$ of X are available, the sample mean and covariance matrix are defined as

$$\hat{\mu}_X = \frac{1}{T}\sum_{t=1}^{T} x_t \, , \hat{\Sigma}_x = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{\mu}_X)(x_t - \hat{\mu}_X)^\top$$

These sample statistics are consistent estimates of their theoretical counterparts provided that the covariance matrix of X exists. In the finance literature, multivariate normal distribution is often used for the log return r_t.

3.3.2.2 The likelihood function

One can use the partition in Equation (3.3.9) to obtain the likelihood function of the log returns $\{r_1, ..., r_T\}$ for the ith asset. Assuming that the conditional distribution $f(r_t|r_{t-1}, ..., r_1; \Theta)$ is normal with mean μ_t and variance σ_t^2, then Θ consists of the parameters μ_t and σ_t^2 and the likelihood function of the data is

$$f(r_1, ..., r_T; \Theta) = f(r_1; \Theta) \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi}\sigma_t} e^{\frac{-(r_t - \mu_t)^2}{2\sigma_t^2}}$$

where $f(r_1; \Theta)$ is the marginal density function of the first observation r_1. The value Θ^* maximising this likelihood function is the maximum likelihood estimate (MLE) of Θ. The log function being monotone, the MLE can be obtained by maximising the log likelihood function

$$\ln f(r_1, ..., r_T; \Theta) = \ln f(r_1; \Theta) - \frac{1}{2}\sum_{t=2}^{T}[\ln 2\pi + \ln(\sigma_t^2) + \frac{(r_t - \mu_t)^2}{\sigma_t^2}] \tag{3.3.10}$$

Note, even if the conditional distribution $f(r_t|r_{t-1}, ..., r_1; \Theta)$ is not normal, one can still compute the log likelihood function of the data.

3.3.3 Testing the series against trend

We have made the assumption of independently distributed returns in Section (3.3.1.4) which is at the heart of the efficient market hypothesis (EMH), but we saw in Section (1.7.6) that the technical community totally rejected the idea of purely random prices. In fact, portfolio returns are significantly autocorrelated leading to contrarian and momentum strategies. One must therefore test time series for trends. When testing against trend we are testing the hypothesis that the members of a sequence of random variables $x_1, .., x_n$ are distributed independently of each other, each with the same distribution. Following the definition of trend given by Mann [1945], a sequence of random variables $x_1, .., x_n$ is said to have a downward trend if the variables are independently distributed so that x_i has the cumulative distribution f_i and $f_i(x) < f_j(x)$ for every $i < j$ and every x. Similarly, an upward trend is defined with $f_i(x) > f_j(x)$ for every $i < j$.

Since all statistical tests involve the type I error (rejecting the null hypothesis when it is true), and the type II error (not rejecting the null hypothesis when it is false), it is important to consider the power of a test, defined as one minus the probability of type II error. A powerful test will reject a false hypothesis with a high probability. Studying

the existence of trends in hydrological time series, Onoz et al. [2002] compared the power of parametric and non-parametric tests for trend detection for various probability distributions estimated by Monte-Carlo simulation. The parametric test considers the linear regression of the random variable Y on time X with the regression coefficient b (or the Pearson correlation coefficient r) computed from the data. The statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b}{\frac{s}{S_q^x}}$$

follows the Student's t distribution with $(n-2)$ degrees of freedom, where n is the sample size, s is the standard deviation of the residuals, and S_q^x is the sums of squares of the independent variable (time in trend analysis). For non-parametric tests, Yue et al. [2002] showed that the Spearman's rho test provided results almost identical to those obtained with the Mann-Kendall test. Hence, we consider only the non-parametric Mann-Kendall test for analysing trends. Kendall [1938] introduced the T-test for testing the independence in a bivariate distribution by counting the number of inequalities $x_i < x_j$ for $i < j$ and computing the distribution of T via a recursive equation. Mann [1945] introduced lower and upper bounds for the power of the T-test. He proposed a trend detection by considering the statistic

$$S_M^n(t) = \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} sign(y_{t-i} - y_{t-j})$$

where each pair of observed values (y_i, y_j) for $i > j$ of the random variable is inspected to find out whether $y_i > y_j$ or $y_i < y_j$. If $P(t)$ is the number of the former type of pairs, and $M(t)$ is the number of the latter type of pairs, the sattistic becomes $S_M^n(t) = P(t) - M(t)$. The variance of the statistic is

$$Var(S_M^n(t)) = \frac{1}{18}n(n-1)(2n+5)$$

so that the statistic is bounded by

$$-\frac{1}{2}n(n+1) \leq S_M^n(t) \leq \frac{1}{2}n(n+1)$$

The bounds are reached when $y_t < y_{t-i}$ (negative trend) or $y_t > y_{t-i}$ (positive trend) for $i \in \mathbb{N}^*$. Hence, we obtain the normalised score

$$\overline{S}_M^n(t) = \frac{2}{n(n+1)} S_M^n(t)$$

where $\overline{S}_M^n(t)$ takes the value 1 (or -1) if we have a perfect positive (or negative) trend. In absence of trend we get $\overline{S}_M^n(t) \approx 0$. Letting that statement be the null hypothesis (no trend), we get

$$Z^n(t) \to_{n\to\infty} N(0,1)$$

with

$$Z^n(t) = \frac{S_M^n(t)}{\sqrt{Var(S_M^n(t))}}$$

The null hypothesis that there is no trend is rejected when $Z^n(t)$ is greater than $z_{\frac{\alpha}{2}}$ in absolute value. Note, parametric tests assume that the random variable is normally distributed and homosedastic (homogeneous variance), while non-parametric tests make no assumption on the probability distribution. The t-test for trend detection is based on linear regression, thus, checks only for a linear trend. There is no such restriction for the Mann-Kendall test. Further, MK is expected to be less affected by outliers as its statistic is based on the sign of the differences, and not directly on the values of the random variable. Plotting the ratio of the power of the t-test to that of the Mann-Kendall test as

function of the slope of the trend of a large number of simulated time series, Yue et al. [2002] showed that the power of the Mann-Kendall trend test was dependent on the distribution types, and was increasing with the coefficient of skewness. Onoz et al. [2002] repeated the experiment on various distributions obtaining a ratio slightly above one for the normal distribution, and for all other (non-normal) distributions the ratio was significantly less than one. For skewed distributions, the Mann-Kendall test was more powerful, especially for high coefficient of skewness.

3.3.4 Testing the assumption of normally distributed returns

We saw in Section (2.3.1.1) that the mean-variance efficient portfolios introduced by Markowitz [1952] require some fairly restrictive assumptions on the class of return distribution, such as the assumption of normally distributed returns. Further, the Sharpe type metrics for performance measures described in Section (2.4.2) depend on returns of individual assets that are jointly normally distributed. Hence, one must be able to assess the suiability of the normal assumption, and to quantitfy the deviations from normality.

3.3.4.1 Testing for the fitness of the Normal distribution

While changes in financial asset prices are known to be non-normally distributed, practitioners still assume that they are normally distributed because they can make predictions on their conditional mean and variance. There has been much discussion about the usefulness of the underlying assumption of normality for return series. One way foreward is to compare directly the predictions made ny the Normal model to what we observe. In general, practitiners use simple heuristics to assess model performances such as measuring

- the difference between observed and predicted frequecies of observations in the tail of the normal distribution.
- the difference between observed and predicted values of these tail observations.

In the case of univariate tail probabilities, we let $R_t = R_{t-1,t} = \frac{S_t - S_{t-1}}{S_{t-1}}$ be the percent return at time t and $\sigma_t = \sigma_{t|t-1}$ be the one day forecast standard deviation of R_t. The theoretical tail probabilities corresponding to the lower and upper tail areas are given by

$$P(R_t < -1.65\sigma_t) = 5\% \text{ and } P(R_t > 1.65\sigma_t) = 5\%$$

Letting T be the total number of returns observed over a given sample period, the observed tail probabilities are given by

$$\frac{1}{T}\sum_{t=1}^{T} I_{\{R_t < -1.65\hat{\sigma}_t\}} \text{ and } \frac{1}{T}\sum_{t=1}^{T} I_{\{R_t > 1.65\hat{\sigma}_t\}}$$

where $\hat{\sigma}_t$ is the estimated standard deviation using a particular model. Having estimated the tail probabilities, we are now interested in the value of the observations falling in the tail area, called tail points. We need to check the predictable quality of the Normal model by comparing the observed tail points and the predicted values. In the case of the lower tail, we first record the value of the observations $R_t < -1.65\hat{\sigma}_t$ and then find the average value of these returns. We want to derive forecasts of these tail points called the predicted values. Since we assumed normally distributed returns, the best guess of any return is simply its expected value. Therefore, the predicted value for an observation falling in the lower tail at time t is

$$E[R_t | R_t < -1.65\sigma_t] = -\sigma_t \lambda(-1.65) \text{ and } \lambda(x) = \frac{\phi(x)}{N(x)}$$

where $\phi(.)$ is the standard normal density, and $N(.)$ is the standard normal cumulative distribution function. The same heuristic test must be performed on correlation. Assuming a portfolio made of two underlyings R_1 and R_2, its daily earning at risks (DEaR) is given by

$$DEaR(R_1, R_2) = \sqrt{V^\top CV}$$

where $V = (DEaR(R_1), DEaR(R_2))^\top$ is a transpose vector, and C is the 2×2 correlation matrix. In the bivariate case, we want to analyse the probabilities associated with the joint distribution of the two return series to assess the performance of the model. We consider the event $P(\frac{R_1(t)}{\sigma_1(t)} < 0$ and $\frac{R_2(t)}{\sigma_2(t)} < -1.65)$ where the choice for $R_1(t)$ being less than zero is arbitrary. The observed values are given by

$$\frac{1}{T}\sum_{t=1}^{T} I_{\{\frac{R_1(t)}{\sigma_1(t)}<0 \text{ and } \frac{R_2(t)}{\sigma_2(t)}\}} \times 100$$

and the predicted probability is obtained by integrating over the bivariate density function

$$B(0, -1.65, \rho) = \int_{-\infty}^{0}\int_{-\infty}^{-1.65} \phi(x_1, x_2, \rho)dx_1dx_2$$

where $\phi(x_1, x_2, \rho)$ is the standard normal bivariate density function, and ρ is the correlation between S_1 and S_2. For any pair of returns, we are now interested in the value of one return when the other is a tail point. The observed values of return S_1 are the average of the $R_1(t)$ when $R_2(t) < -1.65\sigma_2(t)$. Hence, we first record the value of the observations $R_1(t)$ corresponding to $R_2(t) < -1.65\sigma_2(t)$ and then find the average value of these returns. Based on the assumption of normality for the returns we can derive the forecasts of these tail points called the predicted values. Again, the best guess is the expected value of $R_1(t)|R_2(t) < -1.65\sigma_2(t)$ given by

$$E[R_1(t)|R_2(t) < -1.65\sigma_2(t)] = -\sigma_1(t)\rho\lambda(-1.65)$$

In the case of individual returns, assuming an exponentially weighted moving average (EWMA) with decay factor 0.94 for the estimated volatility, Longerstaey et al. [1995a] concluded that the observed tail frequencies and points match up quite well their predictions from the Normal model. Similarly, in the bivariate case (except for money market rates), the Normal model's predictions of frequencies and tail points coincided with the observed ones.

3.3.4.2 Quantifying deviations from a Normal distribution

For more than a century, the problem of testing whether a sample is from a normal population has attracted the attention of leading figures in statistics. The absence of exact solutions for the sampling distributions generated a large number of simulation studies exploring the power of these statistics as both directional and omnibus tests (see D'Agostino et al. [1973]). A wide variety of tests are available for testing goodness of fit to the normal distribution. If the data is grouped into bins, with several counts in each bin, Pearson's chi-square test for goodness of fit maybe applied. In order for the limiting distribution to be chi-square, the parameters must be estimated from the grouped data. On the other hand, departures from normality often take the form of asymmetry, or skewness. It happens out of this that mean and variance are no longer sufficient to give a complete description of returns distribution. In fact, skewness and kurtosis (the third and fourth central moments) have to be taken into account to describe a stock (index) returns' probability distribution entirely. To check whether the skewness and kurtosis statistics of an asset can still be regarded as normal, the Jarque-Bera statistic (see Jarque et al. [1987]) can be applied. We are going to briefly describe the Omnibus test for normality (two sided) where the information in b_1 and b_2 is indicated for a general non-normal alternative. Note, when a specific alternative distribution is indicated, one can use a specific likelihood ratio test, increasing power. For instance, when information on the alternative distribution exists, a directional test using only b_1 and b_2 is preferable. However, the number of cases where such directional tests are available is limited for practical application. Let $X_1, ..., X_n$ be independent random variables with absolutely continuous distributions function F. We wish to test

$$H_0 : F(x) = N(\frac{x-\mu}{\sigma}) \, , \forall x \in \mathbb{R}$$

versus the two sided alternative

$$H_1 : F(x) \neq N(\frac{x-\mu}{\sigma}) \text{ for at least one } x \in \mathbb{R}$$

where $N(.)$ is the cdf of the standard normal distribution and $\sigma(\sigma > 0)$ may be known or unknown. In practice, the null hypothesis of normality is usually specified in composite form where μ and σ are unknown. When performing a test hypothesis, the p-value is found by using the distribution assumed for the test statistic under H_0. However, the accuracy of the p-value depends on how close the assumed distribution is to the true distribution of the test statistic under the null hypothesis.

Suppose that we want to test the null hypothesis that the returns $R_i(1), R_i(2), ..., R_i(n)$ for the ith asset are independent normally distributed random variables with the same mean and variance. A goodness-of-fit test can be based on the coefficient of skewness for the sample of size n

$$b_1 = \frac{\hat{m}_3^c}{(\hat{m}_2^c)^{\frac{3}{2}}} = \frac{\frac{1}{n}\sum_{j=1}^n (R_i(j) - \overline{R}_i)^3}{S^3}$$

where m_2^c and m_3^c are the theoretical second and third central moments, respectively, with its sample estimates

$$\hat{m}_j^c = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^j \ , j = 2, 3, 4$$

The test rejects for large values of $|b_1|$.

***Remark* 3.3.2** *In some articles, the notation for b_1 is sometime slightly different, with b_1 replaced with $\sqrt{b_1} = \frac{\hat{m}_3^c}{(\hat{m}_2^c)^{\frac{3}{2}}}$.*

Note, skewness is a non-dimensional quantity characterising only the shape of the distribution. For the idealised case of a normal distribution, the standad deviation of the skew coefficient b_1 is approximately $\sqrt{\frac{15}{n}}$. Hence, it is good practice to believe in skewness only when they are several or many times as large as this. Departure of returns from the mean may be detected by the coefficient of kurtosis for the sample

$$b_2 = \frac{\hat{m}_4^c}{(\hat{m}_2^c)^2} = \frac{\frac{1}{n}\sum_{j=1}^n (R_i(j) - \overline{R}_i)^4}{S^4}$$

The kurtosis is also a non-dimensional quantity. To test kurtosis we can compute

$$kurt = b_2 - 3$$

to recover the zero-value of a normal distribution. The standard deviation of $kurt$ as an estimator of the kurtosis of an underlying normal distribution is $\sqrt{\frac{96}{n}}$.

The estimates of skewness and kurtosis are used in the Jarque-Bera (JB) statistic (see Jarque et al. [1987]) to analyse time series of returns for the assumption of normality. It is a goodness-of-fit measure with an asymptotic χ^2-distribution with two degrees of freedom (because JB is just the sum of squares of two asymptotically independent standardised normals)

$$JB \sim \chi_2^2 \ , n \to \infty \text{ under } H_0$$

However, in general the χ^2 approximation does not work well due to the slow convergence to the asymptotic results. Jarque et al. showed, with convincing evidence, that convergence of the sampling distributions to asynmptotic results was very slow, especially for b_2. Nonetheless, the JB test can be used to test the null hypothesis that the data are from a normal distribution. That means that H_0 has to be rejected at level α if

$$JB \geq \chi^2_{1-\alpha,2}$$

At the 5% significance level the critical value is equal to 5.9915. The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being 0, as samples from a normal distribution have an expected skewness of 0 and an expected excess kurtosis of 0. As the definition of JB shows, any deviation from this increases the JB statistic

$$JB = \frac{n}{6}(b_1^2 + \frac{(b_2-3)^2}{4})$$

where b_1 is the coefficient of skewness and b_2 is the coefficient of kurtosis. Note, Urzua [1996] introduced a modification to the JB test by standardising the skewness b_1 and the kurtosis b_2 in the JB formula, getting

$$JBU = \frac{b_1^2}{v_S} + \frac{(b_2 - e_K)^2}{v_K}$$

with

$$v_S = \frac{6(n-2)}{(n+1)(n+3)}\ , e_K = \frac{3(n-1)}{(n+1)}\ , v_K = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Note, JB and JBU are asymptotically equivalent, that is, H_0 has to be rejected at level α if $JBU \geq \chi^2_{1-\alpha,2}$. Critical values of tests for various sample sizes n with $\alpha = 0.05$ are

$$n = 50\ JB = 5.0037\ , n = 100\ JB = 5.4479\ , n = 200\ JB = 5.7275\ , n = 500\ JB = 5.8246$$

See Thadewald et al. [2004] for tables. Testing for normality, they investigated the power of several tests by considering independent random variables (model I), and the residual in the classical linear regression (model II). The power comparison was carried out via Monte Carlo simulation with a model of contaminated normal distributions (mixture of normal distributions) with varying parameters μ and σ as well as different proportions of contamination. They found that for the JB test, the approximation of critical values by the chi-square distribution did not work well. The test was superior in power to its competitors for symmetric distributions with medium up to long tails and for slightly skewed distributions with long tails. The power of the JB test was poor for distributions with short tails, especially bimodal shape. Further, testing for normality is problematic in the case of autocorrelated error terms and in the case of heteroscedastic error terms.

3.3.5 The sample moments

We assume that the population is of size N and that associated with each member of the population is a numerical value of interest denoted by $x_1, x_2, .., x_N$. We take a sample with replacement of n values $X_1, ..., X_n$ from the population, where $n < N$ and such that X_i is a random variable. That is, X_i is the value of the ith member of the sample, and x_i is that of the ith member of the population. The population moments and the sample moments are given in Appendix (B.10.1).

3.3.5.1 The population mean and volatility

While volatility is a parameter measuring the risk associated with the returns of the underlying price, local volatility is a parameter measuring the risk associated with the instantaneous variation of the underlying price. It can be deterministic or stochastic. On the other hand, historical volatility is computed by using historical data on observed values of the underlying price (opening, closing, highest value, lowest value etc...). In general, one uses standard estimators of variance per unit of time of the logarithm of the underlying price which is assumed to follow a non-centred Brownian motion (see Section (3.3.1.4)). Considering observed values uniformally allocated in time with time difference δ, the stock price a time $t+1 = (j+1)\delta$ is given by

$$S_{(j+1)\delta} = S_{j\delta} e^{\mu\delta - \sigma(W_{(j+1)\delta} - W_{j\delta})}$$

with

$$\ln \frac{S_{(j+1)\delta}}{S_{j\delta}} = \mu\delta - \sigma(W_{(j+1)\delta} - W_{j\delta})$$

Given N period of time, the first two sample moments are

$$\begin{aligned} \hat{\mu}_N &= \frac{1}{N}\sum_{j=0}^{N-1} \ln \frac{S_{(j+1)\delta}}{S_{j\delta}} \\ \hat{\sigma}_N^2 &= \frac{1}{N}\sum_{j=0}^{N-1} \big(\ln \frac{S_{(j+1)\delta}}{S_{j\delta}} - \hat{\mu}_N\big)^2 \end{aligned}$$

Alternatively, we can consider the return of the underlying price given by

$$R_{(j+1)\delta} = \frac{S_{(j+1)\delta} - S_{j\delta}}{S_{j\delta}} = e^{\mu\delta - \sigma(W_{(j+1)\delta} - W_{j\delta})} - 1$$

Considering the expansion $e^x \approx 1 + x$ for $|x| < 1$ and assuming $\mu\delta - \sigma(W_{(j+1)\delta} - W_{j\delta})$ to be small, we get

$$R_{(j+1)\delta} = \frac{S_{(j+1)\delta} - S_{j\delta}}{S_{j\delta}} \approx \mu\delta - \sigma(W_{(j+1)\delta} - W_{j\delta})$$

and we set

$$\begin{aligned} \tilde{\mu}_N &= \frac{1}{N}\sum_{j=0}^{N-1} \frac{S_{(j+1)\delta} - S_{j\delta}}{S_{j\delta}} \\ \tilde{\sigma}_N^2 &= \frac{1}{N}\sum_{j=0}^{N-1} \big(\frac{S_{(j+1)\delta} - S_{j\delta}}{S_{j\delta}} - \tilde{\mu}_N\big)^2 \end{aligned}$$

In order to use statistical techniques on market data, one must make sure that the data is stationary, and test the assumption of log-normality on the observed values of the underlying price. Note, the underlying prices rarely satisfy the Black-Scholes assumption. However, we are not trying to compute option prices in the risk-neutral measure, but we are estimating the moments of the stock returns under the historical measure.

3.3.5.2 The population skewness and kurtosis

As defined in Appendix (B.4), skewness is a statistical measure of the asymmetry of the probability distribution of a random variable, in this case the return of a stock. Since a normal distribution is symmetrical, it exhibits exactly zero skewness. The more asymmetric and thus unlike a normal distribution, the larger the figure gets in absolute terms. Given N period of time, the sample skewness is

$$\hat{S}_N = \frac{1}{N\hat{\sigma}_N^3}\sum_{j=0}^{N-1} \big(\ln \frac{S_{(j+1)\delta}}{S_{j\delta}} - \hat{\mu}_N\big)^3$$

Kurtosis is a measure of the Peakedness of a probability distribution of random variables. As such, it discloses how concentrated a return distribution is around the mean. Higher kurtosis means more of the variance is due to infrequent

extreme deviations (fat tails), lower kurtosis implies a variance composed of frequent modestly-sized deviations. Normally distributed asset returns exhibit a kurtosis of 3. To measure the excess kurtosis with regard to a normal distribution, a value of 3 is hence subtracted from the kurtosis value. A distribution with positive excess kurtosis is called leptokurtic. We can calculate the sample kurtosis of a single asset class with

$$\hat{K}_N \frac{1}{N\hat{\sigma}_N^4} \sum_{j=0}^{N-1} \left(\ln \frac{S_{(j+1)\delta}}{S_{j\delta}} - \hat{\mu}_N\right)^4$$

Under normality assumption, skew and kurtosis are distributed asymptotically as normal with mean equal to zero and variance being $\frac{6}{N}$ and $\frac{24}{N}$ respectively (see Snedecor et al. [1980]).

3.3.5.3 Annualisation of the first two moments

The sample estimates of the first two moments are often based on monthly, weekly or daily data, but all quantities are usually quoted in annualised terms. Annualisation is often performed on the sample estimates under the assumption that the random variables (returns) are i.i.d. We let the annualised volatility σ be the standard deviation of the instrument's yearly logarithmic returns. The generalised volatility σ_T for time horizon T in years is expressed as $\sigma_T = \sigma\sqrt{T}$. Therefore, if the daily logarithmic returns of a stock have a standard deviation of σ_{SD} and the time period of returns is P (or Δt), the annualised volatility is

$$\sigma = \frac{\sigma_{SD}}{\sqrt{P}}$$

A common assumption for daily returns is that $P = 1/252$ (there are 252 trading days in any given year). Then, if $\sigma_{SD} = 0.01$ the annualised volatility is

$$\sigma = \frac{0.01}{\sqrt{\frac{1}{252}}} = 0.01\sqrt{252}$$

More generally, we have

$$X \times \frac{F}{N}$$

where X is the sum of all values referenced, F is the base rate of return (time period frequency) with 12 monthly, 252 daily, 52 weekly, 4 quaterly, and N is the total number of periods. For example, setting $P = \frac{1}{F}$ the annualised mean and variance becomes

$$\begin{aligned} \mu &= \frac{\mu_{SD}}{P} = \mu_{SD}252 \\ \sigma^2 &= \frac{\sigma_{SD}^2}{P} = \sigma_{SD}^2 252 \end{aligned}$$

This formula to convert returns or volatility measures from one time period to another assume a particular underlying model or process. These formulas are accurate extrapolations of a random walk, or Wiener process, whose steps have finite variance. However, if portfolio returns are autocorrelated, the standard deviation does not obey the square-root-of-time rule (see Section (10.1.1)). Again, F is the time period frequency (number of returns per year), then the annualised mean return is still F times the mean return, but the standard deviation of returns should be calculated using the scaling factor

$$\sqrt{F + 2\frac{Q}{(1-Q)^2}\left[(F-1)(1-Q) - Q(1-Q^{F-1})\right]} \tag{3.3.11}$$

where Q is the first order autocorrelation of the returns (see Alexander [2008]). If the autocorrelation of the returns is positive, then the scaling factor is greater than the square root of F. More generally, for natural stochastic processes, the precise relationship between volatility measures for different time periods is more complicated. Some researchers use the Levy stability exponent α (linked to the Hurst exponent) to extrapolate natural processes

$$\sigma_T = T^{\frac{1}{\alpha}}\sigma$$

If $\alpha = 2$ we get the Wiener process scaling relation, but some people believe $\alpha < 2$ for financial activities such as stocks, indexes and so on. Mandelbrot [1967] followed a Levy alpha-stable distribution with $\alpha = 1.7$. Given our previous example with $P = 1/252$ we get

$$\sigma = 0.01(252)^{\frac{1}{\alpha}}$$

We let $\alpha_W = 2$ for the Wiener process and $\alpha_M = 1.7$ for the Levy alpha-stable distribution. Since $\frac{1}{\alpha_M} > \frac{1}{\alpha_W}$ we get $\hat{\alpha}_M = \frac{1}{\alpha_M} = \hat{\alpha}_W + \xi$ we get $(252)^{\hat{\alpha}_M} = (252)^{\hat{\alpha}_W}(252)^{\xi}$ so that

$$\sigma_M = 0.01(252)^{\hat{\alpha}_N}(252)^{\xi}$$

with $\xi = 0.09$ and $(252)^{\xi} = 1.64$. Hence, we get the Mandelbrot annualised volatility

$$\sigma_M = \sigma_{SD}\sqrt{252}(1 + \epsilon_S) = \sigma_W(1 + \epsilon_S)$$

where $\epsilon_S \geq 0$ is the adjusted volatility.

3.4 Introducing the volatility process

3.4.1 An overview of risk and volatility

3.4.1.1 The need to forecast volatility

When visualising financial time series, one can observe heteroskedasticity [3], with periods of high volatility and periods of low volatility, corresponding to periods of high and low risks, respectively. We also observe returns having very high absolute value compared with their mean, suggesting fat tail distribution for returns, with large events having a larger probability to appear when compared to returns drawn from a Gaussian distribution. Hence, besides the return series introduced in Section (3.3.1), we must also consider the volatility process and the behaviour of extreme returns of an asset (the large positive or negative returns). The negative extreme returns are important in risk management, whereas positive extreme returns are critical to holding a short position. Volatility is important in risk management as it provides a simple approach to calculating the value at risk (VaR) of a financial position. Further, modelling the volatility of a time series can improve the efficiency in parameter estimation and the accuracy in interval forecast. As returns may vary substantially over time and appear in clusters, the volatility process is concerned with the evolution of conditional variance of the return over time. When using risk management models and measures of preference, users must make sure that volatilities and correlations are predictable and that their forecasts incorporate the most useful information available. As the forecasts are based on historical data, the estimators must be flexible enough to account for changing market conditions. One simple approach is to assume that returns are governed by the random walk model described in Section (3.3.1.4), and that the sample standard deviation $\hat{\sigma}_N$ or the sample variance $\hat{\sigma}_N^2$ of returns for N periods of time can be used as a simple forecast of volatility of returns, r_t, over the future period $[t + 1, t + h]$ for some positive integer h. However, volatility has some specific characteristics such as

- volatility clusters: volatility may be high for certain time periods and low for other periods.
- continuity: volatility jumps are rare.

[3] Heteroskedastic means that a time series has a non-constant variance through time.

- mean-reversion: volatility does not diverge to infinity, it varies within some fixed range so that it is often stationary.
- volatility reacts differently to a big price increase or a big price drop.

These properties play an important role in the development of volatility models. As a result, there is a large literature on econometric models available for modelling the volatility of an asset return, called the conditional heteroscedastic (CH) models. Some univariate volatility models include the autoregressive conditional heteroscedastic (ARCH) model of Engle [1982], the generalised ARCH (GARCH) model of Bollerslev [1986], the exponential GARCH (EGARCH) of Nelson [1991], the stochastic volatility (SV) models and many more. Tsay [2002] discussed the advantages and weaknesses of each volatility model and showed some applications of the models. Following his approach, we will describe some of these models in Section (5.6). Unfortunately, stock volatility is not directly observable from returns as in the case of daily volatility where there is only one observation in a trading day. Even though one can use intraday data to estimate daily volatility, accuracy is difficult to obtain. The unobservability of volatility makes it difficult to evaluate the forecasting performance of CH models and heuristics must be developed to estimate volatility on small samples.

3.4.1.2 A first decomposition

As risk is mainly given by the probability of large negative returns in the forthcoming period, risk evaluation is closely related to time series forecasts. The desired quantity is a forecast for the probability distribution (pdf) $\tilde{p}(r)$ of the possible returns r over the risk horizon ΔT. This problem is generally decomposed into forecasts for the mean and variance of the return probability distribution

$$r_{\Delta T} = \mu_{\Delta T} + a_{\Delta T}$$

with

$$a_{\Delta T} = \sigma_{\Delta T}\epsilon$$

where the return $r_{\Delta T}$ over the period ΔT is a random variable, $\mu_{\Delta T}$ is the forecast for the mean return, and $\sigma_{\Delta t}$ is the volatility forecast. The term $a_{\Delta T} = r_{\Delta T} - \mu_{\Delta T}$ is the mean-corrected asset return. The residual ϵ, which corresponds to the unpredictable part, is a random variable distributed according to a pdf $p_{\Delta T}(\epsilon)$. The standard assumption is to let $\epsilon(t)$ be an independent and identically distributed (iid) random variable. In general, a risk methodology will set the mean μ to zero and concentrate on σ and $p(\epsilon)$. To validate the methodology, we set

$$\epsilon = \frac{r - \mu}{\sigma}$$

compute the right hand side on historical data, and obtain a time series for the residual. We can then check that ϵ is independent and distributed according to $p(\epsilon)$. For instance, we can test that ϵ is uncorrelated, and that given a risk threshold α (say, 95%), the number of exceedance behaves as expected. However, when the horizon period ΔT increases, it becomes very difficult to perform back testing due to the lack of data. Alternatively, we can consider a process to model the returns with a time increment δt of one day, computing the forecasts using conditional averages. We can then relate daily data with forecasts at any time horizon, and the forecasts depend only on the process parameters, which are independent of ΔT and are consistent across risk horizon. The quality of the volatility forecasts is the major determinant factor for a risk methodology. The residuals can then be computed and their properties studied.

3.4.2 The structure of volatility models

The above heuristics being poor estimates of the future volatility, one must rely on proper volatility models such as the conditional heteroscedastic (CH) models. Since the early 80s, volatility clustering spawned a large literaturure on a new class of stochastic processes capturing the dependency of second moments in a phenomenological way. As the

lognormal assumption is not consistent with all the properties of historical stock returns, Engle [1982] first introduced the autoregressive conditional heteroscedasticity model (ARCH) which has been generalised to GARCH by Bollerslev [1986]. We let r_t be the log return of an asset at time t, and assume that $\{r_t\}$ is either serially uncorrelated or with minor lower order serial correlations, but it is dependent. Volatility models attempt at capturing such dependence in the return series. We consider the conditional mean and conditional variance of r_t given the filtration $\mathcal{F}_{t-1}$ defined by

$$\mu_t = E[r_t|\mathcal{F}_{t-1}] \, , \sigma_t^2 = Var(r_t|\mathcal{F}_{t-1}) = E[(r_t - \mu_t)^2|\mathcal{F}_{t-1}]$$

Since we assumed that the serial dependence of a stock return series was weak, if it exists at all, μ_t should be simple and we can assume that r_t follows a simple time series model such as a stationary $ARMA(p,q)$ model. That is

$$r_t = \mu_t + a_t \, , \mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{3.4.12}$$

where a_t is the shock or mean-corrected return of an asset return [4]. The model for μ_t is the mean equation for r_t, and the model for σ_t^2 is the volatility equation for r_t.

***Remark* 3.4.1** *Some authors use h_t to denote the conditional variance of r_t, in which case the shock becomes $a_t = \sqrt{h_t}\epsilon_t$.*

The paramereters p and q are non-negative integers, and the order (p,q) of the ARMA model may depend on the frequency of the return series. The excess kurtosis values, measuring deviation from the normality of the returns, are indicative of the long-tailed nature of the process. Hence, one can then compute and plot the autocorrelation functions for the returns process r_t as well as the autocorrelation functions for the squared returns r_t^2.

1. If the securities exhibit a significant positive autocorrelation at lag one and higher lags as well, then large (small) returns tend to be followed by large (small) returns of the same sign. That is, there are trends in the return series. This is evidence against the weakly efficient market hypothesis which asserts that all historical information is fully reflected in prices, implying that historical prices contain no information that could be used to earn a trading profit above that which could be attained with a naive buy-and-hold strategy which implies further that returns should be uncorrelated. In this case, the autocorrelation function would suggest that an autoregressive model should capture much of the behaviour of the returns.

2. The autocorrelation in the squared returns process would suggest that large (small) absolute returns tend to follow each other. That is, large (small) returns are followed by large (small) returns of unpredictable sign. It implies that the returns series exhibits volatility clustering where large (small) returns form clusters throughout the series. As a result, the variance of a return conditioned on past returns is a monotonic function of the past returns, and hence the conditional variance is heteroskedastic and should be properly modelled.

The conditional heteroscedastic (CH) models are capable of dealing with this conditional heteroskedasticity. The variance in the model described in Equation (3.4.12) becomes

$$\sigma_t^2 = Var(r_t|\mathcal{F}_{t-1}) = Var(a_t|\mathcal{F}_{t-1})$$

Since the way in which σ_t^2 evolves over time differentiate one volatility model from another, the CH models are concerned with the evolution of the volatility. Hence, modelling conditional heteroscdasticity (CH) amounts to augmenting a dynamic equation to a time series model to govern the time evolution of the conditional variance of the shock. We distinguish two types or groups of CH models, the first one using an exact function to govern the evolution of σ_t^2, and the second one using a stochastic equation to describe σ_t^2. For instance, the (G)ARCH model belongs to the former, and the stochastic volatility (SV) model belongs to the latter. In general, we estimate the conditional mean and variance equations jointly in empirical studies.

[4] since $a_t = r_t - \mu_t$

3.4.2.1 Benchmark volatility models

The ARCH model, which is the first model providing a systematic framework for volatility modelling, states that

1. the mean-corrected asset return a_t is serially uncorrelated, but dependent
2. the dependence of a_t can be described by a simple quadratic function of its lagged values.

Specifically, setting $\mu_t = 0$ for simplicity, an $ARCH(p)$ model assumes that

$$r_t = h_t^{\frac{1}{2}} \epsilon_t \, , \, h_t = \alpha_0 + \alpha_1 r_{t-1}^2 + ... + \alpha_p r_{t-p}^2$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean zero and variance 1, $\alpha_0 > 0$, and $\alpha_i \geq 0$ for $i > 0$. In practice, ϵ_t follows the standard normal or a standardised Student-t distribution. Generalising the ARCH model, the main idea behind (G)ARCH models is to consider asset returns as a mixture of normal distributions with the current variance being driven by a deterministic difference equation

$$r_t = h_t^{\frac{1}{2}} \epsilon_t \text{ with } \epsilon_t \sim N(0,1) \tag{3.4.13}$$

and

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \, , \, \alpha_0 > 0 \, , \, \alpha_i, \beta_j > 0 \tag{3.4.14}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$. The latter constraint on $\alpha_i + \beta_i$ implies that the unconditional variance of r_t is finite, whereas its conditional variance h_t evolves over time. In general, empirical applications find the $GARCH(1,1)$ model with $p = q = 1$ to be sufficient to model financial time series

$$h_t = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 h_{t-1} \, , \, \alpha_0 > 0 \, , \, \alpha_1, \beta_1 > 0 \tag{3.4.15}$$

When estimated, the sum of the parameters $\alpha_1 + \beta_1$ turns out to be close to the non-stationary case, that is, mostly only a constraint on the parameters prevents them from exceeding 1 in their sum, which would lead to non-stationary behaviour. Different extensions of GARCH were developed in the literature with the objective of better capturing the financial stylised facts. Among them are the Exponential GARCH (EGARCH) model proposed by Nelson [1991] accounting for asymmetric behaviour of returns, the Threshold GARCH (TGARCH) model of Rabemananjara et al. [1993] taking into account the leverage effects, the regime switching GARCH (RS-GARCH) developed by Cai [1994], and the Integrated GARCH (IGARCH) introduced by Engle et al allowing for capturing high persistence observed in returns time series. Nelson [1990] showed that Ito diffusion or jump-diffusion processes could be obtained as a continuous time limit of discrete GARCH sequences. In order to capture stochastic shocks to the variance process, Taylor [1986] introduced the class of stochastic volatility (SV) models whose instantaneous variance is driven by

$$\ln(h_t) = k + \phi \ln(h_{t-1}) + \tau \xi_t \text{ with } \xi_t \sim N(0,1) \tag{3.4.16}$$

This approach has been refined and extended in many ways. The SV process is more flexible than the GARCH model, providing more mixing due to the co-existence of shocks to volatility and return innovations. However, one drawback of the GARCH models and extension to Equation (3.4.16) is their implied exponential decay of the autocorrelations of measures of volatility which is in contrast to the very long autocorrelation discussed in Section (10.4). Both the GARCH and the baseline SV model are only characterised by short-term rather than long-term dependence. In order to capture long memory effects, the GARCH and SV models were expanded by allowing for an infinite number of lagged volatility terms instead of the limited number of lags present in Equations (3.4.14) and (3.4.16). To obtain a compact characterisation of the long memory feature, a fractional differencing operator was used in both extensions, leading to the fractionally integrated GARCH (FIGARCH) model introduced by Baillie et al. [1996], and the long-memory stochastic volatility model of Breidt et al. [1998]. As an intermediate approach, Dacorogna et al. [1998] proposed

the heterogeneous ARCH (HARCH) model, considering returns at different time aggregation levels as determinants of the dynamic law governing current volatility. In this model, we need to replace Equations (3.4.14) with

$$h_t = c_0 + \sum_{j=1}^{n} c_j r^2_{t,t-\Delta t_y}$$

where $r_{t,t-\Delta t_y} = \ln(p_t) - \ln(p_{t-\Delta t_j})$ are returns computed over different frequencies. This model was motivated by the finding that volatility on fine time scales can be explained to a larger extend by coarse-grained volatility than vice versa (see Muller et al. [1997]). Thus, the right hand side of the above equation covers local volatility at various lower frequencies than the time step of the underlying data ($\Delta t_j = 2, 3, ..$). Note, multifractal models have a closely related structure but model the hierarchy of volatility components in a multiplicative rather than additive format.

3.4.2.2 Some practical considerations

As an example, we are breifly discussing the (G)ARCH models, which are assumed to be serially uncorrelated. However, before these models are used one must remove the autocorrelation present in the returns process. One can remove the autocorrelation structure from the returns process by fitting Autoregressive models. For instance, we can consider the $AR(p)$ model

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + \phi_2(r_{t-2} - \mu) + ... + \phi_p(r_{t-p} - \mu) + u_t$$

where μ is the sample mean of the series $\{r_t\}$ and the error $\{u_t\}$ are assumed to be from an i.i.d. process. One can use the Yule-Walker method to estimate the model parameters. To discern between models we use the AIC criterion which while rewarding a model fitting well (large maximum likelihood) also penalises for the inclusion of too many parameter values, that is, overfitting. We then obtain models of order p with associated standard errors of the parameter values. Again one has to study the normality of the errors by computing kurtosis statistics as well as the autocorrelation function. One expect the residuals to be uncorrelated not contradicting the Gaussian white noise hypothesis. However, it may happen that the squares of the residuals are autocorrelated. That is, the $AR(p)$ model did not account for the volatility clustering present in the returns process. Even if the residuals from returns are uncorrelated, but still show some evidence of volatility clustering, we can attempt to model the residuals with (G)ARCH processes. Hence, we can remove the autocorrelation by modelling the first order moment with $AR(p)$ models, and we can model the second order moments using (G)ARCH models. (G)ARCH models can explain the features of small autocorrelations, positive and statistically significant autocorrelations in the squares and excess kurtosis present in the residuals of the AR model. For an adequate fit, the residuals from the (G)ARCH models should have white noise properties. Hence, we fit (G)ARCH models to the residuals of the AR models $\{u_t\}$ getting

$$\mu_t - \nu = h_t^{\frac{1}{2}} \epsilon_t$$

where $\{\epsilon_t\}$ is Gaussian with zero mean and constant variance. Letting ν be the mean, we get

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i (u_{t-i} - \nu)^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$$

For $q = 0$ the process reduces to the $ARCH(p)$ process and for $p = q = 0$ the process is a Gaussian white noise. In the $ARCH(p)$ process the conditional variance is specified as a linear function of past sample variances only, whereas the $GARCH(p, q)$ process allows lagged conditional variances to enter as well. Once the (G)ARCH models have been fitted, we can discern between (nested) competing models using the Likelihood Ratio Test. To discern between two competing models where neither is nested in the other, we resort to examining the residuals of the models. For an adequate model, the residuals should ressemble a white noise process being uncorrelated with constant variance. In addition, the autocorrelation of the squared residuals should also be zero. We should favour simpler models when two models appear adequate from an examination of the residuals.

3.4.3 Forecasting volatility with RiskMetrics methodology

3.4.3.1 The exponential weighted moving average

As volatility of financial markets changes over time, we saw that forecasting volatility could be of practical importance when computing the conditional variance of the log return of the undelying asset. The historical sample variance computed in Section (3.3.5) assigns weights of zero to squared deviations prior to a chosen cut-off date and an equal weight to observations after the cut-off date. In order to benefit from closing auction prices, one should consider close-to close volatility, with a large number of samples to get a good estimate of historical volatility. Letting D be the number of past trading days used to estimate the volatility, the daily log-return at time t is $r(t) = r_{t-1,t} = C(t) - C(t-1)$ where $C(t)$ is the closing log-price at the end of day t, and the annualised D-day variance of return is given by

$$\sigma^2_{STDEV}(t;D) = C_{sf}\frac{1}{D}\sum_{i=0}^{D-1}(r(t-i)-\overline{r}(t))^2$$

where the scalar C_{sf} scales the variance to be annual, and where $\overline{r}(t) = \frac{1}{D}\sum_{i=0}^{D-1} r(t-i)$ is the sample mean. Some authors set $C_{sf} = 252$ while others set it to $C_{sf} = 261$.

While for large sample sizes the standard close to close estimator is best, it can obscure short-term changes in volatility. That is, volatility reacts faster to shocks in the market as recent data carry more weight than data in the distant past. Further, following a shock (a large return), the volatility declines exponentially as the weight of the shock observation falls. As a result, the equally weighted moving average leads to relatively abrupt changes in the standard deviation once the shock falls out of the measurement sample, which can be several months after it occurs. On the other hand, the weighting scheme for GARCH(1,1) in Equation (3.4.15) and the Exponential Weighted Moving Average (EWMA) (see details in Section (5.7.1)) are such that weights decline exponentially in both models. Therefore, RiskMetrics (see Longerstaey et al. [1996]) let the ex-ante volatility σ_t be estimated with the exponentially weighted lagged squared daily returns (similar to a simple univariate GARCH model). It assigns the highest weight to the latest observations and the least to the oldest observations in the volatility estimate. The assignment of these weights enables volatility to react to large returns (jumps) in the market, so that following a jump the volatility declines exponentially as the weight of the jump falls. Specifically, the ex-ante annualised variance σ^2_t is calculated as follows:

$$\sigma^2_t(t;D) = C_{sf}\sum_{i=0}^{\infty}\omega_i(r_{t-1-i}-\overline{r}_t)^2$$

where the weights $\omega_i = (1-\delta)\delta^i$ add up to one, and where

$$\overline{r}_t = \sum_{i=0}^{\infty}\omega_i r(t-i)$$

is the exponentially weighted average return computed similarly. The parameter δ with $0 < \delta < 1$ is called the decay factor and determines the relative weights that are assigned to returns (observations) and the effective amount of data used in estimating volatility. It is chosen so that the center of mass of the weights $\sum_{i=0}^{\infty}(1-\delta)\delta^i i = \frac{\delta}{(1-\delta)}$ is equal to 30 or 60 days (since $\sum_{i=0}^{\infty}(1-\delta)\delta^i = 1$). The volatility model is the same for all assets at all times. To ensure no look-ahead bias contaminates our results, we use the volatility estimates at time $t-1$ applied to time t returns throughout the analysis, that is, $\sigma_t = \sigma_{t-1|t}$.

The formula above being an infinite sum, in practice we estimate the volatility σ with the Exponential Weighted Moving Average model (EWMA) for a given sequence of k returns as

$$\sigma^2_{EWMA}(t)(t;D) = C_{sf}\sum_{i=0}^{k-1}\omega_i(r_{t-i}-\overline{r})^2$$

The latest return has weight $(1-\delta)$ and the second latest $(1-\delta)\delta$ and so on. The oldest return appears with weight $(1-\delta)\delta^{k-1}$. The decay factor δ is chosen to minimise the error between observed volatility and its forecast over some sample period.

3.4.3.2 Forecasting volatility

One advantage of the exponentially weighted estimator is that it can be written in recursive form, allowing for volatility forecasts. If we assume the sample mean $\overline{r}$ is zero and that infinite amounts of data are available, then by using the recursive feature of the exponential weighted moving average (EWMA) estimator, the one-day variance forecast satisfies

$$\sigma^2_{1,t+1|t} = \delta\sigma^2_{1,t|t-1} + (1-\delta)r^2_{1,t}$$

where $\sigma^2_{1,t+1|t}$ denotes 1-day time $t+1$ forecast given information up to time t. It is derived as follow

$$\begin{aligned}
\sigma^2_{1,t+1|t} &= (1-\delta)\sum_{i=0}^{\infty}\delta^i r^2_{1,t-i} \\
&= (1-\delta)\big(r^2_{1,t} + \delta r^2_{1,t-1} + \delta^2 r^2_{1,t-2} + ...\big) \\
&= (1-\delta)r^2_{1,t} + \delta(1-\delta)\big(r^2_{1,t-1} + \delta r^2_{1,t-2} + \delta^2 r^2_{1,t-3} + ...\big) \\
&= \delta\sigma^2_{1,t|t-1} + (1-\delta)r^2_{1,t}
\end{aligned}$$

Taking the square root of both sides of the equation we get the one day volatility forecast $\sigma_{1,t+1|t}$. For two return series, the EWMA estimate of covariance for a given sequence of k returns is given by

$$\sigma^2_{1,2}(t;D) = C_{sf}\sum_{i=0}^{k-1}\omega_i(r_{1,t-i}-\overline{r}_1)(r_{2,t-i}-\overline{r}_2)$$

Similarly to the variance forecast, the covariance forecast can also be written in recursive form. The 1-day covariance forecast between two return series $r_{1,t}$ and $r_{2,t}$ is

$$\sigma^2_{12,t+1|t} = \delta\sigma^2_{12,t|t-1} + (1-\delta)r_{1,t}r_{2,t}$$

In order to get the correlation forecast, we apply the corresponding covariance and volatility forecast. The 1-day correlation forecast is given by

$$\rho_{12,t+1|t} = \frac{\sigma^2_{12,t+1|t}}{\sigma_{1,t+1|t}\sigma_{2,t+1|t}}$$

Using the EWMA model, we can also construct variance and covariance forecast over longer time horizons. The T-period forecasts of the variance and covariance are, respectively,

$$\sigma^2_{1,t+T|t} = T\sigma^2_{1,t+1|t}$$

and

$$\sigma^2_{12,t+T|t} = T\sigma^2_{12,t+1|t}$$

implying that the correlation forecasts remain unchanged irrespective of the forecast horizon, that is, $\rho_{t+T|t} = \rho_{t+1|t}$. We observe that in the EWMA model, multiple day forecasts are simple multiples of one-day forecasts. Note, the square root of time rule results from the assumption that variances are constant. However, the above derivation of volatilities and covariances vary with time. In fact, the EWMA model implicitly assume that the variance process is non-stationary. In the literature, this model is a special case the integrated GARCH model (IGARCH) described in Section (5.6.3). In practice, scaling up volatility estimates prove problematic when

- rates/prices are mean-reverting.
- boundaries limit the potential movements in rates and prices.
- estimates of volatilities optimised to forecast changes over a particular horizon are used for another horizon.

The cluster properties are measured by the lagged correlation of volatility, and the decay of that correlation quantifies the memory shape and magnitude. Based on statistical analysis, Zumbach [2006a] [2006b] found the lagged correlation of volatility to decay logarithmically as $1 - \frac{\log \Delta T}{\log \Delta T_0}$ in the range from 1 day to 1 year for all assets. That is, the memory of the volatility decays very slowly, so that a volatility model must capture its long memory. He considered a multi-scales long memory extention of IGARCH called the Long-Memory ARCH (LMARCH) process. The main idea being to measure the historical volatilities with a set of exponential moving averages (EMA) on a set of time horizons chosen according to a geometric series. The feed-back loop of the historical returns on the next random return is similar to a $GARCH(1,1)$ process, but the volatilities are measured at multiple time scales. Computing analytically the conditional expectations related to the volatility forecasts, we get

$$\sigma^2_{\Delta T|t} = \frac{\Delta T}{\delta t} \sum_i \lambda(\frac{\Delta T}{\delta t}, i) r^2_{t-i\delta t}$$

with weights $\lambda(\frac{\Delta T}{\delta t}, i)$ derived from the process equations and satisfying $\sum_i \lambda(\frac{\Delta T}{\delta t}, i) = 1$. We see that the leading term of the forecast is given by $\sigma_{\Delta T|t} \approx \sqrt{\frac{\Delta T}{\delta t}}$ which is the Normal square root scaling, and $\sum_i \lambda(i) r^2(i)$ is a measure of the past volatility constructed as a weighted sum of the past return square. We saw in Section (3.4.2) that more complicated nonlinear modelling of volatility exists such as GARCH, stochastic volatility, applications of chaotic dynamics etc. We will details all these models in Section (5.6).

3.4.3.3 Assuming zero-drift in volatility calculation

Given the logarithmic return in Equation (3.3.6) and the k-days period, the sample mean return $\overline{r} = \frac{1}{k} \sum_{i=0}^{k-1} r_{t-i}$ is the estimate of the mean μ. An important issue arising in the estimation of the historical variance is the noisy estimate of the mean return. This is due to the fact that the mean logarithmic return depends on the range (length) of the return series in the sense that

$$\overline{r} = \frac{1}{k} \sum_{i=0}^{k-1} r_{t-i} = \frac{1}{k} \sum_{i=0}^{k-1} (L_t - L_{t-i})$$

Thus, the mean return does not take into account the price movements or the number of prices within the period. Therefore, while a standard deviation measures the dispersion of the observations around its mean, in practice, it may be difficult to obtain a good estimate of the mean. As a result, some authors suggested to measure volatility around zero rather than the sample mean. Assuming a EWMA model to estimate volatiliy, Longerstaey et al. [1995a] proposed to study the difference between results given by the sample mean and zero-mean centred estimators. They considered the one-day volatility forecast referred to as the estimated mean estimator, and given by

$$\hat{\sigma}^2_t = \delta \hat{\sigma}^2_{t-1} + \delta(1-\delta)(R_t - \overline{R}_{t-1})^2$$

where R_t is the percentage change return and $\overline{R}_{t-1}$ is an exponentially weighted mean. The zero-men estimator is derived in Section (3.4.3.2) is given by

$$\tilde{\sigma}_t^2 = (1-\delta)R_t^2 + \delta\tilde{\sigma}_{t-1}^2$$

Setting up a Monte Carlo experiment, they studied the forecast difference of the two models $\hat{\sigma}_t^2$ and $\tilde{\sigma}_t^2$ at any time t. Defining the arithmetic diffenence Δ_t as

$$\Delta_t = \tilde{\sigma}_t^2 - \hat{\sigma}_t^2|_{\overline{R}_i=0} \text{ for } i = t, t-1, ...$$

we get

$$\Delta_t = R_t^2(1-\delta)^2$$

so that the one-day volatility forecast for $\delta = 0.94$ becomes $\delta_t = 0.0036R_t^2$. Assuming zero sample mean, for sufficiently small percentage return R_t, the difference Δ_t is negligible, and one should not expect significant differences between the two models. Considering a database consisting of eleven time series, Longerstaey et al. [1995a] concluded that the relative differences between the two models are quite small. Further, investigating the differences of the one-day forecasted correlation between 1990 and 1995, they found very small deviations. They extended the analysis of the difference between the estimated mean and zero-mean estimators to one month horizons and obtained relatively small differences between the two estimates. As a result, the zero-mean estimator is a viable alternative to the estimated mean estimator, which is simpler to compute and not sensitive to short-term trends.

Note, assuming a conditional zero mean of returns is consistent with the financial theory of the efficient market hypothesis (EMH). We will see in Chapter (10) that financial markets are multifractal and that conditional mean of returns experience long term trends. In fact, Zumbach [2006a] [2006b] showed that neglecting the mean return forecast was not correct, particularly for interest rates and stock indexes. For the former, the yields can follow a downward or upward trend for very long periods, of the order of a year or more, and the latter can follow an overall upward trend related to interest rates. These long trends introduce correlations, equivalent to some predictability in the rates themselves. Even though these effects are quantitatively small, they introduce clear deviations from the random walk with ARCH effect on the volatility.

3.4.3.4 Estimating the decay factor

We now need to estimate the sample mean and the exponential decay factor δ. The largest sample size available should be used to reduce the standard error. Choosing a suitable decay factor is a necessity in forecasting volatility and correlations. One important issue in this estimation is the determination of an effective number of days (k) used in forecasting. It is postulated in RiskMetrics that the volatility model should be determined by using the metric

$$\Omega_k = (1-\delta)\sum_{t=k}^{\infty}\delta^t$$

where Ω_k is set to the tolerance level α. We can now solve for k. Expanding the summation we get

$$\alpha = \delta^k(1-\delta)[1+\delta+\delta^2+...]$$

and taking the natural logarithms on both sides we get

$$\ln\alpha = k\ln\delta + \ln(1-\delta) + \ln[1+\delta+\delta^2+...]$$

Since $\log(1\pm x) \approx \pm x - \frac{1}{2}x^2$ for $|x| < 1$ we get

$$k \approx \frac{\ln\alpha}{\ln\delta}$$

In principle, we can find a set of optimal decay factors, one for each covariance can be determined such that the estimated covariance matrix is symmetric and positive definite. RiskMetrics presented a method for choosing one optimal decay factor to be used in estimation of the entire covariance matrix. They found $\delta = 0.94$ to be the optimal decay factor for one-day forecast and $\delta = 0.97$ to be optimal for one month (25 trading days) forecast.

3.4.4 Computing historical volatility

Computing historical volatility is not an easy task as it depends on two parameters, the length of time and the frequency of measurement. As volatility mean revert over a period of months, it is difficult to define the best period of time to obtain a fair value of realised volatility. Further, in presence of rare events, the best estimate of future volatility is not necessary the current historical volatility. While historical volatility can be measured monthly, quarterly or yearly, it is usually measured daily or weekly. In presence of independent stock price returns, then daily and weekly historical volatility should on average be the same. However, when stock price returns are not independent there is a difference due to autocorrelation. If we assume that daily volatility should be preferable to weekly volatility because there are five times as many data points available then intraday volatility should always be preferred. However, intraday volatility is not constant as it is usually greatest just after the market open and just before the market close and falling in the middle of the day, leading to noisy time series. Hence, traders taking into account intraday prices should depend on advanced volatility measures. One approach is to use exponential weighted moving average (EWMA) model described in Section (3.4.3) which avoids volatility collapse of historic volatility. It has the advantage over standard historical volatility (STDEV) to gradually reduce the effect of a spike on volatility. However, EWMA models are rarely used, partly due to the fact that they do not properly handle regular volatility driving events such as earnings. That is, previous earnings jumps will have least weight just before an earnings date (when future volatility is most likely to be high), and most weight just after earnings (when future volatility is most likely to be low).

The availability of intraday data allows one to consider volatility estimators making use of intraday information for more efficient volatility estimates. Using the theory of quadratic variation, Andersen et al. [1998] and Barndorff-Nielsen et al. [2002] introduced the concept of integrated variance and showed that the sum of squared high-frequency intraday log-returns is an efficient estimator of daily variance in the absence of price jumps and serial correlation in the return series. However, market effects such as lack of continuous trading, bid/ask spread, price discretisation, swamp the estimation procedure, and in the limit, microstructure noise dominates the result. The research on high-frequency volatility estimation and the effects of microstructure noise being extremely active, we will just say that intervals between 5 to 30 minutes tend to give satisfactory volatility estimates. We let $N_{day}(t)$ denote the number of active price quotes during the trading day t, so that $S_1(t), .., S_{N_{day}}(t)$ denotes the intraday quotes. Assuming 30 minutes quotes, the realised variance (RV) model is

$$\hat{\sigma}^2_{RV}(t) = \sum_{j=2}^{N_{day}} \left(\log S_j(t) - \log S_j(t-1)\right)^2$$

One simple heuristic to define advanced volatility measures making use of intraday information is to consider range estimators that use some or all of the open (O), high (H), low (L) and close (C). In that setting we define

- opening price: $O(t) = \log S_1(t)$
- closing price: $C(t) = \log S_{N_{day}}(t)$
- high price: $H(t) = \log\left(\max_{j=1,..,N_{day}} S_j(t)\right)$
- low price: $L(t) = \log\left(\min_{j=1,..,N_{day}} S_j(t)\right)$
- normalised opening price: $o(t) = O(t) - C(t-1) = \log \frac{S_1(t)}{S_{N_{day}}(t-1)}$

- normalised closing price: $c(t) = C(t) - O(t) = \log \frac{S_{N_{day}}(t)}{S_1(t)}$
- normalised high price: $h(t) = H(t) - O(t) = \log\left(\max_{j=1,..,N_{day}} \frac{S_j(t)}{S_1(t)}\right)$
- normalised low price: $l(t) = L(t) - O(t) = \log\left(\min_{j=1,..,N_{day}} \frac{S_j(t)}{S_1(t)}\right)$

We are going to discuss a few of them below, introduced by Parkinson [1980], Garman et al. [1980], Rogers et al. [1994], Yang et al. [2000]. We refer the readers to Baltas et al. [2012b] and Bennett [2012] for a more detailed list with formulas.

- close to close (C): the most common type of calculation benefiting only from reliable prices at closing auctions.
- Parkinson (HL): it is the first to propose the use of intraday high and low prices to estimate daily volatility

$$\hat{\sigma}^2_{PK}(t) = \frac{1}{4\log 2}(h(t) - l(t))^2$$

  The model assumes that the asset price follows a driftless diffusion process and it is about 5 times more efficient than STDEV.

- Garman-Klass (OHLC): it is an extention of the PK model including opening and closing prices. It is the most powerful estimate for stocks with Brownian motion, zero drift and no opening jumps. The GK estimator is given by

$$\hat{\sigma}^2_{GK}(t) = \frac{1}{2}(h(t) - l(t))^2 - (2\log 2 - 1)c^2(t)$$

  and it is about 7.4 times more efficient than STDEV.

- Rogers-Satchell (OHLC): being similar to the GK estimate, it benefits from handling non-zero drift in the price process. However, opening jumps are not well handled. The estimator is given by

$$\hat{\sigma}^2_{RS}(t) = h(t)(h(t) - c(t)) + l(t)(l(t) - c(t))$$

  Rogers-Satchell showed that GK is just 1.2 times more efficient than RS.

- Garman-Klass Yang-Zhang extension (OHLC): Yang-Zhang extended the GK method by incorporating the difference between the current opening log-price and the previous day's closing log-price. The estimator becomes robust to the opening jumps, but still assumes zero drift. The estimator is given by

$$\hat{\sigma}_{GKYZ} = \sigma^2_{GK} + (O(t) - C(t-1))^2$$

- Yang-Zhang (OHLC): Having a multi-period specification, it is an unbiased volatility estimator that is independent of both the opening jump and the drift of the price process. It is the most powerful volatility estimator with minimum estimation error. It is a linear combination of the RS estimator, the standard deviation of past daily log-returns (STDEV), and a similar estimator using the normalised opening prices instead of the close-to-close log-returns.

Since these estimators provide daily estimates of variance/volatility, an annualised D-day estimator is therefore given by the average estimate over the past D days

$$\sigma^2_l(t; D) = \frac{C_{sf}}{D}\sum_{i=0}^{D-1} \hat{\sigma}^2_l(t-i)$$

for $l = \{RV, PK, GK, GKYZ, RS\}$. The Yang-Zhang estimator is given by

$$\sigma^2_{YZ}(t:D) = \sigma^2_{open}(t:D) + k\sigma^2_{STDEV}(t:D) + (1-k)\sigma^2_{RS}(t;D)$$

where $\sigma^2_{open}(t;D) = \frac{C_{sf}}{D}\sum_{i=0}^{D-1}(o(t-i) - \overline{o}(t))^2$. The parameter k is chosen so that the variance of the estimator is minimised. YZ showed that for the value

$$k = \frac{0.34}{1.34 + \frac{D+1}{D-1}}$$

their estimator is $1 + \frac{1}{k}$ times more efficient than the ordinary STDEV estimator. Brandt et al. [2005] showed that the range-based volatility estimates are approximately Gaussian, whereas return-based volatility estimates are far from Gaussian, which is is an advantage when calibrating stochastic volatility models with likelihood procedure. Bennett [2012] defined two measures to determine the quality of a volatility measure, namely, the efficiency and the bias. The former is defined by

$$(\sigma^2_e) = \frac{\sigma^2_{STDEV}}{\sigma^2_l}$$

where σ_l is the volatility of the estimate and σ_{STDEV} is the volatility of the standard close to close estimate. It describes the volatility of the estimate, and decreases as the number of samples increases. The latter is the difference between the estimated variance and the average volatility. It depends on the sample size and the type of distribution of the underlying. Generally, for small sample sizes the Yang-Zhang estimator is best overall, and for large sample sizes the standard close to close estimator is best. Setting aside the realised variance, the Yang-Zhang estimator is the most efficient estimator, it exhibits the smallest bias when compared to the realised variance, and it generates the lowest turnover. While the optimal choice for volatility estimation is the realised variance (RV) estimator, the Yang-Zhang estimator constitute an optimal tradeoff between efficiency, turnover, and the necessity of high frequency data, as it only requires daily information on opening, closing, high and low prices.

Part II

Statistical tools applied to finance

Chapter 4

Filtering and smoothing techniques

4.1 Presenting the challenge

4.1.1 Describing the problem

Dynamical systems are characterised by two types of noise, where the first one is called observational or additive noise, and the second one is called dynamical noise. In the former, the system is unaffected by this noise, instead the noise is a measurement problem. The observer has trouble precisely measuring the output of the system, leading to recorded values with added noise increment. This additive noise is external to the process. In the latter, the system interprets the noisy output as an input, leading to dynamical noise because the noise invades the system. Dynamical noise being inherent to financial time series, we are now going to summarise some of the tools proposed in statistical analysis and signal processing to filter it out.

In financial time series analysis, the trend is the component containing the global change, while the local changes are represented by noise. In general, the trend is characterised by a smooth function representing long-term movement. Hence, trends should exhibit slow changes, while noise is assumed to be highly volatile. Trend filtering (TF) attempts at differentiating meaningful information from exogenous noise. The separation between trend and noise lies at the core of modern statistics and time series analysis. TF is generally used to analyse the past by transforming any noisy signal into a smoother one. It can also be used as a predictive tool, but it can not be performed on any time series. For instance, trend following predictions suppose that the last observed trend influences future returns, but the trend may not persist in the future.

A physical process can be described either in the time domain, by the values of some quantity h as a function of time $h(t)$, or in the frequency domain where the process is specified by giving its amplitude H as a function of frequency f, that is, $H(f)$ with $-\infty < f < \infty$. One can think of $h(t)$ and $H(f)$ as being two different representations of the same function, and the Fourier transform equations are a tool to go back and forth between these two representations. There are several reasons to filter digitally a signal, such as applying a high-pass or low-pass filtering to eliminate noise at low or high frequencies, or requiring a bandpass filter if the interesting part of the signal lies only in a certain frequency band. One can either filter data in the frequency domain or in the time domain. While it is very convenient to filter data in the former, in the case where we have a real-time application the latter may be more appropriate.

In the time domain, the main idea behind TF is to replace each data point by some kind of local average of surrounding data points such that averaging reduce the level of noise without biasing too much the value obtained. Observations can be averaged using many different types of weightings, some trend following methods are referred to as linear filtering, while others are classified as nonlinear. Depending on whether trend filtering is performed to

explain past behaviour of asset prices or to forecast future returns, one will consider different estimator and calibration techniques. In the former, the model and parameters can be selected by minimising past prediction error or by considering a benchmark estimator and calibrating another model to be as close as possible to the benchmark. In the later, trend following predictions assume that positive (or negative) trends are more likely to be followed by positive (or negative) returns. That is, trend filtering solve the problem of denoising while taking into account the dynamics of the underlying process.

Bruder et al. [2011] tested the persistence of trends for major financial indices on a period ranging from January 1995 till October 2011 where the average one-month returns for each index is separated into a set including one-month returns immediately following a positive three-month return and another set for negative three-month returns. The results showed that on average, higher returns can be expected after a positive three-month return than after a negative three-month period so that observation of the current trend may have a predictive value for the indices under consideration. Note, on other time scales or for other assets, one may obtain opposite results supporting contrarian strategies.

The main goal of trend filtering in finance is to design portfolio strategies benefiting from these trends. However, before computing an estimate of the trend, one must decide if there is a trend or not in the series. We discussed in Section (3.3.3) the power of statistical tests for trend detection, and concluded that the Mann-Kendall test was more powerful than the parametric t-test for high coefficient of skewness. Given an established trend, one approach is to use the resulting trend indicator to forecast future asset returns for a given horizon and allocate accordingly the portfolio. For instance, an investor could buy assets with positive return forecasts and sell them when the forecasts are negative. The size of each long or short position is a quantitative problem requiring a clear investment process. As explained in Section (), the portfolio allocation should take into account the individual risks, their correlations and the expected return of each asset.

4.1.2 Regression smoothing

A regression curve describes a general relationship between an explanatory variable X, which may be a vector in $\mathbb{R}^d$, and a response variable Y, and the knowledge of this relation is of great interest. Given n data points, a regression curve fitting a relationship between variables $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ is commonly modelled as

$$Y_i = m(X_i) + \epsilon_i \, , \, i = 1, .., n$$

where ϵ is a random variable denoting the variation of Y around $m(X)$, the mean regression curve $E[Y|X = x]$ when we try to approximate the mean response function m. By reducing the observational errors, we can concentrate on important details of the mean dependence of Y on X. This curve approximation is called smoothing. Approximating the mean function can be done in two ways. On one hand the parametric approach assume that the mean curve m has some prespecified functional form (for example a line with unknown slope and intercept). On the other hand we try to estimate m nonparametrically without reference to a specific form. In the former, the functional form is fully described by a finite set of parameters, which is not the case in the latter, offering more flexibility for analysing unknown regression relationship. It can be used to predict observations without referencing to a fixed parametric model, to find spurious observations by studying the influence of isolated points, and to substitute missing values or interpolate between adjacent points. As an example, Engle et al. [1986] considered a nonlinear relationship between electricity sales and temperature using a parametric-nonparametric estimation approach. The prediction of new observations is of particular interest to time series analysis. In general, classical parametric models are too restrictive to give reasonable explanations of observed phenomena. For instance, Ullah [1987] applied kernel smoothing to a time series of stock market prices and estimated certain risk indexes. Deaton [1988] used smoothing methods to examine demand patterns in Thailand and investigated how the knowledge of those patterns affects the assessment of pricing policies.

Smoothing of a data set $\{(X_i, Y_i)\}_{i=1}^n$ involves the approximation of the mean response curve $m(x)$ which should

be any representative point close to the point x. This local averaging procedure, which is the basic idea of smoothing, can be defined as

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_{ni}(x) Y_i$$

where $\{W_{ni}(x)\}_{i=1}^n$ denotes a sequence of weights which may depend on the whole vector $\{X_i\}_{i=1}^n$. The amount of averaging is controlled by the weight sequence which is tuned by a smoothing parameter regulating the size of the neighborhood around x. In the special case where the weights $\{W_{ni}(x)\}_{i=1}^n$ are positive and sum to one for all x, that is,

$$n^{-1} \sum_{i=1}^{n} W_{ni}(x) = 1$$

then $\hat{m}(x)$ is a least squares estimate (LSE) at point x since it is the solution of the minimisation problem

$$\min_{\theta} n^{-1} \sum_{i=1}^{n} W_{ni}(x)(Y_i - \theta)^2 = n^{-1} \sum_{i=1}^{n} W_{ni}(x)(Y_i - \hat{m}(x))^2$$

where the residuals are weighted quadratically. Thus, the basic idea of local averaging is equivalent to the procedure of finding a local weighted least squares estimate. In the random design model, we let $\{(X_i, Y_i)\}_{i=1}^n$ be independent, identically distributed variables, and we concentrate on the average dependence of Y on $X = x$, that is, we try to estimate the conditional mean curve

$$m(x) = E[Y|X = x] = \frac{\int y f(x, y) dy}{f(x)}$$

where $f(x, y)$ is the joint density of (X, Y), and $f(x) = \int f(x, y) dy$ is the marginal density of X. Note that for a normal joint distribution with mean zero, the regression curve is linear and $m(x) = \rho x$ with $\rho = Corr(X, Y)$.

By contrast, the fixed design model is concerned with controlled, non-stochastic X-variables, so that

$$Y_i = m(X_i) + \epsilon_i \, , 1 \leq i \leq n$$

where $\{\epsilon_i\}_{i=1}^n$ denotes zero-mean random variables with variance σ^2. Although the stochastic mechanism is different, the basic idea of smoothing is the same for both random and nonrandom X-variables.

4.1.3 Introducing trend filtering

4.1.3.1 Filtering in frequency

We consider the removal of noise from a corrupted signal by assuming that we want to measure the uncorrupted signal $u(t)$, but that the measurement process is imperfect, leading to the corrupted signal $c(t)$. On one hand the true signal $u(t)$ may be convolved with some known response function $r(t)$ to give a smeared signal $s(t)$

$$s(t) = \int_{-\infty}^{\infty} r(t - \tau) u(\tau) d\tau \text{ or } S(f) = R(f)U(f)$$

where S, R, U are the Fourier transforms of s, r, u respectively. On the other hand the measured signal $c(t)$ may contain an additional component of noise $n(t)$ (dynamical noise)

$$c(t) = s(t) + n(t) \tag{4.1.1}$$

While in the first case we can devide $C(f)$ by $R(f)$ to get a deconvolved signal, in presence of noise we need to find the optimal filter $\phi(t)$ or $\Phi(f)$ producing the signal $\tilde{u}(t)$ or $\tilde{U}(f)$ as close as possible to the uncorrupted signal $u(t)$ or $U(f)$. That is, we want to estimate

$$\tilde{U}(f) = \frac{C(f)\Phi(f)}{R(f)}$$

in the least-square sense, such that

$$\int_{-\infty}^{\infty} |\tilde{u}(t) - u(t)|^2 dt = \int_{-\infty}^{\infty} |\tilde{U}(f) - U(f)|^2 df$$

is minimised. In the frequency domain we get

$$\int_{-\infty}^{\infty} |\frac{(S(f) + N(f))\Phi(f)}{R(f)} - \frac{S(f)}{R(f)}|^2 df = \int_{-\infty}^{\infty} \frac{|S(f)|^2|1 - \Phi(f)|^2 + |N(f)|^2|\Phi(f)|^2}{|R(f)|^2} df$$

if S and N are uncorrelated. We get a minimum if and only if the integrand is minimised with respect to $\Phi(f)$ at every value of f. Differentiating with respect to Φ and setting the result to zero, we get

$$\Phi(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2} \tag{4.1.2}$$

involving the smearded signal S and the noise N but not the true signal U. Since we can not estimate separately S and N from C we need extra information or assumption. One way forward is to sample a long stretch of data $c(t)$ and plot its power spectral density as it is proportional to $|S(f)|^2 + |N(f)|^2$

$$|S(f)|^2 + |N(f)|^2 \approx P_c(f) = |C(f)|^2 \, , 0 \leq f < f_c$$

which is the modulus squared of the discrete Fourier transform of some finite sample. In general, the resulting plot shows the spectral signature of a signal sticking up above a continuous noise spectrum. Drawing a smooth curve through the signal plus noise power, the difference between the two curves is the smooth model of the signal power. After designing a filter with response $\Phi(f)$ and using it to make a respectable guess at the signal $\tilde{U}(f)$ we might regard $\tilde{U}(f)$ as a new signal to improve even further with the same filtering technique. However, the scheme converges to a signal of $S(f) = 0$. Alternatively, we take the whole data record, FFT it, multiply the FFT output by a filter function $\mathcal{H}(f)$ (constructed in the frequency domain), and then do an inverse FFT to get back a filtered data set in the time domain.

4.1.3.2 Filtering in the time domain

As discussed above, even though it is very convenient to filter data in the frequency domain, in the case where we have a real-time application the time domain may be more appropriate. A general linear filter takes a sequence y_k of input points and produce a sequence x_n of output points by the formula

$$x_n = \sum_{k=0}^{M} c_k y_{n-k} + \sum_{j=1}^{N} d_j x_{n-j}$$

where the $M + 1$ coefficients c_k and the N coefficients d_j are fixed and define the filter response. This filter produces each new output value from the current and M previous input values, and from its own N previous output values. In the case where $N = 0$ the filter is called nonrecursive or finite impulse response (FIR), and if $N \neq 0$ it is called recursive or infinite impulse response (IIR). The relation between the c_k's and d_j's and the filter response function $\mathcal{H}(f)$ is

$$\mathcal{H}(f) = \frac{\sum_{k=0}^{M} c_k e^{-2\pi i k(f\Delta)}}{1 - \sum_{j=1}^{N} d_j e^{-2\pi i j(f\Delta)}}$$

where Δ is the sampling interval and $f\Delta$ is the Nyquist interval. To determine a filter we need to find a suitable set of c's and d's from a desired $\mathcal{H}(f)$, but like many inverse problem it has no all-purpose solution since the filter is a continuous function while the short list of the c's and d's represents only a few adjustable parameters. When the denominator in the filter $\mathcal{H}(f)$ is unity, we recover a discrete Fourier transform. Nonrecursive filters have a frequency response that is a polynomial in the variable $\frac{1}{z}$ where

$$z = e^{2\pi i(f\Delta)}$$

while the recursive filter's frequency response is a rational function in $\frac{1}{z}$. However, nonrecursive filters are always stable but recursive filters are not necessarily stable. Hence, the problem of designing recursive filters is an inverse problem with an additional stability constraint. See Press et al. [1992] for a sketch of basic techniques.

4.2 Smooting techniques and nonparametric regression

As opposed to pure parametric curve estimations, smoothing techniques provide flexibility in data analysis. In this section, we are going to consider the statistical aspects of nonparametric regression smoothing by considering the choice of smoothing parameters and the construction of confidence bands. While various smoothing methods exist, all smoothing methods are in an asymptotic sense equivalent to kernel smoothing.

4.2.1 Histogram

The density function f tells us where observations cluster and occur more frequently. Nonparametric approach does not restrict the possible form of the density function by assuming f to belong to a prespecified family of functions. The estimation of the unknown density function f provides a way of understanding and representing the behaviour of a random variable.

4.2.1.1 Definition of the Histogram

Histogram combines neighbouring needles by counting how many fall into a small interval of length h called a bin. The probability for observations of x to fall into the interval $[-\frac{h}{2}, \frac{h}{2})$ equals the shaded area under the density

$$P(X \in [-\frac{h}{2}, \frac{h}{2})) = \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x)dx \tag{4.2.3}$$

Histogram as a frequency counting curve Histogram counts the relative frequency of observations falling into a prescribed mesh and normalised so that the resulting function is a density. The relative frequency of observations in this interval is a good estimate of the probability in Equation (4.2.3), which we divide by the number of observations to get

$$P(X \in [-\frac{h}{2}, \frac{h}{2})) \simeq \frac{1}{n}\#(X_i \in [-\frac{h}{2}, \frac{h}{2}))$$

where n is the sample size. Applying the mean value theorem to Equation (4.2.3), we obtain

$$\int_{-\frac{h}{2}}^{\frac{h}{2}} f(x)dx = f(\xi)h \,, \xi \in [-\frac{h}{2}, \frac{h}{2})$$

so that

$$P(X \in [-\frac{h}{2}, \frac{h}{2})) = \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x)dx = f(\xi)h$$
$$P(X \in [-\frac{h}{2}, \frac{h}{2})) \simeq \frac{1}{n}\#(X_i \in [-\frac{h}{2}, \frac{h}{2}))$$

Equating the two equations, we arrive at the density estimate

$$\hat{f}_h(x) = \frac{1}{nh}\#(X_i \in [-\frac{h}{2}, \frac{h}{2}))\ , x \in [-\frac{h}{2}, \frac{h}{2})$$

The calculation of the histogram is characterised by the following two steps

1. divide the real line into bins
$$B_j = [x_0 + (j-1)h, x_0 + jh)\ , j \in z$$
2. count how many data fall into each bin

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} \sum_j I_{\{X_i \in B_j\}} I_{\{x \in B_j\}}$$

The histogram as a Maximum Likelihood Estimate We want to find a density $\hat{f}$ maximising the likelihood in the observations

$$\prod_{i=1}^{n} \hat{f}(X_i)$$

However, this task is ill-posed, and we must restrict the class of densities in order to obtain a well-defined solution.

Varying the binwidth By varying the binwidth h we can get different shapes for the density $\hat{f}_h(x)$

- $h \to 0$: needleplot, very noisy representation of the data
- $h =$: smoother less noisy density estimate
- $h \to \infty$: box-shaped, overly smooth

We now need to choose the binwidth h in practice. It can be done by working out the statistics of the histogram.

Statistics of the histogram We need to find out if the estimate $\hat{f}_h$ is unbiased, and if it matches on average the unknown density. If $x \in B_j = [(j-1), jh)$, the histogram for x (i.i.d.) is given by

$$\hat{f}_h = (nh)^{-1} \sum_{i=1}^{n} I_{\{X_i \in B_j\}}$$

where n is the sample size.

Bias of the histogram Since the X_i are identically distributed, the expected value of the estimate is given by

$$E[\hat{f}_h(x)] = (nh)^{-1} \sum_{i=1}^{n} E[I_{\{X_i \in B_j\}}] = \frac{1}{h} \int_{(j-1)h}^{jh} f(u)du$$

We define the bias as

$$Bias(\hat{f}_h(x)) = E[\hat{f}_h(x)] - f(x) \tag{4.2.4}$$

which becomes in our setting

$$Bias(\hat{f}_h(x)) = \frac{1}{h} \int_{B_j} f(u)du - f(x)$$

Rewriting the bias of the histogram in terms of the binwidth h and the density f, we get

$$Bias(\hat{f}_h(x)) = ((j - \frac{1}{2})h - x)f^{'}((j - \frac{1}{2})h) + o(h) \text{ , } h \to 0$$

The stability of the estimate is measured by the variance which we are now going to calculate.

Variance of the histogram If the index function $I_{\{X_i \in B_j\}}$ is a Bernoulli variable, the variance is given by

$$\begin{aligned} Var(\hat{f}_h(x)) &= Var((nh)^{-1} \sum_{i=1}^{n} I_{B_j}(X_j)) \\ &= (nh)^{-1} f(x) + o((nh)^{-1}) \text{ , } nh \to \infty \end{aligned}$$

Clearly, the variance decreases when nh increases, so that we have a dilema between the bias and the variance. Hence, we are going to consider the Mean Square Error (MSE) as a measure of accuracy of an estimator.

The Mean Squared Error of the histogram The Mean Squared Error of the estimate is given by

$$\begin{aligned} MSE(\hat{f}_h(x)) &= E[(\hat{f}_h(x) - f(x))^2] \\ &= Var(\hat{f}_h(x)) + Bias^2(\hat{f}_h(x)) \\ &= \frac{1}{nh} f(x) + ((j - \frac{1}{2})h - x)^2 f^{'}((j - \frac{1}{2})h)^2 + o(h) + o(\frac{1}{nh}) \end{aligned} \tag{4.2.5}$$

The minimisation of the MSE with respect to h denies a compromise between the problem of oversmoothing (if we choose a large binwidth h to reduce the variance) and undersmoothing (for the reduction of the bias by decreasing the binwidth h). We can conclude that the histogram $\hat{f}_h(x)$ is a consistent estimator for $f(x)$ since

$$h \to 0 \text{ , } nh \to \infty \text{ implies } MSE(\hat{f}_h(x)) \to 0$$

and

$$\hat{f}_h(x) \xrightarrow{p} f(x)$$

However, the application of the MSE formula is difficult in practice, since it contains the unkown density function f both in the variance and the squared bias.

$$MSE = \text{ estimate in one particular point } x$$

The Mean Integrated Squared Error of the histogram Another measure we can consider is the Mean Integrated Squared Error (MISE). It is a measure of the goodness of fit for the whole histogram defined as

$$MISE(\hat{f}_h) = \int_{-\infty}^{\infty} MSE(\hat{f}_h(x))dx$$

measuring the average MSE deviation. The speed of convergence of the MISE in the asymptotic sense is given by

$$MISE(\hat{f}_h) = (nh)^{-1} + \frac{h^2}{12}\|f^{'}\|_2^2 + o(h^2) + o(\frac{1}{nh})$$

where $\|f(x)\|^2 = \int_{-\infty}^{\infty} |f(x)|^2 dx$. The leading term of the MISE is the asymptotic MISE given by

$$A - MISE(\hat{f}_h) = (nh)^{-1} + \frac{h^2}{12}\|f^{'}\|_2^2$$

We can minimise the A-MISE by differentiating it with respect to h and equating the result to zero

$$\frac{\partial}{\partial h} A - MISE(\hat{f}_h) = 0$$

getting

$$h_0 = \Big(\frac{6}{n\|\hat{f}^{'}\|_2^2}\Big)^{\frac{1}{3}} \tag{4.2.6}$$

Therefore, choosing theoretically

$$h_0 \sim n^{-\frac{1}{3}}$$

we obtain

$$MISE \sim n^{-\frac{1}{3}} >> n^{-1}\ , \text{ for n sufficiently large}$$

The calculation of h_0 requires the knowledge of the unkown parameter $\|f^{'}\|_2^2$. One approach is to use the plug-in method, which mean that we take an estimate of $\|f^{'}\|_2^2$ and plug it into the asymptotic formula in Equation (4.2.6). However, it is difficult to estimate this functional form. A practical solution is for $\|f^{'}\|_2^2$ to take a reference distribution and to calculate h_0 by using this reference distribution.

4.2.1.2 Smoothing the histogram by WARPing

One of the main criticism of the histogram is its dependence on the choice of the origin. This is because if we use one of these histograms for density estimation, the choice is arbitrary. To get rid of this shortcoming, we can average these histograms, which is known in a generalised form as Weighted Averaging of Rounded Points (WARPing). It is based on a smaller bin mesh, by discretising the data first into a finite grid of bins and then smoothing the binned data

$$B_{j,l} = \big[(j - 1 + \frac{l}{M})h, (j + \frac{l}{M})h\big)\ , l \in 0, ..., M-1$$

The bins $B_{j,l}$ are generated by shifting each B_j by the amount of $\frac{lh}{M}$ to the right. We now have M histograms based on these shifted bins

$$\hat{f}_{h,l}(x) = (nh)^{-1} \sum_{i=1}^{n} \big(\sum_j I_{\{x \in B_{j,l}\}} I_{\{X_i \in B_{j,l}\}}\big)$$

The WARPing idea is to calculate the average over all these histograms

$$\hat{f}_h(x) = M^{-1} \sum_{l=0}^{M-1} (nh)^{-1} \sum_{i=1}^{n} \Big(\sum_j I_{\{x \in B_{j,l}\}} I_{\{X_i \in B_{j,l}\}} \Big)$$

We can inspect the double sum over the two index functions more closely by assuming for a moment that x, and X_i are given and fixed, In that case, we get

$$\begin{aligned} & x \in \big[(j-1+\frac{l}{M})h, (j-1+\frac{l+1}{M})h\big) = B^*_{j,l} \\ & X_i \in \big[(j-1+\frac{l+K}{M})h, (j-1+\frac{l+K+1}{M})h\big) = B^*_{j,l+K} \\ & \text{for } j, K \in z \text{ , } l \in 0, ..., M-1 \end{aligned}$$

where $B^*_{j,l}$ is generated by dividing each bin B_j of binwidth h into M subbins of binwidth $\frac{h}{M} = \delta$. Hence,we get

$$\sum_{l=0}^{M-1} \sum_j I_{B_{j,l}}(X_i) I_{B_{j,l}}(x) = \sum_z I_{B^*_z}(x) \sum_{K=1-M}^{M-1} I_{B^*_{z+K}}(X_i)(M - |K|)$$

where $B^*_z = [\frac{z}{M}h, \frac{z+1}{M}h)$. This leads to the WARPed histogram

$$\begin{aligned} \hat{f}_h(x) &= (nMh)^{-1} \sum_{i=0}^{n} \sum_j I_{B^*_j}(x) \sum_{K=1-M}^{M-1} I_{B^*_{j+K}}(X_i)(M - |K|) \\ &= (nh)^{-1} \sum_j I_{B^*_j}(x) \sum_{K=1-M}^{M-1} W_M(K) n_{j+K} \end{aligned}$$

where $n_K = \sum_{i=1}^{n} I_{B^*_K}(X_i)$ and $W_M(K) = 1 - \frac{|K|}{M}$. The explicit specification of the weighting function $W_M(\bullet)$ allows us to approximate a bigger class of density estimators such as the Kernel.

4.2.2 Kernel density estimation

Rosenblatt [1956] proposed putting smooth Kernel weights in each of the observations. That is, around each observation X_i a Kernel function $K_h(\bullet - X_i)$ is centred. Like histogram, instead of averaging needles, one averages Kernel functions, and we have a smoothing parameter, called the bandwidth h, regulating the degree of smoothness for Kernel smoothers.

4.2.2.1 Definition of the Kernel estimate

A general Kernel K is a real function defined as

$$K_h = \frac{1}{h} K(\frac{x}{h})$$

Averaging over these Kernel functions in the observations leads to the Kernel density estimator

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^{n} nK(\frac{x - X_i}{h})$$

See Hardle [1991] for a short survey of some other Kernel functions. Some of the properties of the Kernel functions are

- Kernel functions are symmetric around 0 and integrate to 1
- since the Kernel is a density function, the Kernel estimate is a density too

$$\int K(x)dx = 1 \text{ implies } \int \hat{f}_h(x)dx = 1$$

Varying the Kernel The property of smoothness of the Kernel K is inherited by the corresponding estimate $\hat{f}_h(x)$

- the Uniform Kernel is not continuous in -1 and 1, so that $\hat{f}_h(x)$ is discountinuous and not differentiable in $X_i - h$ and $X_i + h$.
- the Triangle Kernel is continuous but not differentiable in -1, 0, and 1, so that $\hat{f}_h(x)$ is continuous but not differentiable in $X_i - h$, X_i, and $X_i + h$.
- the Quartic Kernel is continuous and differentiable everywhere.

Hence, we can approximate f by using different Kernels which gives qualitatively different estimates $\hat{f}_h$.

Varying the bandwidth We consider the bandwidth h in Kernel smoothing such that

- $h \to 0$: needleplot, very noisy representation of the data
- h small : a smoother, less noisy density estimate
- h big : a very smooth density estimate
- $h \to \infty$: a very flat estimate of roughly the shape of the chosen Kernel

4.2.2.2 Statistics of the Kernel density

Kernel density estimates are based on two parameters

1. the bandwidth h
2. the Kernel density function

We now provide some guidelines on how to choose h and K in order to obtain a good estimate with respect to a given goodness of fit criterion. We are interested in the extent of the uncertainty or at what speed the convergence of the smoother actually happens. In general, the extent of this uncertainty is expressed in terms of the sampling variance of the estimator, but in nonparametric smoothing situation it is not enough as there is also a bias to consider. Hence, we should consider the pointwise mean squared error (MSE), the sum of variance and squared bias. A variety of distance measures exist, both uniform and pointwise, but we will only describe the mean squared error and the mean integrated squared error.

Bias of the Kernel density We first check the asymptotic unbiasedness of $\hat{f}_h(x)$ using Equation (4.2.4). The expected value of the estimate is

$$E[\hat{f}_h(x)] = \frac{1}{n}\sum_{i=1}^{n} E[K_h(x - X_i)]$$

with the property

$$\text{if } h \to 0 \text{ , } E[\hat{f}_h(x)] \to f(x)$$

The estimate is thus asymptotically unbiased, when the bandwidth h converges to zero. We can look at bias analysis by using a Taylor expansion of $f(x+sh)$ in x, getting

$$Bias(\hat{f}_h(x)) = \frac{h^2}{2} f^{''}(x)\mu_2(K) + o(h^2) \text{ , } h \to 0$$

where $\mu_2(K) = \int_{-\infty}^{\infty} u^2 K(u)du$. The bias is quadratic in h. Hence, we have to choose h small enough to reduce the bias. The size of the bias depends on the curvature of f in x, that is, on the absolute value of $f^{''}(x)$.

Variance of the Kernel density We compute the variance of the Kernel density estimation to get insight into the stability of such estimates

$$\begin{aligned} Var(\hat{f}_h(x)) &= n^{-2} Var(\sum_{i=1}^{n} nK_h(x - X_i)) \\ &= (nh)^{-1}\|K\|_2^2 f(x) + o((nh)^{-1}) \text{ , } nh \to \infty \end{aligned}$$

The variance being proportional to $(nh)^{-1}$, we want to choose h large. This contradicts the aim of decreasing the bias by decreasing the bandwidth h. Therefore we must consider h as a compromise of both effects, namely the $MSE(\hat{f}_h(x))$ or $MISE(\hat{f}_h(x))$.

The Mean Squared Error of the Kernel density We are now looking at the Mean Squared Error (MSE) of the Kernel density defined in Equation (4.2.5)

$$MSE(\hat{f}_h(x)) = \frac{1}{nh} f(x)\|K\|_2^2 + \frac{h^4}{4}(f^{''}(x)\mu_2(K))^2 + o((nh)^{-1}) + o(h^4) \text{ , } h \to 0 \text{ , } nh \to \infty$$

The MSE converges to zero if $h \to 0$, $nh \to \infty$. Thus, the Kernel density estimate is consistent

$$\hat{f}_h(x) \xrightarrow{p} f(x)$$

We define the optimal bandwidth h_0 for estimating $f(x)$ by

$$h_0 = \arg\min_h MSE(\hat{f}_h(x))$$

We can rewrite the MSE as

$$MSE(\hat{f}_h(x)) = (nh)^{-1}c_1 + \frac{1}{4}h^4 c_2$$

with

$$\begin{aligned} c_1 &= f(x)\|K\|_2^2 \\ c_2 &= (f^{''}(x))^2(\mu_2(K))^2 \end{aligned}$$

Setting the derivative $\frac{\partial}{\partial h} MSE(\hat{f}_h(x))$ to zero, yields the optimum h_0 as

$$h_0 = (\frac{c_1}{c_2 n})^{\frac{1}{5}} = (\frac{f(x)\|K\|_2^2}{(f''(x))^2(\mu_2(K))^2 n})^{\frac{1}{5}}$$

Thus, the optimal rate of convergence of the MSE is given by

$$MSE(\hat{f}_{h_0}(x)) = \frac{5}{4}\big(\frac{f(x)\|K\|_2^2}{n}(f^{''}(x)\mu_2(K))^2\big)^{\frac{1}{5}}$$

Again, the formula includea the unknown functions $f(\bullet)$ and $f^{''}(\bullet)$.

The Mean Integrated Squared Error of the Kernel density We are now looking at the Mean Integrated Squared Error (MISE) of the Kernel density

$$MISE(\hat{f}_h(x)) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\mu_2(K))^2\|f^{''}\|_2^2 + o((nh)^{-1}) + o(h^4)\ ,\ h \to 0\ ,\ nh \to \infty$$

The optimal bandwidth h_0 which minimise the A-MISE with respect to the parameter h is given by

$$h_0 = \big(\frac{\|K\|_2^2}{\|f^{''}\|_2^2(\mu_2(K))^2 n}\big)^{\frac{1}{5}}$$

The optimal rate rate of convergence of the MISE is given by

$$A - MISE(\hat{f}_{h_0}(x)) = \frac{5}{4}(\|K\|_2^2)^{\frac{4}{5}}\big(\mu_2(K)\|f^{''}\|_2^2\big)^{\frac{2}{5}}n^{-\frac{4}{5}}$$

We have not escaped from the circulus virtuosis of estimating f, by encountering the knowledge of a function of f, here $f^{''}$. Fortunately, there exists ways of computing good bandwidths h, even if we have no knowledge of f. A comparison of the speed of convergence of MISE for histogram and Kernel density estimation is given by Hardle. The speed of convergence is faster for the Kernel density than for the histogram one.

4.2.2.3 Confidence intervals and confidence bands

To obtain confidence intervals, we derive the asymptotic distribution of the kernel smoothers and use either their asymptotic quantiles or bootstrap approximations for these quantiles. The estimate $\hat{f}_h(x)$ is asymptotically normally distributed as n increases and the bandwidth h decreases in the order of $n^{-\frac{1}{2}}$. Using the bias and the variance of the Kernel, we can derive the following theorem.

Theorem 4.2.1 *Suppose that $f^{''}(x)$ exists and $h_n = cn^{-\frac{1}{5}}$. Then the Kernel density estimate $\hat{f}_h(x)$ is asymptotically normally distributed.*

$$n^{\frac{2}{5}}(\hat{f}_{h_n}(x) - f(x)) \to N\big(\frac{c^2}{2}f^{''}(x)\mu_2(K), c^{-1}f(x)\|K\|_2^2\big)\ ,\ n \to \infty$$

This theorem enables us to compute a confidence interval for $f(x)$. An asymptotical $(1-a)$ confidence interval for $f(x)$ is given by

$$\Big[\hat{f}_h(x) - n^{-\frac{2}{5}}\big(\frac{c^2}{2}f^{''}(x)\mu_2(K) + d_a\big), \hat{f}_h(x) - n^{-\frac{2}{5}}\big(\frac{c^2}{2}f^{''}(x)\mu_2(K) - d_a\big)\Big]$$

with $d_a = u_{1-\frac{a}{2}}\sqrt{c^{-1}f(x)\|K\|_2^2}$ and $u_{1-\frac{a}{2}}$ is the $(1-\frac{a}{2})$ quantile of a standard normal distribution. Again, it includes the functions $f(x)$ and $f^{''}(x)$. A more practical way is to replace $f(x)$ and $f^{''}(x)$ with estimates $\hat{f}_h(x)$ and $\hat{f}_g^{''}(x)$ orcorresponding values of reference distributions. The estimate $\hat{f}_g^{''}(x)$ can be defined as $[\hat{f}_g(x)]^{''}$, where the bandwidth g is not the same as h. If we use a value of two as quantile for an asymptotic 95% confidence interval

$$\Big[\hat{f}_h(x) - n^{-\frac{2}{5}}\big(\frac{c^2}{2}\hat{f}_g^{''}(x)\mu_2(K) + 2\sqrt{c^{-1}f(x)\|K\|_2^2}\big), \hat{f}_h(x) - n^{-\frac{2}{5}}\big(\frac{c^2}{2}\hat{f}_g^{''}(x)\mu_2(K) - 2\sqrt{c^{-1}f(x)\|K\|_2^2}\big)\Big]$$

this technique yields a 95% confidence interval for $f(x)$ and not for the whole function. In order to get a confidence band for the whole function Bicket et al. suggested choosing a smaller bandwidth than of order $n^{-\frac{1}{5}}$ to reduce the bias, such that the limiting distribution of the estimate has an expectation equal to $f(x)$.

4.2.3 Bandwidth selection in practice

The choice of the bandwidth h is the main problem of the Kernel density estimation. So far we have derived formulas for optimal bandwidths that minimise the MSE or MISE, but employ the unknown functions $f(\bullet)$ and $f^{''}(\bullet)$. We are now considering how to obtain a reasonable choice of h when we do not know $f(\bullet)$.

4.2.3.1 Kernel estimation using reference distribution

We can adopt the technique using reference distribution for the choice of the binwidth of the histogram described above. We try to estimate $\|f^{''}\|_2^2$, assuming f to belong to a prespecified class of density functions. For example, we can choose the normal distribution with parameter μ and σ

$$\begin{aligned} \|f^{''}\|_2^2 &= \sigma^{-5} \int (Q^{''}(x))^2 dx \\ &= \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0.212\sigma^{-5} \end{aligned}$$

We can then estimate $\|f^{''}\|_2^2$ through an estimator $\hat{\sigma}$ for σ. For instance, if we take the Gaussian Kernel we obtain the following rule of thumb

$$\begin{aligned} \hat{h}_0 &= \left(\frac{\|Q\|_2^2}{\|\hat{f}^{''}\|_2^2 \mu_2^2(Q) n}\right)^{\frac{1}{5}} \\ &= \frac{4\hat{\sigma}^5}{3n} \approx 1.06\hat{\sigma} n^{-\frac{1}{5}} \end{aligned}$$

Note, we can use instead of $\hat{\sigma}$ a more robust estimate for the scale parameter of the distribution. For example, we can take the interquartile range $\hat{R}$ defined as

$$\hat{R} = X_{[0.75n]} - X_{[0.25n]}$$

The rule of thumb is then modified to

$$\hat{h}_0 = 0.79\hat{R} n^{-\frac{1}{5}}$$

We can combine both rules to get the better rule

$$\hat{h}_0 = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}}$$

since for Gaussian data $\hat{R} \approx 1.34\hat{\sigma}$.

4.2.3.2 Plug-in methods

Another approach is to directly estimate $\|f^{''}\|_2^2$, but doing so we move the problem from the estimation of f to the estimation of $f^{''}$. The plug-in procedure is based on the asymptotic expansion of the squared error kernel smoothers.

4.2.3.3 Cross-validation

Maximum likelihood cross-validation We want to test for a specific h the hypothesis

$$\hat{f}_h(x) = f(x) \text{ vs. } \hat{f}_h(x) \neq f(x)$$

The likelihood ratio test would be based on the rest statistic $\frac{f(x)}{\hat{f}_h(x)}$ and should be close to 1, or the average over X, $E_X[\log(\frac{f}{\hat{f}_h})(X)]$ should be 0. Thus, a good bandwidth minimising this measure of accuracy is in effect optimising the Kullback-Leibler information

$$d_{KL}(f, \hat{f}_h) = \int \frac{f}{\hat{f}_h}(x) f(x) dx$$

We are not able to compute $d_{KL}(f, \hat{f}_h)$ from the data, since it requires the knowledge of f. However, from a theoretical point of view, we can investigate this distance for the choice of an appropriate bandwidth h minimising $d_{KL}(\hat{f}, f)$. If we are given additional observations X_i, the likelihood for these observations $\prod_i \hat{f}_h(X_i)$ for different h would indicate which value of h is preferable, since the logarithm of this statistic is close to $d_{KL}(\hat{f}_h, f)$. In the case where we do not have additional observations, we can base the estimate $\hat{f}_h$ on the subset $\{X_j\}_{j \neq i}$ and to calculate the likelihood for X_i. Denote the Leave-One-Out estimate by

$$\hat{f}_{h,i}(X_i) = (n-1)^{-1} h^{-1} \sum_{j \neq i} K(\frac{X_i - X_j}{h})$$

The likelihood is

$$\prod_{i=1}^{n} \hat{f}_{h,i}(X_i) = (n-1)^{-n} h^{-n} \prod_{i=1}^{n} \sum_{j \neq i} K(\frac{X_i - X_j}{h})$$

We take the logarithm of this statistic normalised with the factor n^{-1} to get the maximum likelihood CV

$$CV_{KL}(h) = n^{-1} \sum_{i=1}^{n} \log[\hat{f}_{h,i}(X_i)] + n^{-1} \sum_{i=1}^{n} \log[\sum_{j \neq i} K(\frac{X_i - X_j}{h})] - \log[(n-1)h]$$

so that

$$\hat{h}_{KL} = \arg\max CV_{KL}(h)$$

and

$$E[CV_{KL}] \approx -E[d_{KL}(f, \hat{f}_h)] + \int \log[f(x)] f(x) dx$$

For more details see Hall [1982].

Least-squares cross-validation We consider an alternative distance measure between $\hat{f}$ and f called the Integrated Squared Error (SSE) defined as

$$d_I(h) = \int (\hat{f}_h - f)^2(x) dx$$

which is a quadratic measure of accuracy. Hence, we get

$$d_I(h) - \int f^2(x) dx = \int \hat{f}_h^2(x) dx - 2 \int (\hat{f}_h f)(x) dx$$

and

$$\int (\hat{f}_h f)(x)dx = E_X[\hat{f}_h(x)]$$

The leave-one-out estimate is

$$E_X[\hat{f}_h(x)] = n^{-1}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i)$$

It determines a good bandwidth h minimising the right hand side of the above equation using the leave-one-out estimate. This leads to the least-squares cross-validation

$$CV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i)$$

The bandwidth minimising this function is

$$\hat{h}_{CV} = \arg\min CV(h)$$

Scott et al. [1987] call the function CV an unbiased cross-validation criterion, since

$$E[CV(h)] = MISE(\hat{f}_h) - \|f\|_2^2$$

defines a sequence of bandwidths $\hat{h}_n = h(X_1, .., X_n)$ to be asymptotically optimal if

$$\frac{d_I(h_n)}{\inf_{h\geq 0} d_I(h)} \to 1\ ,\ n \to \infty$$

If the density f is bounded, then $\hat{h}_{CV}$ is asymptotically optimal. Note, the minimisation of $CV(h)$ is independent from the order of differentiability p of f. So, this technique is more general to apply than the plig-in method, which requires f to be exactly of the same order of differentiability p. For the computation of the score function, note that

$$\int \hat{f}_h^2(x)dx = n^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}K * K(\frac{X_j - X_i}{h})$$

where $K * K(u)$ is the convolution of the Kernel function K. As a result, we get

$$\begin{aligned} CV(h) &= n^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}K * K(\frac{X_j - X_i}{h}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,i}(X_i) \\ &= \frac{2}{n^2 h}\Big[\frac{n}{2}K * K(0) + \sum_{i=1}^{n}\sum_{j=1}^{n}K * K(\frac{X_j - X_i}{h}) - \frac{2}{n-1}K(\frac{X_j - X_i}{h})\Big] \end{aligned}$$

Biased cross-validation The biased cross-validation introduced by Scott et al. [1987] is based on the idea of a direct estimate of $A - MISE(\hat{f}_h)$ given by

$$A - MISE(\hat{f}_h) = (nh)^{-1}\|K\|_2^2 + \frac{h^4}{4}\mu_2^2(K)\|f^{''}\|_2^2$$

where we have to estimate $\|f^{''}\|_2^2$. So, we get

$$BCV_1(h) = (nh)^{-1}\|K\|_2^2 + \frac{h^4}{4}\mu_2^2(K)\|\hat{f}^{''}\|_2^2$$

The minimisation of the $MISE(\hat{f}_h)$ requires a sequence of bandwidths proportional to $n^{-\frac{1}{5}}$. We can use a bandwidth of this order for the optimisation of $BCV_1(h)$

$$Var(\hat{f}_h^{''}(x)) = Var(h^{-3}\sum_{i=1}^{n} K^{''}(\frac{x - X_i}{h})) \sim n^{-1}h^{-5}\|K^{''}\|_2^2$$

Hence, the variance of $\hat{f}_h^{''}$ does not converge to zero for this choice of $h \sim n^{-\frac{1}{5}}$ so that the $BCV_1(h)$ can not approximate the $MISE(\hat{f}_h)$. This is because the same bandwidth h is used for the estimation of $\|f^{''}\|_2^2$ and that of f. Hence, we have to employ different bandwidths. For this bias in the estimation of the L_2 norm, the method is called biased CV. We have a formula for the expectation of $\|\hat{f}^{''}\|_2^2$ given by Scott et al. [1987]

$$E[\|\hat{f}_h^{''}\|_2^2] = \|f^{''}\|_2^2 + \frac{1}{nh^5}\|K^{''}\|_2^2 + o(h^2)$$

Therefore, we correct the above bias by

$$\|\hat{f}^{''}\|_2^2 = \|\hat{f}_h^{''}\|_2^2 - \frac{1}{nh^5}\|K^{''}\|_2^2$$

It is asymptotically unbiased when we let $h \sim n^{-\frac{1}{5}} \to 0$. The biased cross-validation is given by

$$BCV(h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\mu_2^2(K)\|\hat{f}_h^{''}\|_2^2 - \frac{1}{nh^5}\|K^{''}\|_2^2$$

where

$$\hat{h}_{BVC} = \arg\min BCV(h)$$

is an estimate for the optimal bandwidth $\hat{h}_0$ minimising $d_I(h)$. Scott et al. [1987] showed that $\hat{h}_{BVC}$ is asymptotically optimal. The optimal bandwidth $\hat{h}_{BVC}$ has a smaller standard deviation than $\hat{h}_{CV}$ and hence, gives satisfying results for the estimation of the A-MISE. On the other hand, for some skewed distributions biased cross-validation tends to oversmooth where $CV(h)$ is still quite close to the A-MISE optimal bandwidth.

4.2.4 Nonparametric regression

A regression curve fitting a relationship between variables $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, where the former is the explanatory variable and the latter is the response variable, is commonly modelled as

$$Y_i = m(X_i) + \epsilon_i \ , \ i = 1, .., n$$

where ϵ is a random variable denoting the variation of Y around $m(X)$, the mean regression curve $E[Y|X = x]$ when we try to approximate the mean response function m. By reducing the observational errors, we can concentrate on important details of the mean dependence of Y on X. This curve approximation is called smoothing. Approximating the mean function can be done in two ways. On one hand the parametric approach assume that the mean curve m has some prespecified functional form (a line with unknown slope and intercept). On the other hand we try to estimate m nonparametrically without reference to a specific form. In the former, the functional form is fully described by a finite set of parameters, which is not the case in the latter offering more flexibility for analysing unknown regression relationship. For regression curve fitting we are interested in weighting the response variable Y in a certain neighbourhood of x. Hence, we weight the observations Y_i depending on the distance of X_i to x using the estimator

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_n(x; X_1, .., X_n) Y_i$$

Since in general most of the weight $W_n(x; X_1, .., X_n)$ is given to the observation X_i, we can abbreviate it to $W_{ni}(x)$, so that

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_{ni}(x) Y_i$$

where $\{W_{ni}(x)\}_{i=1}^n$ denotes a sequence of weights which may depend on the whole vector $\{X_i\}_{i=1}^n$. We call smoother the regression estimator $\hat{m}_h(x)$, and smooth the outcome of the smoothing procedure. We consider the random design model, where the X-variables have been randomly generated, and we let $\{(X_i, Y_i)\}_{i=1}^n$ be independent, identically distributed variables. We concentrate on the average dependence of Y on $X = x$, that is, we try to estimate the conditional mean curve

$$m(x) = E[Y|X = x] = \frac{\int y f(x, y) dy}{f(x)}$$

where $f(x, y)$ is the joint density of (X, Y), and $f(x) = \int f(x, y) dy$ is the marginal density of X. Following Hardle [1990], we now present some common choice for the weights $W_{ni}(\bullet)$.

4.2.4.1 The Nadaraya-Watson estimator

We want to find an estimate of the conditional expectation $m(x) = E[Y|X = x]$ where

$$m(x) = \frac{\int y f(x, y) dy}{\int f(x, y) dy} = \int y f(y|x) dy$$

since $f(x) = \int f(x, y) dy$ and $f(y|x) = \frac{f(x,y)}{f(x)}$ is the conditional density of Y given $X = x$. While various smoothing methods exist, all smoothing methods are in asymptotic sense equivalent to kernel smoothing. We therefore choose the Kernel density K to represent the weight sequence $\{W_{ni}(x)\}_{i=1}^n$. We saw in Section (4.2.2) how to estimate the denominator using the Kernel density estimate. For the numerator, we could estimate the joint density $f(x, y)$ by using the multiplicative Kernel

$$\hat{f}_{h_1,h_2} = n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) K_{h_2}(x - X_i)$$

We can work out an estimate of the numerator as

$$\int y \hat{f}_{h_1,h_2}(x, y) dy = n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) Y_i$$

Hence, employing the same bandwidth h for both estimates, we can estimate the conditional expectation $m(x)$ by combining the estimates of the numerator and the denominator. This method was proposed by Nadaraya [1964] and Watson [1964] and gave the Nadaraya-Watson estimator

$$\hat{m}_h(x) = \frac{n^{-1} \sum_{i=1}^n K_h(x - X_i) Y_i}{n^{-1} \sum_{j=1}^n K_h(x - X_j)}$$

In terms of the general nonparametric regression curve estimate, the weights have the form

$$W_{hi}(x) = \frac{h^{-1}K(\frac{x-X_i}{h})}{\hat{f}_h(x)}$$

where the shape of the Kernel weights is determined by K, and the size of the weights is parametrised by h.

A variety of Kernel functions exist, but both practical and theoretical considerations limit the choice. A commonly used Kernel function is of the parabolic shape with support $[-1, 1]$ (Epanechnikov)

$$K(u) = 0.75(1 - u^2)I_{\{|u|\leq 1\}}$$

but it is not differentiable at $u = \pm 1$. The Kernel smoother is not defined for a bandwidth with $\hat{f}_h(x) = 0$. If such $\frac{0}{0}$ case occurs, one defines $\hat{m}_h(x)$ as being zero. Assuming that the kernel estimator is only evaluated at the observations $\{X_i\}_{i=1}^n$, then as $h \to 0$, we get

$$\hat{m}_h(X_i) \to \frac{K(0)Y_i}{K(0)} = Y_i$$

so that small bandwidths reproduce the data. In the case where $h \to \infty$, and K has support $[-1, 1]$, then $K(\frac{x-X_i}{h}) \to K(0)$, and

$$\hat{m}_h(x) \to \frac{n^{-1}\sum_{i=1}^n K(0)Y_i}{n^{-1}\sum_{i=1}^n K(0)} = n^{-1}\sum_{i=1}^n Y_i$$

resulting in an oversmooth curve, the average of the response variables.

Statistics of the Nadaraya-Watson estimator The numerator and denominator of this statistic are both random variables so that their analysis is done separately. We first define

$$r(x) = \int yf(x, y)dy = m(x)f(x) \tag{4.2.7}$$

The estimate is

$$\hat{r}_h(x) = n^{-1}\sum_{i=1}^n K_h(x - X_i)Y_i$$

The regression curve estimate is thus given by

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)}$$

We already analised the properties of $\hat{f}_h(x)$, and can work out the expectation and variance of $\hat{r}_h(x)$

$$\begin{aligned} E[\hat{r}_h(x)] &= E[n^{-1}\sum_{i=1}^n K_h(x - X_i)Y_i] \\ &= \int K_h(x - u)r(u)du \end{aligned}$$

Similarly to the density estimation with Kernels, we get

$$E[\hat{r}_h(x)] = r(x) + \frac{h^2}{2}r''(x)\mu_2(K) + o(h^2)\,, h \to 0$$

To compute the variance of $\hat{r}_h(x)$ we let $s^2(x) = E[Y^2|X = x]$, so that

$$\begin{aligned} Var(\hat{r}_h(x)) &= Var(n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i) = n^{-2}\sum_{i=1}^{n} Var(K_h(x - X_i)Y_i) \\ &= n^{-1}\{\int K_h^2(x-u)s^2(u)f(u)du - (\int K_h(x-u)r(u)du)^2\} \\ &\approx n^{-1}h^{-1}\int K^2(u)s^2(x+uh)f(x+uh)du \end{aligned}$$

Using the techniques of splitting up integrals, the variance is asymptotically given by

$$Var(\hat{r}_h(x)) = n^{-1}h^{-1}f(x)s^2(x)\|K\|_2^2 + o((nh)^{-1}) \, , nh \to \infty$$

and the variance tends to zero as $nh \to \infty$. Thus, the MSE is given by

$$MSE(\hat{r}_h(x)) = \frac{1}{nh}f(x)s^2(x)\|K\|_2^2 + \frac{h^4}{4}(r''(x)\mu_2(K))^2 + o(h^4) + o((nh)^{-1}) \, , h \to 0 \, , nh \to \infty \tag{4.2.8}$$

Hence, if we let $h \to 0$ such that $nh \to \infty$, we have

$$MSE(\hat{r}_h(x)) \to 0$$

so that the estimate is consistent

$$\hat{r}_h(x) \xrightarrow{p} m(x)f(x) = r(x)$$

The denominator of $\hat{m}_h(x)$, the Kernel density estimate $\hat{f}_h(x)$, is also consistent for the same asymptotics of h. Hence, using Slutzky's theorem (see Schonfeld [1969]) we obtain

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)} \xrightarrow{p} \frac{r(x)}{f(x)} = \frac{m(x)f(x)}{f(x)} = m(x) \, , h \to 0 \, , nh \to \infty$$

and $\hat{m}_h(x)$ is a consistent estimate of the regression curve $m(x)$, if $h \to 0$ and $nh \to \infty$. In order to get more insight into how $\hat{m}_h(x)$ behaves, such as its speed of convergence, we can study the mean squared error

$$d_M(x,h) = E[(\hat{m}_h(x) - m(x))^2]$$

at a point x.

Theorem 4.2.2 *Assume the fixed design model with a one-dimensional predictor variable X, and define*

$$c_K = \int K^2(u)du \, , d_K = \int u^2K(u)du$$

Further, assume K has support $[-1,1]$ with $K(-1) = K(1) = 0$, $m \in \mathcal{C}^2$, $\max_i |X_i - X_{i-1}| = o(n^{-1})$, and $var(\epsilon_i) = \sigma^2$ for $i = 1,..,n$. Then

$$d_M(x,h) \approx (nh)^{-1}\sigma^2 c_K + \frac{h^4}{4}d_K^2(m''(x))^2 \, , h \to 0 \, , nh \to \infty$$

which says that the bias, as a function of h, is increasing whereas the variaiance is decreasing. To understand this result, we note that the estimator $\hat{m}_h(x)$ is a ratio of random variables, such that the central limit theorem can not directly be applied. Thus, we linearise the estimator as follows

$$\begin{aligned}\hat{m}_h(x) - m(x) &= \big(\frac{\hat{r}_h(x)}{\hat{f}_h(x)} - m(x)\big)\big(\frac{\hat{f}_h(x)}{f(x)} + (1 - \frac{\hat{f}_h(x)}{f(x)})\big) \\ &= \frac{\hat{r}_h(x) - m(x)\hat{f}_h(x)}{f(x)} + (\hat{m}_h(x) - m(x))\frac{f(x) - \hat{f}_h(x)}{f(x)}\end{aligned}$$

By the above consistency property of $\hat{m}_h(x)$ we can choose $h \sim n^{-\frac{1}{5}}$. Using this bandwidth we can state

$$\begin{aligned}\hat{r}_h(x) - m(x)\hat{f}_h(x) &= (\hat{r}_h(x) - r(x)) - m(x)(\hat{f}_h(x) - f(x)) \\ &= o_p(n^{-\frac{2}{5}}) + m(x)o_p(n^{-\frac{2}{5}}) \\ &= o_p(n^{-\frac{2}{5}})\end{aligned}$$

such that

$$\begin{aligned}(\hat{m}_h(x) - m(x))(f(x) - \hat{f}_h(x)) &= o_p(1)o_p(n^{-\frac{2}{5}}) \\ &= o_p(n^{-\frac{2}{5}})\end{aligned}$$

The leading term in the distribution of $\hat{m}_h(x) - m(x)$ is

$$(f(x))^{-1}(\hat{r}_h(x) - m(x)\hat{f}_h(x))$$

and the MSE of this leading term is

$$(f(x))^{-2}E[(\hat{r}_h(x) - m(x)\hat{f}_h(x))^2]$$

leading to the approximate mean squared error

$$MSE(\hat{m}_h(x)) = \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\|K\|_2^2 + \frac{h^4}{4}\big(m^{''}(x) + 2\frac{m^{'}(x)f^{'}(x)}{f(x)}\big)^2\mu_2^2(K) + o((nh)^{-1}) + o(h^4)\ , h \to 0\ , nh \to \infty$$

The MSE is of order $o(n^{-\frac{4}{5}})$ when we choose $h \sim n^{-\frac{1}{5}}$. The second summand corresponds to the squared bias of $\hat{m}_h(x)$ and is either dominated by the second derivative $m^{''}(x)$ when we are near to a local extremum of $m(x)$, or by the first derivative $m^{'}(x)$ when we are near to a deflection point of $m(x)$.

Confidence intervals The asymptotic confidence intervals for $m(x)$ is computed using the formulas of the asymptotic variance and bias of $\hat{m}_h(x)$. The asymptotic distribution of $\hat{m}_h(x)$ is given by the following theorem

Theorem 4.2.3 *The Nadaraya-Watson Kernel smoother $\hat{m}_h(x_j)$ at the K different locations $x_1, .., x_K$ converges in distribution to a multivariate normal random vector with mean B and identity covariance matrix*

$$\Big\{(nh)^{\frac{1}{2}}\frac{\hat{m}_h(x_j) - m(x_j)}{(\frac{\sigma^2(x_j)\|K\|_2^2}{f(x_j)})^{\frac{1}{2}}}\Big\}_{j=1}^{K} \to N(B, I)$$

where

$$B = \left\{ \mu_2(K) \left[m^{''}(x_j) + 2 \frac{m^{'}(x_j) f^{'}(x_j)}{f(x_j)} \right] \right\}_{j=1}^{K}$$

We can use this theorem to compute an asymptotic $(1-a)$ confidence interval for $m(x)$, when we employ estimates of the unknown functions $\sigma(x)$, $f(x)$, $f^{'}(x)$, $m(x)$, and $m^{'}(x)$. One way forward is to assume that the bias of $\hat{m}_h(x)$ is of negligible size compared with the variance, so that B is set equal to the zero vector, leaving only $\sigma(x)$ and $f(x)$ to be estimated. Note, $f(x)$ can be estimated with the Kernel density estimator $\hat{f}_h(x)$, and the conditional variance $\sigma^2(x)$ can be defined as

$$\hat{\sigma}^2(x) = n^{-1} \sum_{i=1}^{n} W_{hi}(x)(Y_i - \hat{m}_h(x))^2$$

We then compute the interval $[clo, cup]$ around $\hat{m}_h(x)$ at the K distinct points $x_1, .., x_k$ with

$$\begin{aligned} clo &= \hat{m}_h(x) - c_a c_K^{\frac{1}{2}} \frac{\hat{\sigma}(x)}{(nh\hat{f}_h(x))^{\frac{1}{2}}} \\ cup &= \hat{m}_h(x) + c_a c_K^{\frac{1}{2}} \frac{\hat{\sigma}(x)}{(nh\hat{f}_h(x))^{\frac{1}{2}}} \end{aligned}$$

If the bandwidth is $h \sim n^{-\frac{1}{5}}$, then the computed interval does not lead asymptotically to an exact confidence interval for $m(x)$. A bandwidth sequence of order less than $n^{-\frac{1}{5}}$ must be chosen such that the bias vanishes asymptotically. For simultaneous error bars, we can use the technique based on the golden section bootstrap which is a delicate resampling technique used to approximate the joint-distribution of $\hat{m}_h(x) - m(x)$ at different points x.

Fixed design model This is the case where the density $f(x) = F^{'}(x)$ of the predictor variable is known, so that the Kernel weights become

$$W_{hi}(x) = \frac{K_h(x - X_i)}{f(x)}$$

and the estimate can be written as

$$\hat{m}_h = \frac{\hat{r}_h(x)}{f(x)}$$

We can employ the previous results concerning $\hat{r}_h(x)$ to derive the statistical properties of this smoother. If the X observations are taken at regular distances, we may assume that they are uniformly $U(0,1)$ distributed. In the fixed design model of nearly equispaced, nonrandom $\{X_i\}_{i=1}^n$ on $[0,1]$, Priestley et al. [1972] and Benedetti [1977] introduced the weight sequence

$$W_{hi}(x) = n(X_i - X_{i-1}) K_h(x - X_i) \, , X_0 = 0$$

The spacing $(X_i - X_{i-1})$ can be interpreted as an estimate of $n^{-1}f^{-1}$ from the Kernel weight above. Gasser et al. [1979] considered the weight sequence

$$W_{hi}(x) = n \int_{S_{i-1}}^{S_i} K_h(x - u) du$$

where $X_{i-1} \leq S_{i-1} \leq X_i$ is chosen between the odered X-data. It is related to the convolution smoothing proposed by Clark [1980].

4.2.4.2 Kernel smoothing algorithm

Computing kernel smoothing at N distinct points for a kernel with unbounded support would result in $o(Nn)$ operations. However, using kernels with bounded support, say $[-1,1]$ would result in $o(Nnh)$ operations since about $2nh$ points fall into an interval of length $2h$. One computational approach consists in using the WARPing defined in Section (4.2.1.2). Another one uses the Fourier transforms

$$\tilde{g}(t) = \int g(x)e^{-itx}dx$$

where for $g(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i$, the Fourier transform becomes

$$\tilde{g}(t) = \tilde{K}(th)\sum_{i=1}^{n} e^{-itX_i}Y_i$$

Using the Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$$

we get $\tilde{K}(th) = e^{-\frac{t^2}{2}}$. Decoupling the smoothing operation from the Fourier transform of the data $\sum_{i=1}^{n} e^{-itX_i}Y_i$, we can use the Fast Fourier Transform described in Appendix () with $o(N \log N)$ operations.

4.2.4.3 The K-nearest neighbour

Definition of the K-NN estimate Regression by Kernels is based on local averaging of observations Y_i in a fixed neighbourhood around x. Rather than considering this fixed neighbourhood, the K-NN employs varying neighbourhood in the X-variables which are among the K-nearest neighbours of x in Euclidean distance. Introduced by Loftsgaarden et al. [1965], it is defined in the form of the general nonparametric regression estimate

$$\hat{m}_K(x) = n^{-1}\sum_{i=1}^{n} W_{Ki}(x)Y_i$$

where the weight sequence $\{W_{Ki}(x)\}_{i=1}^{n}$ is defined through the set of indices

$$J_x = \{i : X_i \text{ is one of the K nearest observations to } x\}$$

This set of neighboring observations defines the K-NN weight sequence

$$W_{Ki}(x) = \begin{cases} \frac{n}{K} \text{ if } i \in J_x \\ 0 \text{ otherwise} \end{cases}$$

where the smoothing parameter K regulates the degree of smoothness of the estimated curve. Assuming fixed n, in the case where K becomes larger than n, the the K-NN smoother is equal to the average of the response variables. In the case $K = 1$ we obtain a step function with a jump in the middle between two observations.

Statistics of the K-NN estimate Again, we face a trade-off between a good approximation to the regression function and a good reduction of observational noise. It can be expressed formally by an expansion of the mean squared error of the K-NN estimate. Lai (1977) proposed the following theorem (See Hardle [1990] for references and proofs).

Theorem 4.2.4 *Let $K \to \infty$, $\frac{K}{n} \to 0$, $n \to \infty$. Bias and variance of the K-NN estimate $\hat{m}_K$ with weights as in the K-NN weight sequence, are given by*

$$E[\hat{m}_K(x)] - m(x) \approx \frac{1}{24f^3(x)}\left[(m^{''}f + 2m^{'}f^{'})(x)\right](\frac{K}{n})^2$$

and

$$Var(\hat{m}_K(x)) \approx \frac{\sigma^2(x)}{K}$$

We observe that the bias is increasing and the variance is decreasing in the smoothing parameter K. To balance this trade-off in an asymptotic sense, we should choose $K \sim n^{\frac{4}{5}}$. We then obtain for the mean squared error (MSE) a rate of convergence to zero of the order $K^{-1} = n^{-\frac{4}{5}}$. Hence, for this choice of K, the MSE is of the same order as for the Kernel regression. In addition to the uniform weights above, Stone [1977] defined triangular and quadratic K-NN weights. In general, the weights can be thought of as being generated by a Kernel function

$$W_{Ri}(x) = \frac{K_R(x - X_i)}{\hat{f}_R(x)}$$

where

$$\hat{f}_R(x) = n^{-1}\sum_{i=1}^{n} K_R(x - X_i)$$

is a Kernel density estimate of $f(x)$ with Kernel sequence

$$K_R(u) = R^{-1}K(\frac{u}{R})$$

and R is the distance between x and its kth nearest neighbour.

4.2.5 Bandwidth selection

While the accuracy of kernel smoothers, as estimators of m or of derivatives of m, is a function of the kernel K and the bandwidth h, it mainly depends on the smoothing parameter h. So far in choosing the smoothing parameter h we tried to compute sequences of bandwidths which approximate the A-MISE minimising bandwidth. However, the A-MISE of the Kernel regression smoother $\hat{m}_h(x)$ is not the only candidate for a reasonable measure of discrepancy between the unknown curve $m(x)$ and the approximation $\hat{m}_h(x)$. A list of distance measurements is given in Hardle [1991] together with the Hardle et al. [1986] theorem showing that they all lead asymptotically to the same level of smoothing. For convenience, we will only consider the distance that can be most easily computed, namely the average squared error (ASE).

4.2.5.1 Estimation of the average squared error

A typical representative quadratic measure of accuracy is the Integrated Squared Error (ISE) defined as

$$d_I(m, \hat{m}) = \int (m(x) - \hat{m}(x))^2 f(x)W(x)dx$$

where W denotes a nonnegative weight function. Taking the expectation of d_I with respect to X yields the MISE

$$d_M(m, \hat{m}) = E[d_I(m, \hat{m})]$$

A discrete approximation to d_I is the averaged squared error (ASE) defined as

$$d_A(m, \hat{m}) = ASE(h) = n^{-1}\sum_{i=1}^{n}(m(X_i) - \hat{m}_h(X_i))^2 W(X_i) \tag{4.2.9}$$

To illustrate the distribution of $ASE(h)$ we look at the $MASE(h)$, the conditioned squared error of $\hat{m}_h(x)$, conditioned on the given set of predictor variables $X_1, .., X_n$, and express it in terms of a variance component and a bias component.

$$\begin{aligned} d_C(m, \hat{m}) = MASE(h) &= E[ASE(h)|X_1, .., X_n] \\ &= n^{-1} \sum_{i=1}^{n} \big\{ Var(\hat{m}_h(X_i)|X_1, .., X_n) + Bias^2(\hat{m}_h(X_i)|X_1, .., X_n) \big\} W(X_i) \end{aligned}$$

where distance d_C is a random distance through the distribution of the Xs. The expectation of $ASE(h)$, d_C, contains a variance component $\nu(h)$

$$\nu(h) = n^{-1} \sum_{i=1}^{n} \big[n^{-2} \sum_{j=1}^{n} W_{hj}^2(X_i) \sigma^2(X_j) \big] W(X_i)$$

and a squared bias component $b^2(h)$

$$b^2(h) = n^{-1} \sum_{i=1}^{n} \big[n^{-1} \sum_{j=1}^{n} W_{hj}^2(X_i) m(X_j) - m(X_i) \big]^2 W(X_i)$$

The squared bias $b^2(h)$ increases with h, while $\nu(h)$ proportional to h^{-1} decreases. The sum of both components is $MASE(h)$ which shows a clear minimum. We therefore need to approximate the bandwidth minimising ASE. To do this we consider the averaged squared error (ASE) defined in Equation (4.2.9), and expand it

$$d_A(h) = n^{-1} \sum_{i=1}^{n} m^2(X_i) W(X_i) + n^{-1} \sum_{i=1}^{n} \hat{m}_h^2(X_i) W(X_i) - 2n^{-1} \sum_{i=1}^{n} m(X_i) \hat{m}_h(X_i) W(X_i)$$

where the first term is independent of h, the second term cam be entirely computed from the data, and the third term could be estimated if it was vanishing faster than d_A tends to zero. A naive estimate of this distance can be based on the replacement of the unknown value $m(X_i)$ by Y_i leading to the so-called Resubstitution estimate

$$P(h) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{m}_h(X_i))^2 W(X_i)$$

Unfortunately, $P(h)$ is a biased estimate of $ASE(h)$. The intuitive reason for this bias being that the observation Y_i is used in $\hat{m}_h(X_i)$ to predict itself. To get a deeper insight, denote $\epsilon_i = Y_i - m(X_i)$ the ith error term and consider the expansion

$$P(h) = n^{-1} \sum_{i=1}^{n} \epsilon_i^2 W(X_i) + ASE(h) - 2n^{-1} \sum_{i=1}^{n} \epsilon_i (\hat{m}_h(X_i) - m(X_i))^2 W(X_i)$$

Thus, the approximation of $ASE(h)$ by $P(h)$ would be fine if the last term had an expectation which is asymptotically of negligible size in comparison with the expectation of $ASE(h)$. This is unfortunately not the case as

$$\begin{aligned} E\big[-2n^{-1} \sum_{i=1}^{n} \epsilon_i (\hat{m}_h(X_i) - m(X_i))^2 W(X_i) | X_1, .., X_n \big] &= -2n^{-1} \sum_{i=1}^{n} E[\epsilon_i | X_1, .., X_n] \\ - \big(n^{-1} \sum_{j=1}^{n} W_{hj} m(X_j) - m(X_i) \big) W(X_i) - 2n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{hj}(X_i) E[\epsilon_i \epsilon_j | X_1, .., X_n] W(X_i) \end{aligned}$$

The error ϵ_i are independent random variables with expectation zero and variance $\epsilon^2(X_i)$. Hence,

$$E\big[-2n^{-1}\sum_{i=1}^{n}\epsilon_i(\hat{m}_h(X_i)-m(X_i))^2W(X_i)|X_1,..,X_n\big]=-2n^{-1}\sum_{i=1}^{n}W_{hi}(X_i)\epsilon^2(X_i)W(X_i)$$

This quantity tends to zero at the same rate as the variance component $\nu(h)$ of $ASE(h)$. Thus, $P(h)$ is biased by this additional variance component. We can use this naive estimate $P(h)$ to construct an asymptotically unbiased estimate of $ASE(h)$. We shall discuss two techniques, namely the concept of penalising functions which improve the estimate $P(h)$ by introducing a correcting term for this estimate, and the cross-validation where the computation is based on the leave-one-out estimate $\hat{m}_{hi}(X_i)$, the Kernel smoother without (X_i, Y_i).

4.2.5.2 Penalising functions

With the goal of asymptotically cancelling the bias, the prediction error $P(h)$ is adjusted by the correction term $\Xi(n^{-1}W_{hi}(X_i))$ to give the penalising function selector

$$G(h)=n^{-1}\sum_{i=1}^{n}(Y_i-\hat{m}_h(X_i))^2\Xi(n^{-1}W_{hi}(X_i))W(X_i)$$

The form of the correction term $\Xi(n^{-1}W_{hi}(X_i))$ is restricted by the first order Taylor expansion of Ξ

$$\Xi(u)=1+2u+o(u^2)\ ,\ u\to 0$$

Hence, the correcting term can be written as

$$\begin{aligned}\Xi(n^{-1}W_{hi}(X_i)) &= 1+2n^{-1}W_{hi}(X_i)+o((nh)^{-2})\ ,\ nh\to\infty\\ &= 1+2(nh)^{-1}\frac{K(0)}{\hat{f}_h(X_i)}+o((nh)^{-2})\end{aligned}$$

penalising values of h too low. We can work out the leading terms of $G(h)$ ignoring terms of lower order

$$\begin{aligned}G(h) &= n^{-1}\sum_{i=1}^{n}\epsilon_i^2W(X_i)+ASE(h)\\ &+ 2n^{-1}\sum_{i=1}^{n}\epsilon_i(m(X_i)-\hat{m}_h(X_i))W(X_i)+2n^{-2}\sum_{i=1}^{n}\epsilon_i^2W_{hi}(X_i)W(X_i)\end{aligned}$$

The first term is independent of h, and the expectation of the third summand (in equation above) is the negative expected value of the last term in the leading terms of $G(h)$. Hence, the last two terms cancel asymptotically so that $G(h)$ is roughly equal to $ASE(h)$, and as a result $G(h)$ is an unbiased estimator of $ASE(h)$. This gives rise to a lot of penalising functions which lead to asymptotically unbiased estimates of the ASE minimising bandwidth. The simplest function is of the form

$$\Xi(u)=1+2u$$

The objective of these selector functions is to penalise too small bandwidths. Any sequence of optimising bandwidths of one of these penalising functions is asymptotically optimal, that is, the ratio of the expected loss to the minimum loss tends to one. Denote $\hat{h}$ as the minimising bandwidth of $G(h)$ and $\hat{h}_0$ as the ASE optimal bandwidth. Then

$$\frac{ASE(\hat{h})}{ASE(\hat{h}_0)}\xrightarrow{p}1\ ,\ \frac{\hat{h}}{\hat{h}_0}\xrightarrow{p}1$$

However, the speed of convergence is slow. The relative difference between the estimate $\hat{h}$ and $\hat{h}_0$ is of rate $n^{-\frac{1}{10}}$ and we can not hope to derive a better one, since the relative difference between $\hat{h}_0$ and the MISE optimal bandwidth h_0 is of the same size.

4.2.5.3 Cross-validation

Cross-validation employs the leave-one-out estimates $\hat{m}_{hi}(X_i)$ in the formula of the prediction error instead of the original estimates. That is, one observation, say the ith one, is left out

$$\hat{m}_{hi}(X_i) = n^{-1} \sum_{j \neq i} W_{hj}(X_i) Y_j$$

This leads to the score function of cross-validation

$$CV(h) = n^{-1} \sum_{i=1}^{n} \big(Y_i - \hat{m}_{hi}(X_i)\big)^2 W(X_i)$$

The equation of the leading terms of $G(h)$ showed that $P(h)$ contains a component of roughly the same size as the variance of $ASE(h)$, but with a negative sign, so that the effect of the variance cancels. When we use the leave-one-out estimates of $\hat{m}_h(X_i)$ we arrive at

$$\begin{aligned} & E\big[-2n^{-1} \sum_{i=1}^{n} \epsilon_i \big(\hat{m}_{hi}(X_i) - m(X_i)\big) W(X_i) | X_1, .., X_n\big] \\ = \quad & -2n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} W_{hj}(X_i) E[\epsilon_i \epsilon_j | X_1, .., X_n] W(X_i) = 0 \end{aligned}$$

The cross-validation can also be understood in terms of penalising functions. Assume that $\hat{f}_{hi}(X_i) \neq 0$ and $\hat{m}_{hi}(X_i) \neq Y_i$ for all i, then we note that

$$\begin{aligned} CV(h) &= n^{-1} \sum_{i=1}^{n} \big(Y_i - \hat{m}_h(X_i)\big)^2 \big(\frac{\hat{m}_h(X_i) - Y_i}{\hat{m}_{hi}(X_i) - Y_i}\big)^{-2} W(X_i) \\ &= n^{-1} \sum_{i=1}^{n} \big(Y_i - \hat{m}_h(X_i)\big)^2 \big(1 - n^{-1} W_{hi}(X_i)\big)^{-2} W(X_i) \end{aligned}$$

Thus, the score function for CV can be rewritten as a penalising function with the selector of generalised cross-validation. Hence, a sequence of bandwidths on $CV(h)$ is asymptotically optimal and yields the same speed of convergence as the other techniques.

4.3 Trend filtering in the time domain

4.3.1 Some basic principles

We are now going to filter, in the time domain, the corrupted signal defined in Equation (4.1.1) in the case of financial time series. Slightly modifying notation, we let y_t be a stochastic process made of two different unobservable components, and assume that in its simplest form its dynamics are

$$y_t = x_t + \epsilon_t \tag{4.3.10}$$

where x_t is the trend, and the noise ϵ_t is a stochastic process. We are now concerned with estimating the trend x_t. We let $y = \{..., y_{-1}, y_0, y_1, ...\}$ be the ordered sequence of observations of the process y_t and $\hat{x}_t$ be the estimator of the unobservable underlying trend x_t. A filtering procedure consists in applying a filter $\mathcal{L}$ to the data y

$$\hat{x} = \mathcal{L}(y)$$

with $\hat{x} = \{..., \hat{x}_{-1}, \hat{x}_0, \hat{x}_1, ..\}$. In the case of linear filter we have $\hat{x} = \mathcal{L}y$ with the normalisation condition $1 = \mathcal{L}1$. Further, if the signal y_t is observed at regular dates, we get

$$\hat{x}_t = \sum_{i=-\infty}^{\infty} \mathcal{L}_{t,t-i} y_{t-i}$$

and the linear filter may be viewed as a convolution. Imposing some restriction on the coefficients $\mathcal{L}_{t,t-i}$, to use only past and present values, we get a causal filter. Further, considering only time invariant filters, we get a simple convolution of observed signal y_t with a window function $\mathcal{L}_i$

$$\hat{x}_t = \sum_{i=0}^{n-1} \mathcal{L}_i y_{t-i} \tag{4.3.11}$$

corresponding to the nonrecursive filter (or FIR) in Section (4.1.3.2). That is, a linear filter is characterised by a window kernel $\mathcal{L}_i$ and its support n where the former defines the type of filtering and the latter defines the range of the filter. When it is not possible to express the trend as a linear convolution of the signal and a window function, the filters are called nonlinear filters. Bruder et al. [2011] provide a detailed description of linear and non-linear filter. As an example of linear filter, in the well known moving average (MA), we take a square window on a compact support $[0, T]$ with $T = n\Delta$ [1], the width of the averaging window, and get the kernel

$$\mathcal{L}_i = \frac{1}{n} I_{\{i<n\}}$$

Note, the only calibration parameter is the window support $T = n\Delta$ characterising the smoothness of the filtered signal. In the limit $T \to 0$, the window becomes a Dirac distribution δ_t and the filtered signal becomes the observed one.

As an example, we are now going to use linear filtering to measure the trend of an asset price. We let S_t be asset price, and assume it follows a geometric Brownian motion with dynamics

$$\frac{dS_t}{S_t} = \mu_t dt + \sigma_t dW_t$$

where μ_t is the drift, σ_t is the volatility, and W_t is a standard Brownian motion. Assuming the asset price to be observed in a series of discrete dates $\{t_0, .., t_n\}$ we let $y_t = \ln S_t$ be the signal to be filtered, and let $R_t = \ln S_t - \ln S_{t-1}$ be the realised return at time t over a unit period. That is, $y_t = \ln S_{t-1} + R_t$ where $\ln S_{t-1}$ is known at time t. In the case where μ_t and σ_t are known, the return is given by

$$R_t = \big(\mu_t - \frac{1}{2}\sigma_t^2\big)\Delta + \sigma_t\sqrt{\Delta}a_t$$

where a_t is a standard Gaussian white noise. Note, R_t is a stochastic process with dynamics represented by Equation (4.3.10). In order to filter the drift μ_t from the asset price S_t, we can apply Equation (4.3.11) to the process R_t

$$\big(\hat{\mu}_t - \frac{1}{2}\sigma_t^2\big)\Delta = \sum_{i=0}^{n-1} \mathcal{L}_i R_{t-i}$$

[1] $\Delta = t_{i+1} - t_i$

and get the estimator

$$\hat{\mu}_t = \frac{1}{2}\sigma_t^2 + \frac{1}{\Delta}\sum_{i=0}^{n-1}\mathcal{L}_i R_{t-i}$$

Neglecting the contribution from the term σ_t^2, the estimator for μ_t becomes

$$\hat{\mu}_t \approx \frac{1}{\Delta}\sum_{i=0}^{n-1}\mathcal{L}_i R_{t-i}$$

From the above equation, we see that $\hat{\mu}_t$ is a biased estimator of μ_t and that the bias increases with the volatility σ_t. Alternatively, using Equation (4.3.11) on the process $y_t = \ln S_{t-1} + R_t$, we can also filter μ_t by directly filtering the trend of the process y_t, getting

$$\hat{x}_t = \sum_{i=0}^{n-1}\mathcal{L}_i \ln S_{t-i-1} + \sum_{i=0}^{n-1}\mathcal{L}_i R_{t-i} \approx \sum_{i=0}^{n-1}\mathcal{L}_i \ln S_{t-i-1} + \hat{\mu}_t\Delta$$

so that the estimator of μ_t can be given by

$$\hat{\mu}_t \approx \frac{1}{\Delta}\big(\hat{x}_t - \sum_{i=0}^{n-1}\mathcal{L}_i \ln S_{t-i-1}\big) = \frac{1}{\Delta}\big(\sum_{i=0}^{n-1}\mathcal{L}_i y_{t-i} - \sum_{i=0}^{n-1}\mathcal{L}_i \ln S_{t-i-1}\big)$$

***Remark* 4.3.1** *While the estimator $\hat{x}_t$ is used for econometric methods, the estimator $\hat{\mu}_t$ is more important for trading strategies.*

Defining the derivative of the window function as $l_i = \dot{\mathcal{L}}_i$ we can rewrite the estimator of the drift μ_t as

$$\hat{\mu}_t \approx \frac{1}{\Delta}\sum_{i=0}^{n} l_i y_{t-i}$$

where

$$l_i = \begin{cases} \mathcal{L}_0 \text{ if } i = 0 \\ \mathcal{L}_i - \mathcal{L}_{i-1} \text{ if } i = 1, ..., n-1 \\ -\mathcal{L}_{n-1} \text{ if } i = n \end{cases}$$

Using the derivative operator we see that $\hat{\mu}_t = \frac{d}{dt}\hat{x}_t$.

4.3.2 The local averages

While the simplest trend filtering is the moving average filter, observations can be averaged by using many different types of weightings. We are going to present a few of these weighting averages or linear filters. In the case of the moving average (MA) defined above, the estimator in Equation (4.3.11) simplifies to

$$\hat{x}_t = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i} + \frac{1}{n}\sum_{i=0}^{n-1}\epsilon_{t-i} = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}$$

if we assume that the noise ϵ_t is independent from x_t and is a centred process. If the trend is homogeneous, the average value is located at $t - \frac{1}{2}(n-1)$ and the filtered signal lags the observed signal by a time period being half the window. The main advantage of the MA filter is the reduction of noise due to the central limit theorem. On average, the noisy parts of observations cancel each other out and the trend has a cumulative nature. In the limit case $n \to \infty$, the signal is completely denoised as it corresponds to the average of the value of the trend. Further, the estimator is

also biased and as a result one must simultaneously maximise denoising while minimising bias. Note, we previously expressed the estimator of μ_t by filtering the trend of the returns so that one should use the moving average of returns. If practitioners use the average of the logarithm of the price, the trend should be estimated from the difference between two moving averages over two different time horizons. To conclude, the MA filter can also be directly applied to the signal, and $\hat{\mu}_t$ becomes the cumulative return over the window period which only needs the first and last dates of the period under consideration. Applying directly a uniform moving average to the logarithm of the prices, we get

$$\hat{y}_t^n = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i}$$

In order to estimate the trend μ_t we need to compute the difference between two moving averages over two different time horizons n_1 and n_2. Assuming $n_1 > n_2$, the trend is approximated by

$$\hat{\mu}_t \approx \frac{2}{(n_1 - n_2)\Delta}(\hat{y}_t^{n_2} - \hat{y}_t^{n_1})$$

which is positive when the short-term moving average is higher than the long-term moving average. Hence, the sign of the approximated trend changes when the short-term MA crosses the long-term one. Note, this estimator may be viewed as a weighted moving average of asset returns. Inverting the derivative window l_i we recover the operator $\mathcal{L}_i$ as

$$\mathcal{L}_i = \begin{cases} l_0 \text{ if } i = 0 \\ l_i + \mathcal{L}_{i-1} \text{ if } i = 1, ..., n-1 \\ -l_{n-1} \text{ if } i = n \end{cases}$$

and one can then interpret the estimator in terms of asset returns. The weighting of each return in the estimator forms a triangle where the biggest weighting is given at the horizon of the smallest MA. Hence, the indicator can be focused towards the current trend (if n_2 is small) or towards past trends (if n_2 is as large as $\frac{n_1}{2}$).

In order to improve the uniform MA estimator, we can consider the kernel function

$$l_i = \frac{4}{n^2} sign(\frac{n}{2} - i)$$

where the estimator $\hat{\mu}_t$ takes into account all the dates of the window period. Taking the primitive of the function l_i, the filter becomes

$$\mathcal{L}_i = \frac{4}{n^2}(\frac{n}{2} - |i - \frac{n}{2}|)$$

Other types of MA filter exists which are characterised by an asymmetric form of the convolution kernel. For instance, one can take an asymmetric window function with a triangular form

$$\mathcal{L}_i = \frac{2}{n^2}(n - i)I_{\{i<n\}}$$

and computing the derivative of that window function we get the kernel

$$l_i = \frac{2}{n}(\delta_i - I_{\{i<n\}})$$

leading to the estimator

$$\hat{\mu}_t = \frac{2}{n}\big(x_t - \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}\big)$$

Another approach is to consider the Lanczos generalised derivative which is more general than the traditional derivative and offers more advantages by allowing one to compute a pseudo-derivative at points where the function is not differentiable (see Groetsch [1998]). In the case of interest to us, the traditional derivative of the observable signal y_t does not exists due to the noise ϵ_t, but it exists in the case of the Lanczos derivative. Using the Lanczos's formula to differentiate the trend at the point $t - \frac{T}{2}$ we get

$$\frac{d^L}{dt}\hat{x}_t = \frac{12}{n^3}\sum_{i=0}^{n}(\frac{n}{2} - i)y_{t-i}$$

we obtain the kernel

$$l_i = \frac{12}{n^3}(\frac{n}{2} - i)I_{0\leq i\leq n}$$

and by integrating by parts, we obtain the trend filter

$$\mathcal{L}_i = \frac{6}{n^3}i(n-i)I_{0\leq i\leq n}$$

Note, one can extend these filters by computing the convolution of two or more filters. Also, the choice of n has a big impact on the filtered series. Further, the trader may be more interested in the derivative of the trend than the absolute value of the trend itself in which case the choice of window function is important.

4.3.3 The Savitzky-Golay filter

When measuring a variable that is both slowly varying and also corrupted by random noise, the use of low-pass filters permits to smooth that noisy data. Hence, to find the trend, we can consider the Savitzky-Golay low-pass smoothing filter (also called least-squares or Digital Smoothing Polynomial (DISPO)) which is a kind of generalised moving average. Rather than having their properties defined in the Fourier domain and then translated to the time domain, SG filters derive directly from a particular formulation of the data smoothing problem in the time domain. The filter coefficients are derived by performing an unweighted linear least squares fit using a polynomial of a given degree (see Press et al. [1992]). Assume we want to smooth a series of equally spaced data points $f_i = f(t_i)$ where $t_i = t_0 + i\Delta$ for some constant Δ and $i = ... - 2, -1, 0, 1, 2, ...,$. This filter replaces each data value f_i by a linear combination g_i of itself and some number of nearby neighbours

$$g_i = \sum_{n=-n_L}^{n_R} c_n f_{i+n} \tag{4.3.12}$$

where n_L is the number of points used to the left of a data point i, and n_R is the number of points used to the right. Setting $n_R = 0$ is called a causal filter, while setting $c_n = \frac{1}{n_L+n_R+1}$ we recover the moving window averaging (MA) which always reduces the function value at a local maximum. In the spectrometric application, a narrow spectral line has its height reduced and its width increased, introducing an undesirable bias. However, MA preserves the area under a spectral line (zeroth moment) and also (if $n_L = n_R$) its mean position in time (first moment). Yet, even though the first moment is preserved, the second moment, being equivalent to the line width, is violated. Hence, the main idea of Savitzky-Golay filtering is to find filter coefficients c_n that preseve higher moments by approximating the underlying function within the moving window not by a constant (whose estimate is the average), but by a polynomial of higher order, typically quadratic or quartic. For each point f_i we least-squares fit a polynomial to all $n_L + n_R + 1$ points in the moving window, and then set g_i to be the value of that polynomial at position i. We make no use of the value of the polynomial at any other point. When we move on to the next point f_{i+1}, we do a whole new least-squares fit using a shifted window. As the process of least-squares fitting involves only a linear matrix inversion, the coefficients of a fitted polynomial are themselves linear in the values of the data so that all the fitting can be done in advance (for fictious data made of zeros except for a single one) and then do the fits on the real data just by taking linear combinations. In order to compute g_i we want to fit a polynomial of degree M in i

$$a_0 + a_1 i + ... + a_M i^M$$

to the values $f_{-n_L}, .., f_{n_R}$. Note, M is the order of the smoothing polynomial, which is also equal to the highest conserved moment. Then g_0 will be the value of that polynomial at $i = 0$, namely a_0. The normal equations in matrix form for this least square problem is

$$(A^\top.A).a = A^\top.f \text{ or } a = (A^\top.A)^{-1}.(A^\top.f)$$

where

$$A_{ij} = i^j \ , i = -n_L, .., n_R \ , j = 0, .., M$$

We also have the specific forms

$$\{A^\top.A\}_{ij} = \sum_{k=-n_L}^{n_R} A_{ki}A_{kj} = \sum_{k=-n_L}^{n_R} k^{i+j}$$
$$\{A^\top.f\}_j = \sum_{k=-n_L}^{n_R} A_{kj}f_k = \sum_{k=-n_L}^{n_R} k^j f_k$$

Since the coefficient c_n is the component a_0 when f is replaced by the unit vector e_n for $-n_L \leq n \leq n_R$, we have

$$c_n = \{(A^\top.A)^{-1}.(A^\top.e_n)\}_0 = \sum_{m=0}^{M} \{(A^\top.A)^{-1}\}_{0m} n^m$$

meaning that we only need one row of the inverse matrix (numerically we can get this by LU decomposition with only a single back-substitution). A higher degree polynomial makes it possible to achieve a high level of smoothing without attenuation of real data features. Hence, within limits, the Savitzky-Golay filtering manages to provides smoothing without loss of resolution. When dealing with irregularly sampled data, where the values f_i are not uniformly spaced in time, there is no way to obtain universal filter coefficients applicable to more than one data point. Note, the Savitzky-Golay technique can also be used to compute numerical derivatives. In that case, the desired order is usually $m = M = 4$ or larger where m is the order of the smoothing polynomial, also equal to the highest conserved moment. Numerical experiments are usually done with a 33 point smoothing filter, that is $n_L = n_R = 16$.

4.3.4 The least squares filters

4.3.4.1 The L2 filtering

An alternative to averaging observations is to impose a model on the process y_t and its trend x_t (see Section (4.1.2)). For instance, the Lanczos filter in the previous section may be considered as a local linear regression. Given a model for the process y_t, least squares methods are often used to define trend estimators

$$\{\hat{x}_1, ..., \hat{x}_n\} = arg\min \frac{1}{2} \sum_{t=1}^{n} (y_t - \hat{x}_t)^2$$

but the problem is ill-posed and one must impose some restrictions on the underlying process y_t or on the filtered trend $\hat{x}_t$ to obtain a solution. For instance, we can consider the deterministic constant trend

$$x_t = x_{t-1} + \mu$$

such that the process y_t becomes $y_t = x_{t-1} + \mu + \epsilon_t$. Iterating with $x_0 = 0$ [2], we get the process

$$y_t = \mu t + \epsilon_t \tag{4.3.13}$$

and estimating the filtered trend $\hat{x}_t$ is equivalent to estimating the coefficient μ

$$\hat{\mu} = \frac{\sum_{t=1}^{n} t y_t}{\sum_{t=1}^{n} t^2}$$

In the case where the trend is not constant one can consider the Hodrick-Prescott filter (or L_2 filter) where the objective function is

$$\frac{1}{2}\sum_{t=1}^{n}(y_t - \hat{x}_t)^2 + \lambda \sum_{t=2}^{n-1}(\hat{x}_{t-1} - 2\hat{x}_t + \hat{x}_{t+1})^2$$

where $\lambda > 0$ is a regularisation parameter controlling the trade off between the smoothness of $\hat{x}_t$ and the noise $(y_t - \hat{x}_t)$. Rewriting the objective function in vectorial form, we get

$$\frac{1}{2}\|y - \hat{x}\|_2^2 + \lambda \|D\hat{x}\|_2^2$$

where the operator D is the $(n-2) \times n$ matrix

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & \\ 0 & 0 & 0 & 0 & ... & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & 1 & -2 & 1 \end{bmatrix}$$

so that the estimator becomes

$$\hat{x} = \left(I + 2\lambda D^\top D\right)^{-1} y$$

4.3.4.2 The L1 filtering

One can generalise the Hodrick-Prescott filter to a larger class of filters by using the L_p penalty condition instead of the L_2 one (see Daubechies et al. [2004]). If we consider an L_1 filter, the objective function becomes

$$\frac{1}{2}\sum_{t=1}^{n}(y_t - \hat{x}_t)^2 + \lambda \sum_{t=2}^{n-1}|\hat{x}_{t-1} - 2\hat{x}_t + \hat{x}_{t+1}|$$

which is expressed in vectorial form, as

$$\frac{1}{2}\|y - \hat{x}\|_2^2 + \lambda \|D\hat{x}\|_1$$

Kim et al. [2009] showed that the dual problem of the L_1 filter scheme is a quadratic program with some boundary constraints. Since the L_1 norm imposes the condition that the second derivative of the filtered signal must be zero, we obtain a set of straight trends and breaks. Hence, the smoothing parameter λ plays an important role in detecting the number of breaks.

[2] $y_t = x_{t-n} + n\mu + \epsilon_t$, with $x_{t-n} = x_0$.

4.3.4.3 The Kalman filters

Another approach to estimating the trend is to consider the Kalman filter where the trend μ_t is a hidden process following a given dynamics (see details in Appendix (D.4)). For instance, we can assume it follows the dynamics

$$\begin{aligned} R_t &= \mu_t + \sigma_R \epsilon_R(t) \\ \mu_t &= \mu_{t-1} + \sigma_\mu \epsilon_\mu(t) \end{aligned} \tag{4.3.14}$$

where R_t is the observable signal of realised returns and the hidden process μ_t follows a random walk. Hence, it follows a Markov model. If we let the conditional trend be $\hat{\mu}_{t|t-1} = E_{t-1}[\mu_t]$ and the estimation error be $P_{t|t-1} = E_{t-1}[(\hat{\mu}_{t|t-1} - \mu_t)^2]$, then we get the forecast estimator

$$\hat{\mu}_{t+1|t} = (1 - K_t)\hat{\mu}_{t|t-1} + K_t R_t$$

where

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + \sigma_R^2}$$

which is the Kalman gain. The estimation error is given by the Riccati's equation

$$P_{t+1|t} = P_{t|t-1} + \sigma_\mu^2 - P_{t|t-1}K_t$$

with stationary solution

$$P^* = \frac{1}{2}\sigma_\mu\left(\sigma_\mu + \sqrt{\sigma_\mu^2 + 4\sigma_R^2}\right)$$

and the filter equation becomes

$$\hat{\mu}_{t+1|t} = (1 - \kappa)\hat{\mu}_{t|t-1} + \kappa R_t$$

with

$$\kappa = \frac{2\sigma_\mu}{\left(\sigma_\mu + \sqrt{\sigma_\mu^2 + 4\sigma_R^2}\right)}$$

Note, the Kalman filter in Equation (4.3.14) can be rewritten as an exponential moving average (EMA) filter with parameter $\lambda = -\ln(1 - \kappa)$ for $0 < \kappa < 1$ and $\lambda > 0$. In this setting, the estimator is given by

$$\hat{\mu}_t = (1 - e^{-\lambda}) \sum_{i=0}^{\infty} e^{-\lambda i} R_{t-i}$$

with $\hat{\mu}_t = E_t[\mu_t]$ so that the 1-day forecast estimator is $\hat{\mu}_{t+1|t} = \hat{\mu}_t$. From our discussion above, the filter of the trend $\hat{x}_t$ is given by the equation

$$\hat{x}_t = (1 - e^{-\lambda}) \sum_{i=0}^{\infty} e^{-\lambda i} y_{t-i}$$

and the derivative of the trend may be related to the signal y_t by the following equation

$$\hat{\mu}_t = (1 - e^{-\lambda})y_t - (1 - e^{-\lambda})(e^{\lambda} - 1) \sum_{i=1}^{\infty} e^{-\lambda i} y_{t-i}$$

One can relate the regression model in Equation (4.3.13) with the Markov model in Equation (4.3.14) by notting that they are special cases of the structural models described in Appendix (D.4.12). More precisely, the regression model in Equation (4.3.13) is equivalent to the state space model

$$\begin{aligned} y_t &= x_t + \sigma_y \epsilon_y(t) \\ x_t &= x_{t-1} + \mu \end{aligned}$$

If we let the tend be stochastic we get the local level model

$$\begin{aligned} y_t &= x_t + \sigma_y \epsilon_y(t) \\ x_t &= x_{t-1} + \mu + \sigma_x \epsilon_x(t) \end{aligned}$$

Further, assuming that the slope of the trend is stochastic, we obtain the local linear trend model

$$\begin{aligned} y_t &= x_t + \sigma_y \epsilon_y(t) \\ x_t &= x_{t-1} + \mu_{t-1} + \sigma_x \epsilon_x(t) \\ \mu_t &= \mu_{t-1} + \sigma_\mu \epsilon_\mu(t) \end{aligned} \tag{4.3.15}$$

and setting $\sigma_y = 0$ we recover the Markov model in Equation (4.3.14). These examples are special case of structural models which can be solved by using the Kalman filter. Note, the Kalman filter is optimal in the case of the linear Gaussian model in Equation (), and it can be regarded as an efficient computational solution of the least squares method (see Sorenson [1970]).

***Remark* 4.3.2** *The Kalman filter can be used to solve more sophisticated process than the Markov model, but some nonlinear or non Gaussian models may be too complex for Kalman filtering.*

To conclude, the Kalman filter can be used to derive an optimal smoother as it improves the estimate of $\hat{x}_{t-i}$ by using all the information between $t-i$ and t.

4.3.5 Calibration

When filtering trends from time series, one must consider the calibration of the filtering parameters. We briefly discuss two possible calibration schemes, one where the calibrated parameters incorporate our prediction requirement and the other one where they can be mapped to a known benchmark estimator.

To illustrate the approach of statistical inference we consider the local linear trend model in Equation (4.3.15). We estimate the set of parameters $(\sigma_y, \sigma_x, \sigma_\mu)$ by maximising the log-likelihood function

$$l = \frac{1}{2} \sum_{t=1}^{n} \ln 2\pi + \ln F_t + \frac{v_t^2}{F_t}$$

where $v_t = y_t - E_{t-1}[y_t]$ is the (one-day) innovative process and $F_t = E_{t-1}[v_t^2]$ is the variance of v_t.

In order to look at longer trend, the innovation process becomes $v_t = y_t - E_{t-h}[y_t]$ where h is the horizon time. In that setting, we calibrate the parameters θ by using a cross-validation technique. We divide our historical data into an in-sample set and an out-sample set charaterised by two time parameters T_1 and T_2 where the size of the former controls the precision of the calibration to the parameter θ. We compute the value of the expectation $E_{t-h}[y_t]$ in the in-sample set which are used in the out-sample set to estimate the prediction error

$$e(\theta;h) = \sum_{t=1}^{n-h} \big(y_t - E_{t-h}[y_t]\big)^2$$

which is directly related to the prediction horizon $h = T_2$ for a given strategy. Minimising the prediction error, we get the optimal value θ^* of the filter parameter used to predict the trend for the test set.

The estimator of the slope of the trend $\hat{\mu}_t$ is a random value defined by a probability distribution function, and based on the sample data, takes a value called the estimate of the slope. If we let μ_t^0 be the true value of the slope, the quality of the slope is defined by the mean squared error (MSE)

$$MSE(\hat{\mu}_t) = E[(\hat{\mu}_t - \mu_t^0)^2]$$

The estimator $\hat{\mu}_t^{(1)}$ is more efficient than the estimator $\hat{\mu}_t^{(2)}$ if its MSE is lower

$$\hat{\mu}_t^{(1)} \succ \hat{\mu}_t^{(2)} \Leftrightarrow MSE(\hat{\mu}_t^{(1)}) \leq MSE(\hat{\mu}_t^{(2)})$$

Decomposing the MSE into two components we get

$$MSE(\hat{\mu}_t) = E[(\hat{\mu}_t - E[\hat{\mu}_t])^2] + E[(E[\hat{\mu}_t] - \mu_t^0)^2]$$

where the first component is the variance of the estimator $Var(\hat{\mu}_t)$ while the second one is the square of the bias $B(\hat{\mu}_t)$. When comparing unbiased estimators, we are left with comparing their variances. Hence, the estimate of a trend may not be significant when the variance of the estimator is too large.

4.3.6 Introducing linear prediction

We let $\{y_\alpha^{'}\}$ be a set of measured values for some underlying set of true values of a quantity y, denoted $\{y_\alpha\}$, related to these true values by the addition of random noise

$$y_\alpha^{'} = y_\alpha + \eta_\alpha$$

The Greek subscript indexing the vales indicates that the data points are not necessarily equally spaced along a line, or even ordered. We want to construct the best estimate of the true value of some particular point y_* as a linear combination of the known, noisy values. That is, given

$$y_* = \sum_\alpha d_{*\alpha} y_\alpha^{'} + x_* \tag{4.3.16}$$

we want to find coefficients $d_{*\alpha}$ minimising the discepancy x_*.

***Remark* 4.3.3** *In the case where we let y_* be one of the existing y_α's the problem becomes one of optimal filtering or estimation. On the other hand, if y_* is a completely new point then the problem is that of a linear prediction.*

One way forward is to minimise the discrepancy x_* in the statistical mean square sense. That is, assuming that the noise is uncorrelated with the signal ($< \eta_\alpha y_\beta >= 0$), we seek $d_{*\alpha}$ minimising

$$\begin{aligned} < x_*^2 > &= < \big[\sum_\alpha d_{*\alpha}(y_\alpha + \eta_\alpha) - y_*\big]^2 > \\ &= \sum_{\alpha\beta}(< y_\alpha y_\beta > + < \eta_\alpha \eta_\beta >) d_{*\alpha} d_{*\beta} - 2\sum_\alpha < y_* y_\alpha > d_{*\alpha} + < y_*^2 > \end{aligned}$$

where $<>$ is the statistical average, and β is a subscript to index another member of the set. Note, $< y_\alpha y_\beta >$ and $< y_* y_\alpha >$ describe the autocorrelation structure of the underlying data. For point to point uncorrelated noise we get $< \eta_\alpha \eta_\beta >=< \eta_\alpha^2 > \delta_{\alpha\beta}$ where $\delta_.$ is the Dirac function. One can think of the various correlation quantities as comprising matrices and vectors

$$\phi_{\alpha\beta} =< y_\alpha y_\beta > \ , \phi_{*\alpha} =< y_* y_\alpha > \ , \eta_{\alpha\beta} =< \eta_\alpha \eta_\beta > \text{ or } < \eta_\alpha^2 > \delta_{\alpha\beta}$$

Setting the derivative with respect to the $d_{*\alpha}$'s equal to zero in the above equation, we get the set of linear equations

$$\sum_\beta [\phi_{\alpha\beta} + \eta_{\alpha\beta}] d_{*\beta} = \phi_{*\alpha}$$

Writing the solution as a matrix inverse and omitting the minimised discrepancy x_*, the estimation of Equation (4.3.16) becomes

$$y_* \approx \sum_{\alpha\beta} \phi_{*\alpha} [\phi_{\mu\nu} + \eta_{\mu\nu}]^{-1}_{\alpha\beta} y'_\beta \tag{4.3.17}$$

We can also calculate the expected mean square value of the discrepancy at its minimum

$$< x_*^2 >_0 =< y_*^2 > - \sum_\beta d_{*\beta} \phi_{*\beta} =< y_*^2 > - \sum_\beta \phi_{*\alpha} [\phi_{\mu\nu} + \eta_{\mu\nu}]^{-1}_{\alpha\beta} \phi_{*\beta}$$

Replacing the star with the Greek index γ, the above formulas describe optimal filtering. In the case where the noise amplitudes η_α goes to zero, so does the noise autocorrelations $\eta_{\alpha\beta}$, cancelling a matrix times its inverse, Equation (4.3.17) simply becomes $y_\gamma = y'_\gamma$. In the case where the matrices $\phi_{\alpha\beta}$ and $\eta_{\alpha\beta}$ are diagonal, Equation (4.3.17) becomes

$$y_\gamma = \frac{\phi_{\gamma\gamma}}{\phi_{\gamma\gamma} + \eta_{\gamma\gamma}} y'_\gamma \tag{4.3.18}$$

which is Equation (4.1.2) with $S^2 \to \phi_{\gamma\gamma}$ and $N^2 \to \eta_{\gamma\gamma}$. For the case of equally spaced data points, and in the Fourier domain, autocorrelations simply become squares of Fourier amplitudes (Wiener-Khinchin theorem), and the optimal filter can be constructed algebraically, as Equation (4.3.18), without inverting any matrix. In the time domain, or any other domain, an optimal filter (minimising the square of the discrepancy from the underlying true value in the presence of measurement noise) can be constructed by estimating the autocorrelation matrices $\phi_{\alpha\beta}$ and $\eta_{\alpha\beta}$, and applying Equation (4.3.17) with $* \to \gamma$. Classical linear prediction (LP) specialises to the case where the data points y_β are equally spaced along a line y_i for $i = 1, ., N$ and we want to use M consecutive values of y_i to predict the $M + 1$ value. Note, stationarity is assumed, that is, the autocorrelation $< y_j y_k >$ is assumed to depend only on the difference $|j - k|$, and not on j or k individually, so that the autocorrelation ϕ has only a single index

$$\phi_j =< y_i y_{i+j} >\approx \frac{1}{N-j} \sum_{i=1}^{N-j} y_i y_{i+j}$$

However, there is a better way to estimate the autocorrelation. In that setting the estimation Equation (4.3.16) is

$$y_n = \sum_{j=1}^{M} d_j y_{n-j} + x_n \tag{4.3.19}$$

so that the set of linear equations above becomes the set of M equations for the M unknown d_j's, called the linear prediction (LP) coefficients

$$\sum_{j=1}^{M} \phi_{|j-k|} d_j = \phi_k \ , k = 1, .., M$$

Note, results obtained from linear prediction are remarkably sensitive to exactly how the ϕ_k's are estimated. Even though the noise is not explicitly included in the equations, it is properly accounted for, if it is point-to-point uncorrelated. Note, ϕ_0 above estimates the diagonal part of $\phi_{\alpha\alpha} + \eta_{\alpha\alpha}$, and the mean square discrepancy $< x_n^2 >$ is given by

$$< x_n^2 >= \phi_0 - \phi_1 d_1 - ... - \phi_M d_M$$

Hence, we first compute the d_j's with the equations above, then calculate the mean square discrepancy $< x_n^2 >$. If the discrepancies are small, we continue applying Equation (4.3.19) right on into the future, assuming future discrepancies x_i to be zero. This is a kind of extrapolation formula. Note, Equation (4.3.19) being a special case of the general linear filter, the condition for stability is that the characteristic polynomial

$$z^N - \sum_{j=1}^{N} d_j z^{N-j} = 0$$

has all N of its roots inside the unit circle

$$|z| \leq 1$$

If the data contain many oscillations without any particular trend towards increasing or decreasing amplitude, then the complex roots of the polynomial will generally all be rather close to the unit circle. When the instability is a problem, one should massage the LP coefficients by

1. solving numerically the polynomial for its N complex roots
2. moving the roots to where we think they should be inside or on the unit circle
3. reconstructing the modified LP coefficients

Assuming that the signal is truly a sum of undamped sine and cosine waves, one can simply move each root z_i onto the unit circle

$$z_i \to \frac{z_i}{|z_i|}$$

Alternatively, one can reflect a bad root across the unit circle

$$z_i \to \frac{1}{z_i^*}$$

preserving the amplitude of the output of Equation (4.3.19) when it is driven by a sinusoidal set of x_i's. Note, the choice of M, the number of LP coefficients to use, is an open problem. Linear prediction is successful at extrapolating signals that are smooth and oscillatory, not necessarily periodic.

Chapter 5

Presenting time series analysis

5.1 Basic principles of linear time series

We consider the asset returns (see Section (3.3.1)) to be a collection of random variables over time, obtaining the time series $\{r_t\}$ in the case of log returns. Linear time series analysis is a first step to understanding the dynamic structure of such a series (see Box et al. [1994]). That is, for an asset return r_t, simple models attempt at capturing the linear relationship between r_t and some information available prior to time t. For instance, the information may contain the historical values of r_t and the random vector Y that describes the economic environment under which the asset price is determined. As a result, correlations between the variable of interest and its past values become the focus of linear time series analysis, and are referred to as serial correlations or autocorrelations. Hence, Linear models can be used to analyse the dynamic structure of such a series with the help of autocorrelation function, and forecasting can then be performed (see Brockwell et al. [1996]).

5.1.1 Stationarity

While the foundation of time series analysis is stationarity, autocorrelations are basic tools for studying this stationarity. A time series $\{x_t, \mathbb{Z}\}$ is said to be strongly stationary, or strictly stationary, if the joint distribution of $(x_{t_1}, .., x_{t_k})$ is identical to that of $(x_{t_1+h}, .., y_{t_k+h})$ for all h

$$(x_{t_1}, .., x_{t_k}) = (x_{t_1+h}, .., y_{t_k+h})$$

where k is an arbitrary positive integer and $(t_1, .., t_k)$ is a collection of k positive integers. Thus, strict stationarity requires that the joint distribution of $(x_{t_1}, .., x_{t_k})$ is invariant under time shift. Since this condition is difficult to verify empirically, a weaker version of stationarity is often assumed. The time series $\{x_t, \mathbb{Z}\}$ is weakly stationary if both the mean of x_t and the covariance between x_t and x_{t-k} are time-invariant, where k is an arbitrary integer. That is, $\{x_t\}$ is weakly stationary if

$$E[x_t] = \mu \text{ and } Cov(x_t, x_{t-k}) = \gamma_k$$

where μ is constant and γ_k is independent of t. That is, we assume that the first two moments of x_t are finite. In the special case where x_t is normally distributed, then the weak stationarity is equivalent to strict stationarity. The covariance γ_k is called the lag-k autocovariance of x_t and has the following properties:

- $\gamma_0 = Var(x_t)$
- $\gamma_{-k} = \gamma_k$

The latter holds because $Cov(x_t, x_{t-(-k)}) = Cov(x_{t-(-k)}, x_t) = Cov(x_{t+k}, x_t) = Cov(x_{t_1}, x_{t_1-k})$, where $t_1 = t + k$.

In the finance literature, it is common to assume that an asset return series is weakly stationary since a stationary time series is easy to predict as its statistical properties are constant. However, financial time series such as rates, FX, and equity are non-stationary. The non-stationarity of price series is mainly due to the fact that there is no fixed level for the price which is called unit-root non-stationarity time series. It is well known that these underlyings are prone to different external shocks. In a stationary time series, these shocks should eventually die away, meaning that a shock occurring at time t will have a smaller effect at time $t + 1$, and an even smaller effect at time $t + 2$ gradually dying out. However, if the data is non-stationary, the persistence of shocks will always be infinite, meaning that a shock at time t will not have a smaller effect at time $t + 1$, $t + 2$ and so on.

5.1.2 The autocorrelation function

Studies often mention the problem of timely dependence in returns series of stocks or indices. Typically, estimates for non existent return figures are then set equal to the last reported transaction price. This results in serial correlation for stock prices, which further causes distortions in the parameter estimates, especially the standard deviation. When the linear dependence between x_t and x_{t-i} is of interest, we consider a generalisation of the correlation called the autocorrelation.

Definition 5.1.1 *ACF*
The autocorrelation function (ACF), $\rho(k)$ for a weakly-stationary time series, $\{x_t : t \in \mathbb{N}\}$ is given by

$$\rho(k) = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sigma^2}$$

where $E[x_t]$ is the expectation of x_t, μ is the mean and σ^2 is the variance.

Following Eling [2006], we compute the first order autocorrelation value for all stocks and then use the Ljung-Box statistic (see Ljung et al. [1978]) to check whether this value is statistically significant. It test for high order serial correlation in the residuals. Given two random variables X and Y, the correlation coefficient between these two variables is

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sqrt{E[(X - \mu_x)^2]E[(Y - \mu_y)^2]}}$$

where μ_x and μ_y are the mean of X and Y, and with $-1 \leq \rho_{x,y} \leq 1$ and $\rho_{x,y} = \rho_{y,x}$. Given the sample $\{(x_t, y_t)\}_{t=1}^T$, then the sample correlation can be consistently estimated by

$$\hat{\rho}_{x,y} = \frac{\sum_{t=1}^T (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^T (x_t - \overline{x})^2 \sum_{t=1}^T (y_t - \overline{y})^2}}$$

where $\overline{x} = \frac{1}{T}\sum_{t=1}^T x_t$ and $\overline{y} = \frac{1}{T}\sum_{t=1}^T y_t$ are respectively the sample mean of X and Y. Similarly, given the weakly stationary time seris $\{x_t\}$, the lag-k autocorrelation of x_t is defined by

$$\rho_k = \frac{Cov(x_t, x_{t-k})}{\sqrt{Var(x_t)Var(x_{t-k})}} = \frac{Cov(x_t, x_{t-k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}$$

since $Var(x_t) = Var(x_{t-k})$ for a weakly stationary series. We have $\rho_0 = 1$, $\rho_k = \rho_{-k}$, and $-1 \leq \rho_k \leq 1$. Further, a weakly stationary series x_t is not serially correlated if and only if $\rho_k = 0$ for all $k > 0$. Again, we let $\{x_t\}_{t=1}^T$ be a given sample of X, and estimate the autocorrelation coefficient at lag k with

$$\hat{\rho}(k) = \frac{\frac{1}{T-k-1}\sum_{t=k+1}^{T}(x_t - \overline{x})(x_{t-k} - \overline{x})}{\frac{1}{T-1}\sum_{t=1}^{T}(x_t - \overline{x})^2}, 0 \leq k < T-1$$

where $\overline{x}$ is the sample mean. If $\{x_t\}$ is an iid sequence satisfying $E[x_t^2] < \infty$, then $\hat{\rho}(k)$ is asymptotically normal with mean zero and variance $\frac{1}{T}$ for any fixed positive integer k (see Brockwell et al. [1991]). For finite samples, $\hat{\rho}(k)$ is a biased estimator of $\rho(k)$. For significantly large amount of historical returns, up to 5% of the sample estimate of the autocorrelation function should fall outside the interval

$$\hat{\rho}(k) \in [-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}]$$

5.1.3 The portmanteau test

Financial applications often require to test jointly that several autocorrelations of $R(t)$ are zero. Box and Pierce [1970] proposed the Portmenteau statistic defined by

$$Q^*(h) = N\sum_{j=1}^{h}\hat{\rho}_j^2$$

as a test statistic for the null hypothesis

$$H_0 : \rho_1 = ... = \rho_h$$

versus the alternative hypothesis

$$H_1 : \rho_i \neq 0 \text{ for some } i \in \{1, ..., h\}$$

where N is the number of observations, h is the largest lag and $\hat{\rho}_j = \hat{\rho}(j)$ is the sample autocorrelation function of lag j of an appropriate time series. Under the assumption that $\{R(t)\}$ is an iid sequence with certain moment conditions, $Q^*(h)$ is asymptotically a chi-squared random variable with h degrees of freedom. The Ljung-Box Q-statistic, proposed by Ljung et al. [1978], modifies the $Q^*(h)$ statistic to increase the power of the test in finite samples. It is defined by

$$Q(h) = N(N+2)\sum_{j=1}^{h}\frac{\hat{\rho}_j^2}{N-j}$$

where the function $\hat{\rho}_1, \hat{\rho}_2, ...\hat{\rho}_h$ is called the sample autocorrelation function (ACF) of x_t. As linear time series model can be characterised by its ACF, linear time series modelling makes use of the sample ACF to capture the linear dynamics of the data. Under the null hypothesis that a times series is not autocorrelated, the Ljung-Box Q-Statistic is distributed as chi-squared with h degrees of freedom. The first order $(k = 1)$ autocorrelation value ρ_1^i of asset i is calculated as

$$\hat{\rho}_1^i = \frac{\sum_{t=2}^{N}(R_i(t) - \overline{R}_i)(R_i(t-1) - \overline{R}_i)}{\sum_{t=1}^{N}(R_i(t) - \overline{R}_i)^2}$$

where $\overline{R}_i$ is the mean for asset i over the period covered. Several values of h are often used, but studies suggest that $h \approx \ln N$ provides better power performance. The Ljung-Box test statistic $LB(i)$ is

$$LB(i) = \frac{N(N+2)}{(N-1)}(\hat{\rho}_1^i)^2$$

where $LB(i)$ is χ^2-distributed with one degree of freedom. Geltner [1991] presented a methodology to deal with return series that have been positively tested for significant autocorrelation. He unsmoothed the observed returns to create a new time series which is more volatile and whose characteristics are believed to better depict the true underlying values. The following filter is used

$$\hat{R}_i(t) = \frac{R_i(t) - \rho R_i(t-1)}{(1-\rho)}$$

A time series x_t is called white noise if $\{x_t\}$ is a sequence of independent and identically distributed (iid) random variables with finite mean and variance. If x_t is normally distributed with mean zero and variance σ^2, the series is called a Gaussian white noise. For a white noise series, all the ACFs are zero, so that if all ACFs are close to zero, the series is assumed to be a white noise series.

5.2 Linear time series

5.2.1 Defining time series

We consider a univariate time series x_t observed at equally spaced time intervals and let $\{x_t\}_{t=1}^T$ be the observations where T is the sample size. A purely stochastic time series x_t is said to be linear if it can be written as

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}$$

where μ is a constant, the weights ψ_i are real numbers with $\psi_0 = 1$, and $\{a_t\}$ is a sequence of independent and identically distributed (iid) random variables with a well defined distribution function. We assume that the distribution of a_t is continuous and $E[a_t] = 0$. We can also assume that $Var(a_t) = \sigma^2$ or directly that a_t is Gaussian. Note, if $\sigma^2 \sum_{i=0}^{\infty} \psi_i^2 < \infty$ then x_t is weakly stationary (the first two moments of x_t are time-invariant). In that case, we can obtain its mean and variance by using the independence of $\{a_t\}$ as

$$E[x_t] = \mu \, , Var(x_t) = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2$$

Furthermore, the lag-k autocovariance of x_t is

$$\gamma_k = Cov(x_t, x_{t-k}) = E[\big(\sum_{i=0}^{\infty} \psi_i a_{t-i}\big)\big(\sum_{j=0}^{\infty} \psi_j a_{t-k-j}\big)] = E[\sum_{i,j=0}^{\infty} \psi_i \psi_j a_{t-i} a_{t-k-j}] = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}$$

so that the weights ψ_i are related to the autocorrelations of x_t as follows

$$\rho_k = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+k}}{1 + \sum_{i=1}^{\infty} \psi_i^2} \, , k \geq 0$$

where $\psi_0 = 1$. In general, a purely stochastic time series for x_t is a function of an iid sequence consisting of the current and past shocks written as

$$x_t = f(a_t, a_{t-1}, ..)$$

In the linear model, the function $f(.)$ is a linear function of its arguments, while any nonlinearity in $f(.)$ results in a nonlinear model. Writting the conditional mean and variance of x_t as

$$\mu_t = E[x_t|\mathcal{F}_{t-1}] = g(\mathcal{F}_{t-1}) \, , \sigma_t^2 = Var(x_t|\mathcal{F}_{t-1}) = h(\mathcal{F}_{t-1})$$

where $g(.)$ and $h(.)$ are well defined functions with $h(.) > 0$, then we get the decomposition

$$x_t = g(\mathcal{F}_{t-1}) + \sqrt{h(\mathcal{F}_{t-1})}\epsilon_t$$

where $\epsilon_t = \frac{a_t}{\sigma_t}$ is a standardised shock. In the linear model, $g(.)$ is a linear function of elements of $\mathcal{F}_{t-1}$ and $h(.) = \sigma^2$. Nonlinear models involves making extensions such that if $g(.)$ is nonlinear, x_t is nonlinear in mean, and if $h(.)$ is time-variant, then x_t is nonlinear in variance.

5.2.2 The autoregressive models

5.2.2.1 Definition

Letting $x_t = r_t$ and observing that monthly returns of equity index has a statistically significant lag-1 autocorrelation indicates that the lagged return r_{t-1} may be useful in predicting r_t. A simple autoregressive (AR) model designed to use such a predictive power is

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t$$

where $\{a_t\}$ is a white noise series with mean zero and variance σ^2. This is an $AR(1)$ model where r_t is the dependent variable and r_{t-1} is the explanatory variable. In this model, conditional on the past return r_{t-1}, we have

$$E[r_t|r_{t-1}] = \phi_0 + \phi_1 r_{t-1}\ , Var(r_t|r_{t-1}) = Var(a_t) = \sigma^2$$

This is a Markov property such that conditional on r_{t-1}, the return r_t is not correlated with r_{t-i} for $i > 1$. In the case where r_{t-1} alone can not determine the conditional expectation of r_t, we can use a generalisation of the $AR(1)$ model called $AR(p)$ model defined as

$$r_t = \phi_0 + \phi_1 r_{t-1} + ... + \phi_p r_{t-p} + a_t$$

where p is a non-negative integer. In that model, the past p values $\{r_{t-i}\}_{i=1,..,p}$ jointly determine the conditional expectation of r_t given the past data. The $AR(p)$ model is in the same form as a multiple linear regression model with lagged values serving as explanatory variables.

5.2.2.2 Some properties

Given the conditional mean of the $AR(1)$ model in Section (5.2.2.1), under the stationarity condition we get $E[r_t] = E[r_{t-1}] = \mu$, so that

$$\mu = \phi_0 + \phi_1\mu \text{ or } E[r_t] = \frac{\phi_0}{1-\phi_1}$$

As a result, the mean of r_t exists if $\phi_1 \neq 1$, and it is zero if and only if $\phi_0 = 0$, implying that the term ϕ_0 is related to the mean of r_t. Further, using $\phi_0 = (1-\phi_1)\mu$, we can rewrite the $AR(1)$ model as

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + a_t$$

By repeated substitutions, the prior equation implies

$$r_t - \mu = a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + ... = \sum_{i=0}^{\infty} \phi_1^i a_{t-i}$$

such that $r_t - \mu$ is a linear function of a_{t-i} for $i \geq 0$. From independence of the series $\{a_t\}$, we get $E[(r_t - \mu)a_{t+1}] = 0$, and by the stationarity assumption we get $Cov(r_{t-1}, a_t) = E[(r_{t-1} - \mu)a_t] = 0$. Taking the square, we obtain

$$Var(r_t) = \phi_1^2 Var(r_{t-1}) + \sigma^2$$

since the covariance between r_{t-1} and a_t is zero. Under the stationarity assumption $Var(r_t) = Var(r_{t-1})$, we get

$$Var(r_t) = \frac{\sigma^2}{1 - \phi_1^2}$$

provided that $\phi_1^2 < 1$ which results from the fact that the variance of a random variable is bounded and non-negative. One can show that the $AR(1)$ model is weakly stationary if $|\phi_1| < 1$. Multiplying the equation for $r_t - \mu$ above by a_t, and taking the expectation we get

$$E[a_t(r_t - \mu)] = E[a_t(r_{t-1} - \mu)] + E[a_t^2] = E[a_t^2] = \sigma^2$$

where σ^2 is the variance of a_t. Repeating the process for $(r_{t-k} - \mu)$ and using the prior result, we get

$$\gamma_k \begin{cases} \phi_1\gamma_1 + \sigma^2 \text{ if } k = 0 \\ \phi_1\gamma_{k-1} \text{ if } k > 0 \end{cases}$$

where we use $\gamma_k = \gamma_{-k}$. As a result, we get

$$Var(r_t) = \gamma_0 = \frac{\sigma^2}{1 - \phi_1^2} \text{ and } \gamma_k = \phi_1\gamma_{k-1} \text{ for } k > 0$$

Consequently, the ACF of r_t satisfies

$$\rho_k = \phi_1\rho_{k-1} \text{ for } k \geq 0$$

Since $\rho_0 = 1$, we have $\rho_k = \phi_1^k$ stating that the ACF of a weakly stationary $AR(1)$ series decays exponentially with rate ϕ_1 and starting value $\rho_0 = 1$. Setting $p = 2$ in the $AR(p)$ model and repeating the same technique as that of the $AR(1)$ model, we get

$$E[r_t] = \mu = \frac{\phi_0}{1 - \phi_1 - \phi_2}$$

provided $\phi_1 + \phi_1 \neq 1$. Further, we get

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} \text{ for } k > 0$$

called the moment equation of a stationary $AR(2)$ model. Dividing this equation by γ_0, we get

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} \text{ for } k > 0$$

for the ACF of r_t. It satisfies the second order difference equation

$$\left(1 - \phi_1 B - \phi_2 B^2\right)\rho_k = 0$$

where B is the back-shift operator $B\rho_k = \rho_{k-1}$. Corresponding to the prior difference equation, we can solve a second order polynomial equation leading to the characteristic roots ω_i for $i = 1, 2$. In the case of an $AR(p)$ model, the mean of a stationary series satisfies

$$E[r_t] = \frac{\phi_0}{1 - \phi_1 - ... - \phi_p}$$

provided that the denominator is not zero. The associated polynomial equation of the model is

$$x^p - \phi_1 x^{p-1} - ... - \phi_p = 0$$

so that the series r_t is stationary if all the characteristic roots of this equation are less than one in modulus. For a stationary $AR(p)$ series, the ACF satisfies the difference equation

$$\left(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p\right)\rho_k = 0$$

5.2.2.3 Identifying and estimating AR models

The order determination of AR models consists in specifying empirically the unknown order p of the time series. We briefly discuss two approaches, one using the partial autocorrelation function (PACF), and the other one using some information criterion function. Different approaches to order determination may result in different choices for p, and there is no evidence suggesting that one approach is better than another one in real application.

The partial autocorrelation function One can introduce PACF by considering AR models expressed in the form of a multiple linear regression and arranged in a sequential order, which enable us to apply the idea of partial F test in multiple linear regression analysis

$$\begin{aligned} r_t &= \phi_{0,1} + \phi_{1,1} r_{t-1} + \epsilon_{1t} \\ r_t &= \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-2} + \epsilon_{2t} \\ r_t &= \phi_{0,3} + \phi_{1,3} r_{t-1} + \phi_{2,3} r_{t-2} + \phi_{3,3} r_{t-3} + \epsilon_{3t} \\ ... &= ... \end{aligned}$$

where $\phi_{0,j}$, $\phi_{i,j}$, and $\{\epsilon_{jt}\}$ are the constant term, the coefficient of r_{t-i}, and the error term of an $AR(j)$ model. The estimate $\hat{\phi}_{j,j}$ of the jth equation is called the lag-j sample PACF of r_t. The lag-j PACF shows the added contribution of r_{t-j} to r_t over an $AR(j-1)$ model, and so on. Therefore, for an $AR(p)$ model, the lag-p sample PACF should not be zero, but $\hat{\phi}_{j,j}$ should be close to zero for all $j > p$. Under some regularity conditions, it can be shown that the sample PACF of an $AR(p)$ model has the following properties

- $\hat{\phi}_{p,p}$ converges to ϕ_p as the sample size T goes to infinity.
- $\hat{\phi}_{k,k}$ converges to zero for all $k > p$.
- the asymptotic variance of $\hat{\phi}_{k,k}$ is $\frac{1}{T}$ for $k > T$.

That is, for an $AR(p)$ series, the sample PACF cuts off at lag p.

The information criteria All information criteria available to determine the order p of an AR process are likelihood based. For example, the Akaike Information Criterion (AIC), proposed by Akaike [1973], is defined as

$$AIC = \frac{-2}{T} \ln(\text{likelihood}) + \frac{2}{T}(\text{number of parameters})$$

where the likelihood function is evaluated at the maximum likelihood estimates, and T is the sample size. The second term of the equation above is called the penalty function of the criterion because it penalises a candidate model by the number of parameters used. Different penalty functions result in different information criteria. For a Gaussian $AR(p)$ model, the AIC simplifies to

$$AIC(k) = \ln(\hat{\sigma}_k^2) + \frac{2k}{T}$$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate of σ^2 the variance of a_t. In practice, one computes $AIC(k)$ for $k = 0, ..., P$ where P is a prespecified positive integer, and then selects the order p^* that has the minimum AIC value.

5.2.2.4 Parameter estimation

One usually use the conditional least squares methods when estimating the parameters of an $AR(p)$ model, which starts with the $(p+1)$th observation. Conditioning on the first p observations, we have

$$r_t = \phi_0 + \phi_1 r_{t-1} + ... + \phi_p r_{t-p} + a_t \, , t = p+1, .., T$$

which can be estimated by the least squares method (see details in Section (3.2.4.2)). Denotting $\hat{\phi}_i$ the estimate of ϕ_i, the fitted model is

$$\hat{r}_t = \hat{\phi}_0 + \hat{\phi}_1 r_{t-1} + ... + \hat{\phi}_p r_{t-p}$$

and the associated residual is

$$\hat{a}_t = r_t - \hat{r}_t$$

The series $\{\hat{a}_t\}$ is called the residual series, from which we obtain

$$\hat{\sigma}^2 = \frac{1}{T-2p-1} \sum_{t=p+1}^{T} \hat{a}_t^2$$

If the model is adequate, the residual series should behave as a white noise. If a fitted model is inadequate, it must be refined. The ACF and the Ljung-Box statistics of the residuals can be used to check the closeness of $\hat{a}_t$ to a white noise. In the case of an $AR(p)$ model, the Ljung-Box statistic $Q(h)$ follows asymptotically a chi-squared distribution with $(h-p)$ degrees of freedom.

5.2.3 The moving-average models

One can consider another simple linear model called moving average (MA) which can be treated either as a simple extension of white noise series, or as an infinite order AR model with some parameter constraints. In the latter, we can write

$$r_t = \phi_0 - \theta_1 r_{t-1} - \theta_1^2 r_{t-2} - ... + a_t$$

where the coefficients depend on a single parameter θ_1 via $\phi_i = -\theta_1^i$ for $i \geq 1$. To get stationarity, we must have $|\theta_1| < 1$ since $\theta_1^i \to 0$ as $i \to \infty$. Thus, the contribution of r_{t-i} to r_t decays exponentially as i increases. Writing the above equation in compact form we get

$$r_t + \theta_1 r_{t-1} + \theta_1^2 r_{t-2} + ... = \phi_0 + a_t$$

repeating the process for r_{t-1}, multiplying by θ_1 and subtracting the result from the equation for r_t, we obtain

$$r_t = \phi_0(1-\theta_1) + a_t - \theta_1 a_{t-1}$$

which is a weighted average of shocks a_t and a_{t-1}. This is the $MA(1)$ model with $c_0 = \phi_0(1-\theta_1)$. The $MA(q)$ model is

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q}$$

where $q > 0$. MA models are always weakly stationary since they are finite linear combinations of a white noise sequence for which the first two moments are time-invariant. For example, in the $MA(1)$ model, $E[r_t] = c_0$ and

$$Var(r_t) = \sigma^2 + \theta_1^2 \sigma^2 = (1+\theta_1^2)\sigma^2$$

where a_t and a_{t-1} are uncorrelated. In the $MA(q)$ model, $E[r_t] = c_0$ and

$$Var(r_t) = \big(1 + \theta_1^2 + ... + \theta_q^2\big)\sigma^2$$

Assuming $C_0 = 0$ for an $MA(1)$ model, multiplying the model by r_{t-k} we get

$$r_{t-k}r_t = r_{t-k}a_t - \theta_1 r_{t-k}a_{t-1}$$

and taking expectation

$$\gamma_1 = -\theta_1\sigma^2 \text{ and } \gamma_k = 0 \text{ for } k > 1$$

Given the variance above, we have

$$\rho_0 = 1\ ,\ \rho_1 = \frac{-\theta_1}{1+\theta_1^2}\ ,\ \rho_k = 0 \text{ for } k > 1$$

Thus, the ACF of an $MA(1)$ model cuts off at lag 1. This property generalises to other MA models, so that an $MA(q)$ series is only linearly related to its first q lagged values. Hence, it is a finite-memory model. The maximum likelihood estimation is commonly used to estimate MA models.

5.2.4 The simple ARMA model

We described in Section (5.2.2) the autoregressive models and in Section (5.2.3) the moving-average models. We can then combine the AR and MA models in a compact form so that the number of parameters used is kept small. The general $ARMA(p,q)$ model (see Box et al. [1994]) is defined as

$$y_t = \phi_0 + \sum_{i=1}^{p}\phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q}\theta_i\epsilon_{t-i}$$

where $\{\epsilon_t\}$ is a white noise series and p and q are non-negative integers. Using the back-shift operator, we can rewrite the model as

$$\big(1 - \phi_1 B - ... - \phi_p B^p\big)y_t = \big(1 - \theta_1 B - ... - \theta_q B^q\big)\epsilon_t$$

and we require that there are no common factors between the AR and MA polynomials, so that the order (p,q) of the model can not be reduced. Note, the AR polynomial introduces the characterisic equation of an ARMA model. Hence, if all the solutions of the characteristic equation are less than 1 in absolute value, then the ARMA model is weakly stationary. In this case, the unconditional mean of the model is

$$E[y_t] = \frac{\phi_0}{1 - \phi_1 - ... - \phi_p}$$

Both the ACF and PACF are not informative in determining the order of an ARMA model, but one can use the extended autocorrelation function (EACF) to specify the order of an ARMA process(see Tsay et al. [1984]). It states that if we can obtain a consistent estimate of the AR component of an ARMA model, then we can derive the MA component and use ACF to identify its order. The output of EACF is a two-way table where the rows correspond to AR order p and the columns to MA order q. Once an $ARMA(p,q)$ model is specified, its parameters can be estimated by either the conditional or exact likelihood method. Then the Ljung-Box statistics of the residuals can be used to check the adequacy of the fitted model. In the case where the model is correctly specified, then $Q(h)$ follows asymptotically a chi-squared distribution with $(h-g)$ degrees of freedom, where g is the number of parameters used.

The representation of the $ARMA(p,q)$ model using the back-shift operator is compact and useful in parameter estimation. However, other representations exist using the long division of two polynomials. That is, given two polynomials $\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$ and $\theta(B) = 1 - \sum_{i=1}^{q}\theta_i B^i$, we get by long division

$$\begin{aligned}\frac{\theta(B)}{\phi(B)} &= 1+\psi_1 B+\psi_2 B^2+...=\psi(B)\\ \frac{\phi(B)}{\theta(B)} &= 1-\pi_1 B-\pi_2 B^2-...=\pi(B)\end{aligned}$$

From the definition, $\psi(B)\pi(B)=1$, and making use of the fact that $Bc=c$ for any constant, we have

$$\frac{\phi_0}{\theta(1)}=\frac{\phi_0}{1-\theta_1-...-\theta_q} \text{ and } \frac{\phi_0}{\phi(1)}=\frac{\phi_0}{1-\phi_1-...-\phi_q}$$

Using the results above, the $ARMA(p,q)$ model can be written as an AR model

$$y_t=\frac{\phi_0}{1-\theta_1-...-\theta_q}+\pi_1 y_{t-1}+\pi_2 y_{t-2}+...+\epsilon_t$$

showing the dependence of the current value y_t to the past values y_{t-i} for $i>0$. To show that the contribution of the lagged value y_{t-i} to y_t is diminishing as i increases, the π_i coefficient should decay to zero as i increases. An $ARMA(p,q)$ model having this property is invertible. A sufficient condition for invertibility is that all the zeros of the polynomial $\theta(B)$ are greater than unity in modulus. Using the AR representation, an invertible $ARMA(p,q)$ series y_t is a linear combination of the current shock ϵ_t and a weighted average of the past values, with weights decaying exponentially. Similarly, the $ARMA(p,q)$ model can also be written as an MA model

$$y_t=\mu+\epsilon_t+\psi_1\epsilon_{t-1}+\psi_2\epsilon_{t-2}+...=\mu\psi(B)\epsilon_t$$

where $\mu=E[y_t]=\frac{\phi_0}{1-\phi_1-...-\phi_p}$. It shows explicitly the impact of the past shock ϵ_{t-i}, $i>0$, on the current value y_t. The coefficients $\{\psi_i\}$ are called the impulse response function of the ARMA model. For a weakly stationary series, the ψ_i decay exponentially as i increases. The MA representation provides a simple proof of mean reversion of a stationary time series, since the speed at which $\hat{y}_{t+k|t}$ approaches μ determines the speed of mean reversion.

5.3 Forecasting

The ultimate objective of using a stochastic process to model time series is for forecasting. In the construction of forecasts the idea of conditional expectations is very important. The conditional expectation has the property of being the minimum mean square error (MMSE) forecast. This means that if the model specification is correct, there is no other forecast which will have errors whose squares have a lower expected value

$$\hat{y}_{t+k|t}=E[y_{t+k}|\mathcal{F}_t]$$

with forecast horizon k. Therefore given a set of observations, the optimal predictor k steps ahead is the expected value of y_{t+k} conditional on the information at time t. The predictor k is said to be optimal because it has minimum mean square error. In order to prove this statement, the forecasting error can be split as

$$y_{t+k}-\hat{y}_{t+k|t}=\big(y_{t+k}-E[y_{t+k|t}]\big)+\big(E[y_{t+k|t}]-\hat{y}_{t+k|t}\big)$$

and it follows that

$$MSE(\hat{y}_{t+k|t})=Var(\hat{y}_{t+k|t})+\big(\hat{y}_{t+k}-E[\hat{y}_{t+k|t}]\big)^2$$

The conditional variance of y_{t+k} does not depend on $\hat{y}_{t+k|t}$, and therefore, the MMSE of y_{t+k} is given by the conditional mean.

5.3.1 Forecasting with the AR models

For the $AR(p)$ model, with forecast origin t, and forecast horizon k, we let $\hat{y}_{t+k|t}$ be the forecast of y_{t+k} using the minimum squared error loss function. That is, $\hat{y}_{t+k|t}$ is chosen such that

$$E[y_{t+k} - \hat{y}_{t+k|t}] \leq \min_g E[(y_{t+k} - g)^2]$$

where g is a function of the information available at time t (inclusive). In the $AR(p)$ model, for the 1-step ahead forecast, we have

$$y_{t+1} = \phi_0 + \phi_1 y_t + ... + \phi_p y_{t+1-p} + \epsilon_{t+1}$$

where y_t corresponds to r_t and ϵ_t to a_t when considering returns. Under the minimum squared error loss function, the point forecast of y_{t+1}, given the model and observations up to time t, is the conditional expectation

$$\hat{y}_{t+1|t} = E[y_{t+1}|y_t, y_{t-1}, ...] = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t+1-i}$$

and the associeated forecast error is

$$e_{t+1|t} = y_{t+1} - \hat{y}_{t+1|t} = \epsilon_{t+1}$$

The variance of the 1-step ahead forecast error is $Var(e_{t+1}|t) = Var(\epsilon_{t+1}) = \sigma^2$. Hence, if ϵ_t is normally distributed, then a 95% 1-step ahead interval forecast of y_{t+1} is

$$\hat{y}_{t+1|t} \pm 1.96 \times \sigma$$

Note, ϵ_{t+1} is the 1-step ahead forecast error at the forecast origin t, and it is referred to as the shock of the series at time $t+1$. Further, in practice, estimated parameters are often used to compute point and interval forecast, resulting in a Conditional Forecast since it does not take into consideration the uncertainty in the parameter estimates. Considering parameter uncertainty is a much more involved process. When the sample size used in estimation is sufficiently large, then the conditional forecast is close to the unconditional one. In the general case, we have

$$y_{t+k} = \phi_0 + \phi_1 y_{t+k-1} + ... + \phi_p y_{t+k-p} + \epsilon_{t+k}$$

The k-step ahead forecast based on the minimum squared error loss function is the conditional expectation of y_{t+k} given $\{y_{t-i}\}_{i=0}^{\infty}$ obtained as

$$\hat{y}_{t+k|t} = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{y}_{t+k-i|t}$$

where $\hat{y}_{t+i|t} = y_{t+i}$ for $i \leq 0$. This forecast can be computed recursively using forecast $\hat{y}_{t+i|t}$ for $i = 1, .., k-1$. The k-step ahead forecast error is $e_{t+k|t} = y_{t+k} - \hat{y}_{t+k|t}$. It can be shown that for a stationary $AR(p)$ model, $\hat{y}_{t+k|t}$ converges to $E[y_t]$ as $k \to \infty$, meaning that for such a series, long-term point forecast approaches its unconditional mean. This property is called mean-reversion in the finance literature. The variance of the forecast error approaches the unconditional variance of y_t.

5.3.2 Forecasting with the MA models

Since the MA model has finite memory, its point forecasts go to the mean of the series very quickly. For the 1-step ahead forecast of an $MA(1)$ process at the forecast origin t, we get

$$y_{t+1} = c_0 + \epsilon_{t+1} - \theta_1 \epsilon_t$$

and taking conditional expectation, we have

$$\begin{aligned} \hat{y}_{t+1|t} &= E[y_{t+1}|y_y, y_{t-1}, ..] = c_0 - \theta_1 \epsilon_t \\ e_{t+1|t} &= y_{t+1} - \hat{y}_{t+1|t} = \epsilon_{t+1} \end{aligned}$$

The variance of the 1-step ahead forecast error is $Var(e_{t+1|t}) = \sigma^2$. To compute ϵ_t one can assume that $\epsilon_0 = 0$ and get $\epsilon_1 = y_1 - c_0$, and then compute ϵ_h for $2 \leq h \leq t$ recursively by using $\epsilon_h = y_h - c_0 + \theta_1 \epsilon_{h-1}$. For the 2-step ahead forecast we get $\hat{y}_{t+2|t} = c_0$ and the variance of the forecast error is $Var(e_{t+2|t}) = (1 + \theta_1)^2 \sigma^2$, so that the 2-step ahead forecast of the series is simply the unconditional mean of the model. More generally $\hat{y}_{t+k|t} = c_0$ for $k \geq 2$. Hence, the forecast $\hat{y}_{t+k|t}$ versus k form a horizontal line on a plot after one step. In general, for an $MA(q)$ model, multistep ahead forecasts go to the mean after the first q steps.

5.3.3 Forecasting with the ARMA models

The forecasts of an $ARMA(p, q)$ model have similar characteristics to those of an $AR(p)$ model, after adjusting for the impacts of the MA component on the lower horizon forecasts. For the 1-step ahead forecast of y_{t+1}, with forecast origin t, we have

$$\hat{y}_{t+1|t} = E[y_{t+1}|y_t, y_{t-1}, ...] = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t+1-i} - \sum_{i=1}^{q} \theta_i \epsilon_{t+1-i}$$

and the associated forecast error is $e_{t+1|t} = y_{t+1} - \hat{y}_{t+1|t} = \epsilon_{t+1}$. The variance of the 1-step ahead error is $Var(e_{t+1|t}) = \sigma^2$. For the k-step ahead forecast of $y_{t+k|t}$, with forecast origin t, we have

$$\hat{y}_{t+k|t} = E[y_{t+k}|y_t, y_{t-1}, ...] = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{y}_{t+k-i|t} - \sum_{i=1}^{q} \theta_i \hat{\epsilon}_{t+k-i|t}$$

where $\hat{y}_{t+k-i|t} = y_{t+k-i}$ if $k - i \leq 0$ and

$$\hat{\epsilon}_{t+k-i|t} = \begin{cases} 0 \text{ if } k - i > 0 \\ \epsilon_{t+k-i} \text{ if } k - i \leq 0 \end{cases}$$

Thus, the multi-step ahead forecasts of an ARMA model can be computed recursively, and the associated forecast error is

$$e_{t+k|t} = y_{t+k} - \hat{y}_{t+k|t}$$

5.4 Nonstationarity and serial correlation

5.4.1 Unit-root nonstationarity

When modelling equity stocks, interest rates, or foreign excange rates, the two most considered models for characterising their non-stationarity have been the random walk model with a drift

$$y_t = \mu + \phi y_{t-1} + \epsilon_t \tag{5.4.1}$$

and the trend-stationary process

$$y_t = \alpha + \beta t + \epsilon_t \tag{5.4.2}$$

where $\{\epsilon_t\}$ is a white noise series. In the random walk model the value of ϕ can have different effects on the stock process

1. $\phi < 1 \to \phi^T \to 0$ as $T \to \infty$
2. $\phi = 1 \to \phi^T = 1$ for all T
3. $\phi > 1$

In case (1) the shocks in the system gradually die away which is called the stationarity case. In case (2) the shocks persist in the system and do not die away, leading to

$$y_t = y_0 + \sum_{t=0}^{\infty} \epsilon_t \text{ as } T \to \infty$$

That is, the current value of y is an infinite sum of the past shocks added to the starting value of y. This case is known as the unit root case. In the case (3) the shocks become more influential as time goes on.

5.4.1.1 The random walk

A time series $\{y_t\}$ is a random walk if it satisfies

$$y_t = y_{t-1} + \epsilon_t$$

where the real number y_0 is the starting value of the process and $\{\epsilon_t\}$ is a white noise series. The random walk is a special $AR(1)$ model with coefficient ϕ_1 of y_{t-1} being equal to unity, so that it does not satisfies the weak stationarity of an $AR(1)$ model. Hence, we call it a unit-root nonstationarity time series. Under such a model, the stock price is not predictable or mean reverting. The 1-step ahead forecast of the random walk at the origin t is

$$\hat{y}_{t+1|t} = E[y_{t+1}|y_t, y_{t-1}, ...] = y_t$$

which is the value at the forecast origin, and thus, has no practical value. For any forecast horizon $k > 0$ we have

$$\hat{y}_{t+k|t} = y_t$$

so that point forecasts of a random walk model are simply the value of the series at the forecast origin, and the process is not mean-reverting. The MA representation of the random walk is

$$y_t = \sum_{i=0}^{\infty} \epsilon_{t-i}$$

and the k-step ahead forecast error is

$$e_{t+k|t} = \epsilon_{t+k} + ... + \epsilon_{t+1}$$

with $Var(e_{t+k|t}) = k\sigma^2$, which diverges to infinity as $k \to \infty$. Hence, the usefulness of point forecast $\hat{y}_{t+k|t}$ diminishes as k increases, implying that the model is not predictable. Further, as the variance of the forecast error approaches infinity when k increases, the unconditional variance of y_t is unbounded, meaning that it can take any real value for sufficiently large t which is questionable for indexes. At last, since $\psi_i = 1$ for all i, then the impact of any past shock ϵ_{t-i} on y_t does not decay over time, and the series has a strong memory as it remembers all of the past shocks. That is, the shocks have a permanent effect on the series.

5.4.1.2 The random walk with drift

When a time series experience a small and positive mean, we can consider a random walk with drift

$$y_t = \mu + y_{t-1} + \epsilon_t$$

where $\mu = E[y_t - y_{t-1}]$ is the time-trend of y_t, or drift of the model, and $\{\epsilon_t\}$ is a white noise series. Assuming initial value y_0, we can rewrite the model as

$$y_t = t\mu + y_0 + \epsilon_t + \epsilon_{t-1} + ... + \epsilon_1$$

consisting of the time-trend $t\mu$ and a pure random walk process $\sum_{i=1}^t \epsilon_i$. Also, since $Var(\sum_{i=1}^t \epsilon_i) = t\sigma^2$, the conditional standard deviation of y_t is $\sqrt{t}\sigma$ which grows at a slower rate than the conditional expectation of y_t. Plotting y_t against the time index t, we get a time-trend with slop equal to μ. We can analyse the constant term in the series by noting that for an $MA(q)$ model the constant term is the mean of the series. In the case of a stationary $AR(p)$ model or $ARMA(p,q)$ model, the constant term is related to the mean via

$$\mu = \frac{\phi_0}{1 - \sum_{i=1}^p \phi_i}$$

These differences in interpreting the constant term reflects the difference between the dynamic and linear regression models. In the general case, allowing the AR polynomial to have 1 as a characteristic root, we get the autoregressive integrated moving average $ARIMA$ model which is unit-root nonstationary because its AR polynomial has a unit root. An $ARIMA$ model has a strong memory because the ψ_i coefficients in its MA representation do not decay over time to zero, so that the past shocks ϵ_{t-i} of the model has a permanent effect on the series. A conventional approach for handling unit root nonstationarity is to use differencing.

5.4.1.3 The unit-root test

To test whether the value y_t follows a random walk or a random walk with a drift, the unit-root testing problem (see Dickey et al. [1979]) employs the models

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \epsilon_t \\ y_t &= \phi_0 + \phi_1 y_{t-1} + \epsilon_t \end{aligned}$$

where ϵ_t denotes the error term, and consider the null hypothesis

$$H_0 : \phi_1 = 1$$

versus the alternative hypothesis

$$H_1 : \phi_1 < 1$$

A convenient test statistic is the t ratio of the least squares (LS) estimate of ϕ_1 under the null hypothesis. The LS method fot the first equation above gives

$$\hat{\phi}_1 = \frac{\sum_{t=1}^T y_{t-1} y_t}{\sum_{t=1}^T y_{t-1}^2}, \hat{\sigma}_\epsilon^2 = \frac{1}{T-1}\sum_{t=1}^T (y_t - \hat{\phi}_1 y_{t-1})^2$$

where $y_0 = 0$ and T is the sample size. The t ratio is

$$DF = \text{t-ratio} = \frac{\hat{\phi}_1 - 1}{std(\hat{\phi}_1)} = \frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\hat{\sigma}_\epsilon \sqrt{\sum_{t=1}^T y_{t-1}^2}}$$

which is referred to as the Dickey-Fuller test. If $\{\epsilon_t\}$ is a white noise series with finite moments of order slightly greater than 2, then the DF-statistic converge to a function of the standard Brownian motion as $T \to \infty$ (see Chan et al. [1988]). If $\phi_0 = 0$, but the model is still used, the resulting t ratio for testing $\phi_1 = 1$ will converge to another non-standard asymptotic distribution. If $\phi_0 \neq 0$, and the model is still used, the t ratio for testing $\phi_1 = 1$ is asymptotically normal, but large sample sizes is required.

5.4.2 Regression models with time series

The relationship between two time series is of major interest, for instance the Market model in finance relates the return of an individual stock to the return of a market index. In general, we consider the linear regression

$$r_{1t} = \alpha + \beta r_{2t} + \epsilon_t$$

where r_{it} for $i = 1, 2$ are two time series and ϵ_t is the error term. The least squares (LS) method is often used to estimate the model parameters (see details in Section (3.2.4.2)). If $\{\epsilon_t\}$ is a white noise series, then the least square method (LS) produces consistent estimates. However, in practice the error term $\{\epsilon_t\}$ is often serially correlated, so that we get a regression model with time series errors. Even though this approach is widely used in finance, it is a misused econometric model when the serial dependence in ϵ_t is overlooked. On can look at the time plot and ACF of the two series residuals to detect patterns of a unit-root nonstationarity time series. When two time series are unit-root nonstationary, the behaviour of the residuals indicates that the series are not co-integrated. In that case the data fail to support the hypothesis that there exists a long-term equilibrium between the two series. One way forward for building a linear regression model with time series errors, is to use a simple time series model for the residual series and estimate the whole model jointly. For example, considering the modified series

$$\begin{aligned} c_{1t} &= r_{1t} - r_{1,t-1} = (1-B)r_{1t} \text{ for } t \geq 2 \\ c_{2t} &= r_{2t} - r_{2,t-1} = (1-B)r_{2t} \text{ for } t \geq 2 \end{aligned}$$

we can specify a $MA(1)$ model for the residuals and modify the linear regression model to get

$$c_{2t} = \alpha + \beta c_{1t} + \epsilon_t \text{ , } \epsilon_t = a_t - \theta_1 a_{t-1}$$

where $\{a_t\}$ is assumed to be a white noise series. The $MA(1)$ model is used to capture the serial dependence in the error term. More complex time series models can be added to a linear regression equation to form a general regression model with time series error. One can consider the Cochran-Orcutt estimator to handle the serial dependence in the residuals (see Greene [2000]). When the time series model used is stationary and invertible, one can estimate the model jointly by using the maximum likelihood method (MLM). Note, one can use the Durbin-Watson (DW) statistic to check residuals for serial correlation, but it only consider the lag-1 serial correlation. For a residual series ϵ_t with T observations, the Durbin-Watson statistic is

$$DW = \frac{\sum_{t+2}^{T}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t+1}^{T}\epsilon_t^2}$$

which is approximated by

$$DW \approx 2(1 - \hat{\rho}_1)$$

where $\hat{\rho}_1$ is the lag-1 ACF of $\{\epsilon_t\}$. When residual serial dependence appears at higher order lags (seasonal behaviour), one can use the Ljung-Box statistics.

5.4.3 Long-memory models

Even though for a stationary time series the ACF decays exponentially to zero as lag increases, for a unit-root non-stationary time series, Tiao et al. [1983] showed that the sample ACF converges to 1 for all fixed lags as the sample size increases. There exist some time series, called long-memory time series, whose ACF decays slowly to zero at a polynomial rate as the lag increases. For instance, Hosking [1981] proposed the fractionally differenced process defined by

$$(1-B)^d x_t = a_t \ , \ -\frac{1}{2} < d < \frac{1}{2}$$

where $\{a_t\}$ is a white noise series. We consider some properties of the model below and refer the readers to Section () for more details.

- if $d < \frac{1}{2}$, then x_t is a weakly stationary process and has the infinite MA representation

$$x_t = a_t + \sum_{i=1}^{\infty} \psi_i a_{t-i}$$

  with

$$\psi_k = \frac{d(1+d)...(k-1+d)}{k!} = \frac{(k+d-1)!}{k!(d-1)!}$$

- if $d > -\frac{1}{2}$, then x_t is invertible and has the infinite AR representation

$$x_t = \sum_{i=1}^{\infty} \pi_i x_{t-i} + a_t$$

  with

$$\pi_k = \frac{-d(1-d)...(k-1-d)}{k!} = \frac{(k-d-1)!}{k!(-d-1)!}$$

- for $-\frac{1}{2} < d < \frac{1}{2}$, the ACF of x_t is

$$\rho_k = \frac{d(1+d)...(k-1+d)}{(1-d)(2-d)...(k-d)} \ , \ k = 1, 2, ...$$

  in particular, $\rho_1 = \frac{d}{1-d}$ and

$$\rho_k \approx \frac{(-d)!}{(d-1)!} k^{2d-1} \text{ as } k \to \infty$$

- for $-\frac{1}{2} < d < \frac{1}{2}$, the PACF of x_t is $\phi_{k,k} = \frac{d}{(k-d)}$ for $k = 1, 2, ..$
- for $-\frac{1}{2} < d < \frac{1}{2}$, the spectral density function $f(w)$ of x_t, which is the Fourier transform of the ACF of x_t, satisfies

$$f(w) \sim w^{-2d} \text{ as } w \to 0$$

  where $w \in [0, 2\pi]$ denotes the frequency.

In the case where $d < \frac{1}{2}$, the property of the ACF of x_t says that $\rho_k \sim k^{2d-1}$, which decays at a polynomial rate rather than an exponential one. Note, in the spectral density above, the spectrum diverges to infinity as $w \to 0$, but it is bounded for all $w \in [0, 2\pi]$ in the case of a stationary ARMA process. If the fractionally differenced series $(1-B)^d x_t$ follows an $ARMA(p, q)$ model, then x_t is called an $ARFIMA(p, d, q)$ process, which is a generalised ARIMA model by allowing for noninteger d. One can estimate d by using either a maximum likelihood method or a regression method with logged periodigram at the lower frequency.

5.5 Multivariate time series

5.5.1 Characteristics

For an investor holding multiple assets, the dynamic relationships between returns of the assets play an important role in the process of decision. Vector or multivariate time series analysis are methods used to study jointly multiple return series. As multivariate time series are made of multiple single series referred as components, vector and matrix are the necessary tools. We let $r_t = (r_{1t}, r_{2t}, .., r_{Nt})^\top$ be the log returns of N assets at time t. The series r_t is weakly stationary if its first two moments are time-invariant. Hence, the mean vector and covariance matrix of a weakly stationary series are constant over time. Assuming weakly stationarity of r_t, the mean vector and covariance matrix are given by

$$\mu = E[r_t] \, , \Gamma_0 = E[(r_t - \mu_t)(r_t - \mu_t)^\top]$$

where the expectation is taken element by element over the joint distribution of r_t. The mean μ is a N-dimensional vector, and the covariance matrix Γ_0 is a $N \times N$ matrix. The ith diagonal element of Γ_0 is the variance of r_{it}, and the (i,j)th element of Γ_0 for $i \neq j$ is the covariance between r_{it} and r_{jt}. We then write $\mu = (\mu_1, .., \mu_N)^\top$ and $\Gamma_0 = [\Gamma_{ij}(0)]$ when describing the elements.

We let $D = diag[\sqrt{\Gamma_{11}(0)}, ..., \sqrt{\Gamma_{NN}(0)}]$ be a $N \times N$ diagonal matrix consisting of the standard deviation of r_{it} for $i = 1, .., N$. The lag-zero cross-correlation matrix of r_t is

$$\rho_0 = [\rho_{ij}(0)] = D^{-1}\Gamma_0 D^{-1}$$

with (i,j)th element being

$$\rho_{ij}(0) = \frac{\Gamma_{ij}(0)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{Cov(r_{it}, r_{jt})}{\sigma(r_{it})\sigma(r_{jt})}$$

and corresponding to the correlation between r_{it} and r_{jt} where $\rho_{ij}(0) = \rho_{ji}(0)$, $-1 \leq \rho_{ij}(0) \leq 1$, and $\rho_{ii}(0) = 1$ for $1 \leq i, j \leq N$. In order to understand the lead-lag relationships between component series, the cross-correlation matrices are used to measure the strength of linear dependence between time series. The lag-k cross-covariance matrix of r_t is defined as

$$\Gamma_k = [\Gamma_{ij}(k)] = E[(r_t - \mu)(r_{t-k} - \mu)^\top]$$

where μ is the mean vector of r_t. For a weakly stationary series, the cross-covariance matrix Γ_k is a function of k, but not the time t. The lag-k cross-correlation matrix (CCM) of r_t is defined as

$$\rho_k = [\rho_{ij}(k)] = D^{-1}\Gamma_k D^{-1}$$

with (i,j)th element being

$$\rho_{ij}(k) = \frac{\Gamma_{ij}(k)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{Cov(r_{it}, r_{j,t-k})}{\sigma(r_{it})\sigma(r_{jt})}$$

and corresponding to the correlation between r_{it} and $r_{j,t-k}$. When $k > 0$ it measures the linear dependence between r_{it} and $r_{j,t-k}$ occurring prior to time t such that when $\rho_{ij}(k) \neq 0$ the series r_{jt} leads the series r_{it} at lag k. This result is reversed for $\rho_{ji}(k)$, and the diagonal element of $\rho_{ii}(k)$ is the lag-k autocorrelation of r_{it}. In general, when $k > 0$ we get $\rho_{ij}(k) \neq \rho_{ji}(k)$ for $i \neq j$ because the two correlation coefficients measure different linear relationships between $\{r_{it}\}$ and $\{r_{jt}\}$, and Γ_k and ρ_k are not symmetric. Further, from $Cov(r_{it}, r_{j,t-k}) = Cov(r_{j,t-k}, r_{it})$ and by the weak stationarity assumption

$$Cov(r_{j,t-k}, r_{it}) = Cov(r_{j,t}, r_{i,t+K}) = Cov(r_{jt}, r_{i,t-(-k)})$$

we have $\Gamma_{ij}(k) = \Gamma_{ji}(-k)$. Since $\Gamma_{ji}(-k)$ is the (j,i)th element of the matrix Γ_{-k}, and since the equality holds for $1 \leq i, j \leq N$, we have $\Gamma_k = \Gamma_{-k}^\top$ and $\rho_k = \rho_{-k}^\top$. Hence, unlike the univariate case, $\rho_k \neq \rho_{-k}$ for a general vector time series when $k > 0$. As a result, it suffices to consider the cross-correlation matrices ρ_k for $k \geq 0$. Given the information contained in the cross-correlation matrices $\{\rho_k\}_{k=0,1,2,..}$ of a weakly stationary vector time series, if $\rho_{ij}(k) = 0$ for all $k > 0$, then r_{it} does not depend linearly on any past value $r_{j,t-k}$ of the r_{jt} series.

Given the data $\{r_t\}_{t=1}^T$, the cross-covariance matrix Γ_k is computed as

$$\hat{\Gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (r_t - \overline{r})(r_{t-k} - \overline{r})^\top \ , k \geq 0$$

where $\overline{r} = \frac{1}{T} \sum_{t=1}^T r_t$ is the vector sample mean. The cross-correlation matrix ρ_k is estimated by

$$\hat{\rho}_k = \hat{D}^{-1} \hat{\Gamma}_k \hat{D}^{-1} \ , k \geq 0$$

where $\hat{D}$ is the $N \times N$ diagonal matrix of the sample standard deviation of the component series. The asymptotic properties of the sample cross-correlation matrix $\hat{\rho}_k$ have been investigated under various assumptions (see Fuller [1976]). For asset return series, the presence of conditional heteroscedasticity and high kurtosis complexify the finite sample distribution of $\hat{\rho}_k$, and proper bootstrap resampling methods should be used to get an approximate estimate of the distribution. The univariate Ljung-Box statistic $Q(m)$ has been generalised to the multivariate case by Hosking [1980] and Li et al. [1981].

5.5.2 Introduction to a few models

The vector autoregressive (VAR) model is considered a simple vector model when modelling asset returns. A multivariate time series r_t is a VAR process of order 1 or $VAR(1)$ if it follows the model

$$r_t = \phi_0 + \Phi r_{t-1} + a_t$$

where ϕ_0 is a N-dimentional vector, Φ is a $N \times N$ matrix, and $\{a_t\}$ is a sequence of serially uncorrelated random vectors with mean zero and positive definite covariance matrix Σ. For example, in the bivariate case with $N = 2$, $r_t = (r_{1t}, r_{2t})^\top$ and $a_t = (a_{1t}, a_{2t})^\top$, we get the equations

$$\begin{aligned} r_{1t} &= \phi_{10} + \Phi_{11} r_{1,t-1} + \Phi_{12} r_{2,t-1} + a_{1t} \\ r_{2t} &= \phi_{20} + \Phi_{21} r_{1,t-1} + \Phi_{22} r_{2,t-1} + a_{2t} \end{aligned}$$

where Φ_{ij} is the (i,j)th element of Φ and ϕ_{i0} is the ith element of ϕ_0. The coefficient matrix Φ measures the dynamic dependence of r_t, and the concurrent relationship between r_{1t} and r_{2t} is shown by the off-diagonal element σ_{12} of the covariance matrix Σ of a_t. The $VAR(1)$ model is called a reduced-form model because it does not show explicitly the concurrent dependence between the component series. Assuming weakly stationarity, and using $E[a_t] = 0$, we get

$$E[r_t] = \phi_0 + \Phi E[r_{t-1}]$$

Since $E[r_t]$ is time-invariant, we get

$$\mu = E[r_t] = (I - \Phi)^{-1} \phi_0$$

provided that $I - \Phi$ is non-singular, where I is the $N \times N$ identity matrix. Hence, using $\phi_0 = (I - \Phi)\mu$, we can rewrite the $AR(1)$ model as

$$(r_t - \mu) = \Phi(r_{t-1} - \mu) + a_t$$

Letting $\tilde{r}_t = r_t - \mu$ be the mean-corrected time series, the $VAR(1)$ model becomes

$$\tilde{r}_t = \Phi \tilde{r}_{t-1} + a_t$$

By repeated substitutions, the $VAR(1)$ model becomes

$$\tilde{r}_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + ...$$

characterising the $VAR(1)$ process. We can generalise the $VAR(1)$ model to $VAR(p)$ models and get their characteristics.

Another approach is the generalise univariate ARMA models to handle vector time series and obtain VARMA models. However, these models suffer from identifiability problem as they are not uniquely defined. Hence, building a VARMA model for a given data set requires some attention.

5.5.3 Principal component analysis

Another important statistic in multivariate time series analysis is the covariance (or correlation) structure of the series. Given a N-dimensional random variable $r = (r_1, .., r_N)^\top$ with covariance matrix Σ_r, a principal component analysis (PCA) is concerned with using a few linear combinations of r_i to explain the structure of Σ_r. PCA applies to either the covariance matrix Σ_r or the correlation matrix ρ_r of r. The correlation matrix being the covariance matrix of the standardised random vector $r^* = D^{-1}r$ where D is the diagonal matrix of standard deviations of the components of r, we apply PCA to the covariance matrix. Let $c_i = (c_{i1}, ..., c_{iN})^\top$ be a N-dimensional vector, where $i = 1, .., N$ such that

$$y_i = c_i^\top r = \sum_{j=1}^{N} c_{ij} r_j$$

is a linear combination of the random vector r. In the case where r consists of the simple returns of N stocks, then y_i is the return of a portfolio assigning weight c_{ij} to the jth stock. Without modifying the proportional allocation of the portfolio, we can standardise the vector c_i so that $c_i^\top c_i = \sum_{j=1}^{N} c_{ij}^2 = 1$. Using the properties of a linear combination of random variables, we get

$$\begin{aligned} Var(y_i) &= c_i^\top \Sigma_r c_i \ , i = 1, .., N \\ Cov(y_i, y_j) &= c_i^\top \Sigma_r c_j \ , i, j = 1, .., N \end{aligned}$$

The idea of PCA is to find linear combinations c_i such that y_i and y_j are uncorrelated for $i \neq j$ and the variances of y_i are as large as possible. Specifically

1. the first principal component of r is the linear combination $y_1 = c_1^\top r$ maximising $Var(y_1)$ under the constraint $c_1^\top c_1 = 1$.
2. the second principal component of r is the linear combination $y_2 = c_2^\top r$ maximising $Var(y_2)$ under the constraint $c_2^\top c_2 = 1$ and $Cov(y_1, y_2) = 0$.
3. the ith principal component of r is the linear combination $y_i = c_i^\top r$ maximising $Var(y_i)$ under the constraint $c_i^\top c_i = 1$ and $Cov(y_i, y_j) = 0$ for $j = 1, ..., i-1$.

Since the covariance matrix Σ_r is non-negative definite, it has a spectral decomposition (see Appendix (A.6)). Hence, letting $(\lambda_1, e_1), ..., (\lambda_N, e_N)$ be the eigenvalue-eigenvector pairs of Σ_r, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$, then the ith principal component of r is $y_i = e_i^\top r = \sum_{j=1}^{N} e_{ij} r_j$ for $i = 1, .., N$. Moreover, we get

$$\begin{aligned} Var(y_i) &= e_i^\top \Sigma_r e_i = \lambda_i \, , i = 1, .., N \\ Cov(y_i, y_j) &= e_i^\top \Sigma_r e_j = 0 \, , i \neq j \end{aligned}$$

In the case where some eigenvalues λ_i are equal, the choices of the corresponding eigenvectors e_i and hence y_i are not unique. Further, we have

$$\sum_{i=1}^{N} Var(r_i) = tr(\Sigma_r) = \sum_{i=1}^{N} \lambda_i = \sum_{i=1}^{N} Var(y_i)$$

This result says that

$$\frac{Var(y_i)}{\sum_{i=1}^{N} Var(r_i)} = \frac{\lambda_i}{\lambda_1 + .. + \lambda_N}$$

so that the proportion of total variance in r explained by the ith principal component is simply the ratio between the ith eigenvalue and the sum of all eigenvalues of Σ_r. Since $tr(\rho_r) = N$, the proportion of variance explained by the ith principal component becomes $\frac{\lambda_i}{N}$ when the correlation matrix is used to perform the PCA. A byproduct of the PCA is that a zero eigenvalue of Σ_r or ρ_r indicates the existence of an exact linear relationship between the components of r. For instance, if the smallest eigenvalue $\lambda_N = 0$, then from the previous result $Var(y_N) = 0$, and therefore $y_N = \sum_{j=1}^{N} e_{Nj} r_j$ is a constant and there are only $N - 1$ random quantities in r, reducing the dimension of r. Hence, PCA has been used as a tool for dimension reduction. In practice, the covariance matrix Σ_r and the correlation matrix ρ_r of the return vector r are unknown, but they can be estimated consistently by the sample covariance and correlation matrices under some regularity conditions. Assuming that the returns $\{r_t\}_{t=1}^{T}$ are weakly stationary, we get the estimates

$$\hat{\Sigma}_r = [\hat{\sigma}_{ij,r}] = \frac{1}{T-1} \sum_{t=1}^{T} (r_t - \overline{r})(r_t - \overline{r})^\top \, , \overline{r} = \frac{1}{T} \sum_{t=1}^{T} r_t$$

and

$$\hat{\rho}_r = \hat{D}^{-1} \hat{\Sigma}_r \hat{D}^{-1}$$

where $\hat{D} = diag\{\sqrt{\hat{\sigma}_{11,r}}, .., \sqrt{\hat{\sigma}_{NN,r}}\}$ is the diagonal matrix of sample standard errors of r_t. Methods to compute eigenvalues and eigenvectors of a symmetric matrix can then be used to perform PCA. An informal technique to determine the number of principal components needed in an application is to examine the scree plot, which is the time plot of the eigenvalues $\hat{\lambda}_i$ ordered from the largest to the smallest, that is, a plot of $\hat{\lambda}_i$ versus i. By looking for an elbow in the scree plot, indicating that the remaining eigenvalues are relatively small and all about the same size, one can determine the appropriate number of components. Note, except for the case in which $\lambda_j = 0$ for $j > i$, selecting the first i principal components only provides an approximation to the total variance of the data. If a small i can provide a good approximation, then the simplification becomes valuable.

5.6 Some conditional heteroscedastic models

Following the notation in Section (3.4.2) we are going to describe a few conditional heteroscedastic (CH) models.

5.6.1 The ARCH model

Starting with the ARCH model proposed by Engle, the main idea is that

1. the mean-corrected asset return a_t is serially uncorrelated, but dependent, and
2. the dependence of a_t can be described by a simple quadratic function of its lagged values.

Formally, an $ARCH(m)$ model is given by

$$a_t = \sigma_t \epsilon_t \ , \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + ... + \alpha_m a_{t-m}^2 \tag{5.6.3}$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean zero and variance 1, $\alpha_0 > 0$, and $\alpha_i \geq 0$ for $i > 0$. The coefficients α_i must satisfy some regularity conditions to ensure that the unconditional variance of a_t is finite. In practice, ϵ_t is often assumed to follow the standard normal or a standardised Student-t distribution. From the structure of the model, one can see that large past squared shocks $\{a_{t-i}^2\}_{i=1}^m$ imply a large conditional variance σ_t^2 for the mean-corrected return a_t, so that a_t tends to assume a large value (in modulus). Hence, in the ARCH model large shocks tend to be followed by another large shock similarly to the volatility clustering observed in asset returns. For simplicity of exposition, we consider the $ARCH(1)$ model given by

$$a_t = \sigma_t \epsilon_t \ , \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$. The unconditional mean of a_t remains zero because

$$E[a_t] = E[E[a_t|\mathcal{F}_{t-1}]] = E[\sigma_t E[\epsilon_t]] = 0$$

Further, the unconditional variance of a_t can be obtained as

$$Var(a_t) = E[a_t^2] = E[E[a_t^2|\mathcal{F}_{t-1}]] = E[\alpha_0 + \alpha_1 a_{t-1}^2] = \alpha_0 + \alpha_1 E[a_{t-1}^2]$$

Since a_t is a stationary process with $E[a_t] = 0$, $Var(a_t) = Var(a_{t-1}) = E[a_{t-1}^2]$, so that $Var(a_t) = \alpha_0 + \alpha_1 Var(a_t)$ and $Var(a_t) = \frac{\alpha_0}{1-\alpha_1}$. As the variance of a_t must be positive, we need $0 \leq \alpha_1 < 1$. Note, in some applications, we need higher order moments of a_t to exist, and α_1 must also satisfy additional constraints. For example, when studying the tail behaviour we need the fourth moment of a_t to be finite. Under the normality assumption of ϵ_t we have

$$E[a_t^4|\mathcal{F}_{t-1}] = 3\big(E[a_t^2|\mathcal{F}_{t-1}]\big)^2 = 3\big(\alpha_0 + \alpha_1 a_{t-1}^2\big)^2$$

Therefore, we get

$$E[a_t^4] = E[E[a_t^4|\mathcal{F}_{t-1}]] = 3E[\big(\alpha_0 + \alpha_1 a_{t-1}^2\big)^2] = 3E[\alpha_0^2 + 2\alpha_0\alpha_1 a_{t-1}^2 + \alpha_1^2 a_{t-1}^4]$$

If a_t is fourth-order stationary with $m_4 = E[a_t^4]$, then we have

$$m_4 = 3\big(\alpha_0^2 + 2\alpha_0\alpha_1 Var(a_t) + \alpha_1^2 m_4\big) = 3\alpha_0^2(1 + 2\frac{\alpha_1}{1-\alpha_1}) + 3\alpha_1^2 m_4$$

so that we get

$$m_4 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}$$

As a result,

1. since m_4 is positive, then α_1 must also satisfy the condition $1 - 3\alpha_1^2 > 0$, that is, $0 \leq \alpha_1^2 < \frac{1}{3}$
2. the unconditional kurtosis of a_t is

$$\frac{E[a_t^4]}{Var(a_t^4)} = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)} \frac{(1-\alpha_1)^2}{\alpha_0^2} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3$$

Thus, the excess kurtosis of a_t is positive and its tail distribution is heavier than that of a normal distribution. That is, the shock a_t of a conditional Gaussian $ARCH(1)$ model is more likely than a Gaussian white noise to produce outliers in agreement with the empirical findings on asset returns. These properties continues to hold for general ARCH models. Note, a natural way of achieving positiveness of the conditional variance is to rewrite an $ARCH(m)$ model as

$$a_t = \sigma_t \epsilon_t \ , \sigma_t^2 = \alpha_0 + A_{m,t-1}^\top \Omega A_{m,t-1}$$

where $A_{m,t-1} = (a_{t-1}, .., a_{t-M})^\top$ and Ω is a $m \times m$ non-negative definite matrix. Hence, we see that the $ARCH(m)$ model requires $Omega$ to be diagonal, and that Engle's model uses a parsimonious approach to approximate a quadratic function. A simple way to achieve the diagonality constraint on the matrix Ω is to employ a random coefficient model for a_t as done in the CHARMA and RCA models. Further, ARCH models also have some weaknesses

- the model assumes that positive and negative shocks have the same effects on volatility because it depends on the square of the previous shocks.
- the model is rather restrictive, since in the case of an $ARCH(1)$ the parameter α_1^2 is constraint to be in the interval $[0, \frac{1}{3}]$ for the series to have a finite fourth moment.
- the model is likely to overpredict the volatility because it slowly responds to large isolated shocks to the return series.

A simple way for building an ARCH model consists of three steps

1. build an econometric model (for example an ARMA model) for the return series to remove any linear dependence in the data, and use the residual series to test for ARCH effects
2. specify the ARCH order and perform estimation
3. check carefully the fitted ARCH model and refine it if necessary

To determine the ARCH order, we define $\eta_t = a_t^2 - \sigma_t^2$ since it was shown that $\{\eta_t\}$ is an uncorrelated series with zero mean. The ARCH model then becomes

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + ... + \alpha_m a_{t-m}^2 + \eta_t$$

which is the form of an $AR(m)$ model for a_t^2, except that $\{\eta_t\}$ is not a i.i.d. series. As a result, the least squares estimates of the prior model are consistent, but not efficient. The PACF of a_t^2, which is a useful tool for determining the order m, may not be effective in the case of small sample size.

Forecasts of the ARCH model in Equation (5.6.3) are obtained recursively just like those of an AR model. Given the $ARCH(m)$ model, at the forecast origin h, the 1-step ahead forecast of σ_{h+1}^2 is

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + ... + \alpha_m a_{h+1-m}^2$$

and the l-step ahead forecast for σ_{h+l}^2 is

$$\sigma_h^2(l) = \alpha_0 + \sum_{i=1}^{m} \alpha_i \sigma_h^2(l-i)$$

where $\sigma_h^2(l-i) = a_{h+l-i}^2$ if $l - i \leq 0$.

5.6.2 The GARCH model

In the GARCH model, given a log return series r_t, we assume that the mean equation of the process can be adequately described by an ARMA model. Then, the mean-corrected log return a_t follows a $GARCH(m,s)$ model if

$$a_t = \sigma_t \epsilon_t \, , \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{5.6.4}$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with zero mean and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i) < 1$ implying that the unconditional variance of a_t is finite, whereas its conditional variance σ_t^2 evolves over time. In general, ϵ_t is assumed to be a standard normal or standardised Student-t distribution. To understand the properties of GARCH models we let $\eta_t = a_t^2 - \sigma_t^2$ so that $\sigma_t^2 = a_t^2 - \eta_t$. By plugging $\sigma_{t-i}^2 = a_{t-i}^2 - \eta_{t-i}$ for i=0,..,s into Equation (5.6.4), we get

$$a_t^2 = \alpha_0 + \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t - \sum_{j=1}^{s} \beta_j \eta_{t-j} \tag{5.6.5}$$

Note, while $\{\eta_t\}$ is a martingale difference series ($E[\eta_t] = 0$ and $Cov(\eta_t, \eta_{t-j}) = 0$ for $j \geq 1$), in general it is not an i.i.d. sequence. Since the above equation is an ARMA form for the squared series a_t^2, a GARCH model can be regarded as an application of the ARMA idea to the squared series a_t^2. Hence, using the unconditional mean of an ARMA model, we have

$$E[a_t^2] = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i)}$$

provided that the denominator of the prior fraction is positive. For simplicity of exposition we now consider the $GARCH(1,1)$ model given by

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \, , 0 \leq \alpha_1, \beta_1 \leq 1, (\alpha_1 + \beta_1) < 1$$

We see that a large a_{t-1}^2 or σ_{t-1}^2 gives rise to a large σ_t^2 meaning that a large a_{t-1}^2 tends to be followed by another large a_t^2. It can also be shown that if $1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 > 0$, then

$$\frac{E[a_t^4]}{\big(E[a_t^2]\big)^2} = \frac{3\big(1 - (\alpha_1 + \beta_1)^2\big)}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3$$

so that, similarly to ARCH models, the tail distribution of a $GARCH(1,1)$ process is heavier than that of a normal distribution. Further, the model provides a simple parametric function that can be used for describing the volatility evolution. Forecasts of a GARCH model can be obtained in a similar way to those of the ARMA model. For example, in the $GARCH(1,1)$ model, with forecast origin h, for a 1-step ahead forecast we have

$$\sigma_{h+1}^2 = \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2$$

where a_h and σ_h^2 are known at the time index h. Hence, the 1-step ahead forecast is given by

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2 \tag{5.6.6}$$

In the case of multiple steps ahead, we use $a_t^2 = \sigma_t^2 \epsilon_t^2$ and rewrite the volatility equation as

$$\sigma_{t+1}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_t^2 + \alpha_1 \sigma_t^2 (\epsilon_t^2 - 1)$$

Setting $t = h + 1$, since $E[\epsilon_{h+1}^2 - 1 | \mathcal{F}_h] = 0$, the 2-step ahead volatility forecast becomes

$$\sigma_h^2(2) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(1)$$

and the l-step ahead volatility forecast satisfies the equation

$$\sigma_h^2(l) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(l-1) \ , \ l > 1 \tag{5.6.7}$$

which is the same result as that of an $ARMA(1,1)$ model with AR polynomial $1 - (\alpha_1 + \beta_1)B$. By repeated substitution of the above equation (5.6.7), the l-step ahead forecast can be rewritten as

$$\sigma_h^2(l) = \frac{\alpha_0\big(1 - (\alpha_1 + \beta_1)^{l-1}\big)}{1 - \alpha_1 - \beta_1} + (\alpha_1 + \beta_1)^{l-1}\sigma_h^2(1)$$

such that

$$\sigma_h^2(l) \to \frac{\alpha_0}{1 - \alpha_1 - \beta_1} \text{ as } l \to \infty$$

provided that $\alpha_1 + \beta_1 < 1$. As a result, the multistep ahead volatility forecasts of a $GARCH(1,1)$ model converge to the unconditional variance of a_t as the forecast horizon increases to infinity, provided that $Var(a_t)$ exists. Note, the GARCH models encounter the same weaknesses as the ARCH models, such as responding equally to positive and negative shocks. While the approach used to build ARCH models can be used for building GARCH models, it is difficult to specify the order of the latter. Fortunately, only lower order GARCH models are used in most applications. The conditional maximum likelihood method still applies provided that the starting values of the volatility $\{\sigma_t^2\}$ are assumed to be known. In some applications, the sample variance of a_t is used as a starting value.

5.6.3 The integrated GARCH model

If the AR polynomial of the GARCH representation in Equation (5.6.5) has a unit root, then we have an IGARCH model. That is, IGARCH models are unit-root GARCH models. As for ARIMA models, a key feature of IGARCH models is that the impact of past squared shocks $\eta_{t-i} = a_{t-i}^2 - \sigma_{t-i}^2$ for $i > 0$ on a_t^2 is persistent. For example, the $IGARCH(1,1)$ model is given by

$$a_t = \sigma_t\epsilon_t \ , \ \sigma_t^2 = \alpha_0 + \beta_1\sigma_{t-1}^2 + (1 - \beta_1)a_{t-1}^2$$

where $\{\epsilon_t\}$ is defined as before, and $1 > \beta_1 > 0$. In some applications, the unconditional variance of a_t, hence that of r_t, is not defined in this model. From a theoretical point of view, the IGARCH phenomenon might be caused by occasional level shifts in volatility. When $\alpha_1 + \beta_1 = 1$ in Equation ((5.6.7)), repeated substitutions in the l-step ahead volatility forecast equation of GARCH models gives

$$\sigma_h^2(l) = \sigma_h^2(1) + (l-1)\alpha_0 \ , \ l \geq 1$$

such that the effect of $\sigma_h^2(1)$ on future volatility is also persistent, and the volatility forecasts form a straight line with slope α_0. Note, the process σ_t^2 is a martingale for which some nice results are available (see Nelson [1990]). Under certain conditions, the volatility process is strictly stationary, but not weakly stationary as it does not have the first two moments. Further, in the special case $\alpha_0 = 0$ in the $IGARCH(1,1)$ model, the volatility forecasts are simply $\sigma_h^2(1)$ for all forecast horizons (see Equation (5.6.6)). This is the volatility model used in RiskMetrics (see details in Section (3.4.3)), which is an approach for calculating Value at Risk (VaR) (see details in Scetion (9.5.2)).

5.6.4 The GARCH-M model

In general, asset return should depend on its volatility, and one way forward is to consider the GARCH-M model, which is a GARCH in mean. A simple $GARCH(1,1) - M$ model is given by

$$\begin{aligned} r_t &= \mu + c\sigma_t^2 + a_t \ , \ a_t = \sigma_t\epsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1\sigma_{t-1}^2 \end{aligned}$$

where μ and c are constants. The parameter c is called the risk premium, with a positive value indicating that the return is positively related to its past volatility. The formulation of the above model implies serial correlations in the return series r_t introduced by those in the volatility process σ_t^2. Thus, the existence of risk premium is, therefore, another reason for historical stock returns to have serial correlations.

5.6.5 The exponential GARCH model

Nelson [1991] proposed the exponential GARCH (EGARCH) model allowing for asymmetric effects between positive and negative asset returns. He considered a weighted innovation which can be written as

$$g(\epsilon_t) = \begin{cases} (\theta+\gamma)\epsilon_t - \gamma E[|\epsilon_t|] \text{ if } \epsilon_t \geq 0 \\ (\theta-\gamma)\epsilon_t - \gamma E[|\epsilon_t|] \text{ if } \epsilon_t < 0 \end{cases}$$

where θ and γ are real constants. Both ϵ_t and $|\epsilon_t| - E[|\epsilon_t|]$ are zero-mean i.i.d. sequences with continuous distributions, so that $E[g(\epsilon_t)] = 0$. For the standard Gaussian random variable ϵ_t, we have $E[|\epsilon_t|] = \sqrt{\frac{2}{\pi}}$. An $EGARCH(m,s)$ model can be written as

$$a_t = \sigma_t\epsilon_t \ , \ \ln(\sigma_t^2) = \alpha_0 + \frac{1+\beta_1 B + ... + \beta_s B^s}{1-\alpha_1 B - ... - \alpha_m B^m} g(\epsilon_{t-1})$$

where α_0 is a constant, B is the back-shift (or lag) operator such that $Bg(\epsilon_t) = g(\epsilon_{t-1})$, and both the numerator and denominator above are polynomials with zeros outside the unit circle (absolute values of the zeros are greater than one) and have no common factors. Since the EGARCH model uses the ARMA parametrisation to describe the evolution of the conditional variance of a_t, some properties of the model can be obtained in a similar manner as those of the GARCH model. For instance, the unconditional mean of $\ln(\sigma_t^2)$ is α_0. However, the EGARCH model uses logged conditional variance to relax the positiveness constraint of model coefficients, and the use of $g(\epsilon_t)$ enables the model to respond asymmetrically to positive and negative lagged values of a_t. For example, in the simple $EGARCH(1,0)$ we get

$$a_t = \sigma_t\epsilon_t \ , (1-\alpha B)\ln(\sigma_t^2) = (1-\alpha)\alpha_0 + g(\epsilon_{t-1})$$

where $\{\epsilon_t\}$ are i.i.d. standard normal and the subscript of α_1 is omitted. In this case, $E[|\epsilon_t|] = \sqrt{\frac{2}{\pi}}$ and the model for $\ln(\sigma_t^2)$ becomes

$$(1-\alpha B)\ln(\sigma_t^2) = \begin{cases} \alpha_* + (\theta+\gamma)\epsilon_{t-1} \text{ if } \epsilon_{t-1} \geq 0 \\ \alpha_* + (\theta-\gamma)\epsilon_{t-1} \text{ if } \epsilon_{t-1} < 0 \end{cases}$$

where $\alpha_* = (1-\alpha)\alpha_0 - \sqrt{\frac{2}{\pi}}\gamma$. Note, this is a nonlinear function similar to that of the threshold autoregressive (TAR) model of Tong [1990]. In this model, the conditional variance evolves in a nonlinear manner depending on the sign of a_{t-1}, that is,

$$\sigma_t^2 = \sigma_{t-1}^{2\alpha} e^{\alpha_*} \begin{cases} e^{(\theta+\gamma)\frac{a_{t-1}}{\sqrt{\sigma_{t-1}^2}}} \text{ if } a_{t-1} \geq 0 \\ e^{(\theta-\gamma)\frac{a_{t-1}}{\sqrt{\sigma_{t-1}^2}}} \text{ if } a_{t-1} < 0 \end{cases}$$

The coefficients $(\theta \pm \gamma)$ show the asymmetry in response to positive and negative a_{t-1}, and the model is nonlinear when $\gamma \neq 0$. In presence of higher orders, the nonlinearity becomes much more complicated. Now, given the $EGARCH(1,0)$ model, assuming known model parameters and that the innovations are standard Gaussian, we have

$$\ln(\sigma_t^2) = (1-\alpha_1)\alpha_0 + \alpha_1 \ln(\sigma_{t-1}^2) + g(\epsilon_{t-1})$$
$$g(\epsilon_{t-1}) = \theta\epsilon_{t-1} + \gamma\big(|\epsilon_{t-1}| - \sqrt{\frac{2}{\pi}}\big)$$

Taking exponential, the model becomes

$$\sigma_t^2 = \sigma_{t-1}^{2\alpha_1} e^{(1-\alpha_1)\alpha_0} e^{g(\epsilon_{t-1})}$$

and the 1-step ahead forecast, with forecast origin h, satisfies

$$\sigma_{h+1}^2 = \sigma_h^{2\alpha_1} e^{(1-\alpha_1)\alpha_0} e^{g(\epsilon_h)}$$

where all the quantities on the right-hand side are known. Thus, the 1-step ahead volatility forecast at the forecast origin h is simply $\hat{\sigma}_h^1(1) = \sigma_{h+1}^2$. Repeating for the 2-step ahead forecast, and taking conditional expectation at time h, we get

$$\hat{\sigma}_h^2(2) = \hat{\sigma}_h^{2\alpha_1}(1) e^{(1-\alpha_1)\alpha_0} E_h[e^{g(\epsilon_{h+1})}]$$

where $E_h[\bullet]$ denotes the conditional expectation at the time origin h. After some calculation, the prior expectation is given by

$$E[e^{g(\epsilon)}] = e^{-\gamma\sqrt{\frac{2}{\pi}}}\left(e^{\frac{1}{2}(\theta+\gamma)^2} N(\theta+\gamma) + e^{\frac{1}{2}(\theta-\gamma)^2} N(\theta-\gamma)\right)$$

where $f(\bullet)$ and $N(\bullet)$ are the probability density function and CDF of the standard normal distribution, respectively. As a result, the 2-step ahead volatility forecast becomes

$$\hat{\sigma}_h^2(2) = \hat{\sigma}_h^{2\alpha_1}(1) e^{(1-\alpha_1)\alpha_0} e^{-\gamma\sqrt{\frac{2}{\pi}}}\left(e^{\frac{1}{2}(\theta+\gamma)^2} N(\theta+\gamma) + e^{\frac{1}{2}(\theta-\gamma)^2} N(\theta-\gamma)\right)$$

Repeating this procedure, we obtain a recursive formula for the j-step ahead forecast

$$\hat{\sigma}_h^2(j) = \hat{\sigma}_h^{2\alpha_1}(j-1) e^{w}\left(e^{\frac{1}{2}(\theta+\gamma)^2} N(\theta+\gamma) + e^{\frac{1}{2}(\theta-\gamma)^2} N(\theta-\gamma)\right)$$

where $w = (1-\alpha_1)\alpha_0 - \gamma\sqrt{\frac{2}{\pi}}$.

5.6.6 The stochastic volatility model

An alternative approach for describing the dynamics of volatility is to introduce the innovation v_t to the conditional variance equation of a_t obtaining a stochastic volatility (SV) model (see Melino et al. [1990], Harvey et al. [1994]). Using $\ln(\sigma_t^2)$ to ensure positivity of the conditional variance, a SV model is defined as

$$a_t = \sigma_t \epsilon_t \ , \ \left(1 - \alpha_1 B - ... - \alpha_m B^m\right) \ln(\sigma_t^2) = \alpha_0 + v_t$$

where ϵ_t are i.i.d. $N(0,1)$, v_t are i.i.d. $N(0,\sigma_v^2)$, $\{\epsilon_t\}$ and $\{v_t\}$ are independent, α_0 is a constant, and all zeros of the polynomial $1 - \sum_{i=1}^m \alpha_i B^i$ are greater than 1 in modulus. While the innovation v_t substantially increases the flexibility of the model in describing the evolution of σ_t^2, it also increases the difficulty in parameter estimation since for each shock a_t the model uses two innovations ϵ_t and v_t. To estimate a SV model, one need a quasi-likelihood method via Kalman filtering or a Monte Carlo method. Details on SV models and their parameters estimation can be found in Taylor [1994]. Properties of the model can be found in Jacquier et al. [1994] when $m = 1$. In that setting, we have

$$\ln(\sigma_t^2) \sim N(\frac{\alpha_0}{1-\alpha_1}, \frac{\sigma_v^2}{1-\alpha_1^2}) = N(\mu_h, \sigma_h^2)$$

and

$$E[a_t^2] = e^{\mu_h + \frac{1}{2\sigma_h^2}} \ , \ E[a_t^4] = 3e^{2\mu_h^2 + 2\sigma_h^2} \ , \ Corr(a_t^2, a_{t-i}^2) = \frac{e^{\sigma_h^2 \alpha_1^i} - 1}{3e^{\sigma_h^2} - 1}$$

While SV models often provide improvements in model fitting, their contributions to out-of-sample volatility forecasts received mixed results. Note, using the idea of fractional difference (see Section (5.4.3)), SV models have been extended to allow for long memory in volatility. This extension has been motivated by the fact that autocorrelation function of the squared or absolute-valued series of asset returns often slowly decay, even though the return series has no serial correlation (see Ding et al. [1993]). A simple long-memory stochastic volatility (LMSV) model can be defined as

$$a_t = \sigma_t \epsilon_t \ , \sigma_t = \sigma e^{\frac{1}{2}u_t} \ , (1-B)^d u_t = \eta_t$$

where $\sigma > 0$, ϵ_t are i.i.d. $N(0,1)$, η_t are i.i.d. $N(0,\sigma_\eta^2)$ and independent of ϵ_t, and $0 < d < \frac{1}{2}$. The feature of long memory stems from the fractional difference $(1-B)^d$ implying that the ACF of u_t decays slowly at a hyperbolic, instead of an exponential rate as the lag increases. Using these settings we have

$$\begin{aligned} \ln(a_t^2) &= \ln(\sigma^2) + u_t + \ln(\epsilon_t^2) \\ &= \big(\ln(\sigma^2) + E[\ln(\epsilon_t^2)]\big) + u_t + \big(\ln(\epsilon_t^2) - E[\ln(\epsilon_t^2)]\big) = \mu + u_t + e_t \end{aligned}$$

so that the $\ln(a_t^2)$ series is a Gaussian long-memory signal plus a non-Gaussian white noise (see Breidt et al. [1998]). Estimation of the long-memory stochastic volatility model is difficult, but the fractional difference parameter d can be estimated by using either a quasi-maximum likelihood method or a regression method. Using the log series of squared daily returns for companies in $S\&P$ 500 index, Bollerslev et al. [1999] and Ray et al. [2000] found the median estimate of d to be about 0.38. Further, Ray et al. studied common long-memory components in daily stock volatilities of groups of companies classified according to various characteristics, and found that companies in the same industrial or business sector tend to have more common long-memory components.

5.6.7 Another approach: high-frequency data

Due to the availability of high-frequency financial data, especially in the foreign exchange markets, alternative approach for volatility estimation using high-frequency data to calculate volatility of low frequency returns developed (see French et al. [1987]). For example, considering the monthly volatility of an asset for which daily returns are available, we let r_t^m be the monthly log return of the asset at month t. Assuming n trading days in the month t, the daily log returns of the asset in the month are $\{r_{t,i}\}_{i=1}^n$. Using properties of log returns, we have

$$r_t^m = \sum_{i=1}^n r_{t,i}$$

Assuming that the conditional variance and covariance exist, we have

$$Var(r_t^m|\mathcal{F}_{t-1}) = \sum_{i=1}^n Var(r_{t,i}|\mathcal{F}_{t-1}) + 2\sum_{i<j} Cov(r_{t,i}, r_{t,i}|\mathcal{F}_{t-1})$$

where $\mathcal{F}_{t-1}$ is the filtration at time $t-1$. The prior equation can be simplified if additional assumptions are made. For instance, if we assume that $\{r_{t,i}\}$ is a white noise series, then

$$Var(r_t^m|\mathcal{F}_{t-1}) = nVar(r_{t,1})$$

where $Var(r_{t,1})$ can be estimated from the daily returns $\{r_{t,i}\}_{i=1}^n$ by

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (r_{t,i} - \overline{r}_t)^n$$

where $\overline{r}_t$ is the sample mean of the daily log returns in month t, that is, $\overline{r}_t = \frac{1}{n}\sum_{i=1}^{n}$. The estimated monthly volatility is then

$$\hat{\sigma}_m^2 = \frac{n}{n-1}\sum_{i=1}^{n}(r_{t,i} - \overline{r}_t)^2$$

If $\{r_{t,i}\}$ follows an $MA(1)$ model, then

$$Var(r_t^m|\mathcal{F}_{t-1}) = nVar(r_{t,1}) + 2(n-1)Cov(r_{t,1}, r_{t,2})$$

which can be estimated by

$$\hat{\sigma}_m^2 = \frac{n}{n-1}\sum_{i=1}^{n}(r_{t,i} - \overline{r}_t)^2 + 2\sum_{i=1}^{n-1}(r_{t,i} - \overline{r}_t)(r_{t,i+1} - \overline{r}_t)$$

Note, the model for daily returns $\{r_{t,i}\}$ is unknown which complicates the estimation of covariances. Further, with about 21 trading days in a month, the sample size is small, leading to poor accuracy of the estimated variance and covariance. The accuracy depends on the dynamic structure of $\{r_{t,i}\}$ and their distribution. If the daily log returns have high excess kurtosis and serial correlations, the sample estimates $\hat{\sigma}_m^2$ may not even be consistent.

5.6.8 Forecasting evaluation

Since one can not directly observe the volatility of asset returns, comparing the forecasting performance of different volatility models is a challenge. It is common practice to use out-of-sample forecasts and compare the volatility forecasts $\sigma_h^2(l)$ with the shock a_{h+l}^2 in the forecasting sample to assess the forecasting performance of a volatility model. Doing so, researchers often finds a low correlation coefficient between the two terms. However, this is not surprising because a_{h+l}^2 alone is not an adequate measure of the volatility at time index $h+l$. That is, in the 1-step ahead forecasts, we have $E[a_{h+l}^2|\mathcal{F}_h] = \sigma_{h+1}^2$ so that a_{h+l}^2 is a consistent estimate of σ_{h+l}^2, but it is not a accurate estimate since a single observation of a random variable with known mean value can not provide an accurate estimate of its variance. See Andersen et al. [1998] for more information on the forecasting evaluation of GARCH models.

5.7 Exponential smoothing and forecasting data

Since the original work by Brown [1959] and Holt [1957] on exponential smoothing (ES), a complete statistical rational for ES based on a new class of state-space models with a single source of error developed (see Gardner [2006]). Exponential smoothing can be justified in part through equivalent kernel regression and ARIMA models, and in their entirety through the new class of single source of error (SSOE) state-space models having many theoretical advantages, among which the ability to make the errors dependent on the other components of the time series. This kind of multiplicative error structure being not possible with the ARIMA class, exponential smoothing are a much broader class of models. ES originated in Brown's work as an OR analyst for the US Navy during World War II. During the early 1950's, Brown extended simple exponential smoothing (SES) to discrete data and developed methods for trends and seasonality. Brown [1959] presented his research at a conference of the Operations Research Society of America, and later [1963] developed the general exponential smoothing (GES) methodology. During the 1950's. Holt [1957], with the support from the Logistics Branch of the Office of Naval Research, worked independently of Brown to develop a similar method for ES of additive trends and an entirely different method for smoothing seasonal data. In a landmark article, Winters [1960] tested Holt's methods with empirical data, and they became known as the Holt-Winters forecasting system. At the same time, Muth [1960] was among the first to examine the optimal properties of ES forecasts. The work of Gardner [1985], in particular on damped trend exponential smoothing, led to the use of ES in automatic forecasting. Hyndman et al. [2008] developed a more general class of methods with a uniform approach to calculation of prediction intervals, maximum likelihood estimation and the exact calculation of

model selection criteria such as Akaike's Information Criterion (AIC). In view of capturing the locally constant nature of the linear trend by means of its gradient which can change smoothly or suddenly, Gardner [2009] developed a random coefficient state-space model for which damped trend smoothing provides an optimal approach.

In order to develop an effective forecasting process, we need to understand when and how to use forecasting methods, how to interpret the results, and how to recognise their limitations and the potential for improvement. We start by presenting some heuristic forecasting methods and will move towards procedures relying upon careful modelling of the process being studied.

5.7.1 The moving average

In statistics, a moving average is a type of finite impulse response filter used to analyse a set of data points by creating a series of averages of different subsets of the full data set. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by shifting forward, that is, excluding the first number of the series and including the next number following the original subset in the series. This creates a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plot line connecting all the (fixed) averages is the moving average. A moving average may also use unequal weights for each datum value in the subset to emphasise particular values in the subset. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. Mathematically, a moving average is a type of convolution and so it can be viewed as an example of a low-pass filter used in signal processing.

5.7.1.1 Simple moving average

In financial applications a SMA is the unweighted mean of the previous n datum points. As an example of a simple equally weighted running mean for a n-day sample of closing price is the mean of the previous n days's closing prices. Given the latest day $t = j\delta$, if these prices are $P_j, P_{j-1}, .., P_{j-(n-1)}$ then the formula for the moving average at index j and order n is

$$SMA(j,n) = \frac{1}{n}\sum_{i=0}^{n-1} P_{j-i}$$

where P_j is the price of the latest day.

***Remark* 5.7.1** *Note, it correspond to the Savitzky-Golay low-pass smoothing filter given in Equation (4.3.12) with* $n_R = 0$ *and* $c_n = \frac{1}{n_L+1}$ *(or* $m = 0$*).*

When calculating successive values, a new value comes into the sum and an old value drops out, meaning a full summation each time is unnecessary for this simple case

$$SMA(j,n) = SMA(j-1,n) - \frac{P_{j-n}}{n} + \frac{P_j}{n}$$

The period selected depends on the type of movement of interest, such as short, intermediate, or long term. In financial terms MA levels can be interpreted as support in a rising market, or resistance in a falling market. If data used is not centred around the mean, a SMA lags behind the latest datum by half the sample width. Further, an SMA can be disproportionately influenced by old datum points dropping out or new data coming in. In order to chose the order n we need to evaluate suitable measures of forecast performance.

5.7.1.2 Weighted moving average

A weighted average is any average that has multiplaying factors to give different weights to data at different positions in the sample window. The MA is the convolution of the datum points with a fixed weighting function. In finance, a WMA has the specific meaning of weights that decrease in arithmetical progression. In an n-day WMA the latest day $t = j\delta$ has weight n, the second latest $(n-1)$ etc. down to one

$$WMA(j) = \frac{nP_j + (n-1)P_{j-1} + .. + 2P_{j-n-2} + P_{j-n+1}}{n + (n-1) + ... + 2 + 1}$$

The denominator is a triangle number equal to $\frac{n(n+1)}{2}$. In the more general case the denominator will always be the sum of the individual weights

$$WMA(j) = \frac{1}{norm}\sum_{i=0}^{n-1} \omega_i P_{j-i}$$

with $\omega_i = n - i$ and $norm = \sum_{i=0}^{n-1} \omega_i$. Hence, in the SMA the weights are $\omega_i = 1$ for $i = 0, .., (n-1)$. The difference of the numerator at time $(t+1)$ and t is

$$Num(j+1) - Num(j) = \sum_{i=0}^{n-1} \omega_i P_{(j+1)-i} - \sum_{i=0}^{n-1} \omega_i P_{j-i} = nP_{j+1} - P_j - .. - P_{j-(n-1)}$$

Let's call $T(j) = P_j + ... + P_{j-(n-1)}$ then

$$T(j+1) = T(j) + P_{j+1} - P_{j-(n-1)}$$

As a result

$$Num(j+1) = Num(j) + nP_{j+1} - T_j$$

so that the weighted average at time $t+1$ becomes

$$WMA(j+1) = \frac{1}{norm} Num(j+1)$$

5.7.1.3 Exponential smoothing

Brown [1959] introduced the exponential smoothing as a forecasting method adjusting more smoothly over time than moving averages. We let $\overline{Y}(N)$ be the arithmetic mean of the past N observations where the updated mean $\overline{Y}(N+1)$ for the past $(N+1)$ observations is given by

$$\overline{Y}(N+1) = \frac{Y_1 + .. + Y_{N+1}}{N+1} = \overline{Y}(N) + \frac{Y_{N+1} - \overline{Y}(N)}{N+1}$$

However, as the series length increases the forecasts generated by the latest mean, become increasingly unresponsive to fluctuations in recent values, since each observation has weight $\frac{1}{\text{sample size}}$. The update of the simple MA given in Section (5.7.1.1) avoided this problem and maintained a constant coefficient $\frac{1}{N}$ where N is the sample running period, but at the cost of dumping the oldest observation completely. To overcome this problem, we consider the updating relationship of the Exponential Smoothing (or exponentially weighted moving average EWMA) with basic equation being

$$\overline{Y}(N+1, \alpha) = \overline{Y}(N, \alpha) + \alpha[Y_{N+1} - \overline{Y}(N, \alpha)]$$

where $\alpha \in [0,1]$ is the smoothing constant. The process involves comparing the latest observation with the previous weighted average and making a proportional adjustment, governed by the coefficient α, known as the smoothing constant. Note, the updates require only the latest observation and the previous mean. If we start with the expression for time $(N+1)$ and substitute into it the comparable expression for time N, we obtain

$$\begin{aligned}\overline{Y}(N+1,\alpha) &= (1-\alpha)\overline{Y}(N,\alpha) + \alpha Y_{N+1} \\ &= (1-\alpha)[(1-\alpha)\overline{Y}(N-1,\alpha) + \alpha Y_N] + \alpha Y_{N+1} \\ &= (1-\alpha)^2\overline{Y}(N-1,\alpha) + \alpha[(1-\alpha)Y_N + Y_{N+1}]\end{aligned}$$

Continuing to substitute the earlier means, we eventually arrive back at the start of the series with the expression

$$\overline{Y}(N+1,\alpha) = (1-\alpha)^{N+1}\overline{Y}(0,\alpha) + \alpha[Y_{N+1} + (1-\alpha)Y_N + (1-\alpha)^2 Y_{N-1} + ... + (1-\alpha)^N Y_1] \tag{5.7.8}$$

The right hand side contains a weighted average of the observations and the weights $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, ...$ decay steadily over time in an exponential fashion. The decay is slower for small values of α. Note, the equation depends on a starting value $\overline{Y}(0,\alpha)$. Further, when N and α are both small, the weight attached to the starting value may be high.

We now need to convert these averages into forecasts where we assume that the average level for future observations is best estimated by our current weighted average. Thus, forecasts made at time t for all future time periods will be the same

$$F_{t+1|t} = F_{t+2|t} = ... = F_{t+h|t} = \overline{Y}(t,\alpha)$$

so that as the lead time increases the forecasts will usually be progressively less accurate.

***Remark* 5.7.2** *We note that forecasts based upon the mean or the simple moving average also have the same property of being equal for all future periods.*

Of course, once the next observation is made, the common forecast value changes. The forecast equation for the EWMA is

$$F_{t+1|t} = F_{t|t-1} + \alpha[Y_t - F_{t|t-1}]$$

We define the observed error ϵ_t as the difference between the newly observed value of the series and its previous one-step-ahead forecast

$$\epsilon_t = Y_t - F_t$$

so that the forecast equation becomes

$$F_{t+1|t} = F_{t|t-1} + \alpha\epsilon_t \tag{5.7.9}$$

To set up these calculations, we need to specify the value for α and a starting value F_1. Two principal options are commonly used to resolve the choice of starting values, either to use the first observation as F_1, or to use an average of a number of observations (average of the first six observations). Earlier literature recommended a choice for α in the range $0.1 < \alpha < 0.3$ to allow the EWMA to change relatively slowly. However, as discussed in Section (**??**), we should not rely on an arbitrary pre-set smoothing parameter. One possibility is to minimise the mean squared error (MSE) for the one-step ahead forecasts. Given Equation (5.7.8) we can rewrite the forecast equation as

$$\hat{Y}_{T+\tau}(T) = (1-\alpha)^T\hat{Y}_0 + \alpha[Y_T + (1-\alpha)Y_{T-1} + (1-\alpha)^2 Y_{T-2} + ... + (1-\alpha)^{N-1}Y_{T-(N-1)}]$$

where $\hat{Y}_0$ is an initial estimate of the smoothed series $\{\hat{Y}_t\}_{t\in\mathbb{Z}^+}$ which is simply the average of the first six observations or, if there are less than six observations, the average of all the observations. As a measure of error we can approximate the 95% prediction interval for $\hat{Y}_{T+\tau}(T)$ given by

$$\hat{Y}_{T+\tau}(T) \pm z_{0.25} 1.25 MAD(T)$$

where $z_{0.25} \approx 1.96$.

5.7.1.4 Exponential moving average revisited

An exponential moving average (EMA), also known as an exponentially weighted moving average (EWMA), is a type of infinite impulse response filter that applies weighting factors which decrease exponentially. The weighting for each older datum point decreases exponentially, never reaching zero. According to Hunter (1986), the EMA for a series P may be calculated recursively as

$$E_t = \alpha P_t + (1-\alpha)E_{t-1} \text{ for } t > 1$$

with $E_1 = P_1$ and where P_t is the value at a time period t, and E_t is the value of the EMA at any time period t. The coefficient $\alpha \in [0,1]$ represents the degree of weighting decrease. A higher α discounts older observations faster. Alternatively, α may be expressed in terms of N time periods, where $\alpha = \frac{2}{(N+1)}$. For example, if $N = 19$ it is equivalent to $\alpha = 0.1$, the half-life of the weights is approximately $\frac{N}{2.8854}$. Note, E_1 is undefined, and it may be initialised in a number of different ways, most commonly by setting E_1 to P_1, though other techniques exist, such as setting E_1 to an average of the first 4 or 5 observations. The prominence of the E_1 initialisation's effect on the resultant moving average depends on α; smaller α values make the choice of E_1 relatively more important than larger α values, since a higher α discounts older observations faster. By repeated application of this formula for different times, we can eventually write E_t as a weighted sum of the datum points P_t as:

$$E_t = \alpha\big(P_{t-1} + (1-\alpha)P_{t-2} + (1-\alpha)^2 P_{t-3} + .. + (1-\alpha)^k P_{t-(k+1)}\big) + (1-\alpha)^{k+1} E_{t-(k+1)}$$

for any suitable $k = 0, 1, 2, ...$ The weight of the general datum point P_{t-i} is $\alpha(1-\alpha)^{i-1}$. We can show how the EMA steps towards the latest datum point, but only by a proportion of the difference (each time)

$$E_t = E_{t-1} + \alpha(P_t - E_{t-1})$$

Expanding out E_{t-1} each time results in the following power series, showing how the weighting factor on each datum point $p_1, p_2, ...$ decreases exponentially:

$$E_t = \alpha\big(p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + ...\big)$$

where $p_1 = P_t$, $p_2 = P_{t-1}$, ... Note, the weights $\alpha(1-\alpha)^t$ decrease geometrically, and their sum is unity. Using a property of geometric series, we get

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = \alpha \sum_{i=1}^{t} (1-\alpha)^{i-1} = \alpha\big[\frac{1-(1-\alpha)^t}{1-(1-\alpha)}\big] = 1-(1-\alpha)^t$$

and $\lim_{t\to\infty}(1-\alpha)^t = 0$. As a result, we get

$$\alpha \sum_{i=1}^{\infty} (1-\alpha)^{i-1} = 1$$

Focusing on the term α, we get

$$E_t = \frac{p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + ...}{norm} = \frac{1}{norm}\sum_{i=1}^{\infty}\omega_i p_i$$

where $\omega_i = (1-\alpha)^{i-1}$ and

$$\frac{1}{\alpha} = norm = 1 + (1-\alpha) + (1-\alpha)^2 + ... = \sum_{i=1}^{\infty}(1-\alpha)^{i-1}$$

This is an infinite sum with decreasing terms. The N periods in an N-day EMA only specify the α factor. N is not a stopping point for the calculation in the way it is in an SMA or WMA. For sufficiently large N, the first N datum points in an EMA represent about 86% of the total weight in the calculation

$$\frac{\alpha\big(1 + (1-\alpha) + (1-\alpha)^2 + ... + (1-\alpha)^N\big)}{\alpha\big(1 + (1-\alpha) + (1-\alpha)^2 + ... + (1-\alpha)^\infty\big)} = 1 - (1 - \frac{2}{N+1})^{N+1}$$

and

$$\lim_{N\to\infty}\big[1 - (1 - \frac{2}{N+1})^{N+1}\big] = 1 - e^{-2} \approx 0.8647$$

since $\lim_{n\to\infty}(1 + \frac{x}{n})^n = e^x$. This power formula can give a starting value for a particular day, after which the successive days formula above can be applied. The question of how far back to go for an initial value depends, in the worst case, on the data. Large price values in old data will affect on the total even if their weighting is very small. The weight omitted by stopping after k terms can be used in the fraction

$$\frac{\text{weight omitted by stopping after k terms}}{\text{total weight}} = (1-\alpha)^k$$

For example, to have 99.9% of the weight, set the ratio equal to 0.1% and solve for k:

$$k = \frac{\log(0.001)}{\log(1-\alpha)}$$

As the Taylor series of $\log(1-\alpha) = -\alpha - \frac{1}{2}\alpha^2 - ...$ tends to $-\alpha$ we get the limit $\lim_{N\to\infty}\log(1-\alpha) = -\frac{2}{(N+1)}$, and the computation simplifies to

$$k = -\log(0.001)\frac{(N+1)}{2}$$

for this example and $\frac{1}{2}\log(0.001) = -3.45$.

5.7.2 Introducing exponential smoothing models

Given the recorded observations $Y_1, Y_2, ..., Y_t$ over t time periods, which represent all the data currently available, our interest lies in forecasting the series $Y_{t+1}, .., Y_{t+h}$ over the next h weeks, known as the forecasting horizon. The (point) forecasts for future series are all made at time t, known as the forecast origin, so the first forecast will be made one step ahead, the second two steps ahead, and so on. We let $F_{t+h|t}$ be the forecast for Y_{t+h} made at time t. The subscripts always tell us which time period is being forecast and when the forecast was made. When no ambiguity arises, we will use F_{t+1} to represent the one-step-ahead forecast $F_{t+1|t}$.

5.7.2.1 Linear exponential smoothing

Following Brown [1959] and Holt [1957] on exponential smoothing, to avoid the continuation of global patterns, we are now considering tools that project trends more locally. Given the straight-line equation

$$Y_t = L_0 + Bt$$

where L_t is the level of series at time t and B_t is slope of series at time t. In our example, as we have a constant slope $B_t = B$ then the value of the series starts out at the value L_0 at time zero and increases by an amount B in each time period. Another way of writing the right-hand side of this expression is to state directly that the new level, L_t is obtained from the previous level by adding one unit of slope B

$$L_t = L_0 + Bt = L_{t-1} + B$$

We may then define the variable Y_t in terms of the level and the slope as

$$Y_t = L_{t-1} + B$$

Further, we can see that if we go h periods ahead, we can define the variable at time $(t + h)$ in terms of the level at time $t - 1$ and the appropriate number of slope increments

$$Y_{t+h} = L_0 + B(t + h) = L_t + Bh$$

There is a lot of redundancy in these expressions, since we are considering the error-free case. When we turn back to the real problem with random errors and changes over time in the level and the slope, these equations suggest that we consider forecasts of the form

$$F_{t+h|t} = L_t + hB_t \tag{5.7.10}$$

which is a straight line. Thus, the one-step-ahead forecast made at time $(t - 1)$ is

$$F_{t|t-1} = F_t = L_{t-1} + B_{t-1}$$

We can now consider updating the level and the slope using equations like those we used for SES. We define the observed error ϵ_t as the difference between the newly observed value of the series and its previous one-step-ahead forecast

$$\epsilon_t = Y_t - F_t = Y_t - (L_{t-1} + B_{t-1})$$

Given the latest observation Y_t we update the expressions for the level and the slope by making partial adjustments that depend upon the error

$$\begin{aligned} L_t &= L_{t-1} + B_{t-1} + \alpha\epsilon_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + B_{t-1}) \\ B_t &= B_{t-1} + \alpha\beta\epsilon_t \end{aligned} \tag{5.7.11}$$

The new slope is the old slope plus a partial adjustment (weight $\alpha\beta$) for the error. These equations are known as the error correction form of the updating equations. As may be checked by substitution, the slope update can also be expressed as

$$B_t = B_{t-1} + \beta(L_t - L_{t-1} - B_{t-1})$$

where $L_t - L_{t-1}$ represents the latest estimate of the slope such that $(L_t - L_{t-1} - B_{t-1})$ is the latest error made if the smoothed slope is used as an estimate instead. Hence, a second round of smoothing is applied to estimate the slope which has led some authors to describe the method as double exponential smoothing.

To start the forecast process we need starting values for the level and slope, and values for the two smoothing constants α and β. The smoothing constants may be specified by the user, and conventional wisdom decrees using $0.05 < \alpha < 0.3$ and $0.05 < \alpha < 0.15$. These values are initial values in a procedure to select optimal coefficients by minimising the MSE over some initial sample. As for SES, different programs use a variety of procedures to set starting values (see Gardner [1985]). One can use

$$B_3 = \frac{Y_3 - Y_1}{2}$$

for the slope and

$$L_3 = \frac{Y_1 + Y_2 + Y_3}{3} + \frac{Y_3 - Y_1}{2}$$

for the level, corresponding to fitting a straight line to the first three observations. Thus, the first three observations were used to set initial values for the level and slope. Once the initial values are set, equations (5.7.11) are used to update the level and slope as each new observation becomes available. When time series show a very strong trend, we would expect LES to perform much better than SES.

The LES method requires that the series is locally linear. That is, if the trend for the last few time periods in the series appears to be close to a straight line, the method should work well. However, in many cases this assumption is not realistic. For example, in the case of exponential growth, any linear approximation will undershoot the true function sooner or later. A first approach would be to use a logarithmic transformation. Given

$$Y_t = \gamma Y_{t-1}$$

where γ is some constant, the log transform becomes

$$\ln Y_t = \ln \gamma + \ln Y_{t-1}$$

and the log-transform produces a linear trend to which we can apply LES. We must then transform back to the original series to obtain the forecasts of interest. Writing $Z_t = \ln Y_t$ the reverse transformation is

$$Y_t = e^{Z_t}$$

In general, SES should not be used for strongly trending series; whether to use LES on the original or transformed series, or to use SES on growth rates, remains a question for further examination in any particular study.

5.7.2.2 The damped trend model

Some time series have a history of growth, possibly later followed by a decline, other series may have a strong tendency to increase over time, while other time series relates to the returns on an investment. Based on the findings of the M competition, Makridakis et al. [1982] showed that the practice of projecting a straight line trend indefinitely into the future was often too optimistic (or pessimistic). Hence, one can either convert the series to growth over time and forecast the growth rate, or we need to develop forecasting methods that account for trends. When the growth rate slow down and then decline we can accommodate such effects by modifying the updating equations for the level and slope. Assuming that the series flatten out unless the process encounters some new stimulus, the slope should approach zero. We consider the damped trend model introduced by Gardner and McKenzie [1985] which proved to be very effective (see Makridakis and Hibon [2000]). This is achieved by introducing a dampening factor ϕ in equations (5.7.11), getting

$$\begin{aligned} L_t &= L_{t-1} + \phi B_{t-1} + \alpha \epsilon_t \\ B_t &= \phi B_{t-1} + \alpha\beta\epsilon_t \end{aligned}$$

where $\phi \in [0,1]$ multiplies each slope term B_{t-1} shifting that term towards zero, or dampening it. Computing the forecast function for h-steps ahead, we get

$$F_{t+h|t} = L_t + (\phi + \phi^2 + .. + \phi^h)B_t$$

This forecast levels out over time approaching the limiting value $L_t + \frac{B_t}{1-\phi}$ provided the dampening factor is less than one. This is to contrast with the case $\phi = 1$ when the forecast keeps increasing so that $F_{t+h|t} = L_t + hB_t$.

Robert Brown [1959] was the original developer of exponential smoothing methods where his initial derivation of exponential smoothing used a least squares argument which, for a local linear trend reduces to the use of LES with $\alpha = \beta$. In general, there is no particular benefit to imposing this restriction. However, the discounted least squares approach is particularly useful when complex non-linear functions are involved and updating equations are not readily available. If we set $\beta = 0$ the updating equations (5.7.11) become

$$\begin{aligned} L_t &= L_{t-1} + B + \alpha\epsilon_t \\ B_t &= B_{t-1} = B \end{aligned}$$

which may be referred to as SES with drift, since the level increases by a fixed amount each period (see SES in Equation (5.7.10)). This method being just a special case of LES, the simpler structure makes it easier to derive an optimal value for B using the estimation sample (see Hyndman and Billah [2003]).

Trigg and Leach [1967] introduced the concept of a tracking signal, whereby not only are the level and slope updated each time, but also the smoothing parameters. In the case of SES, we can use the updated value for α given by

$$\hat{\alpha}_t = \frac{\sum E_t}{\sum M_t}, \alpha_t = \frac{E_{t-1}}{M_{t-1}}$$

where E_t and M_t are smoothed values of the error and the absolute error respectively given by

$$\begin{aligned} E_t &= \delta\epsilon_t + (1-\delta)E_{t-1} \\ M_t &= \delta|\epsilon_t| + (1-\delta)M_{t-1} \end{aligned}$$

where $\delta \in [0.1, 0.2]$. If a string of positive errors occurs, the value of α_t increases, to speed up the adjustment process, the reverse occurs for negative errors. A generally preferred approach is to update the parameter estimate regularly, which is no longer much of a computational problem even for large numbers of series.

One alternative to SES is look at the successive differences in the series $Y_t - Y_{t-1}$ and take a moving average of these values to estimate the slope. The net effect is to estimate the slope by $\frac{Y_t - Y_{t-n}}{n}$ for a n-term moving average. Again, LES usually provides better forecasts.

5.7.3 A summary

Gardner [2006] reviewed the state of the art in exponential smoothing (ES) up to date. He classified and gave formulations for the standard methods of ES which can be modified to create state-space models. For each type of trend, and for each type of seasonality, there are two sections of equations. We first consider recurrence forms (used in the original work by Brown [1959] and Holt [1957]) and then we give error-correction forms (notation follows Gardner [1985]) which are simpler and give equivalent forecasts. Note, there is still no agreement on notation for ES. The notation by Hyndman et al. [2002] and extended by Taylor [2003] is helpful in describing the methods. Each

Table 5.1: List of ES models

Trend Component	N (none)	A (additive)	M (multiplicative)
N (none)	NN	NA	NM
A (additive)	AN	AA	AM
DA (damped-additive)	DA-N	DA-A	DA-M
M (multiplicative)	MN	MA	MM
DM (damped-multiplicative)	DM-N	DM-A	DM-M

method is denoted by one or two letters for the trend and one letter for seasonality. Method (N-N) denotes no trend with no seasonality, or simple exponential smoothing (SES). The other nonseasonal methods are additive trend (A-N), damped additive trend (DA-N), multiplicative trend (M-N), and damped multiplicative trend (DM-N).

All seasonal methods are formulated by extending the methods in Winters [1960]. Note that the forecast equations for the seasonal methods are valid only for a forecast horizon (h) less than or equal to the length of the seasonal cycle (p). Given the smoothing parameter for the level of the series $\alpha \in [0,1]$, the smoothing parameter for the trend $\gamma \in [0,1]$, the smoothing parameter for seasonal indices δ, the autoregressive or damping parameter $\phi \in [0,1]$, we let $\epsilon_t = Y_t - F_t$ be the one-step-ahead forecast error with $F_t = F_{t|t-1} = \hat{Y}_{t-1}(1)$ is the one-step ahead forecast, and we get

1. (N-N)

$$\begin{aligned} S_t &= \alpha Y_t + (1-\alpha) S_{t-1} \\ \hat{Y}_t(h) &= F_{t+h|t} = S_t \end{aligned}$$

and

$$\begin{aligned} S_t &= S_{t-1} + \alpha \epsilon_t \\ F_{t+h|t} &= S_t \end{aligned}$$

2. (A-N)

$$\begin{aligned} S_t &= \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1}) \\ T_t &= \gamma(S_t - S_{t-1}) + (1-\gamma) T_{t-1} \\ F_{t+h|t} &= S_t + hT_t \end{aligned}$$

and

$$\begin{aligned} S_t &= S_{t-1} + T_{t-1} + \alpha \epsilon_t \\ T_t &= T_{t-1} + \alpha\gamma \epsilon_t \\ F_{t+h|t} &= S_t + hT_t \end{aligned}$$

3. (DA-N)

$$\begin{aligned} S_t &= \alpha Y_t + (1-\alpha)(S_{t-1} + \phi T_{t-1}) \\ T_t &= \gamma(S_t - S_{t-1}) + (1-\gamma)\phi T_{t-1} \\ F_{t+h|t} &= S_t + \sum_{i=1}^{h} \phi^i T_t \end{aligned}$$

and

$$
\begin{aligned}
S_t &= S_{t-1} + \phi T_{t-1} + \alpha\epsilon_t \\
T_t &= \phi T_{t-1} + \alpha\gamma\epsilon_t \\
F_{t+h|t} &= S_t + \sum_{i=1}^{h} \phi^i T_t
\end{aligned}
$$

4. (M-N)

$$
\begin{aligned}
S_t &= \alpha Y_t + (1-\alpha) S_{t-1} R_{t-1} \\
R_t &= \gamma \frac{S_t}{S_{t-1}} + (1-\gamma) R_{t-1} \\
F_{t+h|t} &= S_t R_t^h
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} T_{t-1} + \alpha\epsilon_t \\
R_t &= R_{t-1} + \alpha\gamma \frac{\epsilon_t}{S_{t-1}} \\
F_{t+h|t} &= S_t R_t^h
\end{aligned}
$$

5. (DM-N)

$$
\begin{aligned}
S_t &= \alpha Y_t + (1-\alpha) S_{t-1} R_{t-1}^{\phi} \\
R_t &= \gamma \frac{S_t}{S_{t-1}} + (1-\gamma) R_{t-1}^{\phi} \\
F_{t+h|t} &= S_t R_t^{\sum_{i=1}^{h} \phi^i}
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} R_{t-1}^{\phi} + \alpha\epsilon_t \\
R_t &= R_{t-1}^{\phi} + \alpha\gamma \frac{\epsilon_t}{S_{t-1}} \\
F_{t+h|t} &= S_t R_t^{\sum_{i=1}^{h} \phi^i}
\end{aligned}
$$

6. (N-A)

$$
\begin{aligned}
S_t &= \alpha(Y_t - I_{t-p}) + (1-\alpha) S_{t-1} \\
I_t &= \delta(Y_t - S_t) + (1-\delta) I_{t-p} \\
F_{t+h|t} &= S_t + I_{t-p+h}
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} + \alpha\epsilon_t \\
I_t &= I_{t-p} + \delta(1-\alpha)\epsilon_t \\
F_{t+h|t} &= S_t + I_{t-p+h}
\end{aligned}
$$

7. (A-A)

$$
\begin{aligned}
S_t &= \alpha(Y_t - I_{t-p}) + (1-\alpha)(S_{t-1} + T_{t-1}) \\
T_t &= \gamma(S_t - S_t) + (1-\gamma)T_{t-1} \\
I_t &= \delta(Y_t - S_t) + (1-\delta)I_{t-p} \\
F_{t+h|t} &= S_t + hT_t + I_{t-p+h}
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} + T_{t-1} + \alpha\epsilon_t \\
T_t &= T_{t-1} + \alpha\gamma\epsilon_t \\
I_t &= I_{t-p} + \delta(1-\alpha)\epsilon_t \\
F_{t+h|t} &= S_t + hT_t + I_{t-p+h}
\end{aligned}
$$

8. (N-M)

$$
\begin{aligned}
S_t &= \alpha\frac{Y_t}{I_{t-p}} + (1-\alpha)S_{t-1} \\
I_t &= \delta\frac{Y_t}{S_t} + (1-\delta)I_{t-p} \\
F_{t+h|t} &= S_t I_{t-p+h}
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} + \alpha\frac{\epsilon_t}{I_{t-p}} \\
I_t &= I_{t-p} + \delta(1-\alpha)\frac{\epsilon_t}{S_t} \\
F_{t+h|t} &= S_t I_{t-p+h}
\end{aligned}
$$

9. (A-M)

$$
\begin{aligned}
S_t &= \alpha\frac{Y_t}{I_{t-p}} + (1-\alpha)(S_{t-1} + T_{t-1}) \\
T_t &= \gamma(S_t - S_t) + (1-\gamma)T_{t-1} \\
I_t &= \delta\frac{Y_t}{S_t} + (1-\delta)I_{t-p} \\
F_{t+h|t} &= (S_t + hT_t)I_{t-p+h}
\end{aligned}
$$

and

$$
\begin{aligned}
S_t &= S_{t-1} + T_{t-1} + \alpha\frac{\epsilon_t}{I_{t-p}} \\
T_t &= T_{t-1} + \alpha\gamma\frac{\epsilon_t}{I_{t-p}} \\
I_t &= I_{t-p} + \delta(1-\alpha)\frac{\epsilon_t}{S_t} \\
F_{t+h|t} &= (S_t + hT_t)I_{t-p+h}
\end{aligned}
$$

where S_t is the smoothed level of the series, T_t is the smoothed additive trend at the end of period t, R_t is the smoothed multiplicative trend, I_t is the smoothed seasonal index at time t, h is the number of periods in the forecast lead-time, and p is the number of periods in the seasonal cycle.

***Remark* 5.7.3** *When forecasting time series with ES, it is generally assumed that the most common time series in business are inherently non-negative. Therefore, it is of interest to consider the properties of the potential stochastic models underlying ES when applied to non-negative data. It is clearly a problem when forecasting financial returns as the multiplicative error models are not well defined if there are zeros or negative values in the data.*

The (DA-N) method can be used to forecast multiplicative trends with the autoregressive or damping parameter ϕ restricted to the range $1 < \phi < 2$, a method sometimes called generalised Holt. In hopes of producing more robust forecasts, Taylor's [2003] methods (DM-N, DM-A, and DM-M) add a damping parameter $\phi < 1$ to Pegels' [1969] multiplicative trends. Each exponential smoothing method above is equivalent to one or more stochastic models. The possibilities include regression, ARIMA, and state-space models. The most important property of exponential smoothing is robustness. Note, the damped multiplicative trends are the only new methods creating new forecast profiles since 1985. The forecast profiles for Taylor's methods will eventually approach a horizontal nonseasonal or seasonally adjusted asymptote, but in the near term, different values of ϕ can produce forecast profiles that are convex, nearly linear, or even concave.

There are many equivalent state-space models for each of the methods described in the above table. In the framework of Hyndman et al. [2002] each ES method in the table (except the DM methods) has two corresponding state-space models, each with a single source of error (SSOE), one with an additive error and the other with a multiplicative error. The methods corresponding to the framework of Hyndman et al. are the same as the ones in the table appart from two exceptions where one has to modify all multiplicative seasonal methods and all damped additive trend methods. Each ES method in the table is equivalent to one or more stochastic models, including regression, ARIMA, and state-space models. In large samples, ES is equivalent to an exponentially-weighted or DLS regression model. General exponential smoothing (GES) also relies on DLS regression with one or two discount factor to fit a variety of functions of time to the data, including polynomials, exponentials, sinusoids, and their sums and products (see Gardner [1985]). Gijbels et al. [1999] and Taylor [2004c] showed that GES can be viewed in a kernel regression framework. For instance simple smoothing $(N-N)$ is a zero-degree local polynomial kernel model. They showed that choosing the minimum-MSE parameter in simple smoothing is equivalent to choosing the regression bandwidth by cross-validation, a procedure that divides the data into two disjoint sets, with the model fitted in one set and validated in another.

All linear exponential smoothing methods have equivalent ARIMA models which can be easily shown through the DA-N method containing at least six ARIMA models as special cases (see Gardner et al. [1988]). If $0 < \theta < 1$ then the DA-N method is equivalent to the ARIMA (1, 1, 2) model, which can be written as

$$(1-B)(1-\phi B)Y_t = \big[1-(1+\phi-\alpha-\phi\alpha\gamma)B-\phi(\alpha-1)B^2\big]\epsilon_t$$

We obtain an $ARIMA(1,1,1)$ model by setting $\alpha = 1$. When $\alpha = \gamma = 1$ the model is $ARIMA(1,1,0)$. When $\phi = 1$ we have a linear trend (A-N) and the model is $ARIMA(0,2,2)$

$$(1-B)^2Y_t = \big[1-(2-\alpha-\alpha\gamma)B-(\alpha-1)B^2\big]\epsilon_t$$

When $\phi = 0$ we have simple smoothing (N-N) and the equivalent $ARIMA(0,1,1)$ model

$$(1-B)Y_t = \big[1-(1-\alpha)\big]\epsilon_t$$

The $ARIMA(0,1,0)$ random walk model can be obtained from the above equation by choosing $\alpha = 1$. Note, ARIMA-equivalent seasonal models for the linear exponential smoothing methods exist.

The equivalent ARIMA models do not extend to the nonlinear exponential smoothing methods. Prior to the work by Ord et al. [1997] (OKS), state-space models for ES were formulated using multiple sources of error (MSOE). For instance, the exponential smoothing (N-N) is optimal for a model with two sources of error (see Muth [1960]) where observation and state equations are given by

$$Y_t = L_t + \nu_t$$
$$L_t = L_{t-1} + \eta_t$$

so that the unobserved state variable L_t denotes the local level at time t, and the error terms ν_t and η_t are generated by independent white noise processes. Various authors showed that simple smoothing SES is optimal with α determined by the ratio of the variances of the noise processes (see Chatfield [1996]). Harvey [1984] also showed that the Kalman filter for the above equations reduces to simple smoothing in the steady state.

Due to the limitation of the MSOE, Ord et al. [1997] created a general, yet simple class of state-space models with a single source of error (SSOE). For example, the SSOE model with additive errors for the (N-N) model is given by

$$Y_t = L_{t-1} + \epsilon_t$$
$$L_t = L_{t-1} + \alpha\epsilon_t$$

where the error term ϵ_t in the observation equation is the one-step ahead forecast error assuming knowledge of the level at time $t-1$. For the multiplicative error (N-N) model, we alter the additive-error SSOE model and get

$$Y_t = L_{t-1} + L_t\epsilon_t$$
$$L_t = L_{t-1}(1+\alpha\epsilon_t) = L_{t-1} + \alpha L_{t-1}\epsilon_t$$

where the one-step ahead forecast error is still $Y_t - L_{t-1}$ which is no-longer the same as ϵ_t. Hence, the above state equation becomes

$$L_t = L_{t-1} + \alpha L_{t-1}\frac{Y_t - L_{t-1}}{L_{t-1}} = L_{t-1} + \alpha(Y_t - L_{t-1})$$

where the multiplicative error state equation can be written in the error correction form of simple smoothing. As a result, the state equations are the same in the additive and multiplicative error cases, and this is true for all SSOE models. Hyndman et al. [2002] extended the class of SSOE models by Ord et al. [1997] to include all the methods of ES in the above table except from the DM methods. The theoretical advantage of the SSOE approach to ES is that the errors can depend on the other components of the time series. That is, each of the linear exponential smoothing (LES) models with additive errors has an ARIMA equivalent, but the linear models with multiplicative errors and the nonlinear models are beyond the scope of the ARIMA class. The equivalent models help explain the general robustness of exponential smoothing. Simple smoothing (N-N) is certainly the most robust forecasting method and has performed well in many types of series not generated by the equivalent $ARIMA(0,1,1)$ process. Such series include the common first-order autoregressive processes and a number of lower-order ARIMA processes. Bossons [1966] showed that simple smoothing is generally insensitive to specification error, especially when the misspecification arises from an incorrect belief in the stationarity of the generating process. Similarly, Hyndman [2001] showed that ARIMA model selection errors can inflate MSEs compared to simple smoothing. Using AIC to select the best model, the ARIMA forecast MSEs were significantly larger than those of simple smoothing due to incorrect model selections, and becoming worse when the errors were non-normal.

5.7.4 Model fitting

When considering method selection, the definitions of aggregate and individual method selection in the work of Fildes [1992] are useful in exponential smoothing. Aggregate selection is the choice of a single method for all time series

in a population, while individual selection is the choice of a method for each series. While in aggregate selection it is difficult to beat the damped-trend version of exponential smoothing, in individual selection it may be possible to beat the damped trend, but it is not clear how one should proceed. Even though individual method selection can be done in a variety of ways, such as time series characteristics, the most sophisticated approach to method selection is through information criteria.

Various expert systems for individual selection have been proposed, among which the Collopy et al. [1992] including 99 rules constructed from time series characteristics and domain knowledge, combining the forecasts from four methods: a random walk, time series regression, double exponential smoothing, and the (A-N) method. This approach requiring considerable human intervention in identifying features of time series, Vokurka et al. [1996] developed a completely automatic expert system selecting from a different set of candidate methods: the (N-N) and (DA-N) methods, classical decomposition, and a combination of all candidates. Testing their systems using 126 annual time series from the $M1$ competition, they concluded that they were more accurate than various alternatives. Gardner [1999] considered the aggregate selection of the (DA-N) method and showed that it was more accurate at all forecast horizons than either version of rule-based forecasting.

Numerous information criteria that can distinguish between additive and multiplicative seasonality are available for selection of an ES method, but the computational burden can be significant. For instance, Hyndman et al. [2002] recommended fitting all models (from their set of 24 alternatives) to time series, then selecting the one minimising the AIC. In the $1,001$ series ($M1$ and $M3$ data), for the average of all forecast horizons, the (DA-N) method was better than individual selection using the AIC. Later work by Billah et al. [2005] compared eight information criteria used to select from four ES methods, including AIC, BIC, and other standards, as well as two Empirical Information Criteria (EIC) (a linear and a non-linear function) penalising the likelihood of the data by a function of the number of parameters in the model.

Although state-space models for exponential smoothing dominate the recent literature, very little has been done on the identification of such models as opposed to selection using information criteria. Koehler et al. [1988] identified and fitted MSOE state-space models to 60 time series from the 111 series in the $M1$ competition with a semi-automatic fitting routine. In general, the identification process was disappointing. Rather than attempt to identify a model, we could attempt to identify the best exponential smoothing method directly. Chatfield et al. [1988] call this a thoughtful use of exponential smoothing methods that are usually regarded as automatic. They gave a common-sense strategy for identifying the most appropriate method for the Holt-Winters class (see also Chatfield [2002]). Gardner 2006 gave the strategy in a nutshell.

1. We plot the series and look for trend, seasonal variation, outliers, and changes in structure that may be slow or sudden and may indicate that ES is not appropriate in the first place. We should examine any outliers, consider making adjustments, and then decide on the form of the trend and seasonal variation. At this point, we should also consider the possibility of transforming the data, either to stabilise the variance or to make the seasonal effect additive.

2. We fit an appropriate method, produce forecasts, and check the adequacy of the method by examining the one-step-ahead forecast errors, particularly their autocorrelation function.

3. The findings may lead to a different method or a modification of the selected method.

In order to implement an ES method, the user must choose parameters, either fixed or adaptive, as well as initial values and loss functions. Parameter selection is not independent of initial values and loss functions. Note, in the trend and seasonal models, the response surface is not necessarily convex so that one need to start any search routine from several different points to evaluate local minima. We hope that our search routine comes to rest at a set of invertible parameters, but this may not happen. Invertible parameters create a model in which each forecast can be written as a

linear combination of all past observations, with the absolute value of the weight on each observation less than one, and with recent observations weighted more heavily than older ones. If we view an ES method as a system of linear difference equations, a stable system has an impulse response that decays to zero over time. The stability region for parameters in control theory is the same as the invertibility region in time series analysis. In the linear non-seasonal methods, the parameters are always invertible if they are chosen in the interval $[0, 1]$. The same conclusion holds for quarterly seasonal methods, but not for monthly seasonal methods, whose invertibility regions are complex (see Sweet [1985]). Non-invertibility usually occurs when one or more parameters fall near boundaries, or when trend and/or seasonal parameters are greater than the level parameter. For all seasonal ES methods, we can test parameters for invertibility using an algorithm by Gardner et al. [1989] assuming that additive and multiplicative invertible regions are identical, but the test may fail to eliminate some troublesome parameters. Archibald [1990] found that some combination of $[0, 1]$ parameters near boundaries fall within the ARIMA invertible region, but the weights on past data diverge. Hence, they concluded that one should be skeptical of parameters near boundaries in all seasonal models.

Once the parameters have been selected, another problem is deciding how frequently they should be updated. Fildes et al. [1998] compared three options for choosing parameters in the (N-N), (A-N), and (DA-N) methods

1. arbitrarily
2. optimise once at the first time origin
3. optimise each time forecasts are made

and found that the best option was to optimise each time forecasts were made. The term adaptive smoothing mean that the parameters are allowed to change automatically in a controlled manner as the characteristics of the time series change. For instance, the Kalman filter can be used to compute the parameter in the (N-N) method. The only adaptive method that has demonstrated significant improvement in forecast accuracy compared to the fixed-parameter (N-N) method is Taylor's [2004a] [2004b] smooth transition exponential smoothing (STES). Smooth transition models are differentiated by at least one parameter that is a continuous function of a transition variable V_t. The formula for the adaptive parameter α_t is a logistic function (see details in Appendix (A.2))

$$\alpha_t = \frac{1}{1 + e^{a+bV_t}}$$

with several possibilities for V_t including ϵ_t, $|\epsilon_t|$, and ϵ_t^2. Whatever the transition variable, the logistic function restricts α_t to $[0, 1]$. The drawback to STES is that model-fitting is required to estimate a and b; thereafter, the method adapts to the data through V_t. In Taylor [2004a], STES was arguably the best method overall in volatility forecasting of stock index data compared to the fixed-parameter version of (N-N) and a range of GARCH and autoregressive models. Note, with financial returns, the mean is often assumed to be zero or a small constant value, and attention turns to predicting the variance. Following the advice of Fildes [1998], Taylor evaluated forecast performance across time. Using the last 18 observations of each series, he computed successive one-step ahead monthly forecasts, for a total of $25,704$ forecasts, and judged by MAPE and median APE, STES was the most accurate method tested, with best results for the MAPE.

Standard ES methods are usually fitted in two steps, by choosing fixed initial values, followed by an independent search for parameters. In contrast, the new state-space methods are usually fitted using maximum likelihood where initial values are less of a concern because they are refined simultaneously with the smoothing parameters during the optimisation process. However, this approach requires significant computation times. The nonlinear programming model introduced by Sergua et al. [2001] optimise initial values and parameters simultaneously. Examining the $M1$ series, Makridakis et al. [1991] measured the effect of different initial values and loss functions in fitting (N-N), (A-N), and (DA-N) methods, using seasonal-adjusted data where appropriate. Initial values were computed by least squares, backcasting, and several simple methods. Loss functions included the MAD, MAPE. median APE, MSE, the sum of cubed errors, and a variety of non-symmetric functions computed by weighting the errors in different ways.

They concluded that initialising by least squares, choosing parameters from the $[0,1]$ interval, and fitting models to minimise the MSE provided satisfactory results.

5.7.5 Prediction intervals and random simulation

Hyndman et al. [2008] [2008b] provided a taxonomy of ES methods with forecasts equivalent to the one from a state space model. This equivalence allows

1. easy calculation of the likelihood, the AIC and other model selection criteria
2. computation of prediction intervals for each method
3. random simulation from the state space model

Following their notation, the ES point forecast equations become

$$\begin{aligned} l_t &= \alpha P_t + (1-\alpha)Q_t \\ b_t &= \beta R_t + (\phi - \beta)b_{t-1} \\ s_t &= \gamma T_t + (1-\gamma)s_{t-m} \end{aligned}$$

where l_t is the series level at time t, b_t is the slope at time t, s_t is the seasonal component of the series and m is the number of seasons in a year. The values of P_t, Q_t, R_t and T_t vary according to which of the cells the method belongs, and α, β, γ and ϕ are constants. For example, for the (N-N) method we get $P_t = Y_t$, $Q_t = l_{t-1}$, $\phi = 1$ and $F_{t+h} = l_t$. Rewriting these equations in their error-correction form, we get

$$\begin{aligned} l_t &= Q_t + \alpha(P_t - Q_t) \\ b_t &= \phi b_{t-1} + \beta(R_t - b_{t-1}) \\ s_t &= s_{t-m} + \gamma(T_t - s_{t-m}) \end{aligned}$$

Setting $\alpha = 0$ we get the method with fixed level (constant over time), setting $\beta = 0$ we get the method with fixed trend, and the method with fixed seasonal pattern is obtained by setting $\gamma = 0$. Hyndman et al. [2008] extended the work of Ord et al. [1997] (OKS) in SSOE to cover all the methods in the classification of the ES and obtained two models, one with additive error and the other one with multiplicative errors giving the same forecasts but different prediction intervals. The general OKS framework involves a state vector X_t and state space equations of the form

$$\begin{aligned} Y_t &= h(X_{t-1}) + k(X_{t-1})\epsilon_t \\ X_t &= f(X_{t-1}) + g(X_{t-1})\epsilon_t \end{aligned}$$

where $\{\epsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ^2. Defining $(l_t, b_t, s_t, s_{t-1}, .., s_{t-(m-1)})$, $e_t = k(X_{t-1})\epsilon_t$ and $\mu_t = h(X_{t-1})$, we get

$$Y_t = \mu_t + e_t$$

For example, in the (N-N) method we get $\mu_t = l_{t-1}$ and $l_t = l_{t-1} + \alpha\epsilon_t$. Note, the model with additive errors is written $Y_t = \mu_t + \epsilon_t$ where $\mu_t = F_{(t-1)+1}$ is the one-step ahead forecast at time $t-1$, so that $k(X_{t-1}) = 1$. The model with multiplicative errors is written as $Y_t = \mu_t(1+\epsilon_t)$ with $k(X_{t-1}) = \mu_t$ and $\epsilon_t = \frac{e_t}{\mu_t} = \frac{(Y_t - \mu_t)}{\mu_t}$ which is a relative error. Note, the multiplicative error models are not well defined if there are zeros or negative values in the data. Further, we should not consider seasonal methods if the data are not quarterly or monthly (or do not have some other seasonal period).

Model parameters are usually estimated with the maximum likelihood function (see Section (3.3.2.2)) while the Akaike Information Criterion (AIC) (Akaike [1973]) and the bias-corrected version (AICC) (Hurvich et al. [1989]) are standard procedures for model selection (see Appendix (D.3)). The use of the OKS enables easy calculation of the likelihood as well as model selection criteria such as the AIC. We let L^* be equal to twice the negative logarithm of the conditional likelihood function

$$L^*(\theta, X_0) = n \log \left(\sum_{t=1}^{n} \frac{e_t^2}{k^2(x_{t-1})}\right) + 2\sum_{t=1}^{n} \log |k(x_{t-1})|$$

with parameters $\theta = (\alpha, \beta, \gamma, \phi)$ and initial states $X_0 = (l_0, b_0, s_0, s_{-1}, .., s_{-m+1})$. They can be estimated by minimising L^*. Estimates can also be obtained by minimising the one-step MSE, the one-step MAPE, the residual variance σ^2 or by using other criterions measuring forecast error. Models are selected by minimising the AIC among all the ES methods

$$AIC = L^*(\hat{\theta}, \hat{X}_0) + 2p$$

where p is the number of parameters in θ, and $\hat{\theta}$ and $\hat{X}_0$ are the estimates of θ and X_0. Note, the AIC penalises against models containing too much parameters, and also provides a method for selecting between the additive and multiplicative error models because it is based on likelihood and not a one-step forecasts. Given initial values for the parameters θ and following a heuristic scheme for the initial state X_0, bla obtained a robust automatic forecasting algorithm (AFA)

- For each series, we apply the appropriate models and optimise the parameters in each case.
- Select the best model according to the AIC.
- Produce forecasts using the best model for a given number of steps ahead
- to obtain prediction intervals, we use a bootstrap method by simulating 5000 future sample paths for $\{Y_{n+1}, .., Y_{n+h}\}$ and finding the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles of the simulated data at each forecasting horizon. The sample paths are generated by using the normal distribution for errors (parametric bootstrap) or by using the resampled errors (ordinary bootstrap).

Application of the AFA to the M and $M3$ competition data showed that the methodology was very good at short term forecasts (up to about 6 periods ahead).

5.7.6 Random coefficient state space model

Gardner [2009] considered a damped linear trend model with additive errors with ES point equations given by

$$\begin{aligned} y_t &= l_{t-1} + \phi b_{t-1} + \epsilon_t \\ l_t &= l_{t-1} + \phi b_{t-1} + (1-\alpha)\epsilon_t \\ b_t &= \phi b_{t-1} + (1-\beta)\epsilon_t \end{aligned} \tag{5.7.12}$$

where $\{y_t\}$ is the observed series, $\{l_t\}$ is its level, $\{b_t\}$ is the gradient of the linear trend, and $\{\epsilon_t\}$ is the single source of error. Note, these notations are slightly different from the ones in Section (5.7.5) to simplify some of the results. In the special case where $\phi = 1$ we recover the linear trend model corresponding to the $ARIMA(0, 2, 2)$ model

$$(1 - B)^2 y_t = \epsilon_t - (\alpha + \beta)\epsilon_{t-1} + \alpha\epsilon_{t-2}$$

where the gradient of the trend is a random walk. Otherwise for $\phi \neq 1$ we get the $ARIMA(1, 1, 2)$ model

$$(1-\phi B)(1-B)y_t = \epsilon_t - (\alpha + \phi\beta)\epsilon_{t-1} + \phi\alpha\epsilon_{t-2}$$

where the gradient of the trend follows an $AR(1)$ process, and as a result, changes in a stationary way. With ϕ close to 1 the linear trend is highly persistent, but as ϕ moves away from 1 towards zero the trend becomes weakly persistent, and for $\phi = 0$ there is absence of a linear trend. Consequently, we can interpret ϕ as a direct measure of the persistence of the linear trend. In order to assume a locally constant model (Brown), Gardner postulated that we can consider the linear trend model ($\phi = 1$) with a revised gradient each time the local segment of the series changes in a sudden way. He modelled the revision gradient as

$$b_t = A_t b_{t-1} + (1-\beta)\epsilon_t$$

where $\{A_t\}$ is a sequence of i.i.d. binary random variates with $P(A_t = 1) = \phi$ and $P(A_t = 0) = (1-\phi)$. In the case of a strongly persistent trend the sequence $\{A_t\}$ will consist of long runs of 1s interrupted by occasional 0s and vice versa if not persistent. We get a mixture when phi is between 0 and 1 with mean length of such runs given by $\frac{\phi}{(1-\phi)}$. We obtain the random coefficient state space model by replacing ϕ with A_t in the above Equation (5.7.12) and setting (α^*, β^*) to distinguish from the two models. We get a stochastic mixture of two well known forms, the $ARIMA(0,2,2)$ with probability ϕ and the $ARIMA(0,1,1)$ with probability $(1-\phi)$. Gardner [2009] proved that the forecasts in the standard damped trend model as well as the ones in the random coefficient state-space model are optimal, with the same parameter value ϕ, but with different values of α and β. That is, the damped trend forecasts are also optimal for such a more general and broader class of models. This reasoning can be applied to similar models with linear trend component such as the additive seasonal model or linear trend models with multiplicative errors.

Chapter 6

Filtering and forecasting with wavelet analysis

6.1 Introducing wavelet analysis

6.1.1 From spectral analysis to wavelet analysis

6.1.1.1 Spectral analysis

We presented in Section (4.3) the basic principles of trend filtering in the time domain and argued that filtering in the frequency domain was more appropriate. That is, another way of estimating the trend x_t in Equation (4.3.10) is to denoise the signal y_t by using spectral analysis. Fourier analysis (see details in Appendix (G.1)) uses sum of sine and cosine functions at different wavelengths to express almost any given periodic function, and therefore any function with a compact support. We can use the Fourier transform, which is an alternative representation of the original signal y_t, expressed by the frequency function

$$y(w) = \sum_{t=1}^{n} y_t e^{-iwt}$$

where $y(w) = \mathcal{F}(y)$ with w a frequency, and such that $y = \mathcal{F}^{-1}(y)$ with $\mathcal{F}^{-1}$ the inverse Fourier transform. Given the sample $\{y_0, .., y_{n-1}\}$ of a time series, and assuming that the mean has been removed before the analysis, from the Parseval's theorem for the discrete Fourier transform, it follows that the sample variance of $\{y_t\}$ is

$$s^2 = \frac{1}{n}\sum_{t=0}^{n-1} y_t^2 = \frac{1}{n}\sum_{j=0}^{n-1} |y(w_j)|^2$$

where $y(w)$ is the discrete Fourier transform of $\{y_t\}$ and $w_j = \frac{2\pi j}{n}$ for $j = 0, 1, .., [\frac{n}{2}]$ [1] are the Fourier frequencies. Hence, the variance of the series can be decomposed into contributions given by a set of frequencies. The expression $\frac{1}{n}|y(w_j)|^2$, as a function of w_j, is the periodogram of the series, which is an estimator of the true spectrum $f(w)$ of the process, providing an alternative way of looking at the series in the frequency domain rather than in the time domain. If the spectrum of the series peaks at the frequency w_0, it can be concluded that in its Fourier decomposition the component with the frequency w_0 accounts for a large part of the variance of the series. Hence, denoising in spectral analysis consists in setting some coefficients $y(w)$ to zero before reconstructing the signal. Selected parts of the frequency spectrum can be manipulated by filtering tools, some can be attenuated and others may be completely

[1] $[x]$ denotes the integer part of x.

removed. Hence, a smoothing signal can be generated by applying a low-pass filter, that is, by removing the higher frequencies. However, while the Fourier analysis remains an important mathematical tool in many fields of science, the decomposition of a function into simple harmonics of the form Ae^{inw} has some drawbacks. One problem with the Fourier transform is the bad time location for low frequency signals and the bad frequency location for the high frequency signals making it difficult to localise when the trend (located in low frequencies) reverses. That is, the Fourier representation of local events related to a series requires many terms of the form Ae^{inw}. Hence, the non-local characteristic of sine and cosine functions implies that we can only consider stationary signals along the time axis (see Oppenheim et al. [2009]). Even though various methods for time-localising a Fourier transform have been proposed to avoid this problem, such as windowed Fourier transform, the real improvement comes with the development of wavelet theory.

6.1.1.2 Wavelet analysis

Wavelets are the building blocks of wavelet transformations (WT) in the same way that the functions e^{inx} are the building blocks of the ordinary Fourier transformation. Haar [1910] constructed the first known wavelet basis by showing that any continuous function $f(y)$ on $[0,1]$ could be approximated by a series of step functions. Later, Grossmann et al. [1984] introduced Wavelet transform in seismic data analysis, as a solution for analysing time series in terms of the time-frequency dimension. They are defined over a finite domain, localised both in time and in scale, allowing for the data to be described into different frequency component for individual analysis. Wavelets can be (or almost can be) supported on an arbitrarily small closed time interval, making them a very powerful tool in dealing with phenomena rapidly changing in time. A wavelet basis is made of a father wavelet representing the smooth baseline trend and a mother wavelet that is dilated and shifted to construct different level of detail. At high scales, the wavelets have small time support, enabling them to zoom on details and short-lived phenomena. Their abilities to switch between time and scale allow them to escape the Heisenberg's curse stating that one can not analyse both time and frequency with high accuracy. One can then separate signal trends and details using different levels of resolution or different sizes/scales of detail. That is, the transform generate a phase space decomposition defined by two parameters, the scale and location, as opposed to the Fourier decomposition. Several methods exists to compute the wavelet coefficients such as the cascade algorithm of Mallat [1989] and the low-pass and high-pass filters of order 6 proposed by Daubechies [1992]. In digital signal processing, Mallat [1989] discovered the relationship between quadrature mirror filters (QMF) and orthonormal wavelet bases leading to multiresolution analysis which builds on an iterative filter algorithm (pyramid algorithm). It is the cornerstone of the fast wavelet transform (FWT). The last important step in the evolution of wavelet theory is due to Daubechies [1988] who constructed consumer-ready wavelets with a preassigned degree of smoothness. Shensa [1992] clarified the relationship between discrete and continuous wavelet transforms, bringing together two separately motivated implementations of the wavelet transform, namely the algorithme a trous for non-orthogonal wavelets (see Holschneider et al. [1989] and Dutilleux [1989]) and the multiresolution approach of Mallat employing orthonormal wavelets.

6.1.2 The discrete wavelet transform

6.1.2.1 The dyadic DWT

A naive approach to obtaining a detailed picture of the underlying process would be to apply to the data a bank of filters with varying frequencies and widths. However, choosing the proper number and type of filters for this, is a very difficult task. Wavelet transforms (WT) provide a sound mathematical principles for designing and spacing filters, while retaining the original relationships in the time series. While the continuous wavelet transform (CWT) of a continuous function produces a continuum of scales as output, calculating wavelet coefficients at every possible scale takes a fair amount of work, and generates an awful lot of data. Fortunately, we can choose only a subset of scales and positions at which to make our calculations by considering discrete wavelet transform (DWT). DWT transforms discrete (digital) signals to discrete coefficients in the wavelet domain (sampled version of CWT), and can take various forms. For instance, a triangle can be used as a result of decimation, or the retaining, of one sample out of every two,

so that just enough information is kept to allow exact reconstruction of the input data (see details in Appendix (G.3)). For many signals, the low-frequency (LF) content is the most important part giving the signal its identity, while the high-frequency (HF) content imparts flavour or nuance. Mallat [1989] developed an efficient way of implementing this scheme, by using filters known as a two-channel subband coder. Given a time series of length n, the dyadic DWT consits of $\log_2(n)$ stages at most. Starting from the original series, the results include two types of wavelet coefficients at each levels, the approximations and details, where

- the approximations are the high-scale, low-frequency components of the signal.
- the details are the low-scale, high-frequency components.

In general, only the approximation coefficients are analysed, the details corresponding to additive noise. Thus, given an original signal, we keep only one point out of two in each of the two samples to get the complete information, producing two sequences cA and cD called the DWT coefficients. This is the notion of Downsampling. Note, cA is the approximation (smooth) and cD is the details (noise). The decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components, leading to the wavelet decomposition tree.

6.1.2.2 The a trous wavelet decomposition

Even though this approach is ideal for data compression, it can not simply relate information at a given time point at the different scales. Further, it is not possible to have shift invariance. We can get around this problem by means of a redundant, or, non-decimated wavelet transform, such as the a trous algorithm (see details in Appendix (G.4)). Since translation invariant wavelet transform can produce a good local representation of the signal both in the time domain and frequency domain, they have been used to preprocess the data (see Aussem et al. [1998], Gonghui et al. [1999]). We assume that a function f is known only through the time series $\{x_t\}$ consisting of discrete measurements at fixed intervals (see Equation (6.1.2)). We define the signal $S_0(t)$ as the scalar product at samples t of the function $f(x)$ with a scaling function $\phi(x)$

$$S_0(t) =< f(x), \phi(x-t) >$$

where the scaling function satisfies the dilation equation (G.3.17). There exists several ways of constructing a redundant discrete wavelet transform. For instance, we can consider that the successive resolution levels are

- formed by convolving with an increasingly dilated wavelet function which looks like a Mexican hat (central bump, symmetric, two negative side lobes).
- constructed by smoothing with an increasingly dilated scaling function looking like a Gaussian function defined on a fixed support (a B_3 spline).
- constructed by taking the difference between successive versions of the data which are smoothed in this way.

In the a trous wavelet transform (see Shensa [1992]), the input data is decomposed into a set of band-pass filtered components, the wavelet details (or coefficients) plus a low-pass filtered version of the data called residual (or smooth). For illustration, we decompose a time series into a residual and some coefficients with the a trous wavelet decomposition algorithm in Figure (6.1). The smoothed data $S_j(t)$, at a given resolution j and position t, is the scalar product

$$S_j(t) = \frac{1}{2^j} < f(x), \phi(\frac{x-t}{2^j}) >$$

which corresponds to Equation (G.3.13) with $m = j$ and $n = t2^j$. It is equivalent to performing successive convolutions with the discrete lowpass filter h

$$S_{j+1}(t) = \sum_{l=-\infty}^{\infty} h(l) S_j(t + 2^j l)$$

where the finest scale is the original series $S_0(t) = x_t$. The distance between levels increases by a factor 2 from one scale to the next. The name a trous (with holes) results from the increase in the distances between the sampled points ($2^j l$). Hence, we can think of the successive convolutions as a moving average of $2^j l$ increasingly distant points. Here, smoothing with a B_3 spline, the lowpass filter, h, is defined as $\left(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}\right)$, which has compact support and is point symmetric. From the sequence of smoothed representations of the signal, we take the difference between successive smoothed versions, obtaining the wavelet details (or wavelet coefficients)

$$d_j(t) = S_{j-1}(t) - S_j(t)$$

such that

$$\sum_{j=1}^{p} d_j(t) = S_0(t) - S_p(t)$$

for a fixed number of scales p. Note, the wavelet details can also be expressed, independently, as

$$d_j(t) = \frac{1}{2^j} < f(x), \psi(\frac{x-t}{2^j}) >$$

corresponding to the discrete wavelet transform for the resolution level j. The original data is then expanded (reconstructed) as

$$x(t) = S_p(t) + \sum_{j=1}^{p} d_j(t) \tag{6.1.1}$$

for a fixed number of scales p. At each scale j, we obtain a set called wavelet scale, having the same number of samples as the original signal.

6.1.3 Defining the decomposition level

6.1.3.1 The sparsity of the wavelet transform

The performance of wavelet transformations (WT) depends largely on the choice of wavelet type and the decomposition level (DL). Since the analysis process is iterative, it could in theory be continued indefinitely. However, the decomposition can proceed only until the individual details consist of a single sample. Thus, obtaining an optimal DL in DWT becomes a crucial issue. We can select a suitable number of levels based either on the nature of the signal, or on a suitable criterion. We are going to discuss an optimal decomposition with respect to a convenient criterion. For simplicity of exposition we assume that the original signal satisfies

$$X_t = Y_t + \epsilon_t \tag{6.1.2}$$

where Y_t is a deterministic function f, $\epsilon_t \sim N(0, \sigma^2)$ is a white noise component and the noise level σ may be known or unknown. Further, applying the wavelet decomposition to the series, we can reconstruct it as in Equation (6.1.1). Note, the coefficients of WT are usually sparse [2], and become non-zero after contamination with noise. In general, the coefficients with small magnitude can be considered as pure noise and should be set to zero to avoid noisy appearance when reconstructing the signal. Thus, the problem becomes that of recovering the coefficients of f which are relatively stronger than the Gaussian white noise background.

[2] nearly zero in a noiseless wavelet transform.

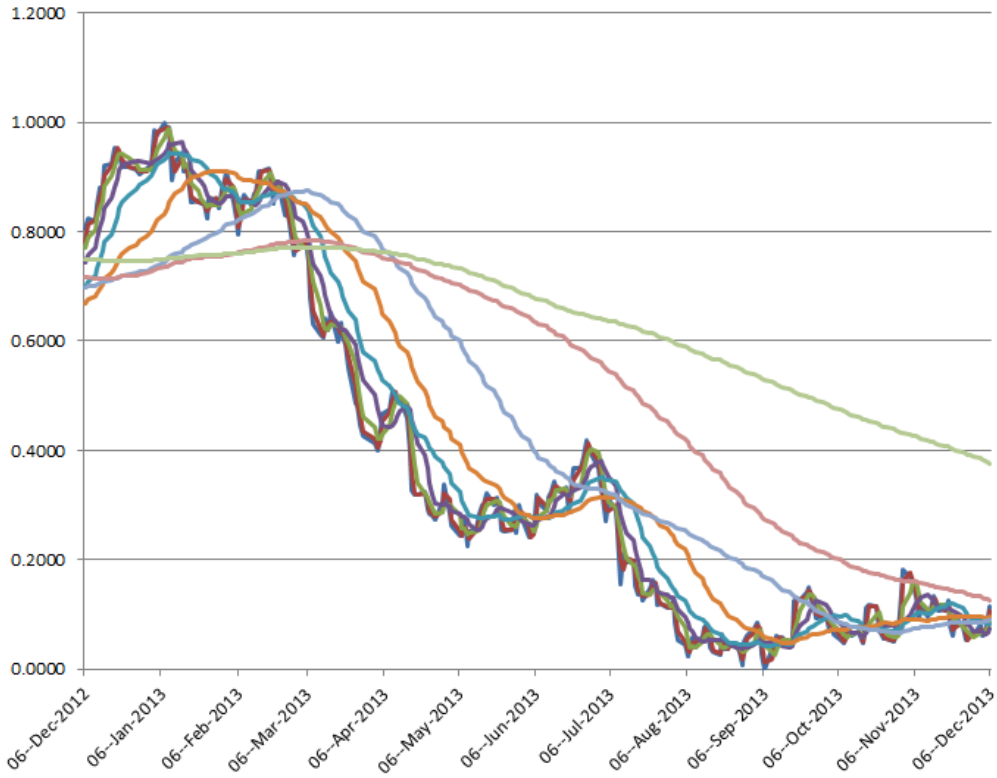


Figure 6.1: Decomposing the original data with the a trous wavelet decomposition algorithm.

***Remark* 6.1.1** *A lot of applications in WT take advantage of the sparseness of the wavelet coefficients to filter white noise. The difference being in the decision process to remove the undesirable coefficients.*

6.1.3.2 The optimal decomposition level

As a first approach we can use the white noise testing method (see Box et al. [1994]), where the noisy series is first separated into sub-signals under different levels, whose autocorrelations are then analised. Once a certain sub-signal fail the test, the corresponding level is considered the best result. However, noisy series usually generates autocorrelated sub-signals, making the WNT results inaccurate and unreliable in many practical situations. On the other hand, functions verifying an additivity-type property are well suited for efficient searching of binary-tree structures and the fundamental splitting. Classical entropy-based criteria match these conditions and describe information-related properties for an accurate representation of a given signal. Sang et al. [2010] proposed to measure the entropy of transformed signal, and let the decomposing process stops when the resultant entropy becomes significantly different from an artificially generated noise series. Decomposing the original signal to extract its noise, they used the wavelet energy entropy (WEE) to describe the variation of the degrees of complexity of noise with DLs. According to the information entropy theory, the WEE charactersises the complexity of the series (see Section (13.2) for details on Information Theory). Assuming J to be the maximum DL, and applying dyadic DWT to the noise series, they reconstructed the sub-signal under each level j and calculated the WEE as

$$WEE(j) = -\sum_{k}^{j} P_k \ln P_k \ , j = 1, ..., J$$

with

$$P_k = \frac{E_k}{\sum_k^j E_k}$$

where $E_k = \sum_{t=1}^n f_k^2(t)$ is the energy at the kth DL, n is the length of the time series, t is the data number and $f_k(t)$ is the sub-signal under the kth DL. Then they chose empirically an appropriate probability distribution to generate normalised noise series and determine its WEE via Monte-Carlo simulations. At last, the values of the WEE obtained using both methods were compared, and the optimal DL identified when they differed. That is, the choice of the optimal DL is based on the difference of energy distributions between noisy series and some defined noise. However, this method requires prior information of the distribution model of the noise, which is generally unknown. Instead, Liao et al. [2013] proposed to exploit the sparsity of the transformed data after DWT has been performed. To quantify the degree of sparseness, they evaluated the percentage of the number of zero/near-zero coefficients among the entire transformed coefficients

$$sp = \frac{N_o}{N-1}$$

where N_o is the number of zero/near-zero coefficients and N is the length of the original signal. They proposed to estimate N_o by regarding a near-zero coefficient as having an absolute value smaller than that of the largest coefficient divided by a constant $K \in [5, 10]$. That is,

$$|d_j(t)| \leq \overline{d}$$

where $\overline{d} = \frac{\max\{\sum_t \sum_j |d_j(t)|\}}{K}$. Note, the constant K depends crucially on the volatility of the original signal. As an alternative, we propose to normalise the coefficients $d_j(t)$ for all j and all t with Equation (12.3.1) in the range $[0, 1]$, and assume J to be the maximum decomposition level. We define the number of zero/near-zero coefficients at scale j as

$$\widehat{d}_j(t) \leq \epsilon$$

where $\widehat{d}_j(t)$ is the normalised coefficient and $\epsilon << 1$. We can then evaluate the probability p_j of the number of zero/near-zero coefficients at the jth scale, among the entire transformed coefficients

$$p_j = \frac{N_j}{\overline{N}}, j = 1, ..., J$$

where N_j is the number of zero/near-zero coefficients at scale j, and $\overline{N}$ is the total number of coefficients in the wavelet decomposition. The sum of propabilities being equal to one, we cumulate the probabilities, getting

$$\widehat{p}_j = \sum_{k=1}^{j} p_k, j = 1, ..., J$$

and stop when $\widehat{p}_j \geq \alpha$ where α is a threshold in $[0, 1]$. If the signal is highly volatile, the high-scale coefficients (low-frequency) will oscilate around zero, thus having a non-negligible impact on the cumulated probabilities. On the other hand, for a slightly volatile signal the high-scale coefficients will hardly cross the x-axis, and only the low-scale coefficients will matter. As an example, we decompose a volatile time series into a residual and some coefficients with the a trous wavelet decomposition algorithm. We show in Figure (6.2) the wavelet decomposition, the coefficients are displayed in Figure (6.3) and the cumulated probabilities are graphed in Figure (6.4).

Figure 6.2: Decomposing volatile signal with the a trous wavelet decomposition algorithm.

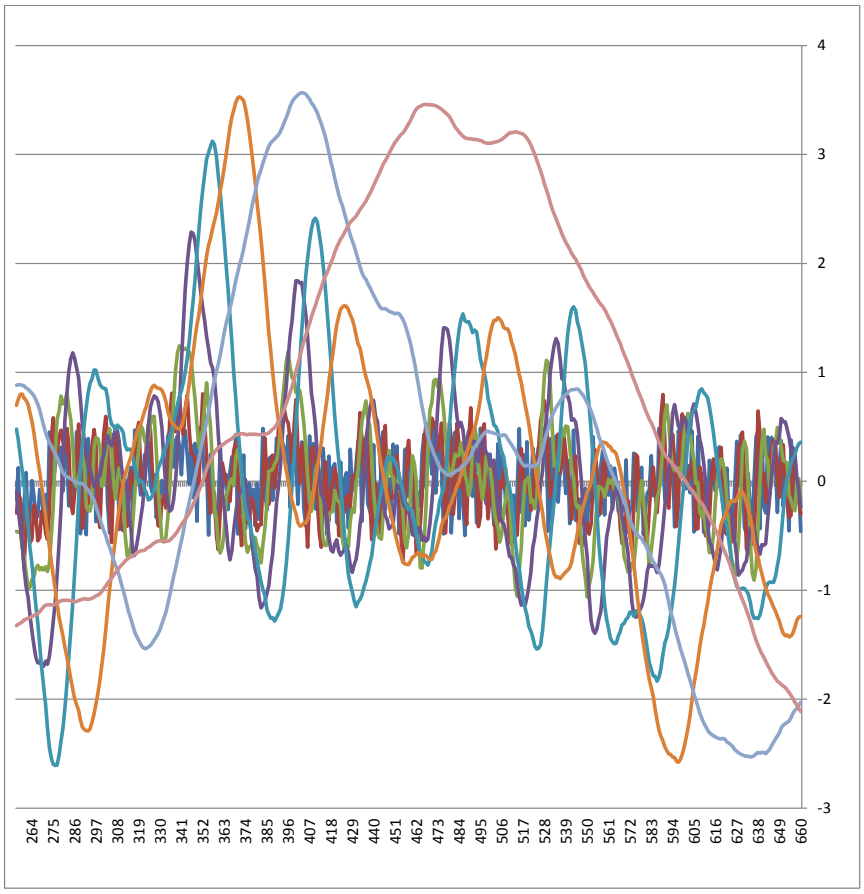


Figure 6.3: The coefficients of the volatile signal.

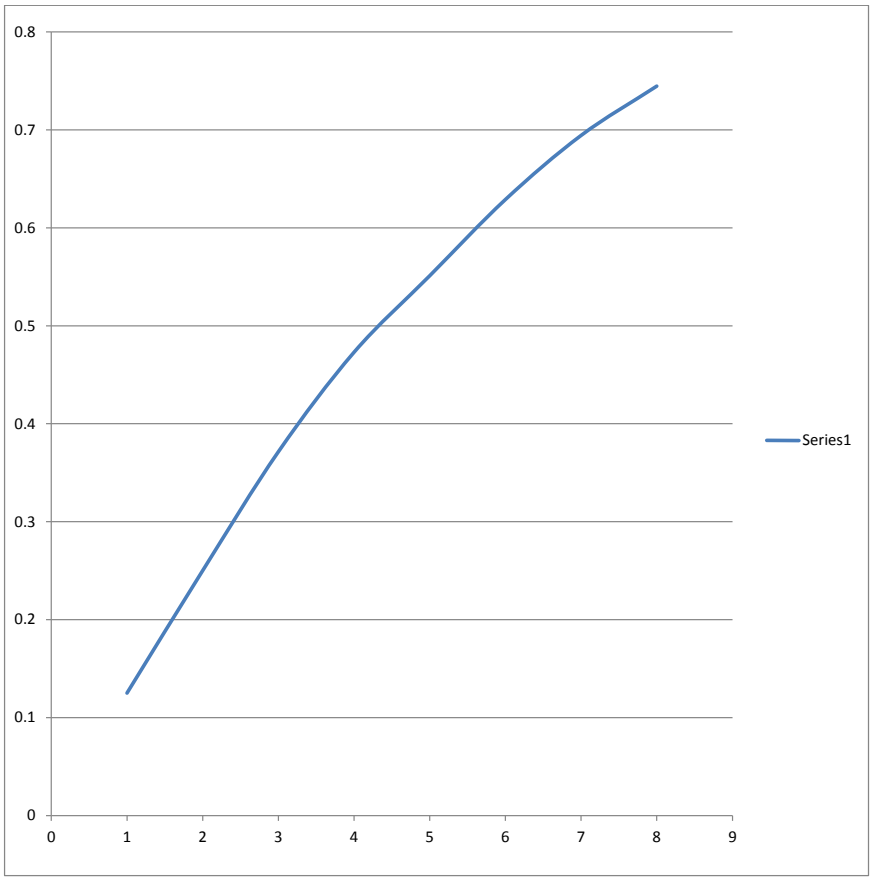


Figure 6.4: The cumulated probabilities versus scales

6.2 Some applications

Even though time series can be corrupted with various types of noise (additive noise, non-addditive noise, Poisson/Laplace noise, ...), we focus on additive white Gaussian noise (AWGN).

6.2.1 A brief review

Wavelets are ideal for frequency domain analysis in time-series econometrics as their capability to simultaneously capture long-term movements and high frequency details are very useful when dealing with non-stationary and complex functions. They can also be used in connection with fractionally integrated process having long-memory properties. It was shown that when decomposing time series with long-term memory, the processes of wavelet coefficients at each scale lack this feature (see Soltani et al. [2000]), enhancing forecasting. Further, decomposing a time series into different scales may reveal details that can be interpreted on theoretical grounds and can be used to improve forecast accuracy. In the former, economic actions and decision making take place at different scales, and in the latter, forecasting seems to improve at the scale level as models like autoregressive moving average (ARMA) or neural networks can extract information from the different scales that are hidden in the aggregate.

Using wavelet transform (WT) we can decompose a time series into a linear combination of different frequencies and then hopefully quantify the influence of patterns with certain frequencies at a certain time, thus, improving the quality of forecasting. For instance, Conejo et al. [2005] decomposed the time series into a sum of processes with different frequencies, and forecasted the individual time series before adding up the results. It is assumed that the motions on different frequencies follow different underlying processes and that treating them separately could increase the forecasting quality. As a result, several approaches have been proposed for time-series filtering and prediction by the wavelet transform (WT), based on neural networks (see Aussem et al. [1998], Zheng et al. [1999]), Kalman filtering (see Cristi et al. [2000]), AR and GARCH models (see Soltani et al. [2000], Renaud et al. [2002]).

There exists different wavelet based forecasting methods such as using WT to eliminate noise in the data or to estimate the components in a structural time series model (STSM) (see Section (3.2.3.1)). Alternatively, we can perform the forecasting directly on the wavelet generated time series decomposition, or we can use locally stationary wavelet processes. We are going to discuss three wavelet based forecasting methods:

1. Wavelet denoising: it is based on the assumption that a data set $(X_t)_{t=1,..,T}$ can be written as in Equation (6.1.2)

$$X_t = Y_t + \epsilon_t$$

   where $\epsilon_t \sim N(0, \sigma^2)$ is a white noise component. Reducing the noise via thresholding yields a modified X_t on which the standard forecasting methods can be applied (see Alrumaih et al. [2002]). More recently, tree-based wavelet denoising methods were developed in the context of image denoising, which exploit tree structures of wavelet coefficients and parent-child correlations.

2. Decomposition tool: we can also decompose the process X_t into components (STSM), such as

$$X_t = T_t + I_t + \epsilon_t$$

   where T_t is a trend and I_t is a seasonality component, and we can do the forecasting by extrapolating from polynomial functions. For example, Wong et al. [2003] used the hidden periodicity analysis to estimate the trend and fitted an $ARIMA(1,0)$ to forecast the noise.

3. Forecasting wavelet coefficients: we can decompose the time series X_t with the wavelet coefficients $T(a,b)$ with $a \in A$, $b = 1,..,T$ where A denotes a scale discretisation. For each a, the corresponding vector $T(a) = (T(a,1),..,T(a,T))$ is treated as a time series, and standard techniques like ARMA-based forecasting are applied to obtain wavelet coefficient forecasts, which are subsequently added to the matrix $T^{'}(a,b)$ (see Conejo et al. [2005]). Note, Renaud et al. [2002] [2005] only used specific coefficients for this forecast which is more efficient but increases the forecasting error. The extended matrix $T^{'}$ is then inverted and we yield a forecast $\hat{X}_{t+1}$ for the value X_t in the series.

6.2.2 Filtering with wavelets

We saw above that denoising could be divided into two types: denoising in the original signal domain (time or space) and denoising in the transform domain (Fourier or wavelet transform). When filtering out white noise in spectral analysis, hard thresholding consists in setting equal to zero all coefficients in the frequency domain below a certain bandwidth. The filtered series are then transformed back into the time domain. Similarly, in the wavelet analysis each coefficient is compared to a threshold level in order to decide if it is a deribable part of the original signal or not. While the method of denoising is the same as for the Fourier analysis, the estimation of the trend x_t can be done in three steps

1. compute the wavelet transform T of the original signal y_t to obtain the wavelet coefficients $w = T(y)$
2. modify the wavelet coefficients according to the denoising rule D, that is,
$$w^* = D(w)$$
3. convert the modified wavelet coefficients into a new signal (de-noised series) using the inverse wavelet transform
$$x = T^{-1}(w^*)$$

The difference between the de-noised series and the original one is called the separated noise. Hence, to perform this method we first need to specify the mother wavelet, the choice of the decomposition level, thresholding estimation and then we must define the denoising rule (see Sang et al. [2009]). Wavelet thresholding and shrinkage in statistics were introduced and explored in a series of papers by Donoho et al. [1995] [1995b]. It consists in shrinking the wavelet image of the original data set and returning the shrunk version of the data domain by the inverse wavelet transformation. This results in the original data being denoised or compressed. More specifically, given the thresholds w^- and w^+ two scalars with $0 < w^- < w^+$ acting as tuning parameters of the wavelet shrinkage, Donoho et al. [1995] defined several shrinkage methods

- Hard shrinkage: consists in setting to 0 all wavelet coefficients having an absolute value lower than a threshold w^+.
$$w_j^* = w_i I_{\{|w_j|>w^+\}}$$
- Soft shrinkage: consists in replacing each wavelet coefficient by the value w^* where
$$w_j^* = sign(w_j)(|w_j| - w^+)^+$$
where $(x)^+ = \max(x, 0)$ and $sign(x) = I_{\{x>0\}} - I_{\{x<0\}}$.
- Semi-soft shrinkage
$$w_j^* = \begin{cases} 0 \text{ if } |w_j| \leq w^- \\ sign(w_j)(w^+ - w^-)^{-1}w^+(|w_j| - w^-) \text{ if } w^- < |w_j| \leq w^+ \\ w_j \text{ if } |w_j| > w^+ \end{cases}$$
- Quantile shrinkage is a hard shrinkage method where w^+ is the qth quantile of the coefficients $|w_j|$

The hard thresholding rule is referred as wavelet thresholding, whereas the soft thresholding rule is referred to as shrinkage as it shrinks the coefficients with high amplitude toward zero. Note, the threshold level w could be a function of the decomposition level j and the index t. For instance, Donoho [1995] introduced the SureShrink which achieves adaptivity through sparsity considerations. It is based on Stein's unbiased risk estimate (SURE) (see Stein [1981]) which consists in estimating the mean vector $\mu = (\mu_1, ..., \mu_n)$ from $X = (X_1, ..., X_n)$ with minimum l_2 risk,

that is, finding $\widehat{\mu} = \min_{\widehat{\mu}} ||\mu - \widehat{\mu}||_2$. Since μ is unknown, for any estimator written as $\widehat{\mu}(X) = X + g(X)$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ and g is weakly differentiable, the l_2 risk can be estimated unbiasedly as

$$SURE(\widehat{\mu}) = E_\mu||\mu - \widehat{\mu}||_2^2 = n + E_\mu[||g(X)||_2^2 + 2\nabla g(X)]$$

where $\nabla g(X) = \sum_{t=1}^n \frac{\partial g_t}{\partial x_t}$ is the divergence of g. Setting $w_j(t)$ as X and assuming the soft thresholding rule, then $w^*(X) = X + g(X)$ where

$$||g(X)||_2^2 = \sum_{t=1}^n \big(\min(|X_t|, w)\big)^2, \nabla g(X) = -\sum_{t=1}^n I_{\{-w,w\}}(X_t)$$

We can then obtain an analytic expression for the SURE as

$$SURE(w; X) = n + \sum_{t=1}^n \big(\min(|X_t|, w)\big)^2 - 2\#(i : |X_t| < w)$$

where $\#$ denotes cardinality. One should choose the threshold level w so as to minimise the SURE

$$w^* = \arg \min_{0 \le w \le \sqrt{2log(n)}} SURE(w; X)$$

which does not take into account values greater

$$w_U = \sqrt{2log(n)}$$

known as the universal bound. Setting w to w_U is called VisuShrink, but it does not adapt to the data. So, one should solve the optimisation problem above. Several thresholds as well as several thresholding policies have been proposed (see Vidakovic [1999]). Other denoising methods were proposed, among which is the NeighBlock presented by Cai et al. [2001], which incorporates information on neighbouring coefficients into the decision process. The main advantage of wavelet shrinkage is that denoising is carried out without smoothing out sharp structures such as spikes and cusps. Review of wavelet shrinkage and wavelet thresholding methods can be found in Antoniadis et al. [2001] and Cohen [2012].

6.2.3 Non-stationarity

The hierarchical construction of wavelets means that non-stationary components of time series are absorbed by the lower scales, while non-lasting disturbances are captured by the higher scales, leading to the whitening property (see Vidakovic [1999]). For example, when generating a sample data from an $ARIMA(2, 1, 1)$, the autocorrelation function shows almost no decay over 20 period, since the process contains a unit root. However, the autocorrelation of the differenced $ARMA(2, 1)$ shows a drop after two periods. Computing the autocorrelation functions of the six highest layers of the original series decomposed with wavelet transform, the ninth scale resemble white noise, while scale six and four show autocorrelation at all included lags.

6.2.4 Decomposition tool for seasonality extraction

One of the fundamental advantages of wavelet analysis is its capacity to decompose time series into different components. Forecasting is one of the reasons for decomposing a series, since it can be easier to forecast the components of the series than the whole series itself. Arino et al. [1995] used the scalogram to decompose the energy into level components in order to detect and separate periodic components in time series. To illustrate their approach, they considered two perfect periodic functions with different frequencies added up together. To filter each component of the combined signal, they used the scalogram and looked at how much energy was contained in each scale of the wavelet transform. Since two peakes were observed, they split the wavelet decomposition d of the time series $\{y_t\}$ into two

new wavelet decompositions, $d^{(1)}$ and $d^{(2)}$, such that the coefficients $d_{j,k}$ of d which are in level j close to the first peak are assigned to $d^{(1)}$ ($d^{(1)}_{j,k} = d_{j,k}$). The corresponding coefficients in d^2 are set to zero ($d^{(2)}_{j,k} = 0$). The same is done for the coefficients close to the second peak. When a level occurs between the two peaks, the coefficients of that level split in two according to two different methods. In the first one, the split is additive with respect to energies but not with respect to wavelet coefficients

$$(d_{j,k})^2 = (d^{(1)}_{j,k})^2 + (d^{(2)}_{j,k})^2$$

It is not additive in the scale domain. In the second one, the split is additive with respect to wavlet coefficients but does not preserve energies

$$d_{j,k} = d^{(1)}_{j,k} + d^{(2)}_{j,k}$$

Following the same approach, Schleicher [2002] illustrated the separation of frequency levels on two sine functions with different frequencies added up. It was done by adding up the squared coefficients in each scale to get a scalogram. Doing so, he observed two spikes, one at the third level and one at the sixth level. Since slow-moving, low frequency components are represented with larger support, he conjectured that level three represented the wavelet transform for function one, and that level six represented that for function two. To filter out function one, he kept only the first four levels and pad the rest of the wavelet transform to zeros, and then took the inverse transform. Conversely, levels five to nine for the second function were kept and pad the first four levels with zeros. As many economic data are likely generated as aggregates of different scales, separating these scales and analysing them individually provides interesting insights and can improve the forecasting accuracy of the aggregate series. Renaud et al. [2003] divided the original time series to multiresolution crystals and forecasted these crystals separately. The forecasts are combined to achieve an aggregate forecast for the original time series. Genacy et al. [2001a] investigated the scaling properties of foreign exchange rates using wavelet methods. They decomposed the variance of the process and found that FX volatilities can be described by different scaling laws on different horizons. Using the maximal overlap discrete wavelet transform (MODWT), Genacy et al. [2001b] constructed a method for seasonality extraction from a time series, which is free of model selection parameters, translationally invariant, and associated with a zero-phase filter. Following the same ideas, Genacy et al. [2003] [2005] proposed a new approach for estimating the systematic risk of an asset and found that the estimation of CAPM could be flawed due to the multiscale nature of risk and return.

6.2.5 Interdependence between variables

Wavelets have been widely used to study interdependence of economic and financial time series. Genacy et al. [2001a] analysed the dependencies between foreign exchange markets and found an increase of correlation from intra-day scale towards the daily timescale stabilising for longer time scale. Fernandez [2005] studied the return spillovers in major stock markets on different time scales and concluded that G7 countries significantly affect global markets but that the reverse reaction is much weaker. Kim et al. [2005] [2006] have conducted many studies in finance using the wavelet variance, wavelet correlation and cross-correlation and found a positive relationship between stock returns and inflation on a scale of one month and 128 months, and a negative relationship between these scales. Studying the relationship between stock and futures markets with the MODWT based estimator of wavelet cross-correlation, In et al. [2006] found a feedback relationship between them on every scale, and correlation increasing with increasing time scale.

6.2.6 Introducing long memory processes

We have seen earlier that it was important to differentiate between stationary $I(0)$ and non-stationary $I(1)$ processes. However, there exist another type of processes, the fractionally integrated $I(d)$ processes, lying between the two sharp-edged alternatives of $I(0)$ and $I(1)$. Long-memory processes, corresponding to $d \in [0, 0.5]$, are processes with finite variance but autocovariance function decaying at a much slower rate than that of a stationary ARMA process. When

$d \in [0.5, 1]$, the variance becomes infinite, but the processes still return to their long-run equilibrium. A fractionally integrated process, $I(d)$, can be defined as

$$(1 - L)^d y(t) = \epsilon(t)$$

where $\epsilon(t)$ is white noise or follows an ARMA process. Since long-memory processes have a very dense covariance matrix, direct maximum likelihood estimation is not feasible for large data sets, and one generally uses a nonparametric approach, which regress the log values of the periodigram on the log Fourier frequencies to estimate d (see Geweke et al. [1983] (GPH)). Alternatively, McCoy et al. [1996] found a log-linear relationship between the variance of the wavelet coefficients and its scale, and developed a maximum likelihood estimator. Jensen [1999] developed an ordinary least square (OLS) estimator based on the observation that for a mean zero $I(d)$ process, $|d| < 0.5$, the wavelet coefficients, w_{jk} (for scale j and translation k), are asymptotically normally distributed with mean zero and variance $\sigma^2 2^{-2jd}$ as j goes to zero. That is, the wavelet transform of these kind of processes have a sparse covariance matrix which can be approximated at high precision with a diagonal matrix, such that the calculation of the likelihood function is of an order smaller than calculations with the exact MLE methods. Taking logs, we can estimate d using the linear relationship

$$\ln R(j) = \ln \sigma^2 - d \ln 2^{2j}$$

where $R(j)$ is the sample estimate of the covariance in each scale. The wavelet estimators have a higher small-sample bias than the GPH estimator, but they have a mean-squared error about six times lower.

6.3 Presenting wavelet-based forecasting methods

6.3.1 Forecasting with the a trous wavelet transform

Knowing the individual time series resulting from the decomposition in Equation (6.1.1), several approaches exist to estimate x_{t+k}, where k is a look-ahead period, from the observations $x_t, x_{t-1}, .., x_1$. For instance,

- if the residual vector S_p is sufficiently smooth, we can use a linear approximation of the data, or a carbon copy ($x_t \to x_{t+k}$) of it.
- we can make independent predictions to the resolution scales d_i and S_p and use the additive property of the reconstruction equation to fuse predictions in an additive manner.
- we can also test a number of short-memory and long-memory predictions at each resolution level, and retain the method performing best.

Note, the symmetric property of the filter function does not support the fact that time is a fundamentally asymmetric variable. In prediction studies, very careful attention must be given to the boundaries of the signal. Assuming a time series of size N, values at times $N, N-1, N-2, ...$ are of great importance. When handling boundary (or edge), any symmetric wavelet function is problematic, as we can not use wavelet coefficients estimated from unknown future data values. One way around is to hypothesise future data based on values in the nearest past. Further, for both symmetric and asymmetric functions, we have to use some variant of the transform to deal with the problem of edges. We can use the mirror or periodic border handling, or even the transformation of the border wavelets and scaling functions (see Cohen et al. [1992]). For instance, Aussem et al. [1998] chose the boundary condition

$$S(N + k) = S(N - k) \tag{6.3.3}$$

and described a novel approach for time-varying data. Two types of feature were considered

1. Decomposition-based approach: wavelet coefficients at a particular time point were taken as a feature vector.

2. Scale-based approach: modelling and prediction were run independently at each resolution level, and the results were combined.

They performed the feature selection with feature vector $x_t = \{d_1(t), d_2(t), .., d_p(t), S_p(t)\}$. Since they are using a wrap-around approach to defining the WT at the boundary region of the data, they considered data up to point $t = t_0$ and used x_{t_0} as a feature vector. The succession of feature vectors, x_{t_0}, corresponding to successive values of t_0, is not the same as a single WT of the input data. While these special coefficients better represent the true wavelet coefficients, they do not sum to zero at each level. However, they do retain the additive decomposition property of the reconstruction equation, and they do not use unknown future data. The scale-based approach used the dynamic recurrent neural network (DRNN) which is endowed with internal memory using additional information on the past time series. They found that the wavelet coefficients at higher frequency levels (lower scales) provided some benefit for estimating variation at less high frequency levels. For instance, to model and predict at scale 2, the target value is $d_2(t-15), d_2(t-14), .., d_2(t)$ combined with the input vector $d_1(t-15), d_1(t-14), .., d_1(t)$. The use of d_1 for prediction of d_2 is of benefit, since the more noisy and irregular the data, the more demanding the prediction task, and the more useful the neural network. On the final smooth trend curve $S_p(t)$, they found that the linear extrapolation $\hat{S}_p(t+5) = S_p(t) + \alpha(S_p(t) - S_p(t-1))$ with $\alpha = 5$ performed better than the NN solution. They considered three performance criteria to test the forecasting method, the normalised mean squared error (NMSE), the directional symmetry (DS), and the direction variation symmetry (DVS) (see details in Section (3.2.2.3)). Recombining all wavelet estimates, they obtained 0.72 for NMSE, 0.73% for the DS, and 0.6% for the DVS.

In conclusion, when forecasting, the features revealed by the individual wavelet coefficient series are meaningful when considered individually. However, when using wavelet coefficients at a given level to forecast the coefficient at the next level, we introduce positive correlation between the prediction residuals, and a single forecast error at a particular level can propagate, impacting the other predictors. Consequently, individual forecasts should be provided by different forecasting models to avoid output discrepancy correlation resulting from model misspecification. Further, the way of dealing with the edge is problematic in prediction applications as it adds artifacts in the most important part of the signal, namely, its right border values. It is interesting to note that years later, Murtagh et al. [2003] did not recommend the use of this algorithm.

6.3.2 The redundant Haar wavelet transform for time-varying data

Smoothing with a B_3 spline to construct the a trous wavelet transform, as described above, is not appropriate for a directed (time-varying) data stream since future data values can not be used in the calculation of the wavelet transform. We could use the Haar wavelet transform due to the asymmetry of the wavelet function, but it is a decimated one. Alternatively, Zheng et al. [1999] developed a non-decimated, or redundant, version of this transform corresponding to the a trous algorithm discussed above, but with a different pair of scaling and wavelet functions. The non-decimated Haar algorithm uses the simple filter $h = (\frac{1}{2}, \frac{1}{2})$, which is non-symmetric, with $l = -1, 0$ in the a trous algorithm. We can then derive the $(j+1)$ wavelet resolution level from the (j) level by convolving the latter with h, getting

$$S_{j+1}(t) = \frac{1}{2}\big(S_j(t-2^j) + S_j(t)\big)$$

and

$$d_{j+1}(t) = S_j(t) - S_{j+1}(t)$$

Hence, at any point, t, we never use information after t when computing the wavelet coefficients (see Percival et al. [2000]), obtaining a computationally straightforward solution to the problem of boundary conditions at time point t. Since at a given time t and scale $(j+1)$ we need two values from the previous scale (j), namely $S_j(t)$ and $S_j(t-2^j)$, the window length must be equal to 2^j for scale (j). Further, the smooth data $S_j(t)$ can be written as a moving average of the original signal as follow

$$S_j(t) = \frac{1}{2^j} \sum_{l=0}^{2^j-1} S_0(t-l)$$

Moreover, we can rewrite the wavelet details as

$$d_{j+1}(t) = \frac{1}{2} S_j(t) - S_{j+1}(t) + \frac{1}{2} S_j(t)$$

such that

$$-S_{j+1}(t) + \frac{1}{2} S_j(t) = -\frac{1}{2^{j+1}} \sum_{l=2^j}^{2^{j+1}-1} S_0(t-l)$$

This method has the following advantages

- The computational requirement is $O(N)$ per scale, and in practice the number of scales is set as a constant.
- Since we do not shift the signal, the wavelet coefficients at any scale j of the signal $(X_1, .., X_t)$ are strictly equal to the first t wavelet coefficients at scale j of the signal $(X_1, .., X_N)$ for $N > t$.

As a result, we get linearity in terms of the mapping of inputs defined by wavelet coefficients vis a vis the output target value. Note the following properties of the multiresolution transform

- all wavelet scales are of zero mean
- the smooth trend is generally much larger-valued than the max-min ranges of the wavelet coefficients

Since the reconstruction of the original signal is linear, given by the Equation (6.1.1), we can easily denoise the series by setting the detail coefficients to zero. For instance, setting $d_1(t) = 0$, we get $\sum_{j=2}^{p} d_j = S_1(t) - S_p(t)$ and the signal becomes

$$\tilde{x}(t) = S_p(t) + \sum_{j=2}^{p} d_j = S_1(t)$$

which is the smoothed data at scale $j = 1$. Similarly, setting $d_1(t) = d_2(t) = 0$, we recover the smoothed data $S_2(t)$ at scale $j = 2$. And so on, until $d_1(t) = d_2(t) = ... = d_p(t) = 0$ where we recover $S_p(t)$. Thus, setting the detail coefficients to zero one after the other produces the smoothed data. However, this approach does not allow for a denoised seried in between the smoothed data. One way forward is to take advantage of the linear property of the reconstruction of the original signal by linearly combining the smoothed data $S_j(t)$. That is, setting

$$\tilde{d}_j(t) = (1 - \alpha_j) d_j(t) \, , \alpha_j \in [0, 1] \, , j = 1, ..., p$$

we get

$$\begin{aligned} \tilde{x}(t) &= S_p(t) + \sum_{j=1}^{p} \tilde{d}_j(t) \\ &= (1-\alpha_1) S_0(t) + (\alpha_1 - \alpha_2) S_1(t) + ... + (\alpha_{p-1} - \alpha_p) S_{p-1}(t) + \alpha_p S_p(t) \end{aligned}$$

where α_j, $j = 1, ..., p$, are the weights of the combination. Two special cases are

$$\alpha_1 = \alpha_2 = ... = \alpha_p = 0 \text{ we recover } S_0(t)$$
$$\alpha_1 = \alpha_2 = ... = \alpha_p = 1 \text{ we recover } S_p(t)$$

For $\alpha_1 = 1$ and $\alpha_2 = ... = \alpha_p = 0$ we get the smoothed data $S_1(t)$, for $\alpha_1 = \alpha_2 = 1$ and $\alpha_3 = ... = \alpha_p = 0$ we get the smoothed data $S_2(t)$, etc. For any other values of α_j we get a comination of smoothed data.

6.3.3 The multiresolution autoregressive model

In order to capture short-range and long-range dependencies of time series, Renaud et al. [2002] proposed a multiresolution autoregressive (MAR) model to forecast time-series. They considered the non-decimated Haar a trous wavelet transform described in Section (6.1.2.2), and used a linear prediction based on some coefficients of the decomposition of the past values. We saw in the previous section that the window length in the Haar WT must be equal to 2^j for scale (j), introducing redundancy. Hence, when selecting the number of coefficients at different scales, we should exclude these redundant points. After some investigation, Renaud et al. found that the wavelet and scaling function coefficients that should be used for the prediction at time $N+1$ should have the form

$$d_{j,N-2^j(k-1)} \text{ and } S_{J,N-2^J(k-1)}$$

for positive value of k. For each N, this subgroup of coefficients is part of an orthogonal transform. We now want to allow for adaptivity in the numbers of wavelet coefficients selected from different resolution scales and used in the prediction.

6.3.3.1 Linear model

Stationary process We let the window size at scale (j) be denoted A_j. Assuming a stationary signal $X = (X_1, .., X_N)$, the one-step forward prediction of an $AR(p)$ process is $\hat{X}_{N+1} = \sum_{k=1}^{p} \hat{\phi}_k X_{N-(k-1)}$ (see details in Section (5.3.3)). In order to use the wavelet decomposition in Equation (6.1.1), Renaud et al. modified the prediction to the AR multiscale prediction as

$$\hat{X}_{N+1} = \sum_{j=1}^{J} \sum_{k=1}^{A_j} \hat{a}_{j,k} d_{j,N-2^j(k-1)} + \sum_{k=1}^{A_{J+1}} \hat{a}_{J+1,k} S_{J,N-2^J(k-1)}$$

where $\mathcal{D} = d_1, .., d_J, S_J$ represents the Haar a trous wavelet transform of X. Note, we have one $AR(p)$ process per scale $j = 1, .., J$ with $\{\hat{a}_{j,k}\}_{1 \leq j \leq J, 1 \leq k \leq A_j}$ parameters for the details and $\{\hat{a}_{J+1,k}\}_{1 \leq k \leq A_{J+1}}$ parameters for the residual. Put another way, if on each scale the lagged coefficients follow an $AR(A_j)$ process, the addition of the predictions on each level would lead to the same prediction formula than the above one. That is, the MAR prediction model is linear. In the special case where $A_j = 1$ for all resolution levels j, the prediction simplifies to

$$\hat{X}_{N+1} = \sum_{j=1}^{J} \hat{a}_j d_{j,N} + \hat{a}_{J+1} S_{J,N}$$

In this model, we need to estimate the $Q = \sum_{j=1}^{J+1} A_j$ unkown parameters which we grouped in the vector α, so that we can solve the equation $A^{'} A\alpha = A^{'} S$ where

$$\begin{aligned}
A^{'} &= (L_{N-1}, .., L_{N-M}) \\
L^{'} &= (d_{1,t}, .., d_{1,t-2A_1}, .., d_{2,t}, .., d_{2,t-2^2A_2}, ..., d_{J,t}, .., d_{J,t-2^JA_J}, S_{J,t}, ..., S_{J,t-2^JA_{J+1}}) \\
\alpha^{'} &= (a_{1,1}, .., a_{1,A_1}, a_{2,1}, .., a_{2,A_2}, .., a_{J,1}, .., a_{J,A_j}, .., a_{J+1,1}, .., a_{J+1,A_{J+1}}) \\
S^{'} &= (X_N, .., X_{t+1}, .., X_{N-M+1})
\end{aligned}$$

where A is a $Q \times M$ matrix (M rows L_t, each with Q elements), α and S are respectively Q and M-size vectors, and Q is larger than M.

Non-stationary process When a trend is present in the time-series, we use the fact that the multiscale decomposition automatically separates the trend from the signal. As a result, we can predict both the trend and the stochastic part within the multiscale decomposition. In general, the trend affects the low frequency components, while the high frequencies are purely stochastic. Hence, we separate the signal X into low (L) and high (H) frequencies, getting

$$L = S_J \text{ and } H = X - L = \sum_{j=1}^{J} d_j$$
$$X_{N+1} = L_{N+1} + H_{N+1}$$

In that setting, the signal H has zero-mean, and the MAR model gives

$$\hat{H}_{N+1} = \sum_{j=1}^{J} \sum_{k=1}^{A_j} a_{j,k} d_{j,N-2^j(k-1)}$$

and the estimation of the Q unkown parameters is as before, except that the coefficients S are not used in L_i and that S is based on H_{t+1}. Since L is very smooth, a polynomial fitting can be used for prediction, or one can use an AR process as for H. Since the frequencies are non-overlapping in each scale, one can select the parameters A_j independently on each scale with AIC, AICC, or BIC methods. Hence, the general method consists in fitting an AR model to each scale of the multiresolution transform. In the case where the true process is AR, thisforecasting procedure will converge to the optimal procedure, since it is asymptotically equivalent to the best forecast.

6.3.3.2 Non-linear model

Note, Murtagh et al. [2003] generalised the MAR formula to the nonlinear case leading to a learning algorithm. Assuming a standard multilayer perceptron with a linear transfer function at the output node, and L hidden layer neurons, they obtained the following

$$\hat{X}_{N+1} = \sum_{l=1}^{L} \hat{a}_l g\Big(\sum_{j=1}^{J} \sum_{k=1}^{A_j} \hat{a}_{j,k} d_{j,N-2^j(k-1)} + \sum_{k=1}^{A_{J+1}} \hat{a}_{J+1,k} S_{J,N-2^J(k-1)} \Big)$$

where the sigmoidal function $g(\bullet)$ is used from the feedforward multilayer perceptron (see details in Section (13.6.1.2)).

6.3.4 The neuro-wavelet hybrid model

Zhang et al. [2001] developed a neuro-wavelet hybrid system incorporating multiscale wavelet analysis into a set of neural networks for a multistage time series prediction. Their approach consists of some three stage prediction scheme. Considering a shift invariant WT, they used the autocorrelation shell representation (ASR) introduced by Beylkin et al. [1992] which we described in Appendix (G.4.3). Then, they performed the prediction of each scale of the wavelet coefficients with the help of a separate feedforward neural network. That is, the prediction results for the wavelet coefficients can either be directly combined from the linear additive reconstruction property of ASR, or, can be combined from another neural network (NN). The main goal of the latter being to adaptively choose the weight of each scale in the final prediction. For the prediction of different scale wavelet coefficients, they applied the Bayesian method of automatic relevance determination (ARD) to learn the different significance of a specific length of past window and wavelet scale.

The additive form of reconstruction of the ASR allows one to combine the predictions in a simple additive manner. In order to deal with the boundary condition of using the most recent data to make predictions, Zhang et al. followed the approach proposed by Aussem et al. [1998] described in Section (6.3.1), and used a time-based a trous filters algorithm on the signal $x_1, x_2, .., x_N$ where N is the present time-point. The steps are as follow:

1. For index k sufficiently large, carry out the a trous transform in Equation (G.4.32) on the signal using a mirror extension of the signal when the filter extends beyond k (see Equation (6.3.3)).
2. Retain the coefficient values (details) as well as the residual values for the kth time point only, that is, $D_k^1, D_k^2, .., D_k^p, S_k^p$. The summation of these values gives x_k.
3. If k is less than N, set k to $k + 1$ and return to step 1).

This process produces an additive decomposition of the signal $x_k, x_{k+1}, ..., x_N$, which is similar to the a trous wavelet transform decomposition on $x_1, x_2, .., x_N$. However, as discussed in Section (6.3.1), the boundary condition in Equation (6.3.3) is not appropriate when forecasting financial data, as we can not use future data in the calculation of the wavelet transform.

In general, the appropriate size for the time-window size of inputs of a regression problem is a difficult choice as there are many possible input variables, some of which may be irrelevant. This problem also applies to time series forecasting with neural networks (NN). Zhang et al. used the ARD method for choosing the length of past windows to train the NN.

6.4 Some wavelets applications to finance

6.4.1 Deriving strategies from wavelet analysis

Given the properties of time series in the frequency and scale domain, we can apply Fourier and wavelet analysis to perform pattern recognition and denoising around time series. As explained above, we want to statistically study wavelets coefficients in order to compare them. We choose to concentrate the information (energy) of market signal in a small number of coefficients at a certain scale. We therefore need to use a mother wavelet minimising the number of levels and coefficients containing significant information.

We can apply wavelet analysis to forecast time series. By decomposing time series into different scales and adding up the squared coefficients within each level, we can measure the energy constant of each scale (power spectral density in the Fourier analysis). Using the properties of the multiscale analysis, we then decompose the time series into two series which we then forecast using an ARIMA model. The forecast of the original series is obtained by aggregating the forecast of the individual series.

Combining wavelet analysis with other tools, such as Neural Network, has now gained wide acceptance in major scientific fields. To apply wavelet network, we first decompose the time series into different scales, and each scale is then used to train a recurrent neural network that will provide the forecast. Aggregating the results we recover the original signal. In both cases, wavelets are used to extract periodic information within individual scales which is used by other techniques.

6.4.2 Literature review

Wavelet analysis can be used on financial time series which are typically highly nonstationary, exhibit high complexity and involve both (pseudo) random processes and intermittent deterministic processes. A good overview of the application of wavelets in economics and finance is given by Ramsey [1999]. Davidson et al. [1998] used the orthogonal dyadic Haar transform to perform semi-nonparametric regression analysis of commodity price behaviour. Ramsey

et al. [1995] searched for evidence of self-similarity in the US stock market price index. Investigating the power law scaling relationship between the wavelet coefficients and scale, they found some evidence of quasi-periodicity in the occurrence of some large amplitude shocks to the system, concluding that there may be a modest amount of predictability in the data. Further, Ramsey et al. [1998] highlighted the importance of timescale decomposition in analysing economic relationships. Wavelet-based methods to remove hidden cycles from within financial time series have been developed by Arino et al. [1995] where they first decompose the signal into its wavelet coefficients and then compute the energy associated with each scale. Defining dominant scales as those with the highest energies, new coefficient sets are produced related to each of the dominant scales by either one of two methods developed by the authors. For the signal containing two dominant scales, two new complete sets of wavelet coefficients are computed which are used to reconstruct two separate signals, corresponding to each dominant scale. Arneodo et al. [1998] found evidence for a cascade mechanism in market dynamics attributed to the heterogeneity of traders and their different time horizons causing an information cascade from long to short timescales, the lag between stock market fluctuations and long-run movements in dividends, and the effect of the release (monthly, quaterly) of major economic indicators which cascades to fine timescales. Aussem et al. [1998] used wavelet transformed financial data as the input to a neural network which was trained to provide five-days ahead forecasts for the $S\&P500$ closing prices. They examined each wavelet series individually to provide separate forecasts for each timescale and recombined these forecasts to form an overall forecast. Ramsey et al. [1996] decomposed the $S\&P500$ index using matching pursuits and found data are characterised by periods of quiet interspersed with intense activity over short periods of time. They found that fewer coefficients were required to specify the data than for a purely random signal, signifying some form of deterministic structure to the signal. Ramsey et al. [1997] also applied matching pursuits to foreign exchange rate data sets, the Deutschmark-US dollar, yen-US dollar and yen-Deutschmark, and found underlying traits of the signal. Even though most of the energy of the system occored in localised burst of activity, they could not predict their occurence and could not improve forecasting. Combining wavelet transforms, genetic algorithms and artificial neural networks, Shin et al. [2000] forecasted daily Korean-US dollar returns one-day ahead of time and showed that the genetic-based wavelet thresholder outperformed cross-validation, best level and best bias. Gencay et al. [2001b] filtered out intraday periodicities in exchange rate time series using the maximal overlap discrete wavelet transform.

Part III

Quantitative trading in inefficient markets

Chapter 7

Introduction to quantitative strategies

The market inefficiency and the asymmetrical inefficiency of the long-only constraint in portfolio construction led some authors to revise the modern portfolio theory developed by Markowitz, Sharpe, Tobin and others. The notion of active, or benchmark-relative, performance and risk was introduced by Grinold [1989] [1994] and the source of excess risk-adjusted return for an investment portfolio was examined.

7.1 Presenting hedge funds

7.1.1 Classifying hedge funds

Due to the various ways of selecting and risk managing a portfolio (see details in Section (2.1)), there is a large number of Hedge Funds in the financial industry classified by the type of strategies used to manage their portfolios. Considering single strategies, we are going to list a few of them.

- Macro: Macro strategies concentrate on forecasting how global macroeconomic and political events affect the valuations of financial instruments. The strategy has a broad investment mandate. With the ability to hold positions in practically any market with any instrument, profits are made by correctly anticipating price movements in global markets.
- Equity Hedge: also known as long/short equity, combine core long holdings of equities with short sales of stock or stock index options and may be anywhere from net long to net short depending on market conditions. The source of return is similar to that of traditional stock picking on the upside, but the use of short selling and hedging attempts to outperform the market on the downside.
- Equity Market Neutral: Using complex valuation models, equity market neutral fund managers strive to identify under/overvalued securities. Accordingly, they are long in undervalued positions while selling overvalued securities short. In contrast to equity hedge, equity neutral has a total net exposure of zero. The strategy intends to neutralise the effect that a systematic change will have on values of the stock market as a whole.
- Relative Value: Generally, a relative value strategy makes Spread Trades in similar or related securities when their values, which are mathematically or historically interrelated, are temporarily distorted. Profits are derived when the skewed relationship between the securities returns to normal.
- Statistical Arbitrage: As a trading strategy, statistical arbitrage is a heavily quantitative and computational approach to equity trading involving data mining and statistical methods, as well as automated trading systems. StatArb evolved out of the simpler pairs trade strategy, but it considers a portfolio of a hundred or more stocks, some long and some short, that are carefully matched by sector and region to eliminate exposure to beta and other risk factors.

7.1.2 Some facts about leverage

7.1.2.1 Defining leverage

There are numerous ways leverage is defined in the investment industry, and there is no consensus on exactly how to measure it. Leverage can be defined as the creation of exposure greater in magnitude than the initial cash amount posted to an investment, where leverage is created through borrowing, investing the proceeds from short sales, or through the use of derivatives. Thus, leverage may be broadly defined as any means of increasing expected return or value without increasing out-of-pocket investment. There are three primary types of leverage

1. Financial Leverage: This is created through borrowing leverage and/or notional leverage, both of which allow investors to gain cash-equivalent risk exposures greater than those that could be funded only by investing the capital in cash instruments.
2. Construction Leverage: This is created by combining securities in a portfolio in a certain manner. The way one constructs a portfolio will have a significant effect on overall portfolio risk, depending on the amount and type of diversification in the portfolio, and the type of hedging applied (e.g., offsetting some or all of the long positions with short positions).
3. Instrument Leverage: This reflects the intrinsic risk of the specific securities selected, as different instruments have different levels of internal leverage.

Leverage allows hedge funds to magnify their exposures and thus magnify their risks and returns. However, a hedge fund's use of leverage must consider margin and collateral requirements at the transaction level, and any credit limits imposed by trading counterparties such as prime brokers. Therefore, hedge funds are often limited in their use of leverage by the willingness of creditors and counterparties to provide the leverage.

7.1.2.2 Different measures of leverage

Leverage may be quoted as a ratio of assets to capital or equity (e.g., 4 to 1), as a percentage (e.g., 400%), or as an incremental percentage (e.g., 300%). The Gross Market Exposure is defined as

$$\text{Gross Market Exposure } = \frac{\text{Long } + \text{ Short}}{\text{Capital or Equity}} 100\%$$

For example, Hedge Fund A has $\$1$ million of capital, borrows $\$250,000$ and invests the full $\$1,250,000$ in a portfolio of stocks (i.e., the Fund is long $\$1.25$ million). At the same time, Hedge Fund A sells short $\$750,000$ of stocks. Then

$$\text{Gross Market Exposure } = \frac{1.25 + 0.75}{1} 100\% = \frac{2}{1} 100\% = 200\%$$

Many investors do not consider the 200% Gross Market Exposure in the above example to be leverage per se. For example, assume Hedge Fund A has capital of $\$1$ million and is $\$1$ million long and $\$1$ million short. This results in Gross Market Exposure of 200%, but Net Market Exposure of zero, which is the typical exposure of an equity market-neutral fund. That is, the Net Market Exposure is defined as

$$\text{Net Market Exposure } = \frac{\text{Long } - \text{ Short}}{\text{Capital or Equity}} 100\%$$

Given the previous example, the Net Market Exposure is

$$\text{Net Market Exposure } = \frac{1.25 - 0.75}{1} 100\% = \frac{1/2}{1} 100\% = 50\%$$

One should ask hedge fund managers the following types of questions regarding leverage:

- Does the manager have prescribed limits for net market exposure and for gross market exposure?
- What drives the decision to go long or short, and to use more or less leverage?
- What leverage and net market exposure was used to generate the manager's track record?
- What is the attribution of return between security selection, market timing and the use of leverage?

7.1.2.3 Leverage and risk

We must distinguish between the concepts of leverage and risk, as there is a common misconception that a levered asset is always riskier than an unlevered asset. In general, risk is defined by the portfolio's stock market risk (beta), and when investors are confronted to several equity portfolios they have to identify the one with the greatest risk. Even though equity portfolios may have the same market risk, as each portfolio has the same aggregate beta, the key point is that the same risk level is achieved through different types of leverage. The relationship between risk and leverage is complex, in particular when comparing different investments, a higher degree of leverage does not necessarily imply a higher degree of risk. Leverage is the link between the underlying or inherent risk of an asset and the actual risk of the investor's exposure to that asset. Thus, the investor's actual risk has two components:

1. The market risk (beta) of the asset being purchased
2. The leverage that is applied to the investment

For example, which is more risky: a fund with low net market exposure and borrowing leverage of 1.5 times capital, or a fund with 100% market exposure and a beta of 1.5 but no borrowing leverage?

For a given capital base, leverage allows investors to build up a larger investment position and thus a higher exposure to specific risks. Buying riskier assets or increasing the leverage ratio applied to a given set of assets increases the risk of the overall investment, and hence the capital base. Therefore, if a portfolio has very low market risk then higher leverage may be more acceptable for these strategies than for strategies that have greater market exposure, such as long-short equity or global macro. In fact, a levered portfolio of low-risk assets may well carry less risk than an unlevered portfolio of high-risk assets. Therefore, investors should not concern themselves with leverage per se, but rather focus on the risk/return relationship that is associated with a particular portfolio construction. In this way, investors can determine the optimal allocation to a specific strategy in a diversified portfolio.

7.2 Different types of strategies

7.2.1 Long-short portfolio

7.2.1.1 The problem with long-only portfolio

Some form of risk constraints are generally placed on fund managers by investors or fund administrators, such as size-neutrality, sector neutrality, value-growth neutrality, maximum total number of positions and long-only constraints. Clarke et al. [2002] found that the long-only constraint is the most significant restriction placed on portfolio managers. While most investors focus on the management of long portfolios and the selection of winning securities, the identification of winning securities ignores by definition a whole class of losing securities. As explained in Section (2.1.1), excess returns come from active security weights, that is, portfolio weights differing from benchmark weights. An active long-only portfolio holds securities expected to perform above average at higher-than-benchmark weights and those expected to perform below average at lower-than-benchmark weights. Without short-selling, it can not underweight many securities by enough to achieve significant negative active weights. Hence, restricting short sales prevents managers from fully implementing their complete information set when constructing their portfolios. As explained by Jacobs et al. [1999], the ability to sell short frees the investor from taking advantage of the full array

of securities and the full complement of investment insights by holding expected winners long and selling expected losers short. Active fund managers expresses their investment view on the assets in their investment universe by holding an over-, neutral or under-weight position $(w_{a,i} > 0, w_{a,i} = 0, w_{a,i} < 0)$ in these assets relative to their assigned benchmark

$$w_{a,i} = w_{f,i} - w_{b,i}$$

where $w_{a,i}$ is the active weight in asset i, $w_{f,i}$ is the weight in asset i,and $w_{b,i}$ is the weight of the benchmark in asset i. Since the conventional long-only fund manager can only expand a negative position to the point of excluding the asset from the fund $(w_{f,i} \geq 0)$, then the most negative active weight possible in any particular asset, in a long-only fund, is the negative of the asset's benchmark weight $(w_{a,i} \geq -w_{b,i})$. It results in a greater scope for expressing positive investment views in each asset than negative views, leading to the asymmetry in the long-only active manager's opportunity set. On the other hand, the unconstrained investor benefits from a symmetrical investment opportunity set with respect to implementing negative active investment weights. The extent to which the fund manager can expand a positive active weight is limited only by a particular mandate restrictions and the ability to finance the total positive active positions in the portfolio with sufficient negative positions in other assets.

7.2.1.2 The benefits of long-short portfolio

The benefits of long-short portfolio are to a large extent dependent on proper portfolio construction, and only an integrated portfolio can maximise the value of investors' insights. Much of the incremental cost associated with a given long-short portfolio reflects the strategy's degree of leverage. Although most existing long-short portfolios are constructed to be neutral to systematic risk, neutrality is neither necessary nor optimal. Further, long-short portfolio do not constitute a separate asset class and can be constructed to include a desired exposure to the return of any existing asset class.

Long-short portfolio should not be considered as a two portfolio strategy but as a one portfolio strategy in which the long and short positions are determined jointly within an optimisation that takes into account the expected returns of the individual securities, the standard deviations of those returns, and the correlations between them, as well as the investor's tolerance for risk (see Jacobs et al. [1995] and [1998]). Within integrated optimisation, there is no need to converge to securities' benchmark weights in order to control risk. Rather, offsetting long and short positions can be used to control portfolio risk. For example, if an investor has some strong insight about oil stocks, some of which are expected to do very well and some other very poorly, he does not need to restrict weights to index-like weights and can allocate much of the portfolio to oil stocks. The offsetting long and short positions control the portfolio's exposure to the oil factor. On the other hand, if he has no insights into oil stock behaviour, the long-short investor can totally exclude oil stocks from the portfolio. The risk is not increased because in that setting it is independent of any security's benchmark weight. The absence of restrictions imposed by securities's benchmark weights enhances the long-short investor's ability to implement investment insights.

An integrated optimisation that considers both long and short positions simultaneously, not only frees the investor from the non-negativity constraint imposed on long-only portfolios, but also frees the long-short portfolio from the restrictions imposed by securities's benchmark weights. To see this we follow Jacobs et al. [1999] and consider an obvious (suboptimal) way of constructing a long-short portfolio. To do so, we combine a long-only portfolio with a short-only portfolio resulting in a long-plus-short portfolio and not a true long-short portfolio. The long side of this portfolio being identical to a long-only portfolio, it offers no benefits in terms of incremental return or reduced risk. Further, the short side is statistically equivalent to the long side, hence to the long-only portfolio. In effect, assuming symmetry of inefficiencies across attractive and unattractive stocks, and, assuming identical and separate portfolio construction for the long and short sides, we get

$$\alpha_L = \alpha_S = \alpha_{LO}$$
$$\sigma_{e,L} = \sigma_{e,S} = \sigma_{e,LO}$$

where α_l for $l = L, S, LO$ is the alpha of the long, short, and long-only portfolio, and $\sigma_{e,l}$ for $l = L, S, LO$ is the residual risk of the respective portfolios. The excess return, or alpha, of the long side of the long-plus-short portfolio will equal the alpha of the short side, which will equal the alpha of the long-only portfolio. This is also true of the residual risk σ_e. It means that all the three portfolios are constructed relative to a benchmark index. Each portfolio is active in pursuing excess return relative to the underlying index only insofar as it holds securities in weights that depart from their index weights. This portfolio construction is index-constrained. Assuming that the beta of the short side equals the beta of the long side, the ratio of the performance of the long-plus-short portfolio to that of the long-only portfolio can be expressed as

$$\frac{IR_{L+S}}{IR_{LO}} = \sqrt{\frac{2}{1+\rho_{L+S}}}$$

where the information ratio IR is a measure of risk-adjusted outperformance. It is the ratio of excess return over the benchmark divided by the residual risk (tracking error) [1]

$$IR = \frac{\alpha}{\sigma_e} \tag{7.2.1}$$

and ρ_{L+S} is the correlation between the alphas of the long and short sides of the long-plus-short portfolio. Hence, the advantage of a long-plus-short portfolio is curtailed by the need to control risk by holding or shorting securities in index-like weights. Benefits only apply if there is a less-than-one correlation between the alphas of its long and short sides. In that case, the long-plus-short portfolio will enjoy greater diversification and reduced risk relative to the long-only portfolio.

Advocates of long-short portfolios also point to the diversification benefits provided by the short side. According to them, a long-short strategy includes a long and a short portfolio; if the two portfolios are uncorrelated, the combined strategy would have a higher information ratio than the two separate portfolios as a result of diversification. Jacobs et al. [1995] addressed the diversification argument by observing that long and short alphas are not separately measurable in an integrated long-short optimisation framework. They suggested that the correlation between the separate long and short portfolios is not relevant. More recently, the centre stage of the long-short debate has focused on whether efficiency gains result from relaxing the long-only constraint. Grinold et al. [2000] showed that information ratios decline when one moves from a long-short to a long-only strategy.

7.2.2 Equity market neutral

An investment strategy or portfolio is considered market-neutral if it seeks to entirely avoid some form of market risk, typically by hedging. A portfolio is truly market-neutral if it exhibits zero correlation with the unwanted source of risk, and it is seldom possible in practice. Equity market-neutral is a hedge fund strategy that seeks to exploit investment opportunities unique to some specific group of stocks while maintaining a neutral exposure to broad groups of stocks defined, for example, by sector, industry, market capitalisation, country, or region. The strategy holds long-short equity positions, with long positions hedged with short positions in the same and related sectors, so that the equity market-neutral investor should be little affected by sector-wide events. For example, a hedge fund manager will go long in the 10 biotech stocks that should outperform and short the 10 biotech stocks that will underperform. Therefore, what the actual market does will not matter (much) because the gains and losses will offset each other. Equivalently, the process of stock picking can be realised with complex valuation models. This places, in essence, a bet that the long positions will outperform their sectors (or the short positions will underperform) regardless of the strength of the sectors.

As an example, a delta neutral strategy describes a portfolio of related financial securities, in which the portfolio value remains unchanged due to small changes in the value of the underlying security. The term delta hedging is the

[1] The tracking error refers to the standard deviation of portfolio returns against the benchmark return. Hence, risk refers to the deviation of the portfolio returns from the benchmark returns.

process of setting or keeping the delta of a portfolio as close to zero as possible. It may be accomplished by buying or selling an amount of the underlier that corresponds to the delta of the portfolio. By adjusting the amount bought or sold on new positions, the portfolio delta can be made to sum to zero, and the portfolio is then delta neutral (see Wilmott et al. [2005]). Another example is the pairs trade or pair trading which corresponds to a market neutral trading strategy enabling traders to profit from virtually any market conditions: uptrend, downtrend, or sideways movement. This strategy is categorised as a statistical arbitrage and convergence trading strategy. The pair trading was pioneered by Gerry Bamberger and later led by Nunzio Tartaglia's quantitative group at Morgan Stanley in the early to mid 1980s (see Gatev et al. [2006], Bookstaber [2007]). The idea was to challenge the Efficient Market Hypothesis and exploit the discrepancies in the stock prices to generate abnormal profits. The strategy monitors performance of two historically correlated securities. When the correlation between the two securities temporarily weakens, that is, one stock moves up while the other moves down, the pairs trade would be to short the outperforming stock and to long the underperforming one, betting that the spread between the two would eventually converge.

There are many ways in which to invest in market neutral strategies, all of which seek to take systematic risk out of the investment equation. Among the most common market neutral approaches is long-short equity. Long-short equity investing has several benefits. The strategy is uncorrelated to other asset classes. The alpha generated by long-short managers is uncorrelated to the alpha generated by index equity managers. Moreover, the alpha generated by long-short managers has low correlation to one another providing an excellent diversifying strategy. Long-short equity also provides flexibility in asset allocation and rebalancing due to the portability of the alpha generated by long-short equity and, moreover, market neutral in general. Over the longer term, long-short equity investing should provide attractive risk adjusted returns as well as greater diversification and flexibility within investment programs.

A portfolio which appears to be market-neutral may exhibit unexpected correlations as market conditions change leading to basis risk. Equity market-neutral managers recognise that the markets are dynamic and take advantage of sophisticated mathematical techniques to explore new opportunities and improve their methodology. The fact that there are many different investment universes globally makes this strategy less susceptible to alpha decay. The abundance of data lends itself well to rigorous back-testing and the development of new algorithms.

7.2.3 Pairs trading

We saw in Section (7.2.2) that Equity Market Neutral is not just a single trading strategy, but it is an umbrella term used for a broad range of quantitative trading strategies such as pairs trading. Pairs trading is one of Wall Street's quantitative methods of speculation which dates back to the mid-1980s (see Vidyamurthy [2004]). Market neutral strategies are generally known for attractive investment properties, such as low exposure to the equity markets and relatively low volatility. The industry practice for market neutral hedge funds is to use a daily sampling frequency and standard cointegration techniques [2] to find matching pairs (see Gatev et al. [2006]). The general description of the technique is that a pair of shares is formed, where the investor is long one share and short another share. The rational is that there is a long-term equilibrium (spread) between the share prices, and thus the share prices fluctuate around that equilibrium level (the spread has a constant mean). The investor evaluates the current position of the spread based on its historical fluctuations and when the current spread deviates from its historical mean by a pre-determined significant amount (measured in standard deviations), the spread is subsequently altered and the legs are adjusted accordingly. Studying the effectiveness of this type of strategy, Gatev et al. [2006] conducted empirical tests on pair trading using common stocks and found that the strategy was profitable even after taking the transaction costs into account. Jurek et al. [2007] improved performance by deriving a mean reversion strategy. Investigating the usefulness of pair trading applied to the energy futures market, Kanamura et al. [2008] obtained high total profits due to strong mean reversion and high volatility in the energy markets.

[2] Cointegration is a quantitative technique based on finding long-term relations between asset prices introduced in a seminal paper by Engle and Granger [1987]. Another approach was developed by Johansen [1988], which can be applied to more than two assets at the same time. The result is a set of cointegrating vectors that can be found in the system. If one only deals with pairs of shares, it is preferable to use the simpler Engle and Granger [1987] methodology.

In practice, the investor bets on the reversion of the current spread to its historical mean by shorting/going long an appropriate amount of each share in pair. That amount is expressed by the variable beta, which tells the investor the number of the shares X he has to short/go long, for each 1 share Y. There are various ways of calculating beta, it can either be fixed, or it can be time-varying. In the latter, one can use rolling ordinary least squares (OLS) regression, double exponential smoothing prediction (DESP) model and the Kalman filter. As an example, Dunis and Shannon [2005] use time adaptive betas with the Kalman filter methodology (see Hamilton [1994] or Harvey [1989] for a detailed description of the Kalman filter implementation). It is a forward looking methodology, as it tries to predict the future position of the parameters as opposed to using a rolling OLS regression (see Bentz [2003]). Later, Dunis et al. [2010] applied a long-short strategy to compare the profit potential of shares sampled at 6 different frequencies, namely 5-minute, 10-minute, 20-minute, 30-minute, 60-minute and daily sampling intervals. They considered an approach enhancing the performance of the basic trading strategy by selecting the pairs for trading based on the best in-sample information ratios and the highest in-sample t-stat of the Augmented Dickey-Fuller (ADF) unit root test of the residuals of the cointegrating regression sampled a daily frequency. As described by Aldridge [2009] one advantage of using the high-frequency data is higher potentially achievable information ratio compared to the use of daily closing prices.

Assuming the pairs belong to the same industry, we follow the description given by Dunis et al. [2010] and calculate the spread between two shares as

$$Z_t = P_t^Y - \beta_t P_t^X$$

where Z_t is the value of the spread at time t, P_t^Y is the price of share Y at time t, P_t^X is the price of share X at time t, and β_t is the adaptive coefficient beta at time t. In general, the spread is normalised by subtracting its mean and dividing by its standard deviation. The mean and the standard deviation are calculated from the in-sample period and are then used to normalise the spread both in the in- and out-of-sample periods. Dunis et al. sell (buy) the spread when it is 2 standard deviations above (below) its mean value and the position is liquidated when the spread is closer than 0.5 standard deviation to its mean. They chose the investment to be money-neutral, so that the amounts of euros to be invested on the long and short side of the trade is the same. They did not assume rebalancing once they entered into the position. Therefore, after an initial entry into the position with equal amounts of euros on both sides of the trade, even when due to price movements both positions stop being money-neutral, they did not rebalance the position. Only two types of transactions were allowed, entry into a new position, and total liquidation of the position they were in previously.

Dunis et al. [2010] explained the different indicators calculated in the in-sample period, trying to find a connecting link with the out-of-sample information ratio and as a consequence proposed a methodology for evaluating the suitability of a given pair for arbitrage trading. All the indicators are calculated in the in-sample period. The objective being to find the indicators with high predictive power of the profitability of the pair in the out-of-sample period. According to Do et al. [2006], the success of pairs trading depends heavily on the modelling and forecasting of the spread time series. For instance, the Ornstein-Uhlenbeck (OU) equation can be used to calculate the speed and strength of mean reversion

$$dZ_t = k\big(\mu - Z_t\big)dt + \sigma dW_t$$

where μ is the long-term mean of the spread, Z_t is the value of the spread at particular point in time, k is the strength of mean reversion, and σ is the standard deviation. The parameters of the process are estimated on the in-sample spread. This SDE is just the supplementary equation from which we calculate the half-life of mean reversion of the pairs. The half-life of mean reversion in number of periods can be calculated as

$$k_{\frac{1}{2}} = -\frac{\ln 2}{k}$$

Intuitively speaking, it is half the average time the pair usually takes to revert back to its mean. Thus, pairs with low half-life should be preferred to high half-lives by traders. The information ratio (IR) gives us an idea of the quality

of the strategy. An annualised information ratio of 2 means that the strategy is profitable almost every month, while strategies with an information ratio around 3 are profitable almost every day (see Chan [2009]). In the case of intraday trading, the annualised information ratio is

$$M_{IR} = \frac{R}{\sigma}\sqrt{h_d \times 252}$$

where h_d is the number of hours traded per day (for a day $h_d \neq 24$). However, it overestimates the true information ratio if returns are autocorrelated (see Alexander [2008]).

Pairs trading is not a risk-free strategy as the difficulty comes when prices of the two securities begin to drift apart, that is, the spread begins to trend instead of reverting to the original mean. Dealing with such adverse situations requires strict risk management rules, which have the trader exit an unprofitable trade as soon as the original setup, a bet for reversion to the mean, has been invalidated. This can be achieved by forecasting the spread and exiting at forecast error bounds. Further, the market-neutral strategies assume that the CAPM model is valid and that beta is a correct estimate of systematic risk. If this is not the case, the hedge may not properly protect us in the event of a shift in the markets. In addition, measures of market risk, such as beta, are historical and could vary from their past behaviour and become be very different in the future. Hence, in a mean reversion strategy where the mean is assumed to remain constant, then a change of mean is referred to as drift.

7.2.4 Statistical arbitrage

Statistical arbitrage (abbreviated as Stat Arb) refers to a particular category of hedge funds based on highly technical short-term mean-reversion strategies involving large numbers of securities (hundreds to thousands, depending on the amount of risk capital), very short holding periods (measured in days to seconds), and substantial computational, trading, and information technology (IT) infrastructure. As a trading strategy, statistical arbitrage is a heavily quantitative and computational approach to equity trading involving data mining and sophisticated statistical methods and mathematical models, as well as automated trading systems to generate a higher than average profit for the traders. StatArb evolved out of the simpler pairs trade strategy (see Section (7.2.3)), but it considers a portfolio of a hundred or more stocks, some long and some short, that are carefully matched by sector and region to eliminate exposure to beta and other risk factors.

Broadly speaking, StatArb is actually any strategy that is bottom-up, beta-neutral in approach and uses statistical/econometric techniques in order to provide signals for execution. The mathematical concepts used in Statistical Arbitrage range from Time Series Analysis, Principal Components Analysis (PCA), Co-integration, neural networks and pattern recognition, covariance matrices and efficient frontier analysis to advanced concepts in particle physics such as free energy and energy minimisation. Signals are often generated through a contrarian mean-reversion principle, but they can also be designed using such factors as lead/lag effects, corporate activity, short-term momentum, etc. This is usually referred to as a multi-factor approach to StatArb. Because of the large number of stocks involved, the high portfolio turnover and the fairly small size of the effects one is trying to capture, the strategy is often implemented in an automated fashion and great attention is placed on reducing trading costs.

As an example, an automated portfolio may consists of two phases. In the scoring phase, each stock in the market is assigned a numeric score or rank reflecting its desirability; high scores indicate stocks that should be held long and low scores indicate stocks that are candidates for shorting. The details of the scoring formula vary and are highly proprietary, but, generally (as in pairs trading), they involve a short term mean reversion principle so that stocks having done unusually well in the past week receive low scores and stocks having underperformed receive high scores. In the second or risk reduction phase, the stocks are combined into a portfolio in carefully matched proportions so as to eliminate, or at least greatly reduce, market and factor risk.

Statistical arbitrage is subject to model weakness as well as stock or security-specific risk. The statistical relationship on which the model is based may be spurious, or may break down due to changes in the distribution of returns

on the underlying assets. Factors, which the model may not be aware of having exposure to, could become the significant drivers of price action in the markets, and the inverse applies also. The existence of the investment based upon model itself may change the underlying relationship, particularly if enough entrants invest with similar principles. The exploitation of arbitrage opportunities themselves increases the efficiency of the market, thereby reducing the scope for arbitrage, so continual updating of models is necessary. Further, StatArb has developed to a point where it is a significant factor in the marketplace, that existing funds have similar positions and are in effect competing for the same returns.

7.2.5 Mean-reversion strategies

Mean reversion strategies have been very popular since 2009. They have performed exceptionally well for the past 10 years, performing well even during the 2008-09 bear market. Different versions have been popularised, notably by Larry Connors and Cezar Alvarez (David Varadi, Michael Stokes). Some of the indicators used are

- the RSI indicator (Relative Strength Index)
- a short term simple moving average
- the boillinger bands

The concept is the same: If price moved up today, it will tend to revert (come down) tomorrow.

Example on RSI (on GSPC index):
Analysing data since 1960, for the last 10 years (2000-2010) the market has changed and has become mean reverting: buy on oversold and sell on overbought. Mean-reverting strategies have not performed as well starting 2010. Let's say we traded the opposite strategy. Buy if short term RSI is high, sell if it's low (trend). As expected, it does well up to 2000, then it's a disaster.

7.2.6 Adaptive strategies

An adaptive strategy depends on a Master Strategy and some Allocation Rules:

1. Master Strategy
   - Instead of deciding on which RSI period and thresholds (sample sizes) to use, we use 6 different versions (RSI(2), RSI(3) and RSI(4), each with different thresholds).
   - One Non-Mean-Reverting strategy: If RSI(2) crosses 50 up then buy. If it crosses below, sell
2. Allocation Rules
   - We measure risk adjusted performance for the last 600 bars for each of the 7 strategies.
   - The top 5 get allocated capital; Best gets 50% of account to trade with, then 2nd gets 40%, 3rd gets 30% etc.

Total allocation is 150%, meaning if all strategies were trading we would have to use 1.5x leverage.

Based on the previous example, up to 2002 the system takes positions mostly in the trend following strategy while starting as early as 1996 mean-reverting strategies start increasing positions and eventually take over by 2004. There is a 3 year period (2000-2003) of continuous draw-down as the environment changes and the strategy tries to adapt. Notice that the trend-following RSI strategy (buy on up, sell on down) briefly started traded in August 2011, after being inactive for 9 years.

7.2.7 Constraints and fees on short-selling

Many complications are related to the use of short-selling in the form of constraints and fees. For instance, a pair trading strategy requiring one to be long one share and short another, is called self-financing strategy (see Alexander et al. [2002]). That is, an investor can borrow the amount he wants to invest, say from a bank, then to be able to short a share, he deposits the borrowed amount with the financial institution as collateral and obtains borrowed shares. Thus, the only cost he has to pay is the difference between borrowing interest rates paid by the investor and lending interest rates paid by the financial institution to the investor. Subsequently, to go short a given share, the investor sells the borrowed share and obtains cash in return. From the cash he finances his long position. On the whole, the only cost is the difference between both interest rates (paid vs. received). A more realistic approach is the situation where an investor does not have to borrow capital from a bank in the beginning (e.g. the case of a hedge fund that disposes of capital from investors) allows us to drop the difference in interest rates. Therefore, a short position would be wholly financed by an investor. However, in that case the investor must establish an account with a prime broker who arranges to borrow stocks for short-selling. The investor may be subject to buy-in and have to cover the short positions. The financial intermediation cost of borrowing including the costs associated with securing and administrating lendable stocks averages 25 to 30 basis points. This cost is incurred as a hair-cut on the short rebate received from the interest earned on the short sale proceeds. Short-sellers may also incur trading opportunity costs because exchange rules delay or prevent short sales. For example, dealing with the 50 most liquid European shares, we can consider conservative total transaction costs of 0.3% one-way in total for both shares (see Alexander et al. [2002]) consisting of transaction costs 0.1% of brokerage fee for each share (thus 0.2% for both shares), plus a bid-ask spread for each share (long and short) which we assume to be 0.05% (0.3% in total for both shares). Long-short portfolio can take advantage of the leverage allowed by regulations (two-to-one leverage) by engaging in about twice as much trading activity as a comparable unlevered long-only strategy. The differential is largely a function of the portfolio's leverage. For example, given a capital of $\$10$ million the investor can choose to invest $\$5$ million long and sell $\$5$ million short. Trading activity for the resulting long-short portfolio will be roughly equivalent to that for a $\$10$ million long-only portfolio. If one considers management fees per dollar of securities positions, rather than per dollar of capital, there should not be much difference between long-short and long-only portfolio. In general, investors should consider the amount of active management provided per dollar of fees. Long-only portfolios have a sizable hidden passive component as only their overweights and underweights relative to the benchmark are truly active. On the other hand, long-short portfolio is entierly active such that in terms of management fees per active dollars, long-short may be substantially less costly than long-only portfolio. Moreover, long-short management is almost always offered on a performance-fee basis. Long-short is viewed as riskier than long-only portfolio due to potentially unlimited losses on short positions. In practice, long-short will incur more risk than long-only portfolio to the extent that it engages in leverage, and/or takes more active positions. Taking full advantage of the leverage available will have at risk roughly double the amount of assets invested in a comparable unlevered long-only strategy. Note, both the portfolio's degree of leverage and its activeness are within the explicit control of the investor.

7.3 Enhanced active strategies

7.3.1 Definition

Enhanced active equity portfolios (EAEP) seek to improve upon the performance of actively managed long-only portfolios by allowing for short-selling and reinvestment of the entire short sales proceeds in incremental long positions. This style advances the pursuit of active equity returns by relaxing the long-only constraint while maintaining full portfolio exposure to market return and risk. Enhanced active equity strategy has short positions equal to some percentage $X\%$ of capital (generally 20% or 30% and possibly 100% or more) and an equal percentage of leveraged long positions $(100 + X)\%$. On a net basis, the portfolio has a 100% exposure to the market and it often has a target beta of one. For example, in a 130-30 portfolio with initial capital of $\$100$, an investor can sell short $\$30$ of securities and use the $\$30$ proceeds along with $\$100$ of capital to purchase $\$130$ of long positions. This way, the 130-30, or active extension, portfolio structure provides fund managers with exposure to market returns unavailable to market neutral

long-short portfolios. A 130/30 strategy has two basic components: forecasts of expected returns, or alphas, for each stock in the portfolio universe, and an estimate of the covariance matrix used to construct an efficient portfolio. With modern prime brokerage structures (called enhanced prime brokerage), the additional long purchases can be accomplished without borrowing on margin, allowing for the management style called enhanced active equity. As a result, the 130-30 products were expected to reach $2 trillion by 2010 (see Tabb et al. [2007]).

7.3.2 Some misconceptions

Since enhanced active strategies differ in some fundamental ways from other active equity strategies, both long-only and long-short, some misconceptions about these strategies formed, which Jacobs et al. [2007a] showed not to survive objective scrutiny. For instance, a portfolio that can sell short can underweight in larger amounts, so that meaningful underweights of most securities can only be achieved if short selling is allowed. Hence, a 120-20 portfolio can take more and/or larger active overweight positions than a long-only portfolio with the same amount of capital. Further, it is not optimum to split a 120-20 equity portfolio into a long-only 100-0 portfolio and a 20-20 long-short portfolio because the real benefits of any long-short portfolio emerge only with an integrated optimisation that considers all long and short positions simultaneously. In that setting, Jacobs et al. [2005] developed a theoretical framework and algorithms for integrated portfolio optimisation and showed that it must satisfies two constraints

1. the sum of the long position weights is $(100 + X)\%$.
2. the sum of the short position weights is $X\%$.

Short-selling, even in limited amounts, can extend portfolio underweights substantially. Opportunities for shorting are not necessarily mirror images of the ones for buying long. It is assumed that overvaluation is more common and larger in magnitude than undervaluation (non-linear relation). Also, price reactions to good and bad news may not be symmetrical. An enhanced active portfolio can take short positions as large as the prime's broker policies on leverage allow. For example, the portfolio could short securities equal to 100% of capital and use the proceeds plus the capital to purchase long positions, resulting in a 200-100 portfolio. Comparing an enhanced active 200-100 portfolio with an equitized market-neutral long-short portfolio with 100% of capital in short positions, 100% in long positions, and 100% in an equity market overlay (stock index futures, swaps, exchange traded funds (ETFs)), we see that they are equivalent with identical active weights and identical market exposures. However, the equity overlay is passive, whereas with an enhanced active equity portfolio, market exposure is established with individual security positions. For each $\$100$ of capital, the investor has $\$300$ in stock positions to use in pursuing return and controlling risk. Further, the cost of both strategies is about the same.

While all enhanced portfolios are in a risky position in terms of potential value added or lost relative to the benchmark index return, losses on unleveraged long positions are limited because a stock price can not drop below zero, but losses on short positions are theoretically unlimited as stock price can rise to infinity. However, this risk can be minimised by diversification and rebalancing so that losses in some positions can be mitigated by gains in others. A 120-20 portfolio is leveraged, in that it has $\$140$ at risk for every $\$100$ of capital invested. The market exposure created by the 20% in leveraged long positions is offset, however, by the 20% sold short. The portfolio has a 100% net exposure to the market, and with appropriate risk control, a marketlike level of systematic risk (a beta of 1). The leverage and added flexibility can be expected to increase excess return and residual risk relative to the benchmark. If the manager is skilled at security selection and portfolio construction, any incremental risk borne by the investor should be compensated for by incremental excess return. Since EAEP have a net market exposure of 100%, any pressures put on individual security prices should net out at the aggregate market level. Turnover in an enhanced active equity portfolio should be roughly proportional to the leverage in the portfolio. With $\$140$ in positions in a 120-20 portfolio, versus $\$100$ in a long-only portfolio, turnover can be expected to be about 40% higher in the 120-20 portfolio. The portfolio optimisation process should account for expected trading costs so that a trade does not occur unless the expected benefit in terms of risk-adjusted return outweighs the expected cost of trading.

Michaud [1993] argued that costs related to short sales are an impediment to efficiency. No investment strategy provides a free lunch. An enhanced active equity strategy has an explicit cost, namely a stock loan fee paid to the prime broker. The prime broker arranges for the investor to borrow the securities that are sold short and handles the collateral for the securities' lenders. The stock loan fee amounts to about 0.5% annually of the market value of the shares shorted (about 10 bps of capital for a 120-20 portfolio). It will usually incur a higher management fee than a long-only portfolio and higher transaction costs, but it offers a more efficient way of managing equities than a long-only strategy allows. The incremental underweights and overweights can lead to better diversification than in a long-only portfolio. Moreover, the enhanced active portfolio may incur more trading costs than a long-only portfolio because, as security prices change, it needs to trade to maintain the balance between its short and long positions relative to the benchmark. For example, assume that a 120-20 portfolio experiences adverse stock price moves so that its long positions lose \$2 (prices drops) and its short positions loose \$3 (prices raise), causing capital to decline from \$100 to \$95. The portfolio now has long positions of \$118 and short positions of \$23, not the desired portfolio proportions (120% of \$95 is \$114 and 20% is \$19). To reestablish portfolio exposures of 120% of capital as long positions and 20% of capital as short positions, the manager needs to rebalance by selling \$4 of long positions and using the proceeds to cover \$4 of short positions. The resulting portfolio restores the 120-20 proportions because the \$114 long and \$19 short are respectively 120% and 20% of the \$95 capital. If an EAEP is properly constructed with the use of integrated optimisation, the performance of the long and short positions can not be meaningfully separated.

The unique characteristics of 130-30 portfolio strategy suggest that the existing indexes such as the $S\&P$ 500 are inappropriate benchmarks for leveraged dynamic portfolios. Lo et al. [2008] provided a new benchmark incorporating the same leverage constraints and the same portfolio construction, but which is otherwise transparent, investable, and passive. They used only information available prior to each rebalancing date to formulate the portfolio weights and obtained a dynamic trading portfolio requiring monthly rebalancing. The introduction of short sales and leverage into the investment process led to dynamic indexes capable of capturing time-varying characteristics.

7.3.3 Some benefits

Recently, the centre stage of the long-short debate focused on whether efficiency gains result from relaxing the long-only constraint. Brush [1997] showed that adding a long-short strategy to a long strategy expands the mean-variance efficient frontier, provided that long-short strategies have positive expected alphas. Grinold et al. [2000b] showed that information ratios (IR) decline when one moves from a long-short to a long-only strategy. Jacobs et al. [1998] [1999] further elaborated on the loss in efficiency occurring as a result of the long-only constraint. Martielli [2005] and Jacobs and Levy [2006] provided an excellent practical perspective on the mechanics of enhanced active equity portfolio construction and a number of operational considerations. They compared the enhanced active equity portfolio (EAEP) with traditional long-only passive and active approaches to portfolio management as well as other long-short approaches including market-neutral and equitized long-short. EAEPs are expected to outperform long-only portfolios based on comparable insights. They afford managers greater flexibility in portfolio construction, allowing for fuller exploitation of investment insights. They also provide managers and investors with a wider choice of risk-return trade-offs. The advantages of enhanced active equity over equitized long-short strategies are summarised in Jacobs et al. [2007b].

Clarke et al. [2002] developed a framework for measuring the impact of constraints on the value added by and the performance analysis of constrained portfolios. Further, Clarke et al. [2004] found that short sale constraints in a long-only portfolio cause the most significant reduction in portfolio efficiency. They showed that lifting this constraint is critical for improving the information transferred from stock selection models to active portfolio weights. Sorensen et al. [2007] used numerical simulations of long-short portfolios to demonstrate the net benefits of shorting and to compute the optimal degree of shorting as a function of alpha (manager skill), desired tracking error (risk target), turnover, leverage, and trading costs. They also found that there was no universal optimal level of short selling in an active extension portfolio, but that level varied according to different factors and market conditions. Johnson et al. [2007] further emphasised the loss in efficiency from the long-only constraint as well as the importance of

the concerted selection of gearing and risk in the execution of long-short portfolios. Adopting several simplifying assumptions regarding the security covariance matrix and the concentration profile of the benchmark, Clarke et al. [2008] derived an equation that shows how the expected short weight for a security depends on the relative size of the security's benchmark weight and its assigned active weight in the absence of constraints. They argue that to maintain a constant level of active risk, the long-short ratio should be allowed to vary over time to accommodate changes in individual security risk, security correlation, and benchmark weight concentration.

7.3.4 The enhanced prime brokerage structures

As explained by Jacobs et al. [2006], with a traditional margin account, the lenders of any securities sold short must be provided with collateral at least equal to the current value of the securities (see details in Section 7.2.7). When the securities are first borrowed, the proceeds from the short sale usually serve as this collateral. As the short positions subsequently rise or fall in value, the investor's account provides to or receives from the securities' lenders cash equal to the change in value. To avoid the need to borrow money from the broker to meet these collateral demands, the account usually maintains a cash buffer. Market-neutral long-short portfolios have traditionally been managed in a margin account, with a cash buffer of 10% typically maintained to meet the daily marks on the short positions. Long positions may sometimes need to be sold to replenish the cash buffer (without earning investment profits).

With the enhanced brokerage structures available today, the investor's account must have sufficient equity to meet the broker's maintenance margin requirements, generally 100% of the value of the shares sold short plus some additional percentage determined by the broker. This collateral requierment is usually covered by the long positions. The investor does not have to meet cash marks to market on the short positions. The broker cover those needs and is compensated by the stock loan fee. Also, dividends received on long positions can be expected to more than offset the amount the account has to pay to reimburse the securities' lenders for dividends on the short positions. The investor thus has little need for a cash buffer in the account. An enhanced active portfolio will generally retain only a small amount of cash, similar to the frictional cash retained in a long-only portfolio.

More formally, the enhanced prime brokerage structures allow investors to establish a stock loan account with a broker where the investor is not a customer of the prime broker, as would be the case with a regular margin account, but rather a counterparty in the stock loan transaction [3]. This is an important distinction for at least four reasons:

1. Investors can use the stock loan account to borrow directly the shares they want to sell short. The shares the investor holds long serve as collateral for the shares borrowed. The broker arranges the collateral for the securities' lenders, providing cash, cash equivalents, securities, or letters of credit. Hence, the proceeds from the short sales are available to the investor to purchase securities long.

2. The shares borrowed are collateralized by securities the investor holds long, rather than by the short sale proceeds, eliminating the need for a cash buffer. All the proceeds of short sale and any other available cash can thus be redirected toward long purchases.

3. A stock loan account in contrast to a margin account provides critical benefits for a tax-exempt investor. The long positions established in excess of the investor's capital are financed by the proceeds from the investor's sale of short positions. The longs are not purchased with borrowed funds.

4. The investor being a counterparty in a stock loan account, the investor's borrowing of shares to sell short is not subject to Federal Reserve Board Regulation T (limits on leverage). Instead, the investor's leverage is limited by the broker's own internal lending policies.

[3] To establish a stock loan account with a prime broker, the manager must meet the criteria for a Qualified Professional Asset Manager. For a registered investment advisor, it means more than $\$85$ million of client assets under management and $\$1$ million of shareholders' equity.

In exchange for its lending services (arranging for the shares to borrow and handling the collateral), the prime broker charges an annual fee equal to about 0.50% of the market value of the shares shorted (fees may be higher for harder-to-borrow shares or smaller accounts). For a 120-20 portfolio with 20% of capital shorted, the fee as a percentage of capital is thus about 0.10%. Generally, the broker also obtains access to the shares the investor holds long, up to the dollar amount the investor has sold short, without paying a lending fee to the investor. Hence, the broker can lend these shares to other investors to sell short, and in turn, the investor can borrow the shares the broker can hypothecate from other investors, as well as the shares the broker holds in its own accounts and the share it can borrow from other lenders.

7.4 Measuring the efficiency of portfolio implementation

7.4.1 Measures of efficiency

There are a number of studies measuring the efficiency of portfolio implementation, such as Grinold [1989], who introduced the Fundamental Law (FL) of active management, given by the equation

$$IR = IC.\sqrt{N}$$

where IR is the observed information ratio given in Equation (7.2.1), IC is the information coefficient (a measure of manager skill) given by the correlation of forecast security returns with the subsequent realised security returns, and N is the number of securities in the investment universe. Even though the FL is an approximation, the main intuition is that returns are a function of information level, breadth of investment universe and portfolio risk. The law was extended by Clarke et al. [2002] who introduced the idea of transfer coefficient (TC) to measure the efficiency of portfolio implementation. It is a measure of how effectively manager information is transferred into portfolio weights. The transfer coefficient is defined as the cross-sectional correlation of the risk-adjusted forecasts across assets and the risk-adjusted active portfolio weights in the same assets

$$TC = \rho(w_a\sigma_e, \frac{\alpha}{\sigma_e})$$

where $w_a\sigma_e$ is a vector of risk adjusted active weights, $\sigma_{e,i}$ is the residual risk for each asset (the risk of each asset not explained by the benchmark portfolio), and α is a vector of forecast active returns (forecast returns in excess of benchmark related return). Hence, the TC measures the manager's ability to invest in a way consistent with their relative views on the assets in their investment universe. While a perfectly consistent investment portfolio has a TC of one, any inconsistency in implementation will reduce the TC below one. Assuming that managers have no restrictions on the construction of a portfolio from the information set they possess, the equation becomes

$$IR = TC.IC.\sqrt{N}$$

where TC acts as a scaling factor on the level of information. In absence of any constraints $TC = 1$, otherwise it is below 1 since the constraints place limits on how efficiently managers can construct portfolios reflecting their forecasts. This result infers that portfolio outperformance is not only driven by the the ability to forecast security returns, but also by the ability to frame those security returns in the form of an efficient portfolio.

Note, we need to know the forecasts and model estimates of residual risk $\sigma_{e,i}$ for every asset i to accurately measure the TC of a fund at a particular time. However, only the portfolio managers themselves should have access to this kind of information. Raubenheimer [2011] proposed a simple metrics, called implied transfer coefficient (ITC), requiring only the weights of the benchmark assets and the investment weight constraints. Nonetheless, we need an understanding of the distribution of likely security weightings in the portfolio. Grinold [1994] proposed the alpha generation formula which was generalised by Clarke er al. [2006]

$$\alpha = IC\sigma^{\frac{1}{2}}S_N$$

where σ is an $N \times N$ estimated covariance matrix of the returns of the securities, and S_N is an $N \times 1$ vector of randomised standard normal scores. This equation presents forecasted excess returns as generated by a random normal process, scaled by skill and risk. Clarke et al. [2008] and Sorensen et al. [2007] argued that, if forecast excess returns follow a random process, the distribution of optimal active weights resulting from these forecasts could be derived accordingly. These simulated distributions of asset weights can provide sound justification for various weight constraints which are appropriate for each asset and across changing investment views. They considered a simplified two-parameter variance-covariance matrix to simulated the active weight distributions by setting all individual asset variances to a single value σ, and all pairwise correlations to the same value ρ. Under these assumptions, they showed that the unconstrained optimal active weights are normally distributed with a mean of zero and a variance proportional to the active risk

$$w_a \sim N\big(0, \frac{\sigma_A}{\sqrt{N}} \frac{1}{\sigma\sqrt{1-\rho}}\big)$$

where σ_A is the target active risk of the portfolio. Increasing volatility and decreasing correlation (increasing cross-sectional variation) result in a narrower distribution of active weights and a lower probability of needing short positions to optimally achieving a particular active risk target. All things being equal, a wider distribution of active weights in each security is required with greater active targets. This increase in active weight spread is exponentially increased by a reduced investment universe (smaller N). Hence, given the same active risk targets, funds managed on a smaller asset universe universe or benchmark will likely be more aggressive in their individual active weights per asset than funds managed in a more diverse universe. Further, the wider the distribution of active weights, the more likely short positions in the smaller stocks will be required in the optimal portfolio construction.

7.4.2 Factors affecting performances

Following Segara et al. [2012] we present some hypotheses relating the performance of active extension portfolios to unique factors:

- Skill levels: Managers with higher skill levels have a greater increase in performance from relaxing the long-only constraint. In the case where a manager has some predictive skill ($IC > 0$), then he will be able to transform larger active weights into greater outperformance, leading to higher level of performance from active extension strategies. However, short selling can only increase up to the point where the additional transaction and financing costs outweigh the marginal benefits.
- Skew in predictive ability: Managers with a higher skew towards picking underperforming stocks can construct active extension portfolios with higher levels of performance (see Gastineau [2008]).
- Risk constraints: Portfolios with higher tracking error targets experience greater performance increase from relaxing the long-only constraint. Portfolio managers must face limits to the size of a tracking error which is a function of portfolio active weights and the variance-covariance matrix. Clarke et al. [2004] found a trade-off between the maximum TC, the target tracking error, and the level of shorting.
- Costs: An increase in costs relative to the skill the manager possesses will at some point lower the performance of active extensions strategies. Transaction, financing and stock borrowing costs increase proportionally to the gross exposure of the fund, which is driven by the level of short selling in the portfolio.
- Volatility: Higher market volatility will increase the performance of active extension strategies. In volatile markets, as greater portfolio concentration may expose a portfolio to higher risk due to lower diversification, an active extension strategy allows for a lower risk target for the same return by using short-side information in a portfolio with added diversification.
- Cross-sectional spread of returns: Active extension portfolios perform better in comparison to long-only portfolios in periods where individual stock returns are more highly correlated. According to Clarke et al. [2008],

in environments of higher correlation between individual security returns, larger active positions are needed to achieve the same target level of outperformance. Hence, a higher level of short selling will allow managers to distribute more efficiently their higher active weights over both long and short positions in the portfolio.

- Market Conditions: The level of outperformance or underperformance of active extension portfolios is equivalent across periods or negative market returns. Having a constant 100% net market exposure and a beta of about 1 when well diversified, an active extension portfolio on average will perform in line with the broader market.

Chapter 8

Describing quantitative strategies

8.1 Time series momentum strategies

Trend-following or momentum investing is about buying assets whose price is rising and selling assets whose price is falling. Cross-sectional momentum strategies in three dimensions (time-series, cross-section, trading frequency), which are the main driver of commodity trading advisors (CTAs) (see Hurst et al. [2010]), were extensively studied and reported to present strong return continuation patterns across different portfolio rebalancing frequencies with high Sharpe ratio (see Jegadeesh et al. [2001], Moskowitz et al. [2012], Baltas et al. [2012a]). Time-series momentum refers to the trading strategy that results from the aggregation of a number of univariate momentum strategies on a volatility-adjusted basis. As opposed to the cross-sectioal momentum strategy which is constructed as a long-short zero-cost portfolio of securities with the best and worst relative performance during the lookback period, the univariate time-series momentum strategy (UTMS) relies heavily on the serial correlation/predictability of the asset's return series. Moskowitz et al. [2012] found strong positive predictability from a security's own past returns across the nearly five dozen futures contracts and several major asset classes studied over the last 25 years. They found that the past 12-month excess return of each instrument is a positive predictor of its future return. This time series momentum effect persists for about a year before partially reversing. Baltas et al. [2012a] showed that time-series momentum strategies have high explanatory power in the time-series of CTA returns. They further documented the existence of strong time-series momentum effects across monthly, weekly and daily frequencies, and confirmed that strategies at different frequencies have low correlation between each other, capturing distinct patterns. This dependence on strong autocorrelation in the individual return series of the contracts poses a substantial challenge to the random walk hypothesis and the market efficiency which was explained by rational and behavioural finance. Using intraday data, Baltas et al. [2012b] explored the profitability of time-series momentum strategies focusing on the momentum trading signals and on the volatility estimation. Results showed that the information content of the price path throughout the lookback period can be used to provide more descriptive indicators of the intertemporal price trends and avoid eminent price reversals. They showed empirically that the volatility adjustment of the constituents of the time-series momentum is critical for the resulting portfolio turnover.

8.1.1 The univariate time-series strategy

Assuming predictability of some market time-series, the univariate time-series momentum strategy (UTMS) is defined as the trading strategy that takes a long/short position on a single asset based on the trading signal $\psi_i(.,.)$ of the recent asset return over a particular lookback period. We let J denote the lookback period over which the asset's past performance is measured and K denote the holding period. In general, both J and K are measured in months, weeks or days depending on the rebalancing frequency of interest. We use the notation M_J^K to denote monthly strategies with a lookback and holding period of J and K months respectively. The notations W_J^K and D_J^K follow similarly for

weekly and daily strategies. Following Moskowitz et al. [2012] (MOP), we construct the return $Y^i_{J,K}(t)$ at time t for the series of the ith available individual strategy as

$$Y^i_{J,K}(t) = \psi_i(t-J,t)R_i(t,t+K) \tag{8.1.1}$$

where $\psi_i(t-J,t)$ is the particular trading signal for the ith asset which is determined during the lookback period and in general takes values in the set $\{-1,0,1\}$ which in turn translates to $\{short, inactive, long\}$. Note, to evaluate the abnormal performance of these strategies, we can compute their alphas from a linear regression of returns (see Equation (8.1.3)) where we control for passive exposures to the three major asset classes on stocks, commodities and bonds.

8.1.2 The momentum signals

Given the return $Y^i_{J,K}(t)$ at time t for the series of the ith available individual strategy in Equation (8.1.1), we consider five different methodologies in order to generate momentum trading signals, all focusing on the asset performance during the lookback period $[t-J,t]$ (see Moskowitz et al. [2012], Baltas et al. [2012b]).

8.1.2.1 Return sign

In that setting, the time-series momentum strategy is defined as the trading strategy that takes a long/short position on a single asset based on the sign of the recent asset return over a particular lookback period. The trading signal is given by

$$\psi_i(t-J,t) = sign_i(t-J,t)$$

where $sign_i(t-J,t)$ is the sign of the J-period past return of the ith asset. That is, a positive (negative) past return dictates a long (short) position. The return of the time-series momentum strategy becomes

$$Y^i_{J,K}(t) = sign_i(t-J,t)R_i(t,t+K) = sign(R_i(t-J,t))R_i(t,t+K) \tag{8.1.2}$$

8.1.2.2 Moving Average

A long (short) position is determined when the J-(month/week) lagging moving average of the price series lies below (above) a past (month/week)'s leading moving average of the price series. Given the price level $S^i(t)$ of an instrument at time t, we let $N_J(t)$ be the number of trading days in the period $[t-J,t]$ and define $A_J(t)$ the average price level during the same time period as

$$A_J(t) = \frac{1}{N_J(t)}\sum_{j=1}^{N_J(t)} S^i(t-N_J(t)+j)$$

The trading signal at time t is determined as

$$MA(t-J,t) = \begin{cases} 1 \text{ if } A_J(t) < A_1(t) \\ -1 \text{ otherwise} \end{cases}$$

Hence, the trading strategy that takes a long/short position on a single asset based on the moving average of the price series over a particular lookback period is

$$Y^i_{J,K}(t) = MA_i(t-J,t)R_i(t,t+K)$$

The idea behind the MA methodology is that when a short-term moving average of the price process lies above a longer-term average then the asset price exhibits an upward trend and therefore a momentum investor should take a

long position. The reverse holds when the relationship between the averages changes. The comparison of the long-term lagging MA with a short-term leading MA gives the MA methodology a market timing feature. The choice of the past month for the short-term horizon is justified, because it captures the most recent trend breaks.

8.1.2.3 EEMD Trend Extraction

This trading signal relies on some extraction of the price trend during the lookback period. We choose to use a recent data-driven signal processing technique, known as the Ensemble Empirical Mode Decomposition (EEMD), which is introduced by Wu et al. [2009] and constitutes an extension of the Empirical Mode Decomposition. The EEMD methodology decomposes a time-series of observations into a finite number of oscillating components and a residual non-cyclical long-term trend of the original series, without virtually imposing any restrictions of stationarity or linearity upon application. That is, the stock price process can be written as the complete summation of an arbitrary number, n of oscillating components $c_k(t)$ for $k = 1, .., n$ and a residual long-term trend $p(t)$

$$S(t) = \sum_{k=1}^{n} c_k(t) + p(t)$$

The focus is on the extracted trend $p(t)$ and therefore an upward (downward) trend during the lookback period determines a long (short) position

$$EEMD(t-J,t) = \begin{cases} 1 \text{ if } p(t) > p(t-J) \\ -1 \text{ otherwise} \end{cases}$$

8.1.2.4 Time-Trend t-statistic

Another way of capturing the trend of a price series is through fitting a linear trend on the J-month price series using least-square. The momentum signal can them be determined based on the significance of the slope coefficient of the fit

$$\frac{S(j)}{S(t-N_J(t))} = \alpha + \beta j + \epsilon(j) \ , j = 1, 2, .., N_J(t)$$

Estimating this model for the asset using all $N_J(t)$ trading days of the lookback period yields an estimate of the time-trend, given by the slope coefficient β. The significance of the trend is determined by the t-statistic of β, denoted as $t(\beta)$, and the cutoff points for the long/short position of the trading signal are chosen to be $+2/-2$ respectively

$$TREND(t-J,t) = \begin{cases} 1 \text{ if } t(\beta) > 2 \\ -1 \text{ if } t(\beta) < -2 \\ 0 \text{ otherwise} \end{cases}$$

In order to account for potential autocorrelation and heteroskedasticity in the price process, Newey et al. [1987] t-statistics are used. Note, the normalisation of the regressand in above equation is done for convenience, since it allows for cross-sectional comparison of the slope coefficient, when necessary. The t-statistic of β is of course unaffected by such scalings.

8.1.2.5 Statistically Meaningful Trend

Bryhn et al. [2011] study the statistical significance of a linear trend and claim that if the number of data points is large, then a trend may be statistically significant even if the data points are very erratically scattered around the trend line. They introduced the term of statistical meaningfulness in order to describe a trend that not only exhibits statistical significance, but also describes the behaviour of the data to a certain degree. They showed that a trend is informative and strong if, except for a significant t-statistic (or equivalently a small p-value where p is the p-value of

the slope coefficient), the R^2 of the linear regression exceeds 65%. Further, they consider a method providing some sort of pre-smoothing in the data before the extraction of the trend. Hence, we split the lookback period in 4 to 10 intervals (i.e. 7 regressions per lookback period per asset) and decide upon a long/short position only if at least one of the regressions satisfies the above criteria

$$SMT(t-J,t) = \begin{cases} 1 \text{ if } t_k(\beta) > 2 \text{ and } R_k^2 \geq 65\% \text{ for some k} \\ -1 \text{ if } t_k(\beta) < -2 \text{ and } R_k^2 \geq 65\% \text{ for some k} \\ 0 \text{ otherwise} \end{cases}$$

where k denotes the kth regression with $k = 1, 2, ..., 7$. Clearly, SMT is a stricter signal than TREND and therefore would lead to more periods of inactivity.

8.1.3 The signal speed

Since the sparse activity could potentially limit the ex-post portfolio mean return, Baltas et al. [2012b] chose to estimate for each contract and for each signal an activity-to-turnover ratio, which is called Signal Speed. It is computed as the square root of the ratio between the time series average of the squared signal value and the time-series average of the squared first-order difference in the signal value

$$(Speed(\psi))^2 = \frac{E[\psi^2]}{E[(\Delta\psi)^2]} = \frac{\frac{1}{T-J}\sum_{t=1}^{T}\psi^2(t-J,t)}{\frac{1}{T-J-1}\sum_{t=1}^{T}\left[\psi(t-J,t)-\psi(t-1-J,t-1)\right]^2}$$

Clearly, the larger the signal activity and the smaller the average difference between consecutive signal values (in other words the smoother the transition between long and short positions), the larger the signal speed. When the signals constantly jump between long $(+1)$ to short (-1) positions the numerator is always equal to 1. Inactive trading allows for smoother transition between long and short positions.

8.1.4 The relative strength index

The relative strength index (RSI) is a technical indicator intended to chart the current and historical strength or weakness of a stock or market based on the closing prices of a recent trading period. The RSI is classified as a momentum oscillator, measuring the velocity and magnitude of directional price movements. Momentum is the rate of the rise or fall in price. The RSI computes momentum as the ratio of higher closes to lower closes: stocks which have had more or stronger positive changes have a higher RSI than stocks which have had more or stronger negative changes. The RSI is most typically used on a 14 day time frame, measured on a scale from 0 to 100, with high and low levels marked at 70 and 30, respectively. For each trading period an upward change U is defined by the close being higher than the previous close

$$U(j\delta) = \begin{cases} S_C(j\delta) - S_C((j-1)\delta) \text{ if } S_C(j\delta) > S_C((j-1)\delta) \\ 0 \text{ otherwise} \end{cases}$$

Similarly, a downward change D si given by

$$D(j\delta) = \begin{cases} S_C((j-1)\delta) - S_C(j\delta) \text{ if } S_C((j-1)\delta) > S_C(j\delta) \\ 0 \text{ otherwise} \end{cases}$$

The average of U and D are calculated by using an n-period Exponential Moving Average (EMA) in the AIQ version but with an equal-weighted moving average in Wilder's original version. The ratio of these averages is the Relative Strength Factor

$$RS = \frac{EMA(U,n)}{EMA(D,n)}$$

The EMA should be appropriately initialised with a simple average using the first n-values in the price series. When the average of D values is zero, the RS value is defined ad 100. The RSF is then converted to a Relative Strength Index in the range $[0, 100]$ as

$$RSI = 100 - \frac{100}{1 + RS}$$

so that when $RS = 100$ then RSI is close to 100 and when $RS = 0$ then $RSI = 0$. The RSI is presented on a graph above or below the price chart. The indicator has an upper line, typically at 70, a lower line at 30 and a dashed mid-line at 50. The inbetween level is considered neutral with the 50 level being a sign of no trend. Wilder posited that when price moves up very rapidly, at some point it is considered overbought, while when price falls very rapidly, at some point it is considered oversold. Failure swings above 70 and below 30 on the RSI are strong indications of market reversals. The slope of the RSI is directly proportional to the velocity of a change in the trend. Cardwell noticed that uptrends generally traded between RSI of 40 and 80 while downtrends traded with an RSI between 60 and 20. When securities change from uptrend to downtrend and vice versa, the RSI will undergo a range shift. Bearish divergence (between stock price and RSI) is a sign confirming an uptrend while bullish divergence is a sign confirming a downtrend. Further, he noted that reversals are the opposite of divergence.

A variation called Cutler's RSI is based on a simple moving average (SMA) of U and D

$$RS = \frac{SMA(U, n)}{SMA(D, n)}$$

When the EMA is used, the RSI value depends upon where in the data file his calculation is started which called the Data Length Dependency. Hence Cutler's RSI is not data length dependent, and it returns consistent results regardless of the length of, or the starting point within a data file. The two measures are similar since SMA and EMA are also similar.

8.1.5 Regression analysis

Before constructing momentum strategies, following Moskowitz et al. [2012], we first assess the amount of return predictability that is inherent in a series of predictors by running a pooled time-series cross-sectional regression of the contemporaneous standardised return on a lagged return predictor. We regress the excess return r_t^i for instrument i in month t on its return lagged h months/weeks/days, where both returns are scaled by their ex-ante volatilities σ_t^i

$$\frac{r_t^i}{\sigma_{t-1}^i} = \alpha + \beta_h Z(t-h) + \epsilon_t^i \tag{8.1.3}$$

where the regressor $Z(t-h)$ is chosen from a broad collection of momentum-related quantities. Note that all regressor choices are normalised, in order to allow for the pooling across the instruments. For example, we can consider the regression

$$\frac{r_t^i}{\sigma_{t-1}^i} = \alpha + \beta_h \frac{r_{t-h}^i}{\sigma_{t-h-1}^i} + \epsilon_t^i$$

Given the vast differences in volatilities we divide all returns by their volatility to put them on the same scale. This is similar to using Generalized Least Squares instead of Ordinary Least Squares (OLS). The regressions are run using lags of $h = 1, 2, .., 60$ months/weeks/days. Another way of looking at time series predictability is to simply focus only on the sign of the past excess return underlying our trading strategies. In a regression setting, this strategy can be captured using the following specification:

$$\frac{r_t^i}{\sigma_{t-1}^i} = \alpha + \beta_h sign(r_{t-h}^i) + \epsilon_t^i \tag{8.1.4}$$

where

$$sign(a) = \begin{cases} 1 \text{ if } a \geq 0 \\ -1 \text{ otherwise} \end{cases}$$

All possible choices are comparable across the various contracts, and refer to a single period $(J = 1)$ to avoid serial autocorrelation in the error term ϵ_t^i. Equation (8.1.4) is estimated for each lag h and regressor Z by pooling all the underlyings together. To allow for the pooling across instruments, all regressor choices are normalised, and the asset returns are normalised. The quantity of interest in these regressions is the t-statistic of the coefficient β_h for each lag h. Large and significant t-statistics essentially support the hypothesis of time-series return predictability. The results are similar across the two regression specifications: strong return continuation for the first year and weaker reversals for the next 4 years (see Moskowitz et al. [2012], Baltas et al. [2012b]).

8.1.6 The momentum profitability

Subsequently, we construct the return series of the aggregate time-series momentum strategy over the investment horizon as the inverse-volatility weighted average return of all available individual momentum strategies

$$R_J^K(t) = \frac{1}{M_t} \sum_{i=1}^{M_t} \frac{C_{sf}}{\sigma_i(t, D)} Y_{J,K}^i(t) \tag{8.1.5}$$

where M_t is the number of available assets at time t, and where C_{sf} is a scaling factor and $\sigma_i(t, D)$ is an estimate at time t of the realised volatility of the ith asset computed using a window of the past D trading days. See Section (3.4) for the description of a family of volatility estimators. This risk-adjustment (use of standardised returns) across instruments allows for a direct comparison and combination of various asset classes with very different return distributions in a single portfolio.

Assuming that the individual time series strategies are mutually independent and that the volatility process is persistent, then the conditional variance of the return can be approximated by

$$Var_t(Y_{J,K}^i(t)) \approx \sigma_i^2(t, D)$$

since $\psi_i^2(t - J, t) = 1$ if we further assume that the frequency of the trading periods when $\psi_i(t - J, t) = 0$ is relatively small. As a result, the conditional variance of the portfolio is approximated by

$$Var_t(R_J^K(t)) \approx \frac{1}{M_t^2} \sum_{i=1}^{M_t} C_{sf}^2$$

This approximation ignores any covariation among the individual momentum strategies as well as any potential changes in the individual volatility processes but it can be used to define the scaling factor C_{sf}. For example, we can consider $D = 60$ trading days. The scaling factor $C_{sf} = 40\%$ is used by MOP in order to achieve an ex-ante volatility equal to 40% for each individual strategy. This is because it results in an ex-post annualised volatility of 12% for their M_{12}^1 strategy roughly matching the level of volatility of several risk factors in their sample period. Baltas et al. [2012b] considered a rolling window of $D = 30$ days and a scaling factor $C_{sf} = 10\% \times \sqrt{M_t}$ to achieve an ex-ante volatility equal to 10%. Regarding the ex-ante volatility adjustment (risk-adjustment), it must be noted that it is compulsory in order to allow us to combine in a single portfolio various contracts of different asset classes with different volatility profiles. Recently, Barroso and Santa-Clara [2012] revised the equity cross-sectional momentum strategy and scaled similarly the winners-minus-losers portfolio in order to form what they call a risk-managed momentum strategy.

Instead of forming a new momentum portfolio every K periods, when the previous portfolio is unwound, we can follow the overlapping methodology of Jegadeesh et al. [2001], and perform portfolio rebalancing at the end of each

month/week/day. The respective monthly/weekly/daily return is then computed as the equally-weighted average across the K active portfolios during the period of interest. Based on this technique, $\frac{1}{K}$-th of the portfolio is only rebalanced every month/week/day. In order to assess the profitability of the portfolio, we consider different momentum trading signal, various out-of-sample performance statistics for the (J, K) time-series momentum strategy. The statistics are all annualised and include the mean portfolio return along with the respective Newey et al. [1987] t-statistic, the portfolio volatility, the dollar growth, the Sharpe ratio and the downside risk Sharpe ratio (see Ziemba [2005]). To analyse the performance of the strategies, we can then plot the annualised Sharpe ratios of these strategies for each stock/futures contract. In most studies, they showed that every single stock/futures contract exhibits positive predictability from past one-year returns. These studies also regressed the strategy for each security on the strategy of always being long, and they got a positive alpha in 90% of the cases. Thus, a time series strategy provides additional returns over and above a passive long position for most instruments.

8.2 Factors analysis

Some theoretical models have been proposed as a framework for researching connections between asset returns and macroeconomic factors. The Arbitrage Pricing Theory (APT) developed by Ross [1976] is one of the theories that relate stock returns to macroeconomic state variables. It argues that the expected future return of a stock can be modelled as a linear function of a variety of economic state variables or some theoretical market indices, where sensitivity of stock returns to changes in every variable is indicated by an economic state variable-specific coefficient. The rate of return provided by the model can then be used for correcting stock pricing. Thus, the current price of a stock should be equal to the expected price at the end of the period discounted by the discount rate that is suggested by the model. The theory suggests that if the current price of a stock diverges from the theoretical price then arbitrage should bring it back into equilibrium. Roll and Ross [1980] argued that the Arbitrage Pricing Theory is an attractive pricing model to researchers because of its modest assumptions and pleasing implications in comparison with the Capital Asset Pricing Model. The vast majority of papers that used APT as a framework attempted to model a short-run relation between the prices of equities and financial and economic variables. They used differenced variables and presumed that the variables were stationary. However, evidence suggests that in the short-run equity prices deviate from their fundamental values and are also driven by non-fundamentals. In this section, we are going to model these non-fundamentals values by assuming they are mean-reverting.

For simplicity, we assume the market is composed of two types of agents, namely, the indexers (mutual fund managers and long-only managers) and the market-neutral agents. The former seek exposure to the entire market or to specific industry sectors with the goal of being generally long the market or sector with appropriate weightings in each stock, whereas the latter seek uncorrelated returns with the market (alpha). In this market, we are going to present a systematic approach to statistical arbitrage defined in Section (7.2.4), and construct market-neutral portfolio strategies based on mean-reversion. This is done by decomposing stock returns into systematic and idiosyncratic components using different definitions of risk factors.

8.2.1 Presenting the factor model

One of the main difficulties in multivariate analysis is the problem of dimensionality, forcing practitioners to use simplifying methods. From an empirical viewpoint, multivariate data often exhibit similar patterns indicating the existence of common structure hidden in the data. Factor analysis is one of those simplifying methods available to the portfolio manager. It aims at identifying a few factors that can account for most of the variations in the covariance or correlation of the data. Traditional factor analysis assumes that the data have no serial correlations. While this assumption is often violated by financial data taken with frequency less than or equal to a week, it becomes more reasonable for asset returns with lower frequencies (monthly returns of stocks or market indexes). If the assumption is violated, one can use parametric models introduced in Section (5.5.2) to remove the linear dynamic dependence of the data and apply factor analysis to the residual series.

Considering factor analysis based on orthogonal factor model, we let $r = (r_1, ..., r_N)^\top$ be the N-dimensional log returns, and assume that the mean and covariance matrix of r are μ and Σ. For a return series, it is equivalent to requiring that r is weakly stationary. The factor model postulates that r is linearly dependent on a few unobservable random variables $F = (f_1, f_2, ..., f_m)^\top$ and N additional noises $\epsilon = (\epsilon_1, .., \epsilon_N)^\top$ where $m < N$, f_i are the common factors, and ϵ_i are the errors. The factor model is given by

$$\begin{aligned} r_1 - \mu_1 &= l_{11}f_1 + ... + l_{1m}f_m + \epsilon_1 \\ r_2 - \mu_2 &= l_{21}f_1 + ... + l_{2m}f_m + \epsilon_2 \\ ... &= ... \\ r_N - \mu_N &= l_{N1}f_1 + ... + l_{Nm}f_m + \epsilon_N \end{aligned}$$

where $r_i - \mu_i$ is the ith mean-corrected value. Equivalently in matrix notation, we get

$$r - \mu = LF + \epsilon \tag{8.2.6}$$

where $L = [l_{ij}]_{N\times m}$ is the matrix of factor loadings, l_{ij} is the loading of the ith variable on the jth factor, and ϵ_i is the specific error of r_i. The above equation is not a multivariate linear regression model as in Section (5.5.2), even though it has a similar appearance, since the m factors f_i and the N errors ϵ_i are unobservable. The factor model is an orthogonal factor model if it satisfies the following assumptions

1. $E[F] = 0$ and $Cov(F) = I_m$, the $m \times m$ identity matrix
2. $E[\epsilon] = 0$ and $Cov(\epsilon) = \Psi = diag\{\psi_1, .., \psi_N\}$ that is, Ψ is a $N \times N$ diagonal matrix
3. F and ϵ are independent so that $Cov(F, \epsilon) = E[F\epsilon^\top] = 0_{m\times N}$

Under these assumptions, we get

$$\begin{aligned} \Sigma &= Cov(r) = E[(r-\mu)(r-\mu)^\top] = E[(LF+\epsilon)(LF+\epsilon)^\top] \\ &= LE[FF^\top]L^\top + E[\epsilon F^\top]L^\top + LE[F\epsilon^\top] + E[\epsilon\epsilon^\top] \\ &= LL^\top + \Psi \end{aligned}$$

and

$$Cov(r, F) = E[(r-\mu)F^\top] = LE[FF^\top] + E[\epsilon F^\top] = L$$

Using these two equations, we get

$$\begin{aligned} Var(r_i) &= l_{i1}^2 + ... + l_{im}^2 + \psi_i \\ Cov(r_i, r_j) &= l_{i1}l_{j1} + ... + l_{im}l_{jm} \\ Cov(r_i, f_j) &= l_{ij} \end{aligned}$$

The quantity $l_{i1}^2 + ... + l_{im}^2$, called the communality, is the portion of the variance of r_i contributed by the m common factors, while the remaining portion ψ_i of the variance of r_i is called the uniqueness or specific variance. The orthogonal factor representation of a random variable r is not unique, and in some cases does not exist. For any $m \times m$ orthogonal matrix P satisfying $PP^\top = P^\top P = I$, we let $L^* = LP$ and $F^* = P^\top F$ and get

$$r - \mu = LF + \epsilon = LPP^\top F + \epsilon = L^*F^* + \epsilon \tag{8.2.7}$$

with $E[F^*] = 0$ and $Cov(F^*) = P^\top Cov(F)P = P^\top P = I$. Thus, L^* and F^* form another orthogonal factor model for r. As a result, the meaning of factor loading is arbitrary, but one can perform rotations to find common factors with nice interpretations. Since P is an orthogonal matrix, the transformation $F^* = P^\top F$ is a rotation in the m-dimensional space.

One can estimate the orthogonal factor model with maximum likelihood methods under the assumption of normal density and prespecified number of common factors. If the common factors F and the specific factors ϵ are jointly normal and the number of common factors are given a priori, then r is multivariate normal with mean μ and covariance matrix $\Sigma_r = LL^\top + \Psi$. One can then use the MLM to get estimates of L and Ψ under the constraint $L^\top \Psi^{-1} L = \Delta$, which is a diagonal matrix.

Alternatively, one can use PCA without requiring the normality of assumption of the data nor the prespecification of the number of common factors, but the solution is often an approximation. Following the description of PCA in Section (5.5.3), we let $(\hat{\lambda}_1, \hat{e}_1), .., (\hat{\lambda}_N, \hat{e}_N)$ be pairs of the eigenvalues and eigenvectors of the sample covariance matrix $\hat{\Sigma}_r$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_N$. Letting $m < N$ be the number of common factor, the matrix of factor loadings is given by

$$\hat{L} = [\hat{l}_{ij}] = \left[\hat{\lambda}_1 \hat{e}_1 | \hat{\lambda}_2 \hat{e}_2 | ... | \hat{\lambda}_m \hat{e}_m\right]$$

The estimated specific variances are the diagonal elements of the matrix $\hat{\Sigma}_r - \hat{L}\hat{L}^\top$, that is $\hat{\Psi} = diag\{\hat{\psi}_1, .., \hat{\psi}_N\}$, where $\hat{\psi}_i = \hat{\sigma}_{ii,r} - \sum_{j=1}^m \hat{l}_{ij}^2$, and $\hat{\sigma}_{ii,r}$ is the (i,i)th element of $\hat{\Sigma}_r$. The communalities are estimated by

$$\hat{c}_i^2 = \hat{l}_{i1}^2 + ... + \hat{l}_{im}^2$$

and the error matrix due to approximation is

$$\hat{\Sigma}_r - (\hat{L}\hat{L}^\top + \hat{\Psi})$$

which we would like close to zero. One can show that the sum of squared elements of the above error matrix is less than or equal to $\hat{\lambda}_{m+1}^2 + ... + \hat{\lambda}_N^2$ so that the approximation error is bounded by the sum of squares of the neglected eigenvalues. Further, the estimated factor loadings based on PCA do not change as the number of common factors m is increased.

For any $m \times m$ orthogonal matrix P, the random variable r can be represented with Equation (8.2.7) and we get

$$LL^\top + \Psi = LPP^\top L^\top + \Psi = L^*(L^*)^\top + \Psi$$

so that the communalities and specific variances remain unchanged under an orthogonal transformation. One would like to find an orthogonal matrix P to transform the factor model so that the common factors have some interpretations. There are infinite possible factor rotations available and some authors proposed criterions to select the best possible rotation (see Kaiser [1958]).

Factor analysis searches common factors to explain the variabilities of the the returns. One must make sure that the assumption of no serial correlations in the data is satisfied which can be done with the multivariate Portmanteau statistics. If serial correlations are found, one can build a VARMA model to remove the dynamic dependence in the data and apply the factor analysis to the residual series.

8.2.2 Some trading applications

8.2.2.1 Pairs-trading

Following the description of pairs trading in Section (7.2.3) we let stocks P and Q be in the same industry or have similar characteristics, and expect the returns of the two stocks to track each other after controlling for beta. Accordingly, if P_t and Q_t denote the corresponding price time series, then we can model the system as

$$\ln \frac{P_t}{P_{t_0}} = \alpha(t - t_0) + \beta \ln \frac{Q_t}{Q_{t_0}} + X_t$$

or, in its differential version

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t \tag{8.2.8}$$

where X_t is a stationary, or mean-reverting process called the cointegration residual, or simply the residual (see Pole [2007]). In many cases of interest, the drift α is small compared to the fluctuations of X_t and can therefore be neglected. This means that, after controlling for beta, the long-short portfolio oscillates near some statistical equilibrium (see Avellaneda et al. [2008]). The model in Equation (8.2.8) suggests a contrarian investment strategy in which we go long 1 dollar of stock P and short β dollars of stock Q if X_t is small and, conversely, go short P and long Q if X_t is large. The portfolio is expected to produce a positive return as valuations converge. The mean-reversion paradigm is typically associated with market over-reaction: assets are temporarily under or over-priced with respect to one or several reference securities (see Lo et al. [1990]).

Generalised pairs-trading, or trading groups of stocks against other groups of stocks, is a natural extension of pairs-trading. The role of the stock Q would be played by an index or exchange-traded fund (ETF) and P would be an arbitrary stock in the portfolio or sector of activity. The analysis of the residuals, based of the magnitude of X_t, suggests typically that some stocks are cheap with respect to the index or sector, others expensive and others fairly priced. A generalised pairs trading book, or statistical arbitrage book, consists of a collection of pair trades of stocks relative to the ETF (or, more generally, factors that explain the systematic stock returns). In some cases, an individual stock may be held long against a short position in ETF, and in others we would short the stock and go long the ETF.

***Remark* 8.2.1** *Due to netting of long and short positions, we expect that the net position in ETFs will represent a small fraction of the total holdings. The trading book will look therefore like a long-short portfolio of single stocks.*

That is, given a set of stocks $S = \{S_1, S_2, .., S_N\}$ we can go long a subset of stocks L_o with h_i dollars invested per ith stock for $i \in L_o$ and short the subset S_o with $h_i\beta_i$ dollars of index or ETF for each stock. Conversely, we can go short a subset of stocks S_o with h_j dollars invested per jth stock for $j \in S_o$ and long the subset L_o with $h_j\beta_j$ dollars of index or ETF for each stock. We can construct the portfolios such that the net position in ETFs will represent a small fraction of the total holdings or even zero by setting $\sum_{k=1}^{N} h_k\beta_k = 0$.

8.2.2.2 Decomposing stock returns

Following the concept of pairs-trading in Section (8.2.2.1), the analysis of residuals will be our starting point. Signals will be based on relative-value pricing within a sector or a group of peers, by decomposing stock returns into systematic and idiosyncratic components and statistically modelling the idiosyncratic part. Here, the emphasis is on the residual that remains after the decomposition is done and not the choice of a set of risk-factors. Given the d-day period discrete return $R_{t-d,t} = \frac{\nabla_d S_t}{S_{t-d}}$ of the underlying process where $\nabla_d S_t = S_t - S_{t-d}$ with period d, we want to explain or predict stock returns. We saw in Section (8.2.1) that one approach is to explain the returns/prices based on some statistical factors

$$R = \sum_{j=1}^{m} \beta_j F_j + \epsilon \, , Corr(F_j, \epsilon) = 0 \, , j = 1, .., m$$

where F_j is the explanatory factor, β_j is the factor loading such that $\sum_{j=1}^{m} \beta_j F_j$ is the explained or systematic portion and ϵ is the residual, or idiosyncratic portion. For example, the CAPM described in Section (2.3.1.2) consider a single explanatory factor called the market portfolio

$$R = \beta F + \epsilon \, , Cov(R, \epsilon) = 0 \, , \; < \epsilon >= 0$$

where F is the returns of a broad-market index (market portfolio). The model implies that if the market is efficient, or in equilibrium, investors will not make money (systematically) by picking individual stocks and shorting the index or vice-versa (assuming uncorrelated residuals) (Sharpe [1964], Lintner [1965]). However, markets may not be efficient, and the residuals may be correlated. In that case, we need additional explanatory factors F to model stock returns (see Ross [1976]). In the multi-factor models above (APT), the factors represent industry returns so that

$$< R >= \sum_{j=1}^{m} \beta_j < F_j >$$

where brackets denote averaging over different stocks. Thus, the problem of correlations of residuals (idiosyncratic risks) will closely depend on the number of explanatory factors in the model.

8.2.3 A systematic approach

8.2.3.1 Modelling returns

A systematic approach in equity when looking at mean-reversion is to look for stock returns devoid of explanatory factors (see Section (1.5.3)), and analyse the corresponding residuals as stochastic processes. Avellaneda et al. [2008] proposed a quantitative approach to stock pricing based on relative performance within industry sectors or PCA factors. They studied how different sets of risk-factors lead to different residuals producing different profit and loss (PnL) for statistical arbitrage strategies. Following their settings, we let $\{R_i\}_{i=1}^{N}$ be the returns of the different stocks in the trading universe over an arbitrary one-day period (from close to close). Considering the econometric factor model in Equation (8.2.1) with the continuous-time model for the evolution of stock prices defined in Equation (1.5.20), we let the return of the kth risky factor be $F_{kt} = \frac{dP_k(t)}{P_k(t)}$

$$\frac{dS_i(t)}{S_i(t)} = \alpha_i dt + \sum_{k=1}^{m} \beta_{ik} \frac{dP_k(t)}{P_k(t)} + dX_i(t)$$

where we let the systematic component of returns $\sum_{k=1}^{m} \beta_{ik} \frac{dP_k(t)}{P_k(t)}$ be driven by the returns of the eigenportfolios or ETFs. Therefore, the factors are either

- eigenportfolios corresponding to significant eigenvalues of the market
- industry ETF, or portfolios of ETFs

The term $dX_i(t)$ is assumed to be the increment of a stationary stochastic process which models price fluctuations corresponding to over-reactions or other idiosyncratic fluctuations in the stock price which are not reflected in the industry sector. Therefore, the approach followed by Avellaneda et al. [2008] was to let the model assumes

- a drift which measures systematic deviations from the sector
- a price fluctuation that is mean-reverting to the overall industry level

Focusing on the residual, they studied how different sets of risk-factors lead to different residuals, and hence, different profit and loss. Market neutrality is achieved via two different approaches, either by extracting risk factors using Principal Component Analysis, or by using industry-sector ETFs as proxies for risk factors.

8.2.3.2 The market neutral portfolio

Following the notation in Section (1.5.2), we consider $h_1, h_2, .., h_N$ to be the dollars invested in different stocks (long or short) and let $S_1, S_2, .., S_N$ be the dividend-adjusted prices. Neglecting transaction costs, we consider the trading portfolio returns given by

$$\sum_{i=1}^{N} h_{i,t} R_{i,t}$$

where $R_{i,t}$ is the expected return on the ith risky security over one period of time. Assuming the stock returns to follow the factor model in Equation (8.2.1) with $\alpha_{i,t} = 0$, the portfolio returns become

$$\sum_{i=1}^{N} h_{i,t} (\sum_{k=1}^{m} \beta_{ik} F_{kt}) + \sum_{i=1}^{N} h_{i,t} X_{i,t}$$

which becomes

$$\sum_{k=1}^{m} (\sum_{i=1}^{N} h_{i,t} \beta_{ik}) F_{kt} + \sum_{i=1}^{N} h_{i,t} X_{i,t}$$

where $\sum_{i=1}^{N} h_{i,t} \beta_{ik}$ is net dollar-beta exposure along factor k and $\sum_{i=1}^{N} h_{i,t}$ is the net dollar exposure of the portfolio.

Definition 8.2.1 *A trading portfolio is said to be market-neutral if the dollar amounts* $\{h_i\}_{i=1}^{N}$ *invested in each of the stocks are such that*

$$\overline{\beta}_k = \sum_{i=1}^{N} h_i \beta_{ik} = 0 \, , \, k = 1, 2, .., m$$

The coefficients $\overline{\beta}_k$ correspond to the portfolio betas, or projections of the portfolio returns on the different factors. A market-neutral portfolio has vanishing portfolio betas; it is uncorrelated with the market portfolio or factors driving the market returns. As a result, cancelling the net dollar-beta exposure $\sum_{i=1}^{N} h_i \beta_{ik}$ for each factor, the portfolio returns become

$$\sum_{i=1}^{N} h_i X_i$$

Thus, a market-neutral portfolio is affected only by idiosyncratic returns (residuals). In G8 economies, stock returns are explained by approximately $m = 15$ factors (or between 10 and 20 factors), and the systematic component of stock returns explains approximately 50% of the variance (see Plerou et al. [2002] and Laloux et al. [2000]).

Further, in this setting, we define the Leverage Ratio as

$$\Lambda = \frac{\sum_{i=1}^{N} |h_{i,t}|}{V_t} \tag{8.2.9}$$

which is also written

$$\Lambda = \frac{\text{Long market value } + | \text{ Short market value } |}{\text{Equity}}$$

Some examples of leverage are

- long-only: $\Lambda = \frac{L}{V}$
- long-only, Reg T: $L \leq 2E$ therefore $\Lambda \leq 2$
- 130-30 investment fund: $L = 1.3V$, $|S| = 0.3V$ therefore $\Lambda = 1.6$
- long-short \$-neutral, Reg T: $L + |S| \leq 2V$ therefore $\Lambda \leq 2$
- long-short equal target position in each stock: $h_i \leq \frac{\Lambda_{max} V}{N}$ therefore $\sum_i |h_i| \leq \Lambda_{max} V$

8.2.4 Estimating the factor model

8.2.4.1 The PCA approach

Although this is very simplistic, the model can be tested on cross-sectional data. Using statistical testing, we can accept or reject the model for each stock in a given list and then construct a trading strategy for those stocks that appear to follow the model and yet for which significant deviations from equilibrium are observed. One of the problem is to find out if the residuals can be fitted to (increments of) OU processes or some other mean-reversion processes? If it is the case, we need to estimate the typical correlation time-scale.

In risk-management, factor analysis is used to measure exposure of a portfolio to a particular industry of market feature. One relies on dimension-reduction technique for the study systems with a large number of degrees of freedom, making the portfolio theory viable in practice. Hence, one can consider PCA for extracting factors from data by using historical stock price data on a cross-section of N stocks going back M days in the past. Considering the time window $t = 0, 1, 2, .., T$ (days) where $\Delta t = \frac{1}{252}$ and a universe of N stocks, we let $\{R_i\}_{i=1}^N$ be the returns of the different stocks in the trading universe over an arbitrary one-day period (from close to close). The returns data is represented by a $T \times N$ matrix $R(i,t)$ for $i = 1, .., N$ with covariance matrix Σ_R and elements

$$\sigma_i^2 = \frac{1}{T-1} \sum_{t=1}^{T} (R(i,t) - \overline{R}_i)^2 \, , \overline{R}_i = \frac{1}{T} \sum_{t=1}^{T} R(i,t)$$

Data centring being an important element of PCA analysis as it helps minimizing the error of mean squared deviation, the standardised returns are given by

$$Y(i,t) = \frac{R(i,t)}{\sigma_i} \text{ or } \overline{Y}(i,t) = \frac{R(i,t) - \overline{R}_i}{\sigma_i}$$

such that the empirical correlation matrix of the data is defined by

$$\Gamma(i,j) = \frac{1}{T-1} \sum_{t=1}^{T} Y(i,t) Y(j,t)$$

where $Rank(\Gamma) \leq \min(N, T)$. So one can consider

$$\Gamma(i,j) = \frac{1}{T-1} \sum_{t=1}^{T} \overline{Y}(i,t) \overline{Y}(j,t)$$

such that for any index i we get

$$\Gamma(i,i) = \frac{1}{T-1}\sum_{t=1}^{T}(\overline{Y}(i,t))^2 = 1$$

One can regularise the correlation matrix as follow

$$C(i,j) = \frac{1}{T-1}\sum_{t=1}^{T}(R(i,t) - \overline{R}_i)(R(j,t) - \overline{R}_j) + \gamma\delta(i,j)\ , \gamma = (10)^{-9}$$

where δ_{ij} is the Kronecker delta, and $C(i,i) \approx \sigma_i^2$. We can obtain the matrix as

$$\Gamma^{reg}(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}$$

This is a positive definite correlation matrix. It is equivalent for all practical purposes to the original one but is numerically stable for inversion and eigenvector analysis. this is especially useful when $T << N$. If we consider daily returns, we are faced with the problem that very long estimation windows $T >> N$ do not make sense because they take into account the distant past which is economically irrelevant. On the other hand, if we just consider the behaviour of the market over the past year, for example, then we are faced with the fact that there are considerably more entries in the correlation matrix than data points. The commonly used solution to extract meaningful information from the data is Principal Components Analysis.

8.2.4.2 The selection of the eigenportfolios

Following Section (5.5.3), we can now let $\lambda_1 > \lambda_2 \geq .. \geq \lambda_N \geq 0$ be the eigenvalues ranked the in decreasing order and $V^{(j)} = (V_1^{(j)}, V_2^{(j)}, .., V_N^{(j)})$ jor $j = 1, 2, .., N$ be the corresponding eigenvectors. We now need to estimate the significant eigenportfolios which can be used as factors. Analysing the density of states of the eigenvalues, we let m to be a fixed number of eigenvalues to extract the factors close to the number of industry sector. In that setting, the eigenportfolio for each index j is

$$F_{jt} = \sum_{i=1}^{N} V_i^{(j)} Y(i,t) = \sum_{i=1}^{N} \frac{V_i^{(j)}}{\sigma_i} R(i,t)\ , j = 1, 2, .., m$$

where $Q_i^{(j)} = \frac{V_i^{(j)}}{\sigma_i}$ is the respective amounts invested in each of the stock. We use the coefficients of the eigenvectors and the volatilities of the stocks to build portfolio weights. It corresponds to the returns of the eigenportfolios which are uncorrelated in the sense that the empirical correlation of F_j and $F_{j'}$ vanishes for $j \neq j'$. These random variables span the same linear space as the original returns. As each stock return in the investment universe can be decomposed into its projection on the m factors and a residual, thus the PCA approach delivers a natural set of risk-factors that can be used to decompose our returns. Assuming that the correlation matrix is invertible, we get

$$< R_i, R_j >= C(i,j) = \sum_{k=1}^{m} \lambda_k V_i^{(k)} V_j^{(k)}$$

with the factors

$$F_k = \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R(i)\ , \tilde{F}_k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R(i)$$

and the norms

$$< F_k^2 >= \lambda_k\ , \ < \tilde{F}_k^2 >= 1\ , \ < \tilde{F}_k \tilde{F}_{k'} >= \delta_{kk'}$$

so that given the return $R_i = \sum_{k=1}^m \beta_{ik} F_k$ we get the coefficient

$$\beta_{ik} = \sigma_i \sqrt{\lambda_k} V_i^{(k)}$$

It is not difficult to verify that this approach corresponds to modelling the correlation matrix of stock returns as a sum of a rank-m matrix corresponding to the significant spectrum and a diagonal matrix of full rank

$$C(i,j) = \sum_{k=1}^m \lambda_k V_i^{(k)} V_j^{(k)} + \epsilon_{ii}^2 \delta_{ij}$$

where δ_{ij} is the Kronecker delta and ϵ_{ii}^2 is given by

$$\epsilon_{ii}^2 = 1 - \sum_{k=1}^m \lambda_k V_i^{(k)} V_i^{(k)}$$

so that $C(i,i) = 1$. This means that we keep only the significant eigenvalues/eigenvectors of the correlation matrix and add a diagonal noise matrix for the purposes of conserving the total variance of the system.

Laloux et al. [2000] pointed out that the dominant eigenvector is associated with the market portfolio, in the sense that all the coefficients $V_i^{(1)}$ for $i = 1, .., N$ are positive. Thus, the eigenportfolio has positive weights $Q_i^{(1)} = \frac{V_i^{(1)}}{\sigma_i}$ which are inversely proportional to the stock's volatility. It is consistent with the capitalisation-weighting, since larger capitalisation companies tend to have smaller volatilities. Note, the remaining eigenvectors must have components that are negative, in order to be orthogonal to $V^{(i)}$. However, contrary to the interest-rate curve analysis, one can not apply the shape analysis to interpret the PCA.

Another method consists in using the returns of sector ETFs as factors. In this approach, we select a sufficiently diverse set of ETFs and perform multiple regression analysis of stock returns on these factors. Unlike the case of eigenportfolios, ETF returns are not uncorrelated, so there can be redundancies: strongly correlated ETFs may lead to large factor loadings with opposing signs for stocks that belong to or are strongly correlated to different ETFs. To remedy this, we can perform a robust version of multiple regression analysis to obtain the coefficients β_{ij} such as the matching pursuit algorithm or the ridge regression. Avellaneda et al. [2008] associated to each stock a sector ETF and performed a regression of the stock returns on the corresponding ETF returns. Letting $I_1, I_2, .., I_m$ be the class of ETFs spanning the main sectors in the economy, and R_{I_j} be the corresponding returns, they decomposed the ETF as

$$R_i = \sum_{j=1}^m \beta_{ij} R_{I_j} + \epsilon_i$$

While we need some prior knowledge of the economy to identify the right ETFs to explain returns, the interpretation of the factor loadings is more intuitive than for PCA. Note, ETF holdings give more weight to large capitalisation companies, whereas PCA has no apriori capitalisation bias.

8.2.5 Strategies based on mean-reversion

8.2.5.1 The mean-reverting model

We consider the evolution of stock prices in Equation (1.5.20) and test the model on cross-sectional data. For instance, in the ETF framework, $P_k(t)$ represents the mid-market price of the k-th ETF used to span the market. In practice, only ETFs that are in the same industry as the stock in question will have significant loadings, so we could also work with the simplified model

$$\beta_{ik} = \begin{cases} \frac{Cov(R_i, R_{P_k})}{Var(R_{P_k})} \text{ if stock } i \text{ is in industry } k \\ 0 \text{ otherwise} \end{cases}$$

where each stock is regressed to a single ETF representing its peers. In order to get a market-neutral portfolio, we we introduce a parametric model for $X_i(t)$, namely, the Ornstein-Uhlembeck (OU) process with SDE

$$dX_i(t) = k_i(m_i - X_i(t))dt + \sigma_i dW_i \ , \ k_i > 0$$

where $k_i > 0$ and $\{dW_i\}_{i=1}^N$ are uncorrelated. This process is stationary and auto-regressive with lag 1 ($AR(1)$ model). See Appendix (D.6.3) for properties of the OU process, details on its discretisation, and the calibration of the $AR(1)$ model. In particular, the increment $dX_i(t)$ has unconditional mean zero and conditional mean equal to

$$E[dX_i(t)|X_i(s), s \leq t] = k_i(m_i - X_i(t))dt$$

The conditional mean, or forecast of expected daily returns, is positive or negative according to the sign of $(m_i - X_i(t))$. In general, assuming that the model parameters vary slowly with respect to the Brownian motion, we estimate the statistics for the residual process on a window of length 60 days, letting the model parameters be constant over the window. This hypothesis is tested for each stock in the universe, by goodness-of-fit of the model and, in particular, by analysing the speed of mean-reversion.

As described in Appendix (D.6.3), if we assume that the parameters of the model are constant, we get the solution

$$X_i(t_0 + \Delta t) = e^{-k_i \Delta t} X_i(t_0) + (1 - e^{-k_i \Delta t}) m_i + \sigma_i \int_{t_0}^{t_0 + \Delta t} e^{-k_i(t_0 + \Delta t - s)} dW_i(s)$$

which is the linear regression

$$X_{n+1} = aX_n + b + \nu_{n+1} \ , \ \{\nu_n\} \text{ iid } N\big(0, \sigma^2(\frac{1 - e^{2k_i \Delta t}}{2k_i})\big)$$

where $a = e^{-k_i \Delta t}$ is the slope and $b = (1 - e^{-k_i \Delta t}) m_i$ is the intercept. Letting Δt tend to infinity, we see that equilibrium probability distribution for the process $X_i(t)$ is normal with

$$E[X_i(t)] = m_i \text{ and } Var(X_i(t)) = \frac{\sigma_i^2}{2k_i} \tag{8.2.10}$$

According to Equation (1.5.20), investment in a market-neutral long-short portfolio in which the agent is long \$1 in the stock and short β_{ik} dollars in the kth ETF has an expected 1-day return

$$\alpha_i dt + k_i(m_i - X_i(t))dt$$

The second term corresponds to the model's prediction for the return based on the position of the stationary process $X_i(t)$. It forecasts a negative return if $X_i(t)$ is sufficiently high and a positive return if $X_i(t)$ is sufficiently low. The parameter k_i is called the speed of mean-reversion and

$$\tau_i = \frac{1}{k_i}$$

represents the characteristic time-scale for mean reversion. If $k >> 1$ the stock reverts quickly to its mean and the effect of the drift is negligible. Hence, we are interested in stocks with fast mean-reversion such that $\tau_i << T_1$ where T_1 is the estimation window.

Based on this simple model, Avellaneda et al. [2008] defined several trading signals. First they considered an estimation window of 60 business days ($T_1 = \frac{60}{252} = 0.24$) incorporating at least one earnings cycle for the company and they selected stocks with mean-reversion times less than $\frac{1}{2}$ period ($k > \frac{252}{30} = 8.4$) with $\tau = 0.12$. That is, $\frac{1}{2}$ period is $\tau = \frac{T_1}{2} = \frac{30}{252}$ and $k = \frac{1}{\tau} = \frac{252}{30}$. To calibrate the model we consider the linear regression above, and deduce that

$$m_i = \frac{b}{1-a}\ , k_i = -\frac{1}{\Delta t}\log a\ , \sigma_i^2 = \frac{2k_i}{1-a^2}Var(\nu)$$

A fast mean-reversion (compared to the 60-days estimation window) requires that $k > \frac{252}{30}$, corresponding to a mean-reversion of the order of 1.5 months at most.

8.2.5.2 Pure mean-reversion

In this section we focus only on the process $X_i(t)$, neglecting the drift α_i. Given Equation (8.2.10) we know that the equilibrium volatility is

$$\sigma_{eq,i} = \frac{\sigma_i}{\sqrt{2k_i}} = \sigma_i\sqrt{\frac{\tau_i}{2}}$$

We can then define the dimensionless variable

$$s_i = \frac{X_i(t) - m_i}{\sigma_{eq,i}}$$

called the s-score. The s-score measures the distance to equilibrium of the cointegrated residual in units standard deviations, that is, how far away a given stock is from the theoretical equilibrium value associated with our model. As a result, one can define a basic trading signal based on mean-reversion as

- buy to open if $s_i < -\overline{s}_{bo}$
- sell to open if $s_i > +\overline{s}_{so}$
- close short position if $s_i < +\overline{s}_{bc}$
- close long position if $s_i > -\overline{s}_{sc}$

where the cutoff values $\overline{s}_l$ for $l = bo, so, bc, sc$ are determined empirically. Entering a trade, that is buy to open, means buying \$1 of the corresponding stock and selling β_i dollars of its sector ETF (pair trading). Similarly, in the case of using multiple factors, we buy β_{i1} dollars of ETF $\#1$, β_{i2} dollars of ETF $\#2$ up to β_{im} dollars of ETF $\#m$. The opposite trade consisting in closing a long position means selling stock and buying ETFs. Since we expressed all quantities in dimensionless variables, we expect the cutoffs $\overline{s}_l$ to be valid across the different stocks. Based on simulating strategies from 2000 to 2004 in the case of ETF factors, Avellaneda et al. [2008] found that a good choice of cutoffs was $\overline{s}_{bo} = \overline{s}_{so} = 1.25$, $\overline{s}_{bc} = 0.75$, and $\overline{s}_{sc} = 0.5$. The rationale for opening trades only when the s-score s_i is far from equilibrium is to trade only when we think that we detected an anomalous excursion of the co-integration residual. Closing trades when the s-score is near zero also makes sense, since we expect most stocks to be near equilibrium most of the time. The trading rule detects stocks with large excursions and trades assuming these excursions will revert to the mean in a period of the order of the mean-reversion time τ_i.

8.2.5.3 Mean-reversion with drift

When ignoring the presence of the drift, we implicitly assume that the effect of the drift is irrelevant in comparison with mean-reversion. Incorporating the drift, the conditional expectation of the residual return over a period of time Δt becomes

$$\alpha_i dt + k_i(m_i - X_i)dt = k_i\big(\frac{\alpha_i}{k_i} + m_i - X_i\big)dt = k_i\big(\frac{\alpha_i}{k_i} - \sigma_{eq,i}s_i\big)dt$$

This suggests that the dimensionless decision variable is the modified s-score

$$s_{mod,i} = s_i - \frac{\alpha_i}{k_i \sigma_{eq,i}} = s_i - \frac{\alpha_i \tau_i}{\sigma_{eq,i}}$$

In the previous framework, we short stock if the s-score is large enough. The modified s-score is larger if α_i is negative, and smaller if α_i is positive. Therefore, it will be harder to generate a short signal if we think that the residual has an upward drift and easier to short if we think that the residual has a downward drift. Since the drift can be interpreted as the slope of a 60-day moving average, we have therefore a built-in momentum strategy in this second signal. A calibration exercise using the training period 2000-2004 showed that the cutoffs defined in the previous strategy are also acceptable for this one. However, Avellaneda et al. [2008] found that the drift parameter had values of the order of 15 basis points and the average expected reversion time was 7 days, whereas the equilibrium volatility of residuals was on the order of 300 bps. The expected average shift for the modified s-score was of the order of $0.15\frac{7}{300} \approx 0.3$. Hence, in practice, the effect of incorporating a drift in these time-scales was minor.

8.2.6 Portfolio optimisation

Following the general portfolio valuation in Section (1.5.2), we consider $h_0, h_1, .., h_N$ to be the dollars invested in different stocks (long or short) and $S_0, S_1, .., S_N$ to be the dividend-adjusted prices. Neglecting transaction costs, the change in portfolio returns becomes

$$dV_t = \sum_{i=1}^{N} h_i \frac{dS_i(t)}{S_i(t)} - (\sum_{i=1}^{N} h_i) r dt + V_t r dt$$

where $R_i(t) = \frac{dS_i(t)}{S_i(t)}$ is the expected return on the ith risky security. Then, given the evolution of stock prices in Equation (1.5.20), the change in portfolio becomes

$$dV_t = \sum_{i=1}^{N} h_i \big(\sum_{k=1}^{m} \beta_{ik} \frac{dP_k}{P_k} + dX_i\big) - (\sum_{i=1}^{N} h_i) r dt + V_t r dt$$

which becomes

$$dV_t = \sum_{i=1}^{N} h_i dX_i + \sum_{k=1}^{m} \big(\sum_{i=1}^{N} h_i \beta_{ik}\big) \frac{dP_k}{P_k} - (\sum_{i=1}^{N} h_i) r dt + V_t r dt$$

where $\sum_{i=1}^{N} h_i \beta_{ik}$ is net dollar-beta exposure along factor k and $\sum_{i=1}^{N} h_i$ the net dollar exposure of the portfolio. Cancelling the net dollar-beta exposure $\sum_{i=1}^{N} h_i \beta_{ik}$ for each factor, the change in portfolio becomes

$$dV_t = \sum_{i=1}^{N} h_i dX_i - (\sum_{i=1}^{N} h_i) r dt + V_t r dt$$

Thus, a market-neutral portfolio is affected only by idiosyncratic returns. Replacing with the residual process, we get

$$dV_t = \sum_{i=1}^{N} h_i \big(k_i(m - X_i) dt + \sigma_i dW_i\big) - (\sum_{i=1}^{N} h_i) r dt + V_t r dt$$

which gives

$$dV_t = \sum_{i=1}^{N} h_i \big(k_i(m - X_i) - r\big) dt + \sum_{i=1}^{N} h_i \sigma_i dW_i + V_t r dt$$

Ignoring the term $V_t r dt$, and taking the conditional expectation, we get

$$E[dV_t|X] = \sum_{i=1}^{N} h_i\big(k_i(m - X_i) - r\big)dt = \sum_{i=1}^{N} h_i\mu_i dt$$

where $\mu_i = k_i(m - X_i) - r$, and the conditional variance

$$Var(dV_t|X) = \sum_{i=1}^{N} h_i^2\sigma_i^2 dt$$

Following the mean-variance approach detailed in Section (2.2.1), we can therefore build the Mean-Variance optimal portfolio. We consider the mean-variance utility function given in Equation (2.2.2) or (9.2.1) where where the expected return of the portfolio is $r_P = \sum_{i=1}^{N} h_i\mu_i$ and the variance of the portfolio's return is $\sigma_P^2 = \sum_{i=1}^{N} h_i^2\sigma_i^2$. In that setting, the optimisation problem is given by

$$\max_h\big(\sum_{i=1}^{N} h_i\mu_i - \frac{1}{2\tau}\sum_{i=1}^{N} h_i^2\sigma_i^2\big)$$

where τ is the investor's risk tolerance. The optimal risky portfolio must satisfy Equation (9.2.4). However, in the case of a beta-neutral portfolio, the constraint in Equation (9.2.1) must be satisfied. In our setting, the optimal weight becomes

$$h_i = \tau\frac{\mu_i}{\sigma_i^2}$$

Replacing the optimal weight h_i in the change of portfolio, we get

$$dV_t = \tau\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\big(k_i(m - X_i) - r\big)^2 dt + \lambda\sum_{i=1}^{N}\frac{(k_i(m - X_i) - r)}{\sigma_i}dW_i$$

Setting $\xi_i = \frac{(m-X_i)}{\sigma_i}\sqrt{2k_i}$ and $r = 0$ we get

$$dV_t = \lambda\sum_{i=1}^{N}\frac{k_i}{2}\xi_i^2 dt + \tau\sum_{i=1}^{N}\sqrt{\frac{k_i}{2}}\xi_i dW_i$$

As a result, we get the norms

$$< dV_t >= \frac{\tau N}{2}\big(\frac{\sum_{i=1}^{N} k_i}{N}\big)dt$$
$$< (dV_t)^2 > - < dV_t >^2 = \frac{\tau^2 N}{2}\big(\frac{\sum_{i=1}^{N} k_i}{N}\big)dt$$

and the annualised sharpe ratio becomes

$$M = \frac{\frac{\tau N}{2}\big(\frac{\sum_{i=1}^{N} k_i}{N}\big)}{\sqrt{\frac{\tau^2 N}{2}\big(\frac{\sum_{i=1}^{N} k_i}{N}\big)}} = \sqrt{\frac{N}{2}}\sqrt{\big(\frac{\sum_{i=1}^{N} k_i}{N}\big)} = \sqrt{\frac{N\bar{k}}{2}}$$

since $\frac{a}{\sqrt{a}} = \sqrt{a}$.

8.2.7 Back-testing

The back-testing experiments consisted in running the signals through historical data, with the estimation of parameters (betas, residuals), signal evaluations and portfolio re-balancing performed daily. That is, we assumed that all trades are done at the closing price of that day. Further, we assume a round-trip transaction cost per trade of 10 basis points, to incorporate an estimate of price slippage and other costs as a single friction coefficient. Given the portfolio dynamics in Equation (1.5.17) where V_t is the portfolio equity at time t, the basic PnL equation for the strategy has the following form

$$\begin{aligned} V_{t+\Delta t} &= V_t + r\Delta t V_t + \sum_{i=1}^{N} h_{i,t}R_{i,t} - r\Delta t(\sum_{i=1}^{N} h_{i,t}) + \sum_{i=1}^{N} h_{i,t}\frac{D_{i,t}}{S_{i,t}} - \epsilon\sum_{i=1}^{N}|h_{i,t+\Delta t} - h_{i,t}| \\ h_{i,t} &= V_t\overline{\Lambda}_t \end{aligned}$$

where $R_{i,t}$ is the stock return on the period $(t, t+\Delta t)$, r represents the interest rate (assuming, for simplicity, no spread between long and short rates), $\Delta t = \frac{1}{252}$, $D_{i,t}$ is the dividend payable to holders of stock i over the period $(t, t+\Delta t)$, $S_{i,t}$ is the price of stock i at time t, and $\epsilon = 0.0005$ is the slippage term alluded to above. At last, $h_{i,t}$ is the dollar investment in stock i at time t which is proportional to the total equity in the portfolio. The proportionality factor $\overline{\Lambda}_t$ is stock-independent and chosen so that the portfolio has a desired level of leverage on average. As a result,

$$\sum_{i=1}^{N}|h_{i,t}| = NE_t\overline{\Lambda}_t$$

Given the definition of the leverage ratio in Equation (8.2.9) we get

$$\Lambda_t = N\overline{\Lambda}_t$$

so that $\overline{\Lambda}_t = \frac{\Lambda_t}{N}$. That is, the weights $\overline{\Lambda}_t$ are uniformly distributed. For example, given $N = 200$, if we have 100 stocks long and 100 short and we wish to have a $(2+2)$ leverage ($\$2$ long and $\$2$ short for $\$1$ of capital), then $\overline{\Lambda}_t = \frac{2}{100}$ and we get $\sum_{i=1}^{N}|Q_{i,t}| = E_t\frac{2}{100}200 = 4E_t$.

***Remark* 8.2.2** *In practice this number is adjusted only for new positions, so as not to incur transaction costs for stock which are already held in the portfolio.*

Hence, it controls the maximum fraction of the equity that can be invested in any stock, and we take this bound to be equal for all stocks.

Given the discrete nature of the signals, the strategy is such that there is no continuous trading. Instead, the full amount is invested on the stock once the signal is active (buy-to-open, short-to-open) and the position is unwound when the s-score indicates a closing signal. This all-or-nothing strategy, which might seem inefficient at first glance, turns out to outperform making continuous portfolio adjustments.

8.3 The meta strategies

8.3.1 Presentation

8.3.1.1 The trading signal

Given the time-series momentum strategy described in Section (8.1), we consider the return $Y^i_{J,K}(t)$ at time t for the series of the ith available individual strategy which is given by Equation (8.1.1). We are now going to consider several possible trading signal $\psi_i(.,.)$ defined in Section (8.1) and characterising the strategies based on the returns of the

underlying process for the period $[t-J,t]$. Contrary to the time-series momentum strategy described in Section (8.1), we do not let the Return Sign or Moving Average be the trading signal $\psi_i(.,t)$ and do not directly follow the strategy characterised with the above return. Instead, we first perform some risk analysis of that strategy based on different risk measures. That is, we no-longer consider the time-series of the stock returns, but instead, we consider the time-series of some associated risk measures and use it to infer some trading signals.

8.3.1.2 The strategies

Return sign Following the return sign strategy discussed in Section (8.1.2.1), we consider the random vector Y_0^i characterising the benchmark strategy of the ith asset where we always follow the previous move of the underlying price returns. For risk management purposes, we set $J=1$ and $K=1$ in Equation (8.1.1) getting the return

$$Y_{1,1}^i(j\delta) = sign_i((j-1)\delta, j\delta)R_i(j\delta,(j+1)\delta) = sign(R_i((j-1)\delta,j\delta))R_i(j\delta,(j+1)\delta) = Y_0^i(j\delta)$$

where the random variables $Y_0^i(j\delta)$ take values in $\mathbb{R}$.

Moving average Following the moving average strategy discussed in Section (8.1.2.2), we consider the random vector Y_0^i characterising the strategy of the ith asset where a long (short) position is determined by a lagging moving average of a price series lying below (above) a past leading moving average. For J lookback periods and K holding periods, the return in Equation (8.1.1) becomes

$$Y_{J,K}^i(j\delta) = MA_i((j-J)\delta, j\delta)R_i(j\delta,(j+K)\delta) = Y_0^i(j\delta)$$

where the random variables $Y_0^i(j\delta)$ take values in $\mathbb{R}$.

8.3.2 The risk measures

8.3.2.1 Conditional expectations

Focusing on the ith asset, we then let $X^i(j\delta)$ for $i=1,..,N$ be a random variable taking on only countably many values and possibly correlated with the random variable $Y_0^i(j\delta)$ defined in Section (14.3.5). One of the main advantage of the theory of conditional expectation (described in Appendix (B.6.1)) is that if we already know the value of $X^i(j\delta)$ we can use this information to calculate the expected value of $Y_0^i(j\delta)$ taking into account the knowledge of $X^i(j\delta)$. That is, suppose we know that the event $\{X^i(j\delta)=k\}$ for some value k has occurred, then the expectation of $Y_0^i(j\delta)$ may change given this knowledge. As a result, the conditional expectation of $Y_0^i(j\delta)$ given the event $\{X^i(j\delta)=k\}$ is defined to be

$$E[Y_0^i(j\delta)|X^i(j\delta)=k] = E^Q[Y_0^i(j\delta)]$$

where Q is the probability given by $Q(\Lambda)=P(\Lambda|X^i(j\delta)=k)$. Further, if the r.v. $Y_0^i(j\delta)$ is countably valued then the conditional expectation becomes

$$E[Y_0^i(j\delta)|X^i(j\delta)=k] = \sum_{l=1}^{\infty} y_0^i(l)P\big(Y_0^i(j\delta)=y_0^i(l)|X^i(j\delta)=k\big)$$

Note, if we denote B_k the event $\{X^i(j\delta)=k\}$, we make sure that the family $B_1,B_2,..,B_n$ is a partition of the sample space Ω (see Appendix (B.6.1)). As a result, we can express the conditional expectation as

$$E[Y_0^i(j\delta)|B_k] = \frac{E[Y_0^i(j\delta)I_{B_k}]}{P(B_k)}$$

For practicality we define the random variable

$$Y_k^i(j\delta) = Y_0^i(j\delta)I_{\{B_k\}}\ , k = 1, 2, ..., n \tag{8.3.11}$$

Using Equation (B.6.2) we can express the expectation of the random variable $Y_0(j\delta)$ characterising the original strategy as a weighted sum of conditional expectation

$$E[Y_0^i(j\delta)] = \sum_{k=1}^{n} E[Y_0^i(j\delta)|B_k]P(B_k) = \sum_{k=1}^{n} E[Y_0^i(j\delta)I_{B_k}] = \sum_{k=1}^{n} E[Y_k^i(j\delta)]$$

8.3.2.2 Some examples

Example 1 For instance, we can define the random variable $X^i(j\delta)$ to characterise the strategy consisting in following the previous move of returns only when that return is either positive or equal to zero, or when that return is negative. That is, the random variable $X^i(j\delta) = sign(R_i((j-1)\delta, j\delta))$ has only two events (or two points), and the state space is given by

$$\Omega = \{R_i((j-1)\delta, j\delta) \geq 0, R_i((j-1)\delta, j\delta) < 0\}$$

with $B_1 = R_i((j-1)\delta, j\delta) \geq 0$ and $B_2 = R_i((j-1)\delta, j\delta) < 0$, so that

$$Y_1^i(j\delta) + Y_2^i(j\delta) = R_t^i\big(I_{\{R_i((j-1)\delta,j\delta)\geq 0\}} - I_{\{R_i((j-1)\delta,j\delta)<0\}}\big) = Y_0^i(j\delta)$$

Since $I_{\{R_i((j-1)\delta,j\delta)\geq 0\}} + I_{\{R_i((j-1)\delta,j\delta)<0\}} = 1$, we get

$$Y_1^i(j\delta) + Y_2^i(j\delta) = R^i(j\delta)\big(2I_{\{R_i((j-1)\delta,j\delta)\geq 0\}} - 1\big) = Y_0^i(j\delta)$$

Example 2 Similarly to the previous example, we can define the random variable $X^i(j\delta)$ to characterise the strategy consisting in following the product of the two previous move of returns only when that product is either positive or equal to zero, or either when that product is negative. That is, the random variable

$$X^i(j\delta) = sign(R_i((j-2)\delta, (j-1)\delta)R_i((j-1)\delta, j\delta))$$

has only two events (or two points), and the state space is given by

$$\Omega = \{R_i((j-2)\delta, (j-1)\delta)R_i((j-1)\delta, j\delta) \geq 0, R_i((j-2)\delta, (j-1)\delta)R_i((j-1)\delta, j\delta)\delta) < 0\}$$

with $B_1 = R_i((j-2)\delta, (j-1)\delta)R_i((j-1)\delta, j\delta) \geq 0$ and $B_2 = R_i((j-2)\delta, (j-1)\delta)R_i((j-1)\delta, j\delta) < 0$, so that

$$Y_1^i(j\delta) + Y_2^i(j\delta) = R_t^i\big(I_{\{R_i((j-2)\delta,(j-1)\delta)R_i((j-1)\delta,j\delta)\geq 0\}} - I_{\{R_i((j-2)\delta,(j-1)\delta)R_i((j-1)\delta,j\delta)<0\}}\big) = Y_0^i(j\delta)$$

Example 3 In this example, we define the random variable $X^i(j\delta)$ to characterise the strategy consisting in following the previous move of returns only when that return is either big, or when that return is small. That is, the random variable $X^i(j\delta) = |R_i((j-1)\delta, j\delta)| \geq c_B$ has only two events (or two points), and the state space is given by

$$\Omega = \{|R_i((j-1)\delta, j\delta)| \geq c_B, |R_i((j-1)\delta, j\delta)| < c_B\}$$

with $B_1 = |R_i((j-1)\delta, j\delta)| \geq c_B$ and $B_2 = |R_i((j-1)\delta, j\delta)| < c_B$, so that

$$Y_1^i(j\delta) + Y_2^i(j\delta) = R_t^i\big(I_{\{|R_i((j-1)\delta,j\delta)|\geq c_B\}} - I_{\{|R_i((j-1)\delta,j\delta)|<c_B\}}\big) = Y_0^i(j\delta)$$

8.3.3 Computing the Sharpe ratio of the strategies

As discussed in Appendix (D) a discrete time series is a set of observations x_t, each one being recorded at a specified fixed time interval. One can assume that each observation x_t is a realised value of a certain random variable X_t. That is, the time series $\{x_t\ , t \in T_0\}$ is a realisation of the family of random variables $\{X_t\ , t \in T_0\}$. Hence, we can model the data as a realisation of a stochastic process $\{X_t\ , t \in T\}$ where $T \supseteq T_0$. Given the definition of the Sharp ratio in Section (2.4.2) and the strategies in Section (8.3.1.2), we can estimate that measure of risk-return for the process $Y_0(t)$ given by

$$M(Y_0(t)) = \frac{E[Y_0(t)]}{\sigma_{Y_0}}$$

where $\sigma_{Y_0} = \sqrt{Var(Y_0(t))}$. Note, in our setting $t = j\delta$ with time interval δ being one day or one week. From the observations $\{y_1, y_2, .., y_n\}$ of the stationary time series $Y_0(t)$ we can calculate the ex-post Sharpe ratio by computing the arithmetic mean and the arithmetic standard deviation.

In the example where the random variable, $X(t) = sign(R_{t-J})$ or $X(t) = MA(t-J,t)$ for J one day or one week, has only two events, then the events B_1 and B_2 form a partition of the state space, and we get the decomposition $Y_0(t) = Y_1(t) + Y_2(t)$. As a result, we get the marginal expectation $E[Y_0(t)] = E[Y_1(t)] + E[Y_2(t)]$ and the marginal variance is $Var(Y_0(t)) = Var(Y_1(t) + Y_2(t))$. Hence, the Sharpe ratio for the random variable $Y_0(t)$ can be decomposed as

$$M(Y_0(t)) = \frac{E[Y_1(t)]}{\sqrt{Var(Y_1(t) + Y_2(t))}} + \frac{E[Y_2(t)]}{\sqrt{Var(Y_1(t) + Y_2(t))}}$$

Since $Var(Y_1(t) + Y_2(t)) > Var(Y_i(t))$ for $i = 1, 2$, when $E[Y_i(t)] > 0$ for $i = 1, 2$ we get

$$M(Y_0(t)) < \frac{E[Y_1(t)]}{\sqrt{Var(Y_1(t))}} + \frac{E[Y_2(t)]}{\sqrt{Var(Y_2(t))}} = M(Y_1) + M(Y_2)$$

and when $E[Y_i(t)] < 0$ for $i = 1, 2$ we get

$$M(Y_0(t)) > \frac{E[Y_1(t)]}{\sqrt{Var(Y_1(t))}} + \frac{E[Y_2(t)]}{\sqrt{Var(Y_2(t))}} = M(Y_1) + M(Y_2)$$

while for $E[Y_i(t)] > 0$ and $E[Y_j(t)] < 0$ for $i = 1, 2$ and for $j = 1, 2$ with $i \neq j$, if $|E[Y_i(t)]| > |E[Y_j(t)]|$ we get

$$M(Y_0(t)) < \frac{E[Y_i(t)]}{\sqrt{Var(Y_i(t))}} + \frac{E[Y_j(t)]}{\sqrt{Var(Y_j(t))}} = M(Y_i) + M(Y_j)$$

Given the definition of the conditional expectation and conditional variance in Appendix (B.6.1), we can consider the Information ratio or conditional Sharpe ratio

$$M(Y_0(t)|B_k) = \frac{E[Y_0(t)I_{\{B_k\}}]/P(B_k)}{\sigma_{Y_0|B_k}} = \frac{E[Y_0(t)|B_k]}{\sigma_{Y_0|B_k}} \tag{8.3.12}$$

where $\sigma_{Y_0|B_k} = \sqrt{Var(Y_0(t)|B_k)}$. Further, given the Definition (B.6.2) of the conditional expectation, we can rewrite the conditional Sharpe ratio as

$$M(Y_0(t)|B_k) = \frac{E^{Q^k}[Y_0(t)]}{\sigma_{Y_0|B_k}}$$

where Q^k is the probability given by $Q^k(\Lambda) = P(\Lambda|X = k)$.

In order to select the appropriate strategy at time t, we need to consider the measure for which the event $B_k(t)$ for $k = 1, 2, ..$ is satisfied. Calling this event $B_k^*(t)$ and its associated measure $M^*(Y_0|B_k)$, if $M^*(Y_0|B_k) > \alpha_k$ we follow the strategy $Y_k(t)$ while if $M^*(X_0|B_k) < \beta_k$ we follow the strategy $-Y_k(t)$ where α_k is a positive constant and β_k is a negative constant.

8.4 Random sampling measures of risk

We assume that the population is of size N and that associated with each member of the population is a numerical value of interest denoted by $x_1, x_2, .., x_N$. We take a sample with replacement of n values $X_1, ..., X_n$ from the population, where $n < N$ and such that X_i is a random variable. That is, X_i is the value of the ith member of the sample, and x_i is that of the ith member of the population. The population moments and the sample moments are given in Appendix (B.10.1).

8.4.1 The sample Sharpe ratio

From an investor's perspective, volatility per se is not a bad feature of a trading strategy. In fact, increases in volatility generated by positive returns are desired. Instead, it is only the part of volatility that is generated by negative returns that is clearly unwanted. There exists different methodologies in describing what is called the downside risk of an investment. Sortino and Van Der Meer [1991] suggested the use of Sortino ratio as a performance evaluation metric in place of the ordinary Sharpe ratio. The Sharpe ratio treats equally positive and negative returns

$$M = \frac{\mu - R_f}{\sigma}$$

where the mean μ is estimated with the sample mean $\overline{X}$, and the variance σ^2 is estimated with the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Whereas the former normalises the average excess returns with the square root of the semi-variance of returns (variance generated by negative returns)

$$M_S = \frac{\mu - R_f}{\sigma^-}$$

where the semi-variance $(\sigma^-)^2$ is estimated with the sample semi-variance

$$(S^-)^2 = \frac{1}{n^- - 1}\sum_{i=1}^{n}(X_i I_{\{X_i<0\}})^2$$

with n^- being the number of periods with a negative return. It is therefore expected that the Sortino ratio will be relatively larger than the ordinary Sharpe ratio for positively skewed distributions

8.4.2 The sample conditional Sharpe ratio

Given the definition of the conditional Sharpe ratio in Equation (8.3.12), we let the probability of the event be $P(B_k) = \frac{n_k}{n}$ where n_k is the number of count in the partition B_k and define the conditional sample mean as

$$\overline{X}_{B_k} = \frac{n}{n_k}\frac{1}{n}\sum_{i=1}^{n} X_i I_{\{B_k\}} = \frac{1}{n_k}\sum_{i=1}^{n} X_i I_{\{B_k\}}$$

which corresponds to the sample mean of the strata k. Given the definition of the local averaging estimates in Equation (9.3.19) we get $\overline{\omega}_{n,i} = I_{\{B_k\}}$ with $norm = n_k$ and we recover the partitioning estimate in Equation (9.3.20)

$$\overline{X}_{B_k} = \frac{1}{norm} \sum_{i=1}^{n} \overline{\omega}_{n,i} X_i$$

Given the definition of the conditional variance in Equation (B.6.3) we get the conditional sample variance as

$$S^2_{B_k} = \frac{n}{n_k} \frac{1}{n} \sum_{i=1}^{n} X_i^2 I_{\{B_k\}} - (\overline{X}_{B_k})^2 = \frac{1}{n_k} \sum_{i=1}^{n} X_i^2 I_{\{B_k\}} - (\overline{X}_{B_k})^2$$

Putting terms together, the sample conditional Sharpe ratio is

$$M(Y_0(t)|B_k) = \frac{\overline{X}_{B_k}}{S_{B_k}}$$

Chapter 9

Portfolio management under constraints

9.1 Introduction

We discussed in Section (1.7.6) the existence of excess returns and showed that the market is inefficient, that is, substantial gain is achievable by rebalancing and predicting market returns based on market's history. As a consequence, we showed in Section (2.1.1) that market inefficiency leads to active equity management where enhanced indexed portfolios were designed to generate attractive risk-adjusted returns through active weights giving rise to active returns. Earlier results in the non-parametric statistics, information theory and economics literature (such as Kelly [1956], Markowitz [1952]) established optimality criterion for long-only, non-leveraged investment. We saw in Section (2.3) that in view of solving the problem of portfolio selection Markowitz [1952] introduced the mean-variance approach which is a simple trade-off between return and uncertainty, where one is left with the choice of one free parameter, the amount of variance acceptable to the individual investor. Similarly, Kelly [1956] introduced an investment theory based on growth by using the role of time in multiplicative processes to solve the problem of portfolio selection. We presented in Section (2.3.1.2) the capital asset pricing model (CAPM) and we introduced in Section (2.3.2) the growth optimal portfolio (POP) as a portfolio having maximal expected growth rate over any time horizon. Going deeper in the analysis, we are going to describe in Section (9.3.16) the growth optimum theory (GOT) as an alternative to the expected utility theory and the mean-variance approaches to asset pricing.

We saw in Section (7) that in view of taking advantage of market excess returns, a large number of hedge funds flourished, using complex valuation models to predict market returns. Specialised quantitative strategies developed along with specific prime brokerage structures. We defined leverage in Section (7.1.2) as any means of increasing expected return or value without increasing out-of-pocket investment. We are now going to introduce portfolio construction in presence of financial leverage and construction leverage such as short selling which is the process of borrowing assets and selling them immediately, with the obligation to rebuy them later. Portfolio optimisation in the long-short context does not differ much from optimisation in the long-only context. Jacobs et al. [1999] showed how short positions could be added to a long portfolio, by removing the greater than zero constraint from the model. In order to optimise a true long-short portfolio, constraints should be added to the model in order to ensure equal (in terms of total exposure) long and short legs. It was shown in Section (2.3.2) that the optimal asymptotic growth rate on non-leveraged, long-only memoryless market (independent identically distributed, i.i.d.) coincides with that of the best constantly rebalanced portfolio (BCRP). In the special case of memoryless assumption on returns, adding leverage through margin buying and short selling, Horvath et al. [2011] derived optimality conditions and generalised the BCRP by establishing no-ruin conditions.

9.2 Robust portfolio allocation

There are some different methods of portfolio optimisation available. We described in Section (2.2.1) the classic Mean-Variance portfolio based on algorithms working with point-estimates of expected returns, variances and covariances.

9.2.1 Long-short mean-variance approach under constraints

Considering proper portfolio optimisation, Jacobs et al. [1999] had a rigorous look at long-short optimality and called into question the goals of dollar, and beta neutrality which is common practice in traditional long-short management. Following their approach we consider the mean-variance [1] utility function (see details in Appendix (A.7.2))

$$U = r_P - \frac{1}{2}\frac{\sigma_P^2}{\tau} \tag{9.2.1}$$

where r_P is the expected return of the portfolio during the investor's horizon, σ_P^2 is the variance of the portfolio's return, and τ is the investor's risk tolerance. We need to choose how to allocate the investor's wealth between a risk-free security and a set of N securities, and we need to choose how to distribute wealth among the N risky securities. We let h_R be the fraction of wealth allocated to the risky portfolio (total wealth is 1), and we let h_i be the fraction of wealth invested in the ith risky security. The three components of capital earning interest at the risk-free rate are

- the wealth allocated to the risk-free security with magnitude of $1 - h_R$
- the balance of the deposit made with the broker after paying for the purchase of shares long with magnitude $h_R - \sum_{i\in L} h_i$ where L is the set of securities held long
- the proceeds of the short sales with magnitude of $\sum_{i\in S} |h_i| = -\sum_{i\in S} h_i$ where S is the set of securities sold short

Since $|x| = x$ for a positive x and $|x| = -x$ for a negative x then h_i is negative for $i \in S$. Summing these three components gives the total amount of capital h_F earning interest at the risk-free rate

$$h_F = 1 - \sum_{i=1}^{N} h_i \tag{9.2.2}$$

which is independent of h_R. We can then make the following observations:

- In case of long-only management where everything is invested in risky assets, the portfolio satisfies $\sum_{i=1}^{N} h_i = 1$ and we get $h_F = 0$,
- while in case of short-only management in which $\sum_{i=1}^{N} h_i = -1$, the quantity h_F is equal to 2 and the investor earns the risk-free rate twice.
- In the case of a dollar-balanced long-short management in which $\sum_{i=1}^{N} h_i = 0$, the investor earns the risk-free rate only once.

Note, if we do not allocate wealth to the risk-free security, the fraction of wealth allocated to the risky portfolio is $h_R = 1$, and the total amount of capital h_F earning interest at the risk-free rate becomes

$$h_F = h_R - \sum_{i=1}^{N} h_i$$

[1] It is only a single-period formulation which is not sensitive to investor wealth.

We now let r_F be the return on the risk-free security, and R_i be the expected return on the ith risky security. As explained in Section (9.2.2.2), in the case of short-selling, when the price of an asset drops, the returns becomes $-R_i$. Since $h_i < 0$, when asset price rises we get $R_i > 0$ and we loose money, and when asset price drops we get $R_i < 0$ and we make money. Putting long and short returns together, the expected return on the investor's total portfolio is

$$r_P = h_F r_F + \sum_{i=1}^{N} h_i R_i$$

We substitute h_F into this equation, the total portfolio return is the sum of a risk-free return and a risky return component

$$r_P = (1 - \sum_{i=1}^{N} h_i) r_F + \sum_{i=1}^{N} h_i R_i = r_F + r_R \tag{9.2.3}$$

where the risky return component is

$$r_R = \sum_{i=1}^{N} h_i r_i$$

where $r_i = R_i - r_F$ is the expected return on the ith risky security in excess of the risk-free rate. It can be expressed in matrix notation as

$$r_R = h^\top r$$

where $h = [h_1, .., h_N]^\top$ and $r = [r_1, .., r_N]^\top$. Similarly to the long-only portfolio in Section (2.2.1), given r_R, the variance of the risky component is

$$\sigma_R^2 = h^\top Q h$$

where Q is the covariance matrix of the risky securities' returns. It is also the variance of the entire portfolio $\sigma_P^2 = \sigma_R^2$. Using these expressions, the utility function in Equation (9.2.1) can be rewritten in terms of controllable variables. We determine the optimal portfolio by maximising the utility function through appropriate choice of these variables under constraints. For instance, the requirement that all the wealth allocated to the risky securities is fully utilised. Some of the most common portfolio constraints are

- beta-neutrality constraint $\sum_{i=1}^{N} h_i \beta_i = 0$
- portfolio constraint $\sum_{i=1}^{N} h_i = 1$ for long-only portfolio and $\sum_{i=1}^{N} h_i = -1$ for short-only portfolio
- leverage constraint $\sum_{i=1}^{N} |h_i| = l$ where l is the leverage or we can set $\sum_{i \in L} |h_i| = l_L$ and $\sum_{i \in S} |h_i| = l_S$ with $l = l_L + l_S$

Most strategies will be dollar and beta neutral, but fewer will be sector/industry, capitalisation and factor neutral. Arguably, the more neutral a long-short portfolio the better, as systematic risk diminishes (as does the residual return correlation of the long and short portfolios) and stock specific risk which is the object of long-short, increases. The solution (provided Q is non-singular) gives the optimal risky portfolio

$$h = \tau Q^{-1} r \tag{9.2.4}$$

corresponding to the minimally constrained portfolio. The expected returns and their covariances must be quantities that the investor expect to be realised over the portfolio's holding period.

***Remark* 9.2.1** *The true statistical distribution of returns being unknown, it leads to different results based on different assumptions. Optimal portfolio holdings will thus differ for each investor, even though investors use the same utility function.*

The optimal holdings in Equation (9.2.4) define a portfolio allowing for short positions because no non-negativity constraints are imposed. The single portfolio exploits the characteristics of individual securities in a single integrated optimisation even though the portfolio can be partitioned artificially into one sub-portfolio of long stocks and the other one of stocks sold short (there is no benefit in doing so). Further, the holdings need not satisfy any arbitrary balance conditions, that is dollar or beta neutrality is not required. The portfolio has no inherent benchmark so that there is no residual risk. The portfolio will exhibit an absolute return and an absolute variance of return. This return is calculated as the weighted spread between the returns to the securities held long and the ones sold short.

Performance attribution can not distinguish between the contribution of the stocks held long and those sold short.

***Remark* 9.2.2** *Separate long and short alpha (and their correlation) are meaningless.*

Long-short portfolio allows investors to be insensitive to chosen exogenous factors such as the return of the equity market. This is done by constructing a portfolio so that the beta of the short positions equals and offsets the beta of the long position, or (more problematically) the dollar amount of securities sold short equals the dollar amount of securities held long. However, market neutrality may exact costs in terms of forgone utility. This is the case if more opportunities exist on the short side than on the long side of the market. One might expect some return sacrifice from a portfolio that is required to hold equal-dollar or equal-beta positions long and shorts. Market neutrality can be achieved by using the appropriate amount of stock index futures, without requiring that long and short security positions be balanced. Nevertheless, investors may prefer long-short balances for mental accounting reasons. Making the portfolio insensitive to equity market return (or to any other factor) constitutes an additional constraint on the portfolio. The optimal neutral portfolio maximise the investor's utility subject to all constraints, including neutrality. However, the optimal solution is no-longer given by Equation (9.2.4).

Definition 9.2.1 *By definition, the risky portfolio is dollar-neutral if the net holding H of risky securities is zero, meaning that*

$$H = \sum_{i=1}^{N} h_i = 0 \tag{9.2.5}$$

This condition is independent of h_R. Applying the condition in Equation (9.2.5) to the optimal weights in Equation (9.2.4) it can be shown that the dollar-neutral portfolio is equal to the minimally constrained optimal portfolio when

$$H \sim \sum_{i=1}^{N} (\xi_i - \overline{\xi}) \frac{r_i}{\sigma_i} = 0$$

where $\xi_i = \frac{1}{\sigma_i}$ is a measure of stability of the return of the stock i and $\overline{\xi}$ is the average return stability of all stocks in the investor's universe. The term $\frac{r_i}{\sigma_i}$ is a risk adjusted return, and $(\xi_i - \overline{\xi})$ can be seen as an excess stability. Highly volatile stocks will have low stabilities and their excess stability will be negative. If the above quantity is positive, the net holding should be long, and if it is negative it should be short.

Once the investor has chosen a benchmark, each security can be modelled in terms of its expected excess return α_i and its beta β_i with respect to that benchmark. If r_B is the expected return of the benchmark, then the expected return of the ith security is

$$r_i = \alpha_i + \beta_i r_B$$

Similarly, the expected return of the portfolio can be modelled in terms of its expected excess return α_P and beta β_P with respect to the benchmark

$$r_P = \alpha_P + \beta_P r_B \tag{9.2.6}$$

where the beta of the portfolio is expressed as a linear combination of the betas of the individual securities

$$\beta_P = \sum_{i=1}^{N} h_i \beta_i$$

This is obtained by replacing the expected return r_i of the ith security in the portfolio return r_P given in Equation (9.2.3). From Equation (9.2.6) it is clear that any portfolio that is insensitive to changes in the expected benchmark return must satisfy the condition

$$\beta_P = 0$$

Applying this condition to the optimal weights in Equation (9.2.4), together with the model for the expected return of the ith security above, it can be shown that the beta-neutral portfolio is equal to the optimal minimally constrained portfolio when

$$\sum_{i=1}^{N} \beta_i \phi_i = 0$$

with weights

$$\phi_i = \frac{r_i}{\sigma_{e,i}^2} \tag{9.2.7}$$

where r_i is the excess return and $\sigma_{e,i}^2$ is the variance of the excess return of the ith security. It is the portfolio net beta-weighted risk-adjusted expected return. When this condition is satisfied the constructed portfolio is unaffected by the return of the chosen benchmark (it is beta-neutral).

9.2.2 Portfolio selection

As an example of long-only, non-leveraged active equity management, the Constantly Rebalanced Portfolio (CRP) is a self-financing portfolio strategy, rebalancing to the same proportional portfolio in each investment period. This means that the investor neither consumes from, nor deposits new cash into his account, but reinvests his capital in each trading period. Using this strategy the investor chooses a proportional portfolio vector $\pi_v = (\pi_v^1, ..., \pi_v^N)$, where π_v^i is defined in Equation (1.5.18), and rebalances his portfolio after each period to correct the price shifts in the market. The idea being that on a frictionless market the investor can rebalance his portfolio for free at each trading period. Hence, asymptotic optimisation on a memoryless market means that the growth optimal (GO) strategy will pick the same portfolio vector at each trading period, leading to CRP strategies. Thus, the one with the highest asymptotic average growth rate is referred to as Best Constantly Rebalanced portfolio (BCRP).

In a memoryless market, leverage is anticipated to have substantial merit in terms of growth rate, while short selling is not expected to yield much better results since companies worth to short in a testing period might already have defaulted. That is, in case of margin buying (the act of borrowing money and increasing market exposure) and short selling, it is easy to default on total initial investment. In this case the asymptotic growth rate becomes minus infinity. Horvath et al. [2011] showed that using leverage through margin buying yields substantially higher growth rate in the case of memoryless assumption on returns. They also established mathematical basis for short selling, that is, creating negative exposure to asset prices. Adding leverage and short selling to the framework, Horvath et al. derived optimality conditions and generalised the BCRP by establishing no-ruin conditions. They further showed that short selling might yield increased profits in case of markets with memory since market is inefficient.

9.2.2.1 Long only investment: non-leveraged

We consider a market consisting of N assets and let the evolution of prices to be represented by a sequence of price vectors $S_1, S_2, .. \in \mathbb{R}_+^N$ where

$$S_n = (S_n^1, S_n^2, ..., S_n^N)$$

S_n^i denotes the price of the ith asset at the end of the nth trading period. We transform the sequence of price vectors $\{S_n\}$ into return vectors

$$x_n = (x_n^1, x_n^2, ..., x_n^N)$$

where

$$x_n^i = \frac{S_n^i}{S_{n-1}^i} \tag{9.2.8}$$

Note, we usually denotes return as

$$R_n^i = x_n^i - 1$$

Note also that expected excess return is the expected absolute return minus the expected risk free rate. A representative example of the dynamic portfolio selection in the long-only case is the constantly rebalanced portfolio (CRP), introduced and studied by Kelly [1956], Latane [1959], and presented in Section (2.3.2). As defined in Equation (1.5.18), we let $\pi_v^i(n) = \frac{\delta^i(n)S^i(n)}{V(n)}$ be the ith component of the vector π_v representing the proportion of the investor's capital invested in the ith asset in the nth trading period.

***Remark* 9.2.3** *In that setting, the proportion of the investor's wealth invested in each asset at the beginning of trading periods is constant.*

The portfolio vector has non-negative components that sum up to 1, and the set of portfolio vectors is denoted by

$$\Delta_N = \left\{\pi_v = (\pi_v^1, ..., \pi_v^N)\,,\, \pi_v^i \geq 0\,,\, \sum_{i=1}^N \pi_v^i = 1\right\}$$

Note, in this example nothing is invested into cash (risk-free rate). Let $V(0)$ denote the investor's initial capital. At the beginning of the first trading period, $n = 1$, $V(0)\pi_v^i$ is invested into asset i, and it results in position size

$$V(1) = V(0)\pi_v^i + V(0)\pi_v^i \frac{S_1^i - S_0^i}{S_0^i} = V(0)\pi_v^i x_1^i$$

after changes in market prices, or equivalently $V(1) = V(0)\pi_v^i(1 + R_1^i)$. That is, we hold $V(0)\pi_v^i$ of asset i and we earn $V(0)\pi_v^i R_1^i$ of interest. Therefore, at the end of the first trading period the investor's wealth becomes

$$V(1) = V(0)\sum_{i=1}^N \pi_v^i x_1^i = V(0) < \pi_v, x_1 >$$

where $< .,. >$ is the inner product. Equivalently, we get

$$V(1) = V(0)\sum_{i=1}^N \pi_v^i(1 + R_1^i) = V(0) < \pi_v, (1 + R_1) >$$

For the second trading period, $n = 2$, $V(1)$ is the new initial capital, and we get

$$V(2) = V(1) < \pi_v, x_2 >= V(0) < \pi_v, x_1 >< \pi_v, x_2 >$$

Note, taking the difference between period $n = 2$ and period $n = 1$ and setting $dS_1^i = S_2^i - S_1^i$, we get

$$V(2) - V(1) = V(1) \sum_{i=1}^{N} \pi_v^i R_2^i = V(1) \sum_{i=1}^{N} \frac{\pi_v^i}{S_1^i} dS_1^i$$

since $\sum_{i=1}^{N} \pi_v^i = 1$. For $\pi_v^i(1) = \frac{\delta^i S_1^i}{V(1)}$ we recover the dynamics of the portfolio. By induction, after n trading periods, the investor's wealth becomes

$$V(n) = V(n-1) < \pi_v, x_n >= V(0) \prod_{j=1}^{n} < \pi_v, x_j >$$

Including cash account into the framework is straight forward by assuming $x_n^i = 1$ (or $R_n^i = 0$) for some i and for all n. The asymptotic average growth rate of the portfolio satisfies

$$W(\pi_v) = \lim_{n\to\infty} \frac{1}{n} \ln V(n) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \ln < \pi_v, X_j >$$

if the limit exists. If the market process $\{X_i\}$ is memoryless, (it is a sequence of independent and identically distributed (i.i.d.) random return vectors) then the asymptotic rate of growth exists almost surely (a.s.), where, with random vector X being distributed as X_i we get

$$W(\pi_v) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \ln < \pi_v, X_j >= E[\ln < \pi_v, X >] = E[\ln < \pi_v, (1+R) >] \tag{9.2.9}$$

given that $E[\ln < \pi_v, X >]$ is finite, due to strong law of large numbers. We can ensure this property by assuming finiteness of $E[\ln X^i]$, that is, $E[|\ln X^i|] < \infty$ for each $i = \{1, 2, .., N\}$. Because of $\pi_v^i > 0$ for some i, we have

$$E[\ln < \pi_v, X >] \geq E[\ln(\pi_v^i X^j)] = \ln \pi_v^i + E[\ln X^j] > -\infty$$

and because of $\pi_v^i \leq 1$ for all i, we have

$$E[\ln < \pi_v, X >] \leq \ln N + \sum_j E[\ln |X^j|] < \infty$$

From Equation (9.2.9), it follows that rebalancing according to the best log-optimal strategy

$$\pi_v^* \in arg \max_{\pi_v \in \Delta_N} E[\ln < \pi_v, X >]$$

is also an asymptotically optimal trading strategy, that is, a strategy with a.s. optimum asymptotic growth

$$W(\pi_v^*) \geq W(\pi_v)$$

for any $\pi_v \in \Delta_N$. The strategy of rebalancing according to π_v^* at the beginning of each trading period, is called best constantly rebalanced portfolio (BCRP). For details on the maximisation of the asymptotic average rate of growth see Bell et al. [1980].

9.2.2.2 Short selling: No ruin constraints

Short selling an asset is usually done by borrowing the asset under consideration and selling it. As collateral the investor has to provide securities of the same value to the lender of the shorted asset. This ensures that if anything goes wrong, the lender still has high recovery rate. While the investor has to provide collateral, after selling the assets having been borrowed, he obtains the price of the shorted asset again. This means that short selling is virtually for free

$$V^{'} = V - C + P$$

where $V^{'}$ is wealth after opening the short position, V is wealth before, C is collateral for borrowing and P is price income of selling the asset being shorted. For simplicity we assume

$$C = P \tag{9.2.10}$$

so that $V^{'} = V$ and short selling is free. Note, in the case of naked short transaction, selling an asset short yields immediate cash (see Cover et al. [1998]).

For example, assume an investor wants to short sell 10 shares of IBM at \$100, and he has \$1000 in cash. First he has to find a lender, the short provider, who is willing to lend the shares. After exchanging the shares and the \$1000 collateral, the investor sells the borrowed shares. After selling the investor has \$1000 in cash again, and the obligation to cover the shorted assets later. If the price drops \$10, he has to cover the short position at \$90, thus he gains $10 \times \$10$ (weight is 10 and return is \$10). If the price rises \$10, he has to cover at \$110, loosing $10 \times \$10$. In this example we assume that our only investment is in asset i and our initial wealth is $V(0)$. We invest a proportion of $\pi_v^i \in (-1, 1)$ of our wealth in the ith risky asset and $\pi_v^0 = 1 - \pi_v^i$ is invested in cash.
If the position is long $(\pi_v^i > 0)$, it results in the wealth in period $n = 1$ as

$$V(1) = V(0)(1 - \pi_v^i) + V(0)\pi_v^i x_1^i = V(0) + V(0)\pi_v^i(X_1^i - 1) = V(0) + \delta^i(S_1^i - S_0^i)$$

where $V(0)\pi_v^0$ is invested in cash. Equivalently, we get

$$V(1) = V(0)(1 - \pi_v^i) + V(0)\pi_v^i(1 + R_1^i) = V(0) + V(0)\pi_v^i R_1^i$$

Again, we hold $V(0)\pi_v^i$ of asset i and we earn $V(0)\pi_v^i R_1^i$ of interest.
While if the position is short $(\pi_v^i < 0)$, we win as much money as price drop of the asset.

***Remark* 9.2.4** *As we are short selling a stock, we do not hold the asset and we do not earn interest out of it. From Equation (9.2.10), we only make a profit when the value of the asset drops.*

Given the return in Equation (9.2.8), when price of asset drops, the return becomes

$$1 - X_n^i = \frac{S_{n-1}^i - S_n^i}{S_{n-1}^i} = -R_n^i \geq 0$$

and the wealth becomes

$$V(1) = V(0) + V(0)|\pi_v^i|(1 - X_1^i) = V(0) - V(0)\pi_v^i(1 - X_1^i) = V(0) + V(0)\pi_v^i(X_1^i - 1)$$

since from Equation (9.2.10) short selling is free. That is, V is used as collateral to borrow stocks, and once sold we receive income (cash) from them. Equivalently, we get

$$V(1) = V(0) + V(0)|\pi_v^i|(-R_1^i) = V(0) + V(0)\pi_v^i R_1^i$$

***Remark* 9.2.5** *Since $\pi_v^i < 0$, when price rises we get positive return $R_1^i > 0$ and we loose money, and when price drops we get negative return $R_1^i < 0$ and we make money.*

Let's consider the general case where $\pi_v = (\pi_v^0, \pi_v^1, ..., \pi_v^N)$ is the portfolio vector such that the 0th component corresponds to cash. From Remark (9.2.4), we can conclude that at the end of the first trading period, $n = 1$, the investor's wealth becomes

$$V(1) = V(0)\Big(\pi_v^0 + \sum_{i=1}^{N}\big[\pi_v^{i^+} X_1^i + \pi_v^{i^-}(X_1^i - 1)\big]\Big)^+$$

where $(.)^-$ denotes the negative part operation. Equivalently, we get

$$V(1) = V(0)\Big(\pi_v^0 + \sum_{i=1}^{N}\big[\pi_v^{i^+}(1 + R_1^i) + \pi_v^{i^-} R_1^i\big]\Big)^+$$

In case of the investor's net wealth falling to zero or below he defaults. However, negative wealth is not allowed in our framework, thus the outer positive part operation. Since only long positions cost money in this setup, we will constrain to portfolios such that

$$\sum_{i=0}^{N} \pi_v^{i^+} = 1$$

leading to $\pi_v^0 = 1 - \sum_{i=1}^{N} \pi_v^{i^+}$. Hence, if the portfolio has no long position, that is, only short position, then from Equation (9.2.10) all the capital is in cash $\pi_v^0 = 1$ since short selling is free. Considering this, we can rewrite the portfolio in period $n = 1$ as

$$\begin{aligned} V(1) &= V(0)\Big(\sum_{i=0}^{N} \pi_v^{i^+} + \sum_{i=1}^{N}\big[\pi_v^{i^+}(X_1^i - 1) + \pi_v^{i^-}(X_1^i - 1)\big]\Big)^+ \\ &= V(0)\Big(1 + \sum_{i=1}^{N}\big[\pi_v^i(X_1^i - 1)\big]\Big)^+ \end{aligned}$$

Equivalently, we get

$$V(1) = V(0)\Big(1 + \sum_{i=1}^{N} \pi_v^i R_1^i\Big)^+$$

This shows that we gain as much as long positions raise and short positions fall (see Remark (9.2.5)). Hence, we can see that short selling is a risky investment, because it is possible to default on total initial wealth without the default of any of the assets in the portfolio. The possibility of this would lead to a growth rate of minus infinity, thus we restrict our market according to

$$1 - B + \delta < X_n^i < 1 + B - \delta\,,\, i = 1, ..., N \tag{9.2.11}$$

or equivalently

$$-B + \delta < R_n^i < B - \delta\,,\, i = 1, ..., N$$

Besides aiming at no-ruin, the role of $\delta > 0$ is ensuring that rate of growth is finite for any portfolio vector. For the usual stock market daily data, there exist $0 < a_1 < 1 < a_2 < \infty$ such that

$$a_1 \leq x_n^i \leq a_2 \text{ or } a_1 - 1 \leq R_n^i \leq a_2 - 1$$

for all $i = 1, ..., N$ and for example $a_1 = 0.7$ and with $a_2 = 1.2$. Thus, we can choose $B = 0.3$. As a result, the maximal loss that we could suffer is

$$B \sum_{i=1}^{N} |\pi_v^i|$$

This value has to be constrained to ensure no-ruin. We denote the set of possible portfolio vectors by

$$\Delta_N^{-B} = \left\{ \pi_v = (\pi_v^0, \pi_v^1, ..., \pi_v^N) \, , \, \pi_v^0 \geq 0 \, , \, \sum_{i=0}^{N} \pi_v^{i^+} = 1 \, , \, B \sum_{i=1}^{N} |\pi_v^i| \leq 1 \right\}$$

where $\sum_{i=0}^{N} \pi_v^{i^+} = 1$ means that we invest all of our initial wealth into some assets - buying long - or cash. By $B \sum_{i=1}^{N} |\pi_v^i| \leq 1$ maximal exposure is limited such that ruin is not possible, and rate of growth it is finite. This is equivalent to

$$\sum_{i=1}^{N} \pi_v^{i^-} \leq \frac{1 - B}{B} = 2.33$$

Since the set of possible portfolio vectors Δ_N^{-B} is not convex we can not apply the Kuhn-Tucker theorem to get the maximum asymptotic average rate of growth. Horvath et al. [2011] proposed to transform the non-convex set Δ_N^{-B} to a convex region $\tilde{\Delta}_N^{-B}$ and showed that in that setting $\tilde{\pi}_v^*$ had the same market exposure as with π_v^*.

9.2.2.3 Long only investment: leveraged

Assuming condition in Equation (9.2.11) then market exposure can be increased over one without the possibility of ruin. Given the portfolio vector $\pi_v = (\pi_v^0, \pi_v^1, ..., \pi_v^N)$ where π_v^0 is the cash component, the no short selling condition implies that $\pi_v^i > 0$ for $i = 1, ..., N$. We assume the investor can borrow money and invest it on the same rate r, and that the maximal investable amount of cash $L_{B,r}$ (relative to initial wealth $V(0)$) is always available for the investor. That is, $L_{B,r} \geq 1$ sometimes called the buying power, is chosen to be the maximal amount, investing of which ruin is not possible given Equation (9.2.11). Because our investor decides over the distribution of his buying power, we get the constraint

$$\sum_{i=0}^{N} \pi_v^i = L_{B,r}$$

so that $\pi_v^0 = L_{B,r} - \sum_{i=1}^{N} \pi_v^i$. Unspent cash earns the same interest r, as the rate of lending. The market vector is defined as

$$X_r = (X^0, X^1, ..., X^N) = (1 + r, X^1, ..., X^N)$$

where $X^0 = 1 + r$. The feasible set of portfolio vectors is

$$\Delta_N^{+B,r} = \left\{ \pi_v = (\pi_v^0, \pi_v^1, ..., \pi_v^N) \in \mathbb{R}_0^{+^{N+1}} \, , \, \sum_{i=0}^{N} \pi_v^i = L_{B,r} \right\}$$

where π_v^0 denotes unspent buying power. Hence, the investor's wealth evolves according to

$$V(1) = V(0)\big(< \pi_v, X_r > -(L_{B,r} - 1)(1 + r) \big)^+$$

where $V(0)r(L_{B,r} - 1)$ is interest on borrowing $(L_{B,r} - 1)$ times the initial wealth $V(0)$. Equivalently, we get

$$V(1) = V(0)\big(\sum_{i=0}^{N} \pi_v^i(1+R^i) - (L_{B,r}-1)(1+r)\big)^+$$

where $\pi_v^0 X^0 = \pi_v^0(1+r)$ so that $R^0 = r$. Given the maximal investable amount of cash $L_{B,r} \geq 1$ relative to initial wealth $V(0)$, we borrow the quantity $(L_{B,r}-1)$ which is invested in the long-only portfolio. However, we do not keep borrowing that quantity on the rolling portfolio and we must subtract it from our rolling portfolio. We can visualise that phenomenon by expending the previous equation

$$V(1) = V(0)\big(L_{B,r} + \sum_{i=0}^{N} \pi_v^i R^i - (L_{B,r}-1)(1+r)\big)^+ = V(0)\big(1 + \sum_{i=0}^{N} \pi_v^i R^i - r(L_{B,r}-1)\big)^+$$

Further, as we borrow the amount $V(0)(L_{B,r}-1)$ from the broker, we must subtract the interest $rV(0)(L_{B,r}-1)$ from our portfolio. To ensure no-ruin and finiteness of growth rate choose

$$L_{B,r} = \frac{1+r}{B+r}$$

This ensures that ruin is not possible:

$$\begin{aligned} &< \pi_v, X_r > -(L_{B,r}-1)(1+r) = \sum_{i=0}^{N} \pi_v^i X^i - (L_{B,r}-1)(1+r) \\ = \quad & \pi_v^0(1+r) + \sum_{i=1}^{N} \pi_v^i X^i - (L_{B,r}-1)(1+r) > \pi_v^0(1+r) + \sum_{i=1}^{N} \pi_v^i(1-B+\delta) - (L_{B,r}-1)(1+r) \end{aligned}$$

and after simplification

$$< \pi_v, X_r > -(L_{B,r}-1)(1+r) > \delta\frac{1+r}{B+r}$$

For details on maximising the asymptotic average rate of growth see Horvath et al. [2011].

9.2.2.4 Short selling and leverage

Assuming short selling and leverage, we combine the results of the previous two sections, so that the investor's wealth evolves according to

$$V(1) = V(0)\Big(\pi_v^0(1+r) + \sum_{i=1}^{N}\big[\pi_v^{i^+} X_1^i + \pi_v^{i^-}(X_1^i - 1 - r)\big] - (L_{B,r}-1)(1+r)\Big)^+$$

where $(.)^-$ denotes the negative part operation. Equivalently, we get

$$V(1) = V(0)\Big(\pi_v^0(1+r) + \sum_{i=1}^{N}\big[\pi_v^{i^+}(1+r_1^i) + \pi_v^{i^-} r_1^i\big] - (L_{B,r}-1)(1+r)\Big)^+ \tag{9.2.12}$$

and expending the equation, the investor's wealth becomes

$$V(1) = V(0)\Big(\pi_v^0(1+r) + \sum_{i=1}^{N}\pi_v^{i^+} + \sum_{i=1}^{N}\big[\pi_v^{i^+} r_1^i + \pi_v^{i^-} r_1^i\big] - (L_{B,r}-1)(1+r)\Big)^+$$

with the buyer power constraint

$$\sum_{i=0}^{N} \pi_v^{i^+} = L_{B,r}$$

so that $\pi_v^0 = L_{B,r} - \sum_{i=1}^{N} \pi_v^{i^+}$. Expending the previous equation, and putting terms together, we get

$$V(1) = V(0)\Big(r\pi_v^0 + 1 + \sum_{i=1}^{N} \pi_v^i r_1^i - r(L_{B,r} - 1)\Big)^+ = V(0)\Big(1 + \sum_{i=0}^{N} \pi_v^i r_1^i - r(L_{B,r} - 1)\Big)^+$$

with $r_1^0 = r$. The feasible set corresponding to the non-convex region is

$$\Delta_N^{\pm B,r} = \Big\{\pi_v = (\pi_v^0, \pi_v^1, ..., \pi_v^N)\ ,\ \sum_{i=0}^{N} \pi_v^{i^+} = L_{B,r}\ ,\ B\sum_{i=0}^{N} |\pi_v^i| \leq 1\Big\} \tag{9.2.13}$$

Again, using the transformation for short selling, Horvath et al. [2011] showed that in that setting $\tilde{\pi}_v^*$ had the same market exposure as with π_v^*.

9.3 Empirical log-optimal portfolio selections

We presented in Section (2.3.2) the growth optimal portfolio (POP) as a portfolio having maximal expected growth rate over any time horizon. Following Gyorfi et al. [2011], we are now going to introduce the growth optimum theory (GOT) as an alternative to the expected utility theory and the mean-variance approaches to asset pricing. Investment strategies are allowed to use information collected from the past of the market, and determine a portfolio at the beginning of a trading period, that is, a way of distributing their current capital among the available assets. The goal of the investor is to maximise his wealth in the long run without knowing the underlying distribution generating the stock prices. Under this assumption the asymptotic rate of growth has a well-defined maximum which can be achieved in full knowledge of the underlying distribution generated by the stock prices. In this section, both static (buy and hold) and dynamic (daily rebalancing) portfolio selections are considered under various assumptions on the behaviour of the market process. While every static portfolio asymptotically approximates the growth rate of the best asset in the study, one can achieve larger growth rate with daily rebalancing. Under memoryless assumption on the underlying process generating the asset prices, the best rebalancing is the log-optimal portfolio, which achieves the maximal asymptotic average growth rate. After presenting the log-optimal portfolio, we will then briefly present the semi-log optimal portfolio selection as an alternative to the log-optimal portfolio.

9.3.1 Static portfolio selection

Keeping the same notation as in Section (2.3.2), the market consists of N assets, represented by an N-dimensional vector process S where

$$S_n = (S_n^0, S_n^1, .., S_n^N)$$

with $S_n^0 = 1$, and such that the ith component S^i of S_n denotes the price of the ith asset on the n-th trading period. Further, we put $S_0^i = 1$. We assume that $\{S_n\}$ has exponential trend:

$$S_n^i = e^{nW_n^i} \approx e^{nW^i}$$

with average growth rate (average yield)

$$W_n^i = \frac{1}{n} \ln S_n^i$$

and with asymptotic average growth rate

$$W^i = \lim_{n\to\infty} \frac{1}{n} \ln S_n^i$$

A static portfolio selection is a single period investment strategy. A portfolio vector is denoted by $\pi_v = (\pi_v^1, .., \pi_v^N)$ where the ith component π_v^i of π_v denotes the proportion of the investor's capital invested in asset i. We assume that the portfolio vector b has non-negative components which sum up to 1, meaning that short selling is not permitted. The set of portfolio vectors is denoted by

$$\Delta_N = \left\{\pi_v = (\pi_v^1, ..., \pi_v^N)\ ,\ \pi_v^i \geq 0\ ,\ \sum_{i=1}^N \pi_v^i = 1\right\}$$

The aim of static portfolio selection is to achieve $\max_{1\leq i\leq N} W^i$. The static portfolio is an index, for example, the $S\&P$ 500 such that at time $n = 0$ we distribute the initial capital $V(0)$ according to a fix portfolio vector π_v, that is, if $V(n)$ denotes the wealth at the trading period n, then

$$V(n) = V(0) \sum_{i=1}^N \pi_v^i S_n^i$$

We apply the following simple bounds

$$V(0) \max_i \pi_v^i S_n^i \leq V(n) \leq NV(0) \max_i \pi_v^i S_n^i$$

If $\pi_v^i > 0$ for all $i = 1, .., N$ then these bounds imply that

$$W = \lim_{n\to\infty} \frac{1}{n} \ln V(n) = \lim_{n\to\infty} \max_i \frac{1}{n} \ln S_n^i = \max_i W^i$$

Thus, any static portfolio selection achieves the growth rate of the best asset in the study, $\max_i W^i$, and so the limit does not depend on the portfolio π_v. In case of uniform portfolio (uniform index) $\pi_v^i = \frac{1}{N}$, and the convergence above is from below:

$$V(0) \max_i \frac{1}{N} S_n^i \leq V(n) \leq V(0) \max_i S_n^i$$

9.3.2 Constantly rebalanced portfolio selection

In order to apply the usual prediction techniques for time series analysis one has to transform the sequence price vectors $\{S_n\}$ into a more or less stationary sequence of return vectors (price relatives) $\{X_n\}$

$$X_n = (X_n^1, .., X_n^N)$$

such that $X_n^i = \frac{S_n^i}{S_{n-1}^i} = R_n^i + 1$. With respect to the static portfolio, one can achieve even higher growth rate for long run investments, if we make rebalancing, that is, if the tuning of the portfolio is allowed dynamically after each trading period. The dynamic portfolio selection is a multi-period investment strategy, where at the beginning of each trading period we can rearrange the wealth among the assets. A representative example of the dynamic portfolio selection is the constantly rebalanced portfolio (CRP), which was introduced and studied by Kelly [1956], Latane [1959], Breiman [1961], Markowitz [1976]. In case of CRP we fix a portfolio vector $b \in \Delta_N$, that is, we are concerned with a hypothetical investor who neither consumes nor deposits new cash into his portfolio, but reinvests his portfolio at each trading period. In fact, neither short selling, nor leverage is allowed. The investor has to rebalance his portfolio after each trading day to corrigate the daily price shifts of the invested stocks.

Let $V(0)$ denote the investor's initial capital. Then at the beginning of the first trading period $V(0)\pi_v^i$ is invested into asset i, and it results in return $V(0)\pi_v^i X_1^i = V(0)b^i(1+R_1^i)$. Therefore at the end of the first trading period the investor's wealth becomes

$$V(1) = V(0)\sum_{i=1}^{N}\pi_v^i X_1^i = V(0)\sum_{i=1}^{N}\pi_v^i(1+R_1^i) = V(0) < \pi_v, X_1 >$$

where $< ., . >$ denotes inner product. For the second trading period, $V(1)$ is the new initial capital

$$V(2) = V(1) < \pi_v, X_2 >= V(0) < \pi_v, X_1 >< \pi_v, X_2 >$$

By induction, for the trading period n the initial capital is $V(n-1)$, therefore

$$V(n) = V(n-1) < \pi_v, X_n >= V(0)\prod_{j=1}^{n} < \pi_v, X_j >$$

The asymptotic average growth rate of this portfolio selection is

$$W = \lim_{n\to\infty}\frac{1}{n}\ln V(n) = \lim_{n\to\infty}\big(\frac{1}{n}\ln V(0) + \frac{1}{n}\sum_{j=1}^{n}\ln < \pi_v, X_j >\big) = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\ln < \pi_v, X_j >$$

therefore without loss of generality one can assume in the sequel that the initial capital $V(0) = 1$.

9.3.2.1 Log-optimal portfolio for memoryless market process

If the market process $\{X_i\}$ is memoryless, that is, it is a sequence of independent and identically distributed (i.i.d.) random return vectors, then one can show that the best constantly rebalanced portfolio (BCRP) is the log-optimal portfolio:

$$\pi_v^* = arg\max_{\pi_v\in\Delta_N} E[\ln < \pi_v, X_1 >] \tag{9.3.14}$$

This optimality means that if $V^*(n) = V(n)(\pi_v^*)$ denotes the capital after day n achieved by a log-optimal portfolio strategy π_v^*, then for any portfolio strategy π_v with finite $E[(\ln < \pi_v, X_1 >)^2]$ and with capital $V(n) = V(n)(\pi_v)$ and for any memoryless market process $\{X_n\}_{-\infty}^{\infty}$

$$W = \lim_{n\to\infty}\frac{1}{n}\ln V(n) \leq \lim_{n\to\infty}\frac{1}{n}\ln V^*(n) \text{ almost surely}$$

and maximal asymptotic average growth rate is

$$\lim_{n\to\infty}\frac{1}{n}\ln V^*(n) = W^* = E[\ln < \pi_v^*, X_1 >] \text{ almost surely}$$

For the proof, set $W(\pi_v) = E[\ln < \pi_v, X_1 >]$ and use the strong law of large numbers (see Gyorfi et al. [2011]). For memoryless or Markovian market process, optimal strategies have been introduced if the distributions of the market process are known.

The principle of log-optimality has the important consequence that

$$V^{\pi_v}(n) \text{ is not close to } E[V^{\pi_v}(n)]$$

The optimality property discribed above means that, for any $\delta > 0$ the event

$$\{-\delta < \frac{1}{n}\ln V^{\pi_v}(n) - E[\ln < \pi_v, X_1 >] < \delta\}$$

has probability close to 1 if n is large enough. On the one hand, we have that

$$\{-\delta < \frac{1}{n}\ln V^{\pi_v}(n) - E[\ln < \pi_v, X_1 >] < \delta\} = \{e^{n(-\delta + E[\ln <\pi_v, X_1>])} < \ln V^{\pi_v}(n) < e^{n(\delta + E[\ln <\pi_v, X_1>])}\}$$

so that

$$V^{\pi_v}(n) \text{ is close to } e^{nE[\ln <\pi_v, X_1>]}$$

On the other hand

$$E[V^{\pi_v}(n)] = E[\prod_{j=1}^{n} < \pi_v, X_j >] = \prod_{j=1}^{n} < \pi_v, E[X_j] >= e^{n \ln <\pi_v, E[X_1]>}$$

By Jensen inequality, we get $\ln < \pi_v, E[X_1] > > E[\ln < \pi_v, X_1 >]$ and therefore

$$V^{\pi_v}(n) \text{ is much less than } E[V^{\pi_v}(n)]$$

Not knowing this fact, one can apply a naive approach

$$arg\max_{\pi_v} E[V^{\pi_v}(n)]$$

Because of

$$E[V^{\pi_v}(n)] =< \pi_v, E[X_1] >^n$$

this naive approach has the equivalent form

$$arg\max_{\pi_v} E[V^{\pi_v}(n)] = arg\max_{\pi_v} < \pi_v, E[X_1] >$$

which is called the Mean approach. In that setting $arg\max_{\pi_v} < \pi_v, E[X_1] >$ is a portfolio vector having 1 at the position, where the vector $E[X_1]$ has the largest component. In his seminal paper, Markowitz [1952] realised that the mean approach was inadequate, that is, it is a dangerous portfolio. In order to avoid this difficulty he suggested a diversification, which is called Mean-Variance portfolio such that

$$\tilde{\pi}_v = arg \max_{\pi_v : Var(<\pi_v, X_1>) \leq \lambda} < \pi_v, E[X_1] > \tag{9.3.15}$$

where $\lambda > 0$ is the investor's risk aversion parameter. For appropriate choice of λ the performance (average growth rate) of $\tilde{\pi}_v$ can be close to the performance of the optimal π_v^*, however, the good choice of λ depends on the (unknown) distribution of the return vector X. The calculation of $\tilde{\pi}_v$ is a quadratic programming (QP) problem, where a linear function is maximised under quadratic constraints.

In order to calculate the log-optimal portfolio π_v^*, one has to know the distribution of X_1. If this distribution is unknown then the empirical log-optimal portfolio can be defined by

$$\pi_v^*(n) = arg\max_{\pi_v} \frac{1}{n}\sum_{j=1}^{n} \ln < \pi_v, X_j > \tag{9.3.16}$$

with linear constraints

$$\sum_{i=1}^{N} \pi_v^i = 1 \text{ and } 0 \leq \pi_v^i \leq 1 \text{ for } i = 1, .., N$$

The behaviour of this empirical portfolio was studied by Mori [1984] [1986]. The calculation of $\pi_v^*(n)$ is a nonlinear programming (NLP) problem (see Cover [1984]).

9.3.2.2 Semi-log-optimal portfolio

Roll [1973], Pulley [1994] and Vajda [2006] suggested an approximation of π_v^* and $\pi_v^*(n)$ using

$$h(z) = z - 1 - \frac{1}{2}(z-1)^2$$

which is the second order Taylor expansion of the function $\ln z$ at $z = 1$. Then, the semi-log-optimal portfolio selection is

$$\overline{\pi}_v = arg \max_{\pi_v} E[h(< \pi_v, X_1 >)]$$

and the empirical semi-log-optimal portfolio is

$$\overline{\pi}_v(n) = arg \max_{\pi_v} \frac{1}{n} \sum_{j=1}^{n} h(< \pi_v, X_j >)$$

In order to compute $\pi_v^*(n)$, one has to make an optimisation over π_v. In each optimisation step the computational complexity is proportional to n. For $\overline{\pi}_v(n)$, this complexity can be reduced so that the running time can be much smaller (see Gyorfi et al. [2011]). The other advantage of the semi-log-optimal portfolio is that it can be calculated via quadratic programming.

9.3.3 Time varying portfolio selection

For a general dynamic portfolio selection, the portfolio vector may depend on the past data. As before $X_j = (X_j^1, .., X_j^N)$ denotes the return vector on trading period j. Let $\pi_v = \pi_v(1)$ be the portfolio vector for the first trading period. For initial capital $V(0)$, we get that

$$V(1) = V(0) < \pi_v(1), X_1 >= V(0) < \pi_v(1), 1 + R_1 >$$

For the second trading period, $n = 2$, $V(1)$ is new initial capital, the portfolio vector is $\pi_v(2) = \pi_v(X_1)$, and

$$V(2) = V(0) < \pi_v(1), X_1 >< \pi_v(X_1), X_2 >$$

For the nth trading period, a portfolio vector is $\pi_v(n) = \pi_v(X_1, .., X_{n-1}) = \pi_v(X_1^{n-1})$ and

$$V(n) = V(0) \prod_{j=1}^{n} < \pi_v(X_1^{j-1}), X_j >= V(0) e^{nW_n(B)}$$

with the average growth rate

$$W_n(B) = \frac{1}{n} \sum_{j=1}^{n} \ln < \pi_v(X_1^{j-1}), X_j >$$

9.3.3.1 Log-optimal portfolio for stationary market process

The fundamental limits, determined in Mori [1982], in Algoet et al. [1988], and in Algoet [1992] [1992], reveal that the so-called (conditionally) log-optimal portfolio $B^* = \{\pi_v^*(.)\}$ is the best possible choice. More precisely, on trading period n, let $\pi_v^*(.)$ be such that

$$E[\ln < \pi_v^*(X_1^{n-1}), X_n > | X_1^{n-1}] = \max_{\pi_v(.)} E[\ln < \pi_v(X_1^{n-1}), X_n > | X_1^{n-1}] \tag{9.3.17}$$

If $V^*(n) = V^{B^*}(n)$ denotes the capital achieved by a log-optimal portfolio strategy B^* after n trading periods, then for any other investment strategy B with capital $V(n) = V^B(n)$ and with

$$\sup_n E[(\ln < \pi_v(X_1^{n-1}), X_n >)^2] < \infty$$

and for any stationary and ergodic process $\{X_n\}_{-\infty}^{\infty}$

$$\lim_{n\to\infty} \sup\big(\frac{1}{n}\ln V(n) - \frac{1}{n}\ln V^*(n)\big) \leq 0 \text{ almost surely}$$

and

$$\lim_{n\to\infty} \frac{1}{n}\ln V^*(n) = W^* \text{ almost surely}$$

where

$$W^* = E\big[\max_{b(.)} E[\ln < \pi_v(X_{-\infty}^{-1}), X_0 > | X_{-\infty}^{-1}]\big]$$

is the maximal possible growth rate of any investment strategy.

***Remark* 9.3.1** *Note that for memoryless markets $W^* = \max_b E[\ln < \pi_v, X_0 >]$ which shows that in this case the log-optimal portfolio is the best constantly rebalanced portfolio.*

9.3.3.2 Empirical portfolio selection

The optimality relations proved above give rise to the following definition:

Definition 9.3.1 *An empirical (data driven) portfolio strategy B is called universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes $\{X_n\}_{-\infty}^{\infty}$ if for each process in the class*

$$\lim_{n\to\infty} \frac{1}{n}\ln V^B(n) = W^* \text{ almost surely}$$

It is not at all obvious that such universally consistent portfolio strategy exists. The surprising fact that there exists a strategy, universal with respect to a class of stationary and ergodic processes was proved by Algoet [1992].

Most of the papers dealing with portfolio selections assume that the distributions of the market process are known. If the distributions are unknown then one can apply a two stage splitting scheme.

1. In the first time period the investor collects data, and estimates the corresponding distributions. In this period there is no investment.
2. In the second time period the investor derives strategies from the distribution estimates and performs the investments.

Gyorfi et al. [2011] showed that there is no need to make any splitting, one can construct sequential algorithms such that the investor can make trading during the whole time period, that is, the estimation and the portfolio selection is made on the whole time period. Given the definition of the conditionally log-optimal portfolio in Equation (9.3.17) for a fixed integer $k > 0$ large enough, we expect that

$$E[\ln < \pi_v(X_1^{n-1}), X_n > | X_1^{n-1}] \approx E[\ln < \pi_v(X_{n-k}^{n-1}), X_n > | X_{n-k}^{n-1}]$$

and

$$\pi_v^*(X_1^{n-1}) \approx b_k(X_{n-k}^{n-1}) = arg\max_{\pi_v(.)} E[\ln < \pi_v(X_{n-k}^{n-1}), X_n > | X_{n-k}^{n-1}]$$

Because of stationarity

$$b_k(x_1^k) = arg\max_{\pi_b} E[\ln < \pi_v, X_{k+1} > | X_1^k = x_1^k]$$

which is the maximisation of the regression function

$$m_b(x_1^k) = E[\ln < b, X_{k+1} > | X_1^k = x_1^k]$$

Thus, a possible way for asymptotically optimal empirical portfolio selection is that, based on the past data, sequentially estimate the regression function $m_b(x_1^k)$ and choose the portfolio vector, which maximises the regression function estimate.

9.3.4 Regression function estimation: The local averaging estimates

We first consider the basics of nonparametric regression function estimation. Let Y be a real valued random variable, and let X denote an observation vector taking values in $\mathbb{R}^d$. The regression function is the conditional expectation of Y given X:

$$m(x) = E[Y|X = x]$$

If the distribution of (X, Y) is unknown, then one has to estimate the regression function from data. The data is a sequence of i.i.d. copies of (X, Y):

$$D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$$

The regression function estimate is of form

$$m_n(x) = m_n(x, D_n)$$

An important class of estimates is the local averaging estimates

$$m_n(x) = \sum_{j=1}^{n} \omega_{n,j}(x; X_1, .., X_n) Y_j \tag{9.3.18}$$

where usually the weights $\omega_{n,j}(x; X_1, .., X_n)$ for $j = 1, .., n$ are non-negative and sum up to 1. Moreover, the weight $\omega_{n,j}(x; X_1, .., X_n)$ is relatively large if x is close to X_j, otherwise it is zero. Note, we can always rewrite the local averaging estimates as

$$m_n(x) = \frac{1}{norm} \sum_{j=1}^{n} \overline{\omega}_{n,j}(x; X_1, .., X_n) Y_j \tag{9.3.19}$$

$norm = \sum_{j=1}^{n} \overline{\omega}_{n,j}(x; X_1, .., X_n)$.

9.3.4.1 The partitioning estimate

An example of such an estimate is the partitioning estimate. Here one chooses a finite or countably infinite partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, ...\}$ of $\mathbb{R}^d$ consisting of cells $A_{n,l} \subset \mathbb{R}^d$ and defines, for $x \in A_{n,l}$ the estimate by averaging the Y_j with the corresponding X_j in $A_{n,j}$. That is

$$m_n(x) = \frac{\sum_{j=1}^n I_{\{X_j \in A_{n,l}\}} Y_j}{\sum_{j=1}^n I_{\{X_j \in A_{n,l}\}}} = \frac{1}{norm} \sum_{j=1}^n I_{\{X_j \in A_{n,l}\}} Y_j \text{ for } x \in A_{n,l} \tag{9.3.20}$$

where $\overline{\omega}_j = I_{\{X_j \in A_{n,l}\}}$ and $norm = \sum_{j=1}^n \overline{\omega}_j$. For notational purpose we use the convention $\frac{0}{0} = 0$. In order to have consistency, on the one hand we need that the cells $A_{n,l}$ should be small, and on the other hand the number of non-zero terms in the denominator of the above ratio should be large. These requirements can be satisfied if the sequences of partition $\mathcal{P}_n$ is asymptotically fine, that is, if

$$diam(A) = \sup_{x,y \in A} ||x - y||$$

denotes the diameter of a set such that $||.||$ is the Euclidean norm, then for each sphere S centred at the origin

$$\lim_{n \to \infty} \max_{l : A_{n,l} \cap S \neq 0} diam(A_{n,l}) = 0$$

and

$$\lim_{n \to \infty} \frac{|\{l : A_{n,l} \cap S \neq 0\}|}{n} = 0$$

For the partition $\mathcal{P}_n$ the most important example is when the cells $A_{n,l}$ are cubes of volume h_n^d. For cubic partition, the consistency conditions above mean that

$$\lim_{n \to \infty} h_n = 0 \text{ and } \lim_{n \to \infty} n h_n^d = \infty \tag{9.3.21}$$

9.3.4.2 The Nadaraya-Watson kernel estimate

The second example of a local averaging estimate is the Nadaraya-Watson kernel estimate. Let $K : \mathbb{R}^d \to \mathbb{R}_+$ be a function called the kernel function, and let $h > 0$ be a bandwidth. The kernel estimate is defined by

$$m_n(x) = \frac{\sum_{j=1}^n K(\frac{x - X_j}{h}) Y_j}{\sum_{j=1}^n K(\frac{x - X_j}{h})} = \frac{1}{norm} \sum_{j=1}^n K(\frac{x - X_j}{h}) Y_j$$

The kernel estimate is a weighted average of the Y_j, where the weight of Y_j (i.e., the influence of Y_j on the value of the estimate at x) depends on the distance between X_j and x. For the bandwidth $h = h_n$, the consistency conditions are given in Equation (9.3.21). If one uses the so-called naive kernel (or window kernel)

$$K(x) = I_{\{||x|| \leq 1\}}$$

then we get

$$m_n(x) = \frac{\sum_{j=1}^n I_{\{||x - X_j|| \leq h\}} Y_j}{\sum_{j=1}^n I_{\{||x - X_j|| \leq h\}}} = \frac{1}{norm} \sum_{j=1}^n I_{\{||x - X_j|| \leq h\}} Y_j$$

where $\overline{\omega}_j = I_{\{||x - X_j|| \leq h\}}$ and $norm = \sum_{j=1}^n \overline{\omega}_j$. That is, one estimates $m(x)$ by averaging the Y_j's such that the distance between X_j and x is not greater than h.

9.3.4.3 The k-nearest neighbour estimate

Another example of local averaging estimates is the k-nearest neighbour (k-NN) estimate. Here one determines the k nearest X_j's to x in terms of the distance $||x - X_j||$ and estimates $m(x)$ by the average of the corresponding Y_j's. That is, for $x \in \mathbb{R}^d$, let

$$(X_{(1)}(x), Y_{(1)}(x)), .., (X_{(n)}(x), Y_{(n)}(x))$$

be a permutation of

$$(X_1, Y_1), .., (X_n, Y_n)$$

such that

$$||x - X_{(1)}(x)|| \leq ... \leq ||x - X_{(n)}(x)||$$

Then, the k-NN estimate is defined by

$$m_n(x) = \frac{1}{k} \sum_{j=1}^{k} Y_{(j)}(x)$$

If $k = k_n \to \infty$ such that $\frac{k_n}{n} \to 0$ then the k-nearest-neighbour regression estimate is consistent.

9.3.4.4 The correspondence

We use the following correspondence between the general regression estimation and portfolio selection:

$$\begin{aligned} X &\sim X_1^k \\ Y &\sim \ln < b, X_{k+1} > \end{aligned}$$

and

$$m(x) = E[Y|X = x] \sim m_b(x_1^k) = E[\ln < b, X_{k+1} > |X_1^k = x_1^k]$$

Note, the theoretical results above hold under the condition of stationarity. Obviously, the real data of returns (relative prices) are not stationary.

9.4 A simple example

9.4.1 A self-financed long-short portfolio

We generally assign weights δ^i to the ith instrument in the portfolio together with a condition of no short sales to guarantee that the sum of the weights equal one. However, as seen in Section (9.2.1), this is no-longer the case when we assume short selling. To overcome this problem, we assume that the index is made of N stocks, and given n periods of time such that $\Delta t = \frac{T}{n}$, we want to build at period j two synthetic portfolios, one made of long stocks and the other one made of short stocks. For each stock $S_i(j)$ with $i = 1, .., N$ we compute the measure $M_i(j)$ observed at time $j\Delta$ and use it to decide weather to buy or sell the stock. We let the Decision Variable $\epsilon_i(j)$ at time $j\Delta$ takes values in the set $\{1, 0, -1\}$. We let $R_i(j)$ be the return of the ith stock in the jth period, and assume that if $M_i(j) > \gamma_U$ we set the Decision Variable to $\epsilon_i(j) = 1$ and we buy the stock, and if $M_i(j) < \gamma_D$ we set $\epsilon_i(j, k) = -1$ and we sell the stock. In the case where $\gamma_D < M_i(j) < \gamma_U$ we choose to do nothing and set $\epsilon_i(j) = 0$.

We let $\delta = (\delta_0, \delta_1, ..., \delta_N)$ be the portfolio vector such that the 0th component corresponds to cash, and we let h_R be the fraction of wealth allocated to the risky portfolio where $1 - h_R$ is allocated to the risk-free security. We then create a weighted portfolio at period j as

$$V^\delta(j) = \sum_{i=1}^{N} \delta_i(j)\epsilon_i(j)S_i(j)$$

with $N \geq N_b + N_s$, where N_b is the number of elements in the long portfolio and N_s is the number of elements in the short portfolio, and where the weights $\delta_i(j)$ are used to allocate the capital C_A. Given the capital C_A, we want to choose the weights $\delta_i^b(\bullet) \geq 0$ such that their sum in the long portfolio equal the percentage ξ_b of capital

$$\sum_{i=0}^{N_b} \delta_i^b(j\delta) = \xi_b$$

where $\xi_b \in [1, 2]$. Note, we let the cash weight δ_0 be part of the long portfolio. Similarly, we want to choose the weights $\delta_i^s(\bullet) < 0$ such that their sum in the short portfolio is a percentage of the capital C_A

$$\sum_{i=1}^{N_s} |\delta_i^s(j\delta)| = \xi_s$$

where $\xi_s \in [0, 1]$. Note, it satisfies the short selling leverage constraint in Equation (9.2.13) as

$$\sum_{i=0}^{N} |\delta_i(j\delta)| = \xi_b + \xi_s = L_{B,r}$$

For example, a 150-50 portfolio has $\xi_b = 1.5$ and $\xi_s = 0.5$ where we sell short 50% of the capital C_A. The daily PnL in the range $[j\Delta, (j+1)\Delta]$ is computed by investing in the portfolio at period j and unwinding it at period $(j+1)$. That is, we let $V^\delta(j)$ be the value of the portfolio at period (j), and $V^\delta(j+1)$ be the price of the portfolio after one period. Then, by using discrete compounding, accounting for fixed costs, and allowing for leverage, we derive (see Section (9.2.2)) the usual expression for a portfolio return as

$$\begin{aligned} V^\delta(j+1) &= V^\delta(j)\Big\{\delta_0^b(j)(1+r) + \sum_{i=1}^{N}\big(\delta_i^b(j)\big[1 + \epsilon_i^b(j)R_i(j, j+1) - c_F^b\big] \\ &+ \delta_i^s(j)\big[\epsilon_i^s(j)R_i(j, j+1) - c_F^s\big]\big) - (\xi_b - 1)\Big\} \end{aligned}$$

where $V^\delta(0) = C_A$ and c_F is a fixed cost. Note, the 0th component $\delta_0^b(j)$ is a fraction of wealth invested in cash corresponding to

$$\delta_0^b(j) = 1 - \sum_{i=1}^{N_b} \delta_i^b(j)$$

Using the constraint on the long portfolio, we can re-express the long-short portfolio as

$$V^\delta(j+1) = V^\delta(j)\Big(1 + \sum_{i=0}^{N} \delta_i(j)\big[\epsilon_i(j)R_i(j, j+1)\big] - \xi_b c_F^b - \xi_s c_F^s\Big)$$

where $R_0(.,.) = r$. We let

$$R_p(j, j+1) = \frac{V^\delta(j+1) - V^\delta(j)}{V^\delta(j)}$$

be the return of the portfolio for the period $(j+1)$ and replace $V^{\delta}(j+1)$ with its value to get

$$R_p(j,j+1) = \sum_{i=0}^{N} \delta_i(j)\epsilon_i(j)R_i(j,j+1) - \xi_b c_F^b - \xi_s c_F^s$$

so that the portfolio at period $(j+1)$ can be expressed in terms of the portfolio return as

$$V^{\delta}(j+1) = V^{\delta}(j)\big(1 + R_p(j,j+1)\big) \tag{9.4.22}$$

Hence, the profit and loss over one period of time becomes

$$\begin{aligned} PnL(j,j+1) &= V^{\delta}(j+1) - V^{\delta}(j) = V^{\delta}(j)R_p(j,j+1) \\ &= V^{\delta}(j)\sum_{i=0}^{N} \delta_i(j)\epsilon_i(j)R_i(j,j+1) - c_F^b V^{\delta}(j) - c_F^s \xi_s V^{\delta}(j) \end{aligned}$$

Given the maturity $T = n\Delta$, the cummulated PnL in the range $[0,T]$ is given by

$$CPnL(0,n) = \sum_{j=0}^{n-1} PnL(j,j+1) = \sum_{j=0}^{n-1} V^{\delta}(j)R_p(j,j+1)$$

That is, the cummulated PnL can be expressed as a weighted sum of portfolio returns

$$CPnL(0,n) = \sum_{j=0}^{n-1} W(j)R_p(j,j+1) \tag{9.4.23}$$

with weights $W(j) = V^{\delta}(j)$ for $j = 0,..,n-1$. Affecting an initial investment of C_A, and assuming no capital inflows and outflows after the initial investment (see Section (9.2.2)), the expression for the portfolio return at time $T = n\Delta$ is

$$V^{\delta}(n) = V^{\delta}(0)\prod_{j=0}^{n-1}\big(1 + R_p(j,j+1)\big) \tag{9.4.24}$$

We can write

$$V^{\delta}(j+1) = V^{\delta}(0)\prod_{k=0}^{j}\big(1 + R_p(k,k+1)\big) = V^{\delta}(j)\big(1 + R_p(j,j+1)\big)$$

and the profit and loss becomes

$$PnL(j,j+1) = V^{\delta}(j+1) - V(j) = V(j)R_p(j,j+1)$$

and the cummulated PnL is

$$\begin{aligned} CPnL(0,n) &= \sum_{j=0}^{n-1} PnL(j,j+1) = V^{\delta}(n) - V^{\delta}(0) \\ &= V^{\delta}(0)\Big[\prod_{j=0}^{n-1}\big(1 + R_p(j,j+1)\big) - 1\Big] \end{aligned}$$

Expending the product term, we get

$$\prod_{j=0}^{n-1}\big(1+R_p(j,j+1)\big) = 1+\sum_{j=0}^{n-1}R_p(j,j+1)+R_p(0,1)\sum_{j=1}^{n-1}R_p(j,j+1)$$
$$+ \quad R_p(1,2)\sum_{j=2}^{n-1}R_p(j,j+1)+...+R_p(n-2,n-1)R_p(n-1,n)+\prod_{j=0}^{n-1}R_p(j,j+1)$$

so that the cummulated PnL can be written as

$$CPnL(0,n) = V^\delta(0)\Big[\sum_{j=0}^{n-1}R_p(j,j+1)+R_p(0,1)\sum_{j=1}^{n-1}R_p(j,j+1)+R_p(1,2)\sum_{j=2}^{n-1}R_p(j,j+1)$$
$$+ \quad ...+R_p(n-2,n-1)R_p(n-1,n)+\prod_{j=0}^{n-1}R_p(j,j+1)\Big]$$

9.4.2 Allowing for capital inflows and outflows

As explained in Section (9.2.2.1), a constantly rebalanced portfolio (CRP) is a self-financing portfolio strategy, rebalancing to the same proportional portfolio in each investment period. That is the investor neither consumes from, nor deposits new cash into his account, but reinvests his capital in each trading period. In the previous Section, given the initial investment C_A, we assumed no capital inflows and outflows after the initial investment and built a long-short portfolio. In the special case where we discretely invest the fixed quantity C_A at each period by locking each daily profit at the risk-free rate in a separate account and borrowing cash from the broker when needed, the portfolio in Equation (9.4.22) simplifies to

$$V^\delta(j+1) = V^\delta(0)\big(1+R_p(j,j+1)\big)$$

At maturity $T = n\Delta t$ the portfolio becomes

$$V^\delta(n) = V^\delta(0)\big(1+R_p(n-1,n)\big)$$

which is to compare with Equation (9.4.24). Each investment being independent from one another, the profit and loss realised in the period $[j, j+1]$ becomes

$$PnL(j,j+1) = V^\delta(j+1)-V^\delta(0) = V^\delta(0)R_p(j,j+1)$$

so that the amount of cash deposited at period $(j+1)$ in a separate account at the risk-free rate is the profit and loss $PnL(j,j+1)$. Hence, the cummulated PnL generated in the range $[0,T]$ is

$$\overline{CPnL}(0,n) = \sum_{j=0}^{n-1}PnL(j,j+1) = V^\delta(0)\sum_{j=0}^{n-1}R_p(j,j+1)$$

In that setting, the weights of the cummulated PnL given in Equation (9.4.23) simplifies to $W(j) = C_A$ and are time-independent. Comparing the two cummulated PnL, their difference is

$$CPnL(0,n) - \overline{CPnL}(0,n) = V^{\delta}(0)\Big[R_p(0,1)\sum_{j=1}^{n-1} R_p(j,j+1) + R_p(1,2)\sum_{j=2}^{n-1} R_p(j,j+1)$$
$$+ \quad ... + R_p(n-2,n-1)R_p(n-1,n) + \prod_{j=0}^{n-1} R_p(j,j+1)\Big]$$

9.4.3 Allocating the weights

As only long positions cost money in the setup, we consider only the portfolio for the buy stocks with N_b stocks and follow the approach described in Section (9.4.4). Note, the 0th component $\omega_0(j\delta)$ is a fraction of wealth invested in cash. In our setting, we need to specify a modified weight $\overline{w}_i$ and then define the quantity $norm$ as

$$norm = \sum_{i=0}^{N_b} |\overline{w}_i(j\delta)|$$

such that the weights becomes $w_i(j\delta) = \xi_b \frac{\overline{w}_i(j\delta)}{norm}$ for $i = 1, .., N_b$ with a sum equal to ξ_b. In the case of the short portfolio we get

$$norm = \sum_{i=1}^{N_s} |\overline{w}_i(j\delta)|$$

with weights $w_i(j\delta) = \xi_s \frac{\overline{w}_i(j\delta)}{norm}$ for $i = 1, .., N_s$. The choice of the modified weight $\overline{w}_i(j\delta)$ is crucial in obtaining the best allocation to the market data.

9.4.3.1 Choosing uniform weights

We assume that all the weights are uniformaly distributed and set $\overline{\omega}_i(j\delta) = 1$ for $i = 1, .., N_b$.

9.4.3.2 Choosing Beta for the weight

We let $\beta_i^a(j\delta)$, taking values in $\mathbb{R}$, be the statistical Beta for the stock $S_i(j\delta)$. We want to define a mapping allocating maximum weight to stocks with $\beta = 0$, and decreasing weight as the β increases. One possibility is to set

$$\beta_i(j\delta) = a + b\beta_i^a(j\delta) \ , i = 1, .., N_b$$

with $a = \frac{1}{3}$ and $b = \frac{2}{3}$. Then, we define the modified weight as

$$\overline{\omega}_i(j\delta) = \frac{1}{\beta_i(j\delta)}$$

An alternative approach is to consider a bell shape for the distribution of the Beta and define the modified weight as

$$\overline{\omega}_i(j\delta) = ae^{-b(\beta_i^a(j\delta))^2}$$

with $a = 3$ and $b = 0.25$. In that setting $\overline{\omega}_i(j\delta) \in [0, a]$ with the property that when $\beta = 0$ we get the maximum value a.

9.4.3.3 Choosing Alpha for the weight

We let $\alpha_i^a(j\delta)$, taking values in $\mathbb{R}$, be the statistical Alpha for the stock $S_i(j\delta)$. We want to define a mapping allocating minimum weight to stocks with $\beta = 0$, and increasing weight as the α increases. One possibility is to define the modified weight as

$$\overline{\omega}_i(j\delta) = a\big(1 - e^{-c\alpha_i^a(j\delta)I_{\{\alpha_i^a(j\delta)\geq 0\}}}\big)$$

with $a = 3$ and $c = 0.3$. In that setting $\overline{\omega}_i(j\delta) \in [0, a]$ with the property that when $\alpha^a = 0$ we get the minimum value 0.

9.4.3.4 Combining Alpha and Beta for the weight

We let $\alpha_i^a(j\delta)$ and $\beta_i^a(j\delta)$ taking values in $\mathbb{R}$ be respectively the statistical Alpha and Beta for the stock $S_i(j\delta)$. We let the blending parameter $p \in [0, 1]$ be the weight representing the quantity of Beta, and define the combined quantity as

$$\overline{\omega}_i(j\delta) = (1-p)\overline{\omega}_i^{\alpha}(j\delta) + p\overline{\omega}_i^{\beta}(j\delta)$$

As an example we set $p = \frac{2}{3}$.

9.4.4 Building a beta neutral portfolio

9.4.4.1 A quasi-beta neutral portfolio

One possibility to satisfy the leverage constraint and to get a quasi-beta-neutral portfolio is to set $h_i = \frac{1}{\beta_i}$. Further, due to the linearity of the leverage constraint, we get $\sum_{i\in L}|h_i| = l_L$ and $\sum_{i\in S}|h_i| = l_S$. Hence, we need to specify a modified weight $h_i^L = \frac{1}{\beta_i}$ for $i \in L$ and then define the quantity $norm_L$ as

$$norm_L = \sum_{i\in L}|h_i^L|$$

such that the weights becomes $h_i = l_L\frac{h_i^L}{norm_L}$ for $i \in L$ with the sum of $|h_i|$ equal to l_L. Similarly, in the case of the short subset, given the modified weight $h_i^S = \frac{1}{\beta_i}$ for $i \in S$ we get

$$norm_S = \sum_{i\in S}|h_i^S|$$

with weights $h_i = l_S\frac{h_i^S}{norm_S}$ for $i \in S$. As the constraint for the beta-neutral portfolio is linear, we need to satisfy

$$\sum_{i\in L} h_i\beta_i = -\sum_{i\in S} h_i\beta_i$$

where $h_i > 0$ for $i \in L$. Replacing the weights h_i with their values in the above equation, we get

$$\sum_{i\in L} h_i\beta_i = \frac{l_L}{norm_L}\sum_{i\in L}\frac{1}{\beta_i}\beta_i = \frac{l_L}{norm_L}N_L$$
$$\sum_{i\in S} h_i\beta_i = \frac{l_S}{norm_S}\sum_{i\in S}\frac{1}{\beta_i}\beta_i = \frac{l_S}{norm_S}N_S$$

with N_L elements in the set L and N_S elements in the set S. In a 100-100 long-short porfolio $l_L = l_S = 1$ so that we are comparing

$$\frac{N_L}{norm_L} = \frac{N_L}{\sum_{i\in L}|\frac{1}{\beta_i}|} \text{ with } \frac{N_S}{norm_S} = \frac{N_S}{\sum_{i\in S}|\frac{1}{\beta_i}|}$$

9.4.4.2 An exact beta-neutral portfolio

One can get an exact beta-neutral portfolio by choosing freely the set of modified weights $(h_2^L, .., h_{N_L}^L)$ within the set L of long stocks, and by constraining the first weight as

$$h_1^L = -\frac{1}{\beta_1}\sum_{i=2}^{N_L} h_i^L \beta_i$$

Again, for $i \in L$ we define the quantity $norm_L$ as

$$norm_L = \sum_{i\in L} |h_i^L|$$

such that the weights becomes $h_i = l_L \frac{h_i^L}{norm_L}$ for $i \in L$ with the sum of $|h_i|$ equal to l_L. The same technique is used to compute the modified weights h_i^S for $i \in S$. Replacing the weights h_i with their values in the beta-neutral portfolio constraint, we get

$$\sum_{i\in L} h_i\beta_i = \frac{l_L}{norm_L}\sum_{i\in L} h_i^L\beta_i = 0$$
$$\sum_{i\in S} h_i\beta_i = \frac{l_S}{norm_S}\sum_{i\in S} h_i^S\beta_i = 0$$

Thus, with one modified weight fixed and the rest chosen to satisfy a particular allocation we can obtain a beta-neutral portfolio independently from the leverage chosen. For instance, the modified weights can be chosen to match the information ratio $\frac{r_i}{\sigma_i^2}$.

9.5 Value at Risk

We are now introducing Value at Risk in the case of portfolio theory and we will discuss Value at Risk for option theory in Section ().

9.5.1 Defining value at risk

9.5.1.1 Some terminology

We will see in Chapter (10) that extreme price movements in the financial markets are not so rare, and do not necessarily correspond to big financial crises. Nonetheless, these large price fluctuations have been associated to risky events and numerous measures of market risk developed such as the value at risk (VaR). The VaR belongs to the class of Monetary risk measures quantifying the risk as a dollar amount and can be viewed as regulatory capital required to cover risk. Various methods for calculating VaR were proposed based on different statistical theories, among which are RiskMetrics. Duffie et al. [1997] described VaR as a single estimate of the amount by which an institution's position in a risk category could decline due to general market movements during a given holding period. Being defined as the maximal loss of a financial position during a given time period for a given probability, VaR is generally used to ensure that a financial institution can still be in business after a catastrophic event. Alternatively, a regulatory committee can define VaR as the minimal loss under extraordinary market circumstances, leading to the same measure. We let $\Delta V(d)$

be the change in value of the assets in a portfolio V from time t to $t+d$ measured in dollars. We let $F_d(x)$ [2] be the cumulative distribution function (CDF) of $\Delta V(d)$ and define the VaR of a long position (loss when $\Delta V(d) < 0$) over the time horizon d with probability p as

$$p = P(\Delta V(d) \leq Var) = F_d(Var) \tag{9.5.25}$$

assuming a negative value for Var when p is small (left side of the distribution). That is, the negative sign means a loss. Hence, the probability that the holder would encounter a loss greater than or equal to VaR over the time horizon d is p. Equivalently, with probability $(1-p)$, the potential loss encountered by the holder of the portfolio over the time period d is less than or equal to VaR. In the case of a short position (loss when $\Delta V(d) > 0$) the Var becomes

$$p = P(\Delta V(d) \geq Var) = 1 - P(\Delta V(d) \leq Var)$$

assuming a positive value for Var when p is small (right side of the distribution). That is, the positive sign means a loss. This definition of the VaR shows that it is concerned with the tail behaviour of the CDF, $F_d(x)$, where the left tail is important for a long position, and the right tail for a short one. Hence, we could use Equation (9.5.25) in the case of a short position by using the distribution of $-\Delta V(d)$. For any univariate CDF, $F_d(x)$, and probability, p, such that $0 < p < 1$, the quantity

$$x_p = \inf\{x \in \mathbb{R} | F_d(x) \geq p\} = F_d^{-1}(p)$$

is called the pth quantile of $F_d(x)$ [3], where inf denotes the smallest real number satisfying $F_d(x) \geq p$. Therefore, for known CDF, $F_d(x)$, the VaR is simply its pth quantile, that is, $VaR = x_p$ or $VaR_p(X) = F^{-1}(p)$. Note, the probability level is about equally often specified as one minus the probability of a VaR break, $(1-\alpha)$, where α is a confidence level (usually 99% or 95%). Further, even though the Var should always represents a loss, it is conventionally reported as a positive number. That is, letting L be the loss of a portfolio, then the VaR of the portfolio at confidence level α is given by the smallest number l such that the probability that the loss L exceeds l is at most $(1-\alpha)$. It is expressed as

$$VaR_\alpha(L) = \inf\{l \in \mathbb{R} | P(L > l) \leq 1 - \alpha\}$$

Hence, assuming a probability distribution, the VaR is the opposite of the $(1-\alpha)$ quantile of the profit and loss

$$VaR_\alpha(X) = -\inf\{x \in \mathbb{R} | F_d(x) \geq 1-\alpha\} = -F_d^{-1}(1-\alpha)$$

When computing the VaR we need to specify the probability of interest p ($p = 0.01$ or $p = 0.05$), the time horizon d (1 day or 10 days), the frequency of the data, the amount of the financial position or the mark-to-market value of the portfolio.

9.5.1.2 The normal assumption

In the special case where the returns of the portfolio R_P are normally distributed with mean μ and standard deviation σ, we can apply the standard transformation $z_p = \frac{Z-\mu}{\sigma}$ with $Z = -VaR$ and $z_p = -z_p$, getting the VaR number

$$VaR(p) = z_p\sigma - \mu$$

For $z \sim N(0,1)$, we are interested in $1-\alpha = P(z \leq z_p)$, and we can read in the standard normal table the number z_p corresponding to the confidence level α. For example, given $\alpha = 95\%$ we get $z_p = 1.65$, and for $\alpha = 99\%$ we get $z_p = 2.33$. The value z_p is the number of standard deviation at $(1-\alpha)$ such that $F(-VaR) = 1-\alpha$ (see Jorion [2001]). Assuming a portfolio with initial price P_0 and rate of return R being normally distributed with mean

[2] Given a random variable X, we get the distribution function $F(x) = P(X \leq x)$ with $F(x) = \int_{-\infty}^{x} f(u)du$ where f is the probability density function of X.

[3] $F^{-1}(u) = \inf\{x \in \mathbb{R} | F(x) \geq u\}$ is the left-continuous generalised inverse of F.

μ and standard deviation σ. Then, after one periode of time the portfolio is $P_1 = P_0(1+R)$ with mean $P_0(1+\mu)$ and standard deviation $P_0\sigma$. We let $P_1^* = P_0^*(1+R^*)$ be the lowest portfolio value at some confidence level α and compute the VaR relative to the expected return as

$$VaR_\mu = E[P_1] - P_1^* = P_0(\mu - R^*)$$

Assuming $\mu = 0$, we get $VaR_0 = -P_0R^*$. Further, letting $R^* = -(z_p\sigma - \mu)$, we get

$$VaR_\mu = P_0 z_p \sigma \, , VaR_0 = P_0(z_p\sigma - \mu)$$

so that for known portfolio value the only variable is the standard deviation of the rate of return. We let $R_P = \sum_{i=1}^N w_i R_i$ be the rate of return of a portfolio with N assets, where R_i is the rate of return of the ith asset, $w_i = \frac{S_i}{P_0}$ is the ith weight with P_0 the porfolio value and S_i the value of the ith asset. If we let w be thecolumn vector of weight and R be the column vector of rate of return, we can rewrite the portfolio return as $R_P = w^\top R$. Since a linear sum of normal variables is normally distributed, then the expected portfolio rate of return μ_P and its variance σ_P^2 satisfies

$$\mu_P = \sum_{i=1}^N w_i\mu_i \text{ and } \sigma_P^2 = \sum_{i,j}^N w_i w_j \sigma_{ij}$$

where $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ is the (i,j) element of the covariance matrix. Using matrix representation, the variance of the rate of return of the portfolio is written as $\sigma_P^2 = w^\top Cw$ where C is the covariance matrix. Assuming $\mu_i = 0$ for $i = 1, .., N$, we get $\sigma_{ij} = Cov(R_i, R_j) = E[R_iR_j]$ and as a result $\sigma_{ij} = \sigma_{ji}$ and $\sigma_{ii} = \sigma_i^2$ so that

$$\sigma_P^2 = \sum_{i=1}^N w_i^2\sigma_i^2 + 2\sum_{i=1}^N \sum_{j<i} w_i w_j \sigma_{ij}.$$

The VaR of a multiple assets portfolio is given by

$$VaR_P = z_p\sigma_P P_0 = z_p\sqrt{X^\top CX}$$

where X is a column vector of asset values S_i for $i = 1, .., N$. This comes from the weight $w_i = \frac{S_i}{P_0}$, such that

$$VaR_P = z_p\sqrt{\sum_{i=1}^N (\frac{S_i}{P_0})^2\sigma_i^2P_0^2 + 2\sum_{i=1}^N\sum_{j<i}\frac{S_iS_j}{P_0^2}\sigma_{ij}P_0^2} = z_p\sqrt{\sum_{i=1}^N S_i^2\sigma_i^2 + 2\sum_{i=1}^N\sum_{j<i} S_iS_j\sigma_{ij}}$$

As a result, we can rewrite the VaR of the portfolio in terms of the VaR of its components as follow

$$VaR_P = \sqrt{\sum_{i=1}^N VaR_i^2 + 2\sum_{i=1}^N\sum_{j<i} VaR_iVaR_j\rho_{ij}}$$

Hence, lower portfolio risk can be achieved through low correlation or large number of assets.

9.5.2 Computing value at risk

Since the CDF is unknown in practice, VaR is essentially concerned with the estimation of the CDF, or its quantile, especially the tail behaviour of the CDF. Consquently, econometric modelling was used to forecast the CDF, where different methods for estimating the CDF gave rise to different approaches for calculating the VaR. Note, given log returns $\{r_t\}$, the VaR calculated from the quantile of the distribution of r_{t+1} given time t is in percentage so that the dollar amount of VaR is the cash value of the portfolio times the VaR of the log return series. While VaR is a prediction concerning possible loss of a portfolio in a given time period, it should be computed by using the predictive distribution of future returns. But, in statistics, predictive distribution takes into account the parameter uncertainty in a properly specified model, which is largely ignored in the available methods for VaR calculation.

9.5.2.1 RiskMetrics

RiskMetrics, developed by Longerstaey et al. [1995b], assumes that the continuously compounded daily return of a portfolio follows a conditional normal distribution $r_t|\mathcal{F}_{t-1} \sim N(\mu_t, \sigma_t^2)$ where μ_t is the conditional mean and σ_t^2 is the conditional variance of r_t. Further, the method assumes that the two quantities evolve over time as

$$\mu_t = 0 \, , \sigma_t^2 = \alpha\sigma_{t-1}^2 + (1-\alpha)r_{t-1}^2 \, , 1 > \alpha > 0 \tag{9.5.26}$$

The method assumes that the logarithm of the daily price, $p = \ln(P_t)$, of the portfolio satisfies the difference equation $p_t - p_{t-1} = a_t$, where $a_t = \sigma_t\epsilon_t$ is an $IGARCH(1,1)$ process without drift. The value of α is usually taken in the range $[0.9, 1]$. For a k-period horizon, the log return from time $t+1$ to time $t+k$ (inclusive) is $r_t(k) = r_{t+1} + ... + r_{t+k-1} + r_{t+k}$. In this model the conditional distribution $r_t(k)|\mathcal{F}_t$ is normal with mean zero and variance $\sigma_t^2(k)$ which can be computed by using the forecasting method presented in Section (5.3.3). Given the assumption of independence of ϵ_t, we have

$$\sigma_t^2 = Var(r_t(k)|\mathcal{F}_t) = \sum_{i=1}^{k} Var(a_{t+i}|\mathcal{F}_t)$$

where $Var(a_{t+i}|\mathcal{F}_t) = E[\sigma_{t+i}^2|\mathcal{F}_t]$ can be obtained recursively. Using $r_{t-1} = a_{t-1} = \sigma_{t-1}\epsilon_{t-1}$, we can rewrite the volatility equation of the $IGARCH(1,1)$ model as

$$\sigma_t^2 = \sigma_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2(\epsilon_{t-1} - 1) \, , \forall t$$

In particular, we have

$$\sigma_{t+i}^2 = \sigma_{t+i-1}^2 + (1-\alpha)\sigma_{t+i-1}^2(\epsilon_{t+i-1} - 1) \text{ for } i = 2, .., k$$

Since $E[(\epsilon_{t+i-1} - 1)|\mathcal{F}_t] = 0$ for $i \geq 2$, the prior equation shows that

$$E[\sigma_{t+i}|\mathcal{F}_t] = E[\sigma_{t+i-1}|\mathcal{F}_t] \text{ for } i = 2, .., k$$

For the 1-step ahead volatility forecast, Equation (9.5.26) shows that $\sigma_{t+1}^2 = \alpha\sigma_t^2 + (1-\alpha)r_t^2$, so that the above equation shows that $Var(r_{t+i}|\mathcal{F}_t) = \sigma_{t+1}^2$ for $i \geq 1$ and hence $\sigma_t^2(k) = k\sigma_{t+1}^2$ such that $r_t(k)|\mathcal{F}_t \sim N(0, k\sigma_{t+1}^2)$. Therefore, in the $IGARCH(1,1)$ model given in Equation (9.5.26), the conditional variance of $r_t(k)$ is proportional to the horizon time k, and the conditional standard deviation of a k-period horizon log return is $\sqrt{k}\sigma_{t+1}$. In the case of a long position with probability set to 5%, the RiskMetrics uses $1.65\sigma_{t+1}$ to measure the risk of the portfolio, that is, it uses the one-sided 5% quantile of a normal distribution with mean zero and standard deviation σ_{t+1} (see Section (9.5.1.2)). Note, the 5% quantile is actually $-1.65\sigma_{t+1}$, but the negative sign is ignored with the understanding that it is a loss. As a result, in the RiskMetrics model, when the standard deviation is measured in percentage, then the daily VaR of the portfolio becomes

$$VaR = W \times 1.65\sigma_{t+1}$$

where W is the amount of the position, and the VaR of a k-day horizon is

$$VaR(k) = W \times 1.65\sqrt{k}\sigma_{t+1}$$

Therefore, we can write the VaR as

$$VaR(k) = \sqrt{k} \times VaR$$

which is the square root of time rule in VaR calculation under RiskMetrics. In the case where the investor holds multiple positions, assuming that daily log returns of each position follow a random-walk $IGARCH(1,1)$ model we

need to estimate the cross-correlation coefficients between the returns. In the case of two positions with VaR being VaR_1 and VaR_2, respectively, and cross-correlation coefficient being

$$\rho_{12} = \frac{Cov(r_{1t}, r_{2t})}{\sqrt{Var(r_{1t})Var(r_{2t})}}$$

then the overall VaR of the investor is

$$VaR = \sqrt{VaR_1^2 + VaR_2^2 + 2\rho_{12}VaR_1VaR_2}$$

In the case of m assets the VaR is generalised to

$$VaR = \sqrt{\sum_{i=1}^{m} VaR_i^2 + 2\sum_{i<j}^{m} \rho_{ij}VaR_iVaR_j}$$

where ρ_{ij} is the cross-correlation coefficient between returns of the ith and jth assets and VaR_i is the VaR of the ith asset.

9.5.2.2 Econometric models to VaR calculation

Note, if either the zero mean assumption or the special $IGARCH(1,1)$ model assumption of the log returns fails, then the rule is invalid. For many heavily traded stocks the assumption that $\mu \neq 0$ holds, so that we can consider the simple model

$$\begin{aligned} r_t &= \mu + a_t \text{ , } a_t = \sigma_t\epsilon_t \text{ , } \mu \neq 0 \\ \sigma_t^2 &= \alpha\sigma_{t-1}^2 + (1-\alpha)a_{t-1}^2 \end{aligned}$$

where $\{\epsilon_t\}$ is a standard Gaussian white noise series. In that model, the distribution of r_{t+1} given $\mathcal{F}_t$ is $N(\mu, \sigma_{t+1}^2)$ and the 5% quantile used to calculate the 1-period horizon VaR becomes $\mu - 1.65\sigma_{t+1}$. For a k-period horizon, the distribution of $r_t(k)$ given $\mathcal{F}_t$ is $N(k\mu, k\sigma_{t+1}^2)$ and the 5% quantile used in k-period horizon VaR calculation is $k\mu - 1.65\sqrt{k}\sigma_{t+1} = \sqrt{k}(\sqrt{k}\mu - 1.65\sigma_{t+1})$ such that $VaR(k) \neq \sqrt{k} \times VaR$ when the mean return is different from zero. This rule also fails when the volatility model of the return is not an $IGARCH(1,1)$ without a drift. More generally, one can use the time series econometric models (linear) where ARMA models can be used to model the mean equation, and GARCH models can be used to model the volatility. For instance, given the log return r_t, a general model can be written as

$$\begin{aligned} r_t &= \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j} \\ a_t &= \sigma_t\epsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{u} \alpha_i a_{t-i}^2 + \sum_{j=1}^{v} \beta_j \sigma_{t-j} \end{aligned}$$

Assuming known parameters, these equations can be used to obtain 1-step ahead forecasts of the conditional mean and the conditional variance of r_t. That is,

$$\begin{aligned}\hat{r}_t(1) &= \phi_0 + \sum_{i=1}^{p} \phi_i r_{t+1-i} - \sum_{j=1}^{q} \theta_j a_{t+1-j} \\ \hat{\sigma}_t^2(1) &= \alpha_0 + \sum_{i=1}^{u} \alpha_i a_{t+1-i}^2 + \sum_{j=1}^{v} \beta_j \sigma_{t+1-j}\end{aligned}$$

Assuming further that ϵ_t is Gaussian, the conditional distribution of r_{t+1} given $\mathcal{F}_t$ is $N(\hat{r}_t(1), \hat{\sigma}_t^2(1))$ and the 5% quantile is $\hat{r}_t(1) - 1.65\hat{\sigma}_t^2(1)$. If we assume that ϵ_t is a standardised Student-t distribution with v degrees of freedom, then the quantile is $\hat{r}_t(1) - t_v^*(p)\hat{\sigma}_t^2(1)$ where $t_v^*(p)$ is the pth quantile of a standardised Student-t distribution with v degrees of freedom. Denoting $t_v(p)$ the pth quantile of a Student-t distribution with v degrees of freedom, we get

$$p = P(t_v \leq q) = P(\frac{t_v}{\sqrt{\frac{v}{(v-2)}}} \leq \frac{q}{\sqrt{\frac{v}{(v-2)}}}) = P(t_v^* \leq \frac{q}{\sqrt{\frac{v}{(v-2)}}})$$

where $v > 2$. It says that if q is the pth quantile of a Student-t distribution with v degrees of freedom, then $\frac{q}{\sqrt{\frac{v}{(v-2)}}}$ is the pth quantile of the standardised distribution. Hence, if ϵ_t of the GARCH model for σ_t^2 is a standardised Student-t distribution with v degrees of freedom and the probability is p, the quantile used to calculate the 1-period horizon VaR at time t is

$$\hat{r}_t(1) - \frac{t_v(p)\hat{\sigma}_t(1)}{\sqrt{\frac{v}{(v-2)}}}$$

We are now considering the k-period log return at the forecast origin h, that is, $r_h(k) = r_{h+1} + ... + r_{h+k}$. We can then use the appropriate forecasting methods to obtain the conditional mean and variance of $r_h(k)$. The conditional mean $E[r_h(k)|\mathcal{F}_h]$ is obtained by using the forecasting method of ARMA models

$$\hat{r}_h(k) = r_h(1) + .. + r_h(k)$$

where $r_h(l)$ is the l-step ahead forecast of the return at the forecast origin h, which can be computed recursively. Using the MA representation of the ARMA model, we can rewrite the l-step ahead forecast error at the forecast origin h as

$$e_h(l) = r_{h+l} - r_h(l) = a_{h+l} + \psi_1 a_{h+l-1} + ... + \psi_{l-1} a_{h+1}$$

The forecast error of the expected k-period return $\hat{r}_h(k)$ is the sum of 1-step to k-step forecast errors of r_t at the forecast origin h and can be written as

$$e_h(k) = e_h(1) + .. + e_h(k) = a_{h+k} + (1+\psi_1)a_{h+k-1} + ... + (\sum_{i=0}^{k-1} \psi_i) a_{h+1}$$

where $\psi_0 = 1$. The volatility forecast of the k-period return at the forecast origin h is the conditional variance of $e_h(k)$ given $\mathcal{F}_h$. Using the independent assumption of ϵ_{t+i} for $i = 1, .., k$ where $a_{t+i} = \sigma_{t+i}\epsilon_{t+i}$, we have

$$\begin{aligned}Var(e_h(k)|\mathcal{F}_h) &= Var(a_{h+k}|\mathcal{F}_h) + (1+\psi_1)^2 Var(a_{h+k-1}|\mathcal{F}_h) + ... + (\sum_{i=0}^{k-1} \psi_i)^2 Var(a_{h+1}|\mathcal{F}_h) \\ &= \sigma_h^2(k) + (1-\psi_1)^2 \sigma_h^2(k-1) + ... + (\sum_{i=0}^{k-1} \psi_i)^2 \sigma_h^2(1)\end{aligned}$$

where $\sigma_h^2(l)$ is the l-step ahead volatility forecast at the forecast origin h. In the GARCH model these volatility forecasts can be obtained recursively. For instance, in the simple model

$$r_t = \mu + a_t \, , a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

we have $\psi_i = 0$ for all $i > 0$. The point forecast of the k-period return at the forecast origin h is $\hat{r}_h(k) = k\mu$ and the associated forecast error is

$$e_h(k) = a_{h+k} + a_{h+k-1} + ... + a_{h+1}$$

As a result, the volatility forecast for the k-period return at the forecast origin h is

$$Var(e_h(k)|\mathcal{F}_h) = \sum_{l=1}^{k} \sigma_h^2(l)$$

Using the forecasting method of $GARCH(1,1)$ models, we have

$$\begin{aligned} \sigma_h^2(1) &= \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2 \\ \sigma_h^2(l) &= \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(l-1) \, , l = 2, .., k \end{aligned}$$

so that $Var(e_h(k)|\mathcal{F}_h)$ can be obtained by the prior recursion. If ϵ_t is Gaussian, then the conditional distribution of $r_h(k)$ given $\mathcal{F}_h$ is normal with mean $k\mu$ and variance $Var(e_h(k)|\mathcal{F}_h)$ and one can compute the quantiles needed in VaR calculation.

9.5.2.3 Quantile estimation to VaR calculation

While quantile estimation provides a nonparametric approach to VaR calculation by making no specific distributional assumption on the return of a portfolio, it assumes that the distribution continues to hold within the prediction period. Two types of quantile methods exict, the first one directly using empirical quantile, and the second one using quantile regression. We consider the collection of n returns $\{r_1, .., r_n\}$ where the minimum return is $r^{(1)}$, the smallest order statistic, and the maximum return is $r^{(n)}$, the maximum order statistic. That is, $r^{(1)} = \min_{1 \leq j \leq n}\{r_j\}$ and $r^{(n)} = \max_{1 \leq j \leq n}\{r_j\}$. Since we have $r^{(n)} = -\min_{1 \leq j \leq n}\{-r_j\} = -r_c^{(1)}$, where $r_{t,c} = -r_t$ with the subscript c denoting sign change, we can concentrate on the minimum return $r^{(1)}$. We further assume that the returns are i.i.d. random variables having a continuous distribution with probability density function (pdf) $f(x)$ and CDF $F(x)$. Then there exists well known asymptotic result for the order statistic $r^{(l)}$ where $l = np$ with $0 < p < 1$ (see Cox et al. [1974]). We let x_p be the pth quantile of $F(x)$, that is $x_p = F^{-1}(p)$ and assume that $f(x)$ is not zero at x_p. Then the order statistic $r^{(l)}$ is asymptotically normal with mean x_p and variance $\frac{p(1-p)}{nf^2(x_p)}$, that is,

$$r^{(l)} \sim N(x_p, \frac{p(1-p)}{nf^2(x_p)}) \, , l = np$$

and it can be used for estimating the quantile x_p. In practice, the probability p may not satisfy np to be a positive integer, and simple interpolation must be used. For noninteger np, we let l_1 and l_2 be two neighbouring positive integers such that $l_1 < np < l_2$ and define $p_i = \frac{l_i}{n}$. Hence, $r^{(l_i)}$ is a consistent estimate of the quantile x_{p_i} since $p_1 < p < p_2$. As a result, the quantile x_p can be estimated by

$$\hat{x}_p = \frac{p_2 - p}{p_2 - p_1} r^{l_1} + \frac{p - p_1}{p_2 - p_1} r^{l_2}$$

Even though this method is simple and use no specific distributional assumption, it assumes stationarity of the distribution. Hence, in the case of tail probability, it implies that the predicted loss can not be greater than the historical one. Further, for extreme quantiles (when p is close to zero or unity), the empirical quantiles are not efficient estimates of the theoretical quantiles. At last, direct quantile estimation fails to take into account the effect of explanatory variables relevant to the portfolio under study. In general, the empirical quantile serves as a lower bound for the actual VaR. In real application, one can rely on explanatory variables having important impacts on the dynamics of asset returns, and should therefore consider the conditional distribution function $r_{t+1}|\mathcal{F}_t$ where the filtration $\mathcal{F}_t$ includes the explanatory variables. The associated quantiles of this distribution are refered to regression quantiles (see Koenker et al. [1978]). We can cast the empirical quantile as an estimation problem where for a given probability p, the pth quantile of $\{r_t\}$ is given by

$$\hat{x}_p = \arg\min_{\beta} \sum_{i=1}^{n} w_p(r_i - \beta)$$

where $w_p(z)$ is defined by

$$w_p(z) = \begin{cases} pz \text{ if } z \geq 0 \\ (p-1)z \text{ otherwise} \end{cases}$$

The regression quantile generalise such an estimate by considering the linear regression

$$r_t = \beta^\top x_t + a_t$$

where β is a k-dimensional vector of parameters and x_t is a vector of predictors that are elements of $\mathcal{F}_{t-1}$. Since $\beta^\top x_t$ is known, then the conditional distribution of r_t given F_{t-1} becomes a translation of the distribution of a_t. Koenker et al. [1978] estimated the conditional quantile $x_p|\mathcal{F}_{t-1}$ of r_t as

$$\hat{x}_p|\mathcal{F}_{t-1} = \inf\{\beta_0^\top x | R_p(\beta_0) = min\}$$

where $R_p(\beta_0) = min$ means that β_0 is obtained by solving

$$\beta_0 = \arg\min_{\beta} \sum_{i=1}^{n} w_p(r_i - \beta^\top x_i)$$

where $w_p(z)$ is defined as before.

9.5.2.4 Extreme value theory to VaR calculation

We assume that returns r_t are serially independent with a common cumulative distribution function $F(x)$ and that the range of the returns is $[l, u]$, where $l = -\infty$ and $u = \infty$ for log returns. Then, the CDF of $r^{(1)}$, denoted by $F_{n,1}(x)$, is given by

$$F_{n,1}(x) = P(r^{(1)} \leq x) = 1 - P(r^{(1)} > x) = 1 - P(r_1 > x, ..., r_n > x)$$

From independence we have

$$F_{n,1}(x) = 1 - \prod_{j=1}^{n} P(r_j > x) = 1 - \prod_{j=1}^{n} \big(1 - P(r_j \leq x)\big)$$

and by common distribution we get

$$F_{n,1}(x) = 1 - \prod_{j=1}^{n} \big(1 - F(x)\big) = 1 - \big(1 - F(x)\big)^n$$

Since the CDF $F(x)$ of r_t is unknown, then $F_{n,1}(x)$ of $r^{(1)}$ is unknown, but as n increases to infinity, $F_{n,1}(x)$ becomes degenerated, that is,

$$F_{n,1}(x) \to \begin{cases} 0 \text{ if } x \leq l \\ 1 \text{ otherwise} \end{cases} \quad \text{as } n \to \infty$$

Since this degenerated CDF has no practical value, the extreme value theory (EVT) is concerned with finding two sequences $\{\alpha_n\}$ and $\{\beta_n\}$, where the former is a series of scaling factors and the latter is a location series with $\alpha_n > 0$ and such that the distribution of $r^{(1_*)} = \frac{(r^{(1)}-\beta_n)}{\alpha_n}$ converges to a nondegenerated distribution as n goes to infinity. Under the independent assumption, the limiting distribution of the normalised minimum $r^{(1_*)}$ is given by

$$F_*(x) = \begin{cases} 1 - e^{-(1+kx)^{\frac{1}{k}}} \text{ if } k \neq 0 \\ 1 - e^{-e^x} \text{ if } k = 0 \end{cases}$$

for $x < -\frac{1}{k}$ if $k < 0$ and for $x > -\frac{1}{k}$ if $k > 0$, where the subscript $*$ signifies the minimum. The case of $k = 0$ is taken as the limit when $k \to 0$, the parameter k is called the shape parameter governing the tail behaviour of the limiting distribution, and the parameter $\alpha = -\frac{1}{k}$ is the tail index of the distribution. The limiting distribution of the above equation is the generalised extreme value distribution of Jenkinson [1955] for the minimum encompassing the three types of limiting distribution of Gnedenko [1943]

1. $k = 0$, the Gumbel family with CDF

$$F_*(x) = 1 - e^{-e^x}, \ -\infty < x < \infty$$

2. $k < 0$, the Frechet family with CDF

$$F_*(x) = \begin{cases} 1 - e^{-(1+kx)^{\frac{1}{k}}} \text{ if } x < -\frac{1}{k} \\ 1 \text{ otherwise} \end{cases}$$

3. $k > 0$, the Weibull family with CDF

$$F_*(x) = \begin{cases} 1 - e^{-(1+kx)^{\frac{1}{k}}} \text{ if } x > -\frac{1}{k} \\ 0 \text{ otherwise} \end{cases}$$

Gnedenko [1943] gave necessary and sufficient conditions for the CDF $F(x)$ of r_t to be associated with one of the three types of limiting distribution. The tail behaviour of $F(x)$ determines the limiting distribution $F_*(x)$ of the minimum. The (left) tail of the distribution declines exponentially for the Gumbel family, by a power function for the Frechet family, and is finite for the Weibull family. For risk management, we are mainly interested in the Frechet family that includes stable and Student-t distributions. The Gumbel family consists of thin-tailed distributions such as normal and lognormal distributions. The probability density function (pdf) of the generalised limiting distribution in Equation (9.5.27) can be obtained by differentiation

$$f_*(x) = \begin{cases} (1+kx)^{\frac{1}{k}-1} e^{-(1+kx)^{\frac{1}{k}}} \text{ if } k \neq 0 \\ e^{x-e^x} \text{ if } k = 0 \end{cases}$$

where $-\infty < x < \infty$ for $k = 0$, $x < -\frac{1}{k}$ for $k < 0$, and $x > -\frac{1}{k}$ for $k > 0$. The two important implications of extreme value theory are

1. the tail behaviour of the CDF $F(x)$ of r_t, not the specific distribution, determines the limiting distribution $F_*(x)$ of the (normalised) minimum, making the theory applicable to a wide range of distributions for the return r_t.

2. Feller [1971] showed that the tail index k does not depend on the time interval of r_t, meaning the tail index (or equivalently the shape parameter) is invariant under time aggregation, being handy in the VaR calculation.

The extreme value theory was extended to serially dependent observations $\{r_t\}_{t=1}^n$ provided that the dependence is weak. Berman [1964] showed that the same form of the limiting extreme value distribution holds for stationary normal sequences provided that the autocorrelation function of r_t is squared summable, that is $\sum_{i=1}^{\infty} \rho_i^2 < \infty$, where ρ_i is the lag-i autocorrelation function of r_t.

We saw that the extreme value distribution contains three parameters k, α_n, and β_n called the shape, scale parameter and location, respectively, and can be estimated by using parametric or nonparametric methods. Since for a given sample one only has a single minimum or maximum, we must divide the sample into subsamples and apply the EVT to the subsamples. Assuming T returns $\{r_j\}_{j=1}^T$, we divide the sample into g non-overlapping subsamples each with n observations (T=ng)

$$\{r_1, .., r_n | r_{n+1}, .., r_{2n} | ... | r_{(g-1)n+1}, .., r_{ng}\}$$

and write the observed returns as r_{in+j} where $1 \leq j \leq n$ and $i = 0, .., g-1$. When n is sufficiently large, we hope that the EVT applies to each subsample. Hence, we let $r_{n,i}$ be the minimum of the ith subsample, so that when n is sufficiently large $x_{n,i} = \frac{(r_{n,i}-\beta_n)}{\alpha_n}$ should follow an extreme value distribution, and the collection of subsample minima $\{r_{n,i}\}_{i=1}^g$ can then be regarded as a sample of g observations from that extreme value distribution. We can define

$$r_{n,i} = \min_{1 \leq j \leq n} \{r_{(i-1)n+j}\}, i = 1, .., g$$

which clearly depends on the choice of the subperiod length n. Alternatively, we can use a parametric approach such as the maximum likelihood method. Assuming that the pdf of $x_{n,i}$ is given above, the pdf of $r_{n,i}$ is obtained by a simple transformation as

$$f(r_{n,i}) = \begin{cases} \frac{1}{\alpha_n}(1 + k_n \frac{(r_{n,i}-\beta_n)}{\alpha_n})^{\frac{1}{k_n}-1} e^{-(1+\frac{k_n(r_{n,i}-\beta_n)}{\alpha_n})^{\frac{1}{k_n}}} & \text{if } k_n \neq 0 \\ \frac{1}{\alpha_n} e^{\frac{(r_{n,i}-\beta_n)}{\alpha_n} - e^{\frac{(r_{n,i}-\beta_n)}{\alpha_n}}} & \text{if } k_n = 0 \end{cases}$$

where $1 + k_n \frac{(r_{n,i}-\beta_n)}{\alpha_n} > 0$ if $k_n \neq 0$. The subscript n added to the shape parameter k means that its estimate depends on the choice of n. Under the assumption of independence, the likelihood function of the subperiod minima is

$$l(r_{n,1}, .., r_{n,g} | k_n, \alpha_n, \beta_n) = \prod_{i=1}^{g} f(r_{n,i})$$

and nonlinear estimation procedures can be used to obtain maximum likelihood estimates of k_n, α_n, and β_n, These estimates are unbiased, asymptotically normal, and of minimum variance under proper assumptions.

9.5.3 The conditional Value at Risk

Artzner et al. [1999] defined a set of properties that must be possessed by a risk measure in order to compute regulatory capital in a sensible risk management system. We let $\mathcal{X}$ be the linear space of possible payoffs containing the constants and define a coherent risk measure as follow

Definition 9.5.1 *A mapping $\rho : \mathcal{X} \to \{+\infty\}$ is called a coherent risk measure if it possesses the following properties*

- *Monotonocity: $X \leq Y$ implies $\rho(X) \geq \rho(Y)$.*
- *Cash invariance: for all $m \in \mathbb{R}$, $\rho(X + m) = \rho(X) - m$.*

- *Subadditivity:* $\rho(X+Y) \geq \rho(X) + \rho(Y)$.
- *Positive homogeneity:* $\rho(\lambda X) = \lambda\rho(X)$ *for* $\lambda \geq 0$.

Note, the positive homogeneity property does not take into account the liquidity risk associated with liquidation costs of large portfolios. Further, under the positive homogeneity property, the subadditivity is equavalent to convexity. Even though VaR possesses the monotonicity, the cash invariance and the positive homogeneity properties, it is not subadditive, and as a result it is not a coherent risk measure. The smallest coherent risk measure dominating the VaR is called the Conditional VaR, or expected shortfall, and is defined as the average VaR with confidence levels between α and 1. It is given by

$$ES_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 VaR_\alpha(X)d\alpha$$

We are now going to discuss the interpretation of ES_α as the expectation of losses in excess of VaR.

Property 9.5.1 *The expected shortfall admits the probabilistic representation*

$$ES_\alpha(X) = VaR_\alpha(X) + \frac{1}{1-\alpha}E[(-VaR_\alpha(X) - X)^+]$$

If the distribution function of X*, denoted by* $F(x)$*, is continuous then in addition*

$$ES_\alpha(X) = -E[X|X < -VaR_\alpha(X)]$$

We are now going to establish a dual representation showing that unlike the VaR, the expected shortfall is convex making it a coherent risk measure.

Property 9.5.2 *The expected shortfall admits the representation*

$$ES_\alpha(X) = \sup\{E^Q[-X] : \mathbb{Q} \in \mathbb{Q}_{1-\alpha}\}$$

where $\mathbb{Q}_{1-\alpha}$ *is the set of probability measures on* $\mathcal{X}$ *satisfying* $\frac{dQ}{dP} \leq \frac{1}{1-\alpha}$.

The expected shortfall encourages diversification, presents computational advantages compared to VaR and does not allow for regulatory arbitrage. It is a more conservative risk measure than VaR, implying higher costs for the bank or hedge fund in terms of regulatory capital.

Part IV

Quantitative trading in multifractal markets

Chapter 10

The fractal market hypothesis

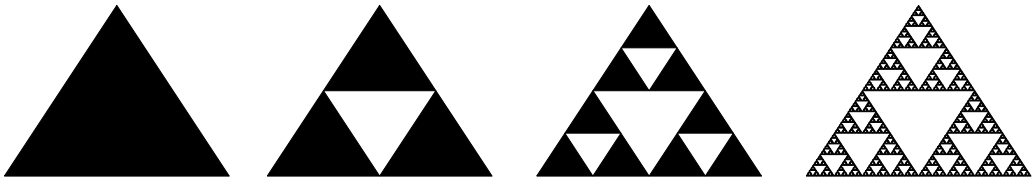
10.1 Fractal structure in the markets

10.1.1 Introducing fractal analysis

10.1.1.1 A brief history

The Hurst Exponent (H) proposed by Hurst [1951] and tested by Hurst et al. [1965] relates to the autocorrelations of the time series, and the rate at which these decrease as the lag between pairs of values increases. It is used as a measure of long-term memory of time series, classifying the series as a random process, a persistent process, or an anti-persistent process. Hurst measured how a reservoir level (Nile River Dam project) fluctuated around its average level over time, and found that the range of the fluctuation would change, depending on the length of time used for measurement. If the series were random, the range would increase with the square root of time $T^{\frac{1}{2}}$. To standardise the measure over time, he created a dimensionless ratio by dividing the range by the standard deviation of the observation, obtaining the rescaled range analysis (R/S analysis). He found that most natural phenomena follow a biased random walk, that is, a trend with noise which could be measured by how the rescaled range scales with time, or how high the exponent H is above $\frac{1}{2}$.

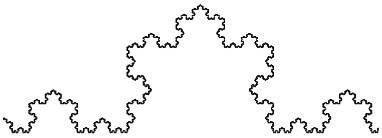
In 1975 Mandelbrot (see Mandelbrot [1975] [2004]) introduced the term fractal as a geometric shape that is self-similar and has fractional dimensions, leading to the Mandelbrot set in 1977 and 1979. An object for which its topological dimension is lower or equal to its Hausdorff-Besicovitch (H-B) dimension is a fractal. By doing so, he was able to show how visual complexity can be created from simple rules. A more general definition states that a fractal is a shape made of parts similar to the whole in some way (see Feder [1988]). Two examples of fractal shapes are the Sierpinski Triange (see Figure (10.1)) and Koch Curve (see Figure (10.2)). The above Figures are bi-dimensional and deterministic fractal. Focusing on financial time series, Figure (10.3) illustrates a one-dimensional non-deterministic fractal representing thirty returns from the CAC40 Index at different time scales. Without any labels on the $X - Y$ scales, it is impossible to know which one is daily, weekly, or monthly returns. The fractal dimension D measures the smoothness of a surface, or, in our case, the smoothness of a time series. As discussed by Peters [1994], the importance of the fractal dimension of a time series is that it recognises that a process can be somewhere between deterministic (a line with fractal dimension of 1) and random (a fractal dimension of 1.50).



(a) Initial triangle (b) Second step (c) Third step (d) Few iterations

Figure 10.1: Sierpinski Triangle

(a) Initial line (b) Second step (c) Third step



(d) Few iterations

Figure 10.2: Koch Curve

Fractal analysis is done by conducting rescaled adjusted range analysis (R/S analysis) of time series (see Mandelbrot [1975b]), where the Hurst Exponent and the fractal dimension of the series are estimated. A major application of the R/S analysis is the study of price fluctuations in financial markets. There, the value of the Hurst Exponent, H, in a time series may be interpreted as an indicator of the irregularity of the price of a commodity, currency or similar quantity. Interval estimation and hypothesis testing for H are central to comparative quantitative analysis.

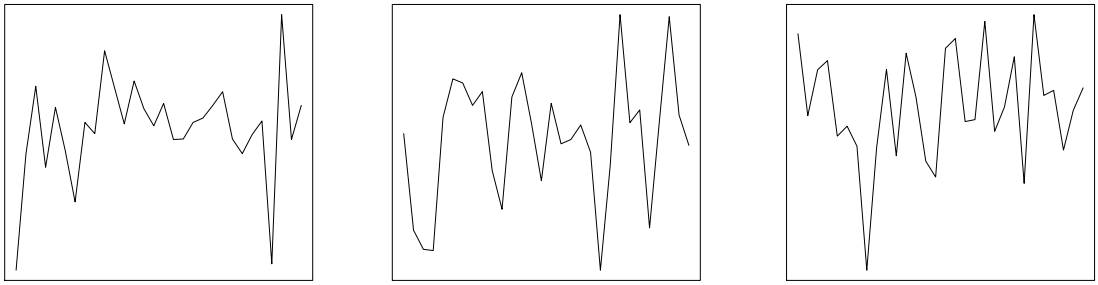


Figure 10.3: Self-similarity in CAC40 Index returns

10.1.1.2 Presenting the results

Working on the Nile River Dam Project, Hurst was concerned with the storage capacity of a reservoir, based on an estimate of the water inflow and of the need for water outflow. Studying a 847-year record from 622 A.D. to 1469 A.D., he found that the record was not random, as larger than average overflows were more likely to be followed by more large overflows. Further, the process would abruptly change to a lower than average overflow, and be followed by more lower overflows. This is the characteristics of cycles with nonperiodic length. Working on Brownian motion, Einstein [1908] found that the distance R that a random particle covers increases with the square root of time T used to measure it, that is,

$$R = T^{\frac{1}{2}} \tag{10.1.1}$$

which is called the T to the one-half rule commonly used in statistics. It is used in financial economics to annualise volatility or standard deviation. However, in a sample data with N elements, we can only use this equation if the time series under consideration is independent of increasing values of N. That is, Equation (10.1.1) only applies to Brownian motion time series. Hurst generalised the concept to non-Brownian motion by taking into account systems that are not independent. Hurst et al. [1965] remedied this drawback by dividing the range by the standard deviation S of the original observations. They proposed a more general relation by calculating the Rescaled Range as

$$\frac{R}{S} = k \times T^H$$

where k is a constant depending on the time series (see Peters [1994] [1991-96]). The $\frac{R}{S}$ value has a mean of zero, is expressed in terms of local standard deviation S, and scales as we increase the time increment T by a power law value equals to the Hurst Exponent (H). By rescaling the data we can compare diverse phenomena and time periods. The value of the Hurst Exponent varies between 0 and 1, with $H = \frac{1}{2}$ implying a random walk or an independent process (see Figure 10.4). Figure (10.5) illustrates the role of the Hurst exponent. For $0 \leq H < \frac{1}{2}$ we have anti-persistence (or ergodicity) where the process covers less distance than a random walk. It is often referred as mean reverting, meaning that if the system has been up in the previous period, it is more likely to be down in the next period. The converse being true. Anti-persistent time series exhibit higher noise and more volatility than a random series because it consists of frequent reversals. For $\frac{1}{2} < H \leq 1$ we have persistence (or trend-reinforcing) where the process covers more distance than a random walk. A persistent series has long memory effects such that the trend at a particular point in time affects the remainder of the time series. If the series has been up (down) in the last period, then the chances are that it will continue to be positive (negative) in the next period.

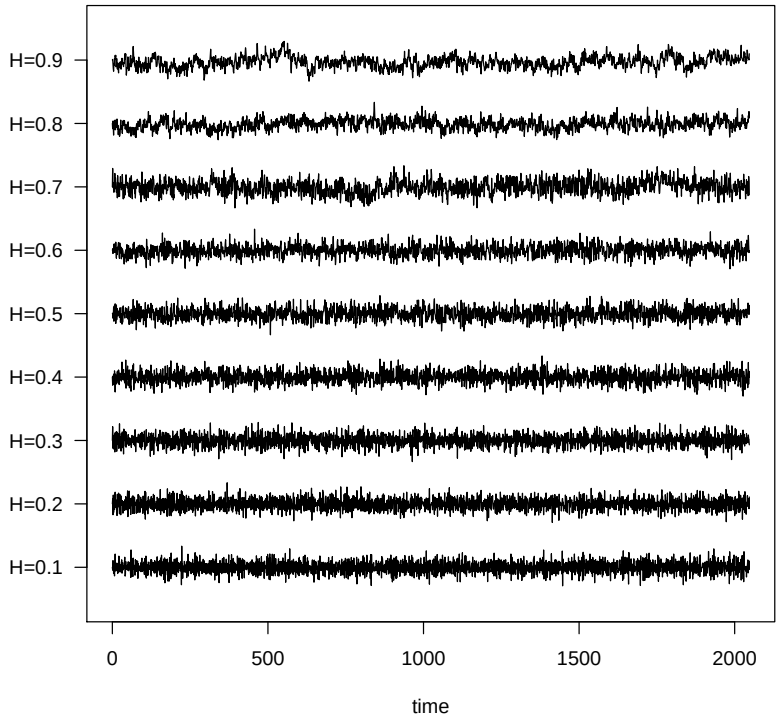
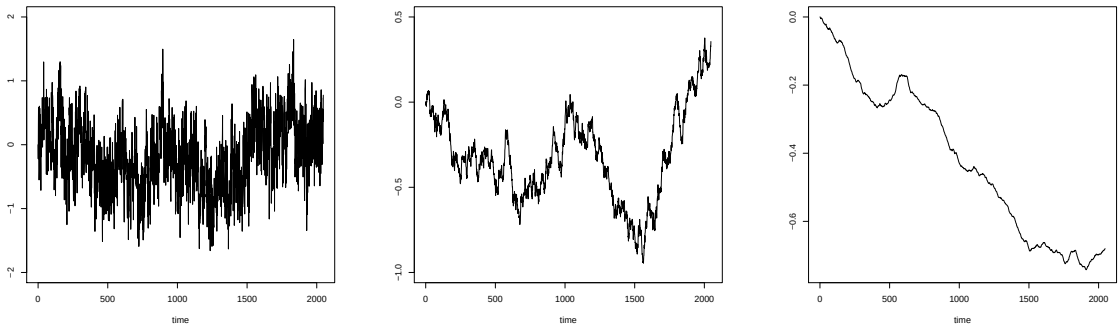


Figure 10.4: Range of Hurst exponents between 0.1 and 0.9



(a) $H = 0.1$ (b) $H = 0.5$ (c) $H = 0.9$

Figure 10.5: Cumulative observations for different values of H

The fractal dimension D of a time series measures how jagged the time series is. In order to calculate the fractal dimension of a time series we count the number of circles of a given, fixed diameter needed to cover the entire time

series. We then increase the diameter and count again the number of circles required to cover the time series. Repeating that process, the number of circles has an exponential relationship with the radius of the circle satisfying

$$N \times d^D = 1$$

where N is the number of circles, and d is the diameter. Transforming the above equation [1], the fractal dimension is given by

$$D = \frac{\log(N)}{\log(\frac{1}{d})} \tag{10.1.2}$$

Thus, the fractal dimension of a time series is a function of scaling in time. Mandelbrot showed that the Hurst Exponent, H, is related to the fractal dimension D for a self-similar surface in n-dimensional space by the relation

$$D = n + 1 - H \tag{10.1.3}$$

where $0 \leq H \leq 1$. While the fractal dimension of a line is 1.0 and the fractal dimension of a geometric plane is 2.0, the fractal dimension of a random walk is half way between the line and the plane at 1.5. A value of $\frac{1}{2} < H \leq 1$ results in a fractal dimension closer to a line, so that a persistent time series would result in a smoother, less jagged line than a random walk. Similarly, for $0 < H < \frac{1}{2}$, a time series is more jagged than a random series, or has more reversals.

Definition 10.1.1 *The Hurst Exponent is the measure of the smoothness of fractal time series based on the asymptotic behaviour of the rescaled range of the process.*

The Hurst Exponent can be estimated by

$$H = \frac{\log(\frac{R}{S})}{\log(T)}$$

where T is the duration of the sample data, and $\frac{R}{S}$ is the corresponding value of the rescaled range.

Definition 10.1.2 *The Rescaled Range is the measure characterising the divergence of time series defined as the range of the mean-centred values for a given duration divided by the standard deviation of that duration.*

Persistent series are fractional Brownian motion, or biased random walk with the strength of the bias depending on how far H is above $\frac{1}{2}$. For $H \neq \frac{1}{2}$ each observation carry a long-term memory, theoretically lasting forever. A system exhibiting Hurst statistics is the result of a long stream of interconnected events where time is important as the future depends on the past. As discussed in Feder [1988], the variance of fractionally integrated processes is proportional to t^{2H}, and the impact of the present on the future can be expressed as a correlation

$$C = 2^{(2H-1)} - 1 \tag{10.1.4}$$

where C is the correlation measure. For $H = \frac{1}{2}$ we get $C = 0$, and events are random and uncorrelated. The R/S analysis can classify an independent series, no matter what is the shape of the underlying distribution. Note, the correlation C is not related to the Auto Correlation Function (ACF) of Gaussian random variables which assumes Gaussian, or near-Gaussian, properties in the underlying distribution.

[1] $$N = (\frac{1}{d})^D$$

10.1.2 Defining random fractals

Mandelbrot et al. [1968] introduced the fractional Brownian motion (fBm) as a generalisation of the ordinary Brownian motion by considering correlations between the increments of the process. The fBm and its corresponding generalised derivative process, the fractional Gaussian noise (fGn), are the benchmark of the Fractal Market Hypothesis (FMH). This is the only family of processes being Gaussian, self-similar, and endowed with stationary increments.

10.1.2.1 The fractional Brownian motion

We consider the one dimensional random process $X(t)$ with values $X(t_1), X(t_2), ...$ such that the increment $X(t_2) - X(t_1)$ has a Gaussian distribution, and the mean square increments have a variance proportional to the time differences,

$$E[|X(t_2) - X(t_1)|^2] \propto |t_2 - t_1|$$

We say that the increments of X are statistically self-similar in the sense that

$$X(t_0 + t) - X(t_0) \text{ and } \frac{1}{\sqrt{r}}(X(t_0 + rt) - X(t_0))$$

have the same finite dimensional joint distribution functions for any t_0 and time scale $r > 0$. The integral of uncorrelated white Gaussian noise W satisfies the above equations, and is given by

$$X(t) = \int_{-\infty}^{t} W(s)ds$$

where the random variables $W(t)$ are uncorrelated and have the same normal distribution. The fractional Brownian motion (fBm) is a generalisation satisfying

$$Var(X(t_2) - X(t_1)) \propto |t_2 - t_1|^{2H}$$

for $0 < H < 1$. In the special case $H = \frac{1}{2}$ we recover the Brownian motion. Again, the increments of X are statistically self-similar with parameter H in the sense that

$$X(t_0 + t) - X(t_0) \text{ and } \frac{1}{r^H}(X(t_0 + rt) - X(t_0))$$

have the same finite dimensional joint distribution functions for any t_0 and $r > 0$. Setting $t_0 = 0$, the accelerated fBm $X(rt)$ is properly rescaled by dividing amplitudes by r^H.

More formally, the fractional Brownian motions, $\{B_H(t), t > 0\}$, is a Gaussian process with zero mean and stationary increments whose variance and covariance are given by

$$\begin{aligned} E[B_H^2(t)] &= \sigma^2 t^{2H} \\ E[B_H(s)B_H(t)] &= \frac{1}{2}\sigma^2\left(s^{2H} + t^{2H} - |t-s|^{2H}\right), t, s > 0 \end{aligned}$$

where $0 < H < 1, \sigma > 0$. Note, fBm are continuous but non-differentiable (in the classical sense). Note, Mandelbrot et al. [1968] defined the above as self-affinity of increments in distribution and proposed the following definition of self-similarity in distribution

Definition 10.1.3 *A process $(X_t)_{-\infty<t<\infty}$ is called self-similar if*

$$X(at) \to a^H X(t)$$

for a positive factor a and non-negative self-similarity parameter H.

A major consequence of this definition is that the moments of X, provided they exist, behave as power laws of time

$$E[|X(t)|^q] = E[|X(1)|^q]|t|^{qH} \tag{10.1.5}$$

In the framework of stochastic processes, such laws could only hold in distribution. In this case, Mandelbrot et al. speak of self-affine processes. The self-affinity in distribution is defined as

Definition 10.1.4 *A process $(X_t)_{-\infty<t<\infty}$, $X(0) = 0$, is called self-affine if*

$$\{X(ct_1), ..., X(ct_k)\} \to \{c^H X(t_1), .., c^H X(t_k)\}$$

for non-negative factors c and k, periods $t_1, .., t_k$, and a positive self-similarity parameter H.

so that self-affinity in distribution is a special case of self-similarity in distribution. Definition (10.1.3) determines the dependency structure of the increments of a process obeying such scaling behaviour as well as their higher moments showing hyperbolic decline of their autocorrelations with an exponent depending linearly on H.

10.1.2.2 The multidimensional fBm

A fractional Brownian function $f(r) : \mathbb{R}^d \to \mathbb{R}$ is charactersised by

- a variance
$$\sigma_H^2 =< [f(r+\lambda) - f(r)]^2 > \propto \|\lambda\|^\alpha \, , \alpha = 2H$$
with $r = (x_1, .., x_d)$, $\lambda = (\lambda_1, .., \lambda_d)$, and $\|\lambda\| = \sqrt{\lambda_1^2 + ... + \lambda_d^2}$.
- a power spectrum
$$S_H \propto \|w\|^{-\beta} \, , \beta = d + 2H$$
with $w = (w_1, .., w_d)$ the angular frequency, and $\|w\| = \sqrt{w_1^2 + ... + w_d^2}$.
- a number of objects N_H of characteristic size ϵ needed to cover the fractal
$$N_H \propto \epsilon^{-D} \, , D = d + 1 - H$$
where D is the fractal dimension of $f(r)$,

In the special case where $d = 1$, the power spectrum for the fractional Brownian motion becomes

$$S_H \propto \frac{1}{|w|^\beta}$$

with $\beta = 2H + 1$ and $1 < \beta < 3$. Similarly, the power spectrum for the fractional Gaussian noise becomes

$$S_H \propto \frac{1}{|w|^\beta}$$

with $\beta = 2H - 1$ and $-1 < \beta < 1$.

10.1.2.3 The fractional Gaussian noise

The implications of self-similarity for time series are not based on distributions but on dynamic properties of the time series which are defined by the autocorrelation function $\gamma(k)$ (see Section ()). We denote by $\{W_H(t), t > 0\}$ the process derived from the increment of fBm, namely

$$W_H(t) = B_H(t+1) - B_H(t)$$

Beran [1994] and Embrechts et al. [2002] showed that self-similar processes have autocorrelation function defined exactly and asymptotically in the following two propositions.

***Proposition* 1** *Let $(X_t)_{t=0}^T$ be self-similar process with $0 < H < 1$ and finite variance $\sigma^2 < \infty$. Then the correlations are given by autocorrelation function*

$$\gamma(k) = E[X_t X_{t+k}] = \frac{(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}}{2}$$

for $k \geq 0$, where H is the Hurst exponent.

***Proposition* 2** *Let $(X_t)_{t=0}^T$ be self-similar process with $0 < H < 1$ and finite variance $\sigma^2 < \infty$. Then the correlations are asymptotically given by autocorrelation function*

$$\gamma(k) \sim H(2H-1)k^{2H-2}$$

for $k \to \infty$, where H is the Hurst exponent.

Note, the second Proposition (2) suggests that if $H = \frac{1}{2}$, we have either an independent process, if the first Proposition (1) holds as well, or a short-term dependent process, if that first proposition is not valid. Thus, an independent process has zero correlations at all non-zero lags, while a short-term dependent process has significantly non-zero correlations at low lags but zero correlations at high lags. For $H > \frac{1}{2}$, the process has significantly positive correlations at all lags. It has hyperbolically decaying correlations which are non-summable and $\sum_{k=0}^{\infty} \gamma(k) = \infty$. For $H < \frac{1}{2}$, the process has significantly negative correlations at all lags decaying hyperbolically. However, the correlations are summable, and we get $0 < \sum_{k=0}^{\infty} \gamma(k) < \infty$ (see Embrechts et al. [2002]). The spectral density (Fourier transform of γ) is

$$f(\lambda) = C_H(2\sin(\frac{\lambda}{2}))^2 \sum_{k=-\infty}^{\infty} \frac{1}{|\lambda + 2\pi k|^{2H+1}} \sim C_H|\lambda|^{1-2H} \text{ as } \lambda \to 0$$

where C_H is a constant.

10.1.2.4 The fractal process and its distribution

Following Mandelbrot et al. [1997] we now define fractal process with the separation between multi-fractal and monofractal processes of Lux [2003].

Definition 10.1.5 *A process $(X_t)_{-\infty<t<\infty}$ is called fractal if*

$$X(ct) \to M(c)X(t) = c^{H(c)}X(t) \tag{10.1.6}$$

for a positive factor c and an arbitrary non-negative function $H(c)$. If $H(c) = H$ is a constant function, the process is said to be monofractal. Otherwise, the process is said to be multi-fractal.

The scaling factor $M(c)$ is a random function with possibly different shape for different scales. One moves from one scale to another via multiplication with a random factor $M(c)$. The last equality emphasise the fact that this variability of scaling laws can be translated into variability of the exponent H. Note, a self-affine process is a special case of a self-similar process which is in turn a special case of a fractal process, all connected by the parameter H showing persistent, anti-persistent, short-term dependent and independent processes. Levy distributions are stable distributions. According to Levy, a distribution function, $F(x)$, is stable if, for all $b_1, b_2 > 0$, there also exists $b > 0$ such that

$$F(\frac{x}{b_1}) \times F(\frac{x}{b_2}) = F(\frac{x}{b})$$

This is a general characteristic of the class of stable distribution. The characteristic functions of F can be expressed in a similar manner

$$f(b_1 t) \times f(b_2 t) = f(bt)$$

so that $f(b_1 t)$, $f(b_2 t)$, and $f(bt)$ all have the same shaped distribution, despite their being products of one another which accounts for their stability. Following Mandelbrot [1964], we are now going to define the family of stable distributions, also called fractal because of their self-similar properties.

Definition 10.1.6 *Stable distributions are determined by characteristic function, natural logarithm of which is defined as*

$$\phi(t) = \ln E[e^{ixt}] = i\delta t - |ct|^\alpha \big(1 - i\beta \frac{t}{|t|} \tan \frac{\pi\alpha}{2}\big)$$

for $\alpha \neq 1$ *and*

$$\phi(t) = i\delta t - |ct|^\alpha \big(1 + i\beta \frac{2}{\pi} \ln |t|\big)$$

for $\alpha = 1$*, where* $0 < \alpha \leq 2$ *is a characteristic exponent determining peakedness at* δ*,* $\beta \leq |1|$ *is a skewness parameter,* $-\infty < \delta < \infty$ *is a location parameter and* $0 \leq c < \infty$ *is a scale parameter.*

Note, $\frac{t}{|t|}$ is the the sign of t. We usually represent the parameters of a stable distribution by $S(x; \alpha, \beta, c, \delta)$. For $\alpha = 2$, $\beta = 0$ we recover the Gaussian distribution with variance equal to $c = 2\sigma^2$. Further, if $\delta = 0$, we get the standardise normal distribution with zero mean and unit variance. Setting $\alpha = 1$, $\beta = 0$, we recover the Cauchy distribution which has infinite variance and mean. The parameter α is crucial for an existence of variance. For $1 < \alpha < 2$, the distribution has infinite or undefined second moment and thus population variance. Moreover, for $0 < \alpha \leq 1$, the distribution has infinite mean as well. Even though there are a number of different ways of measuring the characteristic exponent α (see Fama [1965]), the R/S analysis and spectral analysis offer the most reliable method. Peters [1991-96], Panas [2001] and others showed that the self-similar process is connected to stable distribution of the process by the following relation between the Hurst exponent and the characteristic exponent α

$$\alpha = \frac{1}{H} \tag{10.1.7}$$

While H measures the fractal dimension of the time trace by the fractal dimension $2 - H$ (see Equation (10.1.3) with $n = 1$), it is also related to the statistical self-similarity of the process through Equation (10.1.7). However, $\frac{1}{H}$ measures the fractal dimension of the probability space.

10.1.2.5 An application to finance

Even though the major problem with the family of stable distributions is that they do not lend themselves to closed-for solutions, apart from the normal and Cauchy distributions, they have a number of desirable characteristics

- Stability under addition: the sum of two or more distributions that are fractal with characteristic exponent α keeps the same shape and characteristic exponent α.
- Self-similarity: fractal distributions are infinitely divisible. When the time scale changes, α remains the same.
- They are characterised by high peaks at the mean and by fat tails matching the empirical characteristics of market distributions.

but also some characteristics undesirable from a mathematical point of view, but actually reflecting market behaviour

- Infinite variance: second moments do not exist. Variance is unreliable as a measure of dispersion or risk.
- Jumps: large price changes can be large and discontinuous.

The Brownian motion, the L-stable process (see Mandelbrot [1963]), and the fBm are the main examples of self-affine processes in finance. Long memory is a characteristic feature of fBm, $B_H(t)$, having continuous sample paths, as well as Gaussian increments. Granger et al. [1980] and Hosking [1981] introduced ARFIMA, a discrete-time counterpart of the fBm which was used in economics to introduce long memory (see details in Section (10.1.3.2)). However, neither fBm nor ARFIMA disentangle volatility persistence from long memory in returns. To overcome this problem, Mandelbrot et al. [1997] proposed the multifractal model of asset returns (MMAR) which combines several elements of research on financial time series, such as generating fat tails in the unconditional distribution of returns, generating returns with a finite variance, fluctuations in volatility. Further, the multifractal model has long memory in the absolute value of returns, but the returns themselves have a white spectrum. At last, the multifractal model is built on the concept of trading time. It is constructed by compounding a Brownian motion, $B(t)$, with a random increasing function, $\theta(t)$, called the trading or time-deformation process (see details in Section (11.6)). The MMAR model locally varies as $(dt)^{\alpha(t)}$, where the local scale $\alpha(t)$ can take a continuum of values with relative occurrences summarised in a renormalised probability density called the multifractal spectrum (see details in Section (11.1.2)).

10.1.3 A first approach to generating random fractals

As listed by Coeurjolly [2000] and Dieker [2003], there exists a plethora of methods for generating fractional Brownian motion (fBm). Historically, Mandelbrot et al. [1968] first simulated a fBm by using its stochastic representation with respect to an ordinary Brownian motion. Since this approach makes very rough approximations, Choleski proposed a method using a Choleski decomposition of the covariance matrix of the standard fBm which is exact but highly demanding in computational resource (complexity $O(n^3)$). To circumvent this problem, Davis et al. considered the same approach as Choleski, but used in addition circulant matrix. Later, Wood et al. [1994] generalised the method to end up with an exact simulation of fBm with the complexity $O(n \log n)$. There are many more methods such as Hosking-Levinson-Durbin method (see Hosking [1984]), the Random Midpoint Dispacement (see Lau et al. [1995]), Fast Fourier Transform method (see Paxson [1997]), and wavelet-based method (see Abry et al. [1996]), to name a few. While there is a large number of methods for identifying fractional Brownian objects, Meharabi et al. [1997] compared seven of these methods and found that the wavelet decomposition method offers a highly accurate and efficient tool for characterising long-range correlations in complex distributions and profiles. We are now going to briefly present the Fourier filtering method and the ARFIMA models.

10.1.3.1 Approximating fBm by spectral synthesis

The spectral synthesis method (or Fourier filtering method) for generating fBm is based on the spectral representation of samples of the process $X(t)$. As the Fourier transform of X is generally undefined, we first restrict $X(t)$ to a finite time interval $0 < t < T$, denoted $X(t,T)$. Then, we get

$$X(t,T) = \int_{\infty}^{\infty} F(f,T)e^{2\pi itf}df$$

where $F(f,T)$ is the Fourier transform of $X(t,T)$

$$F(f,T) = \int_0^T X(t)e^{-2\pi itf}dt$$

Note, $|F(f,T)|^2 df$ is the contribution to the total energy of $X(t,T)$ from those components with frequencies between f and $f+df$. Thus, the average power of X (power spectra) contained in the interval $[0,T]$ is given by

$$P_F(f,T) = \frac{1}{T}E_F(f) = \frac{1}{T}\int_\infty^\infty |F(f,T)|^2 df$$

and the power spectral density (PSD) per unit time of $X(t,T)$ (see Equation (G.1.1)) is

$$S(f,T) = \frac{1}{T}|F(t,t)|^2$$

The spectral density of X is then obtained in the limit as $T \to \infty$

$$S(f) = \lim_{T\to\infty} \frac{1}{T}|F(f,T)|^2$$

Note, $S(f)df$ is the average of the contribution to the total power from components in $X(t)$ with frequencies between f and $f+df$. It is a non-negative and even function. In the spectral analysis, $X(t)$ is decomposed into a sum of infinitely many sine and cosine terms of frequencies f whose powers (and amplitudes) are determined by the spectral density $S(f)$.

The main idea behind spectral synthesis is that a prescription of the right kind of spectral density $S(f)$ will give rise to fBm with exponent $0 < H < 1$. For instance, if the random function $X(t)$ contains equal power for all frequencies f, the process is white noise. If $S(f)$ is proportional to $\frac{1}{f^2}$ we obtain the Brownian motion. In general, a process $X(t)$ with a spectral density proportional to $\frac{1}{f^\beta}$ corresponds to fBm with $H = \frac{\beta-1}{2}$. That is,

$$S(f) \propto \frac{1}{f^\beta} \sim \text{ fBm with } \beta = 2H+1$$

Choosing the spectral exponent β between 1 and 3 will generate a graph of fBm with a fractal dimension of

$$D_f = 2 - H = \frac{5-\beta}{2}$$

The relationship between β and H can be derived from the fact that the mean square increments are directly related to the autocorrelation function of X, which in turns defines the spectral density by means of a Fourier transform via the Wiener-Khintchine relation.

To obtain a practical algorithm Saupe [1988] translated the above equations into conditions on the coefficients a_k of the discrete Fourier transform

$$\overline{X}(t) = \sum_{k=0}^{N-1} a_k e^{2\pi ikt}$$

Since the coefficients a_k are in a one-to-one correspondence with the complex values $\overline{X}(t_k)$ for $t_k = \frac{k}{N}$, $k = 0,1,..,N-1$, the condition to be imposed on the coefficients to satisfy $S(f) \propto \frac{1}{f^\beta}$ becomes

$$E[|a_k|^2] \propto \frac{1}{k^\beta}$$

for $0 < k < \frac{N}{2}$ since k denotes the frequency in the above equation. For $k \geq \frac{N}{2}$ we must have $a_k = \overline{a}_{N-k}$ because $\overline{X}$ is a real function. Thus, the method consists of randomly choosing coefficients subject to the above expectation and then computing the inverse Fourier transform to obtain X in the time domain. As the process X need only have real values, it is sufficient to sample real random variables A_k and B_k under the constraint

$$E[A_k^2 + B_k^2] \propto \frac{1}{k^\beta}$$

and then set

$$\overline{X}(t) = \sum_{k=1}^{\frac{N}{2}} \big(A_k \cos kt + B_k \sin kt\big)$$

One can interpret the addition of more and more random Fourier coefficients a_k as a process of adding higher frequencies, thus, increasing the resolution in the frequency domain. However, due to the nature of Fourier transforms, the generated samples are periodic, as one can compute twice or four times as many points as actually needed and then discard a part of the sequence.

10.1.3.2 The ARFIMA models

We saw in Appendix (D.3) that ARIMA models are homogeneous nonstationary systems that can be made stationary by successively differencing the observations. While the differencing parameter, d, was always an integer value, Granger et al. [1980] and Hosking [1981] generalised the $ARIMA(p,d,q)$ value for fractional differencing to yield the autoregressive fractionally integrated moving average (ARFIMA) process where d can be any real value, including fractional value. As a result, these models can generate persistent and antipersistent behaviour where the $ARFIMA(0,d,0)$ process is the fractional Brownian motion introduced by Mandelbrot. The general $ARFIMA(p,d,q)$ process can include short-memory AR or MA processes over a long-memory process. Hence, the very high-frequency terms can be autoregressive, when superimposed over a long-memory Hurst process. Fractional differencing tries to convert a continuous process, fractional Brownian motion into a discrete one by breaking the differencing process into smaller components. Moreover, there is a direct relationship between the Hurst exponent and the fractional operator given by

$$d = H - \frac{1}{2}$$

where $0 < d < \frac{1}{2}$ corresponds to a persistent black noise process, and $-\frac{1}{2} < d < 0$ is equivalent to an antipersistent pink noise system. White noise, the increments of a random walk, corresponds to $d = 0$, and brown noise, the trail of a random walk, corresponds to $d = 1$. We saw in Section () that the discrete time white noise can be represented in terms of a backward shift operator as $B(x_t) = x_{t-1}$ so that

$$\Delta x_t = (1 - B)x_t = a_t$$

where the a_t are i.i.d. random variables. fractional differenced white noise with parameter d is given by the binomial series

$$\begin{aligned} \Delta^d &= (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \\ &= 1 - dB - \frac{1}{2}d(1-d)B^2 - \frac{1}{6}d(1-d)(2-d)B^3 - ... \end{aligned}$$

We are now presenting the relevant characteristic of the fractional noise process developed by Hosking. We let $\{x_t\}$ be an $ARFIMA(0,d,0)$ process where k is the time lag and a_t is a white noise process with mean zero and variance σ_a^2. When $d < \frac{1}{2}$, then $\{x_t\}$ is a stationary process and has infinite moving average representation

$$x_t = \psi(B)a_t = \sum_{k=0}^{\infty} \psi_k a_{t-k}$$

where $\psi_k = \frac{(k+d-1)!}{k!(d-1)!}$. When $d > -\frac{1}{2}$, then $\{x_t\}$ is invertible and has the infinite autoregressive representation

$$\pi(B)x_t = \sum_{k=0}^{\infty} \pi_k x_{t-k}$$

where $\pi_k = \frac{(k-d-1)!}{k!(d-1)!}$. The spectral density of $\{x_t\}$ is

$$s(w) = (2\sin\frac{w}{2})^{-2d}$$

for $0 < w \leq \pi$. The covariance function of $\{x_t\}$ is

$$\gamma_k = E[x_t x_{t-k}] = \frac{(-1)^k(-2d)!}{(k-d)!(-k-d)!}$$

The covariance function of $\{x_t\}$ is

$$\rho_k \sim \frac{(-d)!}{(d-1)!}k^{2d-1}$$

as $k \to \infty$. The inverse correlations of $\{x_t\}$ are

$$\rho_{inv,k} \sim \frac{d!}{(-d-1)!}k^{-1-2d}$$

The partial correlations of $\{x_t\}$ are

$$\phi_{kk} = \frac{d}{k-d}\,, k = 1, 2, ...$$

Note, for $-\frac{1}{2} < d < \frac{1}{2}$, both ϕ_k and π_k decay hyperbolically (according to a power law) rather than exponentially, as they would with an AR process. For $d > 0$, the correlation function is also characterised by power law decay. This implies that $\{x_t\}$ is asymptotically self-similar, or it has a statistical fractal structure. Further, the partial and inverse correlations also decay hyperbolically, unlike the standard $ARIMA(p, 0, q)$. All of the hyperbolic decay behaviour in the correlations is also consistent with a long-memory, stationary process for $d > 0$. In the case where $-\frac{1}{2} < d < 0$, the $ARFIMA(0, d, 0)$ process is antipersistent. The correlations and partial correlations are all negative, except for $\rho_0 = 1$, and decay to zero according toa power law.

The general $ARFIMA(p, d, q)$ process has short-frequency effects superimposed over the low-frequency or long-memory process. Hosking argued that in practice it is likely to be of most interest for small values of p and q. An $ARFIMA(1, d, 0)$ process is defined by

$$(1 - \phi B)\Delta^d y_t = a_t$$

where a_t is a white noise process. We must include the fractional differencing process $x_t = \phi(B)a_t$ with $\Delta x_t = a_t$ so that $x_t = (1 - \phi B)y_t$. The $ARIMA(1, d, 0)$ variable, y_t, is a first-order autoregression with $ARIMA(0, d, 0)$ disturbance, that is, an $ARFIMA(1, d, 0)$ process. Hence, y_t will have short-term behaviour that depends on the coefficient of autoregression, ϕ_t, like a normal $AR(1)$ process, but the long-term behaviour of y_t will be similar to x_t and will exhibit persistence or antipersistence, depending on the value of d. For stationarity and invertibility, we assume $|d| < \frac{1}{2}$ and $|\phi| < 1$.

10.1.4 From efficient to fractal market hypothesis

10.1.4.1 Some limits of the efficient market hypothesis

We saw in Section (1.7.6) that the efficient market hypothesis (EMH) required observed prices to be independent, or at best, to have a short-term memory such that the current change in prices could not be inferred from previous changes. Since efficient markets are priced so that all public information, both fundamental and price history, is already discounted, prices only move when new information is received. That is, only today's unexpected news can cause today's price to change. This could only occur if price changes were random walk and if the best estimate of the future price was the current price. The process would be a martingale, or a fair game. However, stock markets were started so that traders could find a buyer if one of them wanted to sell, and a seller if one of them wanted to buy. This process would create liquidity ensuring that

1. the price investors get is close to what the market consider fair
2. investors with different investment horizons can trade efficiently with one another
3. there are no panics or stampedes occurring when supply and demand become imbalanced

That is, if short-term investor experiences relatively high loss, it becomes a buying opportunity for a long-term investor and vice versa. Investors being considered as rational, after assessing the risks involved, the collective consciousness of the market finds an equilibrium price. Ignoring liquidity, EMH only consider fair price even though completing the trade at any cost may become vital in some market situations (non-stable or illiquid markets). In general, the frequency distribution of returns is a fat-tailed, high-peaked distribution existing at many different investment horizons. Further, the standard deviation of returns increases at a faster rate than the square root of time. As a result, correlations come and go, and volatility is highly unstable. Taking into consideration these observations, Mandelbrot [1960] [1963a] [1963] described the financial market as a system with fat tails, stable distributions and persistence. Consequently, the evidence of long-range memory in financial data causes several drawbacks in modern finance

1. the optimal consumption and portfolio decisions may become extremely sensitive to the investment horizon.
2. the methods used to price financial derivatives based on martingale models are no-longer useful.
3. since the usual tests on the CAPM and arbitrage pricing theory (APT) do not take into account long-range dependence, they can not be applied to series presenting such behaviour.
4. if long-range persistence is present in the returns of financial assets, the random walk hypothesis is no-longer valid and neither is the market efficiency hypothesis.

Hence, long memory effects severely inhibit the validity of econometric models, explaining the poor record economists have had in forecasting. In order to explain discontinuities in the pricing structure, and the fat tails, Shiller [1989] and Miller [1991] proposed that information arrives in a lumpy, discontinuous manner. Investors still react to information homogeneously, preserving the assumption of independence, but the arrival of information is discontinuous. However, investors are not homogeneous, and the importance of information can be considered largely dependent on the investment horizon of the investors. The important information being different at each investment horizons, the source of liquidity can be characterised by investors with different investment horizons, different information sets, and consequently, different concepts of fair price. Further, new information may contribute to increased levels of uncertainty, rather than increased levels of knowledge. We may get increased volatility, or merely a noisy jitter. At last, at longer frequencies, the market reacts to economic and fundamental information in a nonlinear fashion. Studies of macroeconomic indices and simulations showed that the influence of information flow in the formation of economic cycles was highly relevant, and delayed information flow markedly affected economy system evolutions (see Ausloos et al. [2007]) and Miskiewicz et al. [2007]. Hence, while the linear paradigm, built into the rational investor concept, states that investors react to information in a linear fashion, the new paradigms must generalise investor reaction by

accepting the possibility of nonlinear reaction to information. Accounting for these new stylised facts, Larrain [1991] and Vaga [1990] proposed two similar models contradicting some assumptions of EMH, homogeneity of investors, independent identically distributed returns and risk-return tradeoff.

10.1.4.2 The Larrain KZ model

Even though the Larrain's KZ model (LKZ) proposed by Larrain [1991] is a model of real interest rate, its main idea is general enough to be an alternative to the EMH. In the LKZ model, a system can be generated from two separate mechanisms, one based on past behaviour and the other one based on interconnection with other fundamental variables. We let $(X_t; t = 1, .., n)$ be the underlying process, and assume that its dynamics are given by the following two equations

$$X_{t+1} = f(X_{t-n})$$

and

$$X_{t+1} = g(Z)$$

where Z is another stochastic process. The first equation states that the future values X_{t+1} are dependent on present and past values $X_t, .., X_{t-n}$, and the second equation states that the future values X_{t+1} are dependent on fundamental variables represented by Z. As an example, Larrain specified the function $f(\bullet)$, so that the first equation rewrite

$$X_{t+1} = a - cX_t(1 - X_t)$$

where a is an arbitrary constant and $c > 0$ is a positive constant. Putting together the two equations, we get

$$X_{t+1} = a - cX_t(1 - X_t) + g(Z)$$

Testing empirically the model on interest rates with different fundamental variables, Larrain obtained very strong results. The past real rates together with real GNP, nominal money supply, consumer price index, real personal income and real personal consumption all had significant coefficients. Asa result, he concluded that both past prices and fundamental variables are important for describing the dynamical system, so that fundamentalists and technicians can be part of the market and both can influence the behaviour of market prices. Fundamentalists base their estimates on changes of expected cash flows, while technicians base their trading strategy on crowd behaviour, short-term effects or past behaviour of stock prices. This implication strongly contradicts two assumptions of the EMH, namely the homogeneity of the investors and the unusefulness of past prices for future prices prediction.

10.1.4.3 The coherent market hypothesis

Vaga [1990] developed the coherent market hypothesis by considering the theory of coherent systems in natural sciences, which is based on a system with an order parameter summing all external forces driving the system. The main idea is based on Ising model of ferromagnetism where molecules behave randomly (with normally distributed movements) up till order parameter (temperature of an iron bar) reaches certain level where the molecules start to cluster and behave chaotically. The theory has been applied to the behaviour of social groups as well as to the behaviour of investors (see Schobel et al. [2006]). Vaga proposed the following probability formula for annualised return $f(q)$

$$f(q) = c^{-1}Q(q)e^{2\int_{-\frac{1}{2}}^{q}\frac{K(y)}{Q(y)}dy}$$

where

$$K(q) = \sinh(kq + h) - 2q\cosh(kq + h)$$

$$Q(q) = \frac{1}{n}\cosh(kq+h) - 2q\sinh(kq+h)$$

and

$$c^{-1} = \int_{-\frac{1}{2}}^{\frac{1}{2}} Q^{-1}(q) e^{2\int_{-\frac{1}{2}}^{q} \frac{K(y)}{Q(y)} dy} dq$$

where n is the number of degrees of freedom (market participants), k is a degree of crowd behaviour, and h is a fundamental bias. Schobel et al. identified five types of markets with respect to varying parameters k and h, namely,

- Efficient market $(0 \leq k << k_{critical}, h = 0)$ where investors act independently of one another and a random walk is present.
- Coherent market $(k \approx k_{critical}, h << 0 h >> 0)$ where crowd behaviour is in conjunction with strong bullish or bearish fundamentals and creates coherent market where traditional risk-return tradeoff is inverted and investors can earn above-average returns while facing below-average risk.
- chaotic market $(k \approx k_{critical}, h \approx 0)$ where crowd behaviour is in conjunction with only weak bearish or bullish fundamentals and creates the situation of low returns with above-average risk.
- Repelling market $(k < 0, h = 0)$ where opposite of crowd behaviour is present, investors try to avoid having the same opinion as the majority.
- Unstable transitions consist of all market states that can not be assigned to any of the previous states.

Similarly to the LKZ model, markets can exhibit various stages of behaviour by combining fundamental and sentiment influences which contradict the assumptions of EMH.

10.1.4.4 Defining the fractal market hypothesis

Retaining the most important assumptions from the above models, Peters [1994] proposed the Fractal Market Hypothesis (FMH) to model investor behaviour and market price movements, that is,

1. The market is stable when it consists of investors covering a large number of investment horizons, ensuring liquidity for traders.
2. The information set is more related to market sentiment and technical factors in the short term than in the longer term. As investment horizons increase, longer-term fundamental information dominates. Thus, price changes may reflect information important only to that investment horizon.
3. If an event occurs making the validity of fundamental information questionable, long-term investors either stop participating in the market or begin trading based on the short-term information set. When the overall investment horizon of the market shrinks to a uniform level, the market becomes unstable. In that case, there are no long-term investors to stabilise the market by offering liquidity to short-term investors.
4. Prices reflect a combination of short-term technical trading and long-term fundamental valuation, so that short-term price changes are likely to be more volatile, or noisier, than long-term trades. The underlying trend in the market is reflective of changes in expected earnings, based on the changing economic environment. Short-term trends are more likely the result of crowed behaviour. There is no reason to believe that the length of the short-term trends is related to the long-term economic trend.
5. If a security has no tie to the economic cycle, then there will be no long-term trend. Trading, liquidity, and short-term information will dominate.

Those hypothesis state that the different investment horizons value information differently, leading to uneven diffusion of information, so that at any one time, prices only reflect the information important to that investment horizon. Investors with short investment horizon focus on technical information and crowd behaviour of other market participants, while long-term investors base their decisions on fundamental information, which emphasise the heterogeneity of investors. This lead Peters to argue that fair value is not a single price, but a range of prices determined partly by fundamental information (earnings, management, new products, etc.) and by what investors feel other investors will be willing to pay (sentiment component). While the former is using fundamental analysis, the latter uses technical analysis setting a range around the fair value. The combination of information and sentiment results in a bias in the assessment of a stock's value. The range is dynamic as new information about the specific security or the market as a whole can shift the range and cause dramatic reversals in either the market or the single stock. Hence, a single negative information can turn markets into a downward spiral. This is similar with the Coherent Market Hypothesis (CMH) (see Vaga [1990]), and the K-Z model (see Larrain [1991]) where the market assumes different states and can shift between stable and instable regimes. In all cases, the chaotic regime occurs when investors lose faith in long-term fundamental information. EMH works well as long as markets are stable and close to equilibrium, but it cease to work if markets are close to or in turbulence. As opposed to the EMH which restricts the data process to be normal, one important generalisation of the FMH is that it does not restrict the data process to be of any specific distribution. FMH states that the market is stable as long as the returns of different investment horizons are self-similar in their distribution, thus, when it has no characteristic time scale or investment horizon. That is, instability occurs when the market loses its fractal structure and assumes a fairly uniform investment horizon. For fractality measurement, we use the Hurst exponent applied to financial time series.

10.2 The R/S analysis

10.2.1 Defining R/S analysis for financial series

As discussed in Section (10.1.1), the Rescaled Range Analysis (R/S) was developed by Hurst [1951] as a statistical method for analysing long records of natural phenomenon. It has been extended to study economic and capital market time series by defining a range that would be comparable to the fluctuations of the reservoir height levels. That is, the Hurst statistic applied to the Nile River discharge record is replaced by the difference between the maximum and minimum levels of the cumulative deviation over N periods of financial series. In that setting, R/S analysis, which is the central tool for fractal data modelling, is made of

1. the difference between the maximum and the minimum cumulative values
2. the standard deviation from the observed values.

Given a discrete time series (of order of magnitude the market returns) with values $X_1, .., X_n$ where n is the number of observations, we let $\hat{\mu}_n$ and $\hat{\sigma}_n$ be respectively the sample mean and sample standard deviation of the time series. The Rescaled Range is calculated by rescaling the time series by subtracting the sample mean, getting the mean-adjusted series

$$Y_i = X_i - \hat{\mu}_n \, , i = 1, .., n$$

so that the resulting series has zero mean. We let Γ_n be the cumulative deviate time series over n periods with element given by

$$\Gamma_i = \Gamma_{i-1} + Y_i \, , i = 2, .., n \, , \Gamma_1 = Y_1$$

with $\Gamma_n = 0$ since Y has a zero mean [2]. The adjusted range is the difference between the maximum value and the minimum value of Γ_i, that is,

[2] $\Gamma_n = \sum_{i=1}^n X_i - n\hat{\mu}_n = 0$

$$R_n = \max(\Gamma_1, .., \Gamma_n) - \min(\Gamma_1, .., \Gamma_n)$$

with the property that $R_n \geq 0$. The adjusted range R_n is the distance that the system travels for time n. In discrete time, the distance R a particle covers increases with respect to time T according to the general relation

$$(\frac{R}{S})_n = c \times n^H \text{ as } n \to \infty \tag{10.2.8}$$

where the subscript n refers to the duration of the time series, c is a constant, and H is the Hurst Exponent. Taking the logarithm of the Rescaled Range with basis equal to b, we get

$$\log_b(\frac{R}{S})_n = H \log_b(n) + \log_b(c)$$

which indicates the power scaling. Finding the slope of the log/log graph of $\frac{R}{S}$ versus n will give us an estimate of H making no assumptions about the shape of the underlying distribution (see Figure (10.6)). Peters noted that real observations with long scale n could be expected to exhibit properties similar to regular Brownian motion, or pure random walk, as the memory effect dissipates. Finding $H = \frac{1}{2}$ does not prove a Gaussian random walk, but only that there is no long memory process. That is, any independent system, Gaussian or otherwise, produces $H = \frac{1}{2}$. Assuming $c = \frac{1}{2}$, Hurst gave a formula for estimating the exponent H from a single R/S value

$$H = \frac{\log_b(\frac{R}{S})}{\log_b(\frac{n}{2})}$$

where n is the number of observations. For short data sets, where regression is not possible, this empirical law can be used as a reasonable estimate.

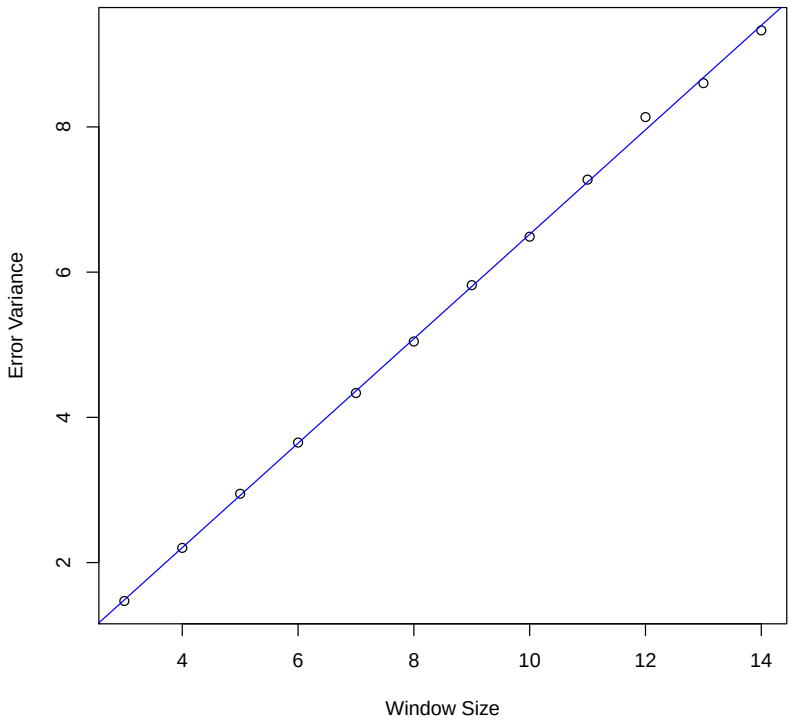


Figure 10.6: Log-regression of the Rescale Range on the scale for a simulated fBm with $H = 0.62$. The slope of the blue line gives the Hurst Exponent

The joker effect Hurst assumed that the water inflow and outflow of the reservoir under study was governed by a biased random walk, and devised an elegant method to simulate such a process. Rather than flipping coins to simulate random walks, Hurst constructed a probability pack of cards with 52 cards numbered $-1, 1, -3, 3, -5, 5, -7, 7$. The numbers were distributed to approximate the normal curve and the random walk was simulated by shuffling, cutting the deck, and noting the cut cards. The process to simulate a biased random walk was more complex, but resulted from randomly cutting the deck. He would first shuffle the deck, cut it once, and note the number, say 3. Replacing the card and reshuffling the deck, he would deal out two hands of 26 cards, obtaining deck A and deck B. He would then take the three highest cards from deck A, place them in deck B, and remove the three lowest cards in deck B. Deck B being biased to a level of 3, a joker is added to deck B which is reshuffled. This biased deck was used as a time series generator, until the joker was cut, in which case a new biased hand was created. Performing 1000 trials of 100 hands, he calculated $H = 0.72$. Yet, no matter how many times the experiment was done, Hurst found the same result. He found that a combination of random events with generating order creates a structure.

10.2.2 A step-by-step guide to R/S analysis

10.2.2.1 A first approach

The R/S analysis being highly data intensive, we must consider a series of executable steps. Given the process $X_{t,N}$, the process for estimating the Hurst Exponent by using the R/S analysis consists of three steps

1. The cumulative total at each point in time, for a time series over a total duration N, is given by

$$\Gamma_{N,k} = \sum_{i=1}^{k}(X_i - \hat{\mu}_N)\ , 0 < k \leq N$$

The range R of Γ is given by

$$R = \max(\Gamma_{N,k}) - \min(\Gamma_{N,k})$$

Given the sample standard deviation $\hat{\sigma}_N$, the Rescaled Range is given by $\frac{R}{S}$ with $S = \hat{\sigma}_N$.

2. We take $N = \frac{N}{2}$ and calculate the new Rescaled Range $\frac{R}{S}$ for each segment, following step 1. The average value of $\frac{R}{S}$ is then computed. We repeat this process for successively smaller intervals over the data set, dividing each segment obtained in each step in two, calculating $\frac{R}{S}$ for each segment and computing the average $\frac{R}{S}$.

3. The Hurst Exponent is estimated by plotting the values of $\log_b(\frac{R}{S})$ versus $\log_b(N)$. Such a graph is called a pox plot. The slope of the best fitting line gives the estimate of the Hurst Exponent.

Note, R/S analysis is calculated for all values of n, such that if there is not enough data left for a full value of n, then it leaves that data off. As n increases, more and more of the data is discarded resulting in the jaggedness of the log/log plots at high value of n. An alternative is to use values of n that are even divisors of the total number of observations so that all of the R/S values use all of the data. One must therefore sometimes reduce the initial data set to a number having the most divisors. Further, one can generate least squares regression line using linear regression method. However, this approach is known to produce biased estimates of the power-law exponent. A more principled approach fits the power law in a maximum-likelihood fashion (see Clauset et al. [2009]).

10.2.2.2 A better step-by-step method

We observe the stochastic process X_t at time points $t \in \mathcal{I} = \{0, .., N\}$, and we let n be a small integer relative to N (asymptoticall, as $\frac{N}{n} \to \infty$), and let A denotes the integer part of $\frac{N}{n}$. Divide the interval $\mathcal{I}$ into A consecutive subintervals, each of length n and with overlapping end points. In every subinterval, correct the original datum X_t for location, using the mean slope of the process in the subinterval, obtaining

$$X_t - \frac{t}{n}(X_{an} - X_{(a-1)n})$$

for all t with $(a-1)n \leq t \leq an$ for all $a = 1, .., A$. Over the a-th subinterval

$$\mathcal{I}_a = \{(a-1)n, (a-1)n+1, .., an\}$$

for $1 \leq a \leq A$, construct the smallest box (with sides parallel to the cordinate axes) such that the box contains all the fluctuations of $X_t - \frac{t}{n}(X_{an} - X_{(a-1)n})$ occurring within $\mathcal{I}_a$. Then, the height of the box equals

$$R_a = \max_{(a-1)n \leq t \leq an}\{X_t - \frac{t}{n}(X_{an} - X_{(a-1)n})\} - \min_{(a-1)n \leq t \leq an}\{X_t - \frac{t}{n}(X_{an} - X_{(a-1)n})\}$$

We let S_a be the sample standard error of the n variables $X_t - X_{t-1}$ for $(a-1)n \leq t \leq an$. If the process X is stationary the S_a varies little with a, otherwise, dividing R_a by S_a corrects for the main effects of scale inhomogeneity in both spatial and temporal domains. The total area of the boxes, corrected for scale, is proportional in n to

$$(\frac{R}{S})_n = A^{-1}\sum_{a=1}^{A}\frac{R_a}{S_a}$$

The slope $\hat{H}$ of the regression of $\log(\frac{R}{S})_n$ on $\log(n)$, for k values of n, may be taken as an estimator of the Hurst constant describing long-range dependence of the process X.

10.2.3 Testing the limits of R/S analysis

We can question the validity of the Hurst estimate itself, as it is sensitive to the amount of data tested. An estimate of exponent H that is significantly different (greater) from $\frac{1}{2}$ has two possible explanations

1. there is a long memory component in the time series studied
2. the analysis itself is flawed.

One way of testing for the validity of the results is to randomly scramble the data so that the order of the observations is completely different from that of the original time series. Repeating the calculation of the Hurst exponent on the scrambled data, if the series is truly independent, the result should remain virtually unchanged because there were no long memory effects. In that case, scrambling the data has no effect on the qualitative aspect of the data. On the other hand, if there was a long memory effect, scrambling the data would destroy the structure of the system. The H estimate would be much lower, and closer to $\frac{1}{2}$, while the frequency distribution remains unchanged. Since an $AR(1)$ process is an infinite-memory process, some authors suggest to take $AR(1)$ residuals off the data before applying R/S analysis to filter out the short-memory process.

In order to assess the significance of R/S analysis, we need some kind of asymptotic theory to define confidence intervals, much like the t-statistics of linear regression. It can be done by first studying the behaviour of R/S analysis when the system is an independent, random one, and compare other processes to the random null hypothesis. Basing his null hypothesis on the binomial distribution and the tossing of coins, Hurst [1951] found the random walk to be a special case of Equation (10.2.8) given by

$$(\frac{R}{S})_n = (n\frac{\pi}{2})^{\frac{1}{2}}$$

where n is the number of observations. While Hurst worked on the rescaled range, Feller [1951] considered the adjusted range $R^{'}$ which is the cumulative deviations with the sample mean deleted, and developed its expected value and its variance. It would avoid the problem of computing the sample standard deviation for small values of n, and at the limit ($n \to \infty$), it would be equivalent to the rescaled range. He obtained the formulas

$$\begin{aligned} E[R^{'}(n)] &= (n\frac{\pi}{2})^{\frac{1}{2}} \\ Var(E[R^{'}(n)]) &= n(\frac{\pi^2}{6} - \frac{\pi}{2}) \end{aligned}$$

Since the R/S values of a random number should be normally distributed, the variance of $R^{'}(n)$ should decrease as we increase samples. Hurst and Feller suggested that the rescaled range also increases with the square root of time and that the variance of the range increases linearly with time.

With the power of computers we can directly apply Monte Carlo method of simulation on Equation (10.2.8) to test the Gaussian hypothesis. Peters considered a pseudo-random number series of 5000 values, scrambled twice, and calculated R/S values for all n that are evenly divisible into 5000. This process was repeated 300 times to generate 300 $(R/S))n$ values for each n. For $n > 20$, results were similar to the formulas by Hurst and Feller, but a consistent deviation was present for smaller values of n. In that setting, the $(R/S))n$ values and their variances were systematically lower. It can be explained because Hurst was calculating an asymptotic relationship holding only for large n. Mandelbrot et al. [1969b] referred to that region of small n as transient because n was not large enough for the proper behaviour to be observed. To circumvent this deviation, Anis et al. [1976] (AL76) developed the following statistic for small n

$$E[(R/S)_n] = \frac{\Gamma(\frac{1}{2}(n-1))}{\sqrt{\pi}\Gamma(\frac{1}{2}n)} \sum_{k=1}^{n-1} \sqrt{\frac{n-k}{k}}$$

However, this equation is less useful for large values of n as the gamma values become too large. Yet, approximating the gamma function $\Gamma(\bullet)$ with $\sqrt{\frac{2}{n}}$, just like the beta function $B(\bullet)$ is used as a substitute of gamma function, Peters approximated the Sterling's function as

$$\frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}n)} \approx (\frac{2}{n})^{\frac{1}{2}}$$

Note that the exact application of the Stirling's approximation yields

$$\frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}n)} \approx (\frac{2}{n-1})^{\frac{1}{2}}$$

Nonetheless, Peters simplified the equation to

$$E[(R/S)_n] = (n\frac{\pi}{2})^{-\frac{1}{2}} \sum_{k=1}^{n-1} \sqrt{\frac{n-k}{k}}$$

which can be used when $n > 300$. Still, the above equation was missing something for values of n less than 20. Multiplying the above equation with an empirical correction factor, Peters [1994] (P94) obtained

$$E[(R/S)_n] = \frac{n-\frac{1}{2}}{n}(n\frac{\pi}{2})^{-\frac{1}{2}} \sum_{k=1}^{n-1} \sqrt{\frac{n-k}{k}} \tag{10.2.9}$$

getting closer to the simulated R/S values. Using this equation, Peters showed that $H = \frac{1}{2}$ for an independent process is an asymptotic limit, and that the Hurst exponent will vary depending on the values of n used to run the regression. We must therefore consider a system under study in relation with the $E[R/S]$ series for the same values of n. Note, Couillard et al. [2005] thoroughly tested the above corrections for finite samples and found that Anis et al. estimates rescaled range for small samples $(n < 500)$ much more accurately and underestimates rescaled range for large samples $(n > 500)$ compared to Peters. But the underestimation is insignificant. They also tested the asymptotic standard deviation of H, denoted $\hat{\sigma}(H)$, and argued that Peters' statement that $\hat{\sigma}(H) \approx \frac{1}{\sqrt{T}}$ is only an asymptotic limit and is significantly biased for finite number of observations. They came up with a new estimate based on simulations up to $T = 10000$, getting $\hat{\sigma}(H) \approx \frac{1}{e^3\sqrt{T}}$.

10.2.4 Improving the R/S analysis

10.2.4.1 Reducing bias

A problem with R/S analysis, which is common to many Hurst parameter estimators, is knowing which values of n to consider. In general, since the R/S analysis is based on range statistic R and standard deviation S, the estimates of Hurst exponent can be biased. For instance, for small n short term correlations dominate and the readings are not valid, while for large n there are a few samples and the value of the statistic will not be accurate. Further, at low scales, n, sample standard deviation can strongly bias the final rescaled range, since standard deviation becomes close to zero, implying infinite rescaled range. Hence, we should run the regression over values of $n \geq 10$, as small values of n produce unstable estimates when sample sizes are small. In addition, range statistic is very sensitive to outliers and its estimate can strongly bias the final rescaled range at high scales, n, as the outliers are not averaged out as it is the case at low scales. Millen et al. [2003] proposed to use a minimum scale of at least 10 observations and a maximum scale of a half of the time series length.

10.2.4.2 Lo's modified R/S statistic

The R/S analysis is problematic when considering heteroskedastic time series and series with short-term memory. The complicated use for heteroskedastic time series, due to the use of sample standard deviation together with a filtration of a constant trend (computation of accumulated deviations from the arithmetic mean), makes R/S analysis sensitive to non-stationarities in the underlying process. One way forward is to consider detrended fluctuation analysis (DFA), where one filters the series not only from a constant trend, but also from higher polynomials such as linear and quadratic (details are given in Section (10.3.2.1)). To deal with short-term dependence in the time series, the modified rescaled range $(M - R/S)$ introduced by Lo [1991] is the mostly used technique. Differing only slightly from the R/S method in the computation of the standard deviation S, and focusing on the single period $n = N$, the $(M-R/S)$ analysis deals with both heteroskedasticity and short-term memory. The modified standard deviation is defined with the use of auto-covariance $\gamma_j = \sum_{i=j+1}^{N}(X_i - \overline{X}_N)(X_{i-j} - \overline{X}_N)$ of the selected sub-period, up to the lag λ, as follow

$$S_\lambda^2 = S^2 + 2\sum_{j=1}^{\lambda}\gamma_j w_j(\lambda)$$

where $w_j(\lambda) = \left(1 - \frac{j}{\lambda+1}\right)$, $\lambda < N$. Lo's modified R/S statistic $V_\lambda(N)$ is defined by

$$V_\lambda(N) = N^{-\frac{1}{2}}(\frac{R}{S_\lambda})_N$$

Note, for $\lambda = 0$ we recover the sample variance S^2 and as a result the R/S method. The weights $w_j(\bullet)$ are chosen such that S_λ^2 is the sample variance of an aggregated or averaged series. If a series has no long-range dependence, Lo showed that given the right choice of λ, the distribution of $V_\lambda(N)$ is asymptotic to that of

$$W_1 = \max_{0\le t\le 1} W_0(t) - \min_{0\le t\le 1} W_0(t)$$

where W_0 is a standard Brownian bridge [3] Since the distribution of the random variable W_1 is known

$$P(W_1 \le x) = 1 - 2\sum_{n=1}^{\infty}(4x^2n^2 - 1)e^{-2x^2n^2}$$

it follows that $P(W_1 \in [0.809, 1.862]) = 0.95$. Hence, Lo used the interval $[0.809, 1.862]$ as the 95% (asymptotic) acceptance region for testing the null hypothesis

$$H_0 = \{\text{ no long-range dependence, } H = \frac{1}{2}\}$$

against the composite alternative

$$H_1 = \{\text{ there is a long-range dependence, } \frac{1}{2} < H < 1\}$$

Unlike the graphical R/S method, which provides a rough estimate of the Hurst parameter, Lo's method only indicates whether long-range dependence is present or not. While Lo's results are asymptotic since they assume $N \to \infty$ and $\lambda \to \infty$, in practice the sample size N is finite and the question of the right choice of λ becomes dominant. The most problematic issue of that new standard deviation is the number of lags (λ) used, since for too small lags it omits lags which may be significant and therefore still biased the estimated Hurst exponent by the short-memory in the time series. On the other hand, if the used lag is too high, the finite sample distribution deviates significantly from its asymptotic limit destroying any long-range effect that might be in the data (see Teverovsky et al. [1999]). Even

[3] $W_0 = B(t) - tB(1)$, where B denotes a standard Brownian motion.

though Lo [1991] proposed an estimator of optimal lag based on the first-order autocorrelation coefficient $\hat{\rho}(1)$, given by,

$$\lambda^* = \left[\left(\frac{3N}{2} \right)^{\frac{1}{3}} \left(\frac{2\hat{\rho}(1)}{1 - (\hat{\rho}(1))^2} \right)^{\frac{2}{3}} \right]$$

where $[]$ is the greatest integer function, the optimal log of Lo is only correct if the underlying process is an $AR(1)$ process. Hence, we focus on the optimal lag devised by Chin [2008] which is based on the length of the sub-period, given by

$$\lambda^* = 4 \left(\frac{n}{100} \right)^{\frac{2}{9}}$$

which is recalculated for each length of specific sub-period n.

10.2.4.3 Removing short-term memory

Since the R/S values are random variables, normally distributed, we would expect that values of H to also be normally distributed. Using Monte Carlo simulation, Peters tested this hypothesis and concluded that $E[H]$ for i.i.d. random variables could be calculated from Equation (10.2.9) with variance $\frac{1}{T}$ where T is the total number of observations in the sample. Unfortunately, the variance for non-normally distributed distributions differs on an individual basis, such that the confidence interval is only valid for i.i.d. random variables. Testing the R/S analysis on different types of time series used to model financial economics, such as ARIMA and ARCH models, Peters [1994] proposed to reduce persistence bias by taking $AR(1)$ residuals on the original (integrated) time series and then follow all the steps of the original procedure with the residuals of the estimated autoregressive process. The $AR(1)$ residuals are calculated as

$$r_n = X_n - (c + c_1 X_{n-1}) \tag{10.2.10}$$

where r_n is the $AR(1)$ residual of X at time n, c is the intercept, and c_1 is the slope. Doing so, he has subtracted out the linear dependence of X_n on X_{n-1}. He found that none of the ARIMA models exhibited the Hurst effect of persistence, once short-term memory processes were filtered out. Further, even though ARCH and GARCH series could not be filtered, they did not exhibit long-term memory either. As a result, the problem of choosing the correct lag, discussed in Section (10.2.4.2), can be partly overcome by short-memory filtration. However, the problem of setting the right lag appears again, as the $AR(1)$ does not need to be satisfied for short-term memory filtration. In which case one should consider applying the $ARIMA(p, 1, q)$ model, but it can be misleading or inefficient on long time series (T > 10000). Consequently, it seems that both R/S and $M - R/S$ analysis can not perfectly separate short-term memory from the long-term one as even a little bias in estimated coefficients can lead to significant break of long-term memory structure. Note, Cajueiro et al. [2004] filtered the data by means of an $AR - GARCH$ procedure, which was intended to filter at the same time the short-range behaviour present in the time series and the volatility of returns. This is to avoid the long-range dependence that may be due to volatility effects, which are known to be persistent in financial time series.

10.2.5 Detecting periodic and nonperiodic cycles

10.2.5.1 The natural period of a system

Some technical analysts assume that there are some regular market cycles, hidden by noise or irregular perturbations, within financial markets. As a result, spectral analysis became a popular tool to break observed financial time series into sine waves. However, if economic cycles are nonperiodic, spectral analysis becomes an inappropriate tool for market cycle analysis, and we need a more robust method that can detect both periodic and nonperiodic cycles. Chaos theory can account for nonperiodic cycles having average duration, but unkown exact duration of a future cycle. Alternatively, R/S analysis can also perform that function. Analysing financial time series, Peters [1991-96] found that when we cross $N = 200$ ($\log(200) = 2.3$), the R/S observations begin to become erratic and random. This

characteristic of the R/S analysis can be used to determine the average cycle length of the system, being the length of time after which knowledge of the initial condition is lost. While long memory processes are supposed to last forever in theory, in practice there is a point in any nonlinear system where memory of initial conditions is lost, corresponding to the end of the natural period of the system. For financial time series with $H \neq \frac{1}{2}$, the long memory effect can impact the future for very long periods and goes across time scales. Since the impact decays with time, the cycle length measures how long it takes for a single period's influence to reduce to unmeasurable amounts. Therefore, one should visually inspect the data to see whether such a transition is occurring. One can assume that the long memory process underlying most systems is finite, and that the length of the memory depends on the composition of the nonlinear dynamic system producing the fractal time series. Hence the importance of visually inspecting the data in the log/log plot before measuring the Hurst exponent. It is then necessary to define the meaning of sufficient or adequate data when estimating the Hurst exponent. Peters [1991-96] suggested that we have enough data when the natural period of the system can easily be discerned. Since Chaos theory suggests that the data from 10 cycles are enough, if we can estimate the cycle length, we can use 10 cycles as a guideline.

10.2.5.2 The V statistic

As a periodic system corresponds to a limit cycle or a similar type of attractor, its phase space portrait is a bounded set. In the case of a sine wave, the time series is bounded by the amplitude of the wave. Since the range can never grow beyond the amplitude, the R/S values should reach a maximum value after one cycle. Mandelbrot et al. [1969a] performed a series of computer simulations and found that R/S analysis could detect periodic components. Considering sine wave, once it has covered a full cycle, its range stops growing since it has reached its maximum amplitude. Repeating the experiment on an infinite sum of series of sine waves with decreasing amplitude (fractal function due to Weirstrass), Peters showed that R/S analysis could find the primary cycle, as well as the underlying ones, as long as the number of subcycles was a small, finite number. In that case, looking at the log/log plot, the end of each frequency cycle and the beginning of the next can be seen as breaks or flattening in the R/S plot. In order to obtain a clear peak when R/S stops scaling at a faster rate than the square root of time, Hurst proposed a simple statistic called V to test for stability, giving a precise measure of the cycle length even in presence of noise. The ratio

$$V_n = \frac{1}{\sqrt{n}}(R/S)_n$$

would result in a horizontal line if the R/S statistic was scaling with the square root of time. This ratio is in general symptomatic of the existence of a periodic or non-periodic cycle. That is, a plot of V versus $\log n$ is flat for an independent, random process, downward sloping for an antipersistent process, and upward sloping for a persistent one. By plotting V on the y axis and $\log n$ on the x axis, the breaks occur when the V chart flattens out, indicating that the long-memory process has dissipated at those points.

10.2.5.3 The Hurst exponent and chaos theory

A nonperiodic cycle has no absolute frequency, but an average one, and can have two sources

1. it can be a statistical cycle, exemplified by the Hurst phenomena of persistence and abrupt changes in direction
2. it can be the result of a nonlinear dynamic system, or deterministic chaos

We saw that the Hurst process can be described as a biased random walk, where the bias can change abruptly in direction or magnitude. The abrupt changes in bias is characterised by the random arrival of the joker in Hurst pack of cards which is not predictable. Hence, the Hurst cycles have no average length, and the log/log plot continues to scale indefinitely. Alternatively, nonlinear dynamical systems are deterministic systems that can exhibit erratic behaviour. In chaos, maps are systems of iterated difference equations, such as the Logistic equation that can generates statistically random numbers, deterministically. Chaotic flows, which are continuous systems of interdependent differential equations, are used to model large ecosystems. The best known system of that type is the Lorenz attractor. A simpler

system of nonlinear equations was derived by Mackey et al. [1977], where over and under production tend to be amplified, resulting in nonperiodic cycles. It is a delay differential equation with an infinite number of degrees of freedom, much like in the markets. It is given by

$$X_t = 0.9X_{t-1} + 0.2X_{t-n}$$

where the degree of irregularity, and as such the fractal dimension, depend on the time lag n. Varying the cycle used, Peters showed that the R/S analysis can detect different cycle length and estimate the average length of a nonperiodic cycle. We consider two types of noise in dynamical systems

1. observational or additive noise where the noise is a measurement problem (the recorded values have noise increment added)
2. dynamical noise when the system interprets the noisy output as an input (the noise invades the system)

Adding one standard deviation of additive noise to the system, Peters showed that the V statistic was unaffected and concluded that R/S analysis was robust with respect to noise.

10.2.6 Possible models for FMH

10.2.6.1 A few points about chaos theory

Fractional Brownian motion (fBm) has some characteristics conforming with FMH such as statistical self-similarity over time and persistence, creating trends and cycles. However, in fBm, there is no reward for long-term investing as the term structure of volatility, in theory, does not stop growing. Further, fBm are more concerned with explaining crowed behaviour than to understand investor expectations about the economy. We have already discussed measurement noise (or observational noise) and system noise (or dynamical noise). Since, at longer frequencies, the market reacts to economic and fundamental information in a nonlinear fashion, some authors added nonlinear dynamical system to fBm to satisfy all aspects of FMH. These systems called chaotic systems allows for nonperiodic cycles and bounded sets (attractors). Chaotic systems are nonlinear feedback systems subject to erratic behaviour, amplification of events, and discontinuities. At least two requirements must be satisfied for a system to be considered chaotic

1. the existence of a fractal dimension
2. a characteristic called sensitive dependence on initial conditions

A chaotic system is analysed in a phase space consisting of one dimension for each factor defining the system. For instance, a pendulum is made of two factors defining its motion, velocity and position. In the theoretical case of no friction, the pendulum would swing back and forth forever, leading to a phase plot being a closed circle. However, in presence of friction, or damping, after each swing the pendulum would slow down with amplitude eventually stopping, and the corresponding phase plot would spiral into the origin, where velocity and position are zero. In the case of the Lorenz attractor, the phase plot never repeats itself, but it is bounded by the owl eyes shape. The lines within the attractor represent a self-similar structure caused by repeated folding of the attractor. As the lines do not intersect, the process will never completely fill the space, and its dimension is fractal. Chaotic systems being characterised by growth and decay factor, the attractor is bounded to a particular region of space. A trip around the attractor is an orbit. In the case of two orbits, as each one of them reaches the outer bound of the attractor, it returns toward the centre and the divergent points will be close together again. This is the property of sensitive dependence on initial conditions. As we can not measure current conditions to an infinite amount of precision, we can not predict where the process will go in the long term. The rate of divergence, or the loss in predictive power, can be characterised by measuring the divergence of nearby orbits in phase space. It is called the Lyapunov exponent. and is measured for each dimension in phase space. While a positive rate means that there are divergent orbits, when combined with a fractal dimension, it means that the system is chaotic. There must also be a negative exponent to measure the folding process, or the return to the attractor. The Lyapunov exponent is given by

$$L_i = \lim_{t\to\infty} \frac{1}{t} \log_2 \frac{p_i(t)}{p_i(0)}$$

where L_i is the Lyapunov exponent for dimension i, and $p_i(t)$ is the position in the ith dimension, at time t. This equation measures how the volume of a sphere grows over time t, by measuring the divergence of the two points $p(t)$ and $p(0)$ in dimension i. There is a certain similarity with the R/S analysis and to the fractal dimension calculation, all being concerned with scaling. However, chaotic attractors have orbits decaying exponentially rather than through power laws. Note, as the phase space includes all of the variables in the system, its dimensionality is dependent on the complexity of the system. The next higher integer to the fractal dimension tells us the minimum number of dynamic variables needed to model the dynamics of the system, placing a lower bound on the number of possible degrees of freedom. We saw that the fractal dimension, D, could be estimated by Equation (10.1.2) for a fractal embedded in two-dimensional space. For a higher-dimensional attractor, we need to use hyperspheres of higher dimensionality. Alternatively, Grassberger et al. [1983] proposed the correlation dimension as an approximation of the fractal dimension which uses the correlation integral $C_m(d)$ where m is the dimension, and d a distance. We let the correlation integral be the probability that any two points are within a certain length, d, apart in phase space. As we increase d, the probability scales according to the fractal dimension of the phase space. It is given by

$$C_{m,N}(d) = \frac{1}{N^2} \sum_{i,j=1}^{N} \mathcal{H}(d - |X_i - X_j|) \, , i \neq j$$

where $\mathcal{H}(x) = 1$ if $d - |X_i - X_j| > 0$ and 0 otherwise, N is the number of observations, d is the distance, and C_m is the correlation integral for dimension m. The function, $\mathcal{H}(\bullet)$, counts the number of points within a distance, d, of one another. Further, the C_m should increase at the rate d^D, with D the correlation dimension of the phase space, which is closely related to the fractal dimension.

10.2.6.2 Using R/S analysis to detect noisy chaos

Studying the attractor by Mackey et al. [1977], Peters showed that R/S analysis was a robust way of detecting noisy chaos. While the continuous, smooth nature of the chaotic flow leads to very high Hurst exponent, gradually adding one standard deviation of white, uniform noise to the system would bring the Hurst exponent down, leading to an index of noise. Note, in markets, system noise is more likely to be a problem than additive noise. Because of the problem of sensitive dependence on initial conditions, system noise increases the problem of prediction. However, the impact of system noise on the Hurst exponent is similar to additive noise. Since the R/S analysis can distinguish cycles, Peters performed R/S analysis on the Mackey-Glass equation with one standard deviation of system noise incorporated and observed a cycle as the log-log plot crossed over to a random walk. This means that

1. the process can be fractional Brownian motion with a long but finite memory, or
2. the system is a noisy chaotic system, and the finite memory length measures the folding of the attractor.

In the latter, the diverging of nearby orbits in phase space means that they become uncorrelated after an orbital period, so that the memory process ceases after an orbital cycle. That is, the finite memory length becomes the length of time it takes the system to forget its initial conditions. Since both explanations are possible, a true cycle should not be dependent on sample size, in which case we should be examining noisy chaos and not fractional noise. If we have sufficient data to obtain ten cycles of observations we can estimate the largest Lyapunov exponent. If that exponent is positive, we have strong evidence that the process is chaotic. Confidence level can be increased if the inverse of the largest Lyapunov exponent is approximately equal to the cycle length. Note, this approach is problematic in presence of small data sets. In that case, Peters proposed to consider both the R/S analysis and the BDS test which measures the statistical significance of the correlation dimension calculations (see Brock et al. [1987]). Even though it is a powerful test for distinguishing random systems from deterministic chaos or nonlinear stochastic systems, it can not distinguish between a nonlinear deterministic system and a nonlinear stochastic system. But it finds nonlinear

dependence. According to the BDS test, the correlation integrals should be normally distributed if the system under study is independent. Note, in real life we do not know the factors involved in the system, and we do not know how many of them there are, as we only observe stock price changes. However, a theorem from Takens [1981] states that we can reconstruct the phase space by lagging the one time series we have for each dimension we think exists. In the case where the number of embedding dimensions is larger than the fractal dimension, then the correlation dimension stabilises to one value. The correlation integral calculates the probability that two points that are part of two different trajectories in phase space, are d units apart. Assuming the X_i in the time series X (with N observations) are independent, we lag the series into n histories by creating a phase space of dimension n from the time series X. We then calculate the correlation integral $C_{n,N}(d)$. As N approaches infinity we get $C_{n,N}(d) \to C_1(d)^n$ with probability 1. This is the typical feature of random processes. The correlation integral simply fills the space of whatever dimension it is placed in. Brock et al. showed that $|C_{n,N}(d) - C_{1,N}(d)^n|\sqrt{N}$ is normally distributed with a mean of 0. The BDS statistic, w, given by

$$w_{n,N}(d) = |C_{n,N}(d) - C_{1,N}(d)^n| \frac{\sqrt{N}}{S_{n,,N}(d)}$$

where $S_{n,N}(d)$ is the standard deviation of the correlation integrals, is also normally distributed. The BDS statistic, w, has a standard normal probability distribution, so that when it is greater than 2 we can reject, with 95% confidence, the null hypothesis that the system under study is random. However, it will find linear as well as nonlinear dependence in the data, so that it is necessary to take $AR(1)$ residuals for this test. Further, the dependence can be stochastic (such as the Hurst process, or GARCH), or it can be deterministic (such as chaos). We must determine the radius, or distance, d, as well as the embedding dimension m. LeBaron [1990] and Hsieh [1989] did extensive tests on stock prices and currencies using $m = 6$ and $d = \frac{1}{2}$ standard deviation of the data. Using these settings Peters tested his approach on the Macky-Glass equation with and without noise (both observational noise and system noise) before applying it to the Dow 20-day and 5-day series obtaining significant results. He concluded that the Dow was chaotic in the long term following economic cycle, and that currencies were fractional noise processes, even in the long term.

10.2.6.3 A unified theory

Peters found that, except for currencies, noisy chaos was consistent with the long-run, fundamental behaviour of markets, and fractional Brownian motion was more consistent with the short-run, trading characteristic. Both being consistent with the fractal market hypothesis (FMH). To unify both theory, Peters examined the relationship between fractal statistics and noisy chaos. He first looked at the frequency distribution of changes on the Mackey-Glass equation adding observational and system noise respectively, and concluded that there was a striking similarity between the system noise frequency distributions and the capital market distributions. Second, looking at the term structure of volatility, when market returns are normally distributed, the volatility should increase with the square root of time. However, stocks, bonds, and currencies all have a volatility term structure increasing at a faster rate, which is consistent with the properties of infinite variance distributions and fBm. Note, for a pure fBm process, such scaling should increase for ever. However, while currencies appear to have no limit to their scaling, Peters showed that US stocks and bonds were bounded at about four years which is similar to the four-year cycle he obtained with the R/S analysis. The connection being that in a chaotic system the attractor is a bounded set. That is, after the system has completed one cycle, changes stop growing. After testing that chaotic systems (using the Mackey-Glass equation with added noise) have bounded volatility term structures, Peters postulated one could test for the presence of nonperiodic cycles by using volatility term structures. The test was repeated with the Lorenz and Rosseler attractors and similar results were obtained. Note, currencies did not have this bounded characteristic. Third, defining the sequential standard deviation as the standard deviation of the time series as we add one observation at a time, assuming normality, the more observations we have the more the sequential standard deviation would tend to the population standard deviation. Similarly, if the mean is stable and finite, the sample mean would eventually converge to the population mean. Testing the Dow Jones, Peters found evidence of convergence after about 100 years of data, meaning that the process is closer to an infinite variance than to a finite one in shorter periods. He also found that the sequential mean converged more

rapidly, and looked more stable. A fractal distribution is a good candidate to reproduce these desired characteristics. Peters repeated the test on chaotic systems and found that, without noise, the Mackey-Glass equation was persistent with unstable mean and variance, while with noise, both observational and system, the system was closer to market series, but not identical. Analysing the Hurst exponent of a chaotic system with noise, Peters obtain $H = 0.64$ compared with $H = 0.72$ using R/S analysis which a big discrepancy. At last, Peters checked cycle lengths of dynamical systems and the self-similarity of noisy chaos and found that it had many desirable attributes. Trying to unify GARCH, fBm, and chaos, Peters postulated

- In the short term, markets are dominated by trading processes, which are fractional noise processes. They are, locally, members of the ARCH family of processes characterised by conditional variances.
- Globally, the process is a stable Levy (fractal) distribution with infinite variance. As the investment horizon increases, it approaches infinite variance behaviour.
- In the very long term (periods longer than four years), the market are characterised by deterministic nonlinear systems or deterministic chaos. Nonperiodic cycles arise from the interdependence of the various capital markets among themselves, as well as from the economy.

That is, short-term trading is dominated by local ARCH and global fractal, while long-term trading is tied to fundamental information and deterministic nonlinearities.

10.2.7 Revisiting the measures of volatility risk

10.2.7.1 The standard deviation

We saw in Section (3.4) that the volatility σ_t is usually defined by

$$r_t = \sigma_t \epsilon_t$$

where r_t is the rate of return and ϵ_t are identical independently distributed random variables with vanishing average and unitary variance. In general, the distribution of the ϵ_t is chosen to be the normal Gaussian, and we also assume the probabilistic independence between σ_t and ϵ_t. Hence, the returns series can be considered as the realisation of a random process based on a zero mean Gaussian, with a standard deviation σ_t changing at each time step. Standard deviation measures the probability that an observation will be a certain distance from the average observation. The larger this number is, the wider the dispersion, meaning that there is a high probability of large swings in returns. In the efficient market hypothesis (EMH), the variance is assumed to be finite, that is, the standard deviation tends to a value that is the population standard deviation. Since the standard deviation is higher if the time series of prices is more jagged, the standard deviation became a measure of the volatility of the stock prices. The scaling feature of the normal distribution being referred to as the $T^{\frac{1}{2}}$ Rule, where T is the increment of time, the investment community annualises risk with that rule. For example, the monthly standard deviation is annualised by multiplying it by $\sqrt{12}$. Turner et al. [1990] found that monthly and quarterly volatility were higher than they should be compared with annual volatility, but that daily volatility was lower. Shiller [1989] found excessive market volatility challenging the idea of rational investors and the concept of market efficiency. Engle [1982] proposed the autoregressive conditional heteroskedastic (ARCH) model where the volatility is conditional upon its previous level. As a result, high volatility levels are followed by more high volatility, and vice versa for low volatility, which is consistent with the observation that the size of price changes (ignoring sign) is correlated. Hence, standard deviation is not a standard measure, at least over the short term.

While variance is stable and finite for the normal distribution alone, it is particularly unstable when the capital markets are described by the Stable Paretian family of distributions described in Section (10.1.2.4). In the family of stable distribution, the normal distribution is a special case that exists when $\alpha = 2$, in which case the population mean and variance do exist. Infinite variance means that there is no population variance that the distribution tends

to at the limit. When we take a sample variance, we do so, under the Gaussian assumption, as an estimate of the unknown population variance. For instance, Sharpe calculated betas from five years' monthly data to get a statistically significant sample variance needed to estimate the population variance. However, five years is statistically significant only if the underlying distribution is Gaussian. Otherwise, for $\alpha < 2$, the sample variance tells nothing about the population variance as there is no population variance. Sample variances would be expected to be unstable and not tend to any value, even as the sample size increases. For $\alpha \leq 1$, the same goes for the mean, which also does not exist in the limit. Various studies showed that the Dow was characterised by a stable mean and infinite memory, in the manner of stable Levy fractal distributions. However, a market characterised by infinite variance does not mean that the variance is truly infinite, but as discussed in Section (10.2.5), there is eventually a time frame where fractal scaling ceases to apply, corresponding to the end of the natural period of the system. Hence, for market returns, there could be a sample size where variance does become finite, but it might be very big (hundreds of years).

10.2.7.2 The fractal dimension as a measure of risk

We saw in the previous section that standard deviation measures the probability that an observation will be a certain distance from the average observation. However, this measure of dispersion is only valid if the underlying system is random. If the observations are correlated (or exhibit serial correlation), then the usefulness of standard deviation as a measure of dispersion is considerably weakened. Hence, while volatility is stated as the statistical measure of standard deviation of returns, that measure of comparative risk is of questionable usefulness. When stock returns are not normally distributed, the standard deviation becomes an invalid measure of risk. For instance, Peters [1991-96] described two stocks, one trendless and the other one trended, with similar volatilities that can have very different patterns of returns. See results in Table (10.1). It would make a lot more sense to compare different stocks by noting their fractal dimensions, D, taking values between 1 and 2.

Table 10.1: Standard deviation and fractal dimension

Observation	S1	S2
1	2	1
2	-1	2
3	-2	3
4	2	4
5	-1	5
6	2	6
cumulative return	1.93	22.83
standard deviation	1.70	1.71
fractal dimension	1.42	1.13

The first stock, $S1$, has a fractal dimension close to that of a random walk, while the second stock, $S2$, has a fractal dimension close to that of a line. Given the relationship between the fractal dimension and the Hurst exponent in Equation (10.1.3), the first stock, $S1$, has the exponent $H = 0.58$ and the second stock, $S2$, has the exponent $H = 0.87$ showing long-memory. If returns in the capital markets exhibit Hurst statistics, then their probability distribution is not normal and the random walk does not apply. As a consequence, much of quantitative analysis collapses, especially the CAPM and the concept of risk as standard deviation or volatility. On the other hand, since the fractal dimension is higher if the time series of prices is more jagged, and since it is well defined for all time series, the fractal dimension is a better measure of the volatility of the stock prices. As a result, assuming the first moment of a time series to be defined, we can replace the standard deviation in the Sharpe type investment statistics described in Section (2.4.2) with the fractal dimension. Hence, the Shape ratio given in Equation (2.4.16) becomes

$$M_{DR} = \frac{E[R_a - R_b]}{D}$$

where R_a is the asset return and R_b is the return of a benchmark asset such as the risk free rate or an index. This ratio, called Fractal Risk-Return, measures the excess return per unit of risk in an investment asset or a trading strategy. It correctly characterise how well the return of an asset compensates the investor for the risk taken. The higher the FRR ratio, the better the combined performance of risk and return. Similarly, we can define the fractal Information ratio as the mean over the fractal dimension of a series of measurements. We can also express the ratio in units of percent returns by modifying the Modigliani risk-adjusted performance measure with D_P the fractal dimension of the asset/portfolio return and D_M the fractal dimension of a benchmark return. Further, we can compute D_+ the fractal dimension of positive returns, and D_- the fractal dimension of negative returns and modify the Omega-Sharpe ratio, getting

$$M_{DoR} = \frac{r_p - r_T}{D_-}$$

where r_T is a minimal acceptable return. In that measure the portfolio manager will only be penalised for variability in negative returns. Similarly, the Sortino's upside potential ratio becomes

$$M_{DUPR} = \frac{\frac{1}{n}\sum_{i=1}^{n} \max(r_i - r_T, 0)}{D_-}$$

where only returns above a target value are considered.

10.3 Hurst exponent estimation methods

We saw in Section (10.1.2) that monofractal and multifractal structures of financial time series are particular kind of scale invariant structures. Since the development of the rescaled range analysis (R/S) detailed in Section (10.2), a large number of fractal quantification methods called Wavelet analysis and Detrending Methods (DMs) have been proposed to accomplish accurate and fast estimates of the Hurst exponent H in order to investigate correlations at different scales in dimension one. While the R/S analysis makes the rough approximation that the trend is the average in each segment under consideration, the former uses spectral analysis, and the latter considers more advanced methods to detrend the series. Examples of the latter are detrended fluctuation analysis (DFA) and detrending moving average analysis (DMA). There exists many more methods for estimating the Hurst exponent such as the aggregated variance, the differenced variance, absolute values, boxed periodigram, Higuchi, modified periodigram, Whittle and local Whittle methods. As the finite sample properties of Hurst estimators revealed quite different from their asymptotic properties, various authors undertook empirical comparisons of etimators of the hurst exponent H, and the differencing parameter d. For instance, Taqqu et al. [1995] studied nine estimators for a single series length of $10,000$ data points, five values of both H and d, and 50 replications. They graphically represented the results in box plots with vertical axis indicating the deviation from the nominal value of H. Later, Taqqu et al. [1999] found that Whittle estimator was the most precise, and Clegg [2006] found that local Whittle and wavelet estimates provided the best estimates. A comparatively smaller number of methods have been proposed to capture spatial correlations operating in $d \geq 2$. We are now going to describe some of these methods.

10.3.1 Estimating the Hurst exponent with wavelet analysis

Wavelet transforms described in Appendix (G.3) can be used to characterise the scaling properties of self-similar fractal and multifractal objects (see Muzy et al. [1991]). In relation to financial time series, wavelet transform analysis has been extensively used in the determination of the scaling properties of fractional Brownian motion (fBm). We saw earlier that the smoothness of the fBm function, denoted $B_H(t)$ increases with the Hurst exponent H varying in the range $0 < H < 1$. Fractional Brownian motions are nonstatinary random processes where the standard deviation, σ, of the fBm trace deviation ΔB_H taken over a sliding window of length T scales as

$$\sigma \propto T^H$$

One method used to determine the exponent H from data suspected of fBm scaling is the Fourier power spectrum $P_F(f)$, for which a fBm should scale as

$$P_F(f) \propto f^{-\beta}, \beta = 2H + 1$$

and a logarithmic plot of power against frequency allows H to be determine from the slope of the spectrum. There has been much research carried out using Fourier transforms. Wavelet power spectra, $P_W(f)$, exhibit similar scaling and can also be used to determine H. However, Simonsen et al. [1998] showed that the wavelet spectrum is in general much smoother due to the finite bandwidth of the spectral components associated with the wavelets, which is a big advantage when only a limited number of data sets are available. Analysing two real data sets with discrete wavelet coefficient based method, Simonsen et al. showed that the mean absolute discrete wavelet coefficient values scale as

$$< |T_{m,n}| >_m \propto a_m^{(H+\frac{1}{2})}$$

where $< |T_{m,n}| >_m$ is the mean absolute value of the wavelet coefficients at scale a_m defined in Appendix (G.3.2). Using dyadic grid Daubechies wavelets, $a_m \propto 2^m$, and setting the maximum a scale to unity, they obtained the average wavelet coefficient (AWC) function. Analysing the stock market index for shares taken from the Milan stock exchange over two and a half year period with both the AWC and the Fourier analysis, they concluded that only when a small number of examples are available, the wavelet method outperforms the Fourier method. Another method for computing H directly from the wavelet coefficients is to use the scaling of the variance $< T_{m,n}^2 >_m$ of discrete wavelet coefficients at scale index m defined in Appendix (G.3.4). Since the coefficient variance is related to the power spectrum as

$$P_W(f_m) \propto < T_{m,n}^2 >_m$$

combining the two expressions and noting that the frequency f is inversely proportional to the wavelet scale $a(= 2^m)$, we get the scaling relationship

$$< T_{m,n}^2 >_m \propto a_m^{(2H+1)}$$

It should be stressed that this result is valid for any self-similar process with stationary increments ($\beta = 2H + 1$) and long-range dependent processes ($\beta = 2H - 1$) corresponding to fBm and fGn, respectively. We let σ_m^2 be the variance of the discrete wavelet coefficients at index m, and rewrite the relationship as

$$\sigma_m^2 \propto a_m^{(2H+1)}$$

and taking the square root on both sides, we get

$$\sigma_m \propto a_m^{(H+\frac{1}{2})}$$

Note, both the mean absolute value of the coefficients and the standard deviation of the coefficients are first order measures of spread. Furthermore, for an orthonormal multiresolution expansion using a dyadic grid, the scale a is proportional to 2^m, so that we can take base 2 logarithms of both sides of the above expression to get

$$\log_2(\sigma_m^2) = (2H + 1)m + \text{ constant}$$

where the constant depends both on the wavelet used and the Hurst exponent. Flandrin [1992] defined the constant and found an explicit form in the case of Haar wavelet. Abry et al. [2000] considered a wavelet-based analysis tool, called the Logscale Diagram consisting of the above log-log plot together with confidence intervals assigned to each $\log_2(\sigma_m^2)$. Note, we can also consider the wavelet scaling energy E_m defined in Appendix (G.3.4) and having the fractal scaling law

$$\log_2(E_m) = 2Hm + \text{ constant}$$

that is

$$E_m \propto a^{2H}$$

which is similar to the scaling of σ, the standard deviation of the fBm trace deviation. This is because E_m is a measure of the scale dependent variance of the signal. Note, the continuous transforms, $T(a,b)$, exhibit the same scaling law

$$E(a) \propto a^{2H}$$

but the a scale parameter is continuous and the slope of the plot of $\log(E(a))$ against $\log(a)$ is used to find H. We briefly describe a methodology proposed by Abry et al. [1993] [1998] to estimate the Hurst exponent

1. Compute the Discrete Wavelet Transform (DWT) of the signal and compute the second-order moment $< T_{m,n}^2 >_m$ at each scale m.
2. Perform a weighted-least squares fit between the scales j_1 and j_2 of the second-order moment on the scales m

$$\hat{H}(j_1,j_2) = \frac{1}{2}\Big[\frac{(\sum_{j=j_1}^{j_2} S_j\eta_j)(\sum_{j=j_1}^{j_2} S_j) - (\sum_{j=j_1}^{j_2} S_j j)(\sum_{j=j_1}^{j_2} S_j\eta_j)}{(\sum_{j=j_1}^{j_2} S_j)(\sum_{j=j_1}^{j_2} S_j j^2) - (\sum_{j=j_1}^{j_2} S_j j)^2} - 1\Big]$$

   where $\eta_j = \log_2(< T_{j,n}^2 >_j)$ and $S_j = \frac{n(\log 2)^2}{2^{j+1}}$ is the inverse of the theoretical asymptotic variance of η_j.

It has been shown that this estimator of H is asymptotically unbiased and efficient. Nonetheless, further improvements have been carried out by Abry et al. [1999], defined as follow

1. At each scale level j, define

$$g_j = \frac{\Psi(2^{N-j-1})}{\log(2)} \text{ and } V_j = \frac{\xi(2,2^{N-j-1})}{(\log(2))^2}$$

   where $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ and $\xi(z,v)$ is the generalised Riemann Zeta function.
2. Perform the least square regression of $\log_2(< T_{j,n}^2 >_j) - g_j$ on the scales j weighted by V_j to find a slightly better estimate of the exponent H.

10.3.2 Detrending methods

Following the R/S analysis, we take the returns $\{x(t)\}$ of a time series of length T, having fluctuations similar to the increments of a (multifractal) random walk. It should be converted to a (multifractal) random walk by subtracting the mean value and integrating the time series. Hence, we construct the integrated series, or profile,

$$X(t) = \sum_{i=1}^{t}\big(x(i) - \overline{x}\big) \tag{10.3.11}$$

for $t = 1, ..., T$ where $\overline{x}$ denotes the mean of the entire time series. We then divide the integrated series into A adjacent sub-periods of length, n, such that $An = T$.

10.3.2.1 Detrended fluctuation analysis

Improving on the R/S analysis, Peng et al. [1994] introduced the detrended fluctuation analysis (DFA) to investigate long-range power-law correlations along DNA sequences. The local detrended fluctuation function is calculated from

$$F^2_{DFA}(a) = \frac{1}{n}\sum_{i=1}^{n} |X(i,a) - Z(i,a)|^2 \, , a = 1, ..., A$$

where $X(i,a) = X((a-1)n+i)$, and $Z(t,a) = c_{1,a}t + c_{2,a}$ is a linear local trend, for sub-period $a = 1, .., A$, computed with a standard linear least-square fit. To improve the computation of the local trend, some authors considered the polynomial fit of order l, noted $X_{n,l}$, of the profile which is estimated for each sub-period. Hence, we get the local detrended signal (or residual variation)

$$Y(t,a) = X(t,a) - X_{n,l}(t,a) \, , a = 1, .., A \tag{10.3.12}$$

where $X_{n,l}(t,a) = \sum_{k=0}^{l} c_{k,a}t^{l-k}$ and $c_{k,a}$ is a constant in the, a, sub-period. Linear, quadratic, cubic, or higher order polynomials can be used in the fitting procedure (conventionally called DFA1, DFA2, DFA3, ...) (see Figure 10.7). The local detrended fluctuation function, $F^2_{DFA}(a)$, becomes

$$F^2_{DFA}(a) = \frac{1}{n}\sum_{i=1}^{n} |Y(i,a)|^2 \, , a = 1, ..., A$$

which is the local mean-square (MS) variation of the time series. $F^2_{DFA}(a)$ is also called the detrended variance of each segment. Note, it is sometime expressed in terms of the scale-dependent measure $\mu_n(t,a) = F_{DFA}(a)$ which is the local root-mean-square (RMS) variation of the time series. Since the detrending of the time series is done by the subtraction of the polynomial fits from the profile, different order DFA differ in their capability of eliminating trends in the series. Thus, one can estimate the type of polynomial trend in a time series by comparing the results for different detrending orders of DFA. Averaging the $F_{DFA}(\bullet)$ over the A sub-periods gives the overall fluctuation, or overall RMS, $F_{DFA}(\bullet)$, as a function of n

$$F_{DFA}(n) = \Big(\frac{1}{A}\sum_{i=1}^{A} F^2_{DFA}(i)\Big)^{\frac{1}{2}}$$

The fast changing fluctuations in the time series X will influence the overall RMS, $F_{DFA}(n)$, for segments with small sample sizes (small scale), whereas slow changing fluctuations will influence $F_{DFA}(n)$ for segments with large sample sizes (large scale), Hence, we should compute $F_{DFA}(n)$ for multiple segments sizes (multiple scales) to emphasise both fast and slow evolving fluctuations influencing the structure of the time series. Doing so, we can be relate $F_{DFA}(n)$ to the Hurst exponent as follow

$$F_{DFA}(n) \sim n^H$$

so that an exponent $H \neq \frac{1}{2}$ in a certain range of n values implies the existence of long-range correlations in that time interval. A straight line can well fit the data between the interval $\log n = 1$ and 2.6 which is called the scaling range. Outside the scaling range, the error bars are larger due to the finite size effects, and/or the lack of numerous data. Hence, the first few points at the low end of the log-log plot of $F_{DFA}(n)$ against n should be discarded as in this region the detrending procedure removes too much of the fluctuation. Similarly, for large n there a few boxes, A, for a proper averaging to be made.

The main advantage of working with fluctuations around the trend rather than a range of signals is that we can analyse non-stationary time series. That is, DFA avoid spurious detection of correlations that are artifacts of nonstationarities in the time series. Even though the DFA can be based on different polynomial fits, trend can be constructed in a variety of ways such as Fourier transforms (see Chianca et al. [2005]), empirical mode decomposition (see Janosi

et al. [2005]), singular value decomposition (see Nagarajan [2006]), different types of moving averages (see Alessio et al. [2002]), and others.

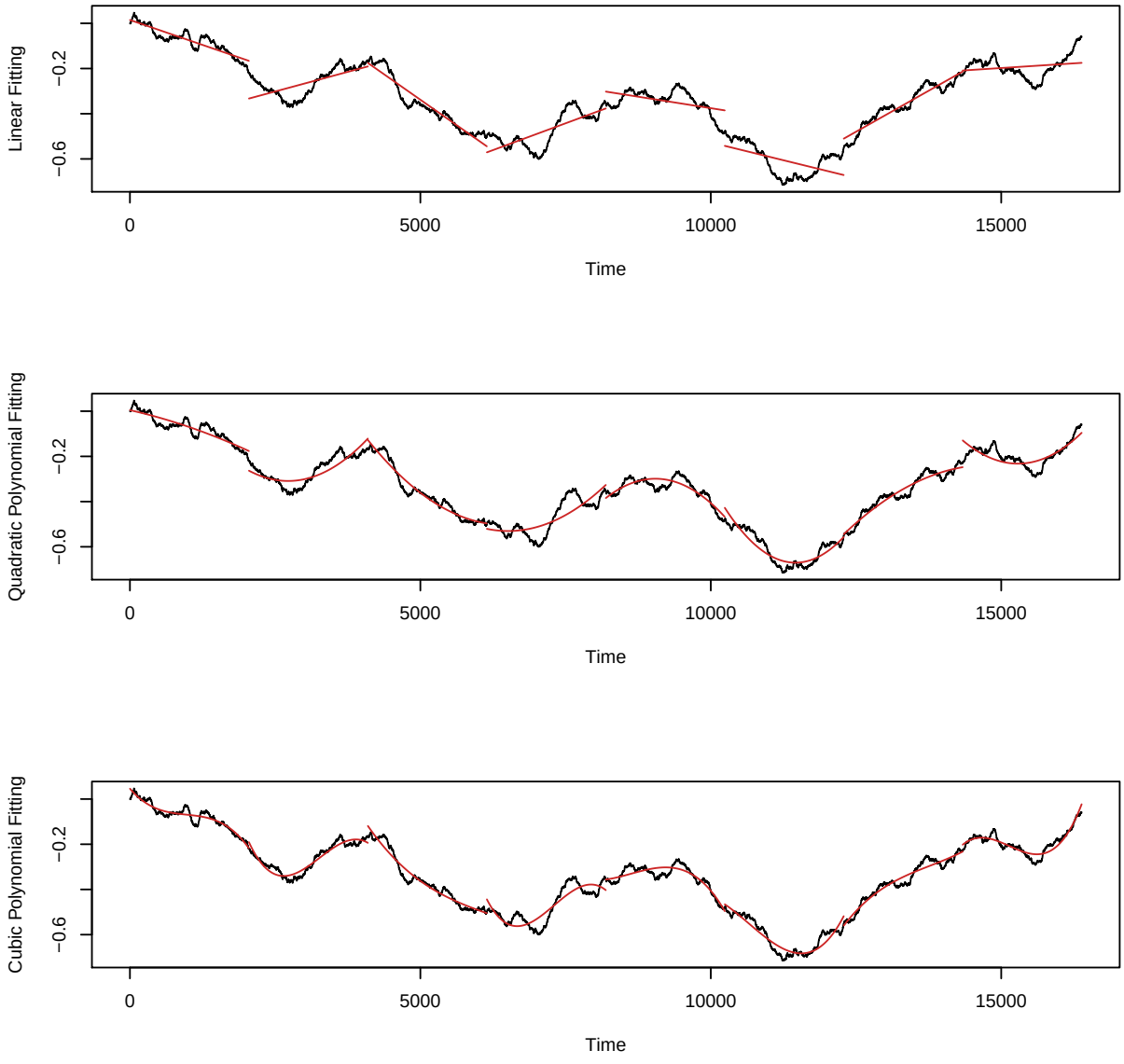


Figure 10.7: Computation of the second-order difference between the time series and its polynomial fit for different polynomial interpolation order.

10.3.2.2 A modified DFA

In the case of a linear polynomial, $X_{n,1}(t,a)$, Taqqu et al. [1995] proved that for large n, the averaged sample variances in the DFA has a resulting number proportional to n^{2H} for fractional Gaussian noise (fGn) and ARFIMA models. That is,

$$E[F^2_{DFA}(a)] = E[\frac{1}{n}\sum_{i=1}^{n}\big(X(t,a) - X_{n,1}(t,a)\big)^2] \sim C_H n^{2H} \text{ as } n \to \infty$$

where

$$C_H = \big(\frac{2}{2H+1} + \frac{1}{H+2} - \frac{2}{H+1}\big)$$

To avoid some of the difficulties associated with choosing an appropriate interval within which to perform the linear fit, Costa et al. [2003] modified the approach described in Section (10.3.2.1). They rescaled the overall fluctuation $F_{DFA}(n)$ by normalising it with the overall standard deviation, S, of the original time series, and adjusted the parameter H so as to obtain the best agreement between the theoretical curve $F_H(n)$ given above in the case of a fBM, and the empirical data for $F_{DFA}(n)$. Rather than running a nonlinear regression, they simply varied H incrementally and decided visually when the theoretical curve best matches the empirical data.

10.3.2.3 Detrending moving average

In signal processing, the moving average $\overline{X}_\lambda(t) = \frac{1}{\lambda}\sum_{k=0}^{\lambda-1} X(t-k)$ of a time series $X(t)$, with the time window λ, is a well-known low-pass filter (see details in Appendix (D.1.2)). In other words, it removes the low frequency motions from the signal. If X increases (resp. decreases) with time, then $\overline{X} < X$ (resp. $\overline{X} > X$), thus, the moving average captures the trend of the signal over the considered time window λ. Vandewalle et al. [1998] observed that the density ρ of crossing points between any two moving averages with time windows respectively equal to λ_1 and λ_2 can be expressed as

$$\rho = \frac{1}{\lambda_2}\big((\Delta\lambda)(1-\Delta\lambda)\big)^{H-1}$$

where $\Delta\lambda = \frac{\lambda_2-\lambda_1}{\lambda_2}$ and $\lambda_2 >> \lambda_1$. This density is fully symmetric, it has a minimum in the middle of the $\Delta\lambda$ interval and diverges for $\Delta\lambda = 0$ and for $\Delta\lambda = 1$, with an exponent which is the Hurst exponent. Hence, using the moving-average method, they could extract the Hurst exponent of correlated time series, obtaining results comparable to DFA. As a result, self-affine signals characterised by the Hurst exponent H can be investigated through mobile average (see Ausloos [2000]). Hence, the density ρ is a measure of long-range power-law correlation in the signal. Since in the limit of $\lambda \to 0$ we get $\overline{X}_\lambda(t) \to X(t)$, then the crossing points correspond to the zeroes of the first-order difference between $X(t)$ and $\overline{X}_\lambda(t)$. Alessio et al. [2002] looked at the properties of the second-order difference between the integrated process $X(t)$ and the moving average $\overline{X}_\lambda(t)$ and proposed the detrending moving average (DMA), where one does not need to divide the time series into sub-periods, but rather use deviations from the moving average of the whole series. Again, we take the returns $\{x(t)\}$ of a time series of length T and compute the integrated series as in Equation (10.3.11). The detrended fluctuation is defined as

$$F^2_{DMA,\lambda} = \frac{1}{T-\lambda}\sum_{t=\lambda}^{T}\big(X(t) - \overline{X}_\lambda(t)\big)^2$$

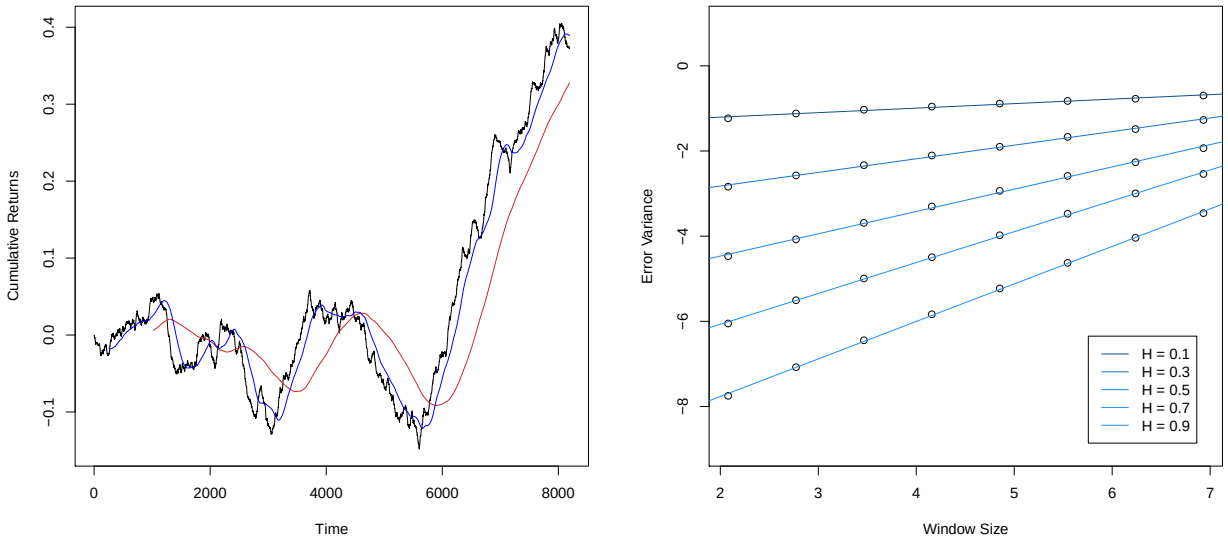
which is the variance of $X(t)$ with respect to the moving average $\overline{X}_\lambda(t)$. Alessio et al. computed the moving averages $\overline{X}_\lambda(t)$ with different values of λ ranging from 2 to λ_{max} where $\lambda_{max} \approx \frac{T}{3}$ (see Figure 10.8(a)). The detrended fluctuation was then calculated over the time interval $[\lambda_{max}, T]$

$$F^2_{DMA,\lambda} = \frac{1}{T - \lambda_{max}} \sum_{t=\lambda_{max}}^{T} \left(X(t) - \overline{X}_\lambda(t)\right)^2$$

Plotting the values of $F_{DMA,\lambda}$ against λ on a log-log axes, they showed that the fluctuation scales as

$$F^2_{DMA,\lambda} \sim \lambda^{2H}$$

Note, the moving average $\overline{X}_\lambda(t)$ can take various forms in centring (backward, forward, centred) and weighting (weighted, not weighted, exponential). Using the above equation, we estimate the Hurst exponent by regressing $\log F_{DMA,\lambda}$ against $\log \lambda$ and computing the slope (see Figure 10.8(b)).



(a) Time series of 8192 points with H=0.8 and two moving averages of size 256 (blue) and 1024 (red)

(b) Log-regression of the error variance on the moving-average window sizes

Figure 10.8: (a) Synthetic time series of 8192 points with H=0.8 and two moving averages of size 256 and 1024. At each window size, we detrend the series with regards to its moving average and we compute the variance of the error. (b) We compute the error variance between the real series and its moving average at different scales and perform a log-regression to get the Hurst exponent

10.3.2.4 DMA in high dimensions

Carbone [2007] proposed an algorithm to estimate the Hurst exponent of high-dimensional fractals, based on a generalised high-dimensional variance around a moving average low-pass filter. The method can capture spatial correlations operating in $d \geq 2$, such as a basket of stocks. For simplicity of exposition we will only describe the case $d = 2$, where the generalised variance is defined as

$$\sigma^2_{DMA} = \frac{1}{(N_1 - n_{1max})(N_2 - n_{2max})} \sum_{i_1=n_1-m_1}^{N_1-m_1} \sum_{i_2=n_2-m_2}^{N_2-m_2} \left[f(i_1, i_2) - \overline{f}_{n_1,n_2}(i_1, i_2)\right]^2$$

where the average $\overline{f}_{n_1,n_2}(i_1, i_2)$ is given by

$$\overline{f}_{n_1,n_2}(i_1, i_2) = \frac{1}{n_1 n_2} \sum_{k_1=-m_1}^{n_1-1-m_1} \sum_{k_2=-m_2}^{n_2-1-m_2} f(i_1 - k_1, i_2 - k_2)$$

where N_l for $l = 1, 2$ is the length of the sequence, n_l for $l = 1, 2$ is the sliding window, and $n_{lmax} = \max\{n_l\} << N_l$. The quantity $m_l = int(n_l\theta_l)$ is the integer part of $n_l\theta_l$ and θ_l is a parameter ranging from 0 to 1. The moving average $\overline{f}_{n_1,n_2}(i_1, i_2)$ is calculated over sub-arrays with different size $n_1 \times n_2$ for different values of the window n_l for $l = 1, 2$, ranging from 2 to the maximum value n_{lmax}. A log-log plot of σ^2_{DMA} gives the relation

$$\sigma^2_{DMA} \sim [\sqrt{n_1^2 + n_2^2}]^{2H} \sim s^H$$

for $s = n_1^2 + n_2^2$, which yields a straight line with slope H.

10.3.2.5 The periodogram and the Whittle estimator

The periodogram, described by Geweke et al. [1983], is defined as

$$I(\lambda) = \frac{1}{2\pi N} |\sum_{j=1}^{N} X_j e^{ij\lambda}|^2$$

where λ is a frequency, N is the number of terms in the series, and X_j is the data. Since for a series with finite variance, the periodogram $I(\lambda)$ is an estimator of the spectral density, a series with long-range dependence should have a periodogram proportional to $|\lambda|^{1-2H}$ close to the origin. Therefore, a regression of the logarithm of the periodogram on the logarithm of the frequency λ should give a coefficient of $1 - 2H$. The Whittle estimator is a maximum likelihood estimator (MLE) which assumes a functional form for $I(\lambda)$ and seeks to minimise parameters based upon this assumption. Using that periodogram, the Whittle estimator (see Fox et al. [1986]) uses the function

$$Q(\eta) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda; \eta)} d\lambda$$

where $f(\lambda; \eta)$ is the spectral density at frequency λ, and where η denotes the vector of unknown parameters. The Whittle estimator is the value of η minimising the function Q. When dealing with fractional Gaussian noise (fGn) or ARFIMA models, η is simply the parameter H or d. If the series is assumed to be $FARIMA(p, d, q)$, then η also includes the unknown coefficients in the autoregressive and moving average parts. Note, this estimator provides confidence intervals, and it is obtained through a non-graphical method. However, it assumes that the spectral density is known. If the user misspecifies the underlying model, then errors may occur. The Local Whittle is a semi-parametric version of the Whittle estimator which only assumes a functional form for the spectral densities at frequencies near zero (see Robinson [1995]).

10.4 Testing for market efficiency

10.4.1 Presenting the main controversy

A time series is described as possessing long-range dependence (LRD) if it has correlations persisting over all time scales. Long-range dependent processes provide an elegant explanation and interpretation of an empirical law commonly referred to as Hurst's law or the Hurst effect. However, the existence of long-term correlations in the fluctuations of a time series implies the possibility of violation of the weak form of market efficiency. Long-range power-law correlations have been discovered in economic systems, and particularly in financial fluctuations (see Mandelbrot [1963]). Historical records of economic and financial data typically exhibit distinct cyclical patterns that are indicative of the presence of significant power at low frequencies (long-range dependence). As such, it has been recognised that systems like exchange markets display scaling properties similar to those of systems in statistical physics. Hence, we can therefore apply multifractal analysis to investigate the dynamics of weak form efficiency of financial market by means of Hurst exponent and fractality degree. We also saw in Section (10.1.2.4) that the characteristic exponent, α, of a Stable distribution was related to the Hurst exponent via Equation (10.1.7), implying that the fractal measure of a process was related to the statistical self-similarity of that process. However, we saw in Section (10.3) that there are a number of different statistics which could be used to estimate the parameter α, or H, which are more or less reliable.

Several authors tried to empirically assess the existence of long-term correlations in financial data using either fractal measures or the statistical self-similarity properties of the process. Hence, various statistical methods have been proposed to measure temporal correlations in financial data and to analyse them. However, due to the lack of a good statistics on the financial data, the problem of fully characterising the mathematical structure of the distributions of asset returns/variations (such as index, foreign exchange, etc.) is still an open problem. Consequently, the statistical investigations performed to test the presence or absence of long-range dependence in economic data have been the subject of an intense debate. As discussed by Willinger [1999] these investigations have often become a source of major controversies, mainly because of the important implications it has on many of the paradigms used in modern financial economics. It is inconsistent with the efficient market hypothesis. We are now going to present some of these authors' findings and discuss the pros and cons of both methods.

10.4.2 Using the Hurst exponent to define the null hypothesis

10.4.2.1 Defining long-range dependence

We let $\{X(t)\}_{t\in\mathbb{N}}$ be a weakly stationary time series (it has a finite mean and the covariance depends only on the lag between two points in the series), and we let $\rho(k)$ be the autocorrelation function (ACF) of $X(t)$. A common definition of LRD in the time domain is as follow

Definition 10.4.1 *The time series $X(t)$ is said to be long-range dependent if $\sum_{k=-\infty}^{\infty}\rho(k)$ diverges.*

In general, in the limit $|k|\to\infty$, the functional form

$$\rho(k)\sim C_\rho|k|^{-\alpha}$$

is assumed, with $C_\rho>0$ and $\alpha\in[0,1]$. It states that the ACF decays so slowly that in the limit $|k|\to\infty$, its sum diverges

$$\int_A^\infty \rho(k)dk=\infty$$

for any $0<A<\infty$. The parameter α is related to the Hurst exponent via the equation $\alpha=2-2H$. This definition can be shown to hold in the frequency domain

Definition 10.4.2 *The weakly stationary time series $X(t)$ is said to be long-range dependent if its spectral density obeys*

$$f(\lambda) \sim C_f|\lambda|^{-\beta}$$

as $\lambda \to 0$, for some $C_f > 0$ and some real $\beta \in [0, 1]$.

The parameter β is related to the parameter α via $\alpha = 1 - \beta$, so that the Hurst exponent is given by $H = \frac{1+\beta}{2}$. Hence, LRD can be thought of in two ways,

1. in the time domain it manifests as a high degree of correlation between distantly separated data points, and
2. in the frequency domain it manifests as a significant level of power at frequencies near zero.

A direct and major practical consequence of LRD is that the estimation becomes very difficult. For instance, the variance of the sample mean μ_N decays as $Var(\mu_N) \sim CN^{-\alpha}$ as $N \to \infty$, which is much slower than the common N^{-1}. In general, for higher order sample moment estimation, such as variance, the estimations are strongly biased and have very slow decreasing variance. Note, LRD relates to a number of other areas of statistics, such as statistical self-similarity. We saw in Section (10.1.2) that for any (discrete) time-dependent self-affine function $X(t)$, we can choose a particular point on the signal and rescale its neighbourhood by a factor, say c, using the Hurst exponent (see Definition (10.1.3)). A major consequence of this definition is that the moments of X behave as power laws of time (see Equation (10.1.5)). An exponent $H < \frac{1}{2}$ involves an antipersistent behaviour, while $H > \frac{1}{2}$ means a persistent signal. Hence, an exponent $H \neq \frac{1}{2}$ in a certain range of t values implies the existence of long-range correlations in that time interval. As a result, the signal X can well be approximated by the fractional Brownian motion law. Mathematically, the correlation (see Equation (10.1.4)) of a future increment $\delta_\theta X(t) = X(t+\theta) - X(t)$ with past increment $\delta_\theta X(0)$ is given by

$$C_\theta(t) = \frac{< \delta_\theta X(t)\delta_\theta X(0) >}{< (\delta_\theta X(0))^2 >} = 2^{2H-1} - 1$$

where the correlations are normalised by the variance of $X(t)$. Then, temporal correlations exist for $H \neq \frac{1}{2}$. The correlation assumes that the distribution of daily fluctuations $(X(t+1) - X(t))$ is symmetric with respect to its zero mean.

10.4.2.2 Defining the null hypothesis

Combining estimated Hurst exponent together with confidence intervals, we can test the null hypothesis that the markets behave independently. The alternative hypothesis being that markets are dependent. This can be done by using different methods of estimation of the Hurst exponent. For $H = \frac{1}{2}$ we get a random walk, and we have weakly efficient market in the sense of Fama (F65). Moreover, from Equation (10.1.7) we know that such process has defined and finite second moment and thus finite variance implying martingale process as well, which in turn indicates efficient market in the sense of Samuelson (S65). Persistent process characterised by $H > \frac{1}{2}$ implies rejection of independence which in turn rejects random walk and consequently efficient market of F65. However, the value of $\frac{1}{2} < H < 1$ implies $1 < \alpha < 2$ which in turn indicates undefined or infinite variance. This leads to infinite or undefined square root of variance, non-existence of martingale process, and in turn rejection of market efficiency in the sense of S65 (see Los [2008]). On the other hand, anti-persistent processes $0 < H < \frac{1}{2}$ do not lead to such strong implications because it implies $2 < \alpha < \infty$ so that the underlying distribution is not stable. Even though the non-stable distributions are not yet well studied, the crucial implication is that the process based on non-stable distribution is not independent with identically distributed innovations (see Der et al. [2006]) so that F65 efficiency is rejected. Nonetheless, non-stable distributions have finite variance so that S65 efficiency can not be rejected (see Da Silva et al. [2005]). Knowing the estimated expected values and standard deviations for each method estimating the Hurst exponent we can test market efficiency of financial time series.

10.4.3 Measuring temporal correlation in financial data

10.4.3.1 Statistical studies

A large number of studies based on market prices (logarithmic prices) is concerned with the existence or not of long-, medium- or/and short-range power-law correlations in various economic systems. Neglecting any bias or trend in the signal X, various authors considered the excursion of the signal X after any time period θ, either in the raw value $\delta_\theta X(t)$ or in the positive values, given by

$$|\delta_\theta X(t)|$$

which is related to the variance σ of the signal around its average value. For a self-affine signal, we have

$$\sigma \sim \theta^H$$

Some authors also considered the average price variation

$$F_\theta(t) = \sqrt{< [\delta_\theta X(t)]^2 >}$$

and suggested that the slope of the function, in double logarithmic scale, was given by the relation $F_\theta(t) \sim \theta^H$, while others computed the histogram $F(x,\theta)$ of price variations $x = \delta_\theta X(t)$ for several values of θ and checked the scaling hypothesis

$$F(x,\theta) = \theta^H F(\theta^H x, 1) = \theta^H g(\theta^H x)$$

by plotting $\frac{F(x,\theta)}{\theta^H}$ versus $\theta^H x$. Some authors performed statistical analysis on daily, weekly, and monthly market prices, while in other less frequent cases high-frequency data was chosen. Other authors considered the same statistics, but rather than taking the price difference as a signal, they analysed the logarithmic price change $\delta_\theta L(t) = L(t+\theta) - L(t)$ where $L(t) = \log X(t)$, obtaining the logarithmic return (see Muller et al. [1990]). This is the most appropriate quantity of investigation if the financial time series are assumed to be associated to multiplicative dynamical processes. However, for high-frequency data, the relative increments of the signal between consecutive times being small, no significative differences are supposed to exist if one performs statistical analysis with linear increments.

While some authors claimed the distributions of stock return and FX price changes to be Paretian stable, or Student distributions, others rejected any single distribution, or even claimed that the process was heteroskedastic (see Diebold et al. [1989]). However, they all agreed on the fact that daily changes were leptokurtic and that there are substantial deviations from Gaussian random walk model. Going further, they now agree that the process is not stable (see Boothe et al. [1987]). Numerous articles were published contradicting the classic financial theory of efficient markets by showing the presence of long-term memory in financial time series. Mostly all of these studies did not find temporal correlations present in the system for price (logarithmic price) changes, but they did for absolute price (logarithmic price) changes, the average of absolute price changes, the square root of the variance, and the interquartile range of the distribution of price (logarithmic price) changes. Taylor [1986] studied the correlations of the transformed returns for 40 series and concluded that the returns process was characterised by substantially more correlation between absolute or squared returns than there was between the returns themselves. Kariya et al. [1990] obtained a similar result when studying Japanese stock prices. Examples of empirical studies identifying anomalous scaling for financial data can be found in Muller et al. [1990] followed by a large quantity of similar results reported in the emerging econophysics literature (see Mantegna et al. [1995], Fisher et al. [1997], Schmitt et al. [1999]). Other authors, such as Ding et al. [1993], Galluccio et al. [1997] to name a few, obtained similar results on the temporal correlations of the underlyings, but presented different conclusions regarding the nature of financial data.

10.4.3.2 An example on foreign exchange rates

Muller et al. [1990] presented an empirical scaling law for mean absolute price changes over a time interval, and found that they were proportional to a power of the interval size. The distributions of price changes are found to be increasingly leptokurtic with decreasing intervals, and hence distinctly unstable. Given a collection of interbank spot prices published by Reuters, they considered two samples labelled 1 and 2 where the former is made of tick-by-tick samples for a period of three years from 1986 to 1989, and the latter is made of daily FX prices recorded at 3 pm NYT for 15 years starting in 1973. Taking the logarithmic middle prices $X_j = \frac{1}{2}(\log P_{ask,j} - \log P_{bid,j})$, such that the price changes ΔX over a time interval Δt correspond to logarithmic returns (see details in Section (3.3.1)), they computed the autocorrelation of hourly changes ΔX of the logarithmic price, their absolute values, and their squares over the whole sample 1. They found that only the last two variables had a significant, strong autocorrelation for small time lags (a few hours) indicating the existence of volatility clusters or patterns. As a result, they studied the average absolute price changes

$$|\overline{\Delta X_i}| = \frac{1}{n_k}\sum_{k=1}^{n_k} |X(t_{i,k} + \Delta t) - X(t_{i,k})|$$

where k is the index of the day or the week, n_k is the total number of days (weeks) in sample 1, i is the index of the interval within the day or the week, and Δt is the interval size (one hour or one day). They found the following empirical law

$$|\overline{\Delta X}| = c(\Delta t)^{\frac{1}{E}}$$

where the bar indicates the average over the whole sample period, and $\frac{1}{E}$ is the drift exponent. This is equivalent to the scaling Equation (11.1.13) with $q = 1$ and $H = \frac{1}{E}$. Note, Mandelbrot [1963] had already taken the mean absolute price change as main scaling parameter which is sometime called volatility (see Glassman [1987]). Since the distributions of price vary over time, the existence of standard deviations is not proved [4]. Nonetheless, Muller et al. also found scaling laws for the square root of the variance $\sqrt{|\Delta X|^2}$ and for the interquartile range of the distribution of ΔX. Both the intervals Δt and the volatilities $|\overline{\Delta X}|$ are plotted on a logarithmic scale producing a straight line which is fitted with a linear regression (see Equation (11.1.14)). Since the $|\overline{\Delta X}|$ values for different intervals Δt are not totally independent, as the larger intervals are aggregates of smaller intervals, the linear regression is an approximation. The scaling law was well obeyed for a large range of time intervals since the correlation coefficients between the logarithms of ΔT and $|\overline{\Delta X}|$ exceeded 0.999 and the standard errors of $\frac{1}{E}$ was less than 1%. The scaling law exponents $\frac{1}{E}$ was cluster around 0.59 for all rates, close together in sample 1 and farther apart in sample 2. No significant asymmetry of positive and negative changes were found.

10.4.4 Applying R/S analysis to financial data

10.4.4.1 A first analysis on the capital markets

In order to apply R/S analysis to the capital markets, Peters [1991-96] used logarithmic returns and considered increments of time $N = 6, 7, 8, ..., 240$ months on monthly time series covering 40 years of data. The stability of the estimate is expected to decrease as N increases, as the number of observations decreases. One must be careful when estimating the exponent H by running a regression of $\log(N)$ versus $\log(\frac{R}{S})$ for the full range of N and computing the slope. Doing so is not correct if the series has a finite memory and begins to follow a random walk. Peters [1991-96] applied the R/S analysis to the $S\&P500$, for monthly data over a 38 year period from January 1959 to July 1988 and found that the long memory process was at work for N for less than approximately 48 months with $H \approx 0.78$. After that point, the graph follow the random walk line $H = \frac{1}{2}$ showing that, on average, returns with more than 48 months apart have little measurable correlation left. Scrambling the series of monthly returns he obtained the new coefficient $H = 0.51$ showing that the scrambling destroyed the long memory structure of the original series and turned it into

[4] We have infinite variance in the stable Paretian distributions.

an independent series. Peters repeated the test on US single stocks and obtained Hurst exponents around 0.7. Stocks grouped by industry tend to have similar values of H and similar cycle lengths (maybe linked to economy cycles of the industry). Note, the $S\&P500$ has a higher value of H than any individual stocks, showing that diversification in a portfolio reduces risk. Similar tests were also applied on the bond market, the currency market, all exhibiting significant Hurst exponents showing that these markets are not random walks. However, a large quantity of market data is necessary for a well defined period.

10.4.4.2 A deeper analysis on the capital markets

Using the tools described in Section (10.2.3), such as significance tests on R/S analysis, Peters [1994] repeated the tests on market time series made in his previous book, and described in Section (10.4.4.1), in order to analyse the type of systems exhibited in the markets. The R/S analysis is now performed on the $AR(1)$ residuals defined in Equation (10.2.10), and follows the step-by-step method described in Section (10.2.2.2). Further, the data is made of 102 years of daily records on the Dow Jones covering the period from January 1888 to December 1990 and containing $26,520$ data points. The time series is sampled at different intervals, 5-day returns, 20-day returns, and 60-day returns. A four-year cycle was found independent of the time increment used for the R/S analysis, and a weaker evidence of a 40-day cycle was revealed. The Hurst exponent was most significant for 20-day returns, and much less for daily returns as the noise in higher frequency data makes the time series more jagged and random-looking. This evidence of persistence in the Dow Jones appeared to be very stable. Peters further studied intraday prices for the $S\&P500$ spanning four year of data from 1989 to 1992, and examined frequencies of 3-minute, 5-minute, and 30-minute. Even though high-frequency data experienced high level of serial correlation that could be reduced by using $AR(1)$ residuals, Peters concluded that any analysis on these frequencies were questionable no matter what significance tests were used. His argument being that in dynamical system analysis a large number of observations covering a short time period may not be as useful as a few points covering a longer time period. This is because the existence of nonperiodic cycles can only be inferred if we average enough cycles together. As a result, data sufficiency depends on the length of a cycle. At these frequencies the level of noise was so high that Peters could barely measure determinism, and concluded that the time series was dominated by a short memory process, impying that traders have short memories and merely react to the last trade. However, this autoregressive process is much less significant once we analyse daily data. That is, information has a different impact at different frequencies, and different horizons can have different structures. While we see pure stochastic processes resembling white noise at high frequencies, as we step back and look at lower frequencies, a global structure becomes apparent. Peters [1994] also performed R/S analysis to test both realised and implied volatility from a daily file of $S\&P$ composite prices from 1928 to 1989. Defining the log return at time t as $r_t = \log \frac{P_t}{P_{t-1}}$, he let the volatility be the standard deviation of contiguous 20-day increments of r_t. Assuming non-overlapping and independent increments, the variance V_n over n days is given by

$$V_n = \frac{1}{n-1}\sum_{t=1}^{n}(r_t - \overline{r})^2$$

where $\overline{r}$ is the average value of r. The logarithm changes in volatility, denoted L_n is then given by $L_n = \log \frac{V_n}{V_{n-1}}$. Both the realised and implied volatility were antipersitent with the exponent $H = 0.31$ and $H = 0.44$, respectively. Antipersistent Hurst exponent is related to the spectral density of turbulent flow described by the stable Levy distributions, which have infinite mean and variance. Turbulent systems having no average or dispersion levels that can be measured, volatility will be unstable. It will have no trends, and will frequently reverse itself.

10.4.4.3 Defining confidence intervals for long-memory analysis

While the methods for estimating the Hurst exponent described in Section (10.3) only work for very long (more than 10,000 observations), or infinite time series (see Weron [2002]), the financial time series are much shorter. That is, in nature, there is no limit to time, and thus the exponent H is non-deterministic, as it may only be estimated based on the observed data. For instance, the most dramatic daily move upwards ever seen in a stock market index can

always be exceeded during some subsequent day. We must therefore study the finite sample properties of described methods, such as R/S, $M - R/S$, and DFA for different degree of detrending. Since the condition for a time series to reject long-term dependence given by, $H = \frac{1}{2}$, is an asymptotic limit, various authors proposed correction for finite samples based on estimating the theoretical $(R/S)_n$ for scale n. Traditionally, the statistical approach is to test the null hypothesis of no or weak dependence against the alternative of strong dependence or long memory at some given significance level. However, to construct such a test we must know the asymptotic distribution of the test statistics, but no asymptotic distribution theory has been derived for the R/S analysis and DFA statistic. With or without known asymptotic properties, one way forward is to use Monte Carlo simulations to construct empirical confidence intervals (see Weron [2002]). The procedure consists in the following few steps

1. for a set of sample lengths $L = 2^N$ for $N = 8, .., 16$, generate a large number (10000) of realisations of an independent or a weakly dependent time series (Gaussian white noise).
2. compute the lower $(0.5\%, 2.5\%, 5\%)$ and upper $(95\%, 97.5\%, 99.5\%)$ sample quantiles for all sample lengths.
3. plot the sample quantiles against sample size and fit them with some functions which will be used to construct confidence intervals.

The 5% and 95% quantiles designate the 90% (two-sided) confidence interval, the 2.5% and 97.5% quantiles designate the 95% confidence interval, and so on. For all estimators considered, the DFA statistic for $L > 500$ with $n > 50$ gave estimated values closest to the initial Hurst exponent $(H = \frac{1}{2})$. Note, for $L = 256$ there is only three divisors greater than 50, $(n = 64, 128, 256)$ leading to large errors in the linear regression.

In view of testing the finite sample properties of the methods used to estimate the Hurst exponent and to compare results of different authors, Kristoufek [2009] performed the original test for time series from $T = 2^9$ up to $T = 2^{17}$ with minimum scale $n_{min} = 16$ trading days and maximum scale $n_{max} = \frac{T}{4}$. All steps of R/S analysis were performed on 10000 time series drawn from standardised normal distribution $N(0, 1)$, and $E_T[H]$ and $\sigma_T(H)$ were computed for all T. At $T = 2^9$, $H = 0.5686$ for AL76 and $H = 0.5992$ for P94, while at T^{17}, $H = 0.5254$ for AL76 and $H = 0.5316$ for P96. Even though the estimates converge to $\frac{1}{2}$, they do not get very close to the asymptotic H even for very high T. Computing the mean and standard deviation of the descriptive statistics together with the Jarque-Bera test for normality, the estimates of Hurst exponent were not equal to $\frac{1}{2}$ as predicted by asymptotic theory. Consequently, one must be careful when accepting or rejecting hypotheses about long-term dependence present in time series solely on its divergence from $\frac{1}{2}$, especially for short time series. Further, since the JB test rejected normality of Hurst exponent estimates for time series lengths of 2^9, 2^{16} and 2^{17}, following Weron [2002], Kristoufek [2009] suggested to use percentiles rather than standard deviations for the estimation of confidence intervals. Still he presented the 95% confidence intervals based on standard deviations for R/S and showed they were quite wide for short time series. For example, for $T = 2^9$, $E[H] = 0.57$ with upper $CI = 0.6843$ and lower $CI = 0.4684$ so that for $H = 0.65$ we can not reject the hypothesis of a martingale process. Kristoufek concluded that AL76 outperformed both P94 and P94c, measured with respect to mean squared errors, and suggested to use AL76 for expected value of H for different T.

10.4.5 Some critics at Lo's modified R/S statistic

Even though the R/S analysis is a simple method for detecting long-range dependence from empirical data which is not reliable for small samples, it is still a highly effective and useful graphical method for large samples. One of its most useful feature is its relative robustness under changes in the marginal distribution of the data, especially if the marginals exhibit heavy tails with infinite variance (see Mandelbrot et al. [1969a] [1969b]). However, we saw in Section (10.2.3) that it is sensitive to the presence of explicit short-range dependence structures, and lack a distribution theory for the underlying statistic. To overcome these shortcomings, Lo [1991] proposed a modified R/S statistic obtained by replacing the sample standard deviation with a consistent estimator of the square root of the variance of the partial sum. He derived the limiting distribution of his statistic under both short-range and long-range dependence,

claiming robustness to short-range dependence (see details in Section (10.2.4.2)). While most of the econometric literature acknowledge Lo's results (see Hauser et al. [1994], Huang et al. [1995], Campbell et al. [1997]), Teverovsky et al. [1999] highlighted a number of problems associated with Lo's method and its use in practice. They used fractional Gaussian noise (fGn) and fractional ARIMA (FARIMA) models to synthetically generate purely long-range dependent observations and hybrid short-range/long-range dependent data. The most important finding is that Lo's method has a strong preference for accepting the null hypothesis of no long-range dependence, irrespective of whether long-range dependence is present in the data or not. As a result, Lo's method should not be used in isolation, and other set of graphical and statistical methods should be considered for checking for long-range dependence (see Abry et al. [1998]). Using R/S analysis in the context of asset returns, Mandelbrot [1967] suggested H-values of around 0.55 to be representative for stock returns. Studying 200 daily stock return series of securities listed on the New York Stock Exchange, Greene et al. [1977] found significant evidence of long-range dependence in many of these series. In contrast, using his modified R/S statistics, Lo did not find evidence of long-range dependence in the Research in Security Prices (CRSP) data. While Lo found strong evidence that the series of absolute values of the CRSP daily stock returns exhibited long-range dependence ($H = 0.85$), he focused on the series itself. Choosing truncation lags $\lambda = 90, 180, 270, 360$, he found that daily stock returns did not exhibit long-range correlation because the values of $V_\lambda(N)$, for $N \approx 6400$, were within the 95% confidence interval. Attributing the findings of Greene et al. to the failure of R/S analysis in presence of short-range dependence, he concluded that the dynamics of asset returns should be described by traditional short-range models. Using the CRSP daily stock return data, Willinger et al. [1999] revisited the question of whether or not actual stock market prices exhibit long-range dependence. Performing an in-depth analysis of the same data sets, they showed that Lo's acceptance of the hypothesis for the CRSP data (no long-range dependence in stock market prices) was less conclusive than expected in the econometric literature. Upon further analysis of the data, they found empirical evidence of long-range dependence in stock price returns, but because the corresponding degree of long-range dependence was typically very low (H-values around 0.6), the evidence was not absolutely conclusive. In real life, the correlations at very large lags are so small that they are very sensitive to slight deviations. That is, financial time series do not have infinite memory, but rather behave like systems with bounded natural periods. Hence, they concluded that present statistical analyses could not be expected to provide a definitive answer to the presence or absence of long-range dependence in asset price returns.

10.4.6 The problem of non-stationary and dependent increments

10.4.6.1 Non-stationary increments

From the definitions of long-range dependence (LRD) given in Section (10.4.2.1), we see that the statistical analysis performed in Section (10.4.3) and Section (10.4.4) are tailored for processes with stationary increments. However, since stock returns and FX rates suffer from systematic effects mainly due to the periodicity of human activities, they can not be considered as processes with stationary increments, and the standard scaling analysis should not be appropriate in this case. It is therefore desirable to get rid of these systematic effects. The scaling analysis has always been applied in statistical physics to processes with stationary time increments. For these processes, the presence of scaling is equivalent to the statement that there is no characteristic scale (or time, in our case) in the system. However, it strongly contrasts with the presence of time-scales associated with days, weeks and months present in financial data. Once these systematic effects are filtered out, we may hope to get clearer statistical properties of the signal. To eliminate problems due to periodic seasonality in the time signal, several authors introduced time transformation based on the use of volatility as an indicator of market activities (see Dacorogna et al. [1993], Galluccio et al. [1997]). They found correlations present in the system and highlighted the multiscaling behaviour of FX rates. They also suggested that the non-stationarity of the signal does affect the results of a statistical analysis, leading to unprecise or even wrong conclusions.

10.4.6.2 Finite sample

In theory, the Hurst exponent only works for very long, or infinite, time series, such that the condition for rejecting LRD is an asymptotic limit. However, financial time series are much shorter, so that the Hurst exponent is non-deterministic, and one must therefore study the finite sample properties of the different methods. Further, financial time series have a finite memory and tend to follow random walk once the time period covered has exceeded the average nonperiodic cycle length. It defines the point where the memory of initial conditions is lost, corresponding to the end of the natural period of the system. We must therefore have sufficient data to detect the natural period when estimating the Hurst exponent. However, in the time domain, LRD is measured only at high lags (strictly at infinite lags) of the ACF, where only a few samples are available and where the measurement errors are largest. Similarly, in the frequency domain, LRD is measured at frequencies near zero, where it is hardest to make measurements. Even though the Hurst exponent is perfectly well defined mathematically, Clegg [2006] showed that is was a very difficult property to measure in real life, since

- the data must be measured at high lags/low frequencies where fewer readings are available
- all estimators are vulnerable to trends in the data, periodicity and other sources of corruption

10.4.6.3 Dependent increments

We saw above that it was very difficult to measure temporal correlations for financial time series. However, even if the series is serially uncorrelated, it can still be dependent. If information comes in bunches, the distribution of the next return will depend on previous returns although they may not be correlated. Even if the returns autocorrelation vanishes, we can not conclude that returns are independent variables, since independence implies that all functions of returns should be uncorrelated variables. However, we saw earlier that numerous studies showed that volatility had a long memory. Examining the autocorrelation of r_t and $|r_t|^d$ for positive d, where r_t is the $S\&P500$ stock return, Ding et al. [1993] found that the sample autocorrelations for absolute returns were greater than the one for squared returns at every lag up to at least 100 lags. It clearly showed that the return process was not an i.i.d. process. They also found from the autocorrelogram that $|r_t|^d$ had the largest autocorrelation up to lag 100 when $d \approx 1$, and that it gets smaller almost monotonically when d goes away from 1. We know from statistical physics that if the time increments are distributed according to a multifractal density, or equivalently, if the distribution of price changes presents different scaling for different time intervals, the first moment $< |X(t)| >$ is larger than the volatility $V(t) = \sqrt{[X(t) - < X(t) >]^2}$. This is due to the convex property of the qth order moments $\xi_q =< |\delta_\theta X(t)|^q >$ as functions of q. Later, Galluccio et al. [1997] showed that FX rates do not have independent increments. They did so by analysing the correlation function of the absolute value of price variations

$$A_\theta(t) =< |\delta_\theta X(t)||\delta_\theta X(0)| > - < |\delta_\theta X(0)| >^2$$

and the sign correlation function

$$S_\theta(t) =< \text{sgn}[\delta_\theta X(t)]\text{sgn}[\delta_\theta X(0)] > - < \text{sgn}[\delta_\theta X(0)] >^2$$

and showed that these correlations were present in the economical system.

10.4.6.4 Applying stress testing

Clegg [2006] tested the R/S parameter, aggregated variance, periodigram, the local Whittle and the wavelet techniques. Real life data is likely to have periodicity, trends, as well as quantisation effects if readings are taken to a given precision. Trial data sets with LRD and a kown Hurst exponent were generated with fractional autoregressive integrated moving average (FARIMA) model and fractional Gaussian noise (FGN). The data was first tested with the various methods listed above and then the same data was corrupted in several ways: addition of zero mean $AR(1)$ model with a high degree of short-range correlation which might be mistaken for LRD, addition of periodic function

(10 complete cycles of a sine wave are added to the signal), addition of linear trend simulating growth in the data. Note, technically the addition of a trend or of a periodic noise makes the time series non-stationary, such that the modified series is no-longer LRD. In addition, real life data were studied to provide an insight on the different measurement methods. For each of the simulation methods chosen, traces were generated with $100,000$ points for each trace, and the Hurst parameter was set to 0.7 and 0.9. Considering fGn models, for $H = 0.7$ all estimators were relatively close when no noise was added, but the addition of $AR(1)$ noise confused all the methods. For $H = 0.9$ the R/S method was under-estimated and all methods performed badly with the addition of $AR(1)$ noise. In all cases, adding sine wave and trend caused trouble to the time domain methods, but the frequency methods were not affected. Testing different FARIMA settings, Clegg obtained similar results. The case $FARIMA(2, d, 1)$ with the AR parameters $\phi_1 = 0.5$, $\phi_2 = 0.2$ and the MA parameter $\theta_1 = 0.1$, indicating strong short-range correlation, was the hardest to estimate. In this simple case, with known theoretical result, all the methods fail to get the correct answer. To get an understanding of this failure, Clegg considered visualising the ACF of the data up to lag 1000 for a data set of $100,000$ points of fGn data with $H = 0.7$. In this case, the log-log plot of the ACF is a straight line and all estimators performed well on that data. Note, at the higher lags the error on the ACF estimate was large. However, when adding $AR(1)$ noise, the log-log plot of the ACF for low lags remained much higher than in the noise free data, with a concave shape making it difficult to fit a straight line. Further, the ACF was heavily perturbated in the log-log plot for lags over fifty. At last, displaying the ACF for the data generated with the $FARIMA(2, d, 1)$, even before adding noise, the shape was not a straight line for low lags and was perturbated for large lags. Clegg concluded that it is impossible to get a good estimate of LRD simply by fitting a straight line to the ACF, and that the addition of highly correlated short range dependent data was vastly changing the nature of the estimation problem. Moreover, when analysing real data with no genuine answer, he could not tell which method was more right than another.

10.4.7 Some results on measuring the Hurst exponent

Since so many papers have been written comparing the different methods to measure the Hurst exponent both in theory and in practice, we are going to repeat some of the tests on R/S analysis, wavelet-based method, DMA and DFA. We will test the methods for a range of H value, different sample sizes, and we will also study the computation time as a function of the sample size for all the methods.

10.4.7.1 Accuracy of the Hurst estimation

We test the different methods for monofractal series of $32,768$ points with different values of Hurst exponent using Wood and Chan algorithm (see Wood et al. [1994]) for the generation of fractional Brownian motion. For each Hurst exponent H, we independently generate 500 different series and we compute the averages and the standard deviations of the results for each method, assuming that the Hurst exponent follows a normal distribution. The results are presented in the table below, and the plots in the Figure (10.9) and Figure (10.10).

From Table (10.2), we notice that all the methods perform well except for the R/S analysis which is significantly less precise than the other methods. Further, the performance of R/S analysis is highly dependent on the level of the Hurst exponent. For instance, the error is incredibly high for $H = 0.1$ but relatively accurate for a Hurst exponent around 0.7.

Figure (10.9) and Figure (10.10) are grouped bar plots representing the absolute error and the standard deviation, respectively, between the Hurst estimates and the theoretical Hurst exponent. As we said, R/S analysis performs badly whereas all the others are fairly accurate. However, we note that the wavelet-based method is much more accurate than the detrending methods (around $3 \cdot 10^{-4}$ versus $6 \cdot 10^{-3}$), but its standard deviation is higher (around $5 \cdot 10^{-2}$ versus $1 \cdot 10^{-2}$).

Among the detrending methods (DMA, DFA1 and DFA2), DMA method is a little bit less accurate than DFA methods and its standard deviation is also slightly higher. We also remark that the second-order polynomial interpola-

Hurst	R/S Analysis	Wavelet-based	DMA	DFA order 1	DFA order 2
0.10	0.1979 ± 0.0077	0.1004 ± 0.0611	0.1099 ± 0.0073	0.1093 ± 0.0052	0.1166 ± 0.0043
0.20	0.2781 ± 0.0103	0.1997 ± 0.0387	0.2106 ± 0.0118	0.2060 ± 0.0087	0.2119 ± 0.0071
0.30	0.3584 ± 0.0127	0.2999 ± 0.0444	0.3107 ± 0.0152	0.3036 ± 0.0115	0.3085 ± 0.0096
0.40	0.4426 ± 0.0158	0.4007 ± 0.0403	0.4109 ± 0.0180	0.4010 ± 0.0140	0.4057 ± 0.0114
0.50	0.5276 ± 0.0182	0.4993 ± 0.0264	0.5120 ± 0.0205	0.4998 ± 0.0157	0.5045 ± 0.0130
0.60	0.6169 ± 0.0203	0.6000 ± 0.0300	0.6127 ± 0.0232	0.5993 ± 0.0170	0.6031 ± 0.0144
0.70	0.7027 ± 0.0231	0.7002 ± 0.0307	0.7120 ± 0.0292	0.6986 ± 0.0199	0.7024 ± 0.0154
0.80	0.7828 ± 0.0231	0.7999 ± 0.0982	0.8070 ± 0.0330	0.7960 ± 0.0181	0.8007 ± 0.0152
0.90	0.8568 ± 0.0236	0.9004 ± 0.0498	0.9004 ± 0.0390	0.8960 ± 0.0222	0.9007 ± 0.0178

Table 10.2: Accuracy of Hurst estimation on fractional Brownian motions with different methods.

tion in the DFA is less accurate that the linear interpolation version of DFA but its standard deviation is slightly lower. However, since the gain in the standard deviation is very small, one would prefer a DFA method of order 1.

Hence, according to these results, one would avoid R/S analysis and DMA methods, and the choice between the wavelet-based method, DFA1 and DFA2 being made according to the user constraints. For instance, if the user needs a Hurst exponent as accurate as possible, he would choose the wavelet-based method. However, if he wants the most stable estimator, he would choose the higher order DFA method. If he wants something both fairly accurate and stable, the DFA method with linear interpolation would be a good compromise. In short, for a large sample size, we have:

$$\text{R/S} \ << \ \text{DMA} \ << \ \text{DFA1} \ << \ \text{DFA2 (stability)} \ \sim \ \text{Wavelets (accuracy)}$$

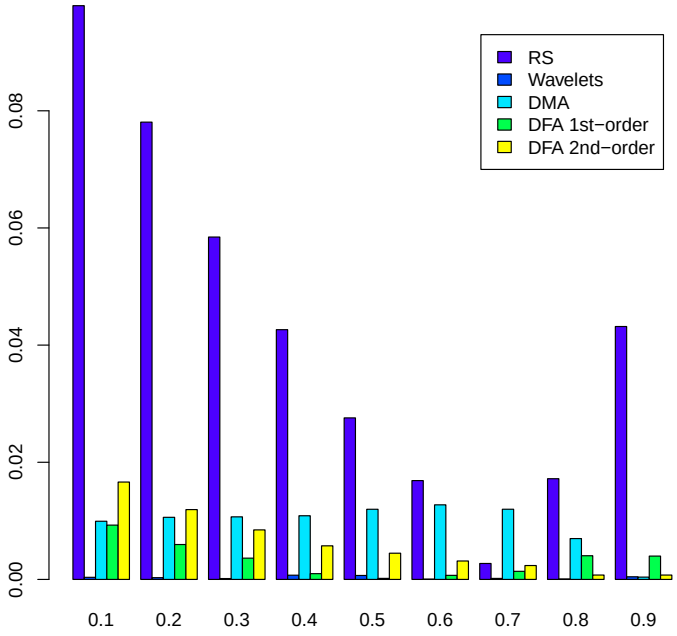


Figure 10.9: Absolute error of Hurst estimates for fractional Brownian motions of 32768 points with different methods for different values of H.

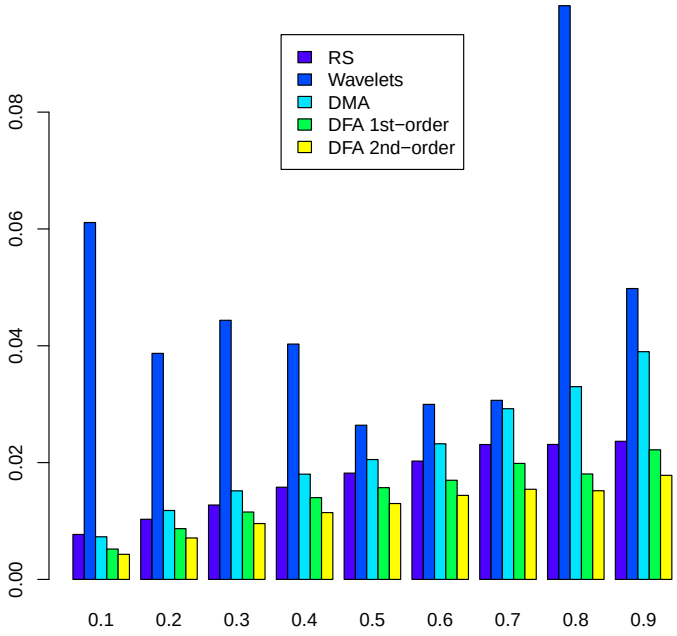


Figure 10.10: Standard deviation of Hurst estimates for fractional Brownian motions of 32768 points with different methods for different values of H.

10.4.7.2 Robustness for various sample size

We are now going to test the different methods against the fractional Brownian motion generated by the Wood and Chan algorithm, for different sample sizes (128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768). For each pair (H, Sample size), we compute the absolute error and the standard deviation of the Hurst estimates. We provide below the detailed table of absolute errors for each pair (H, Sample size) and each method, and we also plot the average absolute errors and standard deviations focusing only on the effect of sample sizes (and not on the Hurst exponent values).

	Size	128	256	512	1024	2048	4096	8192	16384	32768
	R/S	0.1792	0.1629	0.1490	0.1377	0.1277	0.1188	0.1110	0.1039	0.0979
	Wavelets	0.0904	0.1233	0.0850	0.0503	0.0071	0.0001	0.0037	0.0023	0.0003
0.1	DMA	NA	0.0526	0.0387	0.0295	0.0232	0.0185	0.0150	0.0121	0.0099
	DFA1	0.0604	0.0449	0.0326	0.0252	0.0329	0.0255	0.0204	0.0166	0.0092
	DFA2	0.1094	0.0788	0.0582	0.0447	0.0583	0.044	0.0356	0.0290	0.0166
	R/S	0.1544	0.1390	0.125	0.114	0.1055	0.0974	0.0900	0.0836	0.0780
	Wavelets	0.0493	0.050	0.0058	0.011	0.001	0.0007	0.0006	0.0003	0.0003
0.2	DMA	NA	0.0566	0.0403	0.0305	0.0238	0.0188	0.0151	0.0123	0.010
	DFA1	0.0442	0.0320	0.0219	0.0165	0.0232	0.0177	0.0138	0.0110	0.0059
	DFA2	0.0888	0.0625	0.0445	0.0335	0.0451	0.0342	0.0267	0.0213	0.0119
	R/S	0.1275	0.1151	0.1021	0.0924	0.0837	0.0760	0.0694	0.0633	0.0584
	Wavelets	0.0346	0.0265	0.0011	0.0017	0.0040	0.001	0.0005	0.0002	0.0001
0.3	DMA	NA	0.0635	0.0429	0.0325	0.0251	0.0199	0.0159	0.0127	0.0106
	DFA1	0.0308	0.021	0.0141	0.0104	0.016	0.0126	0.0094	0.0074	0.0036
	DFA2	0.0706	0.0476	0.0341	0.0254	0.0358	0.0267	0.0206	0.0163	0.0084
	R/S	0.0994	0.0871	0.0781	0.0703	0.0629	0.0566	0.0510	0.0459	0.0426
	Wavelets	0.0098	0.0143	0.0005	0.0009	0.0000	0.0005	0.0007	0.0062	0.0007
0.4	DMA	NA	0.0624	0.043	0.0333	0.0257	0.0205	0.016	0.0127	0.0108
	DFA1	0.0195	0.0123	0.0077	0.0053	0.0117	0.0087	0.0066	0.0054	0.000
	DFA2	0.0562	0.0370	0.025	0.0186	0.0285	0.0208	0.0160	0.0129	0.0057
	R/S	0.0710	0.0613	0.0551	0.0490	0.0437	0.039	0.035	0.0317	0.0275
	Wavelets	0.0030	0.0022	0.000	0.0020	0.001	0.0038	0.0064	0.0021	0.0006
0.5	DMA	NA	0.0614	0.0461	0.0348	0.0270	0.0211	0.0172	0.0142	0.0119
	DFA1	0.0120	0.0068	0.0035	0.0016	0.0085	0.0058	0.0040	0.0029	0.0001
	DFA2	0.0452	0.0301	0.0194	0.0138	0.0226	0.0168	0.0124	0.0095	0.0044
	R/S	0.0394	0.0324	0.031	0.0279	0.0248	0.0223	0.0206	0.0183	0.0168
	Wavelets	0.0065	0.0036	0.0008	0.0030	0.0012	0.0000	0.0032	0.0020	0.0000
0.6	DMA	NA	0.0620	0.0477	0.0365	0.0283	0.0225	0.018	0.015	0.0127
	DFA1	0.0027	0.0007	0.0005	0.0009	0.0060	0.00	0.0034	0.0026	0.000
	DFA2	0.0349	0.0227	0.0146	0.0102	0.0191	0.0138	0.0109	0.0088	0.0031
	R/S	0.0051	0.0062	0.0042	0.0039	0.0038	0.0036	0.0032	0.0030	0.0027
	Wavelets	0.0158	0.0020	0.003	0.0082	0.0021	0.0036	0.0079	0.0106	0.0001
0.7	DMA	NA	0.0666	0.0473	0.0353	0.027	0.0218	0.0179	0.015	0.0119
	DFA1	0.0015	0.002	0.0034	0.003	0.0050	0.0036	0.0028	0.0026	0.0013
	DFA2	0.0282	0.0176	0.0112	0.0074	0.0167	0.0121	0.0094	0.0075	0.0023
	R/S	0.0346	0.0312	0.0279	0.0250	0.0225	0.0208	0.0195	0.01	0.0171
	Wavelets	0.0248	0.0033	0.0053	0.0019	0.0111	0.0005	0.0005	0.01	0.0000
0.8	DMA	NA	0.059	0.0415	0.0298	0.022	0.0169	0.0130	0.0102	0.0069
	DFA1	0.0064	0.0073	0.0066	0.005	0.0034	0.0020	0.0015	0.0012	0.0040
	DFA2	0.0232	0.0130	0.0077	0.0051	0.0151	0.0103	0.0077	0.0061	0.0007
	R/S	0.0814	0.0751	0.0692	0.064	0.0590	0.0541	0.0505	0.0475	0.0431
	Wavelets	0.0421	0.0152	0.0000	0.0003	0.0003	0.0006	0.0013	0.0004	0.0004
0.9	DMA	NA	0.0521	0.0331	0.0203	0.0125	0.007	0.0040	0.0016	0.0003
	DFA1	0.0094	0.0078	0.0082	0.0075	0.0026	0.0015	0.0008	0.0003	0.003
	DFA2	0.0188	0.0121	0.006	0.0036	0.014	0.0099	0.0073	0.0054	0.0007

Table 10.3: Absolute error of the Hurst estimates for various Hurst exponent, various sample sizes and different methods

From Table (10.3), we logically observe that for any method and any Hurst exponent value, the quality of the Hurst estimates worsen as the sample size get smaller. For example, when $H = 0.1$, the R/S statistic goes from an absolute error of 0.0979 for $32,768$ points to 0.1792 for 128 points; the wavelet method goes from 0.0003 to 0.0904; the DMA goes from 0.0099 to 0.0526 for 256 points; the DFA1 goes from 0.0092 to 0.0604 and the DFA2 goes from 0.0166 to 0.1094.

We also notice that except for the DMA method, the Hurst estimators work better for anti-persistent processes ($H > 0.5$) than for persistent processes ($H < 0.5$). For example, considering the sample size 512, the absolute error is increasing when the Hurst exponent decreases. However, the effect is still less significant compared to the sample size effect on the Hurst estimates.

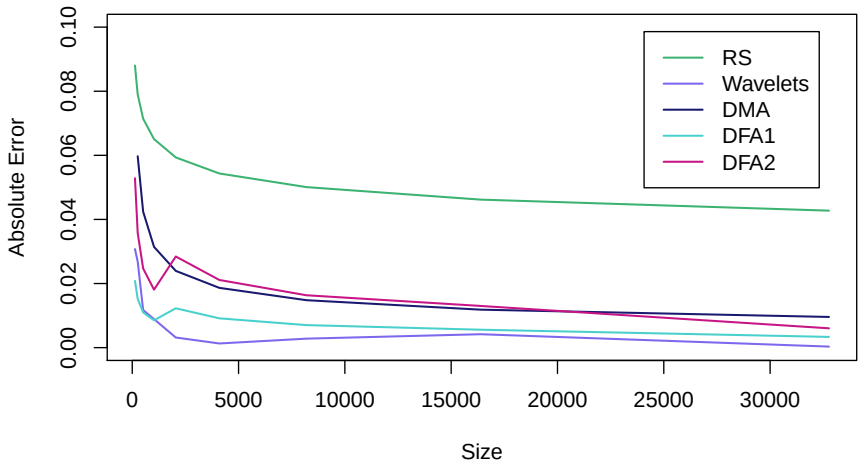


Figure 10.11: Absolute error of the Hurst exponent in function of the sample size.

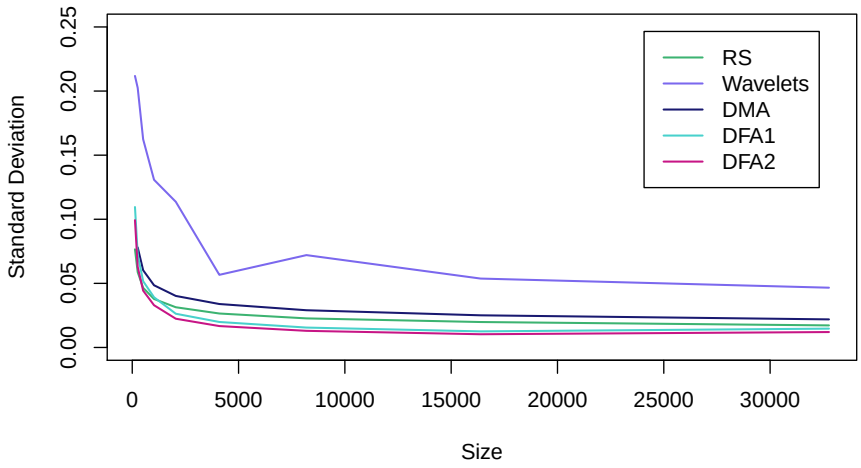


Figure 10.12: Standard deviation of the Hurst exponent in function of the sample size.

Figure (10.11) and Figure (10.12) represent respectively the absolute error and the standard deviation of the Hurst exponent averaged on the value of the Hurst exponent so that we get two plots with those two indicators exclusively in function of the sample sizes. The graphs emphasise our first observation. Both the absolute error of the Hurst exponent and the standard deviation are exponentially decreasing functions of the sample sizes. While the wavelets and DFA1 methods seem to be the two best methods for small sample sizes in terms of absolute error, the former has a standard deviation which is much higher than all the other methods in general, whereas the latter is much more stable.

As a consequence, our previous conclusion still holds true for any sample sizes. The best choice among the methods depends on one's wishes. The R/S analysis and DMA methods seem to be the poorest method compared to the two other methods. If the user does not work on large sample sizes, he has then two choices: either he need accuracy in which case he would choose the wavelet-based method, or, he needs relatively accurate but highly stable estimates and he would consequently choose the DFA method.
The hierarchy is then:

$$\text{R/S} << \text{DMA} << \text{Wavelets (accuracy)} \sim \text{DFA1 (stability)} \sim \text{DFA2 (stability)}$$

10.4.7.3 Computation time

We now wish to apply the long-range dependence analysis (LRD) to the financial industry, which is well-known to be a highly fast-paced and competitive environment where every milli/micro-seconds are critical. It is therefore important to give indications on the computation time of those algorithms. We estimate and plot the computation time of the four methods as a function of the sample size. The following results are computed with a CPU *Intel Core 2 Quad Processor Q9550* (2.83 GHz, 1333 MHz).

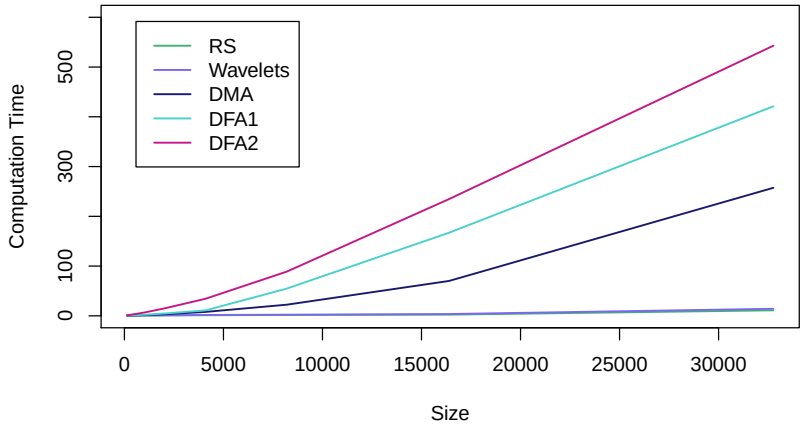


Figure 10.13: Computation time (in milliseconds) of the Hurst estimation methods in function of the sample size.

From Figure (10.13), which represents the computation time in function of the sample size, we immediately distinguish two classes of methods. The wavelets and R/S analysis are linear in function of the sample size, their time complexity are in fact $\mathcal{O}(n)$ which is extremely fast (from 0.62 to 14.18 ms). The second class contains the DMA, DFA1, and DFA2. Their complexity are actually $\mathcal{O}(n^2)$ but the DMA is actually faster than the DFA since computing an average is faster than a linear/polynomial interpolation. The DFA1 is faster than DFA2 since computing a linear interpolation is faster than a 2nd order interpolation. Therefore, in terms of computational time, we have:

$$\text{DFA2} << \text{DFA1} << \text{DMA} << \text{Wavelets} \sim \text{R/S}$$

Chapter 11

The multifractal markets

11.1 Multifractality as a new stylised fact

11.1.1 The multifractal scaling behaviour of time series

11.1.1.1 Analysing complex signals

The use of fractional Brownian motion (fBm) described in Section (10.1.2), which involves a restriction to a certain Holder continuity H of the paths for all times of the process, is too restrictive for many applications, and variable time-dependent Holder continuities of the paths are needed. That is, in the case where the Hurst exponent, H, changes over time, the fractional Brownian motion becomes a somewhat restrictive model, unable to capture more fully the complex dynamics of the series. As explained by Kantelhardt et al. [2002], many series do not exhibit a simple monofractal scaling behaviour described with a single scaling exponent. There may exist crossover (time-) scales $n_\times$ separating regimes with different scaling exponents, such as long-range correlations on small scales, $n << n_\times$, and another type of correlations or uncorrelated behaviour on larger scales $n >> n_\times$. Alternatively, the scaling behaviour may be more complicated, so that different scaling exponents are required for different part of the series. For instance, the scaling behaviour in the first half of the series differ from that in the second half. In complex system, such different scaling behaviour can be observed for many interwoven fractal subsets of the time series, in which case a multitude of scaling exponents is required for a full description of the scaling behaviour, and a multifractal analysis must be applied. While the term fractal was mainly associated with deterministic chaos, originating from a small number of generating equations, in the 80s a new class of stochastic fractals developed termed multifractals, with processes originating in high-dimensional systems (see Schertzer et al. [1991]). In general, two different types of multifractality in time series can be distinguished

1. Multifractality due to a broad probability density function (pdf) for the values of the time series. One can not remove the multifractality by shuffling the series.

2. Multifractality due to different long-range (time-) correlations of the small and large fluctuations. In this case, the pdf of the values can be a regular distribution with finite moments. The corresponding shuffled series will exhibit nonmultifractal scaling, since all long-range correlations are destroyed by the shuffling procedure.

If both kinds of multifractality are present, the shuffled series will show weaker multifractality than the original series. The multifractal formalism (MF) provides a scale invariant mechanism for the analysis and generation of complex signals fitting well with the observed experimental properties of fully developed turbulence (FDT) as well as other physical systems ranging from natural images to heartbeat dynamics or econometric signals. The formalism also allows to highlight relevant dynamical features of the systems under study, and theoretical models can be devised to fit the observed multifractal properties.

11.1.1.2 A direct application to financial time series

Since response time distributions of financial time series have been found to be typically unimodal with fat tails and leptokurtosis, distributions like lognormal, Gamma, Weibull, and power law distributions have been suggested to model asset returns. However, all these models assume that the response times are trial-independent random variables. However, we saw in Section (10.4) that financial time series exhibit non-Gaussian distribution and long-range dependent dynamics. The long-range trial dependency of response time is numerically defined by a scaling exponent obtained by monofractal analyses (see details in Section(10.3)). However, it was observed that this scaling exponent decreased with increasing difficulty and an increased level of external economical perturbation. While conventional monofractal analysis numerically define a long-range dependency as a single scaling exponent, they assume that the response times are Gaussian distributed such that their variations are described by the second-order statistical moment (variance) alone. The presence of a non-Gaussian response time distribution indicates

1. that the variations can not be exclusively described by the scaling of variance alone, but that the scaling of higher order statistical moments such as skewness and kurtosis must be considered,
2. it also indicates intermittent changes in the magnitude of response time variation which might be due to feedback effects, or changes in investor's behaviour.

These intermittent changes in the response time variation provide temporal modulation of both the width and shape of the response time distribution, and consequently, temporal modulation of the scaling exponent. Multifractal analyses estimate a multifractal spectrum of scaling exponents containing the single exponent estimated by the conventional monofractal analyses. Various authors introduced numerical tools to define the scaling exponents of higher order statistical moments and the temporal modulation of a local scaling exponent.

11.1.2 Defining multifractality

We saw in Section (10.1.2.4) that multifractality or anomalous scaling allows for a richer variation of the behaviour of a process across different scales. Equation (10.1.6) shows that this variability of scaling laws can be translated into the variability of the Hurst index, which is no-longer constant. Following Abry et al. [2002] and Wendt [2008] who both gave a synthetic overview of the concepts of scale invariance, scaling analysis, and multifractal analysis, we are going to introduce multifractality.

11.1.2.1 Fractal measures and their singularities

In nonlinear physics, one want to characterise complicated fractal objects and describe the events occurring on them. For example, this characterisation applies to dynamical systems theory, diffusion-limited aggregation, percolation, to name a few. In general, one describe such events by dividing the object into pieces labelled by index i running from 1 to N. The size of the ith piece is l_i and the event occurring upon it is given by the number M_i. For instance, in the droplet theory of Ising model, the magnetisation has a value of the order of

$$M_i \sim l^y$$

where y is a critical index. Since these droplets should fill the entire space, the density of such droplets is

$$\rho(l) \sim \frac{1}{l^d} \tag{11.1.1}$$

where d is the Euclidean dimension of space. Halsey et al. [1986] considered the case where each M_i is a probability that some event will occur upon the ith piece. They assumed the d-dimensional time series $\{X_i\}_{i=1}^N$ with trajectories lying on a strange attractor of dimension D, $D < d$ and tried to find out how many times N_i would the time series visit the ith box. They defined $p_i = \lim_{N\to\infty} \frac{N_i}{N}$ and generated the measure on the attractor $d\mu(X)$, since $p_i = \int_{\text{ith box}} d\mu(X)$. That is, p_i is the probability (integrated measure) in the ith box. If a scaling exponent α [1] is defined

[1] Lipschitz-Holder exponent

by

$$p_i^q \sim l_i^{\alpha q}$$

then α can take on a range of values, corresponding to different regions of the measure. They suggested that the number of times α would take on value between $\alpha^{'}$ and $\alpha^{'} + d\alpha^{'}$ should be of the form

$$d\alpha^{'} \rho(\alpha^{'}) l^{-f(\alpha^{'})}$$

where $f(\alpha^{'})$ is a continuous function. It reflects the differing dimensions of the sets upon which the singularities of strength $\alpha^{'}$ may lie. It is roughly equivalent to Equation (11.1.1) with the dimension d replaced by the fractal dimension $f(\alpha)$ which varies with α. Thus, fractal measures are modelled by interwoven sets of singularities of strength α, each characterised by its own dimension $f(\alpha)$.

The function $f(\alpha)$ must then be related to observable properties of the measure. The most basic property of a strange attractor is its dimension, and the three main ones are the fractional (or similarity) dimension D of the support of the measure (see also box counting dimension), the information dimension σ, and the correlation dimension ν. Hentschel et al. [1983] showed these attractors are characterised by an infinite number of generalised dimension D_q, $q > 0$. Rather than considering droplet theory of Ising model, they let $\{X_i\}_{i=1}^N$ be the points on the attractor covering space with a mesh of d-dimensional cubes of size b^d, and $M(b)$ be the number of cubes containing points of the series. They let $p_\alpha = \frac{N_\alpha}{N}$ be the probability to fall in one of the boxes of type α and defined the rescaling hierarchy based on a box of size l^d on the nth level covered with bins of size b^d where the probability of the ith bin is denoted by $p_i(l, b)$. They showed self-similarity between the box of size l^d on the nth level of the hierarchy and the box of size $((\frac{l}{s})_\alpha)^d$ on the $(n+1)$th level which is used to calculate the dimensions associated with the natural measure. So far, the exponent ν was related to correlations between pairs of point on the fractal, but Hentschel et al. [1983] considered higher order correlation functions or correlation integrals $C_n(l)$ and showed that

$$C_n(l) \approx C_n(l, b) = \sum_{i \in l} p_i^n(l, b)$$

scales like

$$C_n(l, b) = C_n(l) b^{\nu_n}$$

Obtaining D_n for an integer n, they generalised it to an uncountable infinity of quantities D_q with any $q > 0$ defined by

$$D_q = \lim_{l \to 0} \Big(\frac{1}{q-1} \frac{\ln \chi(q)}{\ln l} \Big) \tag{11.1.2}$$

where

$$\chi(q) = \sum_{i \in l} p_i^q$$

is a partition function. Hence, the generalised dimensions D_q correspond to the scaling exponents for the qth moments of the measure. Note, D_0 is the fractional (similarity) dimension, D_1 is the information dimension, and D_2 is the correlation dimension. That is, multifractals have multiple dimension in the D_q versus q spectra, but monofractals stay rather flat in that area. Note, the term

$$S_q = \frac{1}{q-1} \ln \chi(q)$$

is called the Renyi entropy representing a one-parametric generalisation of Shannon's entropy (for which it reduces for $q \to 1$).

Later, Halsey et al. [1986] related the function $f(\alpha)$ to the set of dimensions D_q, and after replacing p_i with its definition above, they obtained

$$\chi(q) = \int d\alpha^{'} \rho(\alpha^{'}) l^{-f(\alpha^{'})} l^{q\alpha^{'}} = \int d\alpha^{'} \rho(\alpha^{'}) l^{q\alpha^{'} - f(\alpha^{'})}$$

Since l is very small, this integral is dominated by the value of $\alpha^{'}$ making $q\alpha^{'} - f(\alpha^{'})$ smallest, provided that $\rho(\alpha^{'})$ is nonzero. To find the minimum, they replaced $\alpha^{'}$ by $\alpha(q)$, which is defined by the external condition

$$\frac{d}{d\alpha^{'}} \left[q\alpha^{'} - f(\alpha^{'}) \right] \Big|_{\alpha^{'} = \alpha(q)} = 0$$

Differentiating one more time, we have

$$\frac{d^2}{d(\alpha^{'})^2} \left[q\alpha^{'} - f(\alpha^{'}) \right] \Big|_{\alpha^{'} = \alpha(q)} > 0$$

such that $f^{'}(\alpha(q)) = q$ and $f^{''}(\alpha(q)) < q$.

***Remark* 11.1.1** *As a result of the optimisation process, $\alpha(q)$ is the value for which the expression $q\alpha - f(\alpha)$ is extremal.*

Then, from the definition of the generalised fractal dimensions in Equation (11.1.2), they obtained

$$D_q = \frac{1}{q-1} \left[q\alpha(q) - f(\alpha(q)) \right] \tag{11.1.3}$$

so that knowing $f(\alpha)$ and the spectrum of α values, one can define D_q. Alternatively, given D_q, one can get $\alpha(q)$ [2] since

$$\alpha(q) = \frac{d}{dq}(q-1)D_q$$

Consequently, the generalised dimensions D_q provide an alternative description of the singular measure. Generalising the definition of the dimension D_q, they considered a strange set S embedded in a finite portion of d-dimensional Euclidean space, and partitioned the set into some number of disjoint pieces $S_1, .., S_N$ in which each piece has a measure p_i and lies within a ball of radius l_l, where each l_i is restricted by $l_i < l$. Then given the partition function

$$\Gamma(q, \tau, \{S_i, l\}) = \sum_{i=1}^{N} \frac{p_i^q}{l_i^\tau}$$

they argued that for large N, it is of the order unity only when $\tau = (q-1)D_q$. Defining

$$\Gamma(q, \tau) = \lim_{l \to 0} \Gamma(q, \tau, l)$$

and arguing that the function $\tau(q)$ is unique, they defined D_q as

$$(q-1)D_q = \tau(q) \tag{11.1.4}$$

which allows to recover $\alpha(q)$ and $f(\alpha(q))$. That is, replacing in Equation (11.1.3), we get

$$\tau(q) = \left[q\alpha(q) - f(\alpha(q)) \right] \tag{11.1.5}$$

and

[2] $\frac{d}{dq}\left[q\alpha(q) - f(\alpha(q)) \right] = \alpha(q) + q\frac{d}{dq}\alpha(q) - \frac{d}{d\alpha}f(\alpha)\frac{d}{dq}\alpha(q) = \alpha(q)$ and $f^{'}(\alpha(q)) = q$

$$\alpha(q) = \frac{d}{dq}\tau(q) \tag{11.1.6}$$

Note, $\tau(q)$ is the Legendre transform of $f(\alpha)$. Further, this definition of D_q precisely makes D_0 the Hausdorff dimension. They considered some simple examples to illustrate the quantities $\tau(q)$ and gain intuition about $\alpha(q)$ and $f(\alpha)$.

This multifractal formalism (MF), which is very useful in the characterisation of singular measures, accounts for the statistical scaling properties of these measures through the determination of their singularity spectrum $f(\alpha)$ which is intimately related to the generalised fractal dimension D_q. It reflects a deep connection with the thermodynamic formalism (TF) of equilibrium statistical mechanics where $\tau(q)$ and q are conjugate theremodynamic variables to $f(\alpha)$ and α. In fact, the concept of multifractality originated from a general class of multiplicative cascade models introduced by Mandelbrot [1974] in the context of fully developed turbulence (FDT). In the past, the D_q has usually been used on real, or, computer experiments, and the $f(\alpha)$ curves were determined from the Legendre transform of the $\tau(q)$ curves, which involved first smoothing the former and then Legendre transforming. A few years later, Chhabra et al. [1989] proposed a direct evaluation of $f(\alpha)$ without resorting to the intermediate Legendre transform. Denoting the probabilities in the boxes of size l as $p_j(l)$, they constructed a one-parameter family of normalised measure $\mu(q)$ given by

$$\mu_i(q,l) = \frac{p_i^q(l)}{\sum_j p_j^q(l)} \tag{11.1.7}$$

and they obtained the singularity spectrum as

$$\begin{aligned} f(q) &= \lim_{l\to 0} \frac{\sum_i \mu_i(q,l)\ln \mu_i(q,l)}{\ln l} \\ \alpha(q) &= \lim_{l\to 0} \frac{\sum_i \mu_i(q,l)\ln p_i(l)}{\ln l} \end{aligned} \tag{11.1.8}$$

which provides a relationship between the Hausdorff dimension f and an average singularity strength α as an implicit function of the parameter q. We are now going to discuss how one can estimate the $f(\alpha)$ spectrum based on the local dissipation.

11.1.2.2 Scaling analysis

As explained in Section (11.1.2.1), processes with spectra obeying a power law within a given (and sufficiently wide) range of frequencies (scales) are often refered to as $\frac{1}{f}$ processes

$$\Gamma_X(\nu) = C_0|\nu|^{-\gamma}\ , \nu_m \leq |\nu| \leq \nu_M$$

where $\Gamma_X(\nu)$ is any standard spectrum estimation procedure, $0 < \gamma < 1$ and C_0 is a positive constant. Two special cases where the scale range is semi-infinite, either at small frequencies, $\nu_m \to 0$ (equivalently large scales) or at large frequencies, $\nu_M \to \infty$ (small scales), define long-range dependent processes and monofractal processes. Scale invariance is generally defined as the power law behaviours of (the time average of the qth power of) multiresolution quantities, denoted $T_X(a,t)$ with respect to the analysis scale a, for a given (large) range of scales $a \in (a_m, a_M)$, $\frac{a_M}{a_m} >> 1$

$$\frac{1}{n_a}\sum_{k=1}^{n_a} |T_X(a,k)|^q \simeq c_q a^{\tau(q)} \tag{11.1.9}$$

where $T_X(a,k)$ describes the content of X around a time position t, and a scale a, such as a wavelet transform (see Muzy et al. [1991], Abry et al. [1995]). As explained in Section (11.1.2.1), $\tau(q)$ is a scaling exponent related to the generalised dimension via Equation (11.1.4). Hence, these quantities can be seen as some kind of spectral estimates.

Definition 11.1.1 *Scaling analysis*
The aim of scaling analysis is to validate the existence of power law behaviours as in Equation (11.1.9), and to measure the scaling exponents $\tau(q)$ that characterise them.

As suggested by Equation (11.1.9), the analysis and estimation procedures consist in tracking straight lines and estimating slopes in log-log plots. The estimated exponents can then be used for the physical understanding of the data or the system producing them. Note, the central theoretical notion of scale invariance is that of self-similarity, where the statistical information obtained from an object is independent of the scale of observation. However, self-similarity is a very demanding property, as it implies exact invariance to dilatation for all scale factors $a > 0$ of all finite dimensional distributions, therefore involving all statistical orders of the process. In practice, investigation of scale invariance is often restricted to the statistical order 2, or only certain range of scale factors. To be more flexible, one definition of scale invariance is obtained by relaxing the self-similarity property. First, scale invariance has to hold only for a restricted range of scaling factors $a \in (a_m, a_M)$, $\frac{a_M}{a_m} >> 1$, rather than all $a > 0$. Second, the (single) self-similarity parameter H is replaced with the function $\tau(q)$, called the scaling function or the scaling exponents of $X(t)$.

Definition 11.1.2 *Scale invariant stationary process*
Suppose $X(t)$ is a process with stationary increments. Then it is scale invariant if

$$E[|X(at)|^q] = |a|^{\tau(q)} E[|X(t)|^q]$$

for a range of scale $a \in (a_m, a_M)$, $\frac{a_M}{a_m} >> 1$ and some range of statistical orders q.

which implies for the increment process

$$E[|\delta_\theta X(t)|^q] = |\theta|^{\tau(q)} E[|\delta_1 X(0)|^q], 0 < \theta_m < \theta < \theta_M < \infty$$

where $\delta_\theta X(t) = X(t+\theta) - X(t)$ is a longitudinal velocity increment over a distance θ. The single parameter H is replaced with a function $\tau(q)$, which is in general non-linear ($\tau(q) \neq qH$), and hence represents a whole collection of parameters for the characterisation of the process.

While the work of Frisch et al. [1985] and Halsey et al. [1986], described in Section (11.1.2.1), showed that there was an infinite hierarchy of exponents allowing for a much more complete representation of the fractal measures, the related ideas of multifractality have only been applied to self-similar sets. Barabasi et al. [1991] extended the concept of multifractality to self-affine fractals, providing a more complete description of fractal surfaces. They investigated the multiscaling properties of the self-affine function $f(X)$ by calculating the qth order height-height correlation function

$$C_q(\Delta t) = \frac{1}{N} \sum_{t=1}^{N} |X(t, \Delta t)|^q$$

where $X(t, \Delta t) = X(t) - X(t - \Delta t)$, $N >> 1$ is the number of points over which the average is taken, and only the terms with $|X(t, \Delta t)| > 0$ are considered. They showed that this correlation function exhibited a nontrivial multiscaling behaviour if

$$C_q(\Delta t) \sim (\Delta t)^{qH(q)} \tag{11.1.10}$$

with $H(q)$ changing continuously with q at least for some region of the q values. As a result, multiscaling in empirical data is typically identified by differences in the scaling behaviour of different (absolute) moments

$$E[|X(t,\Delta t)|^q] = c(q)(\Delta t)^{qH(q)} = c(q)(\Delta t)^{\tau(q)} \tag{11.1.11}$$

where $c(q)$ and $\tau(q)$ are deterministic functions of the order of the moment q. From Equation 11.1.11 we can see that $\tau(q) = qH(q)$, which motivates the definition of the generalised Hurst exponent as

$$H(q) = \frac{\tau(q)}{q} \tag{11.1.12}$$

A similar expression holds for mono-scaling processes such as fBm

$$E[|X(t,\Delta t)|^q] = c^H(\Delta t)^{qH} \tag{11.1.13}$$

***Remark* 11.1.2** *There exists different ways of defining the notion of scale invariance and multifractality in the literature. We will discuss it further in Section (11.2.4.3).*

In terms of the behaviour of moments, monofractal, self-affine processes, are characterised by linear scaling, while multifractality (anomalous scaling) is characterised by non-linear (typically concave) shape. Hence, to diagnose multifractality, we can consider the inspection of the empirical scaling behaviour of an ensemble of moments. The traditional approach in the physics literature consists in extracting the self-similarity exponent, $\tau(q)$, from a chain of linear log-log fits of the behaviour of the various moments q for a certain selection of time aggregation steps Δt. We can therefore use regressions to the temporal scaling of moments of powers q

$$\ln E[|X(t,\Delta t)|^q] = a_0 + a_1 \ln(\Delta t) \tag{11.1.14}$$

and constructs the empirical $\tau(q)$ curve (for a selection of discrete q) from the ensemble of estimated regression coefficients for all q. An alternative and more widespread approach consists in looking directly at the varying scaling coefficients $H(q)$. This is called multifractal analysis which we introduce in Section (11.1.2.3). While the unique coefficient H quantifies a global scaling property of the underlying process, the multiplicity of such coefficients in multifractal processes, called Holder exponents, can be viewed as local scaling rates governing various patches of a time series, leading to a characteristically heterogeneous (or intermittent) appearance of such series. Focusing on the concept of Holder exponents, multifractality analysis amounts to identifying the range of such exponents rather than a degenerate single H as in the case of monofractal processes.

11.1.2.3 Multifractal analysis

So far, we focused on global properties such as moments and autocovariance, we are now going to adopt a more local viewpoint and examine the regularity of realised paths around a given instant. Contrary to scaling analysis which concentrate on the scaling exponent $\tau(q)$, the multifractal analysis (MFA) studies how the (pointwise) local regularity of X fluctuates in time (or space). Local Holder regularity describes the regularity of sample paths of stochastic processes by means of a local comparison against a power law function and is therefore closely related to scaling in the limit of small scales. The exponent of this power law, $h(t)$, is called the (local) Holder exponent and depends on both time and the sample path of X.

Definition 11.1.3 *Pointwise regularity*
A function $f(t)$, $f : \mathbb{R}^d \to \mathbb{R}$ is $C^\alpha(t_0)$ with $\alpha > 0$, denoted as $f \in C^\alpha(t_0)$, if there exist $C > 0$, $\epsilon > 0$ and a polynomial $P_{t_0}(\theta)$ of order strictly smaller than α such that

$$\text{if } |\theta| \leq \epsilon\,,\ |f(t_0+\theta) - P_{t_0}(\theta)| \leq C|\theta|^\alpha$$

If such polynomial $P_{t_0}(\theta)$ exists, it is unique, and its constant part is always given by $P_{t_0}(0) = f(t_0)$.

Definition 11.1.4 *Holder exponent*
The Holder exponent $h_f(t_0)$ of f at t_0 is

$$h_f(t_0) = \sup\{\alpha : f \in C^\alpha(t_0)\}$$

Note, if $P_{t_0}(0) = f(t_0)$ reduces to a constant, the Holder exponent characterises the power law behaviour of the increments at t_0

$$|f(t_0+\theta) - f(t_0)| \leq C(t_0)|\theta|^{h_f(t_0)}$$

where $C(t_0)$ is called the prefactor at time t_0. Put another way, the exponent $h_f(t_0)$ quantifies the scaling properties of the process at time t_0. The Holder exponent generalises the heuristic definition of singularity, so that it is also called the singularity exponent. That is, if f has Holder exponent $h(t_0) = h_f(t_0) < 1$ at t_0, then f has at t_0 either a cusp-type singularity

$$|f(t_0+\theta) - f(t_0)| \sim C|\theta|^{h(t_0)} \tag{11.1.15}$$

or an oscillating singularity

$$|f(t_0+\theta) - f(t_0)| \sim C|\theta|^{h(t_0)} \sin\left(\frac{1}{|\theta|^\beta}\right) \tag{11.1.16}$$

with oscillation exponent $\beta > 0$. Conversely, if f has either a cusp or an oscillating singularity at t_0 and $1 > h(t_0)$, then $h(t_0) = h_f(t_0)$ is the Holder exponent of f at t_0.

Multifractal situations happen when the Holder exponent is no-longer unique, but vary from point to point. More precisely, when the regularity $h(t)$ is itself a highly irregular function of t, possibly even a random process rather than a constant or a fixed deterministic function, the process X is said to be multifractal. The aim of multifractal analysis is to provide a description of the collection of Holder exponents h of the function f. Since the Holder exponent may jump from one point to another, it was understood that describing them for each time step was not of much importance. Hence, researchers focused on a global description of the regularity of the function of f in form of multifractal spectrum (also called the singularity spectrum) reflecting the size of the set of points for which the Holder exponent takes a certain value h. The measure of size most commonly used is the Hausdorff dimension which gives rise to the Hausdorff spectrum (called the multifractal spectrum). It describes the collection of Holder exponents $h(t)$ by mapping to each value of h the Hausdorff dimension $D(h)$ of the collection of points t_i at which $h_f(t_i) = h$. The main idea being that the relative frequency of the local exponents can be represented by a renormalised density called the multifractal spectrum.

Definition 11.1.5 *Iso-Holder sets*
The Iso-Holder sets $I_f(h)$ is the collection of points t_i for which the Holder exponent takes a certain value h.

$$I_f(h) = \{t_i | h_f(t_i) = h\}$$

For defining the Hausdorff dimension, we first need to define the Hausdorff measure.

Definition 11.1.6 *Hausdorff measure*
Let $S \subset \mathbb{R}^d$, $\epsilon > 0$, and let $\gamma_\epsilon(S)$ be ϵ-coverings of S, that is, bounded sets $\{c_n\}_{n\in\mathbb{N}}$ of radius $|c_n| \leq \epsilon$ (maximal distance between two elements of c_n) that cover $S : S \subset \gamma_\epsilon(S)$. Let $C_\epsilon(S)$ be the collection of all ϵ-coverings $\gamma_\epsilon(S)$ of S. The δ-dimensional Hausdorff measure of S is

$$m_\delta(S) = \lim_{\epsilon\to 0} \inf_{C_\epsilon(S)} \sum_{\gamma_\epsilon(S)} |c_n|^\delta$$

It can be shown that either $m_\delta(S) = 0$ if $\delta < \delta_c$, or $m_\delta(S) = \infty$ if $\delta > \delta_c$. The Hausdorff dimension of S is defined as the critical value δ_c.

Definition 11.1.7 *Hausdorff dimension*
The Hausdorff dimension $dim_H(S)$ of $S \subset \mathbb{R}^d$ is given by

$$dim_H(S) = \inf_\delta\{m_\delta(S) = \infty\} = \sup_\delta\{m_\delta(S) = 0\}$$

The multifractal spectrum assigns now to each Holder exponent, as a measure of its geometric importance, the Hausdorff dimension of the set of points that share the same exponent h.

Definition 11.1.8 *Multifractal spectrum*
The multifractal spectrum of a function f is defined as the Hausdorff dimension of the iso-Holder sets $I_f(h)$.

$$D_f(h) = dim_H(I_f(h))$$

We can now review the definition of monofractal and multifractal processes.

Definition 11.1.9 *Monofractal function or process*
A function or process $X(t)$ in $\mathbb{R}^d$ for which $h_X(t)$ is a constant, $\forall t : h_X(t) = H$, is called monofractal.

Definition 11.1.10 *Multifractal function or process*
A function or process $X(t)$ in $\mathbb{R}^d$ is called multifractal if it contains more than one Holder exponent h that is living on a support with non-zero Hausdorff dimension.

All homogeneous processes X with Holder exponents $h_X(x)$ not constant have their exponents which are discontinuous everywhere and hence are highly variable and change widely from point to point. As pointed out by Abry et al. [2002], one of the major consequences of multifractality in processes lies in the fact that quantities called partition functions present power law behaviours in the limit of small scales

$$S_\theta(q) = \sum_{k=1}^{\frac{1}{\theta}} |\delta_\theta X(k\theta)|^q \simeq c(q)|\theta|^{\tau(q)-1}, |\theta| \to 0$$

which correspond to Equation (11.1.10) with $N = \frac{1}{\theta}$. For processes with stationary increments, the time averages $\theta S_\theta(q)$ can be seen as estimators for the statistical averages $E[|\delta_\theta X(t)|^q]$ for $t = k\theta$. As a result, the above equation is highly reminiscent of the fundamental Equation (10.1.5) implied by self-similarity. However, the exponents $\tau(q)$ need not to follow the linear behaviour qH of self-similarity.

11.1.2.4 The wavelet transform and the thermodynamical formalism

Considering the multifractal spectrum of a process to be the multifractal spectrum of each of his realisations, we can speak of multifractal spectrum for both functions and processes. Hence, multifractal analysis determine the multifractal spectrum which describe a local property, the point-wise regularity of a function, globally through the geometrical importance of different Holder exponents, disregarding any information on their precise geometric repartition. Following Parisi et al. [1985], under some assumptions on the homogeneity and isotropy of the statistics of local singularities, it is possible to derive a relation between self-similarity exponents $\tau(q)$ and the singularity spectrum $D(h)$. They showed that the self-similarity exponents $\tau(q)$ could be computed from the the Legendre transformation [3] of the singularity spectrum $D(h)$

[3] The Legendre transformation is a mathematical operation transforming a function of a coordinate, $g(x)$, into a new function $h(y)$ whose argument is the derivative of $g(x)$ with respect to x, i.e., $y = \frac{dg(x)}{dx}$.

$$\tau(q) = \inf_h \{qh + d - D(h)\} \tag{11.1.17}$$

where d is the dimension of the embedding space. This is to be related to the work on fractal measures by Halsey et al. [1986] and described in Section (11.1.2.1).

***Remark* 11.1.3** *It correspond to Equation (11.1.3) with Holder exponent satisfying $\alpha' = h$, that is, before it has been minimised. Once it has been minimised we can replace h with $h(q)$. We also see that the multifractal spectrum $D(h)$ corresponds to the singularity spectrum $f(\alpha(q))$.*

Using this equation, one can then relate statistics and geometry. Note, the singularity spectrum is often estimated by using indirect methods which are numerically more stable, such as the Legendre spectrum. We can invert the equation above since the Legendre spectrum correspond to the Legendre transform of the self-singularity exponents $\tau(q)$. That is, the spectrum of Holder exponents (or multifractal spectrum, or Legendre spectrum) can be obtained by the Legendre transformation of the scaling function $\tau(q)$ as

$$D_L(h) = \inf_q \{qh + d - \tau(q)\}$$

which can be computed numerically since there exists a lot of estimation procedure for the scaling function $\tau(q)$. For example, Calvet et al. [2002] reported the spectrum of multiplicative measures when the random variable V is binomial, Poisson, or Gamma. Note, this method failed to fully characterise the singularity spectrum $D(h)$, since only the strongest singularities are amenable to this analysis. In order to determine the whole singularity spectrum $D(h)$, Muzy et al. [1991] considered the wavelet transform (WT) and established the foundations for a thermodynamical formalism (TF) for singular signals. In that setting, the singularity spectrum $D(h)$ is directly determined from the scaling behaviour of partition functions defined from the wavelet transform modulus maxima (WTMM) (see Section (11.2.2.1) for details). Improving on the standard partition multifractal formalism, the wavelet transform modulus maxima (WTMM) is based on wavelet analysis, and involves tracing the maxima lines in the continuous wavelet transform over all scales (see Arneodo et al. [1995]). Later, under the uniform Holder regularity condition for X, Wendt [2008] showed that $D_L(h)$ with $\tau_L(q)$ estimated from the wavelet Leader structure functions

$$S^L(j,q) = \frac{1}{n_j} \sum_{k=1}^{n_j} L_X^q(j,k)$$

where $L_X(j,k)$ are wavelet Leader coefficients, provided a tight upper bound to the multifractal spectrum $D(h)$

$$D(h) \leq D_L(h)$$

***Remark* 11.1.4** *The structure function $S^L(j,q)$ is a special case of the time average of the qth power of multiresolution quantities $T_X(a,t)$ in Equation (11.1.9), where $T_X(a,t)$ has been replaced with the wavelet Leader coefficients. We will see that other representations of the quantities $T_X(a,t)$ exist, defining alternative models.*

The wavelet leader multifractal formalism (WLMF) asserts that this inequality turns into an equality

$$\forall h\ ,\ D(h) = D_L(h) = \inf_q \{qh + d - \tau(q)\}$$

so that the Legendre spectrum of X can be interpreted in terms of the Holder singularities of X (see Wendt [2008]). One can then compute numerically the Legendre spectrum $D_L(h)$. Hence, the scaling exponents $\tau(q)$ and the multifractal spectrum $D(h)$ are closely related via the Legendre transform. That is, the power law behaviours together with the multifractal spectrum constitute the fundamental relations establishing the connection between scale invariance and multifractality.

11.1.3 Observing multifractality in financial data

As explained in Section (11.1.2.2), when different values of H are found in different regions of the signal, then the signal is multifractal rather than monofractal. These multifractal signals can only be described in terms of an infinite set of exponents and their density distribution. The characterisation of a multifractal process or measure by a distribution of local Holder exponents underlines its heterogeneous nature with alternating calm and turbulent phases. We are going to illustrate how this phenomena was observed in the financial literature.

11.1.3.1 Applying multiscaling analysis

The moment-scaling properties of financial returns have been the object of a growing physics literature confirming that multiscaling is exhibited by many financial time series. While the temporal correlations of FX rates had been analysed on the daily evolution of several FX rates (see Peters [1991-96] [1994]), Galluccio et al. [1997] studied high-frequency data covering one year. Since FX rates suffer from systematic effects mainly due to the periodicity of human activities, they considered FX rates as processes with non-stationary increments, and introduced a time transformation to eliminate these systematic effects. The main idea, proposed by Dacorogna et al. [1993], is to use volatility as an indicator of the activity in the market, and expand times when this activity is large, and shrink times in the opposite situation. Using data on FX rates obtained by the Olsen & Associates Research Institute from 1992 to 1993, they found hidden correlations in the data. In the inner time statistics, in which the signal has stationary increments, the price variations scale in time with Hurst value of $\frac{1}{2}$. However, the FX rates were found to have a far more complex nature than simple random walk, since the sign correlation function and the absolute value correlation function showed correlations in the system, and FX rates also showed multiscaling behaviour. Vandewalle et al. [1997] performed a Detrending Fluctuation Analysis (DFA) (see details in Section (10.3.2.1)) on the daily evolution of several currency exchange rates from 1980 till 1996, where data was collected at 2.30 pm Brussels time. They found that the evolution of the JPY/USD exchange rate had a power law with exponent $H = 0.55 \pm 0.01$ holding over two decades in time. Classifying the behaviour of exchange rates, they also found that the exponent values and the range over which the power law holds varied drastically from one currency exchange rate to another, obtaining three categories, the persistent behaviour, the antipersistent behaviour, and the strictly random one. In addition, they showed in the USD/DEM currency exchange rate that the Hurst exponent was changing with time with successive persistent and antiperisistent sequences. To probe the local nature of the correlations, Vandewalle et al. [1998c] constructed an observation box of length T placed at the beginning of the data and computed its Hurst exponent. They moved the box by a few points (4 weeks) toward the right and again computed the H value. Iterating the procedure for the 1980-1996 period, they obtained a local measurement of the degree of long-range correlations over T. Looking at the evolution of the USD/DEM ratio on a window of size $T = 2$ years, the global exponent was $H = 0.56 \pm 0.01$, and the mobile (local) exponent was varying between 0.4 and 0.6. Repeating the test for the evolution of the GBP/DEM ratio, Vandewalle et al. [1998c] found the global exponent to be at $H = 0.55$, and the local exponent was mostly above $\frac{1}{2}$ which when averaged got back the global value. The multifractality of the foreign exchange rate market was reported in further studies (see Schmitt et al. [1999]). Vandewalle et al. [1998] presented the $H(q)$ spectrum and found the USD/DEM and JPY/USD exchange rates to be multifractal. However, the statistical tools identifying multifractal behaviour have been the subject to dispute as a number of studies showed that scaling in higher moments can easily be obtained in a spurious way without any underlying anomalous diffusion behaviour (see Granger et al. [1999]). For instance, Barndorff-Nielsen et al. [2001] found apparent scaling as a consequence of fat tails in the absence of true scaling. Randomising the temporal structure of financial data, Lux [2004] obtained a non-linear shape of the empirical $\tau(q)$ function, concluding that $\tau(q)$ and $f(\alpha)$ estimators were rather unreliable diagnostic instruments to identify multifractal structure in volatility. Note, even though econometricians did not look at scaling functions and Holder spectrums, the indication of multifractality was mentioned in economics literature. For instance, Ding et al. [1993] found that different powers of returns have different degrees of long-term dependence, and that the intensity of long-term dependence varies non-monotonically with q, which is consistent with concavity of scaling functions and provides evidence for anomalous behaviour.

11.1.3.2 Applying multifractal fluctuation analysis

Analysing the NYSE daily composite index closes from 1966 to 1998 and the USD/DM currency exchange rates from 1989 to 1998, Pasquini et al. [2000] considered the de-averaged daily return defined as

$$r_t = \log \frac{S_{t+1}}{S_t} - < \log \frac{S_{t+1}}{S_t} >$$

where S_t is the index quote or exchange rate quote at time t. They computed the autocorrelation for returns and showed that it was a vanishing quantity for all period L. They repeated the test for powers, γ, of absolute returns and found for $\gamma = 1$ a non-vanishing quantity up to $L = 150$, showing dependent return. Then, they introduced the cumulative returns $\phi_t(L)$ defined as

$$\phi_t(L) = \frac{1}{L} \sum_{i=0}^{L-1} r_{t+i}$$

which is the profile in DFA, and given $\frac{N}{L}$ non overlapping variables of this type, they computed the associated variance $Var(\phi(L))$. Using this tool, they confirmed that returns were uncorrelated. To perform the appropriate scaling analysis, they introduced the generalised cumulative absolute returns defined as

$$\phi_t(L, \gamma) = \frac{1}{L} \sum_{i=0}^{L-1} |r_{t+i}|^\gamma$$

and showed that if the autocorrelation for powers of absolute returns $C(L, \gamma)$ [4] exhibits a power-law with exponent $\alpha(\gamma) \leq 1$ for large L, that is $C(L, \gamma) \sim L^{-\alpha(\gamma)}$, it implies

$$Var(\phi(L, \gamma)) \sim L^{-\alpha(\gamma)}$$

Note, anomalous scaling can not be detected if $|r_t|^\gamma$ are short-range correlated or power-law correlated with an exponent $\alpha(\gamma) > 1$. They found that $\alpha(\gamma)$ was not a constant function of γ, showing the presence of different anomalous scales. Hence, different values of γ select different typical fluctuation sizes, any of them being power-law correlated with a different exponent. It is important to note that a direct analysis of the autocorrelations can not provide clear evidence for multiscale power-law behaviour, since the data show a wide spread compatible with different scaling hypothesis. On the other hand, the scaling analysis introduced by Pasquini et al. proves the power-law behaviour and precisely determines the coefficients $\alpha(\gamma)$. Following a different route, Richards [2000] demonstrated that exchange rates scale as fractals, and that the determinants of exchange rate/interest rates and differential in real rates of return, have also fractal properties. Assuming scale invariance from Definition (11.1.2), he estimated $\tau(q)$ by regressing the linear Equation (11.1.14). Following Schertzer et al. [1991b], he considered the theoretical distribution of $\tau(q)$ given by

$$\begin{aligned} \tau(q) &= qH - \frac{C(1)}{(\alpha - 1)}(q^\alpha - q) \text{ when } \alpha \neq 1 \\ \tau(q) &= qH - C(1) q \ln q \text{ when } \alpha = 1 \end{aligned}$$

where the standard normal is given by $H = \frac{1}{2}$, $\alpha = 2$, $C(1) = 0$ and $\tau(q) = qH$. In the fractal case, $C(1) \neq 0$ and the scaling curve is nonlinear. As $C(1)$ increases above zero and α decreases toward 1, then $\tau(q)$ becomes progressively more curved at higher orders of scaling. The degree of curvature is a measure of the turbulence of the process. Computing the first derivative of $\tau(q)$ at the origin, he estimated α and $C(1)$. Working on time series of 18 currencies from 1971 to 1998, the reported values for H were lying in the range $[0.53, 0.63]$ with a majority between 0.55 and 0.59 showing persistence. The nominal differential showed H in the range $[0.44, 0.45]$ and the real differential showed H

[4] $C(L, \gamma) =< |r_t|^\gamma |r_{t+L}|^\gamma > - < |r_t|^\gamma >< |r_{t+L}|^\gamma >$

in the range $[0.34, 0.35]$. Richards [2000] found a crucial difference between the stochastic fractality found in capital markets and the multifractals identified in physics. The latter exhibit strong scaling symmetries, with proportionality relationships over a wide range of time scales, but in the former, scaling symmetries are much weaker, and hold only over shorter intervals.

11.2 Holder exponent estimation methods

While the monofractal structure of financial signals are defined by single power law exponent, assuming that the scale invariance is independent on time and space, spatial and temporal variation in scale invariant structure often occurs, indicating a multifractal structure of the financial signals. We are now going to describe several methods for the multifractal characterisation of nonstationary financial time series.

***Remark* 11.2.1** *In Section (11.1.2) we denoted the generalised Hurst exponent as $H(q)$ and the local Holder exponent as $h(t)$. In the economic and financial literature, the former is sometime denoted as $h(q)$ and the latter as $H(t)$. When considering an article under study, we will use its notation, and clarify the meaning when necessary.*

11.2.1 Applying the multifractal formalism

Since scaling analysis and multifractal analysis developed, we saw in Section (11.1.3) that various authors performed empirical analysis to identify anomalous scaling in financial data. In order to illustrate the multifractal formalism, we choose to briefly describe the method proposed by Calvet et al. [2002], who investigated the MMAR model, which predicts that the moments of returns vary as a power law of the time horizon, and confirmed this property for Deutsche mark/US dollar exchange rates and several equity series. We let $P(t)$ be the price of a financial asset on the bounded interval $[0, T]$ and define the log-price process as

$$X(t) = \ln P(t) - \ln P(0)$$

Calvet et al. first assumed that $X(t)$ was a compound process $X(t) = B[\theta(t)]$, where $B(t)$ is a Brownian motion, and $\theta(t)$ is a stochastic trading time. Second, they extended the model to get autocorrelated returns by modifying $X(t)$ as $X(t) = B_H[\theta(t)]$, where $B_H(t)$ is a fractional Brownian motion. On a given path, the infinitesimal variation in price around t is of the form

$$|\ln P(t + dt) - \ln P(t)| \sim C(t)(dt)^{\alpha(t)}$$

where $\alpha(t)$ is the local Holder exponent or local scale of the process at time t. Partitioning $[0, T]$ into integer N with intervals of length Δt, the partition function is given by

$$S_q(T, \Delta t) = \sum_{i=0}^{N-1} |X(i\Delta t + \Delta t) - X(i\Delta t)|^q$$

When $X(t)$ is multifractal, the increments are identically distributed, and the scaling law yields

$$E[S_q(T, \Delta t)] = Nc(q)(\Delta t)^{\tau(q)+1}$$

when the qth moment exists, which implies

$$\ln E[S_q(T, \Delta t)] = \tau(q) \ln \Delta t + c^*(q)$$

where $c^*(q) = \ln c(q) + \ln T$ since $T = N\Delta t$. Various methods can be used for constructing an estimator $\hat{\tau}(q)$ from the sample moments of the data. As discussed in Section (11.1.2.4), the Legendre transform $\hat{f}(\alpha)$ is then applied to obtain an estimate of the multifractal spectrum, which can be mapped back into a distribution of the multipliers. In

general, given a set of positive moments q and time scales Δt, the partition functions $S_q(T, \Delta t)$ are calculated from the data, which are then plotted against Δt in logarithmic scales. The linear equation above implies that these plots are approximately linear when the qth moment exists. Regression estimates of the slopes provide the corresponding scaling exponents $\hat{\tau}(q)$. Working with two data sets provided by Olsen and Associates, Calvet et al. [2002] investigated the multifractality of the Deutsche mark/US dollar exchange rates. The former is made of daily data from 1973 to 1996, and the latter contains high-frequency data from 1992 to 1993 which is modified to account for seasonality. Analysing the time series for Δt spanning 15 seconds to 6 months, and five values of q ranging from 1.75 to 2.25, they observed linearity of the partition functions starting at $\Delta t = 1.4$ hours and extending to six months. The slope was zero for q slightly smaller than 2, and they obtained $\hat{H} \approx 0.53$ implying very slight persistence in the DM/USD series. Then, in view of estimating the scaling functions $\hat{\tau}(q)$ for both data sets, they considered a larger range of moments $1.5 \leq q \leq 5$ for Δt between 1.4 hours and 6 months to capture information in the tails of the distribution of returns. Increasing variability of the partition function plots with the time scale Δt was observed. An estimate of the multifractal spectrum $\hat{f}(\alpha)$ was obtained by a Legendre transform of the moments' growth rates, which was concave for daily data. Given the estimated spectrum, they induced a generating mechanism for the trading time based MMAR model. From the shape of the estimated spectrum, nearly quadratic, they inferred a lognormal distribution (of multipliers M) as the primitive of the generating mechanism and estimated its parameters. Simulating the process with MMAR model, they confirmed that the multifractal model replicated the moment behaviour found in the data. Since the estimated value of the most probable local Holder exponent was greater than $\frac{1}{2}$, the estimated multifractal process was therefore more regular than a Brownian motion. But the concavity of the spectrum also implied the existence of lower Holder exponents corresponding to more irregular instants of the price process, contributing disproportionately to volatility. Repeating the analysis in a sample of five major US stocks and an equity index for daily returns from 1962 to 1998, they further demonstrated the scaling behaviour for many scaling financial series.

11.2.2 The multifractal wavelet analysis

We saw in Section (11.1.2.1) that multifractal theory identifies fractal objects that can not be completely described by using a single fractal dimension (monofractal), as they have an infinite number of dimension measures associated with them. The multifractal scaling of an object is characterised by

$$N_\epsilon \propto \epsilon^{-f(\alpha)}$$

where N_ϵ is the number of boxes of length ϵ required to cover the object, and $f(\alpha)$ is the dimension spectrum, which can be interpreted as the fractal dimension of the set of points with scaling index α. We also saw that an alternative multifractal description consisted in extracting the spectrum $D(h)$ of Holder exponents h of the velocity field from the inertial scaling properties of structure functions. However, we saw in Section (11.1.2.4) that there was fundamental drawbacks to these methods characterising the singularity spectrum $f(\alpha)$ and $D(h)$. Considering wavelet decomposition tools, Mallat et al. [1990], Muzy et al. [1991], and Arneodo et al. [1991] generalised the multifractal formalism to singular signals.

11.2.2.1 The wavelet transform modulus maxima

The wavelet transform (WT) can be used as a mathematical microscope to analyse the local regularity of functions. In particular, we get the power law for the WT of the isolated cusp singularity in $f(x_0)$, for fixed position x_0 (or t_0), defined as

$$T(a, x_0) \sim |a|^{h(x_0)} \text{ for } a \to 0$$

in the condition that the wavelet has at least n vanishing moments [5]. Thus, one can extract the exponent $h(x_0)$, for fixed position x_0, from a log-log plot of the WT amplitude versus the scale a. However, the exponent $h(x_0)$ is governed by

[5] It is orthogonal to polynomials up to degree n, that is, $\int_{-\infty}^{\infty} x^m \psi(x) dx = 0 \, \forall m, 0 \leq m < n$.

singularities accumulating at x_0, which results in unavoidable oscillations around the expected power-law behaviour of the WT amplitude, making the exact determination of h uncertain. Since there exist fundamental limitation to the measure of Holder exponents from local scaling behaviour in a finite range of scales, the determination of the singularity spectrum $D(h)$ required a more appropriate investigation of WT local behaviour. An alternative solution is to use the WTMM tree for defining partition function based multifractal formalism (MF). Mallat et al. [1992] showed that for isolated cusp singularities (see Equation (11.1.15)), the location of the singularity could be detected, and the related exponent could be recovered from the scaling of the wavelet transform (WT), along the maxima line, converging towards the singularity. For this particular line, the WT reaches local maximum with respect to the position coordinate. Hence, they are likely to contain all the information on the hierarchical distribution of singularities in the signal. Connecting these local maxima within the continuous wavelet framework, we obtain the entire tree of maxima lines, and restricting oneself to such a collection provides a very useful representation called the wavelet transform modulus maxima (WTMM) of the entire continuous wavelet transform (CWT) (see Mallat et al. [1992b]). This method incorporates the main characteristics of the WT, that is, the ability to reveal the hierarchy of singular features, including the scaling behaviour.

As discussed in Section (11.1.2.1), we find the multifractal spectrum of a signal by partitioning it into N boxes of length ϵ. A probability density, $P(\epsilon, i)$, of the signal in each box, labelled i, is calculated where $P(\epsilon, i)$ is the fraction of the total mass of the object in each box. The qth order moments $M(\epsilon, q)$ are then calculated as

$$M(\epsilon, q) = \sum_{i=1}^{N_\epsilon} P(\epsilon, i)^q \tag{11.2.18}$$

For a multifractal object, this moment function scales as

$$M(\epsilon, q) \propto \epsilon^{\tau(q)}$$

where $\tau(q)$ is a scaling exponent. From this scaling, we saw in Section (11.1.2.1) that both α and the singularity spectrum $f(\alpha)$ can be calculated from

$$\alpha(q) = \frac{d\tau(q)}{dq}$$

and

$$f(\alpha) = q\alpha(q) - \tau(q)$$

In the wavelet based method for calculating the $f(\alpha)$ spectrum, one approach proposed by Muzy et al. [1991] is to let the function in Equation (11.2.18) be replaced by the wavelet based moment function, getting

$$M(a, q) = \sum_{\Omega(a)} |T(a, w_i(a))|^q \sim a^{\tau(q)} \text{ for } a \to 0$$

where $\Omega(a) = \{w_i(a)\}$ is the set of all maxima $w_i(a)$ at the scale a, and $|T(a, w_i(a))|$ is the ith wavelet transform modulus maxima found at scale a. The function $M(a, q)$ is a partition function of the qth moment of the measure distributed over the wavelet transform maxima at the scale a considered. By summing only over the modulus maxima, it directly incorporates the multiplicative structure of the singularity distribution into the calculation of the partition function (see Muzy et al. [1991] [1993]). Since the moment q has the ability to select a desired range of values, small for $q < 0$, or large for $q > 0$, the scaling function $\tau(q)$ globally captures the distribution of the exponent $h(x)$. Further, since we get $\tau(0) = -D_f$, where D_f is the fractal dimension, then D_f can be seen as the ratio of the logarithms of the average maxima multiplicative rate and the average scale factor, respectively. Following Chhabra et al. [1989], we can use the moment function $M(a, q)$ to directly estimate $\alpha(q)$ and $f(\alpha)$ as

$$\alpha(q) = \lim_{a\to 0} \frac{d}{dq} \frac{\log M(a,q)}{\log a} = \lim_{a\to 0} \frac{1}{\log a} \sum_{\Omega(a)} P(a,q,w_i(a)) \log |T(a,w_i(a))|$$

where

$$P(a,q,w_i(a)) = \frac{|T(a,w_i(a))|^q}{\sum_{\Omega(a)} |T(a,w_i(a))|^q}$$

is the weighting measure for the statistical ensemble $\Omega(a)$. Similarly, we get

$$f(\alpha(q)) = \lim_{a\to 0} \frac{1}{\log a} \sum_{\Omega(a)} P(a,q,w_i(a)) \log P(a,q,w_i(a))$$

obtaining the spectrum in a parametric form (the parameter being q) from the log-log plots of the above quantities. Comparing these results with those obtained by Chhabra et al. we see that the probability $p_i^q(l)$ corresponds to $|T(a,w_i(a))|^q$ and that the parametric measure in Equation (11.1.7) corresponds to $P(a,q,w_i(a))$. One drawback of tracking the WTMM is that there may exist extra maxima in the wavelet representation not corresponding to any singularity in the signal. In general, for multifractal processes $\tau(q)$ is an increasing convex nonlinear function of q, and its Legendre transform $D(h)$ is a well-defined unimodal curve, the support of which extends over a finite interval $h_{min} \leq h \leq h_{max}$. The maximum of this curve $D(h(q=0)) = -\tau(0)$ gives the fractal dimension of the support of the set of singularities of f. In general, $\tau(1) \neq 0$, indicating that the cascade of WTMM is not conservative.

Other wavelet formalism exist, mainly consisting in replacing the coefficients in the moment function $M(a,q)$ with other coefficients. The solution relies on the use of the coefficients of either a continuous wavelet transform, or a discrete wavelet transform. For instance, one formalism considered the wavelet coefficients $d_X(j,k)$ leading to the wavelet coefficient based multifractal formalism (WCMF), while another formalism used wavelet Leaders leading to the wavelet Leader multifractal formalism (WLMF) (see Wendt [2008]). Riedi et al. [1999] developed a wavelet-based multifractal model for use in computer traffic network modelling, and also considered applications to finance and geophysics.

11.2.2.2 Wavelet multifractal DFA

We saw in Section (10.3) that DFA and its variants are methods used for analysing the behaviour of the average fluctuations of the data at different scales after removing the local trends. Further, wavelet transforms (WT) are well adapted to evaluate typical self-similarity properties. As a result, some authors (see Murguia et al. [2009], Manimaran et al. [2009]) merged the two approaches obtaining powerful tools for quantifying the scaling properties of the fluctuations. Using the vanishing moment property of wavelets, Manimaran et al. [2005] proposed to separate the trend from the fluctuations of time series with discrete wavelet method. That is, instead of a polynomial fit, we consider the different versions of the low-pass coefficients to calculate the local trend. Again, we work with the profile $X(t)$ given in Equation (10.3.11) and compute the multilevel wavelet decomposition of that profile. For each scale m, we subtract the local trend to data

$$Y(t,m) = X(t) - \overline{X}(t,m)$$

where $\overline{X}(t,m)$ is the reconstructed profile after removal of successive details coefficients at each scale m. We then divide the residual series $Y(t,m)$ into A adjacent sub-periods of length n. For each sub-period, a, we can compute the local detrended fluctuations

$$F^2_{WDFA,q}(a,m) = \Big(\frac{1}{n}\sum_{i=1}^{n} Y_m^2(i,a)\Big)^{\frac{q}{2}}, a = 1, ..., A$$

which is, for $q = 2$, the local root-mean-square (RMS) variation of the time series. Since the detrending of the time series depends of the wavelet decomposition at the scale, m, different level DFA differ in their capability of eliminating the trends in the series. Averaging over the A sub-periods, for different q, the qth order fluctuation function becomes

$$F_{WDFA,q}(n,m) = \Big(\frac{1}{A}\sum_{i=1}^{A} F^2_{WDFA,q}(i,m)\Big)^{\frac{1}{q}}$$

where q can take any real value, except zero. The value for $q = 0$ is computed via a logarithmic averaging procedure

$$F_{WDFA,q}(n,m) = \frac{1}{A}\sum_{i=1}^{A} \ln F^2_{WDFA,2}(i,m)$$

We repeat this calculation to find the fluctuation function $F_{WDFA,q}(n,m)$ for many different box sizes n, which should reveal a power law scaling

$$F_{WDFA,q}(n,m) \sim n^{h(q)}$$

where $h(q)$ is the generalised Hurst exponent.

11.2.3 The multifractal fluctuation analysis

We described in Section (3.2.2.2) the most common measures of variability when summarising the data. When studying the response time distributions of financial return, we saw in Section (10.4.3) that some authors considered the mean absolute price change (ΔX where X is a logarithmic price) as their main scaling parameter, while other considered the square root of the variance and the interquartile range of the distribution of ΔX. An alternative approach is to directly work with the process X and to detrend it with a polynomial, or an averaging function and measure the error with the local root mean square.

11.2.3.1 Direct and indirect procedure

All multifractal analyses of a response time series follow a step-wise procedure, which was described by Ihlen [2013], where the response is first decomposed into both time and scale domains by

1. computing a scale-dependent measure $\mu_n(t_0)$ in a floating trial interval $[t_0 - \frac{n}{2}, t_0 + \frac{n}{2}]$ centred at time t_0, or by
2. computing a scale-dependent measure $\mu_n(t,a)$ in the non-overlapping trial interval $[(a-1)n+1, an]$, defined as

$$\mu_n(t,a) = \Big(\frac{1}{n}\sum_{i=1}^{n}[X(i,a) - \hat{X}_n(i,a)]^2\Big)^{\frac{1}{2}}, a = 1, ..., A \tag{11.2.19}$$

   where $\hat{X}_n(i,a)$ is a scale-dependent trend.

The analyses directly estimating the multifractal spectrum from $\mu_n(t)$ use method 1, while the analyses based on q-order statistics of the measure $\mu_n(t,a)$ use method 2. In the former, the temporal variation in the trial dependency of a response time series can be defined by the local exponents $H(t_0)$ when the measure $\mu_n(t_0)$ satisfies the power law

$$\mu_n(t_0) \sim \lim_{n\to 0} n^{H(t_0)}$$

In general, the local exponent $H(t)$ is estimated as the linear regression slope of log-log plot of the scale-dependent measure $\mu_n(t_0)$ against the scale n. In the latter, the qth order statistical moments $E[\mu^q_n(t,a)]$ satisfy the power law

$$E[\mu_n^q(t,a)] = \frac{1}{A}\sum_{a=1}^{A}\mu_n^q(t,a) \sim n^{\tau(q)}$$

when the response time series has a multifractal structure. The qth order scaling exponent $\tau(q)$ is estimated as the linear regression slope of $E[\mu_n^q(t,a)]$ against n in log-log coordinates.

11.2.3.2 Multifractal detrended fluctuation

Kantelhardt et al. [2002] proposed the multifractal detrended fluctuation analysis (MF-DFA) as a generalisation of the detrended fluctuation analysis (DFA), with the advantage that it can be used for non-stationary multifractal data. Rather than considering deviation from a constant, or, linear trend, in the computation of the range, they considered deviation from a polynomial fit of different moment q. Similarly to the rescaled range analysis, the integrated time series $X(t)$ is divided into sub-periods, and the polynomial fit of order l, noted $X_{n,l}$, of the profile is estimated for each sub-period. For instance, if we set $l = 1$ we get the linear detrending signal $Y(t,a) = X(t,a) - X_{n,1}(t,a)$ for sub-period $a = 1, .., A$. For each sub-period of length, n, the local detrended fluctuation becomes

$$F_{DFA,q}^2(a) = \mu_n^q(t,a) = \Big(\frac{1}{n}\sum_{i=1}^{n} Y^2(i,a)\Big)^{\frac{q}{2}}, a = 1, ..., A \tag{11.2.20}$$

where $\mu_n(t,a) = F_{DFA,1}^2(a)$ corresponds to the scale-dependent measure with $\hat{X}_n(i,a) = X_{n,l}(i,a)$. It is then averaged over the A sub-periods, for different q, to form the qth order fluctuation function

$$F_{DFA,q}(n) = \Big(\frac{1}{A}\sum_{i=1}^{A} F_{DFA,q}^2(i)\Big)^{\frac{1}{q}}$$

where q can take any real value, except zero. The value of $h(0)$, corresponding to the limit of $h(q)$ for $q \to 0$, is computed via a logarithmic averaging procedure

$$F_{DFA,0}(n) = e^{\frac{1}{A}\sum_{i=1}^{A}\ln F_{DFA,2}^2(i)}$$

We repeat this calculation to find the fluctuation function $F_{DFA,q}(n)$ for many different box sizes n, which then scale as

$$F_{DFA,q}(n) \sim n^{h(q)} \tag{11.2.21}$$

where c is a constant independent of n, and $h(q)$ is the generalised Hurst exponent for a particular degree of polynomial. Note, the value of $F_{DFA,q}(n)$ increases as n increases. The Hurst exponent is estimated through ordinary least squares regression on logarithms of both sides

$$\log F_{DFA,q}(n) \approx \log c + h(q)\log n \tag{11.2.22}$$

Note, the authors chose to exclude large scales $n > \frac{N}{4}$ from the fitting procedure as well as small scales $n < 10$. This method has important characteristics when it comes to examining processes with heavy tails since $H(q) \approx \frac{1}{q}$ for $q > \alpha$, and $h(q) \approx \frac{1}{\alpha}$ for $q \leq \alpha$ where α is the parameter of stable distributions. Note, for $q = 2$, and stationary time series, we recover the standard DFA. Also, by comparing the MFDFA results for original series with those for shuffled series, on can distinguish multifractality due to long-range correlations from multifractality due to a broad probability density function. While the MFDFA may be compared to the WTMM method, there is no need to employ a continuously sliding window or to calculate the supremum over all lower scales, since the variances $F_{DFA}^2(n,a)$ will always increase when the segment size, n, is increased. Moreover, note that the MFDFA method can only determine positive generalised Hurst exponents $h(q)$, and it is inaccurate for strongly anti-correlated signals when $h(q)$ is close to zero. One way forward is to integrate the time series before the MFDFA procedure, by replacing

the single summation describing the profile from original data in Equation (10.3.11) with the double summation $\tilde{X}(i) = \sum_{k=1}^{i}[X(k) - < X >]$. In that setting, the fluctuation function scales as

$$\tilde{F}_{DFA,q}(n) \sim n^{\tilde{h}(q)} = n^{h(q)+1}$$

leading to accurate scaling behaviour for $h(q)$ smaller than zero but larger than 1. Note $\frac{\tilde{F}_{DFA,q}(n)}{n}$ corresponds to $F_{DFA,q}(n)$. Since the double summation leads to quadratic trends in the profile $\tilde{X}(i)$, another solution is to use a second order MFDFA to eliminate these trends.

11.2.3.3 Multifractal empirical mode decomposition

The multifractal empirical mode decomposition (MFEMD) is similar to the multifractal detrended fluctuation, but the scale-dependent trend $X_n(t,a)$ is now defined by the intrinsic mode function $d(t,a)$ defined by the empirical mode decomposition of $X(t)$ (see Huang et al. [1998]). That is,

$$X_n(t,a) = \sum_{\lceil a \rceil} d(t,a)$$

where $\lceil a \rceil$ indicates that the sum is taken over all intrinsic mode functions $d(t,a)$ with a scale larger than a. One way of defining the scale of each $d(t,a)$ is by considering the inverse of their mean instant frequency given by $a = \frac{1}{<f_{t,a}>}$ where the frequency $f_{t,a}$ is computed by differentiating the phase angle of the Hilbert transform of $d(t,a)$, that is, $f_{t,a} = \frac{d\theta_{t,a}}{dt}$, using an algorithm suggested by Boashash [1992]. This detrending procedure was proposed by Manimaran et al. [2009].

11.2.3.4 The R/S ananysis extented

Rather than locally detrending multifractal time series by computing the qth order RMS variations and then averaging over equal-sized non-overlapping segments to measure the multifractals of financial time series, Kim et al. [2004] directly considered the qth price-price correlation function

$$F_q(\tau) =< |p(t+\tau) - p(t)|^q > \propto \tau^{qh(q)}$$

where τ is the time lag, and $\{p(t_1), .., p(t_N)\}$ is a price time series of length N. Building a time series of return $\{r_1(\tau), .., r_N(\tau)\}$ where $r_i(\tau) = \ln p(t_i + \tau) - \ln p(t_i)$, and dividing the series into A subseries of length n, they computed the rescaled range analysis with $(R/S)_n(\tau) \propto n^{h(\tau)}$ to study the tick behaviour of the yen-dollar exchange rate. Results for data ranging from January 1971 to June 2003 showed $H(\tau = 1) = 0.6513$ which is significantly different from $H = \frac{1}{2}$. Once the process has been located in the persistence region, they studied the log-log plot of $\frac{F_q(\tau)}{q}$, q = 1, .., 6, against τ which is equivalent to plotting RMS against the scale. They could observe that the generalised Hurst exponent $h(q)$ was a function of q converging as τ got larger. They also computed the probability distribution of returns and showed that the series was consistent with a Lorentz distribution, concluding that the yen-dollar exchange rate was multifractals.

11.2.3.5 Multifractal detrending moving average

Gu et al. [2010] extended the DMA method to multifractal detrending moving average (MFDMA) in order to analyse multifractal time series and multifractal surfaces. The algorithm for the one-dimensional MFDMA is similar to that of the DMA, where the moving average accounts for backward average ($\theta = 0$), centred average $\theta = \frac{1}{2}$, and forward average $\theta = 1$

$$\overline{X}_\lambda(t) = \frac{1}{\lambda} \sum_{k=-\lfloor(\lambda-1)\theta\rfloor}^{\lceil(\lambda-1)(1-\theta)\rceil} X(t-k)$$

where $\lfloor x \rfloor$ is the largest integer not greater than x, $\lceil x \rceil$ is the smallest integer not smaller than x, and θ is the position parameter with values ranging in $[0, 1]$. The moving average considers $\lceil (\lambda - 1)(1 - \theta) \rceil$ data points in the past and $\lfloor (\lambda - 1)\theta \rfloor$ points in the future. We then compute the detrended signal (residual series)

$$Y(t) = X(t) - \overline{X}_\lambda(t)$$

where $\lambda - \lfloor (\lambda - 1)\theta \rfloor \leq t \leq T - \lfloor (\lambda - 1)\theta \rfloor$. The residual series $Y(t)$ is divided into A disjoint segments with the same size, n, where $A = \lfloor \frac{N}{n} - 1 \rfloor$. Each segment is denoted by

$$Y(t, a) = Y((a - 1)n + t)\,, t = 1, .., n$$

The qth function $F_{DMA,q}(a)$ with segment size n is given by

$$F^2_{DMA,q}(a) = \Big(\frac{1}{n}\sum_{i=1}^{n} Y^2(i, a)\Big)^{\frac{q}{2}}\,, a = 1, .., A$$

and the qth order overall fluctuation function satisfies

$$F_{DMA,q}(n) = \Big(\frac{1}{A}\sum_{i=1}^{A} F^2_{DMA,q}(i)\Big)^{\frac{1}{q}}$$

where q can take any real value, except zero. Varying the values of segment size n, we define the power-law relation between the function $F_{DMA,q}(n)$ and the size scale n

$$F_{DMA,q}(n) \sim n^{h(q)} \tag{11.2.23}$$

Note, the window size λ used to determine the moving average function must be identical to the partitioning segment size, otherwise $F_{DMA,q}(n)$ does not show power-law dependence on the scale n. We can then compute the multifractal exponent $\tau(q)$ given in Equation (11.1.17), and the singularity spectrum $f(\alpha)$. Gu et al. concluded that the backward MFDMA with the parameter $\theta = 0$ exhibits the best performance when compared with centred and forward MFDMA. They also showed that it was better than the MFDFA because it gave better power-law scaling in the fluctuations and more accurate estimates of the multifractal scaling exponents and singularity spectrum.

11.2.3.6 Some comments about using MFDFA

Ihlen [2012] presented a step-by-step tutorial of MFDFA in an interactive Matlab session. In order to guide users when applying MFDFA to financial time series, we follow the best practice proposed by Ihlen. We saw in Section (10.3.2.1) that the DFA identifies the monofractal structure of a time series as the power law relation between the overall RMS computed for multiple scales. The slope, H, of the regression line defines a continuum between a noise like time series and a random walk like time series. The Hurst exponent will take values in the interval between 0 and 1 in the former, and it will take values above 1 in the latter. The pink noise $H = 1$ separates between the noises $H < 1$ having more apparent fast evolving fluctuations, and random walks $H > 1$ having more apparent slow evolving fluctuations. Monofractal and multifractal have a long-range dependent structure with Hurst exponent between 0.7 and 0.8. Differentiating between random walk and noise like time series, Eke et al. [2002] proposed to run a monofractal DFA before running MFDFA as it should be performed on a noise like time series. If the series have Hurst exponent between 0.2 and 0.8, it is a noise like series and can be directly employed without transformation, otherwise if the time series are random walk, it can either be differentiated first, or the conversion to a random walk can be eliminated.

Errors in the estimation of the generalised Hurst exponent $h(q)$ occur when the RMS is close to zero because $\log F_{DFA,q}(n)$ for negative q's become infinitely small. Extreme large $h(q)$ will be present for negative q's as output, so that local RMS close to zero will lead to large right tails for the multifractal spectrum. One reason is that the polynomial trend fit $X_{n,l}$ of the time series can be overfitted in segments with small sample sizes (small scale). This

overfitted trend will be close, or similar, to the time series causing the residual fluctuations in Equation (10.3.12) to be close to zero. The sample size of the smallest segment should therefore be much larger than the polynomial order l to prevent from an overfitted trend. Another reason, is that the time series might be smooth with little apparent variation, hence, similar to the polynomial trend even for low order l. In that case, the value of the smallest scales should be raised, and the scale invariance checked carefully. This problem of segments with RMS close to zero can be solved by eliminating RMS below a certain threshold ϵ.

The scale invariance of a time series can be detected when the log-log plot of $F_{DFA,q}(n)$ versus n yield a linear relationship for a given q. If the relationship is curved or S-shaped the qth order Hurst exponent $h(q)$ should not be estimated by a linear regression. One reason for non-linear relation in the log-log plot is an insufficient order l for the polynomial detrending. One solution is to run MFDFA with different l and compare their log-log plot. Another reason is that local fluctuations RMS close to zero for small scales can yield a non-linear dip in the lower end of the log-log plot. This dip can be prevented by elimination of RMS below the threshold ϵ. Further, a non-linear relationship might originate from phenomenon recorded in the time series causing scale invariance to break down at the smallest scales. One way to detect sub-regions with scale invariance is to look for periods with approximately constant $\frac{F_{DFA,q}(n)}{n}$, $q = 1, ..,$ in a log-log plot against the sample size n within the entire scaling range. The scales where $\log \frac{F_{DFA,q}(n)}{n}$ are no-longer constant indicates the segment sizes above and below which the local fluctuations (RMS) are no-longer scale invariant.

Defining the optimal length T of a time series being investigated, as well as the minimum scale n_{min} and its maximum value n_{max}, is still an open problem. Grech et al. [2004] proposed to use the scale range $5 < n < \frac{T}{5}$, and applied DFA procedure on several segments of the time series. They chose 500 segments, and looked at $140 < T < 300$ to define the optimal T on the basis of minimum standard deviation of estimated H exponents, and minimum measure of statistical uncertainty $\frac{E_T[H]}{\hat{\sigma}_T(H)}$. Alternatively, Einstein et al. [2001] proposed to use $n_{min} = 4$ and $n_{max} = \frac{T}{4}$, Weron [2002] proposed to use $n_{min} = 10$ for short time series of 256 or 512 observations and $n_{min} = 50$ for longer series, and Matos et al. [2008] and Kantelhardt proposed to use $n_{max} = \frac{T}{4}$ and suggested to be cautious for small ranges as they can significantly overestimate the resulting Hurst exponent. Even though one usually choose the minimum segment sample size to be $n_{min} \approx 10$, it might be too small for large trend order l. On the other hand, the maximum segment size n_{max} should be small enough to provide a sufficient number of segments in the computation of $F_{DFA,q}(n)$. For instance, a maximum segment size below $\frac{1}{10}$ of the size N of the time series will provide at least 10 segments in the computation of $F_{DFA,q}(n)$. The parameter q, deciding the qth order weighting of the local fluctuation RMS, should consist of both positive and negative values in order to weight the periods with large and small variation in the time series. Since the choice of the qth order should avoid large negative and positive values because they inflict larger numerical errors in the tails of the multifractal spectrum, a sufficient choice would be for q to range between -5 and 5. Further, as the local fluctuation RMS is computed around a polynomial trend whose shape is defined by the order l, a higher order l yield a more complex shape of the trend, but it might lead to overfitting for time series within small segment sizes. Thus, $l = 1 - 3$ are a sufficient choice when the smallest segment sizes contains $10 - 20$ samples.

Note, the MFDFA can be extended to include more adaptive detrending procedure like wavelet decomposition (see Manimaran et al. [2009]), moving average, and empirical mode decomposition. An adaptive fractal analysis is shown to perform better than DFA with polynomial detrending when employed to biomedical time series with strong trends (see Gao et al. [2011]). At last, statistical parameters other than RMS can be used to define the local fluctuation in a time series. For instance, in multifractal analyses based on wavelet transformations, the local fluctuation is defined as the convolution product between the time series and a waveform fitted within local segments of the time series. One can then compare MFDFA with wavelet transformation modulus maxima (WTMM), multifractal analysis with wavelet leaders (see Jaffard et al. [2006]), and gradient modulus wavelet projection. Further, in an entropy-based estimation of the multifractal spectrum, the local fluctuation is defined as the sum of the time series within the local segment relative to the total sum of the entire time series. It uses a qth order entropy function instead of a qth order

RMS, and directly estimates $\alpha(q)$ and $f(\alpha)$ as the regression slope of the qth order entropy functions (see Chhabra et al. [1989]).

11.2.4 General comments on multifractal analysis

11.2.4.1 Characteristics of the generalised Hurst exponent

Even though monofractal and multifractal time series have similar RMS and slopes, H, they have different structures. The generalised Hurst exponent, $h(q)$, is a decreasing function of q for multifractal time series, and a constant for monofractal processes. For monofractal time series with compact support, $h(q)$, is independent of q. Only if small and large fluctuations scale differently, will there be a significant dependence of $h(q)$ on q. For positive (negative) values of q, the exponent $h(q)$ describes the scaling behaviour of segments with large (small) fluctuations. Usually, the large fluctuations are characterised by a smaller scaling exponent $h(q)$ for multifractal series than the small fluctuations. More precisely, multifractal time series have local fluctuations with both extreme small and extreme large magnitudes which are absent in the monofractal time series, resulting in a normal distribution for the monofractal time series (variations are characterised by the first two moments). On the other hand, in the multifractal time series, the behaviour of the local fluctuations is such that the qth order statistical moments must be considered. This is why we introduced the qth order fluctuation in the MFDFA above, where the qth order weights the influence of segments with large and small fluctuations. The local fluctuations $F^2_{DFA,q}(a)$ with large and small magnitudes is graded by the magnitude of the negative or positive q-order, respectively. The midpoint $q = 0$ is neutral to the influence of segments with small and large $F^2_{DFA,q}(a)$. Note, the difference between the qth order RMS for positive and negative q's are more visually apparent at the small segment sizes than at the large ones because the small segments can distinguish between the local periods with large and small fluctuations as they are embedded within these periods. On the other hand, the large segments cross several local periods with both small and large fluctuations averaging out their differences in magnitude. Hence, for the largest segment sizes, the qth order RMS of the multifractal time series converges to that of the monofractal series. As a result, the function $h(q)$ is decreasing with respect to the q's, indicating that the segments with small fluctuations have a random walk like structure whereas segments with large fluctuations have a noise like structure.

11.2.4.2 Characteristics of the multifractal spectrum

Definition We saw in Section (11.1.2.1) that multifractal series are also described by the singularity spectrum $f(\alpha)$ through the Legendre transform (see Peitgen et al. [1992])

$$\alpha(q) = \frac{d\tau(q)}{dq}, f(\alpha) = \arg\min_q (q\alpha - \tau(q))$$

where $f(\alpha)$ denotes the fractal dimension of the series subset characterised by the singularity strength, or Holder exponent, α. The Holder exponent quantifies the local scaling properties (local divergence) of the process at a given point in time, that is, it measures the local regularity of the price process. The spectrum $f(\alpha)$ describes the distribution of the Holder exponents. For monofractal signals, the singularity spectrum produces a single point in the $f(\alpha)$ plane, whereas multifractal processes yield a humped function (see Figure 11.1). From the definition of the generalised Hurst exponent in Equation (11.1.12), we can directly relate the singularity α and spectrum $f(\alpha)$ to $h(q)$ by replacing its value in the above equation

$$\alpha(q) = h(q) + q\frac{dh(q)}{dq} \text{ and } f(\alpha) = q[\alpha(q) - h(q)] + 1$$

Following Schumann et al. [2011], the multifractality degree (or strength of multifractality) in finite limit $[-q, +q]$ can be described by

$$\Delta h_q = h(-q) - h(+q) \tag{11.2.24}$$

Since large fluctuations in the response time series are characterised by smaller scaling exponent $h(q)$ than small fluctuations, then $h(-q)$ are larger than $h(+q)$, so that Δh_q is positively defined. Another quantifier for the multifractality degree (or width) for the same limit is

$$\Delta\alpha = \alpha|_{q=-q} - \alpha|_{q=+q} \tag{11.2.25}$$

We let α_0 be the position of the maximum spectrum $f(\alpha)$ and define the skew parameter as

$$r = \frac{\alpha|_{q=+q} - \alpha_0}{\alpha_0 - \alpha|_{q=-q}}$$

with $r = 1$ for symmetric shapes, $r > 1$ for right-skewed shapes, and $r < 1$ for left-skewed shapes. The width of the spectrum $\Delta\alpha$ measures the degree of multifractality in the series (the wider the range of fractal exponents, the richer the structure of the series). The skew parameter r determines the dominant fractal exponents, that is, fractal exponents describing the scaling of small fluctuations for right-skewed spectrum, or fractal exponents describing the scaling of large fluctuations for left-skewed spectrum. These parameters allow to measure the complexity of a series. A signal having a high value of α_0, a wide range $\Delta\alpha$ of fractal exponents, and a right-skewed shape $r > 1$ may be considered more complex than one with opposite characteristics (see Shimizu et al. [2002]). The parameters $(\alpha_0, \Delta\alpha, r)$ are called the measures of complexity of a series.

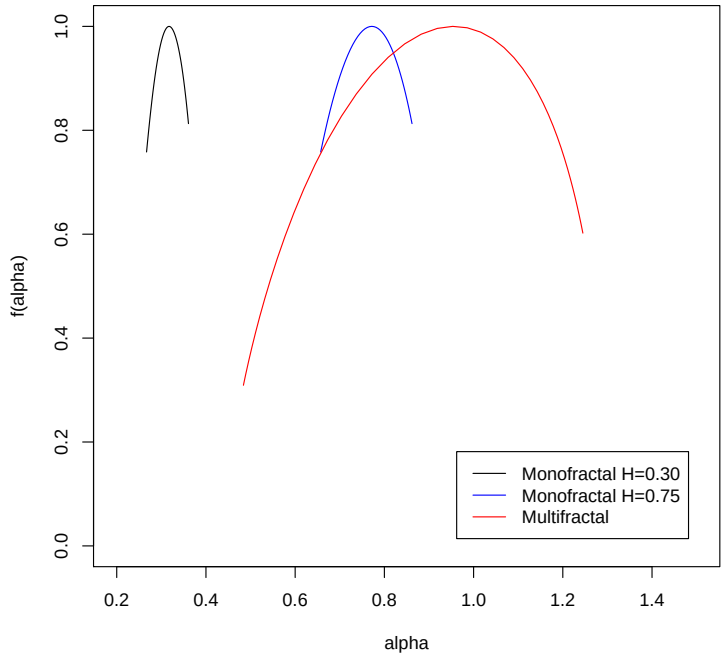


Figure 11.1: Multifractal spectrum $(\alpha(q), f(\alpha))$. The multifractral spectrum is the log-transformation of the normalised probability distribution of local Hurst exponents. Therefore, when the width of the spectrum is wide (red), the series is multifractal. On the contrary, when the width of the spectrum is small (black and blue), the series is monofractal. In this case, the peak is reached on the Hurst exponent.

Interpretation We saw in Section (11.2.4.3) that $h(q)$ is only one of several types of scaling exponents used to parametrise the multifractal structure of time series. The plot of the singularity strength (exponent) $\alpha(q)$ versus the singularity spectrum $f(\alpha)$ is referred to as the multifractal spectrum. The multifractal time series has multifractal exponent $\tau(q)$ with a curved q-dependency, and a decreasing singularity exponent $\alpha(q)$. That is, the resulting multifractal spectrum is a large arc where the difference between the maximum and minimum $\alpha(q)$ are called the multifractal spectrum width (or degree) given in Equation (11.2.25). The Hurst exponent defined by the monofractal DFA represents the average fractal structure of the time series and is closely related to the central tendency of multifractal spectrum. The deviation from average fractal structure for segments with large and small fluctuations is represented by the multifractal spectrum width. Thus, each average fractal structure in the continuum of Hurst exponents has a new continuum of multifractal spectrum widths representing the deviations from the average fractal structure. Moreover, the shape of the multifractal spectrum does not have to be symmetric, it can have either a left or a right truncation originating from a levelling of the q-order Hurst exponent for positive or negative q's, respectively. The spectrum will have a long left tail when the time series have a multifractal structure being insensitive to the local fluctuations with small magnitudes, but it will have long right tail when the structure is insensitive to the local fluctuations with large magnitudes. Consequently, the width and shape of the multifractal spectrum can classify a wide range of different scale invariant structures of time series.

11.2.4.3 Some issues regarding terminology and definition

We discussed in Section (11.1.2.1) the singularity spectrum $f(\alpha)$ associated with the singularity strength α, and we showed how they were derived and how to estimate them. We also discussed in Section (11.1.2.2) how to extract the spectrum $D(h)$ of Holder exponents h. The concepts and terminologies from econophysics have some time been mixed up in finance, leading to confusion or incomplete statements. We present some of the issues regarding terminology and definition used in recent financial articles on detrending fluctuation analysis.

First of all, depending on the way multifractality is defined, the interpretation of the exponents in the fluctuation analysis differs. For instance, when considering the multifractal model of asset returns (MMAR), since the multipliers defined at different stages of the cascade are independent, Calvet et al. [2002] inferred

$$E[\mu(\Delta t)^q] = (\Delta t)^{\tau(q)+1}$$

where $\tau(q) = -\ln E[M^q] - 1$ and by extension to stochastic processes defined multifractality as

Definition 11.2.1 *A stochastic process $\{X(t)\}$ is called multifractal if it has stationary increments and satisfies*

$$E[|X(t)|^q] = c(q)t^{\tau(q)+1} \text{ for all } t \in \mathcal{F}\,,\, q \in \mathcal{Q}$$

where $\tau(q)$ and $c(q)$ are functions with domain $\mathcal{Q}$. Denoting by H the self-affinity index, the invariance condition $X(t) = t^H X(1)$ implies that $E[|X(t)|^q] = t^{qH}E[|X(1)|^q]$ so that $c(q) = E[|X(1)|^q]$ and as a result $\tau(q) = qH - 1$. In this special case, the scaling function $\tau(q)$ is linear and fully determined by its index H. In the more general case, $\tau(q)$ is a nonlinear function of q. That is, $\tau(q)$ is a linear function for monofractal signals, and a nonlinear one for multifractal signals (see Figure 11.2). Note, this definition of multifractality is slightly different from the one given in Equation (11.1.11), where $\tau(q) = qH(q)$, leading to a modified exponent $\tau(q)$.

For example, applying Definition (11.2.1) in finance, Kantelhardt et al. [2002] showed that if we consider only stationary, normalised and positive series [6] defining a measure with compact support, it is possible to omit the detrending procedure in the MFDFA. In that setting, similarly to Barabasi et al. [1991], they showed that the multifractal scaling exponent $h(q)$ was related to the classical multifractal exponent $\tau(q)$ defined by the standard partition multifractal formalism. To prove their statement, they considered the relation

[6] $x_k \geq 0$ and $\sum_{k=1}^N x_k = 1$.

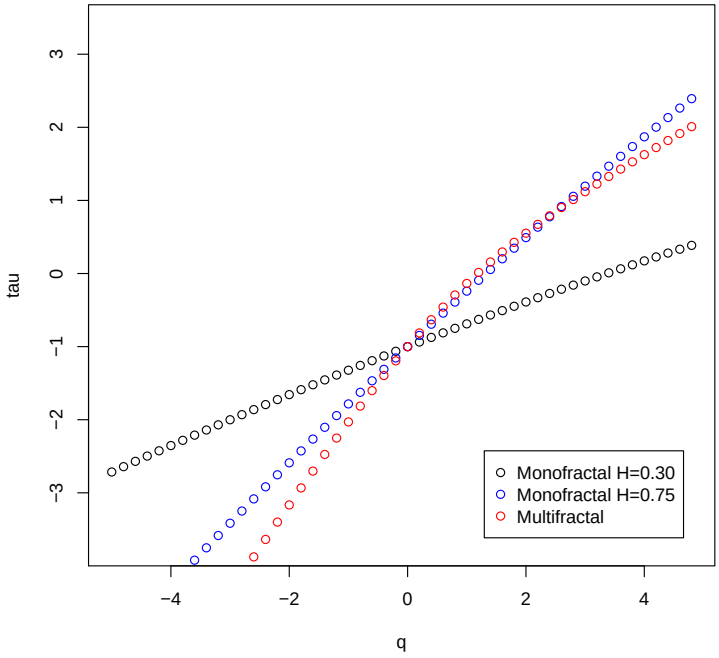


Figure 11.2: Multifractal scaling exponent $\tau(q)$. The multifractal scaling exponent is nonlinear when the time series exhibit a multifractal structure.

$$\Big(\frac{1}{N_s}\sum_{k=1}^{N_s}|X(ks)-X((k-1)s)|^q\Big)^{\frac{1}{q}} \sim s^{h(q)}$$

and assuming that the length N of the series is an integer multiple of the scale s, that is, $N_s = \frac{N}{s}$, the above equation becomes

$$\sum_{k=1}^{\frac{N}{s}}|X(ks)-X((k-1)s)|^q \sim s^{qh(q)-1}$$

Further, since the term $X(ks)-X((k-1)s)$ is identical to the sum of the numbers x_k within each segment k of size s, then, it is known as the box probability $p_s(k)$ in standard multifractal formalism for normalised series (see Section (11.1.2.1)). The scaling exponent $\tau(q)$ being defined via the partition function $Z_q(s)$ as

$$Z_q(s) = \sum_{k=1}^{\frac{N}{s}}|p_s(k)|^q \sim s^{\tau(q)}$$

where $\tau(q)$ is the scaling function related to the multifractal spectrum via the Legendre transform. Relating the two equations they found the relation between the two sets of multifractal scaling exponents to be

$$\tau(q) = qh(q) - 1$$

which correspond to Equation (11.1.17).

11.3 The need for time and scale dependent Hurst exponent

11.3.1 Computing the Hurst exponent on a sliding window

11.3.1.1 Introducing time-dependent Hurst exponent

Even though the Hurst exponent can only be estimated together with a confidence interval, the stability of the Hurst exponent is related to the fact that a process is self-similar only when the coefficient H is well defined. Examining the stability of H on financial time series on the basis of characteristic values, such as rescaled ranges or fluctuations analysis, some authors (see Vandewalle et al. [1998c], Costa et al. [2003], Cajueiro et al. [2004], Grech [2005]) observed that the values of H could be significantly higher or lower for a specific scale. For example, using the DFA method, Costa et al. [2003] analysed the behaviour of the Sao Paulo stock exchange Ibovespa index covering over 30 years of data from 1968 till 2001, amounting to $8,209$ trading days. For the complete time series they obtained $H = 0.6$, indicating persistence in the Ibovespa index. They also computed the local Hurst exponent with a three year sliding window, and found that $H(t)$ varied considerably over time. They found that the local exponent was always greater than $\frac{1}{2}$ before 1990, and then rapidly dropped towards $\frac{1}{2}$ and stayed there afterwards. They interpreted this phenomena as a consequence of the economic plan adopted in March 1990, corresponding to structural reforms in the Brazilian economy. Separate analysis of the two periods gave $H = 0.63$ for the former and $H \approx \frac{1}{2}$ for the latter, indicating multifractality of the process.

Several authors proposed to use methods of long-range analysis such as DFA or DMA to determine the local correlation degree of the series by calculating the local scaling exponent over partially overlapping subsets of the analysed series (see Vandewalle et al. [1998c], Costa et al. [2003], Cajueiro et al. [2004], Carbone et al. [2004], Matos et al. [2008]). The proposed technique examines the local Hurst exponent $H(t)$ around a given instant of time. The ability of DFA and DMA algorithm to perform such analysis relies on the local scaling properties of the scale-dependent measure given in Equation (11.2.19). Following this method, we can calculate the exponent, $H(t)$, without any a priori assumption on the stochastic process and on the probability distribution function of the random variables entering the process. As a result, one can deduce trading strategies from these instabilities of the Hurst exponent H, such as optimal investment horizon (see Lo [1991]). However, the existence of such optimal investment horizons contradicts both EMH and FMH. In the former, it implies potential predictability of the market, while in the latter it implies that self-similar structure of the market breaks down, turning the market into a spiral. Consequently, these authors introduced time-dependent Hurst exponent for which the examination of different scales solely is not possible. In that setting, we can focus on significant changes in values of Hurst exponent as such changes imply significant shift of rescaled ranges or fluctuations at either low or high scales.

11.3.1.2 Describing the sliding window

Again, we consider the integrated returns $X(t)$ of a time series of length T which we divide into, A, adjacent sub-periods of length, n, such that $An = T$. However, rather than computing the Hurst exponent on the whole series, we introduce a sliding window W_s of size N_s, moving along the series with step, δ_s, and compute the scaling exponent on each window obtaining a sequence of Hurst exponent. That is, the window is rolled δ_s points forward, eliminating the first δ_s observations and including the next ones, and the Hurst exponent is re-estimated. In the trivial case of a unique window coinciding with the entire series, the size of the sequence is 1. On the other hand, in the case of a sliding window moving point-by-point along the series, $\delta_s^{min} = 1$, then the size of the sequence is $T - N_s$. To determine a local Hurst exponent H at a particular time t, one has to consider a time interval around t that is considerably smaller than the total time spanned by the data, but still sufficiently large to contain enough points for a meaningful statistics. Starting at the beginning of the time series, one compute the exponent H considering only data points within N_s from the initial point, and then we roll the window, obtaining the curve for $H(t)$ where t denotes the origin of each window.

Alternatively, we can shift the time t so that it denotes the mid-point of each window, that is, $(t - \frac{N_s}{2}, t + \frac{N_s}{2})$. While the question of finding the correct window's size in which to estimate the Hurst exponent is an open problem, there is a tradeoff between the minimum necessary number of points to get proper statistics and the maximum number of points to obtain constant, or almost constant, Hurst exponent. One way forward is for the size of the sliding window to reflect some cycles (economical or political) in the country under study. Another way is to let the minimum size N_s^{min} of each window be defined by the condition that the scaling law in Equation (11.2.21) holds in each window (typically $N_s^{min} \sim 2000 - 3000$). The maximum resolution of the technique is achieved with N_s^{min} and δ_s^{min}. Hence, rather than operating on the global properties of the series by Legendre or Fourier transform of qth order moment (power spectrum analysis), this technique allows for the examination of the local scaling exponent $H(t)$ around a given instant of time.

The main reason why the rolling window approach is favoured to attain the local Hurst exponent is that the Hurst exponent obtained from the whole series does not necessarily imply the presence or absence of long-range correlations, since the exponent obtained could be due to the averaging of those negative and positive correlations. Further, the rolling window approach can also disclose the dynamics of market efficiency. The choice of the length of the rolling sample varies for different purposes. Some authors used window length of $1,008$ business days, or around 4 years of data, to analyse the evolution of market efficiency, while others (see Wang et al. [2011]) used shorter window length of 250 business days, or around 1 year of data, in their researches. While the former is more stable, it is believed that the Hurst exponent can lose its locality if the length of the rolling window is too large (see Grech et al. [2004]). However, for too small periods, the fluctuations in the values of H are too large to be trusted. Costa et al. [2003] found that a 3-year period (736 trading days) was an acceptable compromise between the two opposing demands.

11.3.1.3 Understanding the time-dependent Hurst exponent

In order to understand the time dependence of the Ibovespa Hurst exponent, Costa et al. [2003] considered a theoretical framework based on the multifractal Brownian motion (MFBM) characterised by the Hurst function

$$H(t) = 0.63 - 0.076 \arctan(30t - 24)$$

This example was chosen to mimic the generic trend seen in the Ibovespa where $H(t)$ could be viewed as having (on average) two distinct plateaus, one before 1990 and the other one after. Simulating $N = 2^{13}$ data points from the (bi-fractal) MFBM path, and assuming that the DFA fluctuation function $F_{DFA}(n)$ exhibits an approximate scaling regime given by $F_{DFA}(n) \sim n^{\overline{H}}$, they found $\overline{H} = 0.6$, up to about $n = 160$. Hence, they interpreted the scaling exponent $\overline{H}$ as a kind of average Hurst exponent for the MFBM time series. While the region of small H was only about 20% of the total time series, it had nonetheless a significant weight on the exponent $\overline{H}$, since its value was a lot smaller than the one obtained by simply averaging the curve $H(t)$ ($H_{avg} = 0.68$). Therefore, regions of large fluctuations (smaller H) tend to dominate the behaviour of $F_{DFA}(n)$ for small to intermediate values of n (see Hu et al. [2001]). Hence, the Hurst exponent obtained by applying fluctuation analyses (DFA, DMA etc.) to financial time series may only capture average behaviour over the time period spanned by the data. It was clearly illustrated by their study on the Ibovespa index where $H = 0.6$ on the complete time series (spanning over 30 years of data), whereas during that period $H(t)$ was ranging within $[0.76, 0.42]$. As a result, when dealing with long financial time series, a more localised (in time) analysis is necessary to determine a possible dependence of H on time. In particular, a global Hurst exponent equal to $\frac{1}{2}$ does not necessarily imply absence of correlation, as there could be positive and negative correlations at different periods of time averaging out to yield an effective $H = \frac{1}{2}$.

11.3.1.4 Time and scale Hurst exponent

We saw above that one way of generalising the global behaviour of the Hurst exponent was to estimate the local Hurst exponent $H(t)$ for fixed size widows, N_s, in view of studying its time dependency. Exploring the dichotomy between emerging and mature markets, Matos et al. [2008] considered the local Hurst exponent $H(t, N_s)$ as a function of both time t and scale N_s. To do so, they applied DFA on the interval $(t - \frac{N_s}{2}, t + \frac{N_s}{2})$ in order to recover a power of the

scale inside each sub-interval. The technique called time and scale Hurst exponent (TSH) is performed with maximum scale $N_{s_{max}} = \frac{T}{4}$. In that framework $H(t, N_s)$ is the focus of the analysis, and the DFA or DMA are implementation details. Its goal is to provide a map of the evolution of markets to maturity, including significant episodes (major events affecting markets), but also gradual changes in response times. Considering daily data for a set of worldwide market indices from America, Asia, Africa, Europe and Oceania, they observed several notable features

- They can distinguish mature markets by the stability of H values around $\frac{1}{2}$, and emergent markets by the stability of H values above $\frac{1}{2}$.
- For some periods, a phase transition occurs, sometimes observable across all scales, sometimes across partial scales only. This is reflected by spikes occurring for both lower and larger scales.
- They expected smooth variations of H for large scales since it takes more data into account leading to greater robustness to sudden changes of the data. It was confirmed empirically.
- Some markets evolve in time, where they observed a shift from emergent to mature features. The other markets slowly decreased in the values of the Hurst exponent.
- Significant events, causing marked change in the Hurst exponent behaviour was seen in almost all markets.

We see that studying the Hurst behaviour for multiple time and scale intervals gives more refined details on the time series. Matos et al. proposed a refined classification of financial markets in three states

- mature markets: they have H around $\frac{1}{2}$. The presence of regions with higher values of H is limited to small periods and well defined both in time and scale.
- emergent markets: they have H well above $\frac{1}{2}$. The presence of regions with values of H around $\frac{1}{2}$ is well defined both in time and in scale.
- hybrid markets: unlike the two other cases, the distinction between the mature and emergent phases is not well determined, with behaviour seemingly mixing at all scales.

11.3.2 Testing the markets for multifractality

11.3.2.1 A summary on temporal correlation in financial data

We saw that long-range dependence (LRD) is defined by hyperbolically declining autocorrelations either for a process itself or some functions of it. This property was analysed in the context of fractional integration of Brownian motion (fBm) by Mandelbrot [1967] and Mandelbrot et al. [1968]. Using R/S analysis in the context of asset returns, they suggested H-values of around 0.55 to be representative for stock returns. In contrast, using his modified R/S statistics, Lo [1991] did not find evidence of long-range dependence in the Research in Security Prices (CRSP) data, and attributed the previous findings to the failure of R/S analysis in presence of short-range dependence. Based on the work of Lo [1991], various authors generally believed that there was little evidence of fractional integration in stocks and FX returns. Since then, LRD has been well documented in squared and absolute returns for many financial data sets, but not in raw returns (see Muller et al. [1990], Ding et al. [1993], Dacorogna et al. [1993]).

Many continuous multifractal models, such as the multifractal model of asset returns (MMAR), have been proposed to capture the thick tails and long-memory volatility persistence exhibited in the financial time series. Such models are consistent with economic equilibrium, implying uncorrelated returns and semimartingale prices, thus precluding arbitrage in a standard two-asset economy. Returns have a finite variance, and their highest finite moment can take any value greater than 2. However, the distribution does not need to converge to a Gaussian distribution at low frequencies and never converges to a Gaussian at high frequencies, thus capturing the distributional nonlinearities observed in financial series. These multifractal models have long memory in the absolute value of returns, but the

returns themselves have a white spectrum. That is, there is long memory in volatility, but absence of correlation in returns.

Eveven though it has been assumed little evidence of fractional integration in stock returns (see Lo [1991]), long memory has been identified in the first differences of many economic series. For instance, Maheswaran et al. [1993] suggested potential applications in finance for processes lying outside the class of semimartingales. Later, using the same CRSP data, Willinger et al. [1999] showed that Lo's acceptance of the hypothesis for the CRSP data was strongly biased, and found empirical evidence of long-range dependence in stock price returns. As a result, some authors modelled financial series directly with the fBm or the discrete-time ARFIMA specification. Accounting for autocorrelation in returns together with multifractality in volatility, Calvet et al. [2002] proposed the fractional Brownian motion of multifractal time.

Most of the studies made on LRD were performed in the time or frequency domain by neglecting any bias or trend in the signal, even though the statistical methods used were designed for processes with stationary time increments. However, stock returns and FX rates suffer from systematic effects mainly due to the periodicity of human activities, and can not be considered as processes with stationary increments. Clegg [2006] showed that LRD was a very difficult property to measure in real life because the data must be measured at high lags/low frequencies where fewer readings are available, and all estimators are vulnerable to trends in the data, periodicity and other sources of corruption.

A better approach for analysing the characteristics of financial markets is to consider the detrended fluctuation analysis (DFA) introduced by Peng et al. [1994] to investigate long-range power-law correlations along DNA sequences. In this method, one filters the series not only from a constant trend, but also from higher order polynomials. Constructing an integrated series and estimating the local detrended fluctuation function for a particular sub-period computed with a standard linear least-square fit, they averaged the functions over all sub-periods to get the overall fluctuation. The main advantage of working with fluctuations around the trend rather than a range of signals is that we can analyse non-stationary time series. Analysing the daily evolution of several currency exchange rates with the help of DFA, Vandewalle et al. [1997] found long-range power-law correlations in the data, and they also found that the exponent values and the range over which the power law holds varied drastically from one currency to the next. To probe the local nature of the correlations, they build a sliding window of length T which they moved along the historical data and reported a local exponent varying between 0.4 and 0.6 for some currencies. Similarly, analysing daily returns of the NYSE from 1966 to 1998 by using fluctuation analysis of the generalised cumulative absolute return, Pasquini et al. [2000] found that volatility exhibited power-laws on long time scales with a non-unique exponent. The scaling analysis they introduced proves the power-law behaviour and precisely determines the different coefficients. This is not the case for a direct analysis of the autocorrelations, since the data would show a wide spread compatible with different scaling hypothesis.

11.3.2.2 Applying sliding windows

We saw in Section (11.1.3) that the moment-scaling properties of financial returns were the object of a growing physics literature (see Galluccio et al. [1997], Vandewalle et al. [1998b], Pasquini et al. [2000]), confirming multiscaling in financial time series. Kantelhardt et al. [2002] proposed the multifractal detrended fluctuation analysis (MF-DFA) as a generalisation of the detrended fluctuation analysis (DFA), with the advantage that it can be used for non-stationary multifractal data. Later, measuring multifractality with either DFA, DMA, or wavelet analysis, and computing the local Hurst exponent on sliding windows, a large number of studies confirmed multifractality in stock market indices, commodities and FX markets, such as Matia et al. [2003], Kim et al. [2004], Matos et al. [2004], Norouzzadeh et al. [2005]. Further studies confirmed multifractality in stock market indices such as Zunino et al. [2007] [2008], Yuan et al. [2009], Wang et al. [2009], Barunik et al. [2012], Lye et al. [2012], Kristoufek et al. [2013], Niere [2013], to name but a few. Other studies confirmed multifractality on exchange rates such as Norouzzadeh et al. [2006], Wang et al. [2011b] Barunik et al. [2012], Oh et al. [2012], while some confirmed multifractality on interest rates such as Cajueiro et al. [2007], Lye et al. [2012], as well as on commodity such as Matia et al. [2003], Wang et al. [2011].

Carbone et al. [2004] first tested the feasibility and the accuracy of the dynamic detrending technique (DMA) on artificial series behaving as fBm with assigned Hurst exponent, and then they applied the technique to the log-return of German financial series, the DAX and the BOBL (tick by tick sampled every minute). They chose $T = 2^{20}$ for the artificial series with $H = \frac{1}{2}$, a sliding window with size $N_s = 5,000$ which is rolled $\delta_s = 100$ points, and they let λ varies from 10 to $1,000$ with step 2. The artificial series was characterised by a local variability of the correlation exponent weaker than those of the BOBL and DAX series. Considering the small fluctuations exhibited by the local scaling exponent $H(t)$ of the artificial series as the limits of accuracy of the method, the results showed evidence that the financial returns are a more complex dynamical system with higher standard deviation than the artificial series for about the same mean of Hurst exponent ($H \approx \frac{1}{2}$).

To properly characterise the efficiency of the Latin-American market indices, Zunino et al. [2007] considered two independent ranks, the Hurst and Tsallis parameters. The former was used as a measure of long-range dependence, while the latter was used to measure deviations from the Gaussian hypothesis. The dynamics of the Hurst exponent was obtained via a wavelet rolling sample approach (sliding window). While a small window around a particular time t must be considered, it must also be sufficiently large to contain enough points for a significant statistics. The authors chose a sliding window of size $N_s = 1,024$ (about 4 years) rolled $\delta_s = 16$ points to avoid redundant information. Further, to rank the different countries under study, they build the following measure of inefficiency

$$M_{Ineff} = \frac{|\hat{H} - \frac{1}{2}|}{\sigma_{\hat{H}}}$$

expressed as the distance of the estimated Hurst exponents from the critical value $\frac{1}{2}$ and normalised with the standard deviation. Thus, it can be used to asses whether returns or volatility returns possess LRD. If the score is larger than 2, they confirm the hypothesis of long-range dependence with a 95% confidence interval. They applied this measure to sliding windows on data ranging from January 1995 to February 2007 with about $3,169$ observations per country, and considered the percentage of significant windows as a meaningful indicator for assessing the relative inefficiency of stock markets. They concluded that long memory was categorically present in the volatility measures, but that there was little evidence of it in the returns. Nonetheless, considering the series as a whole, there is stronger long-range dependence in returns for emerging markets than for developed ones. Further, the US return was shown to have a clear antipersistent behaviour. To conclude, the efficiency of stock markets evolve in time, where periods of inefficiency alternate with periods of efficiency.

Zunino et al. [2008] used MFDFA to analyse the multifractality degree of a collection of developed and emerging stock market indices. Collecting daily returns from Bloomberg database for 32 different countries starting in 1995 and ending in 2007, the average multifractality degree Δh for developed markets were 0.336 ± 0.059 ($q \in [-20, 20]$) and 0.257 ± 0.054 ($q \in [-10, 10]$), while they were 0.459 ± 0.137 ($q \in [-20, 20]$) and 0.373 ± 0.131 ($q \in [-10, 10]$) for emerging markets. Further, the width of the multifractal spectrum of the original time series was shown to be a monotonically increasing function of the correlation exponent of the magnitude (volatility) time series. These results provide evidence that the multifractality degree for a broad range of stock markets isassociated with the stage of their development.

Using MFDFA method on the Shanghai stock price daily returns, Yuan et al. [2009] found two different types of sources for multifractality, the fat-tailed probability distributions and nonlinear temporal correlations. To analyse the local Hurst exponent, they considered a sliding window of 240 frequency data in 5 trading days, and found that when the price index fluctuates sharply, a strong variability characterise the generalised Hurst exponent $h(q)$. As a result, two measures were proposed to better understand the complex stock markets.

Gu et al. [2010] analysed the return time series of the Shanghai Stock Exchange Composite (SSEC) Index with the one-dimensional MFDMA model within the time period from January 2003 to April 2008, and confirmed that the series exhibits multifractal nature, not caused by fat-tailedness of the return distribution. They found the generalised

Hurst exponent $h(q)$ to have the values $h(-4) = 0.66 \pm 0.005$, $h(-2) = 0.624 \pm 0.002$, $h(0) = 0.591 \pm 0.001$, $h(2) = 0.53 \pm 0.003$, and $h(4) = 0.493 \pm 0.008$. With $h(2)$ close to $\frac{1}{2}$, and ignoring the nonlinearity in $h(q)$, one could infer that the return time series of the SSEC was almost uncorrelated. However, the span of the multifractal singularity strength function (see Equation (11.2.25)) for the SSEC time series was $\Delta\alpha = 0.72 - 0.39 = 0.33$ indicating that the series has multifractal nature. Shuffling the return time series and repeating the analysis, the singularity spectrum shrunk, implying that the fat-tailedness of the returns played a minor role in the multifractality of the series which was mainly driven by correlations.

Lye et al. [2012] used MFDFA coupled with the rolling window approach to scrutinise the dynamics of weak form efficiency of Malaysian sectoral stock market, and showed that it was adversly affected by both Asian and global financial crises, and also negatively impacted by the capital control implemented by the Malaysian government during the Asian financial crisis. They considered daily closing price indices from a period ranging from 1/11/1993 to 30/06/2011, with a total of $4,609$ observations. Comparing results with the binomial multifractal model, they found multifractality due to both fat-tailed probability distribution and long-range correlations. Since the JB test showed that the local Hurst exponents were not normally distributed, instead of using the mean to rank the sectorial indices, they considered the median, and defined a new measure

$$|\text{ Median} - \frac{1}{2}|$$

to provide a better view of the ranking. Using this measure on the data, the Hurst exponents showed significant deviations way from $H = \frac{1}{2}$ during the Asian and global financial crises, September 11 attacks, dot-com bubble, as well as when the Malaysian Ringgit was pegged to the US dollar, which was on line with the results found by Chin [2008b].

Applying MFDFA to the daily time series data of six composite stock price indices in the Association of Southeast Asian Nation (ASEAN) region covering the period from January 2006 to June 2013, Niere [2013] provided empirical evidence of the presence of multifractality in these time series. To capture the dynamics of local Hurst exponents, a sliding window approach was used in applying MFDFA, with a cubic polynomial, and scale varying from 20 to $\frac{N}{4}$ with a step of four. The window length was 240 days (around 1 year) with a shift between windows equal to 5 trading days. He obtained a spectrum of Hurst exponents in all six indices, manifesting the presence of multifractality. Based on the efficiency measure

$$M_{eff} = \frac{|\overline{H} - \frac{1}{2}|}{\sigma_{\hat{H}}}$$

where $\overline{H}$ is the mean of $h(2)$, an efficiency ranking of stock markets of the six countries under study was provided to guide investors seeking profit opportunity.

Stosic et al. [2014] analysed the transition from managed to independent floating currency rates in eight countries with MFDFA, and found changes in multifractal spectrum indicating an increase in market efficiency. The observed changes were more pronounced for developed countries having an established trading market. Applying the MFDFA method with a second degree polynomial to analyse logarithmic returns of daily closing exchange rates, they obtained parameters $(\alpha_0, \Delta\alpha, r)$ describing multifractal spectrum.

11.4 Local Holder exponent estimation methods

11.4.1 The wavelet analysis

While the partition function-based methodology, described in Section (11.1.2), provides only global estimates of scaling ($\tau(q)$, $h(q)$, or $D(h)$), there are cases when local information about scaling provides more relevant information

than the global spectrum. It generally happens for time series where the scaling properties are non-stationary, due to intrinsic changes in the signal scaling characteristics, or even boundary effects. Until recently, the estimation of both local singularity and their spectra were very unstable and prone to gross numerical errors (see details in Section (11.2.2.1)). However, using the wavelet transform multiscale decomposition, Struzik [1999] proposed stable procedures for both the local exponent and its global spectrum.

11.4.1.1 The effective Holder exponent

Struzik [1999] [2000] started from the fact that for cusp singularities, the location of the singularity could be detected, and the related exponent recovered from the scaling of the wavelet transform (WT), along the maxima line converging towards the singularity. More precisely, the Holder exponent of cusp singularity can be estimated from the slope of the maxima lines approaching isolated singularities (see Section (11.2.2)). In addition, we saw that the scaling of the maxima lines was stable in the log-log plot only below some critical scale. However, it is not possible to estimate the slope of the plot other than in the case of isolated singular structures from the WT. This is because in the case of densely packed singularities, the logaritmic rate of increase, or decay, of the corresponding WT maximum line fluctuates wildly. Struzik [1999] proposed an approach to circumvent this problem, while retaining local information, consisting in modelling the singularities as created in some kind of a collective process of a very generic class. He chose to bound the local Holder exponent by explicitly calculating bounds for the slope locally in scale, and then discard (or ignor) the parts of the maxima lines for which the slope exceeds the imposed bounds (see Struzi [1998]). As an example, he used $|\tilde{h}| < 2$ bound on the local slope $\tilde{h}$ of each maximum. Doing so, he obtained the set of non-diverging values of the maxima lines corresponding to the singularities in the time series. However, even with the diverging parts of the maxima removed, it was still not possible to obtain local estimates of the scaling behaviour other than for isolated singular structures, since the local slope of WT maxima was strongly fluctuating. Nonetheless, the fluctuations carry very relevant information to the scaling properties of the process underlying the origin of the input time series. Hence, Struzik defined the local effective Holder exponent by considering multiplicative cascade model where each point of the cascade is uniquely characterised by a sequence of weights $(a_1, .., a_n)$ taking values from the binary set $\{1, 2\}$, and acting successively along a unique process branch leading to this point. Doing so, he obtained a model-based approximation of the local scaling exponent. Denoting $\kappa(F_i)$ the density of the cascade at the generation level F_i, for $i \in [0, max]$, then it can be written as

$$\kappa(F_{max}) = p_{a_1}, ..., p_{a_n}\kappa(F_0) = P_{F_0}^{F_{max}}\kappa(F_0)$$

such that the local exponent is related to the product $P_{F_0}^{F_{max}}$ of these weights

$$h_{F_0}^{F_{max}} = \frac{\log P_{F_0}^{F_{max}}}{\log(\frac{1}{2})^{max} - \log(\frac{1}{2})^0}$$

Since the weights pi are not known and h_i must be estimated, in multiplicative cascade model, the effective product of the weighting factors is reflected in the difference of logarithmic values of the densities at F_0 and F_{max} along the process branch

$$h_{F_0}^{F_{max}} = \frac{\log\kappa(F_{max}) - \log\kappa(F_0)}{\log(\frac{1}{2})^{max} - \log(\frac{1}{2})^0}$$

The densities $\kappa(F_i)$ can be estimated from the value of the wavelet transform along the maxima lines corresponding to the given process branch. Hence, the estimate of the effective Holder exponent becomes

$$h_{a_{lo}}^{a_{hi}} = \frac{\log T(a_{lo}, w_{pb}) - \log T(a_{hi}, w_{pb})}{\log a_{lo} - \log a_{hi}}$$

where $T(a, w_{pb})$ is the value of the wavelet transform at the scale a, along the maximum line w_{pb} corresponding to the given process branch. The scale a_{lo} corresponds to the generation of F_{max}, while the scale a_{hi} corresponds to that of F_0. In general one would choose

$$a_{hi} \equiv a_{SL} = \log(\text{ Sample Length})$$

to estimate $T(a_{hi}, w_{pb})$, but the wavelet transform coefficients at that scale are heavily distorted by finite size effects. One solution is to estimate the value of $T(a_{hi}, w_{pb})$ with the mean h exponent. For multiplicative cascade, the mean value of the cascade at scale a is given by

$$M(a) = \frac{M(a,1)}{M(a,0)}$$

where $M(a,q)$ is the partition function of the qth moment. We can then estimate the mean value of the local Holder exponent as a linear fit

$$logM(a) = \overline{h}\log(a) + c$$

In order for the Holder exponent to be the local version of the Hurst exponent, we can take the second moment in the partition function to define the mean $\overline{h'}$

$$M^{'}(a) = \sqrt{\frac{M(a,2)}{M(a,0)}}$$

and use the linear fit above, replacing M with $M^{'}$. The estimate of the local Holder exponent $\hat{h}(x_0, a)$ becomes

$$\hat{h}_{a_{lo}}^{a_{SL}} \approx \frac{\log T(a_{lo}, w_{pb}) - (\overline{h'}\log a + c)}{\log a_{lo} - \log a_{SL}} \tag{11.4.26}$$

11.4.1.2 Gradient modulus wavelet projection

The gradient modulus wavelet projection (GMWP), introduced by Turiel et al. [2006], defines a scale-dependent measure with the help of a continuous wavelet transformation. We let N_s be a small segment size, and consider a floating time interval (or sliding window) defined for the range $[t_{ind} - \frac{N_s}{2}, t_{ind} + \frac{N_s}{2}]$ around the time index t_{ind}. The small segment size N_s is assumed to be odd to align the centre of the segments according to a time index. In that setting, the range of the time index satisfies $[\frac{N_s}{2}, N - \frac{N_s}{2}]$, and the local fluctuation for a segment is given by

$$F_{GMWP}(N_s) = \sum_{t_{ind}=\frac{N_s}{2}}^{N-\frac{N_s}{2}} \mu_{N_s}(t_{ind})$$

where the scale-dependent measure is defined as

$$\mu_{N_s}(t_{ind}) = \sum_{i=0}^{N_s} |x_{i,t_{ind}}| \psi_{N_s}(\frac{N_s}{2} - i)$$

and $x_{i,t_{ind}} = x(t_{ind} - \frac{N_s}{2} + i)$. The continuous wavelet transformation is the convolution of the response time series and a waveform $\psi_{N_s}(\bullet)$ scaled to the floating time interval $[-\frac{N_s}{2}, \frac{N_s}{2}]$. Since the detrending of the response series is dependent on the shape of the waveform, the authors used the Laplacian wavelet $\psi_{s,t} = (1 + \frac{t^2}{s})^{-2}$ because it performs well on the smallest scales (see Turiel et al. [2003]).

11.4.1.3 Testing the performances of wavelet multifractal methods

Multifractality is a property verified in the infinitesimal limit only, while empirical data have an inherent discrete nature. As a result, any tools designed to validate the multifractal character of a given signal faces several difficulties linked to the finite size and the discretisation of the data. Moreover, any technique used to validate the multifractal behaviour of a signal necessarily involves some interpolation scheme introducing some bias. As a result, one should know the range of validity, limitations and biases, and the theoretical foundation of any method to determine the degree of reliability of the estimates obtained. Turiel et al. [2006] studied the theoretical foundations, performance and reliability of four different validation techniques for the analysis of multifractality from experimental data. They considered the moment (M) method, the wavelet transform modulus maxima (WTMM) method, the gradient modulus wavelet projection (GMWP) method and the gradient histogram (GH) method, which were used to estimate the singularity spectra of multifractal signals. They showed that all the methods always gave better estimates of the left part of the spectra than of the right part, and as the spectrum width increased the quality on its determination decreased. The main conclusion was that GMWP method got the best overall performance, providing reliable estimates which can be improved with increasing statistics. All the other methods were affected by problems such as the linearisation of the right tail of the spectrum.

11.4.2 The fluctuation analysis

11.4.2.1 Local detrended fluctuation analysis

We saw in Section (11.3) that some researchers tried to estimate the local Hurst exponent at time t as the Hurst exponent computed on a sliding window of size n from $t-n+1$ till t. However, all methods estimating the Hurst exponent work well only in a large size window, and are therefore not capable of detecting local jumps. Following Struzik [2000], who combined a multiplicative cascade model with the wavelet transform modulus maxima (WTMM) method to estimate the local Hurst exponent, Ihlen [2012] detailed a method that retrieve the local Hurst exponent from a small window size using detrended fluctuation analysis (DFA). That is, rather than using the WTMM tree for defining the partition function-based multifractal formalism, he considered the DFA model, obtaining a good estimate of the local Hurst exponent. Following Ihlen [2012], we are now going to explain how to estimate the local Hurst exponent. As opposed to the time independent Hurst exponent estimated by the monofractal DFA, the local Hurst exponent, $H(t)$, estimated for a multifractal time series will strongly fluctuate in time. In order to estimate the local Hurst exponent, $H(t)$, the local mean-square (MS) fluctuation $F^2_{DFA,2}(a)$ given in Equation (11.2.20) has to be defined within a translating segment centred at sample (or time index), $a = t_{ind}$, instead of within non-overlapping segments. That is, given the small segment size, N_s, the sliding window is defined for the range $[t_{ind} - \frac{N_s}{2}, t_{ind} + \frac{N_s}{2}]$. Ihlen [2012] considered the vector of small segment sizes $[7, 9, 11, 13, 15, 17]$ where the segment sizes increases with two samples in order to align the centre of segments according to a time index. Taking the largest segment size, $N_{s,max}$, the time index t_{ind} is defined for the range $[\frac{N_{s,max}}{2}, N - \frac{N_{s,max}}{2}]$. For each small segment size, N_s, we compute the local fluctuations for a translating segment centred at time index t_{ind}, getting

$$F_{DFA,2}(N_s) = \sum_{t_{ind}=\frac{N_{s,max}}{2}}^{N-\frac{N_{s,max}}{2}} F^2_{DFA,1}(t_{ind})$$

where

$$F^2_{DFA,q}(t_{ind}) = \mu^q_{N_s}(t_{ind}) = \big(\frac{1}{N_s}\sum_{i=0}^{N_s} Y^2(i, t_{ind})\big)^{\frac{q}{2}}$$

with

$$Y(i, t_{ind}) = X(i, t_{ind}) - X_{N_s,l}(i, t_{ind})$$

and $X(i,t_{ind}) = X(t_{ind} - \frac{N_s}{2} + i)$. We already saw that the difference between the overall qth order RMS, $F^2_{DFA,q}(n)$, computed with Equation (11.2.22) for q in the range $[-5,5]$, converge toward each other with increasing scale. It forms an envelope with a maximum and minimum variations. The regression line for $q = 0$ is the centre of the spread of RMS in log-coordinates. The same convergence exists for the local RMS, of small segment size, which is used to estimate the local Hurst exponents (linear regression). It is estimated as the slope of the line from local RMS in log-coordinates to the endpoint of the regression at the largest scale N. That is, we fit the line $R_{L,q}$ given by Equation (11.2.22) for $q = 0$ on large scales, getting

$$R_{L,0}(n) = \log F_{DFA,0}(n) \approx \log c + h(0) \log n$$

where $h(0) = h(q)|_{q=0}$. We then extrapolate the fitted line on the small segment sizes, and compute the residual from that line and the local fluctuations, $F^2_{DFA,1}(t_{ind}) = \mu_{N_s}(t_{ind})$, for all time index t_{ind}

$$Res(N_s, t_{ind}) = R_{L,0}(N_s) - \log F^2_{DFA,1}(t_{ind}) \, , t_{ind} = \frac{N_{s,max}}{2}, ..., N - \frac{N_{s,max}}{2}$$

Defining the log-scale for the segment size N_s as $L(N_s) = \log(N) - \log(N_s)$, the coarse-grained singularity exponent, $H_{N_s}(t)$, is obtained by dividing the residuals by the log-scale and adding the slope $h(q)|_{q=0}$ of the regression line

$$H_{N_s}(t_{ind}) = \frac{Res(N_s, t_{ind})}{L(N_s)} + h(q)|_{q=0}$$

Hence, the local Hurst exponent $H(t)$ is estimated from the coarse-grained exponent $H_{N_s}(t)$ in the limit $N_s \to 0$. As explained by Ihlen et al. [2014], the local DFA is conducted in the same way as the conventional DFA with three important differences

1. the RMS of the detrended residuals is computed in a floating time interval across the time series instead of in non-overlapping time intervals as in DFA.
2. the obtained RMS measure $F^2_{DFA,1}(t_{ind}) = \mu_{N_s}(t_{ind})$ is dependent on both time and scale, in contrast to the RMS measure $F^2_{DFA,1}(a)$ of the DFA which only depend on scale.
3. the local DFA scaling exponent $H_{N_s}(t_{ind})$ is numerically defined by the linear slope of $\log F^2_{DFA,1}(t_{ind})$ versus $\log N_s$ for each time instant t_{ind}, instead of the time-independent $\log F_{DFA}(n)$ versus $log(n)$ for the DFA.

Note, the coarse-grained singularity exponent, $H_{N_s}(t)$, can be rewritten as

$$H_{N_s}(t_{ind}) = \frac{Res(N, t_{ind})}{L(N_s)}$$

where N is the length of the time series.

***Remark* 11.4.1** *This expression for the estimated local Holder exponent is to be related to the effective Holder exponent in Equation (11.4.26).*

The multifractal spectrum D_h is defined by the normalised distribution P_h of $H(t_0)$ in log-coordinates

$$D_h = \lim_{\epsilon \to 0} (1 - \frac{1}{\log \epsilon} \log (\frac{P_h}{P_{h,max}})) \tag{11.4.27}$$

where ϵ is the bin size of the histogram used to define P_h, and $P_{h,max}$ is the maximum probability at the mode $h(0)$ of $H(t_0)$ (see Turiel et al. [2008]).

The small $H(t)$ in the periods of the multifractal time series with local fluctuations of large magnitudes reflects the noise like structure of the local fluctuations, while the larger $H(t)$ in the periods with local fluctuations of small

magnitudes reflects the random walk like structure of the local fluctuations. The local Hurst exponent $H(t)$ in periods with fluctuations of small and large magnitudes is therefore consistent with the qth order Hurst exponent $h(q)$ for negative and positive q's, respectively. The advantage of the local Hurst exponent, $H(t)$, compared with the qth order Hurst exponent, $h(q)$, is its ability to identify the instant in time of a structural changes within the time series.

11.4.2.2 The multifractal spectrum and the local Hurst exponent

We have previously discussed how to transform the q-order Hurst exponent $h(q)$ to the mass exponent $\tau(q)$, and finally to the multifractal spectrum $f(\alpha)$ with its associated singularity α. Following Ihlen [2012], we are now going to explain how to estimate directly the multifractal spectrum from the local fluctuations. The temporal variation of the local Hurst exponent can be summarised in a probability distribution, and the multifractal spectrum is just the normalised probability distribution in log-coordinates. Thus, the width and shape of the multifractal spectrum reflect the temporal variation of the local Hurst exponent (the temporal variation in the local scale invariant structure of the time series). More precisely, the temporal variations of the local Hurst exponent $H(t)$ can be summarised in a histogram representing the probability distribution P_h of $H(t)$. The multifractal spectrum $f(\alpha)$ is simply defined by the log-transformation of the normalised probability distribution $\tilde{P}_h$. The probability distribution P_h are computed by dividing the number of local Hurst exponents in each bin by the total number of local Hurst exponents. The multifractal spectrum $f(\alpha)$ are therefore directly related to the distribution P_h of the local fractal structure of the time series via Equation (11.4.27). As an example, we plot the multifractal spectrum and its estimation computed via the local Hurst in Figure (11.3). The probability distribution obtained this way is relatively accurate as it is fairly similar to the multifractal spectrum. However, we can see that the distribution is slightly biased to the right for the monofractal series with $H = 0.3$ and to the left for the series with $H = 0.7$. Note, these differences are minor and the main properties of the distribution are conserved.

11.4.3 Detection and localisation of outliers

In general, outlier difference itself from noise through its inherently isolated and local character, leading to non-stationarity and highly erratic behaviour. Hence, one must determine whether a particular extreme observation belongs to the bulk of the data or should be treated as an outlier. To do so, we need a methodology capable of determining the statistical nature of the non-stationarity process both in a global and local sense. Struzik et al. [2002] used the effective local Holder analysis, described in Section (11.4.26), not only to detect the outliers, but also to localise them in time and space. Computing the singularity $h(q)$ (or $\alpha(q)$) and the spectrum $D(h)$ (or $f(\alpha)$), he found that the spectrum for the clean version was narrow and focused around the mean value of the singularity strength h_{mean}, but that it was very broad in the dirty version, and gradually falling off to zero dimension values for decreasing $h < h_{mean}$. This phenomena corresponds with positive q values having the ability to select exponents of a relatively lower value than the h_{mean} value. Note, the way to distinguish multifractal processes from outliers is through the relative values of $D(h)$ and the spacing between the q values. The former requires dense coverage of $h(q)$ values (dimension 1 on a line), while point-wise events have support 0, so that for $D(h)$ near 0 we have weak support and strong probability of an outlier. Hence, comparing the values of $h(q)$ and $D(h)$ for positive q can allow us to detect the presence of spikes in the time series. To localise the spikes we need the local value of $h(x)$ instead of the global average, which can be estimated with the help of the effective local Holder exponent $\hat{h}(x, a)$. Plotting the local exponents, the value cluster around h_{mean}, but strong singular events drop well below the mean value, such that selecting an appropriate threshold we can filter the outliers from the series. Struzik et al. chose as a threshold the mean value of the local Holder exponent $\overline{h'}$ described in Section (11.4.26). This is because the micro-canonical geometric mean $\overline{h'}$ does not coincide with the h_{mean} mode value of the $D(h)$ distribution in presence of outliers. That is, the probability of outliers is high when the h_{mean} differs significantly from $\overline{h'}$.

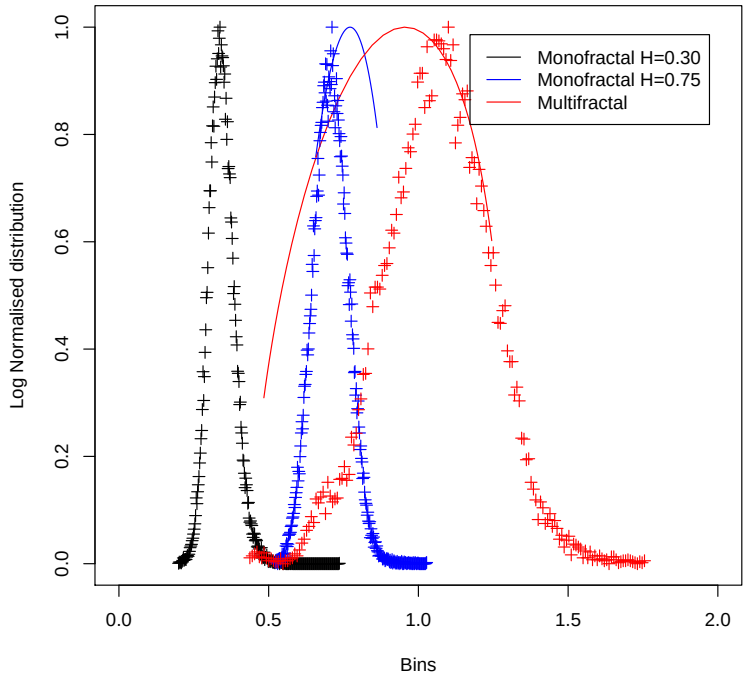


Figure 11.3: Multifractal estimation and its estimation via local Hurst estimates at scale 7 over a samples of size 8192 for series with different fractal structures.

11.4.4 Testing for the validity of the local Hurst exponent

We are now going to test for the validity of the local Hurst exponent observed at time t and computed as the Hurst exponent on a sliding window against the one computed with the local detrended analysis (LDA) described in Section (11.4.2.1). In the following, we will call the former "Local" Hurst exponent with double quotes and the latter Local Hurst exponent without any quote.

11.4.4.1 Local change of fractal structure

In order to perform the test, we construct a fractional Brownian motion (fBm) of size $8,192$ with Hurst exponent $H = \frac{1}{2}$ and we introduce two local changes of size 25 at time $4,096$ and time $6,120$ with Hurst exponent $H = 0.8$ and $H = 0.2$, respectively. The "Local" Hurst exponent is a wavelet-based Hurst estimate on a sliding window of size $1,024$ and the other one is estimated using the LDA algorithm described above. The results are given in Figure (11.4). The "Local" Hurst exponent computed on the sliding window has Hurst exponent comprised between 0.4 and 0.8 without any noticeable change in its structure, whereas the local Hurst exponent computed with the LDA evolves around 0.5 and shows two singular points. The first one peaks at 0.8 around the time $4,096$, and the second one reaches 0.2 approximately at time $6,120$. These two singularities observed in the series are in perfect accordance with the artificial changes we made to the data.

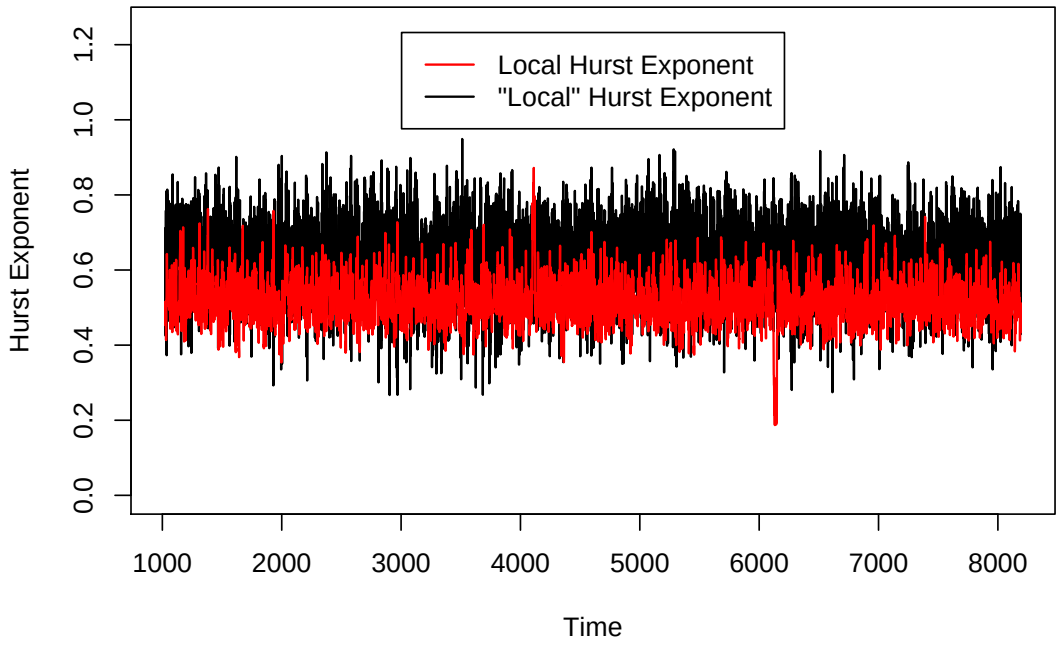


Figure 11.4: Hurst exponents computed on an artificial series of 8192 points containing abrupt local change in structure at time 4096 and 6120.

11.4.4.2 Abrupt change of fractal structure

We now construct a time series using two fractional Brownian motions with different Hurst exponent that we combine together. The first part of the constructed series is a fractional Brownian motion of size $4,096$ with Hurst exponent $H = 0.4$, and the second part of the series is a fractional Brownian motion of size $4,096$ with Hurst exponent $H = 0.7$. As a whole, the final series has a Hurst exponent $H = 0.4$ from time (point) 1 to $4,096$ and an exponent $H = 0.7$ from time (point) $4,097$ to $8,1924$. Similarly to the previous experiment, we compute the Local Hurst exponent with the two methods described above. The first one is wavelet-based Hurst estimates on a moving window of size $1,024$, and the second one uses the algorithm described just above. The results are given in Figure (11.5) below.

It can immediately be seen that the local Hurst exponent computed with the multifractal method shows an important change in values, going from 0.4 to 0.8 at around time $4,096$. It corresponds exactly to the time where we changed the fractal structure of the series. On the other hand, the "Local" Hurst exponent starts to have slight changes at time $5,000$, which is approximately equal to $4,096$ (the exact location of change) plus $1,024$ (sliding window size). In other words, the "Local" Hurst exponent takes the change into account only when it considers a series with a major part containing data of $H = 0.8$ in its sliding window.

11.4.4.3 A simple explanation

The inability for the "Local" Hurst exponent computed on a sliding window to detect large local changes of fractal structure may be easily explained. Assuming that the Hurst exponent of a sample time series is approximately equal

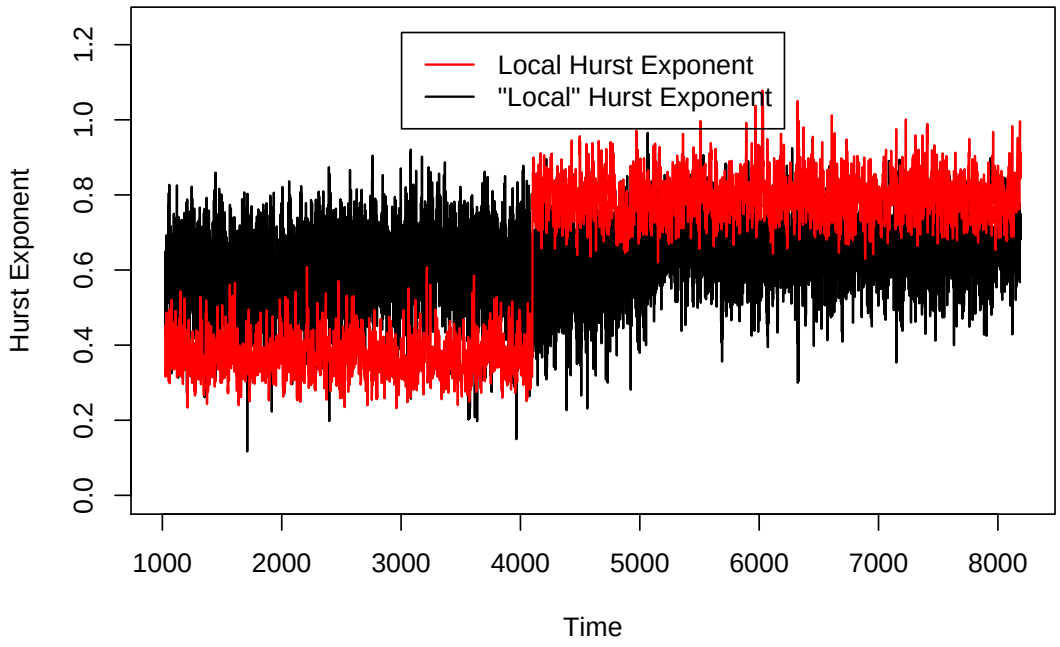


Figure 11.5: Hurst exponents computed on an artificial series of 8192 containing abrupt change in structure at time 4096 with $H = 0.4$ before time 4096 and $H = 0.8$ after time 4096.

to the average of the local Hurst exponent on the sample (see details in Section (11.3.1.3)), noting L the length of the time interval, then the Hurst exponent at time t satisfies

$$H_L(t) = \frac{1}{L}\int_{t-L}^{t} H_t dt$$

Thus, for a local change (pulse) happening on a time interval much smaller than L, the pulse does not have enough weight when averaged to have a significant impact. As a result, even for a large change of fractal structure, the local Hurst exponent computed on a sliding window is not an appropriate estimate of the true local Hurst of a time series. On the other hand, the local Hurst exponent computed with the LDA is perfectly capable of capturing the local structure of the time series. Consequently, the most appropriate estimate of local Hurst exponent is the one introduced by Ihlen [2012], as it can precisely capture abrupt changes of fractal structure in a time series without lagging.

11.5 Analysing the multifractal markets

11.5.1 Describing the method

In order to analyse the markets for multifractality, we consider the S&P 500 Index daily prices provided by Bloomberg from April 4th, 1928 to September 5th, 2014, and compute the local Hurst exponent by using the two different techniques described earlier.

1. We estimate the pseudo local Hurst exponent on a sliding window of $2,048$ days computed using the DFA method as detailed in Section (11.3.1).

2. We compute the proper, or effective, local Hurst exponent following the method described in Section (11.4.2.1).

We saw in Section (11.1.2.3) that the local Hurst exponent $H(t)$ describing multifractal processes was itself a random process. Since it is now well known that financial data have a multifractal nature, we expect the local scaling to change erratically and randomly in time. For presentation purposes, Costa et al. [2003] smoothed out the original curve $H(t)$ by performing a 20-day moving average procedure. Similarly, in order to concentrate on the main characteristics of the estimated local Hurst exponent, we choose to denoise it with the wavelet denoising technique, and only plot the resulting trend. The estimated local Hurst being highly fluctuating, we can therefore gain in readability and focus on its dynamic properties.

The results in Figure (11.6) show an increase in value of the S&P Index from 1928 till 2014, with increasing price fluctuations from the early 1980s onward. As a result, we would expect an important drop in the local Hurst exponent starting from these years. However, even though the pseudo local Hurst exponent does decrease from 1997, going from 0.55 to 0.45, the change in value is clearly not significant. Further, the local Hurst exponent quickly goes back to values around 0.5, suggesting that returns do not have long-memory.

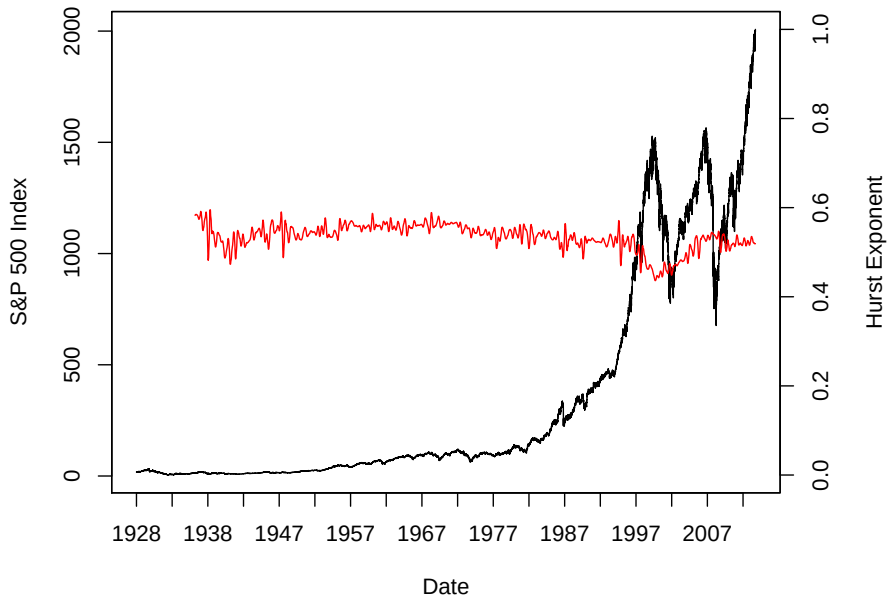


Figure 11.6: S&P 500 Index from April 4th, 1928 to September 5th, 2014 (black line) and its corresponding denoised Hurst exponent by DFA method on a sliding window of 2048 days (red line).

We now compute the proper local Hurst exponent and show the results in Figure (11.7). In this setting, we observe that the Hurst exponent drops substantially from $0.7 - 0.8$ (low fluctuations) in $1928 - 1950$ to 0.2 (high fluctuations) from 1997. As a result, we can infer that the S&P 500 Index had a strong persistent behaviour in the first period, and a strong anti-persistent behaviour in the more recent years. This is therefore a clear evidence of long-range dependence in data series of stock returns. This is on line with the results found in Section (11.4.4) showing the inability of the pseudo local Hurst to detect large changes of fractal structure in financial time series as opposed to the proper local

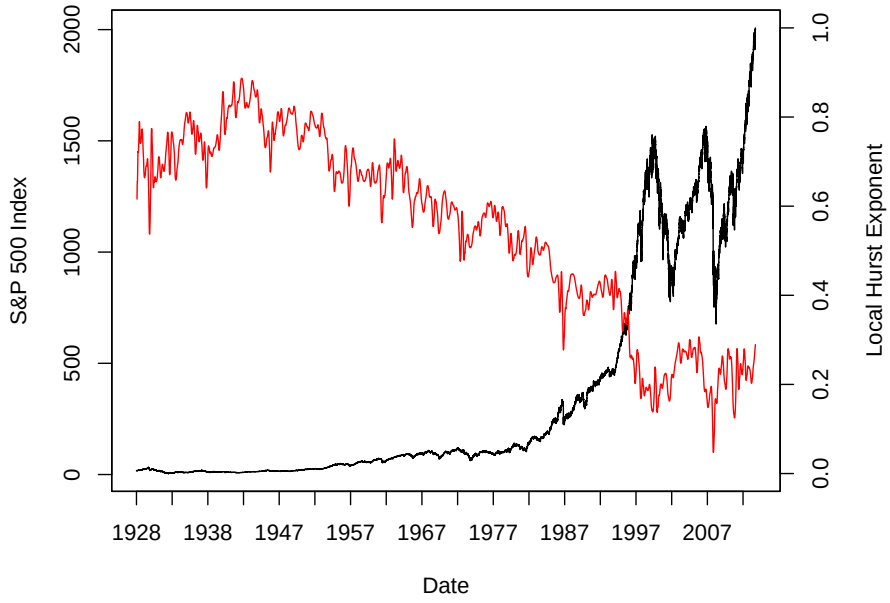


Figure 11.7: S&P 500 Index from April 4th, 1928 to September 5th, 2014 (black line) and its corresponding denoised Local Hurst exponent by MFDFA method (red line).

Hurst technique. As a result, we will only consider the second approach when detecting multifractality in financial markets.

11.5.2 Testing for trend and mean-reversion

11.5.2.1 The equity market

We are now going to consider the price of "Google US" from August 22nd, 2012 to September 5th, 2014, and apply the same techniques as above to detect trends and mean reverting behaviours on a close-to-close strategy (we trade only at the close of the market everyday).

The Figure (11.8) represents Google prices from August 22nd, 2014 to September 5th, 2014. We have also plotted its trend computed using a wavelet denoising method. We observe that the stock prices and the trend coincides in most of the period, except in few time intervals, where it is strongly pushed away and then pulled back. For example, the prices fluctuates around the trend just before 2013, in January 2014 and more significantly between March and May 2014. This phenomena suggest to adapt a trend-following strategy when the price process coincides with the local trend, and on the contrary, we want to apply a mean reverting strategy when the price fluctuates around the trend. To do so, we need a indicator that allows us to take a decision systematically. We observe that the local Hurst exponent coincides exactly with the time interval where we observed strong fluctuations around the trend of the price process. Indeed, just before 2013, we observe a local minimum on the local Hurst exponent corresponding exactly with the fluctuations. The same observations may be made in January 2014. More significantly, between February and May 2014, the local Hurst exponent dropped from 0.55 to 0.35, and again, this phenomena exactly corresponds to the huge fluctuations interval observed in the price process. In addition, the global minimum of the local Hurst exponent occurs exactly when the price jumped from 450 to 500 in August 2013. In that period of time, the price is indeed highly mean reverting, since the difference between the trend (blue) and the actual process (black) is high. Hence, we can infer

with a high probability that the process is more likely to come back to its trend. The autocorrelation is thus highly negative and the local Hurst exponent is subsequently close to zero.

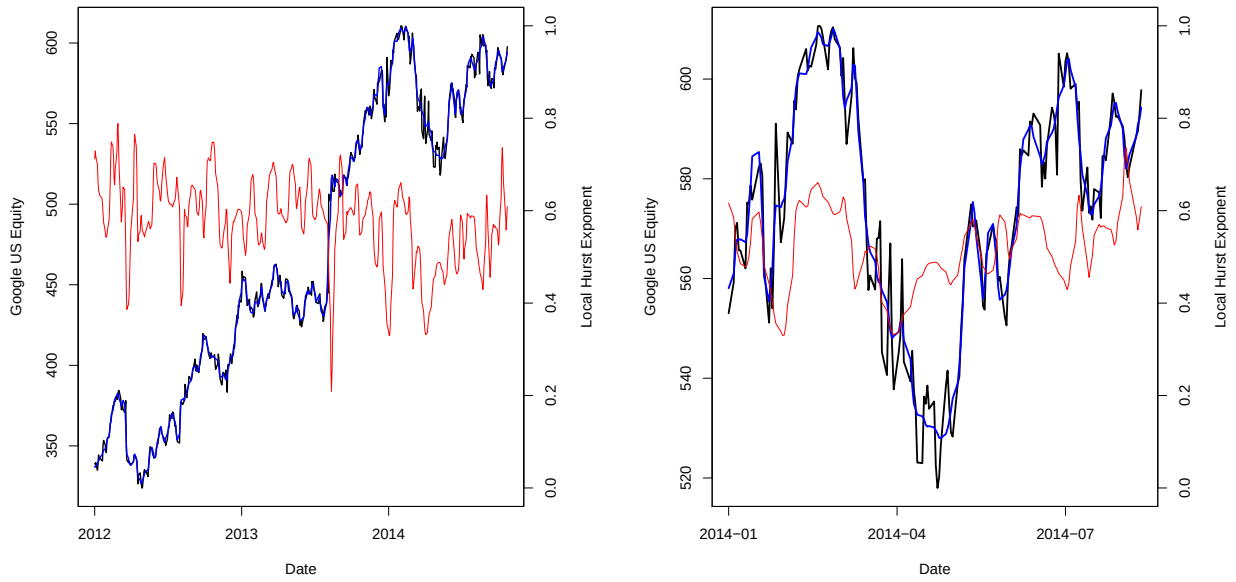


Figure 11.8: Google Equity Price from August 22nd, 2012 to September 5th, 2014 (black line) its trend (blue line) and its corresponding denoised Local Hurst exponent by MFDFA method (red line) and a zoom between January 2014 and August 2014

11.5.2.2 The FX market

The high liquidity in the FX market allows us to define trading strategies at a high-frequency. As illustrated in the figure below, the data is very noisy on this time scale, making it highly difficult to trade directly on the currency pair. One solution could be to use wavelet analysis to extract the trend of the currency pair for different time scales and then build a strategy considering those market imperfections on each time scaling.

Figure (11.9) represents the EUR / GBP currency pair from August 31st to September 5th. We plotted its trend showing its moves of scale 4 (which represents 4 minutes) and its trend for the moves of scale 2 (which depicts 1 minute). While we observe that the two trends are roughly equal in most of the period, we can also see that the red line sometimes slightly fluctuates around the blue line. This is the case on the September 1st at around 10:00, the September 2nd at around 10:00, the September 3rd at around 10:00 and on the September 4th at around 14:00. In those periods, when the red line is below the blue line, the red line is most likely going to increase in the next ticks, and conversely, when the red line is above the blue line, it is most likely going to decrease in the next ticks. This phenomena exactly corresponds to a mean-reverting behaviour of the red process around its local mean, modelled by a longer trend (blue process). Further, this behaviour also coincides exactly with a drop in the local Hurst exponent. Therefore, we can propose a simple trading strategy:

1. When the local Hurst exponent is under a fixed threshold (typically 0.3 here), we compute the two trends (blue and red) of the EUR/GBP currency pair and we check the relative position of the blue line on the red line. If the red line is under the blue line, we buy Euros since its value will most likely increase. On the contrary, if the red line is above the blue line, we short sell Euros since its value will most likely decrease.

2. When the local Hurst exponent is over the fixed threshold, we buy and hold Euros if the trend is upward. On the contrary, we sell Euros if the trend is downward.

Obviously, to get an exact idea of the viability of this strategy in practice, many other factors have to be taken into account, such as taxes, transaction costs, technological means and the total traded amount, since the moves in the price are very low in this frequency (around 1/100 euros per tick).

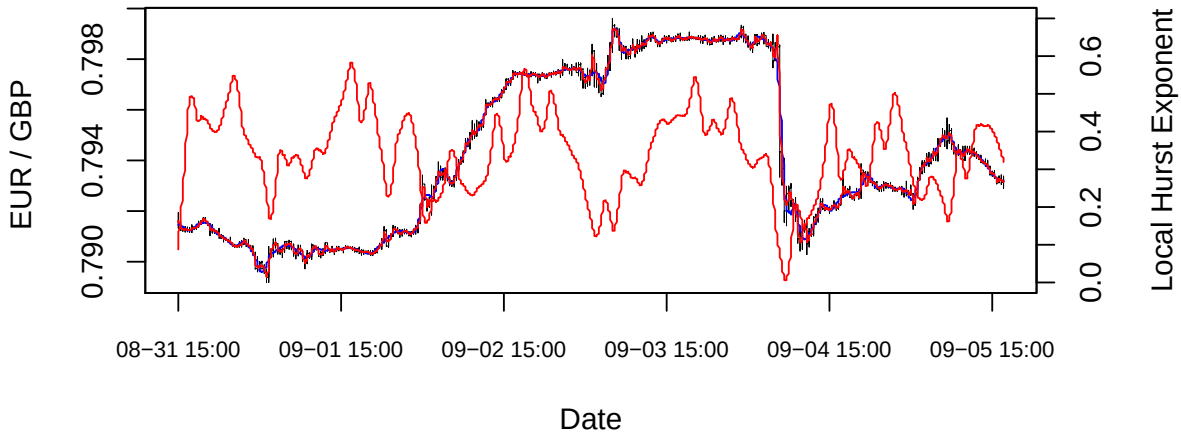


Figure 11.9: Tick-by-tick (20sec) EUR / GBP currency pair from August 31st, 2014 to September 5th, 2014 (grey line), its trend on different scale (red and blue lines) and its corresponding denoised Local Hurst exponent by MFDFA method (thin red line).

11.5.3 Testing for crash prediction

Since around 2003, many articles have been published in the literature on crash prediction in finance using the "local" (pseudo) Hurst exponent. For instance, Grech et al. [2004] studied the "local" Hurst exponent in the Dow Jones Index and showed significant predictability in the 1929 and the 1987 crashes. Czarnecki et al. [2008] obtained the same results in the Warsaw Stock Exchange. Then Xu et al. [2009] also applied the technique to predict sudden drop in the Shanghai Stock Exchange. In this section, we conduct our own investigation in order to predict some of the recent crisis by using the local Hurst exponent defined by Ihlen.

11.5.3.1 The Asian crisis in 1997

The Asian crisis hit East Asia from July 1997, leading to a financial contagion in many emerging countries, such as Argentina and Brasil. It began from the collapse of the Thai Baht (THB) in Thailand, and propagated to all the other countries in East Asia. Indonesia, South Korea, Thailand, Hong Kong, Malaysia, Laos and Philippines were the most affected by the crisis. Japan was also severely hitten as shown by the Nikkei 225 Index dropping from $19,943$ points to $17,019$ (-12.47%) in 1 week from January 7th to 13th.

In Figure (11.10), we plot the THB/USD currency pair from December 14th, 1995 to September 8th, 1999. Before May 1997, we observe that the THB/USD FX rate was slightly increasing with little fluctuations, since the local Hurst exponent is around 1, indicating a very high level of persistence. However, after June of 1997, the THB value has become incredibly low. More precisely, before June 1997, 1 USD could be exchanged for 25 THB, but around February

1998, 1 USD was exchanged for 55 THB. In a period of time of less than a year, the Thai Baht has lost half of his value.

Could it be predicted by using the local Hurst exponent? Analysing the local Hurst exponent, we observe that it was fluctuating around 1 before 1997, indicating that the THB/USD FX rate grew nearly as a pure trend. It suddenly dropped at the beginning of 1997, just before the huge loss of value that occurred during 1997 and 1998. The local Hurst exponent has immediately detected abnormal fluctuations in the FX rate, implying an abrupt change in the fractal structure of the time series. In other word, it has locally detected a huge amount of fluctuations compared to the previous pure trend regime of the Thai Baht, indicating the beginning of a monetary crisis in Thailand.

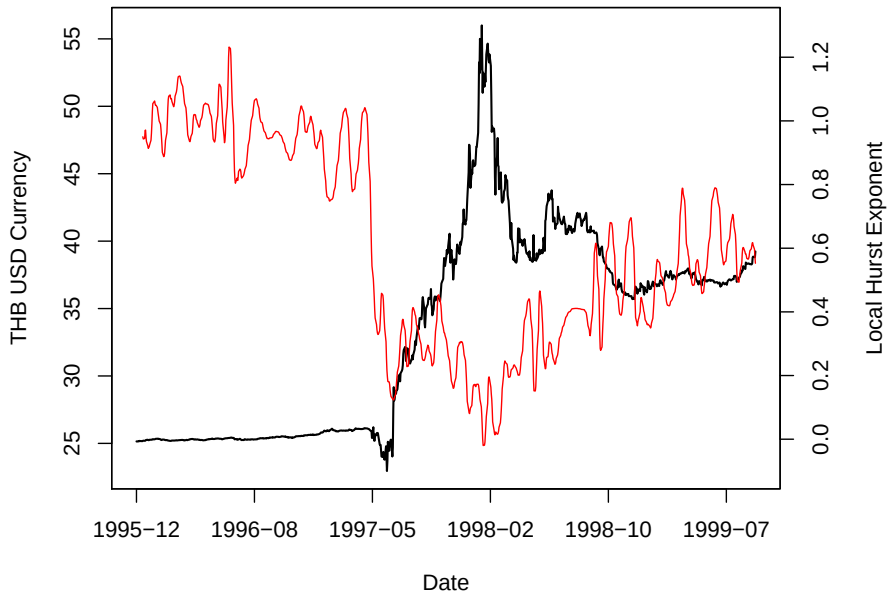


Figure 11.10: Thai BAHT / USD currency pair (black) from December 14th, 1995 to September 08th, 1999 and its local Hurst exponent (red).

11.5.3.2 The dot-com bubble in 2000

The Dot-Com Bubble occurred from 1997 to 2000, and affected all the sectors related to information technology. It has been enhanced by the growth of the World Wide Web which led to speculations on this type of stocks. We are now going to investigate if the local Hurst exponent can provide us with some indications on this phenomena.

Figure (11.11) represents France Telecom SA (former name of Orange SA) stock prices from March 11th, 1998 to September 7th, 2001. Indeed, it shows a growing bubble from 1999 (50 euro per share) till 2000 (180 euro per share) and its burst that occurred in around 2000. The local Hurst exponent fluctuated around 0.6 before 1999, roughly indicating a slight persistent behaviour of the stock prices. It has then suddenly dropped after August 1999, indicating that the market was becoming highly turbulent, since the market prices were increasingly quickly growing. This predicted the forthcoming crash that indeed occurred in the beginning of 2000.

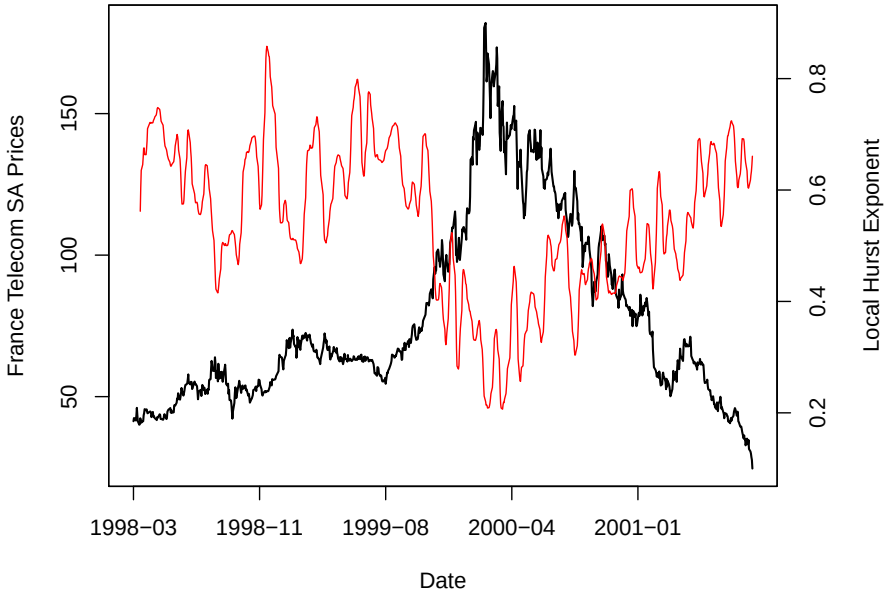


Figure 11.11: France Telecom SA daily prices (black) from March 11th, 1998 to September 07th, 2001 and its local Hurst exponent (red).

11.5.3.3 The financial crisis of 2007

Last but not least, we investigate how the local Hurst exponent varied during the Financial Crisis of 2007. The financial crisis that started in 2007 was characterised by a liquidity and solvency crisis of many financial institutions which propagated to many countries. It is considered by many economists as the worst financial crisis since 1929.

Figure (11.12) represents the NASDAQ Composite Index from 2005 to 2009. We observe that the NASDAQ Index had quite a "normal" behaviour from 2005 till mid-2006, and that from that date onward, more and more erratic fluctuations occurred over time, to finally suddenly drop from $2,500$ points to $1,400$ points at the end of 2008. The local Hurst exponent fluctuated around 0.6 before 2006, then slightly decreased to 0.5 between September 2006 and late 2007, to finally reach 0.4 after 2007. Going further in the analysis, we see that the local Hurst exponent varied more after 2006, indicating a change in its fractal structure. We observe a local minimum of the local Hurst exponent in March 2007, reaching 0.4, which indicates a sudden drop in the index. Again, a similar local minimum is observed in August 2007, and another one in February 2008, to finally reach its global minimum of 0.2 in October 2008, corresponding to a huge loss of $1,000$ points (-56%) in the index. Note, the huge loss in October 2008 did not suddenly appear. We have identified at least three periods of time where the Hurst exponent alarmed us on the possibility of a forthcoming huge drop. We believe that many losses might be avoided by portfolio managers if they correctly interpreted the local Hurst exponent.

As a consequence, through those three experiments on Asian Crisis, Dot-Com Bubble and the 2008 Financial Crisis, we do complete the results from Grech et al., Czarnecki et al. and other authors who claimed that the local Hurst exponent may predict financial crashes. However, we do not pretend that the Hurst exponent is the best way to predict them since the same fluctuations of the Hurst exponent could be observed without leading to any crisis. However, we claim that the fractal structure (and thus the Hurst exponent) do change just before each crisis and as a

result, it provides a good partial predictive indicator of financial crisis. More generally, the local Hurst exponent gives valuable indications on extreme values (upward and downward) in data series.

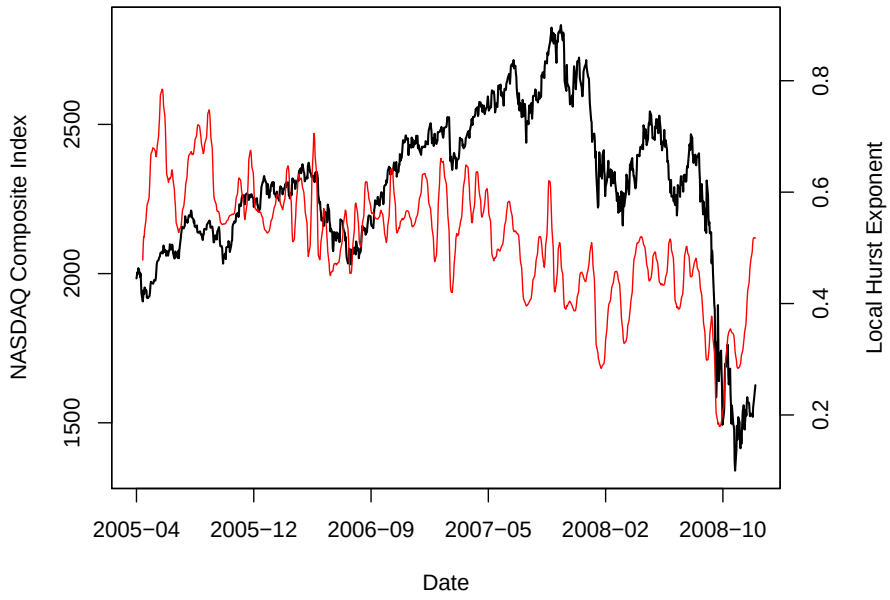


Figure 11.12: NASDAQ Composite Index (black) from April 4th, 2005 to January 05th, 2009 and its local Hurst exponent (red).

11.5.4 Conclusion

While outliers have an inherently isolated and local character, with erratic behaviour (spikes), we saw in Section (11.4.3) that they could be detected and localised in time with the help of the effective local Holder exponents (ELHE) (see Section (11.4.26)). Following the same approach to analyse markets for multifractality, we obtained highly random local Holder exponents, emphasising the strong multifractal nature of financial data. Even the denoised effective exponents obtained with wavelet transform are random with abrupt change in values happening continuously, and extremely fast. That is, the denoised local Holder exponents oscillate around a level of Hurst exponent with a succession of small and large amplitudes (similar to spikes), and is capable of sudden, or abrupt, change to a different level of Hurst exponent in presence of lage extreme price fluctuations related to market crashes. Various authors already suggested that financial markets could be described as a system of a number of coupled oscillators, subject to stochastic regulation (see Matia we al. [2003]). The source of multifractality of financial markets were shown to be due to nonlinear phase ordering as well as the essential contribution of extreme events resulting in fat tails of the event distribution. Even though the width of the multifractal spectra is capable of indicating the presence of large shocks, the oscillating nature of the local Holder exponent characterise the continuously changing dynamics of the response time distribution. Comparing the behaviour of the stock market to that of a healthy heartbeat dominated by two antagonistic systems (sympathetic and parasympathetic) the pessimists who sell stocks and the optimists who buy stocks, leading to a continuously varying number of active agents (degrees of freedom), Struzik [2003] showed that both systems were at work at any moment in time. Hence, the multifractal spectra can detect a change of level of the Hurst exponent (contribution to fat tails), but it can not be used to identify the intermittent changes in the magnitude of response time variation due to, for example, feedback effects. Similarly to outliers, the latter requires a methodology capable of

determining the statistical nature of the non-stationary process both globally and locally, such as the effective local Holder exponent.

11.6 Some multifractal models for asset pricing

In this section we follow Segnon et al. [2013] who explained the construction of simple multifractal measures and detailed multifractal processes designed as models for financial returns. The origins of multifractal theory date back from Kolmogorov [1941] and his work on fully developed turbulence (FDT). Later, Mandelbrot [1974] introduced multifractal measures where he proposed a probabilistic approach for the distribution of energy in turbulent dissipation. Building upon earlier models of energy dissipation from Kolmogorov, he proposed that energy should dissipate in a cascading process on a multifractal set from long to short scales where the set results from operations performed on probability measures. The multifractal cascade starts by assigning uniform probability to a bounded interval, for instance the unit interval $[0,1]$ which is split up into two subintervals receiving fractions m_0 and $1-m_0$, respectively, of the total probability mass of unity of their mother interval. Different choices are possible for the length of the subintervals, such as $\frac{1}{2}$. In the next step, the two subintervals are split up again into similar subintervals of shorter length, receiving again fractions m_0 and $1-m_0$ of the probability mass of their mother intervals. This process being repeated ad infinitum in principle, a heterogeneous, fractal distribution of the overall probability mass results which even in the simplest case has a visual resemblance to time series of returns and volatility in financial markets. This construction reflects the underlying idea of dissipation of energy from the long scales (mother intervals) to the finer scales, preserving the joint influence of all the previous hierarchical levels in the built-up of the cascade. Many variations of this generating mechanism of a simple Binomial multifractal have been proposed, all being implementation of the general form given in Equation (10.1.6) defining multifractality from the scaling behaviour across scales. The recursive construction principles are directly responsible for the multifractal properties of the pertinent limiting measures in accordance to Equation (11.1.11). Denoting by μ a measure defined on $[0,1]$, we get $E[\mu^q(t,t+\Delta t)] \sim c(q)(\Delta t)^{\tau(q)+1}$ so that in the simple case of the Binomial cascade we have $\tau(q) = -\ln E[M^q] - 1$ with $M \in \{m_0, 1-m_0\}$ with probability $\frac{1}{2}$. One must therefore determine the scaling function $\tau(q)$ and the Holder spectrum $f(\alpha)$, as well as the existence of moments in the limit of a cascade with infinite progression.

Multifractal measures have been adapted to asset pricing by using them as a stochastic clock for transformation of chronological time into business time. That is, a time transformation can be represented by stochastic subordination, where $\theta(t)$ is the subordinating process, and the asset price change, $r(t)$, is given by a subordinated process (for example a Brownian motion) measured in transformed time $\theta(t)$. Mandelbrot et al. [1967] introduced the idea of stochastic subordination in financial economics. Later, Mandelbrot et al. [1997] proposed the multifractal model of asset returns (MMAR) where multifractal measure serves as a time transformation from chronological time to business time. It assumes that returns $r(t)$ follow the compound process

$$r(t) = B_H[\theta(t)]$$

where the incremental fBm with Hurst index H, $B_H[\bullet]$, is subordinate to the cumulative distribution function $\theta(t)$ of a multifractal measure constructed as described above. It shares essential regularities observed in financial time series including long tails and long memory in volatility originating from the multifractal measure $\theta(t)$ applied for the transition from chronological time to business time. That is, the heterogeneous sequence of the multifractal measure serves to contract or expand time, and as a result locally contract or expand the homogeneous second moment of the subordinate Brownian motion. Writing $\theta(t) = \int_0^t d\theta(t)$ we see that the incremental multifractal random measure $d\theta(t)$ (which is the limit of $\mu(t,t+\Delta t)$ for $\Delta t \to 0$ and $k \to \infty$ (the number of hierarchical levels)) can be considered as the instantaneous stochastic volatility. Hence, the MMAR essentially applies the multifractal measure to capture the time-dependency and non-homogeneity of volatility. Mandelbrot et al. discussed estimation of the underlying parameters of the model via matching of the $f(\alpha)$ and $\tau(\alpha)$ functions, and showed that the temporal behaviour of various absolute moments of typical financial data squares well with the theoretical results for the multifractal model. Considering the binary cascade described above, the obtained subintervals are assigned fractions of the probability

mass of their mother interval drawn from different types of random distributions. Calvet et al. [2002] discussed Binomial, Lognormal, Poisson and Gamma distributions and associated $\tau(\alpha)$ and $f(\alpha)$ functions. Note, the functions $\tau(\alpha)$ and $f(\alpha)$ capture various moments of the data, so that using them for determination of parameters is similar to moment matching. Estimation of alternative multifractal models has made use of efficient moment estimators as well as other more standard statistical techniques. As the MMAR suffered from the combinatorial nature of the subordinator $\theta(t)$ and its non-stationarity due to the restriction of this measure to a bounded interval, analogous iterative time series models were introduced (see Calvet et al. [2001] and [2004]). Rather than using the grid-based binary splitting of the underlying interval, they assumed that $\theta(t)$ was obtained in a grid-free way by determining a Poisson sequence of change points for the multipliers at each hierarchical level of the cascade. In the limit $k \to \infty$ the Poisson multifractal exhibits typical anomalous scaling, carrying over from the time transformation $\theta(t)$ to the subordinate process for asset returns, $B_H[\theta(t)]$. As opposed to the MMAR, the Poisson multifractal possesses a Markov structure allowing for better statistical tractability. This model motivated the development of the discrete Markov-switching multifractal model (MSM) that is wildly applied in empirical finance. Later, Lux et al. [2013] interpreted the Poisson MMAR as a regime-switching diffusion process with 2^k different volatility states. In the discrete version, the volatility dynamics can be interpreted as a discrete time Markov-switching process with a large number of states. Returns follow Equation (3.4.13) with innovations ϵ_t drawn from a standard normal distribution $N(0,1)$ and instantaneous volatility being determined by the product of k volatility components or multipliers $M_t^1, ..., M_t^k$ and a constant scale factor σ, that is,

$$r_t = \sigma_t \epsilon_t$$

with

$$\sigma_t^2 = \sigma^2 \prod_{i=1}^{k} M_t^i$$

such that the volatility components M_t^i are persistent, non-negative and satisfy $E[M_t^i] = 1$. Further, the volatility components at time t are statistically independent. Each M_t^i is renewed at time t with probability γ_i depending on its rank within the hierarchy of multipliers and remains unchanged with probability $1 - \gamma_i$. Assuming

$$\gamma_i = 1 - (1 - \gamma_1)^{(b^{i-1})}$$

with $\gamma_1 \in [0,1]$ the component at the lowest frequency that subsumes the Poisson intensity parameter λ, and $b \in (1, \infty)$, the discretised Poisson multifractal converges to the continuous time limit for $\Delta t \to 0$. Calvet et al. [2004] assumed a Binomial distribution for M_t^i with parameters m_0 and $2 - m_0$ such that $E[M_t^i] = 1$ for all i. Ignoring convergence to the limit, one can choose the simpler parametrisation $\gamma_i = b^{-i}$. Note, the lognormal distribution for the distribution of the multipliers M_t^i was also considered (see Liu et al. [2007]), with parameter λ and s,

$$M_t^i \sim Ln(-\lambda, s^2)$$

such that for $s^2 = 2\lambda$ we get $E[M_t^i] = 1$. However, the binomial and lognormal multipliers have almost identical results showing that the former is sufficiently flexible.

In the econophysics literature, the multifractal random walk (MRW) developed almost simultaneously as a different type of causal, iterative process. It is essentially a Gaussian process with built-in multifractal scaling via an appropriately defined correlation function. Even though various distributions for the multipliers as the guideline for the construction of different versions of MRW replicating their autocorrelation structures can be used, the literature concentrated on the lognormal distribution. For instance, Bacry et al. [2001] defined the MRW as a Gaussian process with a stochastic variance given by

$$r_{\Delta t}(\tau) = e^{w_{\Delta t}(\tau)} \epsilon_{\Delta t}(\tau)$$

where Δt is a small discretisation step, $\epsilon_{\Delta t}(\bullet)$ is a Gaussian variable with mean zero and variance $\sigma^2\Delta t$, and $w_{\Delta t}(\bullet)$ is the logarithm of the stochastic variance with τ a multiple of Δt along the time axis.In the special case where $w_{\Delta t}(\bullet)$ follows a Gaussian distribution, we get lognormal volatility draws. For longer discretisation steps (for example daily unit time intervals), we get the returns as

$$r_{\Delta t}(t) = \sum_{i=1}^{\frac{t}{\Delta t}} e^{w_{\Delta t}(i)}\epsilon_{\Delta t}(i)$$

To mimic the dependency structure of a lognormal cascade, the returns are assumed to have covariances

$$Cov(w_{\Delta t}(t)w_{\Delta t}(t+h)) = \lambda^2 \ln(\rho_{\Delta t}(h))$$

with

$$\rho_{\Delta t}(h) = \begin{cases} \frac{T}{(|h|+1)\Delta t} \text{ for } |h| \leq \frac{T}{\Delta t} - 1 \\ 0 \text{ otherwise} \end{cases}$$

where T is the assumed finite correlation length (a parameter to be estimated) and λ^2 is the intermittency coefficient characterising the strength of the correlation. For the variance of $r_{\Delta t}(t)$ to converge, one must assume that $w_{\Delta t}(\bullet)$ obey

$$E[w_{\Delta t}(i)] = -\lambda^2 \ln(\frac{T}{\Delta t}) = -Var(w_{\Delta t}(i))$$

Assuming a finite decorrelation scale, rather than a monotonic hyperbolic decay of the autocorrelation, serves to guarantee stationary of the multifractal random walk. Hence, the MRW model does not obey an exact scaling function in the limit $t \to \infty$, as it is only characterised by apparent long-term dependence over a bounded interval. Nonetheless, it possesses nice asymptotic properties facilitating applications of many standard tools of statistical inference. Note, Bacry et al. [2008] showed that the continuous limit of MRW can also be interpreted as a time transformation of a Brownian motion subordinate to a lognormal multifractal random measure. Hence, the MRW can be reformulated like the MMAR model as

$$r(t) = B[\theta(t)] \text{ for all } t \geq 0$$

where $\theta(t)$ is a random measure for the transformation of chronological time to business time, and $B(t)$ is a Brownian motion independent of $\theta(t)$. The business time $\theta(t)$ is obtained along the lines of the MRW model as

$$\theta(t) = \lim_{\Delta \to 0} \int_0^t e^{2w_\Delta(u)} du$$

where $w_\Delta(u)$ is the stochastic integral of Gaussian white noise $dW(s,t)$ over a continuum of scales, s, truncated at the smallest and largest scales Δ and T leading to a cone-like structure defining $w_\Delta(u)$ as the area delimited in time (over the correlation length) and a continuum of scales, s, in the (t,s) plane

$$w_\Delta(u) = \int_\Delta^T \int_{u-s}^{u+s} dW(v,s)$$

A particular correlation structure of the Gaussian elements $dW(v,s)$ must be imposed in order to replicate the weight structure of the multipliers in discrete multifractal models. That is, the multifractal properties are obtained for the following choices of expectation and covariances of $dW(v,s)$

$$Cov(dW(v,s), dW(v^{'}, s^{'})) = \lambda^2 \delta(v - v^{'})\delta(s - s^{'})\frac{1}{s^2} dvds$$

and

$$E[dW(v,s)] = -\lambda^2 \frac{1}{s^2} dvds$$

Bacry et al. [2003] showed that the limiting continuous-time process exists and possesses multifractal properties. Later, Bacry et al. [2013] provided results for the unconditional distribution of returns obtained from this process, demonstrating that it is characterised by fat tails and that it becomes less heavy tailed under time aggregation. They also showed that standard estimators of tail indices are ill-behaved for data from a MRW data-generating process due to the high dependency of adjacent observations. Note, a similar mismatch between implied and empirical tail indices applies to other multifractal models.

While price fluctuations in assert markets exhibit a well known degree of asymmetry due in part to leverage effects, the models described so far assumed complete symmetry for both positive and negative returns. To remedy this problem Pochart et al. [2002] proposed the discrete time skewed multifractal random walk (DSMRW) as an extended version of the MRW by incorporating a direct influence of past realisations on contemporaneous volatility

$$\tilde{w}_{\Delta t}(i) = w_{\Delta t}(i) - \sum_{k<i} K(k,i)\epsilon_{\Delta t}(k)$$

where $K(k,i) = \frac{K_0}{(i-k)^\alpha (\Delta t)^\beta}$ is a positive definite kernel for the influence of returns on subsequent volatility. Bacry et al. [2012] proposed a continuous time skewed multifractal model incorporating also the leverage effect. Eisler et al. [2004] expended the MSM model in a similar way where asymmetry comes in via the renewal probabilities. Later, another asymmetric MSM model was introduced by Calvet et al. [2013] where a multifractal cascade was embedded into a stochastic volatility model. The product of multipliers enters as a time-varying long-run anchor for the volatility dynamics while at the same time governing a jump component in returns relating positive volatility shocks to negative return shocks.

It is interesting to note that Calvet et al. [2006] also introduced a bivariate MF model in the case of a portfolio made of two assets α and β. We let r_t be the vector of log-returns of the portfolio with r_t^α and r_t^β the individual log-returns of the two assets, respectively. Then, the portfolio return is given by

$$r_t = [g(M_t)]^{\frac{1}{2}} \epsilon_t$$

where $g(M_t)$ is a 2×1 vector, $M_{1,t} \times M_{2,t} \times ... \times M_{k,t}$ denotes element by element multiplication, and the column vectors $\epsilon_t \in \mathbb{R}^2$ are i.i.d. Gaussian $N(0, \Sigma)$ with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho_\epsilon \sigma_\alpha \sigma_\beta \\ \rho_\epsilon \sigma_\alpha \sigma_\beta & \sigma_\beta^2 \end{pmatrix}$$

where ρ_ϵ is the unconditional correlation between the residuals as the first source of correlation between both returns. The period t volatility state is characterised by a $2 \times k$ matrix $M_t = (M_{1,t}, M_{2,t}, .., M_{k,t})$ and the vector of the components at the ith frequency is $M_{i,t} = (M_{i,t}^\alpha, M_{i,t}^\beta)$. The volatility vectors $M_{i,t}$ are non-negative and satisfy $E[M_{i,t}] = I$ where $I = (1,1)^\top$. The choice of the dynamics for each vector $M_{i,t}$ is that volatility arrivals are correlated but not necessarily simultaneous across markets, depending on the correlation coefficient λ. Considering two random variables $I_{i,t}^\alpha$ and $I_{i,t}^\beta$ being equal to 1 if each series $c \in \{\alpha, \beta\}$ is hit by an information arrival with probability γ_i, and equal to zero otherwise, Calvet et al. specified the arrival vector to be i.i.d. and assumed its unconditional distribution to satisfy three conditions. First, the arrival vector is symmetrically distributed, $(I_{i,t}^\alpha, I_{i,t}^\beta) = (I_{i,t}^\beta, I_{i,t}^\alpha)$ in distribution. Second, the switching probabilities of both series are equal for each level i, $P(I_{i,t}^\alpha = 1) = P(I_{i,t}^\beta = 1) = \gamma_i$, with γ_i as in the univariate MSM. Third, there exists $\lambda \in [0,1]$ such that

$$P(I_{i,t}^\alpha = 1 | I_{i,t}^\beta = 1) = (1-\lambda)\gamma_i + \lambda$$

These there conditions define a unique distribution of $(I_{i,t}^{\alpha}, I_{i,t}^{\beta})$ whose joint switching probabilities can easily be determined. The univariate dynamics of each series coincides with a univariate MSM model. Liu [2008] considered a closely related bivariate multifractal model based on the assumption that two time series have a certain number of joint cascade levels in common, while the remaining one are chosen independently. The returns are given by

$$r_{q,t} = \big[(\prod_{i}^{k} M_{i,t})(\prod_{l=k+1}^{n} M_{l,t})\big]^{\frac{1}{2}} \epsilon_t$$

where $q = 1, 2$ refers to the two time series, both having an overall number of n levels of their volatility cascades, and they share a number k of joint cascade levels which govern the strength of their volatility correlation. These bivariate models have been generalised for more than two assets in various ways. Similarly, Bacry et al. [2000] proposed a generalisation of the MRW, called the multivariate multifractal random walk (MMRW).

Peltier et al. [1995] proposed the multifractal Brownian motion (MFBM) in which the scaling exponent H is allowed to vary in time. We only present the main properties of the MFBM. Let $H(t)$ be a Holder continuous function in the interval $t \in [0, 1]$ with Holder exponent $\beta > 0$, such that for any $t > 0$ we have $0 < H(t) < \min(1, \beta)$. The MFBM $\{W_{H_t}(t), t > 0\}$ is the Gaussian process defined by

$$W_{H_t}(t) = \frac{1}{\Gamma(H_t + \frac{1}{2})} \int_{-\infty}^{t} \big[(t-s)_+^{H_t - \frac{1}{2}} - (-s)_+^{H_t - \frac{1}{2}}\big] dB(t)$$

where $\Gamma(x)$ is the Gamma function, $(x)_+$ equals x if $x > 0$ and 0 otherwise, and $B(t)$ is a Brownian motion. One important aspect of that process is that its increments are no-longer stationary since it can be shown that

$$E[\big(W_{H_{t+k}}(t+k) - W_{H_t}(t)\big)^2] \approx k^{2H(t)} \text{ as } k \to 0$$

Because of its nonstationarity, the MFBM is no-longer a self-similar process either, but it is possible to define the concept of locally asymptotically similarity at the point $t_0 > 0$ in the following way

$$\frac{W_{H_{t_0+at}}(t_0 + at) - W_{H_{t_0}}(t_0)}{a^{H_{t_0}}} = W_{H_{t_0}}(t) \text{ as } a \to 0$$

Hence, one can think of the process $W_{H_t}(t)$ as a process that resemble, at local time t, a fractional Brownian motion $B_{H_t}(t)$. Peltier et al. presented a practical way to generate a MFBM path, and provided an algorithm for its implementation. Different properties of this process have been studied such as Holder continuity of the paths and Hausdorff dimension, local times of MFBM, and estimates for the local Hurst parameters (see Bertrand et al. [2013]). In white noise analysis, MFBM was considered by Lebovits et al. [2014] together with its respective stochastic calculus.

Chapter 12

Systematic trading

This chapter is under construction and will be completed in the next version.

12.1 Introduction

Through out this book we saw that the equilibrium theory of market prices could not explain the qualitative aspect coming from human decision making process and the resulting market predictability. New tools developed to analyse financial data, such as the wavelet analysis (see Chapter (6)) and the fractal analysis (see Chapter (10)), highlighting the multifractal nature of financial markets (see Chapter (11)). As a result, new theories emerged to explain financial markets, among which is a multitude of interacting agents forming a complex system characterised by a high level of uncertainty. Complexity theory deals with processes where a large number of seemingly independent agents act coherently. Multiple interacting agent systems are subject to contagion and propagation phenomena generating feedbacks and producing fat tails. Real feedback systems involve long-term correlations and trends since memories of long-past events can still affect the decisions made in the present. Most complex, natural systems, can be modelled by nonlinear differential, or difference, equations. These systems are characterised by a high level of uncertainty which is embedded in the probabilistic structure of models. As a result, econometrics can now supply the empirical foundation of economics. For instance, science being highly stratified, one can build complex theories on the foundation of simpler theories. That is, starting with a collection of econometric data, we model it and analyse it, obtaining statistical facts of an empirical nature that provide us with the building blocks of future theoretical development.

New techniques combining elements of learning, evolution and adaptation from the field of Computational Intelligence developed, aiming at generating profitable portfolios by using technical analysis indicators in an automated way. In particular, subjects such as Neural Networks, Swarm Intelligence, Fuzzy Systems and Evolutionary Computation can be applied to financial markets in a variety of ways such as predicting the future movement of stock's price or optimising a collection of investment assets (funds and portfolios). These techniques assume that there exist patterns in stock returns and that they can be exploited by analysis of the history of stock prices, returns, and other key indicators (see Schwager [1996]). With the fast increase of technology in computer science, new techniques can be applied to financial markets in view of developing applications capable of automatically manage a portfolio. Consequently, there is substantial interest and possible incentive in developing automated programs that would trade in the market much like a technical trader would, and have it relatively autonomous. A mechanical trading systems (MTS), founded on technical analysis, is a mathematically defined algorithm designed to help the user make objective trading decisions based on historically reoccurring events.

With the growing quantity of data available, machine-learning methods that have been successfully applied in science are now applied to mining the markets. Data mining and more recent machine-learning methodologies provide a

range of general techniques for the classification, prediction, and optimisation of structured and unstructured data (see details in Chapter (13)). Neural networks, classification and decision trees, k-nearest neighbour methods, and support vector machines (SVM) are some of the more common classification and prediction techniques used in machine learning. Further, combinatorial optimisation, genetic algorithms and reinforced learning are now widespread. Using these automated techniques, one can describe financial markets through degrees of freedom which may be both qualitative and quantitative in nature, each node being the siege of complicated mathematical entity. One could use a matrix form to represent interactions between the various degrees of freedom of the different nodes, each link having a weight and a direction. Further, time delays should be taken into account, leading to non-symmetric matrix (see Ausloos [2010]). For instance, conditional relationships can efficiently be described in a hierarchical fashion like a decision tree. A decision tree is a simple set of if-then-else rules, making it intuitive and easy to analyse, where the most important relationships are first considered, and the less significant ones are considered farther out in the branches of the tree. Sorensen et al. [1998] applied a classification and regression technique (CART) to construct optimal decision trees to model the relative performance of the $S\&P$ 500 with respect to cash. The model assigns different probabilities to three market states: outperform, underperform, and neutral. Since the CART allows for the variables to have non-linear behaviour and conditional relationships, the explanatory variables used can be capital market data such as steepness of the yield curve, credit spread, equity risk premium, dividend yield. The final result is a decision tree of non-linear if-then-else rules where each rule is conditioned on previous rules.

12.2 Technical analysis

12.2.1 Definition

While trading involves the study of technical factors governing short-term market movements together with the behaviour of the market, technical trading rules involve the use of technical analysis (TA) (see Section (2.1.5)) to design indicators (called technical indicators) helping a trader determine whether current behaviour is indicative of a particular trend, together with the timing of a potential future trade. TA is the search for recurrent and predictable patterns in the stock prices by using the past price or volume data in order to help investors anticipate what is most likely to happen to the prices over time. Two important points from the Dow Theory are

1. prices discount everything: current price of stock fully reflects all the information. TA utilise the information captured by the price to interpret what the market is saying with the purpose of forming a view on the future 18
2. price movements are not totally random: most technicians believe that there are inter spread period of trending prices in between random fluctuations. The aim of the technicians is to identify the trend and then make use of it to trade or invest.

As a result, in order to apply TA, we must assume that the historical data in the markets forms appropriate indications about the market future performance. Hence, by analysing financial data and studying charts, we can anticipate the way the market is most likely to go. Due to the difficulty of managing complex models of the market and using rigorous, adaptive decision techniques, day traders tend to use simpler and more intuitive decision rules. This rule of thumb approach is quite effective (see Ramamoorthy [2003]) and has been considered a good candidate for automation (see Dempster et al. [1999]). The main assumption behind TA is that a robust strategy can be designed by composing multiple intuitive strategies. Robustness and relatively complex behaviours can be achieved by synthesising multiple, intuitive strategies. For instance, if an agent were to buy a share at a low price and sell it at a higher price, then a profit of high price - low price would be made, and numerous such trades over the day would accumulate profit for the trader. However, the future prices are unknown and it is therefore not clear whether a decision to buy or sell in anticipation of favourable movements in the future would yield profit. This decision is further complicated by the strict need to unwind, as it is the case in intraday trading. Unwinding is the process of selling excessive shares, if in excess, and buying the number of shares required to make up the deficit, when short.

12.2.2 Technical indicator

As there is no single kind of technical indicator (TI) which can work for all the stocks in the market, there is no optimal way of combining the different TI, so that technical analysis (TA) is more of an art than a science. This led technicians to use a large number of individual TI, and open source library for TI developed. For instance, Chauhan [2008] used 14 TI and coded his own ones to introduce some form of heuristic and generate more robust buy or sell signals. TI is defined as a series of data points that are derived by applying a formula to the price data of security which can be combination of the open, high, low or close over a period of time 18. These data points can be used to generate buy or sell signal. TI can provide unique viewpoint on the strength and direction of the underlying price action. In general, it is difficult to combine TI since they should complement each other rather than moving in unison and generate the same signal 18. Hence, we must make sure that the combination we choose provides different perspective towards the underlying price or volume movement. In addition, a combination that might work for one stock might not work for another one. Further, for each combination of TI, there are many parameters which need to be optimised (the size of the window of observation).

12.2.3 Optimising portfolio selection

The areas of artificial intelligence and its application to technical trading and finance have seen a significant growth in interest over the last ten years, especially the use of genetic algorithms (GA) and genetic programming (GP). These approaches have found financial applications in option pricing (see Allen et al. [1999]), the calibration of implied volatility (see Bloch et al. [2011]), and as an optimisation tool in technical trading (see Dempster et al. [2001]). Evolutionary learning algorithms derive inspiration from Darwinian evolution. GPs are a variation of the standard genetic algorithm, wherein string lengths may vary within the solution space. While GAs are population-based optimisation algorithms, GPs are an extension of this idea proposed by Koza [1992] with a view to evolving computer programs. A simple starting point with using GAs is to represent the solution space as a finite number of strings or binary digits. Binary strings are an effective form of representation since complex statements as well as numerical values of parameters can be represented in this form. The resulting search space is finite when parameters take only discrete values to yield a binary representation as a string of fixed length. We then need to evaluate the fitness of the constituents of the solution space, that is, the suitability of each potential solution in terms of its performance. When selecting trading rules the fitness can be viewed as the profitability of the rule tested over a time series of historical price data, or a function of this variable. Unlike in GAs, solutions in GP can be seen as non-recombining decision trees (see Papagelis et al. [2000]) with non-terminal nodes as functions and the root as the function output. These are usually the optimisation algorithm of choice in cases of evolving strategies based on Boolean operators, when the solution may be evolved with varying depth in the tree (see Dempster et al. [2001b]). It is inherently more flexible than the GA, but care needs to be taken in representation to avoid over-fitting (the phenomenon where a classifier is trained too minutely to fit the training data, causing diminished performance on data outside of the training sample called the out-sample data). Measures calculated over a period of time, such as the Sharpe ratio and the Sortino ratio, are used as the fitness function to evaluate the performance of each member of a population in every generation. Hence, the goal of the evolution process is to find the combination of weights maximising the fitness functions for a given set of training days (in sample). However, even though we can find such weights, there is no guarantee that we will maximise profit over the out of sample data.

One must design trading strategies that resemble a technical trader who would systematically, and based on a set of pre-specified evaluation criteria, choose a subset of trading strategies from a larger set of trading rules, and evaluate the performance of these strategies in various competitive scenarios. We should combine trading strategies and rules with the use of GAs and GPs and evolve strategies that are aimed to be profitable and perform well under the different evaluation criteria involved. In an attempt at making a profitable, robust strategy, by using simple intuitive laws that appeal to the human trader, we use multiple technical trading rules in a weighted combination to produce a unified strategy. The generation of effective strategies using complete, comprehensible indicator strategies may help in the understanding of these component strategies, their effects and limitations. In this formulation, we use a number of

basic (indicator) strategies in a weighted combination to produce a cumulative trading action at every tick.

12.2.3.1 Classifying strategies

For example, the algorithm can consider the principles from the weighted majority algorithm (see Littlestone et al. [1992]) in the use of suggestions from the component strategies and combining them using a weighted majority. They considered the use of a similar suggestion or voting mechanism wherein each indicator would signal a buy, sell, or do nothing action. The steps involved in trading in this environment include getting the raw order book data, evaluation of a recommended action by each of the indicator strategies and combining these indicators using the respective weights giving us a cumulative suggestion (a weighted majority). This is followed by a multiple model control mechanism (CM) that determines the mode it should trade in, based on the current holdings of the agent. The control mechanism is like a faucet acting as the final regulator on the trading decision and volume suggested by the composite strategy so far. It has the power to veto a trading decision suggested by the composite algorithms, when in the safe mode or allow the trade to continue unaltered in the regular mode.

Assuming that a trader would behave differently when in an extremely long or short position as opposed to a neutral share position, the agent can evaluate its current share position, and if necessary, it can keep away from following a possibly unidirectional market trend to the end of the day, and reaching a very long or short position. This mechanism is a control measure to ensure the agent achieves a share position as close to neutral (zero accumulation or deficit) as possible. This multiple model scheme examines the agent's share position and provides a mode of operation that the agent should follow. This regulation is combined with the decision that is output by the composit strategy to arrive at a final trading decision to buy, sell or do nothing together with the volume to be traded. The indicators, although complete strategies in themselves, only provide the system with suggestions to buy or sell. Weights $(\omega_1, .., \omega_n)$ are applied to the indicators to arrive at a weighted suggestion $W \times I$. The final decision depends on this weighted suggestion together with the output of multiple model control mechanism which gives the final action to be performed by the agent. The weighted suggestion to trade a certain way determines the volume of shares to be traded at a tick. The weights $(\omega_1, .., \omega_n)$, represented as two discrete bits in a bit string, are tuned offline with a genetic algorithm. A sign bit associated with each weight is used to evaluate the effect of the weight on the rule.

Moody et al. [2001] used an adaptive algorithm called Recurrent Reinforcement Learning (RRL) to optimise a portfolio and showed that direct reinforcement can be used to optimise risk adjusted returns. The RRL algorithm learns profitable trading strategies by maximising risk adjusted returns measured by the Sharpe ratio, and avoid the downside risk by maximising the Downside Deviation (DD) ratio. RRL trader performed far better than Q trader and enables a simpler problem representation.

The extended classifier system (XCS) is an accuracy based classifier system where classified fitness is derived from estimated accuracy of reward prediction rather than from reward prediction themselves. It is an online learning machine (OLM) which improves its behaviour with time by learning through reinforcement. That is, if it does well we give it positive reward, else we penalise it in some form. Chauhan [2008] improved the learning of the classifier system by incorporating different reinforcement learning algorithm. Assuming that price movements follow some patterns of ups and downs, he tried to model financial forecasting as a multi step process and implemented $Q(1)$ and $Q(\lambda)$ RL algorithm. In his settings the agents present in the current system converts the information given by the technical indicators into input (binary string) for the XCS. The system further tries to learn the optimum decision it should take when faced with a particular combination of binary bits. The quality of the learning process for the system to be robust and produces reliable decision (buy, sell or hold) depends on the way we combine the technical indicator information in the system. As a result, one must experiment with the combination of the technical indicators as well as with the reward mechanism of the system.

12.2.3.2 Examples of multiple rules

Many traders aim to practise TA as systematically as possible without automation (rule construction) while others use TA as the basis for constructing systems that automatically recommend trade positions (system construction) (see Pictet et al. [1992]). Even though there exists a large number of technical trading strategies in intraday stock trading, most of them are too simplistic and too coarse, but combining them with genetic algorithm (GA) to tune the relative merits of the individual rules, much like traders would do, provides much better results. Various attempts at synthesising multiple rules to come up with a composite trading strategy have been tried before, with the component rules being everything from complete rules (see Li et al. [1999]) to very basic predicates and operators that are combined to generate complex rules (see Oussaidene et al. [1997]). Refenes [1995] considered the method of genetic-based global learning in a trading system in order to find the best combination of indicators for prediction and trading. Similarly, Dunis et al. [1998] used a genetic algorithm on a currency exchange system to optimise parameters based on an ensemble of simple technical trading indicators. In the same spirit Neely et al. [1997] used GP to discover profitable trading rules. Generalising the approach, Dempster et al. [2001] proposed a framework for systematic trading system construction and adaptation based on genetic programs. Work in using a multiple model approach for the development of an effective intraday trading strategy was proposed by Ramamoorthy [2003]. Attempts at designing a strategy in the PLAT [1] domain using reinforcement learning and hill climbing as well as a market maker to be used in completion in the domain have been explored by Sherstov [2003]. Subramanian [2004] aimed at synthesising a strategy where the component rules are independent strategies in themselves, so that they would appeal to a human trader. He tried to see if existing, simple, intuitive strategies could be composed in some way to make a robust, profitable strategy. Subramanian studied the effect of combining multiple intuitive trading rules within the framework of PLAT by exploring two schemes of operation (adding weights to the trading suggestions or deleting certain rules) wherein the automated agent trades based on a combination of signals it receives from the various simplistic rules. Subramanian showed that technical indicators are not as profitable when used alone as they are when used in conjunction with other technical trading rules. Ramamoorthy et al. [2004] designed safe strategies for autonomous trading agents by considering a qualitative characterisation of the stochastic dynamics of some simple trading rules.

12.3 Forecasting financial series with neural networks

12.3.1 Generalised nonlinear nonparametric models

12.3.1.1 Presentation

Among the various techniques to forecast and classify financial time series, we saw in Chapter (5) that fundamental and technical analysis were the most popular ones. Even though statistical procedures were widely used for pattern recognition, the effectiveness of these methods relies both on model's assumptions and prior knowledge on data properties. To remedy these pitfalls, several classifiers developed, using various data mining and computational intelligence methods such as rule induction, fuzzy rule induction, decision trees, neural networks etc. For instance, the best recognised tools in the currency markets is the artificial neural networks (ANNs), supported by numerous empirical studies (see Ahmed et al. [2010]). The foremost reason for using ANNs is that there is some nonlinear aspect to the forecasting problem under consideration, taking the form of a complex nonlinear relationship between the independent and dependent variables. The characteristics of financial time series, such as equity stock or currency markets, are influenced by the psychology of traders (behavioural finance) and are strongly non-linear and hardly predictable (see Maknickiene et al. [2011]).

Among the different networks existing, the artificial neural networks (ANNs) and the artificial recurrent neural networks (RNNs) are computational models designed by more or less detailed analogy with biological brain modules.

[1] The PLAT domain uses the Penn Exchange Server (PXS) (real world, real time stock market data for simulated automated trading) to which the trading agents can plug in (see 19). PXS works in a maner very similar to the Island electronic crossing network (ECN).

The former is presented in Section (13.6.1.2) and the latter is introduced in Section (13.7.1). ANNs, formalised by McCulloch et al. [1943], are a class of generalised nonlinear nonparametric models inspired by studies of the human brain. They get their intelligence from learning process, giving them the capability of auto-adaptability, association, and memory to perform certain tasks. The backpropagation network, which is the most popular and the most widely implemented neural network in the financial industry, is based on a multi-layered feedforward topology with supervised learning. The network is fully connected, with every node in the lower layer linked to every node in the next higher layer via weight values. The learning of the backpropagation neural network is based on an error minimisation procedure where the weights are modified according to an error function comparing the neural network output with the training targets. We evaluate the derivatives of the error function with respect to the weights which are used to compute the adjustment to be made to the weights. The simplest such techniques involves gradient descent. In the case of time series prediction, a set of input-target pairs is created, forming the training samples. Every time a new time series is generated, new observations are added to the set and the oldest ones are dropped out. Alternatively, artificial recurrent neural networks (RNNs) are a class of feedback artificial neural network architecture that uses iterative loops to store information, which is inspired by the cyclical connectivity of neurons in the brain. The existence of cycles allows RNNs to develop a self-sustained temporal activation dynamics along its recurrent connection pathways, even in the absence of input, making them a dynamical system. This feature of the RNNs make them attractive for processing serially correlated time series. Unlike the ANNs and traditional time series models, the RNNs are flexible enough, and avoids taking the size of sliding windows into account because they can decide what to store and what to ignore during the learning process.

One advantage in using an ANN is that the researcher does not need to know a priori the type of functional relationship existing between the independent and dependent variables (see Darbellay et al. [2000]). A vast literature demonstrated that neural network performs better than conventional statistic approaches in financial forecasting (see Refenes et al. [1994], Adya et al. [1998], Abu-Mostafa et al. [2001]). For many financial forecasting problems, classification models work better than point prediction. Further, Qi [2001] argued that due to the continually changing nature of financial relationships, ANNs are more likely to outperform traditional techniques when the input data is kept as current as possible. This is done by recursive modeling, where the researcher adds new observations and drops the oldest ones each time a new time series forecast is made (sliding window). Olson et al. [2003] compared neural network forecasts of one-year ahead Canadian stock returns with the forecasts obtained by using ordinary least squares (OLS) and logistic regression techniques and showed that the backpropagation algorithm outperformed the best regression alternatives for both point estimation and in classifying firms expected to have either high or low returns This superiority of the NNs translated into greater profitability using various trading rules. Using data from four major stock market indexes, Fok et al. [2008] compared linear regression and neural network backpropagation by testing their forecasting performances. They showed that the latter had better prediction accuracy.

12.3.1.2 Describing the models

Given a time series $\{x_t, x_{t-1}, ...\}$, we aim to learn from the known data to predict future values. That is, we want to estimate x at some future time

$$\widehat{x}_{t+h} = f(x_t, x_{t-1}, ..., x_{t-p}, z_1, z_2, .., z_q)$$

where h is the horizon prediction, and z_q is the qth other explanatory variable. Before training the network, the input data should be normalised into the interval $[b_L, b_H]$, with b_L and b_H being lower and upper bonds, by using the equation

$$\widehat{x}_i = (b_H - b_L)\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} + b_L \text{ for } i = 1, ..., N \tag{12.3.1}$$

where $\min(x_i)$ is the minimal value of all x_i, $\max(x_i)$ is the maximal value of all x_i, and $\widehat{x}_i$ is the normalised value of x_i, respectively. Note, we sometime need to perform operations on the normalised data within a sample window.

When comparing the results among different windows we must always convert them back into the original order of magnitude using

$$x_i = [\max(x_i) - \min(x_i)]\frac{\widehat{x}_i - b_L}{(b_H - b_L)} + \min(x_i) \text{ for } i = 1, ..., N$$

The statistical approach to forecasting time series involves the construction of stochastic models to predict the value $\hat{x}_{t+h}$ given previous observations. One approach is to use an $ARMA(p,q)$ model, but it lacks the ability to capture the nonlinear features of financial time series. We can extend linear models to nonlinear ones. For example, we can generalise an $AR(p)$ model to become a *nonlinear autoregressive* (NAR) model (see Connor et al. [1994])

$$x_t = h(x_{t-1}, x_{t-2}, \cdots, x_{t-p}) + e_t, \tag{12.3.2}$$

where h is an unknown smooth function. Assuming that $E(e_t|x_{t-1}, x_{t-2}, \cdots) = 0$ and e_t has finite second moment, then the minimum mean square error optimal predictor of x_t with the knowledge of $x_{t-1}, ..., x_{t-p}$ is the conditional mean:

$$\widehat{x}_t = E(x_t|x_{t-1}, \cdots, x_{t-p}) = h(x_{t-1}, \cdots, x_{t-p}), t > p$$

Alternatively, as discussed above we can use ANNs. Lapedes [1987] showed that feedforward networks are NAR models for time series prediction. The predictor is given by the approximation of h as follow

$$\widehat{x}_t = \widehat{h}(x_{t-1}, \cdots, x_{t-p})$$

with

$$\widehat{h}(x_{t-1}, \cdots, x_{t-p}) = \sum_{i=1}^{N_h} w_i f(\sum_{j=1}^{p} w_{ij} x_{t-j} + b_i) \tag{12.3.3}$$

Figure 12.1 shows the general architecture of the NAR as an extension of the $AR(p)$ in a feedforward network. The weights, $\{w_i\}$ and $\{w_{ij}\}$ in Equation (12.3.3) can be seen as knobs defining the input-output function, $\widehat{h}$, of the network. N_h is the number of neurons in the second layer, f is a monototic, smooth, and bounded function, b_i is a bias. The weights of the network, $\{w_i\}$ and $\{w_{ij}\}$ are estimated through supervised learning process given a time series sample $x_0, x_1, \cdots, x_n$, and the input-target pairs in the training set are $((x_{t-1}, x_{t-2}, \cdots, x_{t-p}), x_t)$, where $p > 0, t \geq p$.

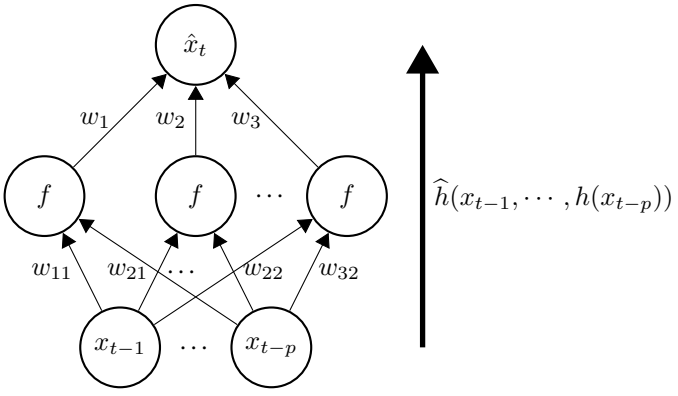


Figure 12.1: Feedforward network for nonlinear autoregressive model.

More generally, to illustrate the nonlinearity of ANNs, we consider a multilayer network with $p = 1, .., P$ input-output training pairs where $p = P$ is the most recent observation, and assume a set of artificial neurons $\{f_{i,j,l}\}_{k=0}^{K}$

where the multilayer subscript $k = 0$ corresponds to the set of inputs $\{x_i\}_{i=1}^{N_0}$ and $k = K$ corresponds to the set of outputs $\{y_i\}_{i=1}^{N_K}$. For simplicity of exposition, we focus on a system with one hidden layer where the subscripts i and j represent the nodes on the input ($(K-2)$ layer) and the $(K-1)$th hidden layer, and the subscript s represent the nodes on the kth layer, respectively. In that setting, the output $O^p_{s,K}(x,w)$ can be expressed as

$$O^p_{s,K}(x,w) = g_{N_K}\Big(\sum_{j=1}^{N_{K-1}} g_{N_{K-1}}\big(\sum_{i=1}^{N_{K-2}} O^p_{i,K-2}(x,w)w^{k-2}_{i,j} + b_{j,K-1}\big)w^{k-1}_{j,s} + b_{s,K}\Big)\ , s = 1,..,N_K$$

where $O^p_{i,K-2}(x,w) = x_{p,i}$ is the ith element of the vector input, and $g_{N_K}(\bullet)$ and $g_{N_{K-1}}(\bullet)$ are two possibly different functions. In the special case where $g_{N_K}(\bullet)$ is a linear function and $g_{N_{K-1}}(\bullet)$ is a sigmoid function, the system simplifies to

$$O^p_{s,K}(x,w) = \sum_{j=1}^{N_{K-1}} g_{n_{K-1}}\big(\sum_{i=1}^{N_{K-2}} O^p_{i,K-2}(x,w)w^{k-2}_{i,j} + b_{j,K-1}\big)w^{k-1}_{j,s} + b_{s,K}\ , s = 1,..,N_K$$

which can be compared to the ordinary least squares regression

$$y^p_s = \alpha + \sum_{i=1}^{N_{K-2}} \beta_i x_{p,i} + \epsilon_s\ , s = 1,..,N_K$$

Hence, we see that the OLS regressor variables y^p_s are put through a transformation, or squashing function, given by $g_{N_{K-1}}(\bullet)$. By iterating within sample and examining the error terms, the neural network modify the input weights $\{w^{k-2}_{i,j}\}$ and the output weights $\{w^{k-1}_{j,s}\}$ to minimise within-sample measure for point prediction models. It can also maximise the classification rate of correct predictions for classification models.

12.3.2 Accounting for time and earning profits

In traditional time series forecasting, the criteria for assessing model performance are error functions based on the goodness of fit of target and predicted values. Since neural networks are similar to conventional regression estimators, except for their nonlinearity, error functions are also used to judge the goodness of fit of the model. However, normalised mean square error (NMSE) and other error functions are not the most appropriate measures in finance.

12.3.2.1 Time factor

For instance, weighting all data equally (ordinary leasy squares) are less accurate than discounted least squares, which weight the most recent data more heavily (see Makridakis et al. [1982]). Incorporating time factor to neural networks, Refenes et al. [1997] proposed the discounted least squares (DLS) neural network model. Assuming a feed-forward network with n input and a single output unit ($m = 1$) with k hidden layers, we let the ordinary least-squares criterion of the network be given by

$$E_{LS} = \frac{1}{2P}\sum_{p=1}^{P}(O_p - d_p)^2$$

where O_p and d_p denote the pth output and target value, respectively.

***Remark* 12.3.1** *In general, when considering a training set consiting of sequential data, or time-varying input, we represent them as $\{(\overline{x}_1, \overline{d}_1), ..., (\overline{x}_T, \overline{d}_T)\}$ with T ordered pairs. To be consistent with the articles described we keep the P ordered pairs, and we let $p = P$ be the most recent observation.*

Refenes et al. [1997] proposed a simple modification to the error backpropagation procedure taking into account gradually changing input-output pairs. The procedure is based on the principle of discounted least-squares whereby learning is biased towards more recent observations with long term effects experiencing exponential decay through time. The cumulative error calculated by the DLS procedure is given by

$$E_{DLS} = \frac{1}{2P}\sum_{p=1}^{P}\phi(p)(O_p - d_p)^2$$

where $\phi(p)$ is an adjustment of the contribution of observation p to the overall error. They chose the simple sigmoidal decay function

$$\phi_{DLS}(p) = \frac{1}{1+e^{a-bp}}$$

where $b = \frac{2a}{P}$ and a is the discount rate. The learning rule is derived in the usual way by repeatedly changing the weights by an amount proportional to

$$\frac{\partial E_{DLS}}{\partial W} = \phi_{DLS}(p)\frac{\partial E_{LS}}{\partial W}$$

since the discount factor $\phi(p)$ is a function of the recency of the observation which is independent of the actual order of pattern presentation within the learning process. Hence, the algorithm is not affected by randomising the order in which patterns are presented and can be applied to both batch and stochastic update rules. Using a sine wave with changing amplitude and a non-trivial application in exposure analysis of stock returns to multiple factors, Refenes et al. [1997] compared the performance of the two cost functions on a backpropagation network with a single hidden layer of 12 units and batch update. They showed that DLS was a more efficient procedure for weakly non-stationary data series.

12.3.2.2 Direction measures

Further, we are usualy more interested in earning profits than in the quality of the forecast itself. When classifying returns, it is possible to use a direction measure to test the number of times a prediction neural network predicts the direction of the predicted return movement correctly. Merton [1981] proposed a modified direction given by

$$\text{Modified Direction} = \frac{\sharp \text{ of correct up predictions}}{\sharp \text{ of times index up}} + \frac{\sharp \text{ of correct down predictions}}{\sharp \text{ of times index down}} - 1$$

In general, direction is defined as the number of times the prediction followed the up and down movement of the underlying stock/index. Harvey et al. [2002] and Castiglione et al. [2001] proposed the following direction measure

$$\xi = \frac{1}{|T|}\sum_{t\in T}\mathcal{H}(R_t\hat{R}_t) + 1 - (|R_t| + |\hat{R}_t|)$$

where $R_t = \frac{P_t}{P_{t-1}}$ is the percentage return on the underlying at time step $t \in T$, $\hat{R}_t$ is the forecasted percentage return, T is the number of days in the validation period, and $\mathcal{H}(\bullet)$ is the Heaviside function. Alternatively, we can write the direction measure as

$$\xi = \frac{1}{|T|}\sum_{t\in T}\mathcal{H}(\Delta P_t\Delta\hat{P}_t) + 1 - (|\Delta P_t| + |\Delta\hat{P}_{i+1}|)$$

where $\Delta P_t = P_t - P_{t-1}$ is the price change at step $t \in T$ and P_{t-1} is assumed to be known. These two direction measures provide a summary of how well the predicted time series and the actual ones move together at any point in time. However, we want to implement measurement errors taking into account the simultaneous behaviour of trend

and magnitude within the trend. Hence, to reflect this point when evaluating the performance of a forecasting model, Yao et al. [1996] used the correctness of trend and a paper profit. They proposed a profit based adjusted weight factor for backpropagation network training by adding a factor containing profit, direction, and time information to the error function.

12.3.2.3 Time dependent direction profit

In order to take profit gain into account, Caldwell [1995] proposed a weighted directional symmetry (WDS) function which penalise more heavily the incorrectly predicted directions than the correct ones. The cumulative error function is given by

$$E_{WDS} = \frac{100}{P}\sum_{p=1}^{P}\phi(p)|O_p - d_p|$$

with

$$\phi(p) = \begin{cases} g \text{ if } \Delta d_p \times \Delta O_p \leq 0 \\ h \text{ otherwise} \end{cases}$$

where $\Delta d_p = d_p - d_{p-1}$ and $\Delta O_p = O_p - O_{p-1}$, and g and h are constants or some function of d_p. The values $g = 1.5$ and $h = 0.5$ were suggested. Yao et al. [1996] argued that the profit driven procedure was not only a function of the direction, but also of the amount of change. That is, the penalty on WDS weights should be increased in case of a wrongly forecasted direction for a big change of values, and the penalty should be further reduced if the direction is correctly forecasted for a big change in value. They proposed a modified directional profit (DP) adjustment factor defined as

$$\phi_{DP}(p) = \begin{cases} a_1 \text{ if } \Delta d_p \times \Delta O_p > 0 \text{ and } |\Delta d_p| \leq \sigma \\ a_2 \text{ if } \Delta d_p \times \Delta O_p > 0 \text{ and } |\Delta d_p| > \sigma \\ a_3 \text{ if } \Delta d_p \times \Delta O_p < 0 \text{ and } |\Delta d_p| \leq \sigma \\ a_4 \text{ if } \Delta d_p \times \Delta O_p < 0 \text{ and } |\Delta d_p| > \sigma \end{cases}$$

where σ is a threshold for the changes in sample values generally taken as the standard deviation of the training set

$$\sigma^2 = \frac{1}{P}\sum_{p=1}^{P}(d_p - \mu_d)^2$$

where μ_d is the mean of the target series. The values $a_1 = 0.5$, $a_2 = 0.8$, $a_3 = 1.2$, and $a_4 = 1.5$ were suggested. While the DLS model focus on a time factor $\phi_{DLS}(p)$ and the DP model focus on a profit factor $\phi_{DP}(p)$, Yao et al. [1998] [2000] combined the two approaches obtaining a time dependent directional profit (TDP) model given by

$$\phi_{TDP}(p) = \phi_{DLS}(p) \times \phi_{DP}(p)$$

The three models, DLS, DP, and TDP were tested and compared to the benchmarked OLS model on four major Asian stock indices and the US Dow Jones Industrials Index (DJ). They considered a one hidden layer network with learning rate $\eta = 0.25$ and momentum rate $\gamma = 0.9$. The structure of the neural network was $30 - 10 - 1$, that is, 30 input, 10 nodes in the hidden layer and one output. Thus, 30 consecutive days of data are fed to the network to forecast the index of the following day. Considering a paper profit measure for the model performance, they showed that the preference of DLS over OLS was of 60%, and that the preference of DP over OLS was of 80%. In this combined model, they obtained up to 27.6% excess annual return above the OLS model with at least 2.8% excess annual return in the worst case. Lu Dang Khoa et al. [2006] argued that big changes in prices between two consecutive periods are rare to occur, leading to change between two consecutive prices smaller than the standard deviation of the price. They replaced the inequality $|\Delta d_p| > \sigma$ with $\frac{\Delta d_p}{d_{p-1}} > a_5$ and obtained the following adjustment factor

$$\phi_{DPN}(p) = \begin{cases} a_1 \text{ if } \Delta d_p \times \Delta O_p > 0 \text{ and } \frac{\Delta d_p}{d_{p-1}} < a_5 \\ a_2 \text{ if } \Delta d_p \times \Delta O_p > 0 \text{ and } \frac{\Delta d_p}{d_{p-1}} > a_5 \\ a_3 \text{ if } \Delta d_p \times \Delta O_p < 0 \text{ and } \frac{\Delta d_p}{d_{p-1}} < a_5 \\ a_4 \text{ if } \Delta d_p \times \Delta O_p < 0 \text{ and } \frac{\Delta d_p}{d_{p-1}} > a_5 \end{cases}$$

with $a_5 = 0.05$. They used that model to predict the $S\&P$ 500 stock index one month in the future without much improvements compared to the traditional FFN algorithm. They considered a mixed of input based on fundamental factors such as the prices of oil, the interest rates, inflation, and technical indicators such as volatility, relative strength index, directional index etc.

12.3.2.4 The problem with direction profit

A large number of studies using backpropagation neural networks to forecast financial time series are considering the directional profit (DP) or time dependent directional profit (TDP) models when defining the error function in the system (see Zhang et al. [2005], Lu Dang Khoa et al. [2006], Wang et al. [2006]). In these studies the function $\phi(\bullet)$ is assumed to be a penalising coefficient against the wrongly forecasted direction, which is independent of the actual order of pattern presentation within the learning process. As a result, the standard steepest descent on the error function is performed, and the weight change is given by

$$\Delta W(t+1) = -\phi_{TDP}(p)\eta\frac{\partial E_{LS}}{\partial W} + \gamma\Delta W(t)$$

However, the step functions $\phi(\bullet)$ above can be decomposed into a series of Heaviside function $\mathcal{H}(\bullet)$ with the help of Equation (A.4.2) where the variable is ΔO_p. Hence, the condition for switching on and off the error function depends on the output values O_p and O_{p-1}, which are function of the weights of the network. Even though it is less of an issue when using a global optimiser (see Dash et al. [2003]), the step functions $\phi(\bullet)$ must be differentiated with respect to the weights when computing the gradient of the error function in order to perform the steepest descent. In that setting the learning rule is modified, and the weight change for the pth training pair is given by

$$\Delta W(t+1) = -\eta\phi_{TDP}(\Delta O_p)\frac{\partial E_p}{\partial W} - \eta\frac{1}{2}\frac{\partial\phi_{TDP}(\Delta O_p)}{\partial W}(O_p - d_p)^2 + \gamma\Delta W(t)$$

Unfortunately, the derivative of the Heaviside function is the Dirac delta function which is zero for $\Delta d_p \times \Delta O_p \neq 0$ and infinite at the point $\Delta d_p \times \Delta O_p = 0$. One solution consists in approximating the Heaviside function $\phi(\bullet)$ with a differentiable function $\psi(\bullet)$, such as a sigmoid function, and properly derive the gradient of the error function. We choose the translated sigmoid $f_{b,c}(\bullet)$ described in Section (13.6.3.3) with $b_L = 1.4$ and $b_H = 0.2$ and display the curves in Figure (12.2) and their derivatives in Figure (12.3) with $c = \frac{1}{p}$ for $p = 0.5, 1, 1.5$.

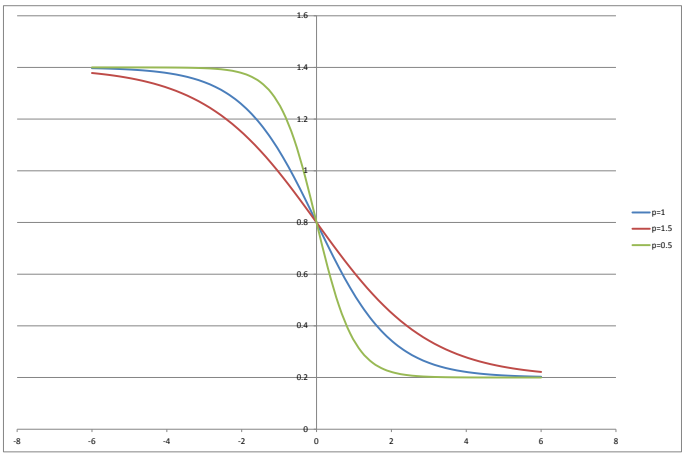


Figure 12.2: Translated logistic function.

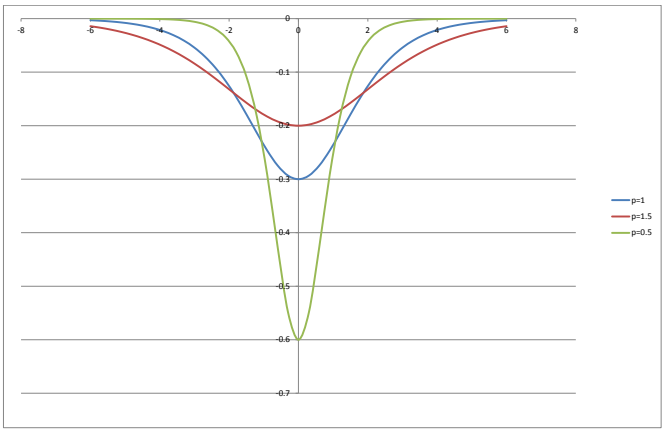


Figure 12.3: Derivative of the translated logistic function.

12.3.3 Minimising the error function with direction profit

We follow the notation in Section (13.6.1.2) and consider the multilayer network described in Section (13.6.4) in the case of $p = 1, .., P$ input-output pair. Assuming a feed-forward network with n input and m output units with k hidden layers, we let the error function of the network be given by

$$E = \frac{1}{2} \sum_{p=1}^{P} \sum_{j=1}^{m} \psi(Z_j^p)(O_j^p - d_j^p)^2$$

where O_j^p and d_j^p denote the jth component of the output vector $\overline{O}_p$ and the target vector $\overline{d}_p$, respectively. Further, we define the spread for the jth output of the pth training sample as

$$Z_j^p = f(O_j^p, O_j^{p-1})$$

To take profit gain into account, we set the spread as $Z_j^p = \alpha(O_j^p - O_j^{p-1})$ with $\alpha = \Delta d_j^p$. In addition we assume that the function $\psi(\bullet)$ at each node of the network is continuous and differentiable given by the sigmoid $f_{b,c}(\bullet)$. We want to minimise E by using an iterative process of gradient descent where we need to compute the gradient ∇E. Each weight is updated by using the increment

$$\Delta w_{i,j} = w_{i,j}^{l+1} - w_{i,j}^l = -\eta \frac{\partial E}{\partial w_{i,j}^l}$$

where l is the iteration index, and η is a learning rate. In that setting, we define the output function of the jth node for the kth layer as

$$O_{j,k}(x, w) = g(\tilde{A}_{j,k-1}(x, w)) \, , j = 1, .., n_k$$

where $\tilde{A}_{j,k-1}(x, w) = A_{j,k-1}(x, w) + b_{j,k}$ with $b_{j,k}$ a constant, and we let the corresponding activation function satisfies

$$A_{j,k-1}(x, w) = \sum_{i=1}^{n_{k-1}} O_{i,k-1}(x, w) w_{i,j}^{k-1} \text{ and } O_{j,0}(x, w) = A_j(x, w) = \sum_{i=1}^{n_0} x_i w_{i,j}^0$$

where $w_{i,j}^{k-1}$ is the weight going from the ith node in layer $k-1$ to the jth node in layer k. Note, since $\frac{d\tilde{A}_{j,k-1}(x,w)}{dA_{j,k-1}(x,w)} = 1$ we get $\frac{\partial O_{j,k}(x,w)}{\partial A_{j,k-1}(x,w)} = \frac{\partial O_{j,k}(x,w)}{\partial \tilde{A}_{j,k-1}(x,w)}$.

12.3.3.1 The output layer

We start with the output layer k and compute the gradient $\nabla E(w_{i,j}^{k-1})$ for the weight $w_{i,j}^{k-1}$ going from the ith node in layer $k-1$ to the jth node in layer k. From the definition of the total error, the gradient satisfies

$$\nabla E_p(w_{i,j}^{k-1}) = \psi(Z_{j,k}^p)\delta_{j,k}^p \frac{\partial}{\partial w_{i,j}^{k-1}}(O_{j,k}^p - d_j) + \frac{\partial}{\partial Z_{j,k}^p}\psi(Z_{j,k}^p)\frac{\partial Z_{j,k}^p}{\partial w_{i,j}^{k-1}}(\delta_{j,k}^p)^2$$

where $\delta_{j,k}^p = (O_{j,k}^p - d_j)$. Note, to simplify notation we define

$$\epsilon_Z = \frac{\partial}{\partial Z_{j,k}^p}\psi(Z_{j,k}^p)(\delta_{j,k}^p)^2$$

where the index p is fixed for a given training pair. Differentiating the output function with respect to the weight, we get

$$\frac{\partial O_{j,k}^p}{\partial w_{i,j}^{k-1}} = \frac{\partial O_{j,k}^p}{\partial A_{j,k-1}^p}\frac{\partial A_{j,k-1}^p}{\partial w_{i,j}^{k-1}}$$

where $\frac{\partial O_{j,k}^p}{\partial A_{j,k-1}^p} = \frac{\partial}{\partial \tilde{A}_{j,k-1}^p} g(\tilde{A}_{j,k-1}^p(x, w))$. From the definition of $A_{j,k-1}^p$ it simplifies to

$$\frac{\partial O_{j,k}^p}{\partial w_{i,j}^{k-1}} = \frac{\partial O_{j,k}^p}{\partial A_{j,k-1}^p} O_{i,k-1}^p$$

Differentiating the spread function with respect to the weight, we get

$$\frac{\partial Z^p_{j,k}}{\partial w^{k-1}_{i,j}} = \alpha \frac{\partial O^p_{j,k}}{\partial w^{k-1}_{i,j}} - \alpha \frac{\partial O^{p-1}_{j,k}}{\partial w^{k-1}_{i,j}} = \alpha \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} O^p_{i,k-1} - \alpha \frac{\partial O^{p-1}_{j,k}}{\partial A^{p-1}_{j,k-1}} O^{p-1}_{i,k-1}$$

Replacing these terms in the gradient, we get

$$\nabla E_p(w^{k-1}_{i,j}) = \psi(Z^p_{j,k})\delta^p_{j,k} \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} O^p_{i,k-1} + \epsilon_Z \alpha \big(\frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} O^p_{i,k-1} - \frac{\partial O^{p-1}_{j,k}}{\partial A^{p-1}_{j,k-1}} O^{p-1}_{i,k-1} \big)$$

We let $e^p_{j,k} = \psi(Z^p_{j,k})\delta^p_{j,k} \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}}$ and $e^p_{Z,j,k} = \epsilon_Z \alpha \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}}$ be the error signals and rewrite the gradient as

$$\nabla E_p(w^{k-1}_{i,j}) = e^p_{j,k} O^p_{i,k-1} + e^p_{Z,j,k} O^p_{i,k-1} - e^{p-1}_{Z,j,k} O^{p-1}_{i,k-1}$$

so that the stochastic gradient descent rule for output units bebome

$$\Delta w^{k-1}_{i,j} = -\eta \nabla E_p(w^{k-1}_{i,j}) = -\eta e^p_{j,k} O^p_{i,k-1} - \eta e^p_{Z,j,k} O^p_{i,k-1} + \eta e^{p-1}_{Z,j,k} O^{p-1}_{i,k-1}$$

We now concentrate on the derivative of the error function E_p with respect to the constant $b_{j,k}$. From the definition of the total error, the gradient satisfies

$$\nabla E_p(b_{j,k}) = \psi(Z^p_{j,k})\delta^p_{j,k} \frac{\partial}{\partial b_{j,k}} (O^p_{j,k} - d_j) + \epsilon_Z \frac{\partial Z^p_{j,k}}{\partial b_{j,k}}$$

Differentiating the output function with respect to the constant $b_{j,k}$, we get

$$\frac{\partial O^p_{j,k}}{\partial b_{j,k}} = \frac{\partial O^p_{j,k}}{\partial \tilde{A}^p_{j,k-1}} \frac{\partial \tilde{A}^p_{j,k-1}}{\partial b_{j,k}} = \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}}$$

and differentiating the spread function with respect to the constant $b_{j,k}$, we get

$$\frac{\partial Z^p_{j,k}}{\partial b_{j,k}} = \alpha \frac{\partial O^p_{j,k}}{\partial b_{j,k}} - \alpha \frac{\partial O^{p-1}_{j,k}}{\partial b_{j,k}} = \alpha \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} - \alpha \frac{\partial O^{p-1}_{j,k}}{\partial A^{p-1}_{j,k-1}}$$

Replacing these terms in the gradient, we get

$$\nabla E_p(b_{j,k}) = \psi(Z^p_{j,k})\delta^p_{j,k} \frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} + \epsilon_Z \alpha \big(\frac{\partial O^p_{j,k}}{\partial A^p_{j,k-1}} - \frac{\partial O^{p-1}_{j,k}}{\partial A^{p-1}_{j,k-1}} \big)$$

which simplifies to

$$\nabla E_p(b_{j,k}) = e^p_{j,k} + e^p_{Z,j,k} - e^{p-1}_{Z,j,k}$$

We observe that the gradient with respect to the constant, $\nabla E_p(b_{j,k})$, is similar to that with respect to the weight, $\nabla E_p(w^{k-1}_{i,j})$, except for the link with the input values $O^p_{i,k-1}$ and $O^{p-1}_{i,k-1}$.

12.3.3.2 The first hidden layer

The next step is to consider the hidden layer $k-1$ and compute the gradient for the weight $w^{k-2}_{i,j}$ going from the ith node on layer $k-2$ to the jth node in layer $k-1$. Note, since the subscripts i and j are taken, for notation purpose, we use the subscript s to represent the nodes on the kth layer. The gradient $\nabla E(w^{k-2}_{i,j})$ for the weight $w^{k-2}_{i,j}$ becomes

$$\nabla E_p(w^{k-2}_{i,j}) = \sum_{s=1}^{n_k} \psi(Z^p_{s,k})\delta^p_{s,k} \frac{\partial}{\partial w^{k-2}_{i,j}} (O^p_{s,k} - d_s) + \sum_{s=1}^{n_k} \frac{\partial}{\partial Z^p_{s,k}} \psi(Z^p_{s,k}) \frac{\partial Z^p_{s,k}}{\partial w^{k-2}_{i,j}} (\delta^p_{s,k})^2$$

where $\delta^p_{s,k} = (O^p_{s,k} - d_s)$ and $\epsilon_{Z,s,k} = \frac{\partial}{\partial Z^p_{s,k}}\psi(Z^p_{s,k})(\delta^p_{s,k})^2$. Differentiating the output function with respect to the weight, we get

$$\frac{\partial O^p_{s,k}}{\partial w^{k-2}_{i,j}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} \frac{\partial A^p_{s,k-1}}{\partial w^{k-2}_{i,j}}$$

Given $A_{s,k-1}(x,w) = \sum_{j=1}^{n_{k-1}} O_{j,k-1}(x,w) w^{k-1}_{j,s}$, then

$$\frac{\partial A^p_{s,k-1}}{\partial w^{k-2}_{i,j}} = \frac{\partial}{\partial w^{k-2}_{i,j}} O^p_{j,k-1} w^{k-1}_{j,s}$$

Since $O_{j,k-1}(x,w) = g(A_{j,k-2}(x,w) + b_{j,k-1})$, we get

$$\frac{\partial O^p_{j,k-1}}{\partial w^{k-2}_{i,j}} = \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} \frac{\partial A^p_{j,k-2}}{\partial w^{k-2}_{i,j}}$$

and the derivative of the output function becomes

$$\frac{\partial O^p_{s,k}}{\partial w^{k-2}_{i,j}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} \frac{\partial A^p_{j,k-2}}{\partial w^{k-2}_{i,j}}$$

Further, given $A^p_{j,k-2}(x,w) = \sum_{i=1}^{n_{k-2}} O^p_{i,k-2}(x,w) w^{k-2}_{i,j}$, the derivative of the output function simplifies to

$$\frac{\partial O^p_{s,k}}{\partial w^{k-2}_{i,j}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} O^p_{i,k-2}$$

We can then differentiate the spread function with respect to the weight, getting

$$\frac{\partial Z^p_{s,k}}{\partial w^{k-2}_{i,j}} = \alpha \frac{\partial O^p_{s,k}}{\partial w^{k-2}_{i,j}} - \alpha \frac{\partial O^{p-1}_{s,k}}{\partial w^{k-2}_{i,j}} = \alpha \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} O^p_{i,k-2} - \alpha \frac{\partial O^{p-1}_{s,k}}{\partial A^{p-1}_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^{p-1}_{j,k-1}}{\partial A^{p-1}_{j,k-2}} O^{p-1}_{i,k-2}$$

Replacing in the gradient term, we get

$$\nabla E_p(w^{k-2}_{i,j}) = \sum_{s=1}^{n_k} e^p_{s,k} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} O^p_{i,k-2} + \sum_{s=1}^{n_k} \epsilon_{Z,s,k} \alpha \Big(\frac{\partial O^p_{s,k}}{\partial w^{k-2}_{i,j}} - \frac{\partial O^{p-1}_{s,k}}{\partial w^{k-2}_{i,j}} \Big)$$

We then set $\delta^p_{j,k-1} = \sum_{s=1}^{n_k} e^p_{s,k} w^{k-1}_{j,s}$ and $\delta^p_{Z,j,k-1} = \sum_{s=1}^{n_k} e^p_{Z,s,k} w^{k-1}_{j,s}$, and we express the error signal as

$$e^p_{j,k-1} = \delta^p_{j,k-1} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} \text{ and } e^p_{Z,j,k-1} = \delta^p_{Z,j,k-1} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}}$$

so that the gradient becomes

$$\nabla E_p(w^{k-2}_{i,j}) = e^p_{j,k-1} O^p_{i,k-2} + e^p_{Z,j,k-1} O^p_{i,k-2} - e^{p-1}_{Z,j,k-1} O^{p-1}_{i,k-2}$$

We now concentrate on the derivative of the error function E_p with respect to the constant $b_{j,k-1}$. The gradient $\nabla E(b_{j,k-1})$ for the constant $b_{j,k-1}$ becomes

$$\nabla E_p(b_{j,k-1}) = \sum_{s=1}^{n_k} \psi(Z^p_{s,k}) \delta^p_{s,k} \frac{\partial}{\partial b_{j,k-1}} (O^p_{s,k} - d_s) + \sum_{s=1}^{n_k} \epsilon_{Z,s,k} \psi(Z^p_{s,k}) \frac{\partial Z^p_{s,k}}{\partial b_{j,k-1}}$$

Differentiating the output function with respect to the constant, we get

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-1}} = \frac{\partial O^p_{s,k}}{\partial \tilde{A}^p_{s,k-1}} \frac{\partial \tilde{A}^p_{s,k-1}}{\partial b_{j,k-1}}$$

Given $\tilde{A}_{s,k-1}(x,w) = \sum_{j=1}^{n_{k-1}} O_{j,k-1}(x,w) w^{k-1}_{j,s} + b_{s,k}$, then

$$\frac{\partial \tilde{A}^p_{s,k-1}}{\partial b_{j,k-1}} = \frac{\partial}{\partial b_{j,k-1}} O^p_{j,k-1} w^{k-1}_{j,s}$$

Since $O_{j,k-1}(x,w) = g(A_{j,k-2}(x,w) + b_{j,k-1})$, we get

$$\frac{\partial O^p_{j,k-1}}{\partial b_{j,k-1}} = \frac{\partial O^p_{j,k-1}}{\partial \tilde{A}^p_{j,k-2}} \frac{\partial \tilde{A}^p_{j,k-2}}{\partial b_{j,k-1}} = \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}}$$

and the derivative of the output function becomes

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-1}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}}$$

Differentiating the spread function with respect to the constant, we get

$$\frac{\partial Z^p_{s,k}}{\partial b_{j,k-1}} = \alpha \frac{\partial O^p_{s,k}}{\partial b_{j,k-1}} - \alpha \frac{\partial O^{p-1}_{s,k}}{\partial b_{j,k-1}} = \alpha \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} - \alpha \frac{\partial O^{p-1}_{s,k}}{\partial A^{p-1}_{s,k-1}} w^{k-1}_{j,s} \frac{\partial O^{p-1}_{j,k-1}}{\partial A^{p-1}_{j,k-2}}$$

Replacing in the gradient term, we get

$$\nabla E_p(b_{j,k-1}) = \sum_{s=1}^{n_k} e^p_{s,k} w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} + \sum_{s=1}^{n_k} e^p_{Z,s,k} \big(w^{k-1}_{j,s} \frac{\partial O^p_{j,k-1}}{\partial A^p_{j,k-2}} - w^{k-1}_{j,s} \frac{\partial O^{p-1}_{j,k-1}}{\partial A^{p-1}_{j,k-2}} \big)$$

and after simplification, the gradient becomes

$$\nabla E_p(b_{j,k-1}) = e^p_{j,k-1} + e^p_{Z,j,k-1} - e^{p-1}_{Z,j,k-1}$$

Again, the gradient with respect to the constant, $\nabla E_p(b_{j,k-1})$, is similar to that with respect to the weight, $\nabla E_p(w^{k-2}_{i,j})$, except for the link with the input values $O^p_{i,k-2}$ and $O^{p-1}_{i,k-2}$.

12.3.3.3 The next hidden layer

Going one step backward, we consider the hidden layer $k-2$ and compute the gradient for the weight $w^{k-3}_{i,j}$ going from the ith node on layer $k-3$ to the jth node in layer $k-2$. We will also assume that it corresponds to the input layer. Note, since the subscripts i and j are taken, for notation purpose, we use the subscript s to represent the nodes on the kth layer and t to represent the nodes on the $(k-1)$th layer. The gradient $\nabla E(w^{k-2}_{i,j})$ for the weight $w^{k-2}_{i,j}$ becomes

$$\nabla E_p(w^{k-3}_{i,j}) = \sum_{s=1}^{n_k} \psi(Z^p_{s,k}) \delta^p_{s,k} \frac{\partial}{\partial w^{k-3}_{i,j}} (O^p_{s,k} - d_s) + \sum_{s=1}^{n_k} \frac{\partial}{\partial Z^p_{s,k}} \psi(Z^p_{s,k}) \frac{\partial Z^p_{s,k}}{\partial w^{k-3}_{i,j}} (\delta^p_{s,k})^2$$

where $\delta^p_{s,k} = (O^p_{s,k} - d_s)$ and $\epsilon_{Z,s,k} = \frac{\partial}{\partial Z^p_{s,k}} \psi(Z^p_{s,k}) (\delta^p_{s,k})^2$. Differentiating the output function with respect to the weight, we get

$$\frac{\partial O^p_{s,k}}{\partial w^{k-3}_{i,j}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} \frac{\partial A^p_{s,k-1}}{\partial w^{k-3}_{i,j}}$$

Given $A_{s,k-1}(x,w) = \sum_{t=1}^{n_{k-1}} O_{t,k-1}(x,w) w_{t,s}^{k-1}$, then

$$\frac{\partial A_{s,k-1}^p}{\partial w_{i,j}^{k-3}} = \sum_{t=1}^{n_{k-1}} \frac{\partial}{\partial w_{i,j}^{k-3}} O_{t,k-1}^p w_{t,s}^{k-1}$$

Replacing in the above equation, we get

$$\frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} = \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial}{\partial w_{i,j}^{k-3}} O_{t,k-1}^p w_{t,s}^{k-1}$$

Since $O_{t,k-1}(x,w) = g(A_{t,k-2}(x,w) + b_{t,k-1})$, we get

$$\frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} = \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} \frac{\partial A_{t,k-2}^p}{\partial w_{i,j}^{k-3}} w_{t,s}^{k-1}$$

Given $A_{t,k-2}^p(x,w) = \sum_{j=1}^{n_{k-2}} O_{j,k-2}^p(x,w) w_{j,t}^{k-2}$, the derivative of the output function becomes

$$\frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} = \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^p}{\partial w_{i,j}^{k-3}} w_{j,t}^{k-2}$$

Further, $O_{j,k-2}(x,w) = g(A_{j,k-3}(x,w) + b_{j,k-2})$, and the derivative of the output function becomes

$$\frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} = \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} w_{j,t}^{k-2} \frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} \frac{\partial A_{j,k-3}^p}{\partial w_{i,j}^{k-3}}$$

At last, $A_{j,k-3}^p(x,w) = \sum_{i=1}^{n_{k-3}} O_{i,k-3}^p(x,w) w_{i,j}^{k-3}$, and the derivative of the output function simplifies to

$$\frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} = \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} w_{j,t}^{k-2} O_{i,k-3}^p(x,w)$$

Again, the derivative of the spread function with respect to the weight is given by

$$\begin{aligned} \frac{\partial Z_{s,k}^p}{\partial w_{i,j}^{k-3}} = \alpha \frac{\partial O_{s,k}^p}{\partial w_{i,j}^{k-3}} - \alpha \frac{\partial O_{s,k}^{p-1}}{\partial w_{i,j}^{k-3}} &= \alpha \frac{\partial O_{s,k}^p}{\partial A_{s,k-1}^p} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} w_{j,t}^{k-2} O_{i,k-3}^p(x,w) \\ &- \alpha \frac{\partial O_{s,k}^{p-1}}{\partial A_{s,k-1}^{p-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^{p-1}}{\partial A_{t,k-2}^{p-1}} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^{p-1}}{\partial A_{j,k-3}^{p-1}} w_{j,t}^{k-2} O_{i,k-3}^{p-1}(x,w) \end{aligned}$$

Replacing in the gradient term, we get

$$\begin{aligned} \nabla E_p(w_{i,j}^{k-3}) &= \sum_{s=1}^{n_k} e_{s,k}^p \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} w_{j,t}^{k-2} O_{i,k-3}^p(x,w) \\ &+ \sum_{s=1}^{n_k} e_{Z,s,k}^p \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} w_{j,t}^{k-2} O_{i,k-3}^p(x,w) \\ &- \sum_{s=1}^{n_k} e_{Z,s,k}^{p-1} \sum_{t=1}^{n_{k-1}} \frac{\partial O_{t,k-1}^{p-1}}{\partial A_{t,k-2}^{p-1}} w_{t,s}^{k-1} \frac{\partial O_{j,k-2}^{p-1}}{\partial A_{j,k-3}^{p-1}} w_{j,t}^{k-2} O_{i,k-3}^{p-1}(x,w) \end{aligned}$$

which we can rewrite as

$$\begin{aligned}\nabla E_p(w_{i,j}^{k-3}) &= \sum_{t=1}^{n_{k-1}}\sum_{s=1}^{n_k} e_{s,k}^p w_{t,s}^{k-1}\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p}w_{j,t}^{k-2}O_{i,k-3}^p(x,w)\\ &+ \sum_{t=1}^{n_{k-1}}\sum_{s=1}^{n_k} e_{Z,s,k}^p w_{t,s}^{k-1}\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p}w_{j,t}^{k-2}O_{i,k-3}^p(x,w)\\ &- \sum_{t=1}^{n_{k-1}}\sum_{s=1}^{n_k} e_{Z,s,k}^{p-1} w_{t,s}^{k-1}\frac{\partial O_{t,k-1}^{p-1}}{\partial A_{t,k-2}^{p-1}}\frac{\partial O_{j,k-2}^{p-1}}{\partial A_{j,k-3}^{p-1}}w_{j,t}^{k-2}O_{i,k-3}^{p-1}(x,w)\end{aligned}$$

which leads to

$$\begin{aligned}\nabla E_p(w_{i,j}^{k-3}) &= \sum_{t=1}^{n_{k-1}} \delta_{t,k-1}^p\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}w_{j,t}^{k-2}\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p}O_{i,k-3}^p(x,w)\\ &+ \sum_{t=1}^{n_{k-1}} \delta_{Z,t,k-1}^p\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p}w_{j,t}^{k-2}O_{i,k-3}^p(x,w)\\ &- \sum_{t=1}^{n_{k-1}} \delta_{Z,t,k-1}^{p-1}\frac{\partial O_{t,k-1}^{p-1}}{\partial A_{t,k-2}^{p-1}}\frac{\partial O_{j,k-2}^{p-1}}{\partial A_{j,k-3}^{p-1}}w_{j,t}^{k-2}O_{i,k-3}^{p-1}(x,w)\end{aligned}$$

Setting

$$\begin{aligned}\delta_{j,k-2}^p &= \sum_{t=1}^{n_{k-1}} \delta_{t,k-1}^p\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}w_{j,t}^{k-2} = \sum_{t=1}^{n_{k-1}} e_{t,k-1}^p w_{j,t}^{k-2}\\ \delta_{Z,j,k-2}^p &= \sum_{t=1}^{n_{k-1}} \delta_{Z,t,k-1}^p\frac{\partial O_{t,k-1}^p}{\partial A_{t,k-2}^p}w_{j,t}^{k-2} = \sum_{t=1}^{n_{k-1}} e_{Z,t,k-1}^p w_{j,t}^{k-2}\end{aligned}$$

we express the error signals as

$$e_{j,k-2}^p = \delta_{j,k-2}^p\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p} \text{ and } e_{Z,j,k-2}^p = \delta_{Z,j,k-2}^p\frac{\partial O_{j,k-2}^p}{\partial A_{j,k-3}^p}$$

so that the gradient becomes

$$\nabla E_p(w_{i,j}^{k-3}) = e_{j,k-2}^p O_{i,k-3}^p + e_{Z,j,k-2}^p O_{i,k-3}^p - e_{Z,j,k-2}^{p-1}O_{i,k-3}^{p-1}$$

We now concentrate on the derivative of the error function E_p with respect to the constant $b_{j,k-2}$. The gradient $\nabla E(b_{j,k-2})$ for the constant $b_{j,k-2}$ becomes

$$\nabla E_p(b_{j,k-2}) = \sum_{s=1}^{n_k}\psi(Z_{s,k}^p)\delta_{s,k}^p\frac{\partial}{\partial b_{j,k-2}}(O_{s,k}^p - d_s) + \sum_{s=1}^{n_k}\epsilon_{Z,s,k}\psi(Z_{s,k}^p)\frac{\partial Z_{s,k}^p}{\partial b_{j,k-2}}$$

Differentiating the output function with respect to the constant, we get

$$\frac{\partial O_{s,k}^p}{\partial b_{j,k-2}} = \frac{\partial O_{s,k}^p}{\partial \tilde{A}_{s,k-1}^p}\frac{\partial \tilde{A}_{s,k-1}^p}{\partial b_{j,k-2}}$$

Given $\tilde{A}_{s,k-1}(x,w) = \sum_{t=1}^{n_{k-1}} O_{t,k-1}(x,w)w_{t,s}^{k-1} + b_{s,k}$, then

$$\frac{\partial \tilde{A}^p_{s,k-1}}{\partial b_{j,k-2}} = \sum_{t=1}^{n_{k-1}} \frac{\partial}{\partial b_{j,k-2}} O^p_{t,k-1} w^{k-1}_{t,s}$$

Replacing in the above equation, we get

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-2}} = \frac{\partial O^p_{s,k}}{\partial \tilde{A}^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial}{\partial b_{j,k-2}} O^p_{t,k-1} w^{k-1}_{t,s}$$

Since $O_{t,k-1}(x,w) = g(A_{t,k-2}(x,w) + b_{t,k-1})$, we get

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-2}} = \frac{\partial O^p_{s,k}}{\partial \tilde{A}^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^p_{t,k-1}}{\partial \tilde{A}^p_{t,k-2}} \frac{\partial \tilde{A}^p_{t,k-2}}{\partial b_{j,k-2}} w^{k-1}_{t,s}$$

Given $A^p_{t,k-2}(x,w) = \sum_{j=1}^{n_{k-2}} O^p_{j,k-2}(x,w) w^{k-2}_{j,t}$, the derivative of the output function becomes

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-2}} = \frac{\partial O^p_{s,k}}{\partial \tilde{A}^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^p_{t,k-1}}{\partial \tilde{A}^p_{t,k-2}} w^{k-1}_{t,s} \frac{\partial O^p_{j,k-2}}{\partial b_{j,k-2}} w^{k-2}_{j,t}$$

Further, $O_{j,k-2}(x,w) = g(A_{j,k-3}(x,w) + b_{j,k-2})$, and the derivative of the output function becomes

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-2}} = \frac{\partial O^p_{s,k}}{\partial \tilde{A}^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^p_{t,k-1}}{\partial \tilde{A}^p_{t,k-2}} w^{k-1}_{t,s} w^{k-2}_{j,t} \frac{\partial O^p_{j,k-2}}{\partial \tilde{A}^p_{j,k-3}} \frac{\partial \tilde{A}^p_{j,k-3}}{\partial b_{j,k-2}}$$

which simplifies to

$$\frac{\partial O^p_{s,k}}{\partial b_{j,k-2}} = \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^p_{t,k-1}}{\partial A^p_{t,k-2}} w^{k-1}_{t,s} w^{k-2}_{j,t} \frac{\partial O^p_{j,k-2}}{\partial A^p_{j,k-3}}$$

Similarly, the derivative of the spread function with respect to the weight is given by

$$\frac{\partial Z^p_{s,k}}{\partial b_{j,k-2}} = \alpha \frac{\partial O^p_{s,k}}{\partial A^p_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^p_{t,k-1}}{\partial A^p_{t,k-2}} w^{k-1}_{t,s} \frac{\partial O^p_{j,k-2}}{\partial A^p_{j,k-3}} w^{k-2}_{j,t} - \alpha \frac{\partial O^{p-1}_{s,k}}{\partial A^{p-1}_{s,k-1}} \sum_{t=1}^{n_{k-1}} \frac{\partial O^{p-1}_{t,k-1}}{\partial A^{p-1}_{t,k-2}} w^{k-1}_{t,s} \frac{\partial O^{p-1}_{j,k-2}}{\partial A^{p-1}_{j,k-3}} w^{k-2}_{j,t}$$

and after simplification the gradient becomes

$$\nabla E_p(b_{j,k-2}) = e^p_{j,k-2} + e^p_{Z,j,k-2} - e^{p-1}_{Z,j,k-2}$$

12.3.4 Some results

As an example, we consider a training sample made of $P = 50$ pairs with $n = 4$ input $X(t), X(t-1), X(t-2), X(t-3)$ and $m = 1$ output. The input and target value $d(t)$ are described in Table (12.1) and Table (12.2), where the index $1, 2, .., 50$ corresponds to time $t, t-1, ..., t-49$. The training sample is plotted in Figure (12.4). We use the translated sigmoid $f_{b,c}(\bullet)$ described in Section (13.6.3.3) with $b_L = -1.5$ and $b_H = 1.5$. Considering a fixed time window, we test the forecasting power of the model on an out of sample series made of 75 values. We estimate the point forecast $O(t+h)$ for $h = 1$ and display the results in Table (12.3) with index $1, 2, ..., 75$ corresponding to time $t+1, t+2, ..., t+75$. Accounting for the continually changing nature of financial relationships, we deliberately choose an out of sample series having a very different regime of volatility than the training sample. The true series and the output of the model are plotted in Figure (12.5).

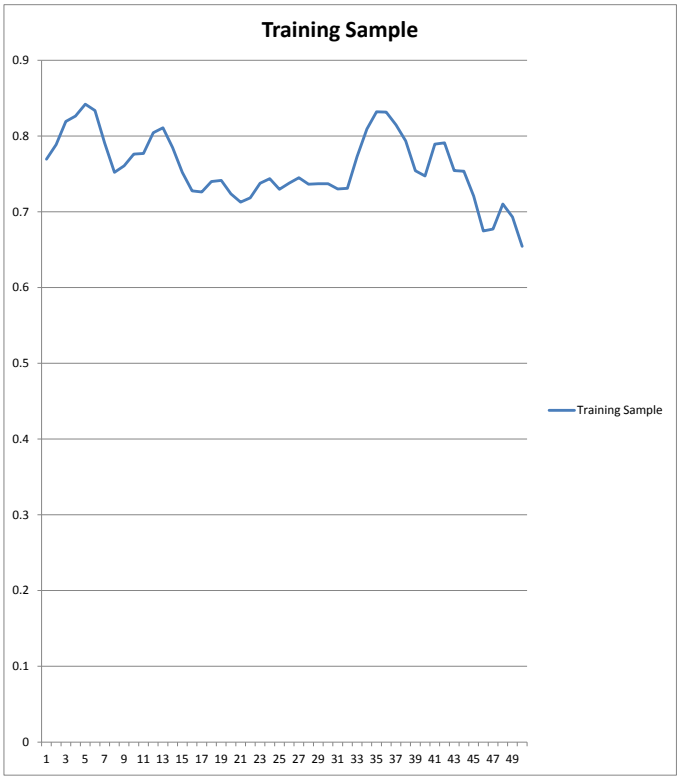


Figure 12.4: Training sample

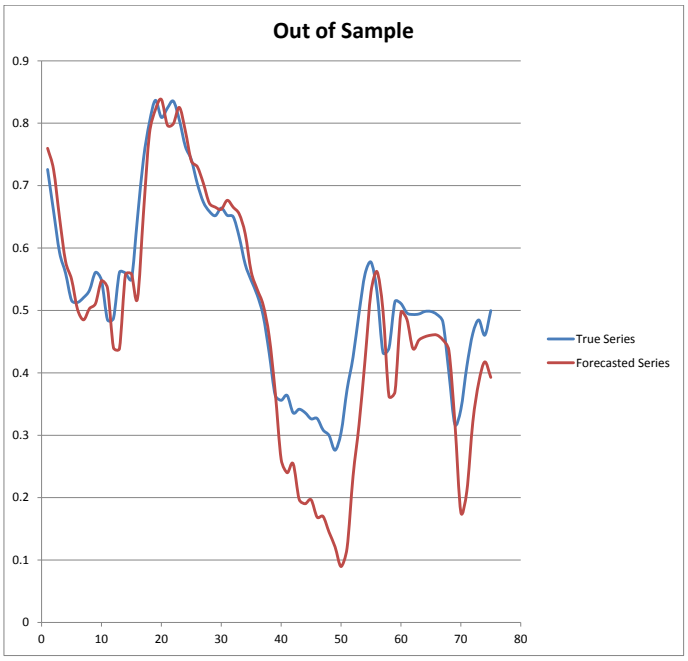


Figure 12.5: Out of sample

Index	$X(t)$	$X(t-1)$	$X(t-2)$	$X(t-3)$	$d(t)$
1	0.788672545	0.819228865	0.826253555	0.841926068	0.769329565
2	0.819228865	0.826253555	0.841926068	0.833534796	0.788672545
3	0.826253555	0.841926068	0.833534796	0.790644945	0.819228865
4	0.841926068	0.833534796	0.790644945	0.752050429	0.826253555
5	0.833534796	0.790644945	0.752050429	0.760675391	0.841926068
6	0.790644945	0.752050429	0.760675391	0.776147235	0.833534796
7	0.752050429	0.760675391	0.776147235	0.776933401	0.790644945
8	0.760675391	0.776147235	0.776933401	0.804244724	0.752050429
9	0.776147235	0.776933401	0.804244724	0.810720748	0.760675391
10	0.776933401	0.804244724	0.810720748	0.78459947	0.776147235
11	0.804244724	0.810720748	0.78459947	0.752037728	0.776933401
12	0.810720748	0.78459947	0.752037728	0.727641188	0.804244724
13	0.78459947	0.752037728	0.727641188	0.726119659	0.810720748
14	0.752037728	0.727641188	0.726119659	0.739826123	0.78459947
15	0.727641188	0.726119659	0.739826123	0.741280339	0.752037728
16	0.726119659	0.739826123	0.741280339	0.723616375	0.727641188
17	0.739826123	0.741280339	0.723616375	0.712822154	0.726119659
18	0.741280339	0.723616375	0.712822154	0.718273242	0.739826123
19	0.723616375	0.712822154	0.718273242	0.737587011	0.741280339
20	0.712822154	0.718273242	0.737587011	0.74372266	0.723616375
21	0.718273242	0.737587011	0.74372266	0.729752024	0.712822154
22	0.737587011	0.74372266	0.729752024	0.737747038	0.718273242
23	0.74372266	0.729752024	0.737747038	0.744911434	0.737587011
24	0.729752024	0.737747038	0.744911434	0.736302982	0.74372266
25	0.737747038	0.744911434	0.736302982	0.737085338	0.729752024
26	0.744911434	0.736302982	0.737085338	0.737085338	0.737747038
27	0.736302982	0.737085338	0.737085338	0.729985715	0.744911434
28	0.737085338	0.737085338	0.729985715	0.730823953	0.736302982
29	0.737085338	0.729985715	0.730823953	0.772753641	0.737085338
30	0.729985715	0.730823953	0.772753641	0.809067133	0.737085338

Table 12.1: Training sample: part I

Index	$X(t)$	$X(t-1)$	$X(t-2)$	$X(t-3)$	$d(t)$
31	0.730823953	0.772753641	0.809067133	0.831910392	0.729985715
32	0.772753641	0.809067133	0.831910392	0.831573827	0.730823953
33	0.809067133	0.831910392	0.831573827	0.814590614	0.772753641
34	0.831910392	0.831573827	0.814590614	0.79339843	0.809067133
35	0.831573827	0.814590614	0.79339843	0.75406093	0.831910392
36	0.814590614	0.79339843	0.75406093	0.747257231	0.831573827
37	0.79339843	0.75406093	0.747257231	0.789236451	0.814590614
38	0.75406093	0.747257231	0.789236451	0.790877365	0.79339843
39	0.747257231	0.789236451	0.790877365	0.75428065	0.75406093
40	0.789236451	0.790877365	0.75428065	0.753528776	0.747257231
41	0.790877365	0.75428065	0.753528776	0.720586017	0.789236451
42	0.75428065	0.753528776	0.720586017	0.674656918	0.790877365
43	0.753528776	0.720586017	0.674656918	0.677087808	0.75428065
44	0.720586017	0.674656918	0.677087808	0.710189324	0.753528776
45	0.674656918	0.677087808	0.710189324	0.693028304	0.720586017
46	0.677087808	0.710189324	0.693028304	0.654421087	0.674656918
47	0.710189324	0.693028304	0.654421087	0.642499055	0.677087808
48	0.693028304	0.654421087	0.642499055	0.659823913	0.710189324
49	0.654421087	0.642499055	0.659823913	0.693947825	0.693028304
50	0.642499055	0.659823913	0.693947825	0.627771465	0.654421087

Table 12.2: Training sample: part II

Index	$d(t+h)$	$O(t+h)$	**Index**	$d(t+h)$	$O(t+h)$
1	0.725833896	0.759841479	38	0.431818059	0.458493194
2	0.658762145	0.725531746	39	0.364310678	0.369775829
3	0.592861386	0.648226008	40	0.356070543	0.261225425
4	0.559709068	0.578630556	41	0.363811545	0.239966892
5	0.516167677	0.55007335	42	0.33600744	0.254313198
6	0.512887118	0.501576787	43	0.341760802	0.197432782
7	0.520546836	0.484970232	44	0.336206839	0.190287143
8	0.532453628	0.503114832	45	0.326323249	0.196626274
9	0.561027388	0.51103254	46	0.326954468	0.168800845
10	0.548293788	0.547520903	47	0.308491638	0.169951838
11	0.484995379	0.535858088	48	0.299671087	0.144593108
12	0.487370387	0.440256917	49	0.276093734	0.121155165
13	0.561018497	0.438887984	50	0.304183602	0.089651619
14	0.559884336	0.55732989	51	0.373128689	0.119159444
15	0.549979155	0.558809799	52	0.425084213	0.234302589
16	0.647468791	0.518685285	53	0.495738797	0.316813804
17	0.741889967	0.654962133	54	0.55822437	0.4202089
18	0.801006076	0.782883837	55	0.577394623	0.529769824
19	0.836565154	0.821748405	56	0.529735704	0.562255484
20	0.809244941	0.838048926	57	0.432632166	0.499735263
21	0.824034764	0.797030768	58	0.439667016	0.362630249
22	0.835232864	0.799250472	59	0.514148285	0.369600935
23	0.805128684	0.825323756	60	0.511089986	0.496528384
24	0.762821789	0.787730918	61	0.495961057	0.485527621
25	0.741475928	0.739709041	62	0.493537787	0.438942178
26	0.703096052	0.730362726	63	0.494335384	0.452528648
27	0.673695484	0.70400469	64	0.498362737	0.457860919
28	0.658820567	0.672047032	65	0.498362737	0.460368694
29	0.651808578	0.665542527	66	0.493555568	0.460578216
30	0.663961761	0.662086644	67	0.48150907	0.452780763
31	0.652267069	0.676479544	68	0.400688943	0.436101157
32	0.649941593	0.665272323	69	0.31776941	0.321101473
33	0.617519558	0.654438011	70	0.341298501	0.176715074
34	0.573786388	0.622658325	71	0.410313441	0.210539523
35	0.548079149	0.561943842	72	0.463235479	0.321180044
36	0.524658013	0.533771078	73	0.484740097	0.385126151
37	0.491506965	0.508523427	74	0.46040833	0.417596243
			75	0.499672166	0.392765747

Table 12.3: Out of sample values and forecasted output

Part V

Numerical Analysis

In this part we present a few numerical tools to perform the necessary computation when performing quantitative trading strategies.

Chapter 13

Presenting some machine-learning methods

Machine learning is about programming computers to learn and to improve automatically with experience. While computers can not yet learn as well as people, algorithms have been devised that are effective for certain types of learning tasks, and a theoretical understanding of learning has emerged. Computer programs developed, exhibiting useful types of learning, especially in speech recognition and data mining (see Mitchell [1997]). Machine learning is inherently a multidisciplinary field drawing results from artificial intelligence, probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology, neurobiology, and other fields. We are going to introduce some of these fields as they have been extensively used in the financial industry to devise systematic trading strategies.

13.1 Some facts on machine-learning

13.1.1 Introduction to data mining

The rapid growth and integration of databases provided scientists, engineers, and business people with a vast new resource that can be analysed to make scientific discoveries, optimise industrial systems, and uncover financially valuable patterns. New methods targeted at large data mining problems have been developed. Data Mining (DM) is the analysis of (large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner (see Hand et al. [2001]). There exists several functionalities to DM, such that data characterisation summarising the general characteristics or features of a target class of data, and data discrimination comparing the general features of target class data objects with the general features of objects from a set of contrasting classes. Further, association analysis is the discovery of association rules showing attribute value conditions occurring frequently together in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Machine learning have been traditionally used for classification, where models such as decision trees, SVM, neural networks, Bayesian belief networks, genetic algorithm etc have been considered. The ability to perform classification and be able to learn to classify objects is paramount to the process of decision making. One classification task, called supervised learning, consists of a specified set of classes, and example objects labelled with the appropriate class. The goal being to learn from the training objects, enabling novel objects to be identified as belonging to one of the classes. The process of seeking relationships within a data set involves a number of steps

1. Model or pattern structure: determining the nature and structure of the representation to be used.
2. Score function: deciding how to quantify and compare how well different representations fit the data (choosing a score function).

3. Optimisation and search method: choosing an algorithmic process optimising the score function.

4. Data management strategy: deciding what principles of data management are required for implementing the algorithms efficiently.

While a model structure is a global summary of a data set making statements about any point in the full measurement space, pattern structures only make statements about restricted regions of the space spanned by the variables. Score functions judge the quality of a fitted model, and should precisely reflect the utility of a particular predictive model. While the task of finding the best values of parameters in models is cast as an optimisation problem, the task of finding interesting patterns (such as rules) from a large family of potential patterns is cast as a combinatorial problem, and is often accomplished using heuristic search techniques. At last, data management strategy is about the ways in which the data are stored, indexed, and accessed. The success of classification learning is heavily dependent on the quality of the data provided for training, as the learner only has the input to learn from. On the other hand, we want to avoid overfitting the given data set, and would rather find models or patterns generalising potential future data. Note, even though data mining is an interdisciplinary exercise, it is a process relying heavily on statistical models and methodologies. The main difference being the large size of the data sets to manipulate, requiring sophisticated search and examination methods. Further difficulties arise when there are many variables (curse of dimensionality), and often the data is constantly evolving. Recently, a lot of advances have been made on machine learning strategies mimicking human learning, and we refer the reader to Battula et al. [2013] for more details.

13.1.2 The challenges of computational learning

While the ability to perform classification and be able to learn to classify objects is paramount to the process of decision making, there exists several types of learning (see Mitchell [1997]), such as

- Supervised learning: learning a function from example data made of pairs of input and correct output.
- Unsupervised learning: learning from patterns without corresponding output values
- Reinforcement learning: learning with no knowledge of an exact output for a given input. Nonetheless, online or delayed feedback on the desirability of the types of behaviour can be used to help adaptation of the learning process.
- Active learning: learning through queries and responses.

More formally, these types of learning are part of what is called inductive learning where conclusions are made from specific instances to more general statements. That is, examples are provided in the form of input-output pairs $[X, f(X)]$ and the learning process consists of finding a function h (called hypothesis) which approximates a set of samples generated by the function f. The search for the function h is formulated in such a way that it can be solved by using search and optimisation algorithms. Some of the challenges of computational learning can be sumarised as follow:

- Identifying a suitable hypothesis can be computationally difficult.
- Since the function f is unknown, it is not easy to tell if the hypothesis h generated by the learning algorithm is a good approximation.
- The choice of a hypothesis space describing the set of hypotheses under consideration is not trivial.

As a result, a simple hypothesis consistent with all observations is more likely to be correct than a complex one. In the case where multiple hypotheses (an Ensemble) are generated, it is possible to combine their predictions with the aim of reducing generalisation error. For instance, boosting works as follow

- Examples in the training set are associated with different weights.
- The weights of incorrectly classified examples are increased, and the learning algorithm generates a new hypothesis from this new weighted training set. The process is repeated with an associated stopping criterion.
- The final hypothesis is a weighted-majority of all the generated hypotheses which can be based on different mixture of expert rules.

The difficult part being to know when to stop the iterative process and how to define a proper measure of error. One way forward is to consider the Probably Approximately Correct (PAC) learning which can be described as follow:

- A hypothesis is called approximately correct if its error insample lies within a small constant of the true error.
- By learning from a sufficient number of examples, one can calculate if a hypothesis has a high probability of being approximately correc.
- There is a connection between the past (seen) and the future (unseen) via an assumption stating that the training and test datasets come from the same probability distribution. It follows from the common sense that non-representative samples do not help learning.

More formally, a concept class C is said to be PAC learnable using a hypothesis class H if there exists a learning algorithm L such that for all concepts in C, for all instance distributions D on an instance space X,

$$\forall \epsilon\ ,\ \delta\ 0 < \epsilon\ ,\ \delta < 1$$

when given access to the example set, produces with probability at least $(1-\delta)$, a hypothesis h from H with error no-more than ϵ. To specify the problem we get a set of instances X, a set of hypotheses H, a set of possible target concepts C, training instances generated by a fixed, unknown probability distribution $\mathcal{D}$ over X, a target value $c(x)$, some training examples $< x, c(x) >$, and a hypothesis h estimating c. Then, the error of a hypothesis h satisfies

$$error_D = P_{x\in\mathcal{D}}(c(x) \neq h(x))$$

and the deviation of the true error from the training error satisfies

- Training error: $h(x) \neq c(x)$ over training instances.
- True error: $h(x) \neq c(x)$ over future random instances.

We must now measure the difference between the true error and the training error. Any hypothesis h is consistent when for all training samples,

$$h(x) = c(x)$$

If the hypothesis space H is finite, and $\mathcal{D}$ is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than ϵ is less than

$$|H|e^{-\epsilon m}$$

and $P(1of|H|hyps.consistentwithmexs.) < |H|e^{-\epsilon m}$. Considering a bounded sample size, for the probability to be at most δ, that is, $|H|e^{-\epsilon m} \leq \delta$, then

$$m \geq \frac{1}{\epsilon}\big(\ln|H| + \ln\frac{1}{\delta}\big)$$

More can be found on agnostic learning and infinite hypothesis space in books by Mitchell [1997] and Bishop [2006].

13.2 Introduction to information theory

13.2.1 Presenting a few concepts

Shannon [1948] introduced Information Theory (IT) by proposing some fundamental limits to the representation and transmission of information. Since then, IT has provided the theoretical motivations for many of the outstanding advances in digital communications and digital storage. One generally uses the following digital communication model in the transfer of information

1. The source is the source of (digital) data
2. The source encoder serves the purpose of removing as much redundancy as possible from the data. It is called the data compression portion.
3. The channel coder puts a modest amount of redundancy back in order to perform error detection or correction.
4. The channel is what the data passes through, possibly becoming corrupted along the way.
5. The channel decoder performs error correction or detection.
6. The source decoder undoes what is necessary to recover the data back.

There are other blocks that could be inserted in that model, such as

- a block enforcing channel constraints sometime called a line coder.
- a block performing encryption/decryption.
- a block performing lossy compression.

One of the key concept in IT is that information is conveyed by randomness, that is, information is defined in some mathematical sense, which is not identical to the one used by humans. However, it is not too difficult to make the connection between randomness and information. Another important concept in IT is that of typical sequences. In a sequence of bits of length n, there are some sequences which are typical. For example, in a sequence of coin-tossing outcomes for a fair coin, we would expect the number of heads and tails to be approximately equal. A sequence not following this trend is thus atypical. A good part of IT is capturing this concept of typicality as precisely as possible and using it to concluding how many bits are needed to represent sequence of data. The main idea being to use bits to represent only the typical sequences, since the other ones are rare events. This concept of typical sequences corresponds to the asymptotic equipartition property. Given a discrete random variable X, with x a particular outcome occurring with probability $p(x)$. Then, we assign to that event x the information that it conveys via the uncertainty measure

$$\text{uncertainty } = -\log p(x)$$

where the base of the logarithms determines the units of information. The units of $\log_2$ are in bits, that of $\log_e$ are in nats. For example, a random variable having two outcomes, 0 and 1, occurring with probability $p(0) = p(1) = \frac{1}{2}$, then each outcome conveys 1 bit of information. However, if $p(0) = 1$ and $p(1) = 0$, then the information conveyed when 0 happens is 0. We get no information from it, as we knew all along that it would happen. Hence, the information that 1 happens is ∞, as we are totally surprised by its occurrence. In general, one commonly use the average uncertainty provided by a random variable X taking value in a space χ, leading to the notion of entropy.

13.2.2 Some facts on entropy in information theory

The entropy functional, defined on a Markov diffusion process, plays an important roll in the theory of information and statistical physics, informational macrodynamics and control systems. In information theory (IT), entropy is the average amount of information contained in each message received. That is, entropy is a measure of unpredictability of information content. Named after Boltzmann's H-theorem, Shannon [1948] defined the entropy H of a discrete random variable X with possible values $\{x_1, .., x_n\}$ and probability mass function $P(X)$ as

$$H(X) = E[I(X)] = E[-\log P(X)]$$

where $I(X)$ is the information content of X, which is itself a random variable. When taken from a finite sample, the entropy can explicitly be written as

Definition 13.2.1 *The entropy $H(X)$ of a discrete random variable X taking values $\{x_1, .., x_n\}$ is*

$$H(X) = \sum_{i=1}^{n} p(x_i)I(x_i) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

where b is the base of the logarithm used.

The unit of entropy is bit for $b = 2$, nat for $b = e$ (where e is Euler's number), and dit (or digit) for $b = 10$. Given the well-known limit $\lim_{p\to 0_+} p\log p = 0$, in the case where $P(x_i) = 0$ for some i, the value of the corresponding summand is taken to be 0. For example, taking a fair coin, we get

$$H(X) = -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) = 1 \text{ bit}$$

For a biased coin with $p(0) = 0.9$, we get

$$H(X) = -(0.9\log 0.9 + 0.1\log 0.1) = 0.469 \text{ bit}$$

If X is a binary random variable

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases}$$

then the entropy of X is

$$H(X) = -p\log p - (1-p)\log(1-p)$$

For some function $g(X)$ of a random variable, we get $E[g(X)] = \sum_{x\in\mathcal{X}} g(x)p(x)$ so that for $g(x) = \log\frac{1}{p(x)}$ we get

$$H(X) = E[g(X)] = E[\log\frac{1}{p(X)}]$$

We are now interested in the entropy of pairs of random variables (X, Y).

Definition 13.2.2 *If X and Y are jointly distributed according to $p(X, Y)$, then the joint entropy $H(X, Y)$ is*

$$H(X,Y) = -E[\log p(X,Y)] = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

Definition 13.2.3 *If $(X, Y) \sim p(x, y)$, the conditional entropy of two events X and Y is given by*

$$H(Y|X) = -E_{p(x,y)}[\log p(Y|X)] = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

This quantity should be understood as the amount of randomness in the random variable Y given that you know the value of X. The conditional entropy can also be written as

$$\begin{aligned} E[Y|X] &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \end{aligned}$$

Theorem 13.2.1 *chain rule*

$$H(X,Y) = H(X) + H(Y|X)$$

We can also have a joint entropy with a conditioning on it.

***Corollary* 2**

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

The inspiration for adopting the word entropy in information theory came from the close resemblance between Shannon's formula and very similar known formulae from statistical mechanics. In statistical thermodynamics the most general formula for the thermodynamic entropy S of a thermodynamic system is the Gibbs entropy

$$S = -k_B \sum_i p_i \ln p_i$$

where k_B is the Boltzmann constant, and p_i is the probability of a microstate. The Gibbs entropy translates over almost unchanged into the world of quantum physics to give the von Neumann entropy, introduced by John von Neumann in 1927

$$S = -k_B Tr(\rho \ln \rho)$$

where ρ is the density matrix of the quantum mechanical system and $Tr()$ is the trace. Note, at a multidisciplinary level, connections can be made between thermodynamic and informational entropy. In the view of Jaynes (1957), thermodynamic entropy, as explained by statistical mechanics, should be seen as an application of Shannon's information theory: the thermodynamic entropy is interpreted as being proportional to the amount of further Shannon information needed to define the detailed microscopic state of the system, that remains uncommunicated by a description solely in terms of the macroscopic variables of classical thermodynamics, with the constant of proportionality being just the Boltzmann constant.

13.2.3 Relative entropy and mutual information

Another useful measure of entropy that works equally well in the discrete and the continuous case is the relative entropy of a distribution. Given a random variable with true distribution p, which we do not know due to incomplete information, instead we assume the distribution q. Then, the code will need more bits to represent the random variable, and the difference in bits is denoted by $D(p||q)$. The relative entropy is defined as the Kullback-Leibler divergence from the distribution to the reference measure q as follows.

Definition 13.2.4 *The relative entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D_{KL}(p||q) = E_p[\log \frac{p(x)}{q(x)}] = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

This form is not symmetric, and q appears only in the denominator. Alternatively, assume that a probability distribution p is absolutely continuous with respect to a measure q, that is, it is of the form $p(dx) = f(x)q(dx)$ for some non-negative q-integrable function f with q-integral 1, then the relative entropy can be defined as

$$D_{KL}(p||q) = \int \log f(x)p(dx) = \int f(x) \log f(x)q(dx)$$

In this form the relative entropy generalises (up to change in sign) both the discrete entropy, where the measure q is the counting measure, and the differential entropy, where the measure q is the Lebesgue measure. If the measure q is itself a probability distribution, the relative entropy is non-negative, and zero if $p = q$ as measures. It is defined for any measure space, hence coordinate independent and invariant under co-ordinate reparametrisations if one properly takes into account the transformation of the measure q. The relative entropy, and implicitly entropy and differential entropy, do depend on the reference measure q.

Another important concept, called mutual information, describes the amount of information a random variable tells about another one. That is, observing the output of a channel, we want to know what information was sent. The channel coding theorem is a statement about mutual information.

Definition 13.2.5 *Let X and Y be random variables with joint distribution $p(X,Y)$ and marginal distributions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution*

$$\begin{aligned} I(X;Y) &= D(p(x,y)||p(x)p(y)) \\ &= \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Note, when X and Y are independent, $p(x,y) = p(x)p(y)$, and we get $I(X;Y) = 0$. That is, in case of independence, Y, can not tell us anything about X. An important interpretation of mutual information comes from the following theorem.

Theorem 13.2.2

$$I(X;Y) = H(X) - H(X|Y)$$

which states that the information that Y tells us about X is the reduction in uncertainty about X due to the knowledge of Y. By symmetry we get

$$I(X;Y) = H(Y) - H(Y|X) = I(Y;X)$$

Using $H(X,Y) = H(X) + H(Y|X)$, we get

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

The information that X tells about Y is the uncertainty in X plus the uncertainty about Y minus the uncertainty in both X and Y. We can summarise statements about entropy as follows

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ I(X;Y) &= H(Y) - H(Y|X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(X;Y) &= I(Y;X) \\ I(X;X) &= H(X) \end{aligned}$$

When dealing with the sequence of random variables $X_1, .., X_n$ drawn from the joint distribution $p(x_1, .., x_n)$, a variety of chain rules have been developed.

Theorem 13.2.3 *The joint entropy of $X_1, .., X_n$ is*

$$H(X_1, .., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, .., X_1)$$

The chain rule for entropy leads us to a chain rule for mutual information.

Theorem 13.2.4

$$I(X_1, .., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, ..., X_1)$$

13.2.4 Bounding performance measures

A large part of information theory consists in finding bounds on certain performance measures. One of the most important inequality used in IT is the Jensen's inequality. We are interested in convex functions because it is known that over the interval of convexity, there is only one minimum (for details see Appendix (A.1)). The following theorem describe the information inequality.

Theorem 13.2.5 $\log x \leq x - 1$, *with equality if and only if* $x = 1$.

We can now characterise some of the information measures defined above.

Theorem 13.2.6 $D(p||q) \geq 0$, *with equality if and only if* $p(x) = q(x)$ *for all* x.

***Corollary* 3** *Mutual information is positive,* $I(X; Y) \geq 0$, *with equality if and only* if X *and* Y *are independent.*

We let the random variable X takes values in the set $\mathcal{X}$, and we let $|\mathcal{X}|$ denotes the number of elements in that set. For discrete random variables, the uniform distribution over the range $\mathcal{X}$ has the maximum entropy.

Theorem 13.2.7 $H(X) \leq \log |\mathcal{X}|$, *with equality if and only if* X *has a uniform distribution.*

Note, if you can show that some performance criterion is upper-bounded by some function, then by showing how to achieve that upper bound, we have found an optimum. We see that the more we know, the less uncertainty there is.

Theorem 13.2.8 *Condition reduces entropy:*

$$H(X|Y) \leq H(X)$$

with equality if and only if X *and* Y *are independent.*

Theorem 13.2.9

$$H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if X_i *are independent.*

The following theorem allow us to deduce the concavity (or convexity) of many useful functions.

Theorem 13.2.10 *Log-sum inequality*
For non-negative numbers $a_1, ..., a_n$ *and* $b_1, .., b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i} = \ constant$.

The conditions on this theorem are much weaker than the one for Jensen's inequality, since it is not necessary to have the sets of numbers add up to 1. Using this inequality, one can prove a convexity statement about the relative entropy function.

Theorem 13.2.11 *If* (p_1, q_1) *and* (p_2, q_2) *are pairs of probability mass function, then*

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda q_2)) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

for all $0 \leq \lambda \leq 1$. *That is,* $D(p||q)$ *is convex in the pair* (p, q).

Further,

Theorem 13.2.12 $H(p)$ *is a concave function of* p.

Theorem 13.2.13 *Let* $(X, Y) \sim p(x, y) = p(x)p(y|x)$. *The mutual information* $I(X;Y)$ *is a concave function of* $p(x)$ *for fixed* $p(y|x)$ *and a convex function of* $p(y|x)$ *for fixed* $p(x)$.

The data processing inequality states that no matter what processing we perform on some data, we can not get more information out of a set of data than was there to begin with. In a sense, it provides a bound on how much can be accomplished with signal processing.

Definition 13.2.6 *Random variable* X, Y, *and* Z *are said to form a Markov chain in that order, denoted by* $X \to Y \to Z$ *if the conditional distribution of* Z *depends only on* Y *and is independent of* X. *(That is, if we know* Y, *then knowing* X *also does not tell us any more than if we only know* Y*). if* X, Y, *and* Z *form a Markov chain, then the joint distribution can be written*

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

The concept of a state is that knowing the present state, the future of the system is independent of the past. The conditional independence idea means

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

Note, if $Z = f(Y)$ then $X \to Y \to Z$.

Theorem 13.2.14 *Data processing inequality*
If $X \to Y \to Z$, *then*

$$I(X;Y) \geq I(X;Z)$$

If we think of Z as being the result of some processing done on the data Y, that is, $Z = f(Y)$ for some deterministic or random function, then there is no function that can increase the amount of information that Y tells about X.

13.2.5 Feature selection

A fundamental problem of machine learning is to approximate the functional relationship $f(\bullet)$ between an input $X = \{x_1, x_2, .., x_M\}$ and an output Y, based on a memory of data points, $\{X_i, Y_i\}$, $i = 1, .., N$, where the X_i are vectors of reals and the Y_i are real numbers. There are some cases where the output Y is not determined by the complete set of the input features $X = \{x_1, x_2, .., x_M\}$, but it is only decided by a subset of them $\{x_{(1)}, x_{(2)}, .., x_{(m)}\}$, where $m < M$. When we have sufficient data and time, we can use all the input features, including the irrelevant ones, to approximate the underlying function between the input and the output. However, in practice the irrelevant features raise two problems in the learning process.

1. The irrelevant input features will induce greater computational cost.
2. The irrelevant input features may lead to overfitting

As a result, it is reasonable and important to ignore those input features with little effect on the output, so as to keep the size of the approximator model small. While the feature selection problem has been studied by the statistics and machine learning communities for decades, it has received much attention in the field of data mining. Some researchers call it filter models, while other classify it as wrapped around methods. It is also known as subset selection in the statistics community.

Since the computational cost of brute-force feature selection method is prohibitively high, with considerable danger of overfitting, people resort to greedy methods, such as forward selection. We must first clarify the problem of performance evaluation of a set of input features. Even if the feature sets are evaluated by test-set cross-validation or leave-one-out cross validation, an exhaustive search of possible feature sets is likely to find a misleadingly well scoring feature set by chance. To prevent this from happening, we can use a cascaded cross-validation procedure, which selects from increasingly large sets of features (and thus from increasingly large model classes).

1. Shuffle the data set and split into a training set of 70% of the data and a test-set of the remaining 30%.
2. Let j vary among feature set sizes $j = (0, 1, .., m)$
   (a) Let f_{s_j} = best feature set of j, where best is measured as the minimiser of the leave-one-out cross-validation error over the training set.
   (b) Let $Testscore_j$ = the RMS prediction error of feature set f_{s_j} on the test-set.
3. Select the feature set f_{s_j} for which the test-set score is minimised.

The score for the best feature set of a given size is computed by independent cross-validation from the score for the best size of feature set. Note, this procedure does not describe how the search for the best feature set of size j in step 2a) is done. Further, the performance evaluation of a feature selection algorithm is more complicated than the evaluation of the feature set. This is because we must first ask the algorithm to find the best feature subset. We must then give a fair estimate of how well the feature selection algorithm performs by trying the first step on different data sets. Hence, the full procedure of evaluating the performance of a feature selection algorithm (described above) should have two layers of loops (see algorithm below). The inner loop using an algorithm to find the best subset of features, and the outer loop evaluating the performance of the algorithm with different data sets.

1. Collect a training data set from the specific domain.
2. Shuffle the data set.
3. Break it into P partition, say $P = 20$.
4. For each partition ($i = 0, 1, .., P - 1$)

(a) Let $OuterTrainset(i) =$ all partitions except i.

(b) Let $OuterTestset(i) =$ the ith partition.

(c) Let $InnerTrain(i) =$ randomly chosen 70% of the $OuterTrainset(i)$.

(d) Let $InnerTest(i) =$ the remaining 30% of the $OuterTrainset(i)$.

(e) For $j = 0, 1, .., m$
Search for the best feature set with j components $f_{s_{ij}}$ using leave-one-out on $InnerTrain(i)$.
Let $InnerTestScore_{ij} =$ RMS score of $f_{s_{ij}}$ on $InnerTest(i)$
End loop of (j).

(f) Select the $f_{s_{ij}}$ with the best inner test score.

(g) Let $OuterScore_i =$ RMS score of the selected feature set on $OuterTestset(i)$.

End of loop of (i).

5. Return the Mean Outer Score

We are now going to introduce the forward feature selection algorithm, and explore three greedy variants of this algorithm improving the computational efficiency without sacrifying too much accuracy.

The forward feature selection procedure begins by evaluating all feature subsets consisting of only one input attribute. We start by measuring the leave-one-out cross-validation (LOOCV) error of the one-component subsets $\{X_1\}, \{X_2\}, ..\{X_M\}$, where M is the input dimensionality, so that we can find the best individual feature $X_{(1)}$. Next, the forward selection finds the best subset consisting of two components, $X_{(1)}$, and one other feature from the remaining $M - 1$ input attributes. Hence, there are a total of $M - 1$ pairs. We let $X_{(2)}$ be the other attribute in the best pair besides $X_{(1)}$. Then, the input subsets with three, four, and more features are evaluated. According to the forward selection, the best subset with m features is the m-tuple consisting of $\{X_1\}, \{X_2\}, ..\{X_m\}$, while overall, the best feature set is the winner out of all the M steps. Assuming that the cost of a LOOCV evaluation with i features is $C(i)$, then the computational cost of forward selection searching for a feature subset of size m out of M total input attributes will be

$$MC(1) + (M - 1)C(2) + ... + (M - m + 1)C(m)$$

For example, the cost of one prediction with one-nearest neighbour as the function approximator, using a kd-tree with j inputs, is $\mathcal{O}(j \log N)$ where N is the number of data points. Thus, the cost of computing the mean leave-one-out error, which involves N predictions, is $\mathcal{O}(jN \log N)$. Hence, the total cost of feature selection using the above formula is $\mathcal{O}(m^2 MN \log N)$.

An alternative solution to finding the total best input feature set is to employ exhaustive search. This method starts with searching the best one-component subset of the input features, which is the same in the forward selection algorithm. Then, we need to find the best two-component feature subset which may consist of any pairs of the input features. Afterwards, it consists in finding the best triple out of all the combinations of any three input features, etc. One can see that the cost of exhaustive search is

$$MC(1) + \binom{M}{2} C(2) + ... + \binom{M}{m} C(m)$$

and we see that the forward selection is much cheaper than the exhaustive search. However, the forward selection may suffer due to its greediness. For example, if $X_{(1)}$ is the best individual feature, it does not guarantee that either $\{X_{(1)}, X_{(2)}\}$ or $\{X_{(1)}, X_{(3)}\}$ must be better than $\{X_{(2)}, X_{(3)}\}$. As a result, a forward selection algorithm may select a feature set different from that selected by the exhaustive searching. Hence, with a bad selection of the input features, the prediction Y_q of a query $X_q = \{x_1, x_2, ..., x_M\}$ may be significantly different from the true Y_q.

In the case where the greediness of forward selection does not have a significantly negative effect on accuracy, we need to know how to modify the forward selection algorithm to be greedier in order to further improve the efficiency. There exists several greedier feature selection algorithms whose goal is to select no more than m features from a total of M input attributes, and with tolerable loss of prediction accuracy. We briefly discuss three of these algorithms

1. The super greedy algorithm: do all the 1-attribute LOOCV calculations, sort the individual features according to their LOOCV mean error, then take the m best features as the selected subset. Thus, we do M computations involving one feature and one computation involving m features. If the nearest neighbour is the function approximator, the cost of super greedy algorithm is $\mathcal{O}((M+m)N\log N)$.

2. The greedy algorithm: do all the 1-attribute LOOCV and sort them, take the best two individual features and evaluate their LOOCV error, then take the best three individual features, and so on, until m features have been evaluated. Compared with the super greedy algorithm, this algorithm may conclude at a subset whose size is smaller than m but whose inner tested error is smaller than that of the m component feature set. Hence, the greedy algorithm may end up with a better feature set than the super greedy one does. The cost of the greedy algorithm for the nearest neighbour is $\mathcal{O}((M+m^2)N\log N)$.

3. The restricted forward selection (RFS):

   (a) caculate all the 1-feature set LOOCV errors, and sort the features according to the corresponding LOOCV errors. We let the features ranking from the most important to the least important be $X_{(1)}, X_{(2)}, ..., X_{(M)}$.

   (b) do the LOOCV of 2-feature subsets consisting of the winner of the first round, $X_{(1)}$, along with another feature, either $X_{(2)}$ or $X_{(3)}$, or any other one until $X_{(\frac{M}{2})}$. There are $\frac{M}{2}$ of these pairs. The winner of this round will be the best 2-component feature subset chosen by RFS.

   (c) calculate the LOOCV errors of $\frac{M}{3}$ subsets consisting of the winner of the second round, along with the other $\frac{M}{3}$ features at the top of the remaining rank. In this way, the RFS will select its best feature triple.

   (d) continue this procedure until RFS has found the best m-component feature set.

   (e) From step 1) to step 4), the RFS has found m feature sets whose sizes range from 1 to m. By comparing their LOOCV errors, the RFS can find the best overall feature set.

One of the difference between RFS and conventional forward selection (FS) is that at each step, when inserting an additional feature into the subset, the FS considers all the remaining features, while the RFS only tries the part of them which seems the more promising. The cost of RFS for the nearest neighbour is $\mathcal{O}(MmN\log N)$.

We are left with finding out how cheap and how accurate all these varieties of forward selection are compared with the conventional forward selection method. Using real world data sets from StatLib/CMU and UCI's machine learning data repository coming from different domains such as biology, sociology, robotics etc, it was demonstrated that Exhaustive Search (ES) is prohibitively time consuming. Even though the features selected by FS may differ from the result of ES because some of the input features are not mutually independent, ES is far more expensive than the FS, while it is not significantly more accurate than FS. In order to investigate the influence of greediness on the above three greedy algorithms, we consider

1. the probabilities for these algorithms to select any useless features

2. the prediction errors using the feature set selected by these algorithms

3. the time cost for these algorithms to find their feature sets

It was found that FS does not eliminate more useless features than the greedier competitors except for the Super Greedy one. However, the greedier the algorithm, the more easily confused by the relevant but corrupted features it becomes. Further, the three greedier feature selection algorithms do not suffer great loss in accuracy, and RFS performs almost

as well as the FS. As expected, the greedier algorithms improve the efficiency. It was found that the super greedy algorithm (Super) was ten time faster than the FS, while the greedy algorithm (Greedy) was seven times faster, and the restricted forward selection (RFS) was three times faster. Finally, the RFS performed better than the conventional FS in all aspects. Inserting more independent random noise and corrupted features to the data sets, the probability for any corrupted feature to be selected remained almost the same, while that of independent noise reduced greatly. To conclude, while in theory the greediness of feature selection algorithms may lead to great reduction in the accuracy of function approximation, in practice it does not often happen. The three greedier algorithms discussed above improve the efficiency of the forward selection algorithm, especially for larger data sets with high input dimensionalities, without significant loss in accuracy. Even in the case where the accuracy is more crucial than the efficiency, restricted forward selection is more competitive than the conventional forward selection.

13.3 Online learning and regret-minimising algorithms

13.3.1 Simple online algorithms

First, we introduce some notation which will be used through out this section. We use $\text{relint}\,(S)$ to refer to the relative interior of a convex set S, which is the set S minus all of the points on the relative boundary. We use $\text{closure}\,(S)$ to refer to the closure of S, the smallest closed set containing all of the limit points of S. For any subset S of $\mathbb{R}^d$, we let $\mathcal{H}(S)$ denote the convex hull of S. We let $\Delta_n = \{x \in \mathbb{R}^n : \sum_i^n x_i = 1, x_i \geq 0 \forall i\}$ be the n-dimensional probability simplex.

13.3.1.1 The Halving algorithm

To introduce online learning we are first going to present the Halving algorithm where a player has access to the prediction of N experts denoted by

$$f_{1,t}, ..., f_{N,t} \in \{0, 1\}$$

At each time $t = 1, .., T$, we observe $f_{i,t}$, i=1,..,N, and predict $p_t \in \{0, 1\}$. We then observe $y_t \in \{0, 1\}$ and suffer loss $I_{\{p_t \neq y_t\}}$. Suppose $\exists j$ such that $f_{j,t} = y_t$ for all $t \in [T]$. The the Halving theorem predict $p_t = majority(C_t)$, where $C_1 = [N]$ and $C_t \subseteq [N]$ is defined below for $t > 1$.

Theorem 13.3.1 *If* $p_t = majority(C_t)$ *and*

$$C_{t+1} = \{i \in C_t : f_{i,t} = y_t\}$$

then we will make at most $\log_2 N$ *mistakes.*

In fact, for every t at which there is a mistake, at least half of the experts in C_t are wrong, so that $|C_{t+1}| \leq \frac{|C_t|}{2}$. It follows that $|C_T| \leq \frac{|C_1|}{2^M}$ where M is the tolal number of mistakes. Further, since there is a perfect expert, $|C_T| \geq 1$. As a result, recalling that $C_1 = [N]$, then $1 \leq \frac{N}{2^M}$, such that, after rearranging we get $M \leq \log_2 N$.

13.3.1.2 The weighted majority algorithm

To illustrate online, no-regret learning, we consider the problem of learning from expert advice where an algorithm must make a sequence of predictions based on the advice of a set of N experts and receive a corresponding sequence of losses. The aim of the algorithm being to achieve a cumulative loss that is almost as low as the cumulative loss of the best performing expert in hindsight. No statistical assumptions are made about these losses. At every time step $t \in \{1, .., T\}$, every expert $i \in \{1, .., N\}$ predict $f_{i,t} \in [0, 1]$ and receives a loss $l_{i,t} = l(f_{i,t}, y_t) \in [0, 1]$ such that the cumulative loss of expert i at time T is given by $L_{i,T} = \sum_{t=1}^T l_{i,t}$. We let the algorithm $\mathcal{A}$ (or player) maintains a

weight $w_{i,t}$ for each expert i at time t, where $\sum_{i=1}^{N} w_{i,t} = 1$. Hence, these weights can be seen as a distribution over the experts. The algorithm then receives its own instantaneous loss

$$l_{\mathcal{A},t} = \sum_{i=1}^{N} w_{i,t} l_{i,t}$$

which can be interpreted as the expected loss the algorithm would receive if it always chose an expert to follow according to the current distribution. The cumulative loss of $\mathcal{A}$ up to time T is defined as

$$L_{\mathcal{A},T} = \sum_{t=1}^{T} l_{\mathcal{A},t} = \sum_{t=1}^{T} \sum_{i=1}^{N} w_{i,t} l_{i,t}$$

For simplicity of exposition we use l_t, L_t, and w_t to refer to the vector of losses, vector of cumulative loss, and vector of weights, respectively, for each expert on round t. We will use the dot product to relate these vectors. Since the algorithm will not achieve a small cumulative loss if none of the experts perform well, we generally measure the performance of an algorithm in terms of its regret, defined as the difference between the cumulative loss of the algorithm and the loss of the best performing expert

$$R_T = L_{\mathcal{A},T} - \min_{i \in \{1,...,N\}} L_{i,T}$$

Hence, the algorithm is said to have no-regret if the average per time step regret approaches 0 as $T \to \infty$. The Randomized Weighted Majority (WM) algorithm (see Littlestone et al. [1994]) is an example of a no-regret algorithm, where the Weighted Majority uses weights

$$w_{i,t} = \frac{e^{-\eta L_{i,t-1}}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}$$

where $\eta > 0$ is a learning rate parameter and the player's predict is $p_t = \sum_{i=1}^{N} w_{i,t} f_{i,t}$. The pseudo code for the weighted majority algorithm is

1. for $t = 1, .., t$ do
    - player observes $f_{1,t}, .., f_{N,t}$
    - player predicts p_t
    - adversary reveals outcome y_t
    - player suffers loss $l(p_t, y_t)$
    - experts suffer loss $l(f_{i,t}, y_t)$
2. end for

After T trials, the regret of WM can be bounded as

$$R_T = L_{WM(\eta),T} - \min_{i \in \{1,...,N\}} L_{i,T} \leq \eta T + \frac{\log N}{\eta}$$

such that if T is known in advance, setting $\eta = \sqrt{\frac{\log N}{T}}$ yields the standard $\mathcal{O}(\sqrt{T \log N})$ regret bound. Further, setting $\eta = \sqrt{8\frac{\log N}{T}}$ we get the bound

$$R_T \leq \sqrt{\frac{T}{2} \log N}$$

13.3.2 The online convex optimisation

13.3.2.1 The online linear optimisation problem

Given the probability simplex Δ_N, the pseudo code for the online linear optimisation is

1. for $t = 1, .., t$ do
   - player predicts $w_t \in \Delta_N$ (w_t is essentially a probability distribution)
   - adversary reveals $l_t \in \mathbb{R}^N$
   - player suffers loss $w_t.l_t$ where $l_{i,t} = l(f_{i,t}, y_t)$
2. end for

The learner suffers regret

$$R_T = \sum_{t=1}^{T} w_t.l_t - \min_{w \in \Delta_N} \sum_{t=1}^{T} w.l_t$$

where $l_t(w) = w.l_t$. Note, for the simplex, the distribution w will place all the probability on the best expert. That is, the experts setting is just a special case of the online linear optimisation, where the set $\mathcal{K}$ is the N-simplex Δ_n.

It was proved that the weights chosen by WM are precisely those minimising a combination of empirical loss and an entropic regularisation term (see Kivinen et al. [1997]). That is, the weight vector w_t at time t is the solution to the following minimisation problem

$$\min_{w \in \Delta_N} w.L_{t-1} - \frac{1}{\eta} H(w)$$

where L_{t-1} is the vector of cumulative loss at time $t-1$, Δ_N is the probability simplex, and $H(\bullet)$ is the entropy function (see Section (13.2.2))

$$H(w) = -\sum_{i=1}^{N} w_i \log(w_i)$$

***Remark* 13.3.1** *The entropy function acts as a regularisation function for the weight vector w.*

13.3.2.2 Considering Bergmen divergence

Given the online linear optimisation problem described in Section (13.3.2.1), we consider the minimisation problem for the weight vector w_{t+1}

$$w_{t+1} \arg\min_{w \in \mathcal{K}} \eta \sum_{s=1}^{t} l_t(w) + R(w)$$

for some convex function $R(\bullet)$. We define $\Phi_0(w) = R(w)$ and $\Phi_t(w) = \Phi_{t-1}(w) + \eta l_t(w)$. We also let $D_f(x, y)$ be the Bergman divergence between x and y with respect to the function f (see details in Appendix (A.1.6)).

Lemma 13.3.1 *Suppose $\mathcal{K} = \mathbb{R}^N$, then for any $u \in \mathcal{K}$ we get*

$$\eta \sum_{t=1}^{T} [l_t(w) - l_t(u)] = D_{\Phi_0}(u, w_1) - D_{\Phi_T}(u, w_{T+1}) + \sum_{t=1}^{T} D_{\Phi_T}(w_t, w_{t+1})$$

and we also get

$$\sum_{t=1}^{T} l_t(w) \leq \inf_{u\in\mathcal{K}}\left[\sum l_t(u) + \eta^{-1} D_R(u, w_1)\right] + \eta \sum_{t=1}^{T} D_{\Phi_t}(w_t, w_{t+1})$$

Now, suppose that $\nabla R(w_1) = 0$ and $\mathbb{R}^N = \mathcal{K}$, then

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^N}\left[\eta l_t(w) + D_{\Phi_{t-1}}(w, w_t)\right]$$

and the two minimisation problems are equivalent. That is,

$$\begin{aligned} \eta l_t(w) &= \Phi_t(w) - \Phi_{t-1}(w) \\ \eta l_t(w) + D_{\Phi_{t-1}}(w, w_t) &= \Phi_t(w) - \Phi_{t-1}(w) + D_{\Phi_{t-1}}(w, w_t) \end{aligned}$$

Assuming that the two equations are equivalent for $\tau \leq t$, and w minimises Φ_{t-1}

$$\begin{aligned} &\nabla_w D_{\Phi_{t-1}}(w, w_t) = \nabla_w \Phi_{t-1}(w, w_t) - \nabla_w \Phi_{t-1}(w_t) \\ &\nabla \Phi_t(w_{t+1}) = \nabla \Phi_{t-1}(w_t) = ... = \nabla \mathbb{R}(w_1) = 0 \end{aligned}$$

thus, $w_{t+1} = \arg\min_{w\in\mathcal{K}} \Phi_t(w)$. Assuming l_t's are linear functions, and letting R^* be the Legendre dual of the function $R(\bullet)$ (see details in Appendix (A.1.4)), we get

Corollary 4 *1.* $\eta\left(\sum l_t.w_t - \sum l_t.u\right) = D_R(u, w_1) + D_R(u, w_{t+1}) + \sum D_R(w_t, w_t - 1)$ *for any* $u \in \mathbb{R}^N$

2. $w_{t+1} = \nabla R^*(\nabla R(w_t) - \eta l_t)$

Recall, the online gradient descent is $w_{t+1} = w_t - \eta l_t$. Further, if $R = \frac{1}{2}||\bullet||^{\frac{1}{2}}$, $\nabla R(w) = w$, $\nabla R^*(w) = w$, and if $l_t(\bullet)$ are convex (but not necessary linear), then

Lemma 13.3.2 *If we choose* $w_{t+1} = \arg\min_{w\in\mathbb{R}^N}\left[\eta \nabla l_t(w)^\top w + D_R(w, w_t)\right]$
(or equivalently $w_{t+1} = \arg\min_{w\in\mathbb{R}^N} \eta \sum\left[\nabla l_t(w)^\top w + R(x)\right]$*), then*

$$\sum_{t=1}^{T}\left[l_t(w_t) - l_t(u)\right] \leq \eta^{-1} D_R(u, w_1) + \sum_{t=1}^{T} D_R(w_t, w_{t+1})$$

13.3.2.3 More on the online convex optimisation problem

The WM is an example of a broader class of algorithms collectively known as Follow the Regularized Leader (FTRL) algorithm (see Hazan et al. [2008]). The FTRL template can be applied to a wide class of learning problems falling under the general framework of Online Convex Optimisation (see Zinkevich [2003]). Other problems falling into this framework include online linear pattern classification, online Gaussian density estimation, and online portfolio selection. Further, the online linear optimisation problem is an extension of the expert setting where the weights w_t are chosen from a fixed bounded convex action space $\mathcal{K} \subset \mathbb{R}^N$ (see Rakhlin [2009]). The algorithm follows

1. Input: convex compact decision set $\mathcal{K} \subset \mathbb{R}^N$.
2. Input: strictly convex differentiable regularisation function $\mathcal{R}(\bullet)$ defined on $\mathcal{K}$.
3. Parameter: $\eta > 0$.

4. Initialise: $L_1 =< 0, .., 0 >$.
5. for $t = 1, .., T$ do
6. The learner selects action $w_t \in \mathcal{K}$ according to

$$w_t = \arg\min_{w \in \mathcal{K}} w.L_{t-1} + \frac{1}{\eta}\mathcal{R}(w)$$

7. Nature reveals l_t, learner suffers loss $l_t.w_t$
8. The learner updates $L_t = L_{t-1} + l_l$
9. end for

Abernethy et al. [2012] showed that the FTRL algorithms for online learning and the problem of pricing securities in a prediction market have a strong syntactic correspondence. To do so they needed to make two assumptions

Assumption 2 *For each time step t,* $||l_t|| \leq 1$.

Assumption 3 *The regulasier* $\mathcal{R}(\bullet)$ *has the Legendre property (see Cesa-Bianchi et al. [2006]).* $\mathcal{R}$ *is strictly convex on relint* $(\mathcal{K})$ *and* $||\nabla\mathcal{R}(w)|| \to \infty$ *as* $w \to$ *relint* $(\mathcal{K})$.

As a result, the solution to the above algorithm will always occur in the relative interior of $\mathcal{K}$ such that the optimisation is unconstrained.

13.4 Presenting the problem of automated market making

We consider a securities market offering a set of contingent securities (or claims) such as Arrow-Debreu prices, or, more generally oprtion prices. In general, a securities market consists of the tuple $\left(\mathcal{O}, \rho, \Pi, R\right)$ where $\mathcal{O}$ is the outcome space, ρ is the payoff function, $\Pi \subseteq \mathbb{R}^d$ is a convex compact set of feasible prices, and $R : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function with domain Π. The cost function C of the market is assumed to be the conjugate of R with respect to the set Π. The market is complete if it offers at least $|\mathcal{O}| - 1$ linearly independent securities over the outcomes set $\mathcal{O}$.

13.4.1 The market neutral case

Various authors proposed a framework to design automated market makers for such markets, where an automated market maker is a market institution setting prices for each security, and always willing to accept trades at these prices. For instance, the Logarithmic Market Scoring Rule (LMSR) market maker is a popular standard prediction market mechanism over combinatorial outcome spaces (see Hanson [2003]). To illustrate the main idea we consider a market maker offering $|\mathcal{O}|$ Arrow-Debreu securities [1], each corresponding to a potential outcome. He uses a differentiable cost function, $C : \mathbb{R}^{|\mathcal{O}|} \to \mathbb{R}$ to determine the cost of each security. We let q_o be the number of shares of security o held by traders, and assume that a trader wants to purchase a bundle of r_o shares [2] for each security $o \in \mathcal{O}$. The trader should pay to the market maker the cost of a bundle r as $C(q + r) - C(q)$, where $q = r_1 + .. + r_t$ is the vector of previous purchases. We let $p_o(q)$ be the instantaneous price of security o given by $\frac{\partial C(q)}{\partial q_o}$. The cost function used in the LMSR model is

$$C(q) = b \log \sum_{o \in \mathcal{O}} e^{\frac{q_o}{b}}$$

[1] $\rho_i(0)$ equals 1 if o is the ith outcome and 0 otherwise.
[2] where each r_o can be positive for a purchase, negative for a sale, or zero if not traded.

where $b > 0$ is a parameter controlling the rate at which prices change. Hence, the corresponding price function for each security o is

$$p_o(q) = \frac{\partial C(q)}{\partial q_o} = \frac{e^{\frac{q_o}{b}}}{\sum_{o' \in \mathcal{O}} e^{\frac{q_{o'}}{b}}}$$

In that setting, the monetary loss of an automated market maker is upper-bounded by $b \log |\mathcal{O}|$.

To design an automated market maker we need first to determine an appropriate set of properties that the market maker should satisfy. In the case of the LMSR defined above, we get

1. The cost function is differentiable everywhere, so that the instantaneous price $p_o(q)$ can always be obtained for the security associated with any outcome o.
2. The market incorporate information from the traders, since the purchase of a security corresponding to outcome o causes p_o to increase.
3. The market does not provide explicit opportunities for arbitrage. The sum of the instantaneous prices of the securities $p_o(q)$ is always 1 (see Remark below).
4. the market is expressive in the sense that a trader with sufficient funds is always able to set the market prices to reflect his beliefs about the probability of each outcome.

***Remark* 13.4.1** *In addition to preventing arbitrage, these properties ensure that prices can be interpreted naturally as probabilities. As a result, prices can represent the market's current estimate of the distribution over outcomes.*

***Remark* 13.4.2** *The conditions of no-arbitrage and the properties of market prices are well known and have been extensively studied (see Appendix (F.4)).*

These conditions lead to some natural mathematical restrictions on the costs of security bundles. Further, we assume that each trader may have his own information about the future, which we represent as a distribution $p \in \Delta_{|\mathcal{O}|}$ over the outcome space, where $\Delta_n = \{x \in \mathbb{R}^n_{\geq 0}\}$ is the n-simplex. While the pricing mechanism incentivise the trader to reveal p, it also avoid providing arbitrage opportunities. In a complete market offering n Arrow-Debreu securities for the n mutually exclusive and exhaustive outcomes, we define a set of market makers by equating the set of allowable prices Π to the n-simplex Δ_n. A market maker satisfying the above conditions can use the cost function

$$C(q) = \sup_{x \in \Delta_n} x.q - R(x) \tag{13.4.1}$$

for a strictly convex function $R(\bullet)$. Hence, the market price $x(q) = \nabla C(q)$ is the optimal solution to the convex optimisation. Further, when $R(x) = b \sum_{i=1}^n x_i \log x_i$, which is the negative entropy function, we recover the LMSR market maker. Note, Agarwal et al. [2011] solved the problem of automated market making by formulating a minimisation problem with an associated penality function playing a similar role to the convex function $R(\bullet)$, but without using the conjugate duality.

13.4.2 The case of infinite outcome space

When $|\mathcal{O}|$ is large or infinite, calculating the cost of a purchase becomes intractable. To address this problem Abernethy et al. [2012] restricted the market maker to offer only K securities for some reasonably sized K, and assumed that the payoff of each security could be described by an arbitrary (but computable) function $\rho : \mathcal{O} \to \mathbb{R}^n_{\geq 0}$. Such a security space is called complex. They formalised the properties of a reasonable market, and precisely characterised the set of all cost functions satisfying these conditions. They obtained the following conditions

1. Existence of instantaneous prices: C is continuous and differentiable everywhere on $\mathbb{R}^K$.
2. Information incorporation: For any q and $r \in \mathbb{R}^K$,
$$C(q+2r) - C(q+r) \geq C(q+r) - C(q) \tag{13.4.2}$$
3. No arbitrage: For all q and $r \in \mathbb{R}^K$, there exists an $o \in \mathcal{O}$ such that
$$C(q+r) - C(q) \geq r.\rho(o) \tag{13.4.3}$$
4. Expressiveness: For any $p \in \Delta_{|\mathcal{O}|}$ we write
$$x_p = E_{o\sim p}[\rho(o)]$$
Then for any $p \in \Delta_{|\mathcal{O}|}$ and any $\epsilon > 0$ there is some $q \in \mathbb{R}^K$ for which $||\nabla C(q) - x_p|| \leq \epsilon$.

We can interpret these conditions as

1. Condition 1: The gradient of C, denoted $\nabla C(q)$ is well defined.
2. Condition 2: There can not be more than one way of expressing one's information. This translate in the convexity of the cost function given by Equation (13.4.2).
3. Condition 3: It is never possible for a trader to purchase a security bundle r and receive a positive profit regardless of the outcome.
4. Condition 4: The gradient $\nabla C(q)$ represents the traders' current estimates of the expected payoff of each security denoted by x_p.

Note, for a given time t, we have $C(q_{t+1}) - C(q_t) \approx \nabla C(q_t).(q_{t+1} - q_t) = x_t.r_t$, where $q_t = r_1 + ... + r_t$ is the vector of previous purchases. Hence, we see that the No-Arbitrage Equation (13.4.3) implies that the instantaneous price x_p must be greater or equal to the payoff $\rho(o)$ for each security $o \in \mathcal{O}$ (depending on the future outcome o). Strong of these conditions, Abernethy et al. formalised the pricing framework as

1. a pricing mechanism can always be described precisely in terms of a convex cost function C.
2. the set of reachable prices of a mechanism, that is the set $\{\nabla C(q) : q \in \mathbb{R}^K\}$, must be identically the convex hull of the payoff vectors for each outcome $\mathcal{H}(\rho(\mathcal{O}))$ except possibly differing at the relative boundary of $\mathcal{H}(\rho(\mathcal{O}))$.

For complete markets, it implies that the set of achievable prices should be the convex hull of the n standard basis vectors. The conditions introduced in the complex space imply that in this setting, we should use a cost function based market with a convex, differentiable cost function such that

$$\text{closure } (\{\nabla C(q) : q \in \mathbb{R}^K\}) = \text{ closure } (\mathcal{H}(\rho(\mathcal{O})))$$

The condition 2 on information incorporation ensures the convexity of the cost function via Equation (13.4.2). Recall, the set of derivatives $\{\nabla C(q) : q \in \mathbb{R}^K\}$ of any convex function C must form a convex set. Further, condition 4 is equivalent to the statement that every element $x_p \in \mathcal{H}(\rho(\mathcal{O}))$ is a limit point of the set $\{\nabla C(q) : q \in \mathbb{R}^K\}$. Since $\{\nabla C(q) : q \in \mathbb{R}^K\} \subseteq$ closure $(\mathcal{H}(\rho(\mathcal{O})))$ the only case where x_p does not equal $\nabla C(q)$ for some q is when x_p lies on the relative boundary of $\mathcal{H}(\rho(\mathcal{O}))$.

Abernethy et al. used convex analysis, and in particular, the notion of conjugate duality (see Appendix (A.1.4)) to design and compare properties of cost functions satisfying these criteria. The result being, pick any closed strictly convex function R with domain containing $\mathcal{H}(\rho(\mathcal{O}))$, and set $C := R^*$. We get the following steps

- Input: security space $\mathbb{R}^k$ and a bounded payoff function $\rho : \mathcal{O} \to \mathbb{R}^K$.
- Input: convex compact price space Π, typically assumed to be $\mathcal{H}(\rho(\mathcal{O}))$.
- Input: closed strictly convex and differentiable R with relint $(\Pi) \subseteq$ dom (R).
- Output cost function $C : \mathbb{R}^K \to \mathbb{R}$ defined by

$$C(q) = \sup_{x \in \text{ relint } (\Pi)} x.q - R(X)$$

Hence, the space of feasible price vectors should be $\Pi = \mathcal{H}(\rho(\mathcal{O}))$, the convex hull of the payoff vectors for each outcome. Note, the duality based approach leads to markets that are efficient to implement whenever $\mathcal{H}(\rho(\mathcal{O}))$ can be described by a polynomial number of simple constraints. Further, since the construction of C is based on convex programming, the problem of automated market making is reduced to the problem of optimisation. At last, this methods yields simple formulas for properties of markets, such as the worst-case monetary loss and the worst-case information loss. That is, the choice of the conjugate function R impacts market properties, and in many situations, an ideal choice would be

$$R(x) = \frac{\lambda}{2}||x - x_0||^2 \tag{13.4.4}$$

which is the squared Euclidean distance between x and an initial price vector $x_0 \in \Pi$, scaled by $\frac{\lambda}{2}$. The market maker can tune λ appropriately according to the desired tradeoff between worst-case market depth and worst-case loss. However, the tradeoff tighten when R has a Hessian that is uniformly a scaled identity matrix, or when R takes the form in Equation (13.4.4). Requiring the conjugate function R to be a pseudo-barrier, we make sure that the instantaneous price vector $\nabla C(q)$ always lies in relint (Π), and does not become a constant near the boundary. For instance, the negative entropy function

$$H(x) = \sum_i x_i \log x_i$$

defined on the n-simplex Δ_n is an example of a convex function that is simultaneously bounded and a pseudo-barrier. Note, the LMSR can be described by the choice

$$R(x) = bH(x)$$

where the price space is $\Pi = \Delta_n$. Note, assuming $\Pi = \mathcal{H}(\rho(\mathcal{O}))$ we can find scenarios for which this hull has a polynomial number of constraints, allowing to efficiently set prices via convex optimisation. However, this may not always be the case, especially when $\mathcal{H}(\rho(\mathcal{O}))$ has exponentially (or infinitely) many constraints. In this case, Abernethy et al. proposed to use an efficient separation oracle to get alternative methods for optimisation. If such a tool is not available, they suggested to modify $\mathcal{H}(\rho(\mathcal{O}))$ to get an alternate price space Π which could work more efficiently. In order to obtain some relaxation of the price space, they allowed for Π to be distinct from $\mathcal{H}(\rho(\mathcal{O}))$, and showed that the no-arbitrage condition 3 could be relaxed but not the expressiveness condition 4.

Theorem 13.4.1 *For any duality-based cost function market maker, the worst-case loss of the market maker is unbounded if* $\rho(\mathcal{O}) \nsubseteq \Pi$.

That is, condition 4 is necessary for the market maker to avoid unbounded loss. If o is the final outcome and $\rho(o) \notin \Pi$, then there exists a $k > 0$ such that $||\rho(0) - \nabla C(q)|| \geq k$, $\forall k$, and it is possible to make an infinite sequence of trades such that each trade causes a constant amount of loss to the market maker. On the other hand, one can choose π to be a superset of $\mathcal{H}(\rho(\mathcal{O}))$, since expanding Π do not hurt the market maker. As long as the initial price vector lies in $\mathcal{H}(\rho(\mathcal{O}))$, any such situations where a trader can earn a guaranteed profit are effectively created (and paid for) by other traders. If the final price vector $\nabla C(q)$ falls outside the convex hull, the divergence term will be strictly positive, improving the bound.

13.4.3 Relating market design to machine learning

Even tough the problem of learning in an online environment is semantically distinct from the problem of pricing securities in a prediction market, the tools developed are similar. A learning algorithm receives losses and selects weights according to the following steps

- the learner is given access to a fixed space of weights $\mathcal{K}$
- the learning algorithm must select a weight vector $w \in \mathcal{K}$
- the learner uses a convex regulariser $\mathcal{R}(\bullet)$, which is a parameter of FTRL
- the learner receives loss vectors l_t
- the learning algorithm maintains a cumulative loss vector L_t and updates according to

$$L_{t+1} \leftarrow L_t + l_l$$

- FTRL selects the weight vector by solving

$$w_{t+1} = \arg\min_{w \in \mathcal{K}} w.L_t + \frac{1}{\eta}\mathcal{R}(w)$$

- the learner suffers regret

$$R_T = \sum_{t=1}^{T} w_t.l_t - \min_{w \in \mathcal{K}} w.L_T$$

On the other hand, a market maker manages trades and set prices according to the following steps

- the market maker has an outcome space $\mathcal{O}$ and a payoff function $\rho : \mathcal{O} \to \mathbb{R}^K$, which define a feasible price space $\Pi = \mathcal{H}(\rho(\mathcal{O}))$
- the market maker must select instantaneous security prices $x \in \Pi$
- the market maker uses a convex conjugate $R(\bullet)$, which is a parameter of the pricing function $C(\bullet)$
- the market maker receives security bundle purchases r_t
- the market maker maintains a quantity vector q_t and updates according to

$$q_{t+1} \leftarrow q_t + r_t$$

- the market mechanism sets prices via

$$x_{t+1} = \arg\max_{x \in \Pi} x.q_t - R(x)$$

- the market maker suffers worst-case loss

$$C(q_0) - C(q_T) + \max_{x \in \Pi} x.q_T$$

These algorithms emphasise the fact that we can identify the objects Π, $R(\bullet)$, and $\{r_t\}$ with the objects $\mathcal{K}$, $\frac{\mathcal{R}(\bullet)}{\eta}$, and $\{-l_t\}$, respectively. As a result, the mechanisms for choosing an instantaneous price vector $x_t \in \Pi$ and selecting a weight vector $w_t \in \mathcal{K}$ are identical. That is, considering the security bundles r_t as the negative loss vectors l_t, the market mechanism becomes the FTRL in the above algorithm. Note, in the last pair of statements, the FTRL regret and the market maker's worst case loss seems very different but they are not so far apart. First, the term $\max_{x\in\Pi} x.q_T$ matches the term $-\min_{w\in\mathcal{K}} w.L_T$. Second, doing a first-order approximation on the first term we get

$$C(q_T) - C(q_0) = \sum_{t=1}^{T} C(q_{t+1}) - C(q_t) \approx \sum_{t=1}^{T} \nabla C(q_t).(q_{t+1} - q_t) = \sum_{t=1}^{T} x_t.r_t$$

since the instantaneous price vector x_t is equal to $\nabla C(q_t)$. Hence, the total earned by the market maker $C(q_T)-C(q_0)$ is roughly the sum of these payments over all trades. To conclude, the expert setting ($\mathcal{K} = \Delta_n$) correspond to complete markets. Weighted Majority corresponds to the LMSR with the learning rate η playing the same role as the parameter b (see Chen et al. [2008]).

13.4.4 The assumptions of market completeness

Even though no particular distributions have been specified for the determination of contingent securities, we saw that a set of conditions, or axioms, was necessary in order to define the expected behaviour of a market. These conditions have naturally been selected to satisfy the complete market theory underlying the efficient market hypothesis (EMH) discussed throughout this guide. The EMH lead to the notion of independent identically distributed prices and martingale processes. While path independence helps reducing arbitrage opportunities, it also helps reducing the strategic play of traders, since traders need not reason about the optimal path leading to some target position. One major consequence of the assumption of path independence on the security purchases was that they could be represented by a convex cost function. We saw in Section (13.4) that the space of feasible price vectors should be $\Pi = \mathcal{H}(\rho(\mathcal{O}))$, the convex hull of the payoff vectors for each outcome. These results have been formalised a long time ago in the option pricing theory (OPT) and extensively commented since then (see Appendix (F.3.3)). In a complete market, we must assume that contingent claims have convex payoffs and that the volatility depends only on time and the current stock price. As a result, the price of the contingent claim is a convex function of the price of the stock. However, if one of these assumptions is violated, contingent claims can be non-increasing, non-convex. Abernethy et al. [2012] showed that including additional price vectors in Π does not impact the market maker's worst-case loss, as long as the initial price vector lies in $\mathcal{H}(\rho(\mathcal{O}))$, even though the no-arbitrage condition is violated. However, traders will incur the costs.

13.5 Presenting scoring rules

13.5.1 Describing a few scoring rules

13.5.1.1 The proper scoring rules

While Sroring Rules (SR) have been used in the evaluation of probabilistic forecasts, in the context of information elicitation SR are used to encourage individuals to make careful assessments and truthfully report their beliefs. In the context of machine learning, SR are used as loss functions to evaluate and compare the performance of different algorithms (see Reid et al. [2009]). We let $\{1, ..., n\}$ be a set of mutually exclusive and exhaustive outcomes of a future event. A scoring rule s maps a probability distribution p over outcomes to a score $s_i(p)$ for each outcome i, with $s_i(p)$ taking values in the range $[-\infty, \infty]$. This score represents the reward received by a forecaster for predicting the distribution p if the outcome turns out to be i. A scoring rule is said to be regular relative to the probability simplex Δ_n if $\sum_{i=1}^n p_i s_i(p^{'}) \in [-\infty, \infty)$ for all p, $p^{'} \in \Delta_n$, with $\sum_{i=1}^n p_i s_i(p) \in (-\infty, \infty)$, implying that $s_i(p)$ is finite whenever $p_i > 0$. Hence, a scoring rule is said to be proper if a risk-neutral forecaster believing the true distribution over outcomes is p has no incentive to report any alternate distribution $p^{'}$, that is, if $\sum_{i=1}^n p_i s_i(p) \geq \sum_{i=1}^n p_i s_i(p^{'})$

for all distributions $p^{'}$. The rule is strictly proper if this inequality holds with equality only when $p = p^{'}$. For instance, the quadratic scoring rule (see Brier [1950])

$$s_i(p) = a_i + b\big(2p_i - \sum_{i=1}^{n} p_i^2\big)$$

and the logarithmic scoring rule (see Good [1952])

$$s_i(p) = a_i + b \log p_i \tag{13.5.5}$$

where $b > 0$ and a_i for $i = 1, .., n$ are parameters, are examples of regular, strictly proper scoring rules used both in information elicitation and machine learning. The following characterisation theorem of Gneiting et al. [2007] gives the precise relationship between convex functions and proper scoring rules.

Theorem 13.5.1 *A regular scoring rule is (strictly) proper if and only if there exists a (strictly) convex function $G : \Delta_n \to \mathbb{R}$ such that for all $i \in \{1, .., n\}$,*

$$s_i(p) = G(p) - G^{'}(p).p + G_i^{'}(p)$$

where $G^{'}(p)$ is any subgradient of G at the point p, and $G_i^{'}(p)$ is the ith element of $G^{'}(p)$.

Note, for a scoring rule defined in terms of a function G,

$$\sum_{i=1}^{n} p_i s_i(p) = \sum_{i=1}^{n} p_i\big(G(p) - G^{'}(p).p + G_i^{'}(p)\big) = G(p)$$

the above theorem indicates that a regular scoring rule is (strictly) proper if and only if its expected score function $G(p)$ is (strictly) convex on Δ_n, and the vector with elements $s_i(p)$ is a subgradient of G at the point p. Therefore, every bounded convex function G over Δ_n induces a proper scoring rule. Gneiting et al. [2007] detail the properties and characterisations of proper scoring rules. Some important results states that if we define $S(\tilde{p}, p) = \sum_{i=1}^{n} p_i s_i(\tilde{p})$ as the expected score of a forecaster with belief p but predicts $\tilde{p}$, then $G(p) = S(p, p)$. Further, if a scoring rule is regular and proper, then $d(\tilde{p}, p) = S(p, p) - S(\tilde{p}, p)$ is the associated divergence function that captures the expected score loss ifa forecaster predicts $\tilde{p}$ rather than his true belief p. Also, it is known that if $G(p)$ is differentiable, the divergence function is the Bregman divergence (see Appendix (A.1.6)) for G, that is, $d(\tilde{p}, p) = D_G(\tilde{p}, p)$.

13.5.1.2 The market scoring rules

The Market Scoring Rules (MSR) introduced by Hanson [2003] [2007] are sequentially shared scoring rules. More precisely, the market maintains a current probability distribution p, and at any time t, a trader can enter the market and change this distribution to an arbitrary distribution $p^{'}$ of his choice. If the outcome turns out to be i, the trader receives the (possibly negative) payoff $s_i(p^{'}) - s_i(p)$. For instance, when using the logarithmic scoring rule in Equation (13.5.5), a trader changing the distribution from p to $p^{'}$ receives the payoff $b \log \frac{p_i^{'}}{p_i}$, which is equivalent to the cost function based formulation of the LMSR in the sense that a trader changing the market probabilities from p to $p^{'}$ in the MSR formulation receives the same payoff for every outcome i as a trader changing the quantity vectors from any q to $q^{'}$, such that market prices become $x(q) = p$ and $x(q^{'}) = p^{'}$ in the cost function based formulation. We see that using proper scoring rules, MSR preserve the nice incentive compatible property of proper scoring rules for myopic traders. Hence, a trader believing the true distribution to be p, and only caring about the payoff of his action, maximises his expected payoff by changing the market's distribution to p.

One advantage of the MSR is the simplicity to bound the market maker's worst case loss, since each trader is responsible for paying the previous trader's score, so that the market maker is only responsible for paying the score of

the final trader. If we let p_0 be the initial probability distribution of the market, then the worst case loss of the market maker is given by

$$\max_{i\in\{1,..,n\}} \sup_{p\in\Delta_n} \big(s_i(p) - s_i(p_0)\big)$$

Note, the LMSR market maker is not the only market which can be defined either as a market scoring rule, or as a cost function based market. In fact, Chen et al. [2007] noted a correspondence between certain market scoring rules and certain cost function based markets. They showed that the MSR with scoring function s and the cost function based market with cost function C are equivalent if for all q and all outcomes i, we get $C(q) = q_i - s_i(x(q))$.

13.5.2 Relating MSR to cost function based market makers

Abernethy et al. [2012] proposed very general conditions under which a market scoring rule (MSR) is equivalent to a cost function based market, and they provided a way of translating a MSR to a cost function based market, and vice versa. Given the cost function in Equation (13.4.1) we let R_C denote the function R corresponding to the cost function C. According to Theorem (13.5.1), there is one-to-one and onto mapping between strictly convex and differentiable R_C and strictly proper, regular scoring rules with differentiable scoring functions $s_i(x)$, where for evry pair we have

$$R_C(x) = \sum_{i=1}^{n} x_i s_i(x)$$

and

$$s_i(x) = R_C(x) - \sum_{j=1}^{n} \frac{\partial R_C(x)}{\partial x_j} x_j + \frac{\partial R_C(x)}{\partial x_i}$$

Abernethy et al. proved the following theorem showing that the cost function based market using R_C and the MSR using $s_i(x)$ are equivalent in term's of trader's profit and reachable price vectors.

Theorem 13.5.2 *Given a pair of strictly convex, differentiable $R_C(x)$ and strictly proper, regular scoring rule with differentiable scoring functions $s_i(x)$ defined above, the corresponding cost function based market and market scoring rule market are equivalent in the following two aspects*

- *The profit of a trade is the same in the two markets if the trade starts with the same market prices and results in the same market prices and the prices for all outcomes are positive before and after the trade.*
- *Every price vector p achievable in the market scoring rule is achievable in the cost function based market.*

13.6 Introduction to artificial neural networks

13.6.1 Neural networks

We saw in Section (6.1.1.2) that we could use wavelet analysis to decompose complex system into simpler elements, in order to understand them. We are now going to show how to gather simple elements to produce a complex system. Networks is one approach among others achieving that goal. They are characterised by a set of interconnected nodes seen as computational units receiving inputs and processing them to obtain an output. The connections between nodes, which can be unidirectional or bidirectional, determine the information flow between them. We obtain a global behaviour of the network, called emergent, since the abilities of the network supercede the one of its elements, making networks a very powerful tool. Since Neural Nets have been widely studied by computer scientists, electronic engineers, biologists, psychologists etc, they have been given many different names such as Artificial Neural Networks

(ANNs), Connectionism or Connectionist Models, Multi-layer Percepetrons (MLPs), or Parallel Distributed Processing (PDP) to name a few. However, a small group of classic networks emerged as dominant such as Back Propagation, Hopfield Networks (see Hopfield [1982]), Competitive Networks and networks using Spiky Neurons.

Among the different networks existing, the artificial neural networks (ANNs) and the artificial recurrent neural networks (RNNs) are computational models designed by more or less detailed analogy with biological brain modules. They are inspired from natural neurons receiving signals through synapses located on the dendrites, or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal through the axon, which might be sent to another synapse, and might activate other neurons. ANN is a high level abstraction of real neurons consisting of inputs (like synapses) multiplied by weights (strength of the respective signals), and computed by a mathematical function determining the activation of the neuron. Another function computes the output of the artificial neuron (sometimes in dependence of a certain threshold). Depending on the weights (which can be negative), the computation of the neuron will be different, such that by adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. In presence of a large number of neurons we must rely on an algorithm to adjust the weights of the ANN to get the desired output from the network. This process of adjusting the weights is called learning or training. The aim of training the network is to obtain the optimal network minimising the difference between the output and a target, given some input data. Then, the output of the optimal network becomes the optimal predictor. The optimisation process is usually done by gradient descent method, where the weights are updated by an amount being equal to the opposit of the gradient. This way of updating weights is called the standard back propagation.

13.6.1.1 The mathematical formalism

In machine learning, a task consist in learning a functional relation between a given input $U(p) \in \mathbb{R}^{N_i}$ and a desired output $\widehat{Y}(p) \in \mathbb{R}^{N_o}$, for the set $p = 1, ..., P$ where P is the number of data points in the training data set $\{(U(p), \widehat{Y}(p))\}$. In a non-temporal task, the data points are independent of each other and the goal is to learn a function

$$Y(p) = f(U(p))$$

minimising an error measure $E = E(Y, \widehat{Y})$. In a temporal task, the input signal and target signal are set in a discrete time domain $t = 1, ..., T$ and the function we are trying to learn has memory. We are going to discuss both non-temporal and temporal tasks, but we first choose to describe the mathematical formalism in the case of temporal tasks. We let the time-varying input signal be an N_ith order column vector $U(t) = [u_i(t)]$, the generated output is an N_oth order column vector $Y(t) = [y_o(t)]$ and the target output $\widehat{Y}(t)$ is an N_oth order column vector, where $t = 1, .., T$ and T is the number of data points in the training set $\{(U(t), \widehat{Y}_t)\}$. The goal is to learn a function

$$Y(t) = f(..., U(t-1), U(t))$$

such that $E(Y, \widehat{Y})$ is minimised, where E is an error measure. The relation between the input $U(t)$ and the desired output $\widehat{Y}(t)$ can either be solved by a linear model as

$$Y(t) = W^\top U(t)$$

where $W \in \mathbb{R}^{N_i \times N_o}$, or by a nonlinear model.

***Remark* 13.6.1** *In general, the weight matrix is defined as $W \in \mathbb{R}^{N_o \times N_i}$ where the weight $[w_{ij}]$ goes from the jth input node to the ith output node. Instead, we define $W \in \mathbb{R}^{N_i \times N_o}$ and transpose it, to let the weight $[w_{ij}]$ goes from the ith input node to the jth output node.*

In nonlinear models, we generally expand nonlinearly the input $U(t)$ into a high dimensional feature vector $X(t) \in \mathbb{R}^{N_x}$, and use linear models to get a reasonable output vector $Y(t)$. The output vector can be written as

$$Y(t) = f_{out}(W_{out}^{\top} X(t)) = f_{out}(W_{out}^{\top} \phi(..., U(t-1), U(t)))$$

where $f_{out}(\bullet)$ is the output function (identity, sigmoid, or other), and $W_{out} \in \mathbb{R}^{N_x \times N_o}$ are the trained output weights. The functions

$$X(t) = \phi(..., U(t-1), U(t))$$

transforming the current input $U(t)$ and its history $U(t-1), ...$ into a higher dimensional vector $X(t)$ are called kernels, and we refer to them in machine learning as expansion methods. Since these kernels have an unbounded number of parameters, we define them recursively as

$$X(t) = \phi(X(t-1), U(t)) \tag{13.6.6}$$

Expansion methods includs, among others, Support Vector Machines, Feedforward Neural Networks, Radial Basis Function approximators, Slow Feature Analysis, various Probability Mixture models. As an example of such kernels, the recurrent neural networks (RNNs) can be written as

$$X(t+1) = f\big(W_{res}^{\top}.X(t) + W_{in}^{\top}.U(t+1) + W_{fb}^{\top}.Y(t)\big) \tag{13.6.7}$$

where $f(\bullet)$ is an activation function, $W_{in} \in \mathbb{R}^{N_i \times N_x}$ is an input weight matrix, $W_{res} \in \mathbb{R}^{N_x \times N_x}$ is a connection weight matrix, and $W_{fb} \in \mathbb{R}^{N_o \times N_x}$ is a feedback weight matrix. In addition to the function $f(\bullet)$, leaky integrator neurons performs a leaky integration of its activation from previous time steps, which can be applied before or after the activation function $f(\bullet)$. In the latter case, assuming no feedback, we get

$$X(t+1) = (1 - a\Delta t)X(t) + \Delta t f\big(W_{in}^{\top} U(t+1) + W_{res}^{\top} X(t)\big)$$

where Δt is a compound time gap between two consecutive time steps divided by the time constant of the system, and a is the delay (or leakage) rate. Setting $a = 1$ and redefining Δt as a constant α controlling the speed of the dynamics, we get

$$X(t+1) = (1 - \alpha)X(t) + \alpha f\big(W_{in}^{\top} U(t+1) + W_{res}^{\top} X(t)\big)$$

which is an exponential moving average. In general, we set $a\Delta t \in [0, 1]$ in the first equation and $\alpha \in [0, 1]$ in the second such that a neuron neither retain, nor leak, more activation than it had. Thus, with one extra parameter we make sure that neuron activations $X(t)$ never go outside the boundary defined by $f(\bullet)$. In fact, the neuron activation performs a low-pass filtering of its activations with the cutoff frequency

$$f_c = \frac{a}{2\pi(1-a)\Delta t}$$

where Δt is the discretisation time step, allowing to tune the reservoirs for particular frequencies. Non-temporal tasks using feedforward networks are functions, while RNNs may develop self-sustained temporal activation dynamics making them dynamical systems. Generally, supervised training of RNNs, such as gradient descent, adapt iteratively all weights according to their estimated gradients $\frac{\partial E}{\partial W}$ to minimise the output error $E = E(Y, \widehat{Y})$. One can adapt backpropagation (BP) methods from feedforward neural networks in the case of RNNs, by propagating the gradient through network connections and time. For example, the Backpropagation Through Time (BPTT) has a runtime complexity of $O(N_x^2)$ per weight update per time step for a single output ($N_o = 1$). Alternatively, the Real-Time Recurrent Learning (RTRL) estimate the gradients recurrently, forward in time, but it has a runtime complexity of $O(N_x^4)$. Improvements in standard RNNs design were proposed independently by Maass et al. [2002] under the name of Liquid State Machines and Jaeger [2001] under the name of Echo State Networks. Over time, these types of models became known as Reservoir Computing.

13.6.1.2 Presentating ANNs

Artificial neural networks (ANNs) provide a general, practical method for learning real-valued, discrete-valued, and vector-valued functions from examples. The basic architecture of an ANN is a network comprised of nodes (called *neurons* in biology) interacting with each other via incoming and outgoing weighted connections indicating the strength degree of dendrite between neurons. Given some input, the network will be activated and this activation signal will be passed throughout the rest of the network through the connections. Generally, an ANN model is specified by its topology, node characteristics and the training algorithms. There are a variety of ANNs with various topologies, node properties and training algorithms. Two of the most important distinct types of ANNs are feedforward and feedback network. ANNs with cycles are *feedback* networks, which are also referred to as *recurrent* neural networks (RNNs). Those networks with acyclic connections are called *feedforward* neural networks (FNNs). Algorithms such as back-propagation use gradient descent to tune network parameters to best fit a training set of input-output pairs, making the learning process robust to errors in the training data. Various successful applications to practical problems developed, such as learning to recognise handwritten characters (see LeCun et al. [1989]) and spoken words (see Lang [1990]), learning to detect fraudulent use of credit cards, drive autonomous vehicles on public highways (see Pomerleau [1993]). Rumelhart et al. [1994] provided a survey of practical applications. Since McCulloch et al. [1943] introduced the first ANN model, various models where developed with different functions, accepted values, topology, learning algorithms, hybrid models where each neuron has a larger set of properties etc. While there is various types of classifiers, neural network based classifiers dominate the literature. Yet, compared to traditional NNs, higher order neural networks (HONNs) have several unique characteristics, including

- stronger approximation with faster convergence property
- greater storage capacity
- higher fault tolerance capability

However, its major drawback is the exponential growth in the number of weights with the increasing order of the network. As a special case of the feedforward HONN, the Pi-Sigma networks (PSNs) introduced by Shin et al. [1991] have the capability of higher order neural networks, but using a smaller number of weights. In order to enhance the learning capability of neural network, many researchers have improved the system by combining other techniques such as fuzzy logic (see Zadeh [1994]), genetic algorithm (see Harrald et al. [1997]).

For simplicity of exposition we are briefly going to describe the back-propagation algorithm proposed by Rumelhart et al. [1986]. The *multilayer perceptron* (MLP) is formed in layers, the idea being that multilayer ANN can approximate any continuous function. This algorithm is a layer feed-forward ANN, since the artificial neurons are organised in layers with signals sent forward and with errors propagated backwards. The network receives inputs via neurons in the input layer, and the output of the network is given via neurons on the output layers. There may be one or more intermediate hidden layers. For simplicity, we consider a network with three layers, illustrated in Figure 13.1, where the bottom layer is the *input layer*. Then the information is propagated to the middle layer, called the *hidden layer*. Finally, the output layer receives the incoming value. This process is termed the *forward pass*. Note that the S-shaped curve is the squashing function which forces the output of the unit to fall between certain bounded interval. Except for the input layer, the output value of each unit in the other layers should be passed through the squashing function before transferring to the next layer. The back-propagation algorithm uses supervised learning where we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. Defining the network function f_{out} as a particular implementation of a composite function from input to output space, the learning problem consists in finding the optimal combination of weights so that f_{out} approximates a given function f as closely as possible. However, in practice the function f is not given explicitly but only implicitly through some examples. Honik et al. [1989] showed that a MLP with one hidden layer and sufficient nonlinear units could approximate any continuous function on a compact input domain to arbitrary precision. Therefore, MLP are recognised as universal approximators.

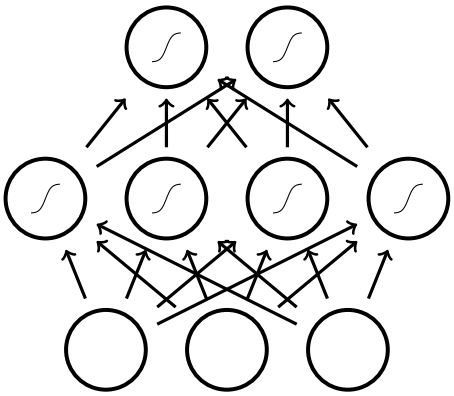


Figure 13.1: **A multilayer perceptron.** The three layers are input, hidden, output layers repectively from bottom to top. The S-shaped curves in the hidden and output layers are the sigmoid squashing functions applying to the net values.

13.6.2 Gradient descent and the delta rule

One way forward to finding the optimal combination of weights is to consider the delta rule which uses gradient descent to search the hypothesis space of possible weight vectors to find the weights best fitting the training examples. The gradient descent provides the basis for the backpropagation algorithm, which can learn networks with many interconnected units. Given a vector of input $x \in \mathbb{R}^n$ together with a weight vector $w \in \mathbb{R}^n$, we consider the task of training an unthresholded perceptron, that is, a linear unit for which the output O is given by

$$O(x) = w.x$$

A linear unit is the first stage of a perceptron without the threshold. We must specify a measure for the training error of a hypothesis (weight vector), relative to the set $\mathcal{D}$ of training examples. A common measure is to use

$$E(w) = \frac{1}{2}\sum_{p=1}^{P}(O_p - d_p)^2$$

where d_p is the target output for training example p and O_p is the output of the linear unit for training example p. One can show that under certain conditions, the hypothesis minimising E is also the most probable hypothesis in H given the training data. The entire hypothesis space of possible weight vectors and their associated E values produce an error surface. Given the way in which we defined E, for linear units, this error surface must always be parabolic with a single global minimum. The gradient descent search determines a weight vector minimising E by starting with an arbitrary initial weight vector, and then repeatedly modifying it in small steps. At each step, the weight vector is altered in the direction that produces the steepest descent along the error surface, until a global minimum error is reached (see details in Appendix (A.9)). This direction is found by computing the derivative of E with respect to each component of the vector w, called the gradient of E with respect to w, given by

$$\nabla E(w) = [\frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_n}]$$

The gradient specifies the direction that produces the steepest increase in E, and its negative produces the steepest decrease. Hence, the training rule for gradient descent is

$$w \leftarrow w + \Delta w$$

where

$$\Delta w = -\eta \nabla E(w)$$

where $\eta > 0$ is the learning rate determining the step size in the gradient descent search. The negative sign implies that we move the weight vector in the direction that decreases E. Given the definition of $E(w)$ we can easily differentiate the vector of $\frac{\partial E}{\partial w_i}$ derivatives as

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{p=1}^{P} (O_p - d_p)^2 \ , i = 1, ..., n$$

which gives after some calculation

$$\frac{\partial E}{\partial w_i} = \sum_{p=1}^{P} (O_p - d_p) x_{i,p}$$

where $x_{i,p}$ is the single input component x_i for training example p. The weight update rule for gradient descent is

$$\Delta w_i = -\eta \sum_{p=1}^{P} (O_p - d_p) x_{i,p} = \eta \sum_{p=1}^{P} (d_p - O_p) x_{i,p}$$

In the case where η is too large, the gradient descent search may overstep the minimum in the error surface rather than settling into it. One solution is to gradually reduce the value of η as the number of gradient descent steps grows. Gradient descent is a strategy for searching through a large or infinite hypothesis space which can be applied whenever

1. the hypothesis space contains continuously parameterised hypotheses.
2. the error can be differentiated with respect to these hypothesis parameters.

The key practical difficulties in applying gradient descent are

1. converging to a local minimum can sometimes be quite slow.
2. if there are multiple local minima in the error surface, then there is no guarantee that the procedure will find the global minimum.

While the gradient descent training rule computes weight updates after summing over all the training examples in $\mathcal{D}$, the stochastic gradient descent approximate this gradient descent search by updating weights incrementally, following the calculation of the error for each individual example. Hence, as we iterate through each training example, we update the weight according to

$$\Delta w_i = \eta (d - O) x_i$$

where the subscript i represents the ith element for the training example in question. One way of viewing this stochastic gradient descent is to consider a distinct error function $E_p(w)$ defined for each individual training example p as follow

$$E_p(w) = \frac{1}{2} (d_p - O_p)^2$$

where d_p and O_p are the target value and the unit output value for training example p. We therefore iterates over the training examples p in $\mathcal{D}$, at each iteration altering the weights according to the gradient with respect to $E_p(w)$. The sequence of these weight updates provides a reasonable approximation to descending the gradient with respect to the original error function $E(w)$. This training rule is known as the delta rule, or sometimes the least-mean-square (LMS) rule. Note, the delta rule converges only asymptotically toward the minimum error hypothesis, but it converges regardless of whether the training data are linearly separable.

13.6.3 Introducing multilayer networks

While a single perceptron can only express linear decision surfaces, multilayer networks are capable of expressing a rich variety of nonlinear decision surfaces. Since multiple layers of cascaded linear units can only produce linear functions, we need to introduce another unit to represent highly nonlinear functions. That is, we need a unit whose output is a nonlinear function of its inputs, but which is also a differentiable function of its input. We are now going to discuss such a unit and then describe how to apply the gradient descent in the case of multilayer networks.

13.6.3.1 Describing the problem

We consider a feedforward network with N_i input units, N_o output units and N_k hidden layers which can exhibit any desired feedforward connection pattern. We are also given a training set $\{(\overline{x}_1, \overline{d}_1), ..., (\overline{x}_P, \overline{d}_P)\}$ consisting of P ordered pairs of n- and m-dimensional vectors called the input and output patterns, respectively. Note, the training set may consits of sequential data, or time-varying input, which we then represent as $\{(\overline{x}_1, \overline{d}_1), ..., (\overline{x}_T, \overline{d}_T)\}$ with T ordered pairs. We assume that the primitive functions at each node of the network is continuous and differentiable, and that the weights of the edges are real numbers selected randomly so that the output $\overline{O}_p$ of the network is initially different from the target $\overline{d}_p$. The idea being to minimise the distance between $\overline{O}_p$ and $\overline{d}_p$ for $p = 1, ..., P$ by using a learning algorithm searching in a large hypothesis space defined by all possible weight values for all the units in the network. There are different measures of error (see Section (14.2.1.2)), and for simplicity we let the error function of the network be given by

$$E = \frac{1}{2} \sum_{p=1}^{P} ||\overline{O}_p - \overline{d}_p||_2^2$$

which is a sum of L_2 norms. After minimising the function with a training set, we can consider new unknown patterns and use the network to interpolate it. We use the backpropagation algorithm to find a local minimum to the error function by computing recursively the gradient of the error function and correcting the initial weights. Every one of the j output units of the network is connected to a node evaluating the function $\frac{1}{2}(O_{p,j} - d_{p,j})^2$ where $O_{p,j}$ and $d_{p,j}$ denote the jth component of the output vector $\overline{O}_p$ and the target vector $\overline{d}_p$, respectively. The outputs of the additional m nodes are collected at a node which adds them up and gives the sum E_p as its output. The same network extension is built for each pattern $\overline{d}_p$. We can then collect all the quadratic errors and output their sum, obtaining the total error for a given training set $E = \sum_{p=1}^{P} E_p$. Since E is computed exclusively through composition of the node functions, it is a continuous and differentiable function of the q weights $w_1, .., w_q$ in the network. Therefore, we can minimise E by using an iterative process of gradient descent where we need to calculate the gradient

$$\nabla E = (\frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_q})$$

and each weight is updated with the increment

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i} \text{ for } i = 1, .., q$$

where η is a learning constant, that is, a proportionality parameter defining the step length of each iteration in the negative gradient direction. Hence, once we have a method for computing the gradient,we can adjust the weights iteratively until we find a minimum of the error function where $\nabla E \approx 0$.

***Remark* 13.6.2** *In the case of multilayer networks, the error surface can have multiple local minima. Hence, the minimisation by steepest descent will only produce a local minimum of the error function, and not necessarily the global minimum error.*

13.6.3.2 Describing the algorithm

For simplicity of exposition we first describe the algorithm in the case where there is no hidden layer and the training set consists of a single input-output pair ($P = 1$). We assume a set of artificial neurons $\{f_{i,j}\}$, each receiving $\{x_i\}_{i=1}^{N_i}$ input and computing $\{O_j\}_{j=1}^{N_o}$ output, and we focus our attention on one of the weights, $w_{i,j}$, going from input i to neuron j in the network. We let W denote the $N_i \times N_o$ weight matrix with element $w_{i,j}$ at the ith row and the jth column and we let $w(j)$ be the $N_i \times 1$ row vector for the jth column. In the back-propagation algorithm, given the input vector $x = (x_1, .., x_{N_i})$, the activation function (also called net_j) for the jth neuron satisfies the weighted sum

$$A_j(x, w) = x.w(j) = \sum_{i=1}^{N_i} x_i w_{i,j} \ , j = 1, .., N_o$$

which only depends on the inputs and the weights $w_{i,j}$ from input i to neuron j. We let the output function (or threshold box, or activation function) be a function g of the activation function, getting

$$O_j(x, w) = g(\tilde{A}_j(x, w)) \ , j = 1, .., N_o$$

where $\tilde{A}_j(x, w) = A_j(x, w) + b_j$ and b_j is a bias value. In compact form, the output vector of all units is

$$O(x, w) = g(xW)$$

using the convention that we apply the function $g(\bullet)$ to each component of the argument vector. The simplest output function is the identity function. When using a threshold activation function, if the previous output of the neuron is greater than the threshold of the neuron, the output of the neuron will be one, and zero otherwise. Further, to simplify computation, the threshold can be equated to an extra weight. The error being the difference between the actual and the desired output, it only depends on the weights. Hence, to minimise the error by adjusting the weights, we define the error function for the output of each neuron. The error function for the jth neuron satisfies

$$E_j(x, w, d) = (O_j(x, w) - d_j)^2 \ , j = 1, .., N_o$$

where d_j is the jth element the desired target vector $\overline{d}$. In that setting, the total error of the network is simply the sum of the errors of all the neurons in the output layer

$$E(x, w, d) = \frac{1}{2}||\overline{O}(x, w) - \overline{d}(x)||_2^2 = \frac{1}{2}\sum_{j=1}^{N_o} E_j(x, w, d)$$

To minimise the error, we will update the weights of the network so that the expected error in the next iteration is lower using the method of gradient descent. Each weight is updated by using the increment

$$\Delta w_{i,j} = w_{i,j}^{l+1} - w_{i,j}^{l} = -\eta \frac{\partial E}{\partial w_{i,j}^{l}}$$

where l is the iteration counter, and $\eta \in (0, 1]$ is a learning rate. The size of the adjustment depends on η and on the contribution of the weight to the error of the function. The adjustment will be largest for the weight contributing the most to the error. We repeat the process until the error is minimal.

13.6.3.3 Describing the nonlinear transformation

We use the chain rule to compute the gradient of the error function. Since this method requires computation of the gradient of the error function at each iteration step, we must guarantee the continuity and differentiability of the error function. One way forward is to use the sigmoid, or logistic function (see details in Appendix (A.2)), a real function $f_c : \mathbb{R} \to (0, 1)$ defined as

$$f_c(x) = \frac{1}{1+e^{-cx}}$$

where the constant c can be chosen arbitrarily, and its reciprocal $\frac{1}{c}$ is called the temperature parameter. For $c \to \infty$ the sigmoid converges to a step function at the origin. Further, $\lim_{x\to\infty} f_c(x) = 1$, $\lim_{x\to-\infty} f_c(x) = 0$, and $\lim_{x\to 0} f_c(x) = \frac{1}{2}$. The first derivative of the sigmoid with respect to x is

$$\frac{d}{dx} f_c(x) = \frac{ce^{-cx}}{(1+e^{-cx})^2} = cf_c(x)(1-f_c(x)) \tag{13.6.8}$$

and the second derivative of the sigmoid with respect to x is

$$\frac{d^2}{dx^2} f_c(x) = cf_c^{'}(x) - 2cf_c(x)f_c^{'}(x)$$

where $f_c^{'}(x) = \frac{d}{dx} f_c(x)$. To get a symmetric output function, we can consider the symmetrical sigmoid $f_s(x)$ defined as

$$f_s(x) = 2f_1(x) - 1 = \frac{1-e^{-x}}{1+e^{-x}}$$

which is the *hyperbolic tangent* for the argument $\frac{x}{2}$, written $tanh(\frac{x}{2})$. An alternative solution is simply to linearly translate the domain of definition of the logistic function in the range $[b_L, b_H]$ where b_L is the lower bound and b_H the upper bound. Hence, the real function $f_{b,c} : \mathbb{R} \to (b_L, b_H)$ is defined as

$$f_{b,c}(x) = b_L + (b_H - b_L)f_c(x)$$

with $\lim_{x\to\infty} f_c(x) = b_H$, $\lim_{x\to-\infty} f_c(x) = b_L$, and $\lim_{x\to 0} f_c(x) = \frac{1}{2}(b_L + b_H)$. The derivative of this function with respect to x is

$$\frac{d}{dx} f_{b,c}(x) = (b_H - b_L)\frac{d}{dx} f_c(x) = (b_H - b_L)cf_c(x)(1-f_c(x))$$

Hence, in order to get a symmetric output function we can set $b_L = -1.5$ and $b_H = 1.5$. Example of translated logistic function and its derivative are given in Figure (13.2) and Figure (13.3) for $b_L = -0.6$ and $b_H = 0.6$ with $c = \frac{1}{p}$ for $p = 0.5, 1, 1.5$. Several other output functions can be used and have been proposed in the back-propagation algorithm. However, smoothed output function can lead to local minima in the error function which would not be there if the Heavised function had been used. Further, the slope of the sigmoid function given in Equation (13.6.8) takes values in the range $[0, \frac{1}{4}c]$ where the maximum is reached at $x = 0$. Thus, the steepest slope of the sigmoid function occurs at the origin, and depending on the size of the constant c, the slope can become greater than 1, leading to gradient explosion. In addition, the derivative of the sigmoid function is asymptotically zero for large values of x, such that the gradient will decay much more when the net value is taking large of small values. The decaying or explosion of the gradient are both called the vanishing gradient problem.

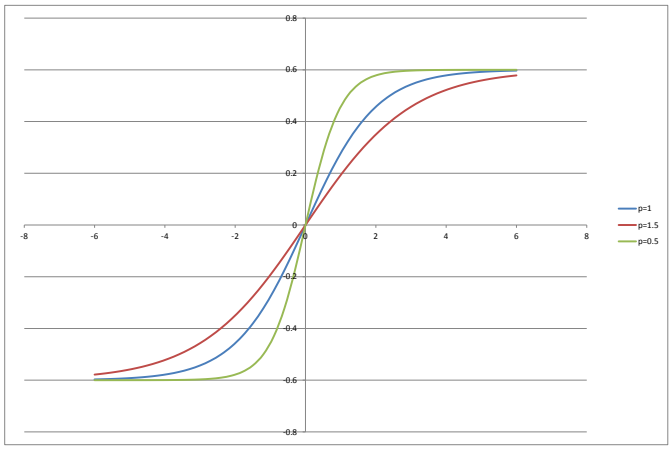


Figure 13.2: Translated logistic function for $b_L = -0.6$ and $b_H = 0.6$.

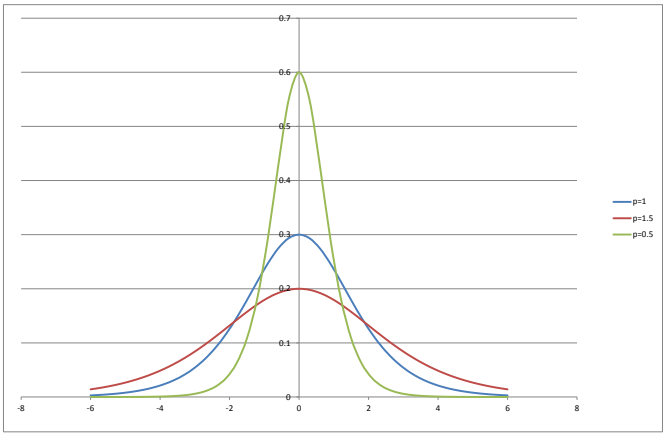


Figure 13.3: Derivative of the translated logistic function for $b_L = -0.6$ and $b_H = 0.6$.

13.6.3.4 A simple example

To explain the method, we first consider a two-layers ANN model, where the most common output function used with a threshold is the sigmoidal function

$$O_j(x, w) = \frac{1}{1 + e^{-A_j(x,w)}}$$

with the limits $\lim_{A_j \to \infty} = 1$, $\lim_{A_j \to -\infty} = 0$, and $\lim_{A_j \to 0} = \frac{1}{2}$ allowing for a smooth transition between the low and high output of the neuron. We compute the gradient of the total error with respect to the weight $w_{i,j}$, noted

$\nabla E(w_{i,j})$, where it is understood that the subscripts i and j are fixed integers within $i = 1, ..., N_i$ and $j = 1, .., N_o$. From the linearity of the total error function, the gradient is given by

$$\frac{\partial E}{\partial w_{i,j}} = \frac{1}{2}\sum_{j=1}^{N_o} \frac{\partial E_j}{\partial w_{i,j}} \, , i = 1, ..., N_i$$

where the subscripts i and j are fixed integers in the partial derivatives. We first differentiate the total error with respect to the output function

$$\frac{\partial E}{\partial O_j} = (O_j - d_j)$$

and then differentiate the output function with respect to the weights

$$\frac{\partial O_j}{\partial w_{i,j}} = \frac{\partial O_j}{\partial A_j}\frac{\partial A_j}{\partial w_{i,j}} = O_j(1 - O_j)x_i$$

since $\frac{\partial A_j}{\partial w_{i,j}} = x_i$. Putting terms together, the adjustment to each weight becomes

$$\Delta w_{i,j} = -\eta(O_j - d_j)O_j(1 - O_j)x_i = \eta(d_j - O_j)O_j(1 - O_j)x_i$$

We can use the above equation to train an ANN with two layers. Given a training set with p input-output pairs, the error function can be computed by creating p similar networks and adding the outputs of all of them to obtain the total error of the set.

13.6.4 Multi-layer back propagation

In the case of a multilayer network, again we consider a single input-output pair and assume a set of artificial neurons $\{f_{i,j,k}\}_{k=0}^{K}$ where the multilayer subscript $k = 0$ corresponds to the set of inputs $\{x_i\}_{i=1}^{N_i}$, and the remaining subscript k corresponds to the set of outputs $\{y_i\}_{i=1}^{N_{o,k}}$, where $N_{o,k}$ is the number of output in the k-th layer. In that setting, we define the output function of the jth node for the k-th layer as

$$O_{j,k}(x, w) = g(\tilde{A}_{j,k-1}(x, w)) \, , j = 1, .., N_{o,k}$$

where $\tilde{A}_{j,k-1}(x, w) = A_{j,k-1}(x, w) + b_{j,k}$ with the bias $b_{j,k}$, and we let the corresponding activation function satisfies

$$A_{j,k-1}(x, w) = \sum_{i=1}^{N_{o,k-1}} O_{i,k-1}(x, w)w_{i,j}^{k-1} \text{ and } O_{j,0}(x, w) = A_{j,-1}(x, w) = \sum_{i=1}^{N_i} x_i$$

where $w_{i,j}^{k-1}$ is the weight going from the ith node in layer $k-1$ to the jth node in layer k. Note, since $\frac{d\tilde{A}_{j,k-1}(x,w)}{dA_{j,k-1}(x,w)} = 1$, we get $\frac{\partial O_{j,k}(x,w)}{\partial A_{j,k-1}(x,w)} = \frac{\partial O_{j,k}(x,w)}{\partial \tilde{A}_{j,k-1}(x,w)}$ where $\frac{\partial O_{j,k}(x,w)}{\partial \tilde{A}_{j,k-1}(x,w)} = \frac{\partial}{\partial A_{j,K-1}} g(\tilde{A}_{j,K-1}(x, w))$.

13.6.4.1 The output layer

We start with the output layer K and compute the gradient $\nabla E(w_{i,j}^{k-1})$ for the weight $w_{i,j}^{k-1}$ going from the ith node in layer $(K - 1)$ to the jth node in layer K. From the definition of the total error, the gradient satisfies

$$\nabla E(w_{i,j}^{k-1}) = \frac{\partial}{\partial w_{i,j}^{k-1}} E(x, w, d) = (O_{j,K} - d_j)\frac{\partial}{\partial w_{i,j}^{k-1}}(O_{j,K} - d_j)$$

since the subsripts i and j are fixed integers. Expending the output function $O_{j,K}$, we get

$$\frac{\partial}{\partial w_{i,j}^{k-1}}E(x,w,d) = (O_{j,K} - d_j)\frac{\partial}{\partial w_{i,j}^{k-1}}g\big(A_{j,K-1}(x,w) + b_{j,K}\big)$$

Using once again the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-1}}E(x,w,d) = (O_{j,K} - d_j)\frac{\partial}{\partial A_{j,K-1}}g(\tilde{A}_{j,K-1}(x,w))\frac{\partial}{\partial w_{i,j}^{k-1}}A_{j,K-1}(x,w)$$

where $\tilde{A}_{j,K-1}(x,w) = A_{j,K-1}(x,w) + b_{j,K}$. The only term that depends on $w_{i,j}^{k-1}$ in the activation function $A_{j,K-1}(x,w)$ is $O_{i,K-1}(x,w)w_{i,j}^{k-1}$, and the rest of the sum will zero out after derivation. Therefore, the gradient becomes

$$\begin{aligned}\frac{\partial}{\partial w_{i,j}^{k-1}}E(x,w,d) &= (O_{j,K} - d_j)\frac{\partial}{\partial A_{j,K-1}}g(\tilde{A}_{j,K-1}(x,w))\frac{\partial}{\partial w_{i,j}^{k-1}}O_{i,K-1}(x,w)w_{i,j}^{k-1} \\ &= (O_{j,K} - d_j)\frac{\partial}{\partial A_{j,K-1}}g(\tilde{A}_{j,K-1}(x,w))O_{i,K-1}(x,w)\end{aligned}$$

and we obtain the adjustment to each weight $\Delta w_{i,j}^{k-1}$. We let $e_{j,K} = \frac{\partial}{\partial O_{j,K}}E = (O_{j,K} - d_j)$ be the pre-error signal and rewrite the partial derivative as

$$\frac{\partial}{\partial w_{i,j}^{k-1}}E(x,w,d) = e_{j,K}\frac{\partial}{\partial A_{j,K-1}}g(\tilde{A}_{j,K-1}(x,w))O_{i,K-1}(x,w)$$

Further, setting the error signal as $\delta_{j,K} = \frac{\partial}{\partial net_{j,K}}E$ where $net_{j,K} = \tilde{A}_{j,K-1}(x,w)$, we get

$$\delta_{j,K} = e_{j,K}\frac{\partial}{\partial A_{j,K-1}}g(\tilde{A}_{j,K-1}(x,w))$$

and the gradient simplifies to

$$\nabla E(w_{i,j}^{k-1}) = \delta_{j,K}O_{i,K-1}(x,w)$$

so that the stochastic gradient descent rule for output units bebomes

$$\Delta w_{i,j}^{k-1} = -\eta\frac{\partial}{\partial w_{i,j}^{k-1}}E = -\eta\delta_{j,K}O_{i,K-1}(x,w)$$

We observe that the weight update rule $\Delta w_{i,j}^{k-1}$ is a multiplication of the error introduced to the output times the gradient of the output function of the current neurons input times this neurons input.

13.6.4.2 The first hidden layer

The next step is to consider the hidden layer $(K-1)$ and compute the gradient for the weight $w_{i,j}^{k-2}$ going from the ith node on layer $(K-2)$ to the jth node in layer $(K-1)$. Note, since the subscripts i and j are taken, for notation purpose, we use the subscript s to represent the nodes on the K-th layer. The gradient $\nabla E(w_{i,j}^{k-2})$ for the weight $w_{i,j}^{k-2}$ becomes

$$\nabla E(w_{i,j}^{k-2}) = \frac{\partial}{\partial w_{i,j}^{k-2}}E(x,w,d) = \sum_{s=1}^{N_{o,K}}(O_{s,K} - d_s)\frac{\partial}{\partial w_{i,j}^{k-2}}(O_{s,K} - d_s)$$

which gives

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial w_{i,j}^{k-2}} g\big(A_{s,K-1}(x,w) + b_{s,K}\big)$$

Using once again the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) \frac{\partial}{\partial w_{i,j}^{k-2}} A_{s,K-1}(x,w)$$

where $A_{s,K-1}(x,w) = \sum_{j=1}^{N_{o,K-1}} O_{j,K-1}(x,w) w_{j,s}^{k-1}$. The only term that depends on $w_{i,j}^{k-2}$ in the activation function $A_{s,K-1}(x,w)$ is $O_{j,K-1}(x,w) w_{j,s}^{k-1}$, and the rest of the sum will zero out after derivation. Therefore, the gradient simplifies to

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,k-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} \frac{\partial}{\partial w_{i,j}^{k-2}} O_{j,K-1}(x,w)$$

which becomes

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} \frac{\partial}{\partial w_{i,j}^{k-2}} g(A_{j,K-2}(x,w) + b_{j,K-1})$$

Using once again the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} \frac{\partial}{\partial A_{j,K-2}} g(\tilde{A}_{j,K-2}(x,w)) \frac{\partial}{\partial w_{i,j}^{k-2}} A_{j,K-2}(x,w)$$

where $A_{j,K-2}(x,w) = \sum_{i=1}^{N_{o,K-2}} O_{i,K-2}(x,w) w_{i,j}^{k-2}$. The only term that depends on $w_{i,j}^{k-2}$ in the activation function $A_{j,K-2}(x,w)$ is $O_{i,K-2}(x,w) w_{i,j}^{k-2}$, and the rest of the sum will zero out after derivation. Therefore, given the pre-error signal $e_{s,K} = (O_{s,K} - d_s)$, we get

$$\begin{aligned}
&\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) \\
= &\sum_{s=1}^{N_{o,K}} e_{s,K} \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} \frac{\partial}{\partial A_{j,K-2}} g(\tilde{A}_{j,K-2}(x,w)) \frac{\partial}{\partial w_{i,j}^{k-2}} O_{i,K-2}(x,w) w_{i,j}^{k-2} \\
= &\sum_{s=1}^{N_{o,K}} e_{s,K} \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} \frac{\partial}{\partial A_{j,K-2}} g(\tilde{A}_{j,K-2}(x,w)) O_{i,K-2}(x,w)
\end{aligned}$$

and we obtain the adjustment to each weight $\Delta w_{i,j}^{k-2}$. Thinking in terms of the pre-error signal, the error from the previous layer $e_{j,K-1} = \frac{\partial}{\partial O_{j,K-1}} E$ is given by

$$e_{j,K-1} = \sum_{s=1}^{N_{o,K}} e_{s,K} \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) w_{j,s}^{k-1} = \sum_{s=1}^{N_{o,K}} \delta_{s,K} w_{j,s}^{k-1}$$

***Remark* 13.6.3** *Some authors refer to the term Downstream of unit j to describe all units whose direct inputs include the output of unit j.*

Hence, we see that the error from the previous layer is scaled down in proportion to the amount of how the previous layer influenced it. Again, the weight update rule $\Delta w_{i,j}^{k-2}$ is a multiplication of the error introduced to the output times the gradient of the output function of the current neurons input times this neurons input. That is,

$$\frac{\partial}{\partial w_{i,j}^{k-2}} E(x,w,d) = e_{j,K-1} \frac{\partial}{\partial A_{j,K-2}} g(\tilde{A}_{j,K-2}(x,w)) O_{i,K-2}(x,w)$$

Setting the error signal $\delta_{j,K-1} = \frac{\partial}{\partial net_{j,K-1}} E$ where $net_{j,K-1} = \tilde{A}_{j,K-2}(x,w)$, as

$$\delta_{j,K-1} = e_{j,K-1} \frac{\partial}{\partial A_{j,K-2}} g(\tilde{A}_{j,K-2}(x,w))$$

the gradient simplifies to

$$\nabla(w_{i,j}^{k-2}) = \delta_{j,K-1} O_{i,K-2}(x,w)$$

so that the stochastic gradient descent rule for output units bebomes

$$\Delta w_{i,j}^{k-2} = -\eta \frac{\partial}{\partial w_{i,j}^{k-2}} E = -\eta \delta_{j,K-1} O_{i,K-2}(x,w)$$

By iteration, we repeat this procedure until we reach $\Delta w_{i,j}^{1}$. Note, we differentiate the error function with respect to the bias value $b_{j,K}$ in the same way.

13.6.4.3 The next hidden layer

Going one step backward, we consider the hidden layer $(K-2)$ and compute the gradient for the weight $w_{i,j}^{k-3}$ going from the ith node on layer $(K-3)$ to the jth node in layer $(K-2)$. We will also assume that it corresponds to the input layer. Note, since the subscripts i and j are taken, for notation purpose, we use the subscript s to represent the nodes on the K-th layer and t to represent the nodes on the $(K-1)$-th layer. We get

$$\frac{\partial}{\partial w_{i,j}^{k-3}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial w_{i,j}^{k-3}} (O_{s,K} - d_s)$$

which gives

$$\frac{\partial}{\partial w_{i,j}^{k-3}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial w_{i,j}^{k-3}} g\big(A_{s,K-1}(x,w) + b_{s,K}\big)$$

Using once again the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-3}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) \frac{\partial}{\partial w_{i,j}^{k-3}} A_{s,K-1}(x,w)$$

where $A_{s,K-1}(x,w) = \sum_{t=1}^{N_{o,K-1}} O_{t,K-1}(x,w) w_{t,s}^{k-1}$. Replacing in the equation we get

$$\frac{\partial}{\partial w_{i,j}^{k-3}} E(x,w,d) = \sum_{s=1}^{N_{o,K}} (O_{s,K} - d_s) \frac{\partial}{\partial A_{s,K-1}} g(\tilde{A}_{s,K-1}(x,w)) \sum_{t=1}^{N_{o,K-1}} \frac{\partial}{\partial w_{i,j}^{k-3}} O_{t,K-1}(x,w) w_{t,s}^{k-1}$$

which gives

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{s=1}^{N_{o,K}}(O_{s,K}-d_s)\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))\sum_{t=1}^{N_{o,K-1}}\frac{\partial}{\partial w_{i,j}^{k-3}}g\big(A_{t,K-2}(x,w)+b_{t,K-1}\big)w_{t,s}^{k-1}$$

Using the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) =$$
$$\sum_{s=1}^{N_{o,K}}(O_{s,K}-d_s)\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))\sum_{t=1}^{N_{o,K-1}}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}A_{t,K-2}(x,w)w_{t,s}^{k-1}$$

where $A_{t,K-2}(x,w)=\sum_{j=1}^{N_{o,K-2}}O_{j,K-2}(x,w)w_{j,t}^{k-2}$. The only term that depends on $w_{i,j}^{k-3}$ in the activation function $A_{t,K-2}(x,w)$ is $O_{j,K-2}(x,w)w_{j,t}^{k-2}$, and the rest of the sum will zero out after derivation. Therefore, given the pre-error signal $e_{s,K}=(O_{s,K}-d_s)$, we get

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{s=1}^{N_{o,K}}e_{s,K}\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))\sum_{t=1}^{N_{o,K-1}}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}O_{j,K-2}(x,w)w_{j,t}^{k-2}w_{t,s}^{k-1}$$

which becomes

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) =$$
$$\sum_{s=1}^{N_{o,K}}e_{s,K}\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))\sum_{t=1}^{N_{o,K-1}}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}g(A_{j,K-3}(x,w)+b_{j,K-2})w_{j,t}^{k-2}w_{t,s}^{k-1}$$

Setting

$$\hat{e}_{t,K-1}=e_{s,K}\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))w_{t,s}^{k-1}$$

we get

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{s=1}^{N_{o,K}}\sum_{t=1}^{N_{o,K-1}}\hat{e}_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}g(A_{j,K-3}(x,w)+b_{j,K-2})w_{j,t}^{k-2}$$

Note, from linearity we can interchange the summation operators and rewrite the above equation as

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{t=1}^{N_{o,K-1}}\sum_{s=1}^{N_{o,K}}\hat{e}_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}g(A_{j,K-3}(x,w)+b_{j,K-2})w_{j,t}^{k-2}$$

We then recover the pre-error term $e_{t,K-1}=\frac{\partial}{\partial O_{t,K-1}}E$ as

$$e_{t,K-1}=\sum_{s=1}^{N_{o,K}}\hat{e}_{s,K}=\sum_{s=1}^{N_{o,K}}e_{s,K}\frac{\partial}{\partial A_{s,K-1}}g(\tilde{A}_{s,K-1}(x,w))w_{t,s}^{k-1}$$

such that the equation simplifies to

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{t=1}^{N_{o,K-1}} e_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}g(A_{j,K-3}(x,w)+b_{j,K-2})w_{j,t}^{k-2}$$

Using once again the chain rule, we obtain

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = \sum_{t=1}^{N_{o,K-1}} e_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial A_{j,K-3}}g(\tilde{A}_{j,K-3}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}A_{j,K-3}(x,w)w_{j,t}^{k-2}$$

where $A_{j,K-3}(x,w) = \sum_{i=1}^{N_{o,K-3}} O_{i,K-3}(x,w)w_{i,j}^{k-3}$. The only term that depends on $w_{i,j}^{k-3}$ in the activation function $A_{j,K-3}(x,w)$ is $O_{i,K-3}(x,w)w_{i,j}^{k-3}$, and the rest of the sum will zero out after derivation. Therefore, the gradient simplifies to

$$\begin{aligned}
&\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) \\
=& \sum_{t=1}^{N_{o,K-1}} e_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial A_{j,K-3}}g(\tilde{A}_{j,K-3}(x,w))\frac{\partial}{\partial w_{i,j}^{k-3}}O_{i,K-3}(x,w)w_{i,j}^{k-3}w_{j,t}^{k-2} \\
=& \sum_{t=1}^{N_{o,K-1}} e_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))\frac{\partial}{\partial A_{j,K-3}}g(\tilde{A}_{j,K-3}(x,w))O_{i,K-3}(x,w)w_{j,t}^{k-2}
\end{aligned}$$

Writing the error from the previous layer in terms of the Downstream of unit j as

$$e_{j,K-2} = \sum_{t=1}^{N_{o,K-1}} e_{t,K-1}\frac{\partial}{\partial A_{t,K-2}}g(\tilde{A}_{t,K-2}(x,w))w_{j,t}^{k-2}$$

where $e_{j,K-2} = \frac{\partial}{\partial O_{j,K-2}}E$. Then, the gradient becomes

$$\frac{\partial}{\partial w_{i,j}^{k-3}}E(x,w,d) = e_{j,K-2}\frac{\partial}{\partial A_{j,K-3}}g(\tilde{A}_{j,K-3}(x,w))O_{i,K-3}(x,w)$$

which corresponds to the multiplication of the error introduced to the output times the gradient of the output function of the current neurons input times this neurons input. Further, setting the error signal $\delta_{j,K-2} = \frac{\partial}{\partial net_{j,K-2}}E$ where $net_{j,K-2} = \tilde{A}_{j,K-3}(x,w)$, as

$$\delta_{j,K-2} = e_{j,K-2}\frac{\partial}{\partial A_{j,K-3}}g(\tilde{A}_{j,K-3}(x,w))$$

the gradient simplifies to

$$\nabla(w_{i,j}^{k-3}) = \delta_{j,K-2}O_{i,K-3}(x,w)$$

and the stochastic gradient descent rule for the next hidden layer units bebomes

$$\Delta w_{i,j}^{k-3} = -\eta\frac{\partial}{\partial w_{i,j}^{k-3}}E = -\eta\delta_{j,K-2}O_{i,K-3}(x,w)$$

This is the general rule for updating internal unit weights in arbitrary multilayer networks. Hence, the error signal travels from the output layer to the input layer. Further, the weights influence the error by some degree, and they must be taken into consideration when propagating the error.

13.6.4.4 Some remarks

Since the error surface for multipayer networks may contain many different local minima, the backpropagation algorithm can only converge toward some local minima in E which is not necessarily the global minimum error. Nonetheless, when the gradient descent falls into a local minimum with respect to one of these weights, it will not necessarily be in a local minimum with respect to other weights. Hence, higher dimensions might provide escape routes to the steepest descent to continue searching the space of possible network weights. In addition, the sigmoid threshold function being approximately linear when the weights are close to zero, we can initialise the network weights to values near zero so that in the early steps the network will represent a very smooth function approximately linear in its inputs. There exists several heuristic to avoid being stuck in a local minima such as adding a momentum term to the weight-update rule or using a stochastic gradient descent. One of the best approach is to train multiple networks using the same data, but initialising each network with different random weights. One can then select the best network according to one of these two methods

1. select the network with the best performance over a separate validation data set.
2. all networks can be retained and treated as a committee of networks whose output is the weighted average of the individual network outputs.

Various authors investigated the backpropagation algorithm to find out which function classes could be described by which types of networks. Three general results are known

- Boolean functions: every boolean function can be represented exactly by some network with two layers of units, although the number of hidden units required grows exponentially in the worst case with the number of network inputs.
- Continuous functions: every bounded continuous function can be approximated with arbitrarily small error (under a finite norm) by a network with two layers of units (see Cybenko [1989]).
- Arbitrary functions: any function can br approximated to arbitrary accuracy by a network with three layers of units (see Cybenko [1988]). This is because any function can be approximated by a linear combination of many localised functions having value 0 everywhere except for some small region, and that two layers of sigmoid units are sufficient to produce good local approximations.

The hypothesis space for backpropagation is the n-dimensional Euclidean space of the n network weights. Further, as opposed to decision tree where the hypothesis space is discrete , it is continuous for backpropagation leading to a well-defined error gradient. Also, the inductive bias of backpropagation learning is characterised by smooth interpolation between data points. An important property of backpropagation is its ability to discover useful intermediate representations at the hidden unit layers inside the network. This ability of multilayer networks to automatically discover useful representations at the hidden layers is a key feature of ANN learning as it provides extra degree of flexibility to invent features not explicitly introduced by the human designer.

13.6.5 Summarising the feedforward ANN

As observed in Section (13.6.4), when running back propapation, the error signal travels from the output layer to the input layer. Given a hidden layer with subsrit k, and focusing on neuron with index h, we define the pre-error signal and error signal as

$$\begin{aligned} e_{h,k} &= \frac{\partial}{\partial O_{h,k}} E \\ \delta_{h,k} &= \frac{\partial}{\partial \tilde{A}_{h,k-1}(x,w)} E = \frac{\partial}{\partial net_{h,k}} E \end{aligned}$$

Using these results, we are going to summarise the feedforward neural network by considering the multilayer perceptron (MLP).

13.6.5.1 Forward pass

We consider an MLP consisting of K hidden layers with N_i input units, $N_{o,k}$ output units in the k-th hidden layer and $N_{o,K}$ output units in the output layer. In each unit of the hidden layer, or in the output layer, we first calculate the weighted sum of the incoming values, which is referred to as the *net value* of the input unit. The *activation function*, denoted by f, is then applied to the net value. The value of the activation function, y, is the output of the unit. The activation function is required to be bounded, differentiable and monotonous. The two most common functions used in machine learning are the *hyperbolic tangent* and the *logistic sigmoid* functions, which are both nonlinear. This important feature makes it possible for the network to model nonlinear equations. As the number of hidden layers increases, the network can approximate more complex nonlinear functions. Methods using a network with a large number of hidden layers are referred to as *deep learning* network. Since the input domain of the activation function is infinite, while the output domain is finite, the activation functions are also termed as *squashing functions*. Letting h be the index of the first hidden unit, the net value and output function satisfy

$$\begin{aligned} net_{h,1} &= \sum_{i=1}^{N_i} w_{ih} x_i \\ y_h &= f(net_{h,1}) \end{aligned} \tag{13.6.9}$$

Considering two adjacent layers, denoted by $k-1$ and k, with index h and h', respectively, the summation and the activation process are similar to that of Equation (13.6.9), given by

$$\begin{aligned} net_{h',k} &= \sum_{h=1}^{N_{o,k-1}} w_{hh'} y_h \\ y_{h'} &= f(net_{h',k}) \end{aligned} \tag{13.6.10}$$

The net value and activation function of the output layer are calculated in the same way as those in the hidden layer. We let s be the index ranging over the output layer, and $N_{o,K-1}$ is the number of neurons in the $(K-1)$-th hidden layer closest to the output layer. We get

$$\begin{aligned} net_{s,K} &= \sum_{h=1}^{N_{o,K-1}} w_{hs} y_h \\ y_s &= f(net_{s,K}) \end{aligned} \tag{13.6.11}$$

13.6.5.2 Backward pass

In general, after a forward pass the network output is not the target input, and we need to measure the distance between the actual output and the target output to serve as a measurement of the network performance. This distance is called the *error function*. Given the input-target pair (x, z), and the network output y, then the error function is given by

$$E(x,z) = \frac{1}{2} \sum_{s:output} (y_s - z_s)^2, \tag{13.6.12}$$

where $y = (y_1, \cdots, y_s, \cdots, y_K)$, $z = (z_1, \cdots, z_s, \cdots, z_K)$, and $N_{o,K}$ is the number of units in the output layer. As long as the error function is larger than a given tolerence level, we need to modify the weights of the network to further decrease the measure E. The algorithm designed for stepwise updating the weights that minimise the error function is called the *training algorithm*, or *learning algorithm* of the network. It is natural to relate the weights updates to

the *gradient*, which is a vector of derivatives of the error function with respect to all the weights. It costs the network certain number of steps to reach a tolerably small error. Putting all the weights in a vector, the weight vector at step n is denoted by w^n, which is then updated by an amount Δw^n. We denote the gradient as $\nabla E = \frac{\partial E}{\partial w^n}$ and apply the *gradient descent* repeatedly using the chain rule. By working backward, from computing the derivative with respect to the output layer to computing the derivatives with respect to all the internal weights, the error is back propagated in the network. This procedure is called the standard *back propagation*, discovered independently by different researchers (see Werbos [1974], Parker [1985], Rumelhart et al. [1986b]). In standard back propagation, the weight update of the n-th step is

$$\Delta w^n = -\eta \frac{\partial E}{\partial w^n}, \tag{13.6.13}$$

where η is the *learning rate*, a real number between 0 and 1. Following the previous notations, the detailed procedure to calculate the gradient $\frac{\partial E}{\partial w_{ij}}$ is as follows:

1. Calculate the derivatives of the error function with respect to the output units:

$$\frac{\partial E}{\partial y_s} = y_s - z_s \tag{13.6.14}$$

$$\frac{\partial E}{\partial net_{s,K}} = \frac{\partial E}{\partial y_s} \frac{\partial y_s}{\partial net_{s,K}} \tag{13.6.15}$$

2. Calculate the derivatives of the error function with respect to the hidden units. Work backward through hidden layers, using the chain rule. To do so, we introduce the *error signal*:

$$\delta_j := \frac{\partial E}{\partial net_j}, \tag{13.6.16}$$

where j is the index of an arbitrary unit in the network.

(a) Calculate the error signals of the last hidden layer:
Let j be the index of a unit in the last hidden layer which is the closest to the output layer. Since the error E depends on hidden unit j only through its connection to the output units, we have

$$\begin{aligned} \delta_j &= \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial net_{j,K-1}} \\ &= \frac{\partial y_j}{\partial net_{j,K-1}} \sum_{s=1}^{N_{o,K}} \frac{\partial E}{\partial net_{s,K}} \frac{\partial net_{s,K}}{\partial y_j} \end{aligned} \tag{13.6.17}$$

From Equation (13.6.10) and (13.6.11), and given the definition of the error signal in Equation (13.6.16), we have,

$$\delta_j = f^{'}(net_{j,K-1}) \sum_{s=1}^{N_{o,K}} \delta_s w_{js} \tag{13.6.18}$$

(b) Calculate the error signals for hidden layers before the last hidden layer:
The error signals for the hidden units in hidden layer k ($k < K$) are based on the calculation of the error signal of the $(k+1)$-th layer due to the chain rule. That is to say, except for the units of the last hidden layer, the error signal of each unit in other hidden unit can be computed recursively as

$$\delta_i = f^{'}(net_{i,k}) \sum_{j=1}^{N_{o,k+1}} \delta_j w_{ij} \tag{13.6.19}$$

where i and j are indicators of unit in the hidden layers k and $k+1$, respectively, before the last hidden layer.

Having computed all the error signals for all the hidden units as well as the one for the output units, the derivatives of error function with respect to the weights w_{ij} can be written as:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_{j,k+1}} \frac{\partial net_{j,k+1}}{\partial w_{ij}} = \delta_j y_i \tag{13.6.20}$$

Then, the change of weight from unit i to unit j is given by

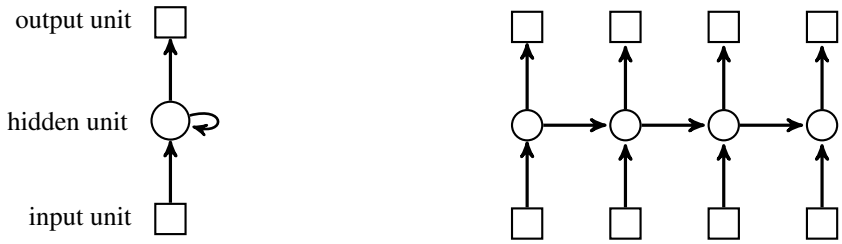
$$\Delta w_{ij} = -\alpha \delta_j y_i \tag{13.6.21}$$

Thus, introducing the concept of error signal makes the back propagation process more comprehensible, as only the error signal propagates backward through the network. The changing amount of the weight is proportional to the product of the error signal of its destination unit and the output of its source unit. To sumarise, the forward pass calculates the network output, while the backward pass updates the weights to minimise the error. The method used in backward pass is called back propagation, and a forward pass together with a backward pass is regarded as one *loop*. The training process continue until some *stopping criteria* is met, such as the error function is small enough, or, it stops to decrease after a certain number of loops.

13.7 Introduction to artificial recurrent neural networks

13.7.1 Presenting recurrent neural networks

While the feedforward neural networks discussed in Section (13.6.4) have no cycles for connections, a *Recurrent Neural Network* (RNN) has feedback connection, meanning that the nodes of the network have cyclical connections between them (see Figure 13.4a). The existence of cycles allows RNNs to develop a self-sustained temporal activation dynamics along its recurrent connection pathways, even in the absence of input, making them a dynamical system. There are two main classes of RNNs, the first one being characterised by an energy-minimising stochastic dynamics and symmetric connections (see Taylor et al. [2007]), and the second featuring a deterministic update dynamics and directed connections. We are going to concentrate on the latter. As shown in Figure 13.4b, RNNs can be visualised by unfolding the RNNs along the whole input sequence. That is, in the case of time series, the RNN are unfolded through time. The unfolded graph has no cycles, which is the same as FNNs, so that the forward pass and backward pass of MLP can be applied. Recall that the MLP is an approximator for nonlinear functions, and since RNN can be unfolded to a deep feedforward network, it has the advantage of the MLP, making it a better approximator. In fact, RNNs have proved to be an attractive form for modelling non-linearity due to their ability to approximate any dynamical system with arbitrary precision (see Siegelmann et al. [1991]). Further, Funahashi et al. [1993] showed that under mild and general assumptions, RNNs are universal approximatiors of dynamical systems. Indeed, as an equivalent theory of universal approximation for MLPs, it is said that a single hidden layer RNN with sufficient hidden units can approximate any sequence to arbitrary accuracy (see Hammer [2000]). For simplicity of exposition, we focus on a simple RNN with one self-connected hidden layer, as shown in Figure 13.4a. As trivial the difference in topology between FNNs and RNNs may seems, the advantage of the RNNs over the FNNs is profound. The latter can only map a limited number of inputs to a limited number of outputs, while a simple RNN can map the entire time series to some outputs. This is significant for time series prediction, as the long history can be fed to the network, and the existence of the recursive connection allows the network to memorise information of previous time steps, which is useful considering that financial time series are mostly serially correlated.



(a) A simple RNN (b) RNN unfolded

Figure 13.4: Recurrent Neural Network

13.7.1.1 Forward pass

We now consider a sequence of length T with N_i inputs, one hidden layer ($K = 2$) with $N_{o,K-1}$ hidden units and N_K output units. The i-th external input at time t is denoted by $x_i(t)$, and we let $net_{h,K-1}(t)$ and $y_h(t)$ be the *net value* (summation of the value received) and the output of unit h, respectively, in the hidden layer at time t. After unfolding the network, the forward pass of the RNN is similar to that of a MLP. The only difference being that the unit in the hidden layer receives values from both the current external inputs and the previous output of all hidden units. Mathematically, the net value of a hidden unit is given by

$$net_{h,K-1}(t) = \sum_{i=1}^{N_i} w_{ih} x_i(t) + \sum_{h'=1}^{N_{o,K-1}} w_{h'h} y_{h'}(t-1) \tag{13.7.22}$$

where $h^{'}$ is a unit on the previous hidden layer. Then the output value of the hidden unit is calculated by applying the sigmoid activation f to the net value

$$y_h(t) = f(net_{h,K-1}(t)) \tag{13.7.23}$$

such that the output of hidden units at each time step can be computed recursively, starting from the first time step $t = 1$. Initial value of $y_h(0), \forall h$ is commonly set to be zero, meaning that the network has not received information before the beginning of forward pass. Still, some researchers tend to have nonzero intial value by which they found that the network is more stable (see Zimmermann [2006]). The net value of the output unit, $net_{s,K}$, depends only on the output value of the hidden layer so that the output units are synchronised with the hidden units

$$net_{s,K}(t) = \sum_{h=1}^{N_{o,K-1}} w_{hs} y_h(t) \tag{13.7.24}$$

The error function can be defined as

$$E = \sum_{t=1}^{T} \sum_{s=1}^{N_K} \frac{1}{2} (y_s(t) - z_s(t))^2 \tag{13.7.25}$$

where $z_s(t)$ is the target value of the output unit s at time t. However, this is not the unique form of error function, as it should be based on the training algorithms of the RNN, which we will discuss in the next section.

13.7.1.2 Backward pass

We explained in Section (13.6.5) how to calculate the derivatives of the error function with respect to the weights in an MLP. In the case of RNNs, there are mainly two algorithms to calculate the weight derivatives: the Real Time Recurrent Learning (RTRL) (see Robinson et al. [1987], Williams et al. [1989]) and the Backpropagation Through

Time (BPTT) (see Werbos [1990], Williams et al [1992]). In ordrer to describe the learning algorithm of the RNN and detail its associated problems, we focus on the BPTT, as it is simple conceptually and more efficient in computation time (see Graves [2005]). Figure 13.5 illustrates the scheme for updating the weights. Note that the same weights are used in every time step. The error function is defined in Equation (13.7.25), which include all the time steps.

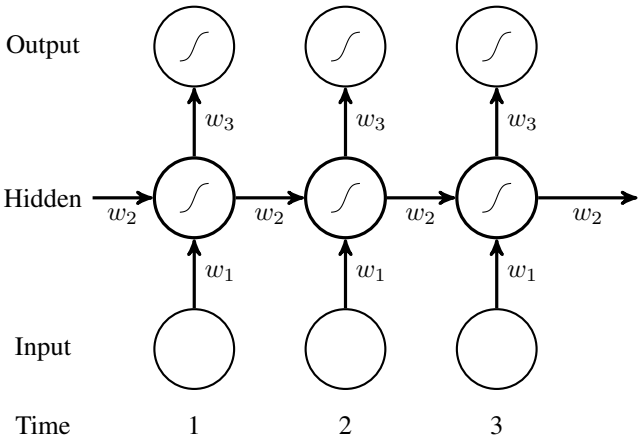


Figure 13.5: Weights in BPTT

For MLPs, the error signals are computed in Equations (13.6.15), (13.6.17) and (13.6.19) according to the type of unit. However, what is common for the error signals of the hidden units is that they depend on the sum of error signals that flow back to them. Since unfolded RNNs are equivalent to MLPs, such rule can also be applied. As illustrated in Figure 13.5, we consider the hidden unit h_2 at time $t = 2$, which has two outgoing connections. We can see that the error signal of hidden unit h_3 at time $t = 3$, and that of the output unit o_2 at time $t = 2$ would flow back to unit h_2 during back propagation through time. The red dashed arrows in Figure 13.6 illustrates the flow of error signals backward to a particular hidden unit. Hence, the error signal at unit h_2 can be computed as

$$\delta_{h_2} = \delta_{h_3} + \delta_{o_2}$$

Generally, the error function depends on the output of hidden layer not only through the current output layer, but also through the hidden layer next time. Thus, we have

$$\delta_h(t) = f^{'}(net_{h,K-1}(t))\Big(\sum_{s=1}^{N_K} \delta_s(t) w_{hs} + \sum_{h'=1}^{N_o} \delta_{h'}(t+1) w_{hh'}\Big) \tag{13.7.26}$$

where $\delta_j(T+1) = 0, \forall j$. Thus, starting the backward pass at the last time T, we apply Equation (13.7.26) recursively. Having computed the error signals, we can now calculate the derivative of the error function with respect to the weight via chain rule, getting

$$\frac{\partial E}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial E}{\partial net_j(t)} \frac{\partial net_j(t)}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j(t) y_i(t) \tag{13.7.27}$$

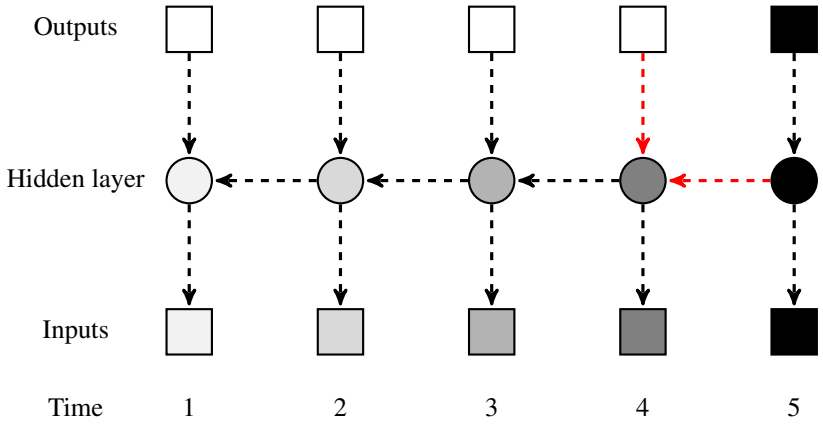


Figure 13.6: The flow of error signals and vanishing gradient problem in back propagation. The darkness of the shading of the nodes in the unfolded recurrent network indicates the degree of influence (sensitivity) of error signal of output unit at time $t = 5$. The darker the shade represents higher sensitivity. The sensitivity decays as the error signal is back propagated through time.

13.7.2 The long short-term memory

The impact of RNNs in nonlinear modelling has been limited because they are difficult to train by gradient-descent methods. The gradual change of network parameters during the learning process leads the network to bifurcations, where the gradient information degenerates and may become ill-defined (see Doya [1992]). Further, many update cycles may be necessary to obtain convergence of a few parameters, leading to long training times. In addition, when dealing with long-range memory, the gradient information may dissolves exponentially over time. One remedy is to use the Long Short-Term Memory networks (LSTM), which we are now going to describe.

13.7.2.1 The vanishing gradient problem

Even though recurrent neural networks (RNNs) are capable of processing serially correlated sequences, the length of sequence that a standard recurrent neural network can access is actually very limited. It results from the fact that the gradients would either decay or blow up when cycling around the recursive connections for too many times, which is usually referred to as *vanishing gradient problem* (see Bengio et al. [1994]). We can see in Equation (13.7.27) that the derivative of the error with respect to a weight depends on the error signal of the weight's destination unit, j, which itself depends on the derivative of the activation function. As discussed in Section (13.6.3.3), the derivative of the sigmoid function can explode or vanish depending on the slope of the function. As the number of time steps get larger and larger, so does the number of layers in the unfolded RNN. Assuming $c = 1$ in the sigmoid function, the error signal of a hidden unit at time $t = T$ would be propagated back through $(T - 1)$ layers, until a hidden layer at time $t = 1$. That is, it would be multiplied by $(T - 1)$ sigmoid activations' derivatives, all in the range $[0, \frac{1}{4}]$. For T not too large, the vanishing problem does not have much impacts on the network, but for larger T the gradient would decay exponentially. Similarly, for large enough c, the derivative of the sigmoid function near the origin ($x = 0$) would be larger than 1, leading to gradient explosion. This property of sigmoid function is illustrated in Figure 13.7. For these reasons, training RNNs with standard gradient descent algorithm is only feasible for small time steps. For longer time dependencies, the gradient vanishes as the error signal is propagated back through time, so that the network weights are never adjusted correctly when taking the events far back in the past into account (see Hochreiter et al. [2001]).

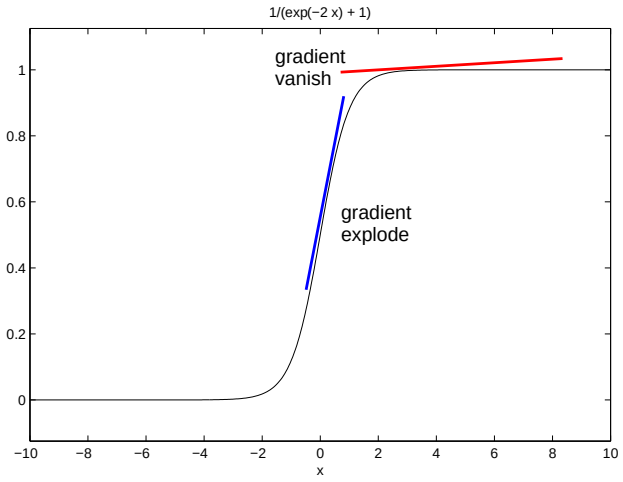


Figure 13.7: Plot of $\frac{1}{1+e^{-2x}}$ and its tangent lines. The black curve is the sigmoid function ranged $[0, 1]$ with $c = 2$. The red line is the flat tangent line.

13.7.2.2 The constant error carousel

One consequence of the vanishing, or exploding, gradient problem described above is that the traditional RNN will fail when involved with large time series. Since the vanishing gradient problem stems from the fact that the tangent line of the sigmoid function is either flat or steep, it is reasonable for some units to have constant gradient equating to 1. This kind of unit has been presented as the *Constant Error Carousel* (CEC) (see Hochreiter et al. [1997]). To describe the advantageous properties of CEC, we consider a single unit, j, with a single connection to itself. According to the rule of calculation of error signals, the jth error signal satisfies

$$\delta_j(t) = f_j^{'}(net_j(t))\delta_j(t+1)w_{jj}$$

where $f_j(\bullet)$ is the activation function of unit j. A constant error flow implies that $\delta_j(t) = \delta_j(t+1)$, leading to

$$f_j^{'}(net_j(t))w_{jj} = 1$$

This is an ordinary differential equation (ODE)

$$\frac{\partial f_j(net_j(t))}{\partial net_j(t)} = \frac{1}{w_{jj}}$$

which we can integrate, getting

$$f_j(net_j(t)) = \frac{net_j(t)}{w_{jj}}$$

for arbitrary $net_j(t)$. Thus, the activation function f_j for the jth unit should be **linear**

$$y_j(t+1) = f_j(net_j(t+1))$$

Since unit j has only one connection to itself, we get

$$net_j(t+1) = w_{jj}f_j(t)$$

and from the equation of $f_j(net_j(t))$, we get

$$f_j(w_{jj}y_j(t)) = \frac{w_{jj}y_j(t)}{w_{jj}} = y_j(t)$$

We can therefore set the activation function to be the identity function $f_j(x) = x, \forall x$, obtaining $w_{jj} = 1$. The unit j above can be extended to CEC by adding some extra features. A multiplicative *input gate unit* and a multiplicative *output gate unit* are introduced to control the input and output of the CEC, respectively. The input gate unit can prevent the memory of unit j from being overwritten by irrelevant inputs, while the output gate unit avoid perturbation of other units by controlling the output of unit j. The CEC ensures that the error signal arriving at the memory cell would not be scaled up or down during back propagation, and thus can avoid exponential gradient.

13.7.2.3 Network architecture

A unit which include the input gate unit, the output gate unit and the CEC is called a *memory cell* (see Figure 13.8). We are now going to describe the *Long Short-Term Memory* (see Hochreiter et al. [1997]) which is a recurrent neural network consisting of set of *memory blocks* where each block contains one or more memory cells. In fact, the LSTM network is the same as that of a standard recurrent neural network, except that the conventional hidden units are replaced by a memory block with cells and gates encapsulated in. Surely, the hidden layer of the LSTM network can be a mixture of conventional hidden units and memory blocks, but the former is not necessary. The v-th memory cell of the j-th memory block is denoted by c_j^v, the net value of the cell, the input gate and the output gate are $net_{c_j^v}$, net_{in_j} and net_{out_j}, respectively, the output value of the cell, the input gate and the output gate are $y^{c_j^v}$, y^{in_j} and y^{out_j}, respectively. Note that memory cells in the same memory block are controlled by the same input gate and output gate (see Figure 13.8).

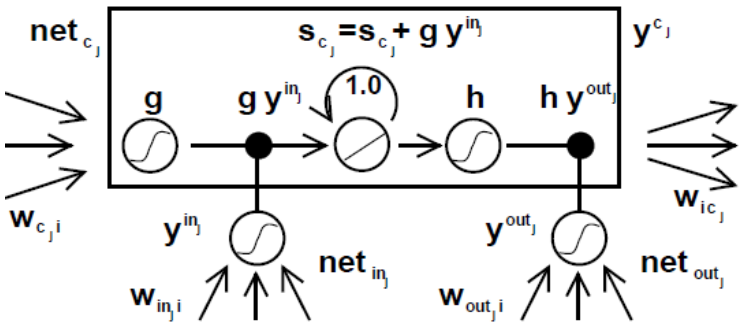


Figure 13.8: Architecture of a memory block with one memory cell. The CEC is a central linear unit with a self-connected weight 1.0. gate units use inputs from other units to decide whether to access or discard certain information.

As with a standard recurrent network, we have

$$y^{in_j}(t) = f_{in_j}(net_{in_j}(t))$$

$$y^{out_j} = f_{out_j}(net_{out_j}(t))$$

where f_{in_j} and f_{out_j} are the activation function of the input gate and output gate of memory block i, and

$$
\begin{aligned}
net_{in_j}(t) &= \sum_u w_{in_j u} y^u(t-1) + \sum_{n=1}^{N_i} w_{in_j n} x_n(t) \\
net_{out_j}(t) &= \sum_u w_{out_j u} y^u(t-1) + \sum_{n=1}^{N_i} w_{out_j n} x_n(t) \\
net_{c_j^v}(t) &= \sum_u w_{c_j^v u} y^u(t-1) + \sum_{n=1}^{N_i} w_{c_j^v n} x_n(t)
\end{aligned}
$$

where the superscript u stands for memory blocks and other traditional hidden unit. The output of the memory cell is different from that of a conventional hidden layer. An additional variable $s_{c_j^v}$, called *internal state*, should be considered

$$
\begin{aligned}
s_{c_j^v}(t) &= s_{c_j^v}(t-1) + y^{in_j}(t) g(net_{c_j^v}(t)),\ t > 0 \\
s_{c_j^v}(0) &= 0,
\end{aligned}
\tag{13.7.28}
$$

Actually, the first line of Equation (13.7.28) can be written as

$$
s_{c_j^v}(t) = 1.0 \times s_{c_j^v}(t-1) + y^{in_j}(t) g(net_{c_j^v}(t))
$$

where 1.0 corresponds to the self-recursive connection with weight 1.0 in Figure 13.8, which is essentially w_{jj} obtained by the introduction of a simple self-connected unit j without the vanishing gradient problem. So the output of the cell can be computed as

$$
y^{c_j^v}(t) = y^{out_j}(t) h(s_{c_j^v}(t))
$$

where g and h are the squashing function of the net value of memory cell and the internal sate $s_{c_j^v}$.

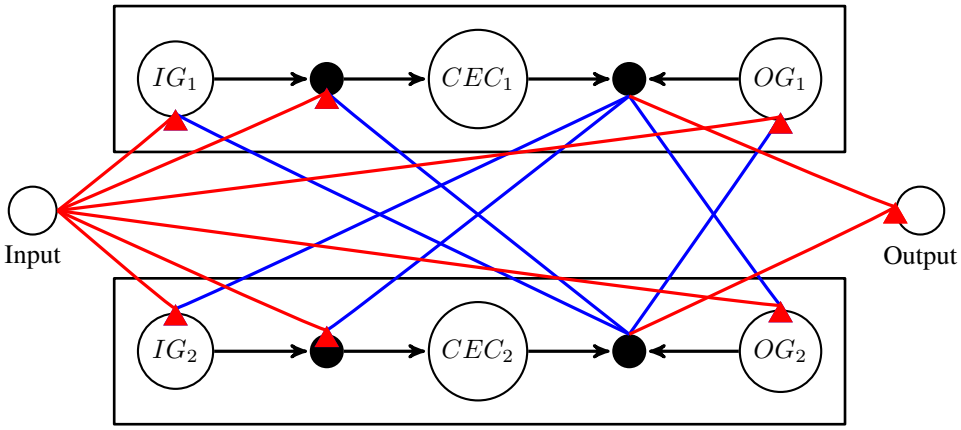


Figure 13.9: Connections in LSTM network. The example network consists of one input unit, a hidden layer of two single-cell LSTM memory blocks and one output units. Suppose the input is at time t. The the sources of the blue connections are from the previous one time step $t-1$. while the red connections convey information of current time step t. IG:input gate, OG: output gate, CEC:constant error carrousel

It is important that the gate units and the internal states are only visible within the cell, and that only the output of the cell can connect to other blocks (including gates and cell) in the hidden layer. Therefore, the net value is the

summation of current input values and the output of memory cells $y^{c_i^j}$ from previous time step. Further, Graves [2005] made some modifications to the original model of Schmidhuber and Hochreiter by modifying the rule connecting the units. Only the outputs memory cells are allowed to connect to other blocks, and the CECs and outputs of gate units are only visible within the memory blocks they are located in. As an example, Figure 13.9 shows the types of connections between two blocks. The gates units enable the memory cells in LSTM to store as well as to access information for a long period of time, and thus alleviate the vanishing gradient problem. For instance, if the activation function of the input gate is zero (the input gate is closed), then the memory of the cell will not be overwritten by the current input, thereby making it accessible to the network at a later time, as long as the output gate is open. However, the original LSTM has a weakness. When the sequence of time is long and have not been segmented to reset the network at certain time, Equation (13.7.28) implies that the internal state will grow infinitely and the network will collapse. To remedy this problem, Gers et al. [2000] proposed to add a *forget gate* to the self-connection with weight 1.0 in the original LSTM cell (see Figure 13.8). Then the revised equation for the internal state is extended from Equation (13.7.28) as follow

$$\begin{aligned} s_{c_j^v}(t) &= y^{\phi_j}(t) s_{c_j^v}(t-1) + y^{in_j}(t) g(net_{c_j^v}(t)),\ t > 0 \\ s_{c_j^v}(0) &= 0. \end{aligned} \tag{13.7.29}$$

where $y^{\phi_i}(t)$ is the value of the forget gate of the i-th memory block, which is squashed between zero and one. The forget gate is analogous to the reset operation of the memory cell. So far, the gate units are not entitled to control the internal state, as the gates have merely two sources of input: from the current input units and the previous output of all memory cells. In other words, the gates unit can only observe the cells' output. Once the output gate is closed, there is no way for gates to access the CEC they are supposed to control, which results in insufficient information that do harm to the performance of the network. In order to avoid the lack of information of internal states, another augmentation of the LSTM with forget gate was also introduced by Gers et al. [2002]. Weighted *peephole weights* are added in order to connect from the CEC to all gates of the same memory block. This effective remedy ensures that the all gates can "inspect" the internal state currently, even if the output gate is closed (see the dashed arrow in Figure 13.10).

We can summarise the types of connections of the modern LSTM as follow

- Outgoing connections
  - outside memory block
    Cells' output can feed to any types of units in any blocks in the hidden layer and the output units and they are the only output that can be connected to other blocks and output units.
  - inside memory block
    The gate unit can only pass value forward to cells of the block where the gate belong to. The internal state can connect to all types of gate within the same block.
- Incoming connections
  The memory cell's input can receive outputs from the hidden layer and the input layer;
  - The gate units can receive from the outputs from the hidden layer, input layer and the value of internal states.
  - The output unit can only receive value from the outputs of conventional hidden units and memory cells.

By including the forget gate and peephole weights, the traditional LSTM evolved to the modern LSTM and the update scheme in Hochreiter et al. [1997] needs refinement.

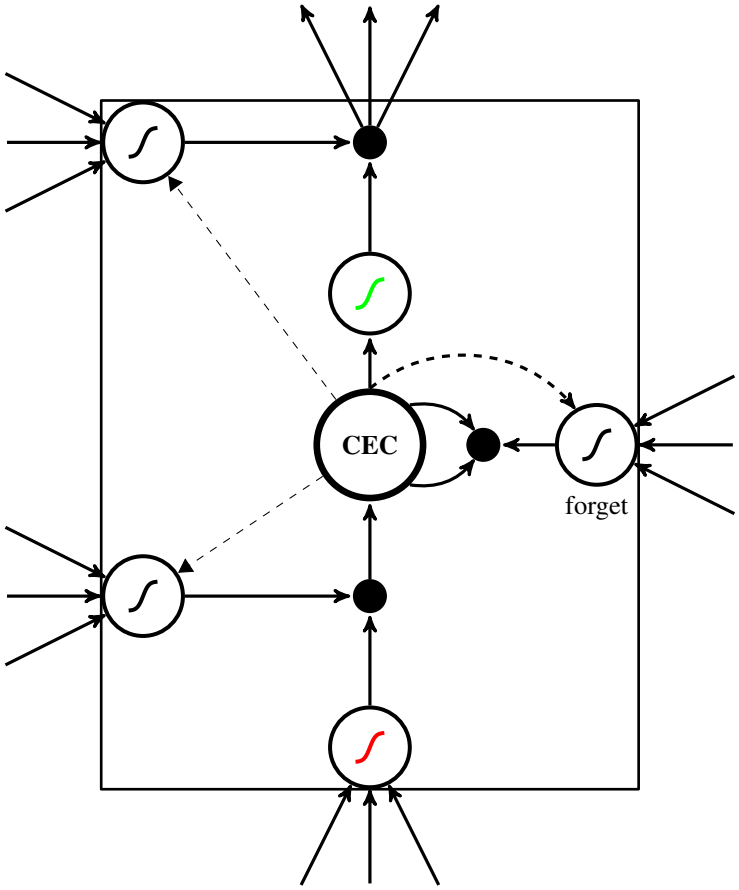


Figure 13.10: A modern LSTM memory block with one cell. The input, output and forget gates are the analogous of write read and reset operations for the cells. The three gates are nonlinear summation units that collect outputs from outside and inside the block. The small black dots are multiplicative units by which the activation of the cell is controlled. The gate units have sigmoid function f with range $[0, 1]$, so that 0 corresponds to the gate being closed and 1 to the gate being open. The unit with red S-shaped curve is the activation function that squashes the net value of cells' input, and the unit with green S-shaped curve is the activation function of cells' output. Usually, the functions g and h are task-specific and are encouraged to have ranges different from $[0, 1]$. The input and output gates multiply the input and output of the cell, and the forget gate multiply the previous internal state. Peephole weights from internal states to gates are shown by dashed arrow. Note, except for the self-recursive connection of the CEC, connections within the blocks have fixed weights of 1.0. The only outputs from the memory block to the rest of the network is from the upper multiplicative unit.

13.7.2.4 The learning algorithm

Details and proofs of the learning algorithm of the LSTM can be found in Hochreiter et al. [1997]. We display the learning algorithm in Figure 13.11. We emphasise here that the update scheme involves truncated derivative which are used to enhance the efficiency of the network. Note that the *training error* is the average of the $|z_t - y_t|$, $t = 1, \cdots, T$, in Figure 13.11. The weight update scheme is illustrated in Figure 13.12. The updated algorithm for the backward pass is a combination of truncated BPTT and customised RTRL. The former refers to BPTT using truncated derivatives.

```
for i = 1 to n do
    Initialize.  Initialize all weights {w_ij(1)} randomly.
    while stopping criteria not met do
        for t = 1 to T − 1 do
            Feedforward from x_t to output y_t using weights {w_ij(t)}.
            Back propagatation.  calculate error signals based on y_t − z_t.  Take down |y_t − z_t|.
            Update weights.  {w_ij(t)} → {w_ij(t + 1)}
        end for
        Reset network except for its weights.
    end while
end for
```

Figure 13.11: Pseudo-code of LSTM. The subscripts i and j standards for any units that have connections. y_t and z_t denote the actual output and target output of the network at time t. $z_t = x_{t+1}$. n is the number of simulations.

To be specific, standard BPTT is applied to output units, truncated BPTT is implemented by output gates, while weights related to input gates, forget gates and memory cells use a truncated RTRL. By comparing Figure 13.5 with Figure 13.12 we can get an idea of the differences between the full BPTT and the truncated one used by LSTM. In the BPTT algorithm for standard RNN, Equation (13.7.26) shows that the information of all time steps has to be saved for the update of weight and the weights remain unchanged during the entire feedforward process. Whereas the LSTM network updates its weights at each time step during the forward pass, so that there is no need to store all values. To summarise, the BPTT with full gradient change the weight for one time by passing forward from time $t = 1$ to $t = T$ and passing backward from $t = T$ to $t = 1$ while the truncated BPTT passes forward from $t = 1$ to $t = T$ and the weigts are updated for T times.

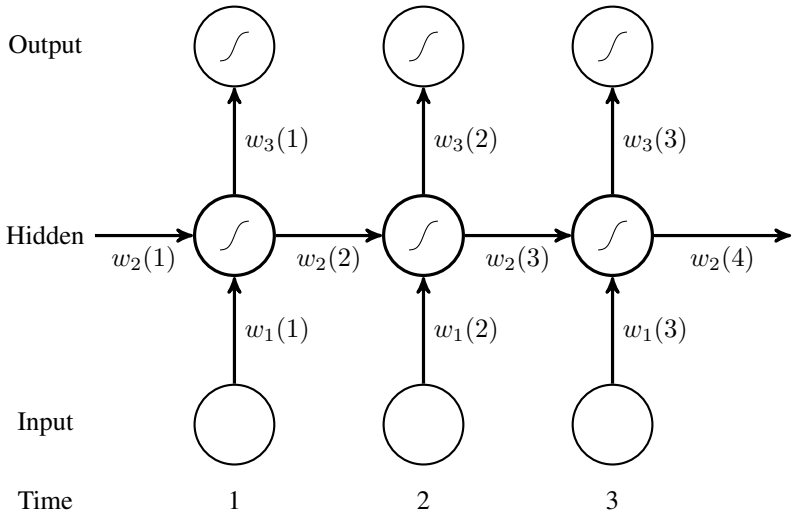


Figure 13.12: Algorithm of weight update of LSTM. Note that the memory cell is regarded as a normal hidden unit for simplicity. Only one hidden unit is shown.The number in bracket represents the time step.Weights are updated at each time step.

Computational Complexity of LSTM. The LSTM algorithm is efficient and its computational complexity is of

order $O(W)$ per time step, where W is the number of weights (see Hochreiter et al. [1997]). Compared with BPTT, LSTM is more economical in terms of space. In fact, calculating the full gradient of LSTM also has advantage since it is easy to debug and can be checked by numerical approximations (see Graves et al. [2005]). With the CEC, a LSTM block may be considered as a smart network unit, compared to the conventional hidden unit, to store information for arbitrary length of time. Therefore, the LSTM is well-suited to process and predict time series when there are unknown size of long time lags between important events. Since financial time series has long term memory, predicting volatility by LSTM network may be rather promising. While the traditional RNN with weight update algorithm BPTT (Back Propagation Through Time) (see Williams et al. [1990]) and RTRL (Real-time Recurrent Learning) (see Robinson et al. [1987] and Williams et al. [1992]) and the combinations of the former two (see Schmidhuber [1992]) have been proved to occur learning failure when processing sequences with only 10 time steps (see Bengio et al. [1994], Hochreiter et al. [1997], Gers et al. [2000], Hochreiter et al. HochreiterEtAl01), the LSTM can deal with 1000 time steps and even more, outperforming those traditional RNN algorithms. In fact, except for prediction, LSTM outperforms other RNNs in numerous aspects such as the best performance in speech recognition (see Graves et al. [2011]) and the ICDAR handwritting competition in 2009.

13.7.3 Reservoir computing

13.7.3.1 Describing the Reservoir methods

Reservoir Computing is a special RNN originating from Echo State Networks (ESNs) (see Jaeger [2001]) and Liquid State Machines (LSMs) (see Maass et al. [2002]), which assumes that supervised adaptation of all inter-connection weights are not necessary, and only training a memoryless supervised readout from it is sufficient to obtain good results. Thus, it avoids the shortcomings of the gradient-descent training of RNNs. RC is based on the computational seperation between a dynamic reservoir and a recurence-free readout. The former is an RNN as a nonlinear temporal expansion function, randomly created and unchanged during training. The latter produces the desired output from the expansion. This is a very simple approach avoiding the problem of bifurcations encountered during the training of RNNs, and the use of complex memory cells when learning long-term dependencies, as in the LSTM described in Section (13.7.2). Even though several studies aimed at understanding RC and the factors affecting its performances, none have been completely satisfactory due to the complexity of the reservoir, often resulting in contradictory conclusions. Introduction to the concepts and methodologies can be found in the literature (see Lukosevicius et al. [2009], [2012]). Goudarzi et al. [2014] compared the performance of three methods, the delay line, the NARX network, and the ESN and concluded that the first two have higher memorisation capability, but fall short of the generalisation power of the latter. Thus, we are going to briefly describe the ESN. We let the time-varying input signal be an N_i-th order column vector $U(t) = [u_i(t)]$, the reservoir state is an N_x-th order column vector $X(t) = [x_j(t)]$, and the generated output is an N_o-th order column vector $Y(t) = [y_o(t)]$. A Reservoir Computer is a collection of internal nodes, whose state vector $X(t)$ evolve in discrete time t according to the nonlinear map in Equation (13.6.6) of the form

$$X(t+1) = f\big(W_{res}^{\top}.X(t) + W_{in}^{\top}.U(t+1) + W_{fb}^{\top}.Y(t)\big) \tag{13.7.30}$$

for some activation function $f(\bullet)$, where the input weight matrix is an $N_i \times N_x$ matrix $W^{in} = [w_{ij}^{in}]$ where w_{ij}^{in} is the weigth of the connection from input node i to reservoir node j. The connection weights inside the reservoir are represented by an $N_x \times N_x$ matrix $W_{res} = [w_{jk}^{res}]$ where w_{jk}^{res} is weight from node j to node k in the reservoir. In presence of a bias, the output matrix is an $(N_x + 1) \times N_o$ matrix $W_{out} = [w_{ko}^{out}]$, where w_{ko}^{out} is the weight of the connection from the reservoir node k to the output node o. In the case where the output nodes are also connected to the input nodes and to themselves the size of the matrix becomes $(N_x + N_i + N_o) \times N_o$. The feedback matrix is an $N_x \times N_o$ matrix $W_{fb} = [w_{ko}^{fb}]$, where w_{ko}^{fb} is the weight of the connection from the reservoir node k to the output node o. If no output feedback is needed, then W_{fb} is null. The generated output is given by

$$Y(t) = f_{out}(W_{out}^{\top}.X(t))$$

where $f_{out}(\bullet)$ is an output activation function, typically the identity or a sigmoid. The output weights are trained to minimise the squared output error

$$E = ||Y(t) - \widehat{Y}(t)||^2$$

given the target output $\widehat{Y}(t)$. Once we know the optimum weights of the network, we can use the model to perform a forecast. To do so, we let T be a network state update, where

$$X(t+h) = T(X(t), \overline{U}^h)$$

denote the network state resulting from an iterated application of Equation (13.7.30) when the input sequence $\overline{U}^h = U(t+1), ..., U(t+h)$ is fed to the network being in state $X(t)$ at time t.

Echo State Networks (ENSs) assume that if a random RNN possesses certain algebriac properties, it suffises to train a linear readout from it to obtain good performances. Even though there are several readout methods, the most popular one is the linear regression where W_{out} solve a system of linear equations. From Remark (13.6.1), we get

$$W_{out}^\top X_m = \widehat{Y}_m$$

where both the matrix $X_m \in \mathbb{R}^{N_x \times T}$ and the matrix $\widehat{Y}_m \in \mathbb{R}^{N_o \times T}$ [3], over the training period $t = 1, ..., T$, have a column for every training time step t. The output weights can be estimated with direct pseudo-inverse calculations, or they can be estimated with the ordinary linear regression (Wiener-Hopf solution). Thus, we get

$$W_{out}^\top X_m . X_m^\top = \widehat{Y}_m . X_m^\top$$

and the weight matrix becomes

$$W_{out}^\top = \widehat{Y}_m . X_m^\top (X_m . X_m^\top)^{-1}$$

where $R = X_m . X_m^\top$ is the correlation matrix of the reservoir states, and $P = \widehat{Y}_m . X_m^\top$ is the cross-correlation matrix between the states and the desired outputs. Note, $P \in \mathbb{R}^{N_o \times N_x}$ and $R \in \mathbb{R}^{N_x \times N_x}$ do not depend on the training length T and can be calculated incrementally. When R is ill-conditioned the method is numerically unstable, but computing the Moore-Penrose pseudo-inverse R^+ instead of R can improve the solution. Alternatively, we can decompose the matrix R into two triangular matrices with Cholesky or the LU decomposition (see Press et al. [1992]) and solve

$$W_{out}^\top R = P$$

by two steps of substitution. Evolutionary search can also be used for training the linear readouts. There exists several bias to reduce the error in the validation set. One can smooth the model with regularisation functions with the Ridge regression

$$W_{out}^\top = P(R + \alpha^2 I)^{-1}$$

where α controls the smoothing effect, and $I \in \mathbb{R}^{N_x \times N_x}$ is the identity matrix. We can also add noise $\epsilon(t)$ (sampled from uniform or Gaussian distribution) to the reservoir states

$$X(t+1) = f\big(W_{res}^\top . X(t) + W_{in}^\top . U(t+1) + W_{fb}^\top . Y(t)\big) + \epsilon(t)$$

to stabilise solutions in the model (see Jaeger [2007]). We can further adjust global control parameters to make the echo state network dynamically similar to the system we model. For instance, we can use fully connected reservoirs or sparsely connected ones.

[3] These matrices are transposed compared to the conventional notation.

In order to produce a reservoir with a rich enough set of dynamics the number of internal connections N_x should be large, the weight matrix W should be sparse, and the weights of the connections should be generated randomly from a uniform distribution symmetric around the zero value. Further, the network should have the echo state property (ESP), which relates asymptotic properties of the excited reservoir dynamics to the driving signal (see Jaeger [2001]). It states that the effect of a previous state $X(t)$ and a previous input $U(t)$ on a future state $X(t+h)$ should vanish gradually as $h \to \infty$, and not persist or be amplified. In reservoirs using the $tanh$ squashing function, and for zero input, the reservoir weight matrix W^{res} must be scaled so that its spectral radius [4] $\rho(W^{res})$ satisfies $\rho(W^{res}) < 1$. For any kind of inputs (including zero) and state vectors, we require $\sigma_{max}(W^{res}) < 1$ where $\sigma_{max}(W^{res})$ is the largest singular value of W^{res}. If the input comes from a stationary source, the ESP holds with probability 1 or 0. Due to the auto-feedback nature of RNNs, the reservoir states $X(t)$ reflect traces of the past input history, which can be seen as a dynamical short-term memory. Assuming a single input ESN, the short-term capacity is given by

$$C = \sum_i r^2(U(t-i), Y_i(t))$$

where $r^2(\bullet, \bullet)$ is the squared correlation coefficient between the input signal delayed by i and an output signal $Y_i(t)$ trained to memorise $U(t-i)$ on the input signal $U(t)$. For an i.i.d. input, the memory capacity C of an echo state network of size N is bounded by N. Thus, we can not train ESN on tasks requiering unbounded-time memory.

13.7.3.2 Some improvements

Separating the reservoir from the readout training allows for two research directions to be pursued independently,

1. the generation of the reservoir, and
2. the output training.

There is no reasons why the reservoir should be randomly generated and alternative methods could be used to obtain optimal reservoir design. However, no single type of reservoir can be optimal for all types of problems (no free lunch principle). Several methods have been proposed for generating the reservoir (see Lukosevicius et al. [2009]), which can be classified as

1. generic methods for generating $RNNs$ with different neuron models, connectivity patterns and dynamics.
2. unsupervised adaptation of the reservoir based on the input data $U(t)$ but not the target value $\widehat{Y}(t)$.
3. supervised learning, adaptation of the reservoir using task-specific information from both $U(t)$ and $\widehat{Y}(t)$.

In order to deal with different time scales simultaneously, one can divide the reservoir into decoupled sub-reservoirs and introduce inhibitory connections among all the sub-reservoirs. However, the inhibitory connections should be heuristically computed from the rest of W and W_{fb} such that they predict the activations of the sub-reservoirs one time step ahead (see Xu et al. [2007]). Alternatively, the Evolvino transfers the idea of ESNs to a LSTM type of RNNs where the LSTM RNN used for its reservoir consists of specific small memory-holding modules. In that model, the weights of the reservoir are trained using evolutionary methods (see Schmidhuber et al. [2007]). Further, ESNs like any other RNNs has only a single layer of neurons (see Figure 13.4a), making it unsuitable for some types of problems requiering multilayers. A solution is to use Layered ESNs (see Lukosevicius [2007]), where part of the reservoir connections are instantaneous, and the rest takes one time step for the signals to propagate as in normal ESNs. One can also add leaky integrator neurons to ESNs, getting leaky integrator ESNs (Li-ESNs) performing at least as well as the simple ESN (see Lukosevicius et al. [2006]). Since the parameters a and Δt control the speed of the reservoir dynamics, small values result in reservoirs reacting slowly to the input. Note, depending on the speed at which the input $U(t)$ changes, we can vary Δt on-the-fly, getting a warping invariant ESNs (TWIESNs).

[4] the largest absolute eigenvalue

Since checking the performance of a resulting ESN is relatively inexpensive, evolutionary methods for pre-training the reservoir developed (see Ishii et al. [2004]). Generally, one separate the topology and weight sizes of W_{res} to reduce the search space (see Bush et al. [2005]). Jiang et al. [2008] showed that by only adapting the slopes of the reservoir unit activation functions $f(\bullet)$ with an evolutionary algorithm, and having W_{out} random and fixed, a very good prediction performance of an ESN could be achieved. Note, evolutionary algorithms can also be used to train the readouts. One can increase the expressiveness of the ESN by having k linear readouts trained and an online switching mechanism among them, or by averaging outputs over several instances of ESNs (see Bush et al. [2006]).

Chapter 14

Introducing Differential Evolution

14.1 Introduction

While almost any problem found in everyday life can be thought as an optimisation problem, we saw that in classical economics the agent decision making process was represented as the maximisation of some expected utility functions. Assuming randomness, portfolio selection and fair price were introduced, such as CAPM and BS formula, and quantitative optimisation techniques could be used. Further, econometric models were devised to forecast price processes in view of either computing the CAPM/BS formula, or, taking advantage of market inefficiencies. In the former, simplicity led to a single optimal solution, while in the latter, complexity led to a range of fair values. With the growing quantity of data available, machine learning methods that have been successfully applied in science are now applied to mining the markets. Data mining and more recent machine-learning methodologies provide a range of general techniques for the classification, prediction, and optimisation of structured and unstructured data. All of these methods require the use of quantitative optimisation techniques known as stochastic optimisation algorithms, such as combinatorial optimisation, simulated annealing (SA), genetic algorithms (GA), or reinforced learning. While stochastic methods have some degree of randomness when operating, heuristic methods incorporate additional strategies or knowledge to their operation. Some of these algorithms use heuristics inspired by real life processes. For instance, genetic algorithms are inspired by the process of evolution, and simulated annealing is inspired by the process of annealing metals. We are going to consider an Evolutionary Algorithm (EA) which we will illustrate with the problem of model calibration to a finite set of option prices.

14.2 Calibration to implied volatility

14.2.1 Introducing calibration

14.2.1.1 The general idea

For every parametric model that one can define, such as regression models or option pricing models, we need to estimate the model parameters from the market prices or the implied volatility surface. It leads to an ill-posed inverse problem because the inversion is not stable and amplifies market data errors in the solution. To be more precise, letting x be the vector of model parameters and y be the vector of market prices, we want to solve

$$Tx = y$$

where $T : X \to Y$ is a (non-linear) operator between reflexive Banach spaces X, Y, with inverse T^{-1} which is not continuous. We further assume that only noisy data y^δ with

$$||y^\delta - y|| \leq \delta$$

is available. The operator T^{-1} can be found by solving a measure among a set of measures, such as

$$||Tx - y^\delta||^2$$

The ill-posness of the problem means that a small error on market data y can lead to a big error on the solution (model parameters) x. Regularisation techniques allow one to capture the maximum information on x from the set y in a stable fashion. See Tankov [2005] for details on regularisation techniques.

14.2.1.2 Measures of pricing errors

We choose to estimate implicitly the vector Ψ of model parameters by minimising, given a measure, the distance between the liquid market price $P_t(T_i, K_i)$ and the model price $C_t(T_i, K_i)$ for $i \in I$ where I is the total number of market prices considered. Note, the market price for the ith stock S_i is denoted by $P_i(t)$ while the option price for maturity T_i and strike K_i is denoted by $P_t(T_i, K_i)$. We consider some benchmark instruments with payoff $(H_i)_{i\in I}$ with obsereved market prices $(P_i^*)_{i\in I}$ in the range $P_i^* \in [P_i^b, P_i^a]$ representing the bid/ask prices. In the option world, we also need to consider a set of arbitrage free model $\mathcal{Q}$ such that the discounted asset price $(\overline{S}_t)_{t\in[0,T]}$ is a martingale under each $\mathbb{Q} \in \mathcal{Q}$ with respect to its own history $\mathcal{F}_t$ and

$$\forall \mathbb{Q} \in \mathcal{Q}\,, \forall i \in I\,, E^{\mathbb{Q}}[|H_i|] < \infty\,, E^{\mathbb{Q}}[H_i] = P_i^*$$

Since the market price P_i^* is only defined up to the bid-ask spread we get

$$\forall \mathbb{Q} \in \mathcal{Q}\,, \forall i \in I\,, E^{\mathbb{Q}}[|H_i|] < \infty\,, E^{\mathbb{Q}}[H_i] \in [P_i^b, P_i^a]$$

Further, different choices of norms for the vector $(P_i^* - E^{\mathbb{Q}}[H_i])_{i\in I}$ lead to different measures for the calibration error. For example, we have

$$\begin{aligned}
||P^* - E^{\mathbb{Q}}[H]||_\infty &= \sup_{i\in I} |P_i^* - E^{\mathbb{Q}}[H_i]| \\
||P^* - E^{\mathbb{Q}}[H]||_1 &= \sum_{i\in I} |P_i^* - E^{\mathbb{Q}}[H_i]| \\
||P^* - E^{\mathbb{Q}}[H]||_p &= \Big(\sum_{i\in I} |P_i^* - E^{\mathbb{Q}}[H_i]|^p\Big)^{\frac{1}{p}}
\end{aligned}$$

Obviously we need to choose a measure among this set of measures, and traditionally practitioners consider the Price Norm

$$f_1(i) = \big|P_t(T_i, K_i) - C_t(T_i, K_i; \Psi)\big|^2$$

or the Relative Price Norm

$$f_2(i) = \Big|\frac{P_t(T_i, K_i) - C_t(T_i, K_i; \Psi)}{P_t(T_i, K_i)}\Big|^2$$

14.2.2 The calibration problem

Since the Black-Scholes model [1973] and the celebrated BS-formula, market prices of index options and foreign exchange options have reached a high degree of liquidity such that they became the benchmark to mark to market or calibrate option pricing models for pricing and hedging exotic options. More formally, using market prices, we need to estimate the vector Ψ of model parameters in order to price exotic options. This amount to solving the following inverse problem.

Problem 1 *Given prices $C_t(T_i, K_i)$ for $i \in I$ where I is the total number of market prices considered, find the vector Ψ of model parameters such that the discounted asset price $\hat{S}_t = e^{-rt}S_t$ is a martingale and the observed option prices are given by their risk-neutral expectations*

$$\forall i \in I \ , \ C_t(T_i, K_i) = e^{-r(T-t)}E^{\Psi}[(S(T_i) - K_i)^+ | S_t = S]$$

That is, we need to retrieve the risk-neutral process and not just the conditional densities which is equivalent to a moment problem for the process S. However, in practice we do not know the call and put prices for all strike prices but only for a finite number of them so that extrapolation and interpolation is needed, resulting in solutions at best approximately verifying the constraints. It is typically an ill posed problem as there may be either no solution at all or an infinite number of solutions (see Cont et al. [2002] for more details). Therefore, one need to use additional criterions for choosing a solution. It means that we need to reformulate the calibration as an approximation problem, for instance minimising the in-sample quadratic pricing error

$$\begin{aligned} \Psi^* &= \arg\inf_{\Psi} \mathcal{J}(\Psi) \\ \mathcal{J}(\Psi) &= \sum_{i=1}^{n} w_i \left| P_t(T_i, K_i) - C_t(T_i, K_i; \Psi) \right|^2 \end{aligned} \tag{14.2.1}$$

where w_i is a weight associated to each market option price. Market practice is to solve the optimisation problem with a gradient-based minimisation method to locate the minima. In that case we can always find a solution but the minimisation function is not convex and the gradient descent may not succeed in locating the minimum. However, the number of parameters to calibrate a model is less important, from a numerical point of view, than the convexity of the objective function to minimise in a gradient based method. Non-Linear Least Square solution does not resolve the uniqueness and stability issues, and the inverse problem remains ill-posed. In order to circumvent these difficulties various authors proposed different regularisation methods all consisting in adding to the objective function a penalisation criterion. As a result, it makes the problem well-posed and allow for gradient-based optimisation algorithm.

For example, we choose to estimate the vector Ψ of model parameters implicitly by minimising, given a measure, the distance between the liquid market price $P_t(T_i, K_i)$ and the model price $C_t(T_i, K_i)$ for $i \in I$ where I is the total number of market prices considered. Obviously we need to define a measure, and for simplicity, we will only expose the optimisation problem. One way of achieving our objective is to minimise a Tikhonov-type functional (see Tikhonov et al. [1998])

$$\begin{aligned} \Psi^* &= \arg\inf_{\Psi} \mathcal{J}(\Psi) \\ \mathcal{J}(\Psi) &= \sum_{i=1}^{n} w_i \left| P_t(T_i, K_i) - C_t(T_i, K_i; \Psi) \right|^2 + \alpha H(\mathbb{Q}, \mathbb{Q}_0) \end{aligned} \tag{14.2.2}$$

where H is a measure of closeness of the model $\mathbb{Q}$ to a prior $\mathbb{Q}_0$. Many choices are possible for the penalisation function H. To get uniqueness and stability of the solution, the function should be convex with respect to the model parameters. So, the target function is made of two components

1. penalisation function convex in model parameters Ψ
2. quadratic pricing error measuring the precision of calibration

The coefficient α is the regularisation parameter and defines the relative importance of the two terms such that when $\alpha \to 0$ we recover the standard Least Square Error which is no-longer a convex function. If α is large enough, the target function inherits the convexity properties of the penalising function and the problem becomes well-posed. The correct choice of α is important and can not be fixed in advance since its optimal value depends on the data and on the level of error δ one wants to achieve.

14.2.3 The regularisation function

The regularisation function $H(\Psi, \Psi_0)$ is a positive function satisfying $H(\Psi, \Psi_0) \geq 0$ and $H(\Psi_0, \Psi_0) = 0$ and it must be a convex function. So, for sufficiently large α the optimisation function $\mathcal{J}(\Psi)$ can be made globally convex, improving the performance of most numerical optimisation schemes using a gradient based method. However, the penalty may influence the loss function too much, leading to biased parameters. There are many choices to get such a function, but to obtain stability in model prices from one day to the next we can choose to relate the vector of model parameters to its previously estimated values at time $t-1$. Setting $\Psi_0 = \Psi_{t-1}$, the regularisation function becomes

$$H(\Psi, \Psi_{t-1}) = |\Psi - \Psi_{t-1}|_{L_2}$$

which is the L_2 norm of the difference of the two vector of model parameters. This penalty should improve hedging as the paths of the estimated parameters are now smoothed over time.

An alternative approach to solving for the vector of model parameters Ψ_0 is to solve the standard least squares method in Equation (14.2.1) getting Ψ_0^*. The objective function not being convex, a simple gradient procedure will not give the global minimum. However, the solution $(\Psi_0, \mathbb{Q}_0)$ will be iteratively improved and should be viewed as a way to regularise the optimisation problem in Equation (14.2.2). We can define a measure of model error $\epsilon(\Psi_0, \mathbb{Q}_0) = \epsilon_0$ which represent the distance of market prices to model prices and gives an a priori level of quadratic pricing error. This error is typically positive, that is $\epsilon_0 > 0$. Now, given the regularisation parameter $\alpha > 0$ we get the solution $(\Psi_\alpha, \mathbb{Q}_\alpha)$ with the a posteriori quadratic pricing error given by $\epsilon(\Psi_\alpha, \mathbb{Q}_\alpha)$. Note we expect $\epsilon_\alpha > \epsilon_0$ since by adding the entropy term we have given up some precision to gain in stability. The Morozov discrepancy principle is an example of an a posteriori parameter choice rule (see Morozov [1984]). It consists in minimising this loss of precision through regularisation by choosing α^* such that

$$\epsilon(\Psi_\alpha, \mathbb{Q}_\alpha) \approx \epsilon_0$$

Practically, the a priori error is

$$\epsilon_0^2 = \inf_{\Psi_0} \sum_{i=1}^{n} w_i \left| P_t(T_i, K_i) - C_t(T_i, K_i; \Psi_0) \right|^2$$

So, for $\delta > 1$, for example $\delta = 1.1$ we solve

$$\epsilon(\Psi_\alpha, \mathbb{Q}_\alpha) = \delta\epsilon_0$$

Since the optimisation problem in Equation (14.2.2) is a differentiable function of α, we can get the solution α^* with a small number of iterations using the Newton-Raphson method.

14.2.4 Beyond deterministic optimisation method

The difference between normal (deterministic) and stochastic optimisation methods (SOM) is that normal methods use some knowledge of the problem such as the derivative of a function (gradient), the continuity of the function etc., and the SOM do not assume any external knowledge. While the additional knowledge makes deterministic optimisation methods more powerful, they are also less robust. In general, we do not know with certainty if the objective function has a unique global minimum, and if it exists, if it can be reached with a gradient based method. Many model parameters can reproduce the call prices with equal precision due to many local minima or flat region (low sensitivity to variations in model parameters). This means that the model is very sensitive to the input prices and the starting point in the algorithm. In addition, in an incomplete market, a deterministic optimisation method will at best locate one of the local minima of the fitting criterion but will not guarantee the global minima and will not acknowledge the multiplicity of solutions of the initial calibration problem. To overcome these issues, we are going to detail an alternative approach to solving non-linear programming problems under constraints that do not require computing the gradient of the model. To do so, we will need to consider an evolutionary algorithm that handle constraints in a simple and efficient way.

14.3 Nonlinear programming problems with constraints

14.3.1 Introducing evolutionary algorithms

14.3.1.1 A brief history

Evolutionary algorithms (EAs) introduced by Holland [1962] [1975] and Fogel [1966] are robust and efficient optimisation algorithms based on the theory of evolution proposed by Darwin [1882], where a biological population evolves over generations to adapt to an environment by mutation, recombination and selection. They are stochastic search algorithms, searching from multiple points in space instead of moving from a single point like gradient-based methods do. These algorithm are typically initiated with a population of potential solutions, that may be drawn randomly or specified prior to the beginning of the search. Iteration on the population is based on the principles of natural selection, with each iteration, or generation, improving in fitness as defined by some pre-determined measure. The methods by which each generation is determined are specific to the particular algorithm in question, however, all evolutionary algorithms rely on classes of stochastic operator known as selection and reproduction operators. The selection operator acts to ensure that individuals with greater fitness in each generation are selected as parents for the next generation, whilst the reproduction operator determines how the next generation is derived from the parents selected. Moreover, they work on function evaluation alone (fitness) and do not require derivatives or gradients of the objective functions. McKay [2008] claimed that despite their simplicity, given modest resources and a relatively tough optimisation problem, Evolutionary Algorithms reliably converge to good solutions, and what's more, they are suited to parallel implementation due to their speed scaling almost linearly with the number of processors.

There is a large literature describing different evolutionary algorithms (EAs) commonly used to solve constrained nonlinear programming problems (CNOPs) such as evolutionary programming, evolution strategies, genetic algorithms (GAs), differential evolution (DE) and many more. For instance, GAs are general purpose search algorithms based on an evolutionary paradigm where the population members are represented by strings, corresponding to chromosomes. Search starts with a population of randomly selected strings, and, from these, the next generation is created by using genetic operators (mutation). At each iteration individual strings are evaluated with respect to a performance criteria and assigned a fitness value. Strings are randomly selected using these fitness values to either survive or to mate to produce children for the next generation. However, among the different EA's commonly used DE became very popular. DE is a population-based approach to function optimisation generating a new position for an individual by calculating vector differences between other randomly selected members of the population. The DE algorithm is found to be a powerful evolutionary algorithm for global optimisation in many real problems. As a result, since the original

article of Storn and Price [1995] many authors improved the DE model to increase the exploration and exploitation capabilities of the DE algorithm when solving optimisation problems.

Since EAs are search engines working in unconstrained search spaces they lacked until recently of a mechanism to deal with the constraints of the problems. The first attempts to handle the constraints were either to incorporate methods from mathematical programming algorithms within EAs such as penalty functions, or, to exploit the mathematical structure of the constraints. Then, a considerable amount of research proposed alternative methods to improve the search of the feasible global optimum solution. Most of the research on DE focused on solving CNOPs by using a sole DE variant, a combination of variants or combining DE with another search method. One of the most popular constraint handling mechanisms is the use of the three feasibility rules proposed by Deb [2000] on genetic algorithms. Using some of the improvements to the DE algorithm combined with simple and robust constraint handling mechanisms we propose a modified algorithm for solving our optimisation problem under constraints which greatly improves its performances.

14.3.1.2 Defining the problems

We consider a system with the real-valued properties

$$g_m \text{ for } m = 0, .., P-1$$

making the objectives of the system to be optimised. Given a N-dimensional vector of real-valued parameter $X \subset \mathbb{R}^N$ the optimisation problem can always be written as

$$\min f_m(X)$$

where $f_m(.)$ is a function by which g_m is calculated and where each element $X(i)$ of the vector is bounded by lower and upper limits $L_i \leq X(i) \leq U_i$ which define the search space $\mathcal{S}$. We follow Lueder in [1990] who showed that all functions $f_m(.)$ can be combined in a single objective function $H : X \subset \mathbb{R}^N \to \mathbb{R}$ expressed as the weighted sum

$$H(X) = \sum_{m=1}^{P} w_m f_m(X)$$

where the weighting factors w_m define the importance of each objective of the system. Hence, the optimisation problem becomes

$$\min H(X)$$

so that all the local and global minima (when the region of eligibility in X is convex) can be found. However, since the problems of calibration in finance involves a single objective function, the optimisation function simplifies. Most complex search problems such as optimisation problems are constrained numerical problem (CNOP) more commonly called general nonlinear programming problems with constraints given by

$$\begin{aligned} g_i(X) &\leq 0 \, , i = 1, .., p \\ h_j(X) &= 0 \, , j = 1, .., q \end{aligned}$$

Equality constraints are usually transformed into inequality constraints by

$$|h_j(X)| - \epsilon \leq 0$$

where ϵ is the tolerance allowed. Given the search space $\mathcal{S} \subset \mathbb{R}^N$, we let $\mathcal{F}$ be the set of all solutions satisfying the constraints of the problems called the feasible region. It is defined by the intersection of $\mathcal{S}$ and the set of $p + q$

additional constraints. At any point $X \in \mathcal{F}$, the constraints $g_i(.)$ that satisfy $g_i(X) = 0$ are active constraints at X while equality constraints $h_j(.)$ are active at all points of $\mathcal{F}$. Many practical problems have objective functions that are non-differentiable, non-continuous, non-linear, noisy, multi-dimensional and have many local minima. This is the case of the calibration problem defined in Section (14.2.2).

14.3.2 Some optimisation methods

Before detailing differential evolution (DE), we introduce a few alternative optimisation methods (see Witkowski [2011]). For simplicity of exposition we define a few operators used in several optimisation methods. We let $R_U(a, b)$ to return a random uniformally distributed value between a (inclusive) and b (exclusive), and $R_N(\mu, \sigma)$ to return a random normally distributed value drawn from a normal distribution with mean μ and standard deviation σ. Further, we define the mutation operator $M(X, \delta)$, where X is a vector of size N and δ is the mutation strength, as follow

```
Begin
 i ← R_U(1, N)
 X_i ← ± R_U(0, 1). δ
return X
End
```

14.3.2.1 Random optimisation

Random optimisation is an iterative optimisation method designed for single objective problem which consists in assigning uniformally distributed random values to an individual. In each iteration, the individual is modified by adding a normally distributed vector to it, in the case where the resulting individual is better than the source individual. We let $f : \mathbb{R}^N \to \mathbb{R}$ be the fitness function subject to optimisation, and we let $X \in \mathbb{R}^N$ be a position in the search space. The vector X is initialised as follow

$$X \leftarrow [R_{U,1}(0, 1), ..., R_{U,N}(0, 1)]$$

and the algorithm satisfies

```
Begin
while not terminationCriterion() do
    X^' ← X + [R_N, 1(μ, σ), ..., R_N, N(μ, σ)]
   if f(X^') > f(X) then
       X ← X^'
   end if
end while
End
```

When the value of an individual is subject to a constraint, we consider four strategies where one or more genes do not fit the constraints.

- Dropping the individual altogether: the algorithm leaves the old value of the individual.
- Dropping only the offending gene: the old gene is left.
- Trimming the gene to the constraint: trimming either to the left or to theright value of the constraint.
- Bounce back: if g_{min} and g_{max} are the minimum and maximum values of the gene, respectively, then do

```
Begin
while not notConstraints(g) do
   if g > g_max then
       g ← g_max − |g − g_max|
   else
       g ← g_min + |g_min − g|
   end if
end while
End
```

14.3.2.2 Harmony search

The Harmony search is a type of evolutionary algorithm inspired by the process of improving jazz musicians. We consider a band participating in improvisation over chord changes of a song. The overall available harmony can be imagined as a search space of some problem where better sounding harmonies are fitter than the bad ones. Each musician represents a decision variable where at each and every practice session (one iteration) he generates some new note (a value of the decision variable). When practicing, musicians either make something up as they play along (representing a randomly created value), or use/modify some note that they have found good in the previous practice session. Notes are taken out of the memory with some given probability $P(chooseFromMemory)$ and can be modified according to some probability $P(pitchAdjust)$. Further, we define MS as the memory size, $memory$ as the harmony search memory where the memory should be at all times sorted so that

$$\forall i = 1, .., N-1 \; f(memory_i) \geq f(memory_{i+1})$$

also, δ is the pitch adjustment strength, N is the size of the decision variable, $f(\bullet)$ is the fitness function, and $memory_{MS}$ is the worst element from $memory$ according to the fitness function. The algorithm is as follow: First we initialise the memory

```
Begin
for i = 1 to MS do
    memory_i ← [R_U, 1(0, 1), ..., R_U, N(0, 1)]
end for
End
```

then we perform

```
Begin
while not terminationCriterion() do
   for i = 1 to N do
      if R_U(0, 1) ≤ P(chooseFromMemory) then
          X ← memory(R_U(0, N))
          X^′ ← X + R_N(μ, σ)
        if R_U(0, 1) ≤ P(pitchAdjust) then
            X_i ← M(X_i, δ)
        end if
      else
          X_i^′ ← R_U(0, 1)
      end if
   end for
   if f(X^′) > f(memory_MS) then
       memory_MS ← X^′
```

```
   end if
end while
End
```

14.3.2.3 Particle swarm optimisation

Particle swarm optimisation is an optimisation algorithm simulating swarming/social behaviour of agents (particles). In this scheme, a single particle is a solution candidate flying through the search space with some velocity, remembering its best position, and learning the best position known to its neighbour particles. While traveling the search space, the particles adjust their speed (both direction and value) based on their personal experiences and on the knowledge of their neighbourhood particles. Since various schemes for determining particle neighbourhood can be imagined, we can discern

- the global neighbourhood: all particles are neighbours to each other.
- the neighbourhood determined by Euclidean distance.
- the neighbourhood determined by normalised Euclidean distance.

where the normalisation process adjust the size of the neighbourhood so that the dimension of the search space is accounted for. It is a parallel direct search method using NP parameter vector

$$X_i \text{ for } i = 1, .., NP$$

as a population with vector of velocity V. The initialisation is given by

$$X_0...X_{NP} \leftarrow [R_{U,1}(0,1), ..., R_{U,N}(0,1)]$$
$$V_0...V_{NP} \leftarrow [R_{U,1}(0,1), ..., R_{U,N}(0,1)]$$

and the algorithm is as follow

```
Begin
while not terminationCriterion() do
   for i = 1 to NP do
      for j = 1 to N do
          φ_1 ← R_U(0,1)
          φ_2 ← R_U(0,1)
          V_ij^' ← w V_ij + σ ( φ_1 C_1 (best(p_i)_j − X_ij)
          + φ_2 C_2 (best_Nhood(p_i)_j − X_ij) )
      end for
      if V^' > V_max then
          V^' ← V_max
      end if
      if V^' < −V_max then
          V^' ← −V_max
      end if
       X_i^' ← X_i + V_i^'
   end for
 X ← X^'
 V ← v^'
end while
End
```

where N is the number of dimension of the search space, X is a vector of particle positions, V is a vector of particle velocities (each velocity is a vector with N elements), V_{max} is the maximum possible velocity, w is the velocity-weight or inertia factor (how much the particle will base its next velocity on its previous one), σ is the position weight determining the importance of the particles position to the next particle velocity, ϕ_i for $i = 1, 2$ are uniform random variables taken as an additional weight when determining the next particle velocity, C_1 is the self confidence weight (or self learning rate), C_2 is the swarm confidence weight (or neighbourhood learning rate), $best_{Nhood}(p)$ is an operator returning the best (most fit) position known to the particle and its neighbourhood, and $best(p)$ is an operator returning the most fit position that a particle has visited.

14.3.2.4 Cross entropy optimisation

The cross entropy optimisation consists of two steps

1. generating a sample population from a distribution.
2. updating the parameters of the random mechanism to produce a better (fitter) sample in the next population.

This behaviour conceptually substitutes the problem of finding an optimal individual to that of iteratively finding a random distribution that generate good individuals. We start by using a random Gaussian distribution to produce variables, and then improve the distribution by means of importance sampling, finding a new distribution from the best samples. We let N be the number of dimensions of the search space, N_{size} be the sample size, $(X_1, .., X_N)$ denotes the optimisation population, and we assume that the population is sorted so that

$$\forall i = 1, .., N-1 \, , f(X_i) \geq f(X_{i+1})$$

We then define the mean and standard deviation as

$$\begin{aligned} mean(X, N_{sample}) &= \frac{1}{N_{sample}} \sum_{i=0}^{N_{sample}} X_i \\ Std(X, N_{sample}) &= \sqrt{\frac{1}{N_{sample}} \sum_{i=0}^{N_{sample}} (X_i - mean(X, N_{sample}))^2} \end{aligned}$$

Since X is a vector, then $mean(X)$ and $Std(X)$ work on a set of vectors and return a vector of elements, that is, a vector of means and standard deviations of columns of the given input matrix. Additional importance sampling with the size of N_{sample} is performed on the given set. The initialisation is as follow

$$X_0 ... X_{NP} \leftarrow [R_{U,1}(0, 1), ..., R_{U,N}(0, 1)]$$

and

$$\begin{aligned} \mu &\leftarrow mean(X, N_{sample}) \\ \sigma &\leftarrow Std(X, N_{sample}) \end{aligned}$$

The algorithm is as follow

```
Begin
while not terminationCriterion() do
   for i = 1 to NP do
       X_i ← [R_N,1(μ_1, σ_1), ..., R_N,N(μ_N, σ_N)]
```

```
    end for
  μ ← mean(X, N_sample)
  σ ← Std(X, N_sample)
end while
End
```

In addition, the values of the mean μ and standard deviation σ can be smoothed over time (from iteration to iteration), improving the convergence of the algorithm.

14.3.2.5 Simulated annealing

Simulated annealing is a probabilistic optimisation method inspired by the physical process of annealing metals where a metal is slowly cooled so that its structure is frozen in a minimal energy configuration (see Tsitsiklis et al. [1993]). The algorithm is similar to hill climbing where an individual will iteratively go to a better neighbourhood position (picked at random), additionally an individual may go to a worse position with a probability proportional to its temperature. The probability that an individual will go to a worse position is calculated by the Boltzmann probability factor $e^{-\frac{E(pos)}{k_B^T}}$ where $E(pos)$ is the energy at the new position (calculated as a difference of fitness of two positions), k_B is the Botzmann constant $(1.38065052410^{-23}J/K)$, and T is the temperature of the solid. The rate at which the solid is frozen is called a cooling schedule, and we consider two variants

1. $$getTemp(t) = T_{start}(a^t)$$
   where T_{start} is the starting temperature, t is the current iteration, and a is a parameter of the cooling schedule. Further, $a > 0 \wedge a \approx 0$.

2. $$getTemp(t) = T_{start}(1-\epsilon)^{\frac{t}{m}}$$
   where ϵ and m are parameters of the cooling schedule. Further, $0 < \epsilon \leq 1$.

The initialisation is as follow:

$$X \leftarrow [R_{U,1}(0,1), .., R_{U,N}(0,1)]$$

and

$$X_{best} \leftarrow X$$

We let $f : \mathbb{R}^N \to \mathbb{R}$, and $M(X, \delta)$ be the mutation operator and define the algorithm as follow:

```
Begin
while not terminationCriterion() do
    X^' ← M(X, δ)
    Δ E ← f(X) − f(X^')
      if Δ E ≤ 0 then
          X ← X^'
         if (f(X)) > f(X_best) then
             X_best ← X
         end if
      else
          T ← getTemp(t)
         if R_U(0,1) < e^(Δ E / k_B^t) then
```

```
            X ← X^'
        end if
    end if
t ← t+1
End
```

14.3.3 The DE algorithm

According to Feoktistov [2006], Differential Evolution (DE) is first and foremost an optimisation algorithm, and in particular, one of the most powerful tools for global optimisation, regardless of it simplicity. It is so named because it identifies differences in individuals through the use of a simple and fast linear operator (differentiation), and in doing so, realises the evolution of a population of individuals in some intelligent manner. That is, the main characteristic of DE is an adaptive scaling of step sizes resulting in fast convergence behaviour. The Differential Evolution (DE) proposed by Storn and Price [1995] is an algorithm that can find approximate solutions to nonlinear programming problems. It is a parallel direct search method using NP parameter vectors, where each vector (or individual) is of dimension D equals the number of objective function parameters. The vectors

$$X_{i,G} \text{ for } i = 0, .., NP - 1$$

forms a population for each generation G, like any evolutionary algorithms. The initial vector population is chosen randomly, covering the entire parameter space. The number of vectors NP is a function of the dimension D, such as $NP \in [5D, 10D]$ but NP must be at least 4 to ensure that DE will have enough mutually different vectors to play with (see Storn et al. [1997]). Each of the NP parameter vectors undergoes mutation, recombination and selection.

14.3.3.1 The mutation

The role of mutation is to explore the parameter space by expanding the search space giving its name to the DE algorithm. For a given parameter vector $X_{i,G}$ called the Target vector, the DE generates a Donor vector V made of three or more independent parent vectors $X_{r_l,G}$ for $l = 1, 2, ..$ where r_l is an integer chosen randomly from the interval $[0, NP - 1]$ and different from the running index i. In the spirit of Wright [1991], the main idea is to perturbate a Base vector $\widehat{V}$ with a weighted difference vector (called differential vectors)

$$V = \widehat{V} + F \sum_{l=1} \left(X_{r_{2l-1},G} - X_{r_{2l},G} \right) \tag{14.3.3}$$

where the mutation factor F is a constant taking values in $[0, 2]$ and scaling the influence of the set of pairs of solutions selected to calculate the mutation value. Most of the time, the Base vector is defined as the arithmetical crossover operator

$$\widehat{V} = \lambda X_{best,G} + (1 - \lambda) X_{r_1,G}$$

where $\lambda \in [0, 1]$ allows for a linear combination between the best element $X_{best,G}$ of the parent population vectors and a randomly selected vector $X_{r_1,G}$. It is called a global selection when $\lambda = 1$ while when $\lambda = 0$ the base vector is the same as the target vector, $X_{r_1,G} = X_{i,G}$ and we get a local selection. In the special case where the mutation factor is set to zero, the mutation operator becomes a crossover operator. Figure (14.1) displays graphically the solution space and how the mutation vector V is realised. Note, the parameter vector X_{r_3} that currently resides outside of the solution space demonstrates how the mutation procedure allows for exploration of alternative solution spaces and therefore, how the algorithm is able to reliably converge to global minima.

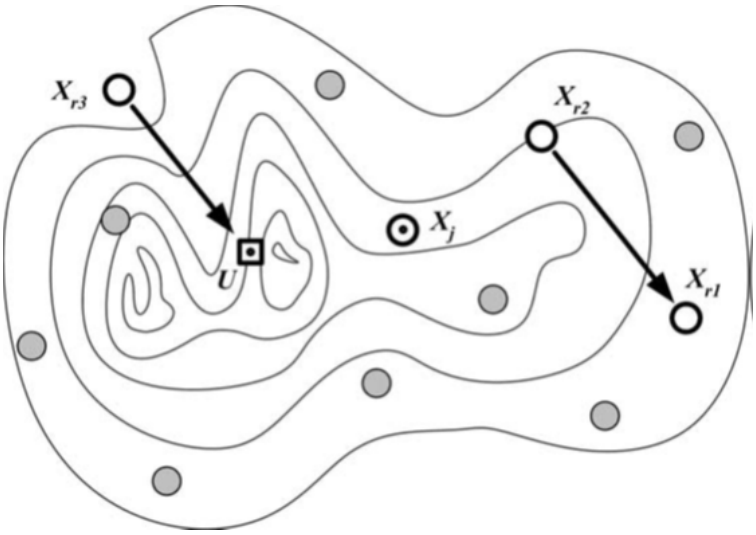


Figure 14.1: Creation of differential mutation vector. (Feoktistov [2006])

14.3.3.2 The recombination

Recombination incorporates successful solutions from the previous generation. That is, according to a rule, we combine elements of the Target vector $X_{i,G}$ with elements of the Donor vector $V_{i,G}$ to create an offspring called the Trial vector $U_{i,G}$. In order to increase the diversity of the parameter vectors, elements of the Donor vector enter the Trial vector with probability CR. In the DE algorithm, each element of the Trial vector satisfies

$$U_{i,G}(j) = \begin{cases} V_{i,G}(j) \text{ for } j = n_r \mod dim, (n_r+1) \mod dim, ..., (n_r+L-1) \mod dim \\ X_{i,G}(j) \text{ for all other } j \in [0, .., NP-1] \end{cases}$$

where $< n_r >_{dim} = n_r \mod dim$ is the modulo of n_r with modulus dim, dim is the dimension of the vector V (here $dim = N$), and the starting index n_r is a randomly chosen integer from the interval $[0, dim - 1]$. Hence, a certain sequence of the element of U is equal to the element of V, while the other elements get the original element of $X_{i,G}$. We only choose a subgroup of parameters for recombination, enhancing the search in parameter space. The integer L denotes the number of parameters that are going to be exchanged and is drawn from the interval $[1, dim]$ with probability

$$P(L > \nu) = (CR)^{\nu}, \nu > 0$$

The random decisions for both n_r and L are made anew at each new generation G. The term $CR \in [0, 1]$ is the crossover factor controlling the influence of the parent in the generation of the offspring. A higher value means less influence from the parent. Generally, values of CR in the range $[0.8, 1]$ lead to good results (see Lampinen et al. [2004]). Most of the time, the mutation operator in Section (14.3.3.1) is sufficient and one can directly set the Trial vector equal to the Donor vector.

14.3.3.3 The selection

Unlike the previous two procedures, the selection procedure is typically deterministic under Differential Evolution. The tournament selection only needs part of the whole population to calculate an individual selection probability where subgroups may contain two or more individuals. In the DE algorithm, the selection is deterministic between the parent and the child. The best of them remain in the next population. We compute the objective function with the original

vector $X_{i,G}$ and the newly created vector $U_{i,G}$. If the value of the latter is smaller than that of the former, the new Target vector $X_{i,G+1}$ is set to $U_{i,G}$ otherwise $X_{i,G}$ is retained

$$X_{i,G+1} = \begin{cases} U_{i,G} \text{ if } H(U_{i,G}) \leq H(X_{i,G}) \text{ , } i = 0, .., NP-1 \\ X_{i,G} \text{ otherwise} \end{cases}$$

This method, as applied in many Evolutionary Algorithms, ensures that the population fitness will always increase or remain constant through each generation. Mutation, recombination and selection continue until some stopping criterion is reached. The mutation-selection cycle is similar to the prediction-correction step in the EM algorithm or in the filtering problems. Figure (14.2) gives a graphical account of the tournament selection process.

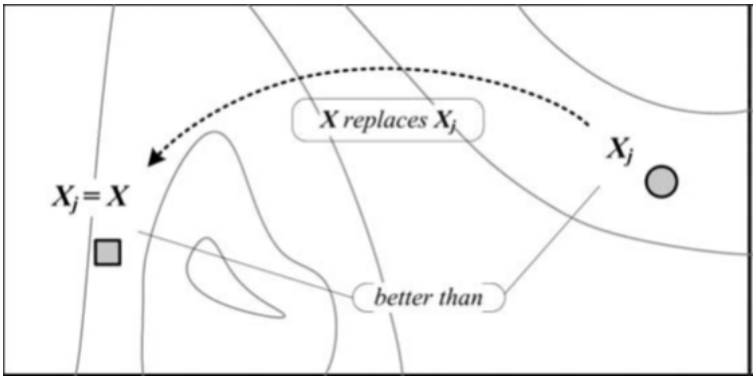


Figure 14.2: Tournament selection. (Feoktistov, [[2006]])

14.3.3.4 Simple convergence criteria

We allow for different convergence criterion in such a way that if one of them is reached, the algorithm terminates. We let $f^* = f_{min}(G)$ be the fittest design in the population so far, and we define

$$f_{a,G} = \frac{1}{NP} \sum_{i=0}^{NP-1} f(X_{i,G})$$

as the average objective value at generation G. The variance of the objective value at generation G is given by

$$f_{v,G} = \frac{1}{NP} \sum_{i=0}^{NP-1} \left(f(X_{i,G}) - f_{a,G}\right)^2 = \frac{1}{NP} \sum_{i=0}^{NP-1} f^2(X_{i,G}) - f_{a,G}^2$$

Then, when the percentage difference between the average value and the best design reaches a specified small value ϵ_1

$$\frac{|f_{a,G} - f_{min}(G)|}{|f_{a,G}|} \times 100 \leq \epsilon_1$$

we terminate the algorithm. Also, we let $f_{min}(G-1)$ be the fittest design in the previous generation $(G-1)$, and consider as a criterion the difference

$$|f_{min}(G-1) - f_{min}(G)| < \epsilon_2$$

where ϵ_2 is user defined. In that case, the DE algorithm will continue until there is no appreciable improvement in the minimum fitness value or some predefined maximum number of iterations is reached.

14.3.4 Pseudocode

We now present the pseudo code of a standard DE algorithm.

```
Initialise vectors of the population NP
Evaluate the cost of each vector
    for i=0 to Gmax do
        repeat
            Select some distinct vectors randomly
            Perform mutation
            Perform recombination
            Perform selection
                if offspring is better than main parent then
                replace main parent in the population
                end if
        until population is completed
        Apply convergence criterions
    next i
```

14.3.5 The strategies

Over the years, Storn and Price as well as a large number of other authors made improvement to the DE model so that there are now many different DE models (see Price et al. [2005], Storn [2008]). These models vary in the type of recombination operator used as well as in the number and type of solutions used to calculate the mutation values. We are now going to list the main schemes perturbating the base vector for mutation.

14.3.5.1 Scheme DE1

To improve convergence on a set of optimisation problems using the Scheme DE4 introduced in Section (14.3.5.4), Storn and Price considered for each vector $X_{i,G}$ a trial vector V generated according to the rule

$$V = X_{r_1,G} + F\big(X_{r_2,G} - X_{r_3,G}\big)$$

where the integer r_l for $l = 1, 2, 3$ are chosen randomly in the interval $[0, NP - 1]$ and are different from the running index i. Also, F is a scalar controlling the amplification of the differential variation $X_{r_2,G} - X_{r_3,G}$. We can slightly modify the scheme by writing

$$V = X_{r_1,G} + F\big(X_{r_2,G} + X_{r_3,G} - X_{r_4,G} - X_{r_5,G}\big)$$

More generally, the base vector can be expressed as a linear combination of other distinct vectors, as in Equation (14.3.3). As a simple rule, the differential weight F is usually in the range $[0.5, 1]$ (see Lampinen et al. [2004]). The population size should be between $3N$ and $10N$ and generally we increase NP if misconvergence happens. In the case where NP is increased we should decrease F.

14.3.5.2 Scheme DE2

Similarly to the Scheme DE1 in Section (14.3.5.1), the trial vector V is generated according to the rule

$$V = X_{i,G} + \lambda\big(X_{best,G} - X_{i,G}\big) + F\big(X_{r_2,G} - X_{r_3,G}\big)$$

where λ is a scalar controlling the amplification of the differential variation $X_{best,G} - X_{i,G}$. It enhances the greediness of the scheme by introducing the current best vector $X_{best,G}$. It is useful for non-critical objective functions, that is, when the global minimum is relatively easy to find. It gives a balance between robustness and fast convergence.

14.3.5.3 Scheme DE3

In the same spirit, Mezura-Montes et al. [2006] also modified the Scheme DE1 in Section (14.3.5.1) by incorporating information of the best solution as well as information of the current parent in the current population to define the new search direction.

$$V = X_{r_3,G} + \lambda\big(X_{best,G} - X_{r_2,G}\big) + F\big(X_{i,G} - X_{r_1,G}\big)$$

where λ is a scalar controlling the amplification of the differential variation $X_{best,G} - X_{r_2,G}$. This scheme has the same properties as the Scheme DE2 in Section (14.3.5.2).

14.3.5.4 Scheme DE4

The oldest strategy developed by Storn and Price [1997] is for the trial vector V to be generated according to the rule

$$V = X_{best,G} + F\big(X_{r_1,G} - X_{r_2,G}\big)$$

where the weight F is a scalar. However, in that setting they found several optimisation problems where misconvergence occurred. To improve the convergence, Price et al. [2005] proposed to perturb the differential weight F by introducing the Dither and Jitter schemes. In the former, Karaboga et al. [2004] randomised the weight as follow

$$\lambda_G = F_l + U_G(0,1)\big(F_u - F_l\big)$$

where F_l is a lower weight, F_u is an upper weight, and $U_G(0,1)$ is uniformly distributed in the range $[0,1]$ for each generation G. In the latter, the random weight is picked for each element $j = 0, ..., D-1$ of the vector being updated

$$\lambda_j = F \times \big[1 + \delta \times (U_j(0,1) - \frac{1}{2})\big]$$

where δ determines the scale of perturbation. It must be small, and is usually set to $\delta = 0.0001$. For example, we can consider the strategy

$$V = X_{best,G} + \lambda_j\big(X_{r_1,G} - X_{r_2,G}\big)$$

It is a jitter which add fluctuation to the random target. The jitter is the time variation of a periodic signal in electronics and telecommunications (swing dancer). It is tailored for small population sizes and fast convergence. We can also modify the scheme by doing

$$V = X_{best,G} + \lambda_j\big(X_{r_1,G} + X_{r_2,G} - X_{r_3,G} - X_{r_4,G}\big)$$

Going one step further, Storn [2000] combined the jitter scheme with the dither one, getting

$$\lambda_{j,G} = \Big(F_l + U_G(0,1)\big(F_u - F_l\big)\Big)\big[1 + \delta \times (U_j(0,1) - \frac{1}{2})\big]$$

14.3.5.5 Scheme DE5

Das et al. [2005] improved the DE's convergence by applying the dither scheme to every difference vector. Similarly to the Scheme DE1 in Section (14.3.5.1), the trial vector V is generated according to the rule

$$V = X_{r_1,G} + \lambda_{d,i}\big(X_{r_2,G} - X_{r_3,G}\big)$$

where $\lambda_{d,i}$ is a computer dithering factor

$$\lambda_{d,i} = F + U_i(0,1) \times (1 - F)\,,\, i = 0, ..., NP - 1$$

also written

$$\lambda_{d,i} = F_l + U_i(0,1)\big(F_u - F_l\big)\ , i = 0, ..., NP-1$$

where $U_i(0,1)$ is uniformly distributed in the range $[0,1]$. It is a per-vector dither, making the Scheme DE1 more robust. Note, as discussed in Section (14.3.5.4) we can also have

$$\lambda_{d,G} = F + U_G(0,1) \times (1-F)$$

which is a per-generation dither factor. In that algorithm, choosing $F = 0.3$ is a good start. As explained by Pedersen [2010], the dither and jitter schemes are similar, except that the dither draws a random weight once for each agent-vector to be updated and the jitter draws a random weight for each element of that vector. To see this, we simply set $F_l = F \times (1 - \frac{\delta}{2})$ and $F_u = F \times (1 + \frac{\delta}{2})$. The dither can simply be rewritten as

$$\lambda_{d,i} \sim U_i(F_l, F_u)\ , i = 0, ..., NP-1$$

where $U_i(F_l, F_u)$ is uniformly distributed in the range $[F_l, F_u]$, and the jitter can simply be rewritten as

$$\lambda_j \sim U_j(F \times (1 - \frac{\delta}{2}), F \times (1 + \frac{\delta}{2}))\ , j = 0, ..., D-1$$

where $U_i(F \times (1 - \frac{\delta}{2}), F \times (1 + \frac{\delta}{2}))$ is uniformly distributed in the range $[F \times (1 - \frac{\delta}{2}), F \times (1 + \frac{\delta}{2})]$. Defining a midpoint F_{mid} and a range F_{range}, the dither becomes

$$\lambda_{d,i} \sim U_i(F_{mid} - F_{range}, F_{mid} + F_{range})\ , i = 0, ..., NP-1$$

and the jitter becomes

$$\lambda_j \sim U_j(F_{mid} - F_{range}, F_{mid} + F_{range})\ , j = 0, ..., D-1$$

We therefore need to select two parameters, F_{mid} and F_{range}, to determine the limits of perturbation for the weight F. Pedersen chose the midpoint in the interval $F_{mid} \in [0,2]$ and the range in the interval $F_{range} \in [0,3]$ allowing for negative differential weights to occur. Thus, perturbing behavioural parameters introduce new parameters in the form of the boundaries for the stochastic sampling ranges. Further, when the behavioural parameters are completely random, it takes statistically a lot of samples before finding an optimum.

14.3.5.6 Scheme DE6

A more complex algorithm than the Scheme DE1 in Section (14.3.5.1) is to introduce the choice between two strategies. In the either-or algorithm, the trial vector V is generated according to the rule

$$V = \begin{cases} \Big(X_{r_1,G} + F\big(X_{r_2,G} - X_{r_3,G}\big)\Big) I_{(\delta < \frac{1}{2})} \\ \Big(X_{r_1,G} + \frac{1}{2}(\lambda+1)\big(X_{r_2,G} + X_{r_3,G} - 2X_{r_1,G}\big)\Big) I_{(\delta \geq \frac{1}{2})} \end{cases}$$

where δ is as above. It alternates between differential mutation and three-point recombination.

14.3.5.7 Scheme DE7

Alternatively, we have the two possible strategies using an either-or algorithm. If we favor the r_1 event we have

$$V = \Big(X_{r_1,G} + F\big(X_{r_2,G} - X_{r_3,G}\big)\Big) I_{(\delta < P_e)}$$

while if we favor the best event we have

$$V = \Big(X_{best,G} + F\big(X_{r_2,G} - X_{r_3,G}\big)\Big) I_{(\delta<P_e)}$$

It alternates between differential mutation and doing nothing.

14.3.5.8 Scheme DE8

Still another more complex algorithm than the Scheme DE1 in Section (14.3.5.1) is to introduce two strategies V_l for $l = 1, 2$. We first build the Base vector $\widehat{V}_1$ as the linear combination of the two original base vector $X_{r_1,G}$ and $X_{best,G}$, getting

$$\widehat{V}_1 = \lambda_{M,G} X_{best,G} + \overline{\lambda}_{M,G} X_{r_1,G}$$

where $\overline{\lambda}_{M,G} = 1 - \lambda_{M,G}$. This time $\lambda_{M,G}$ is a Gaussian variable per-generation with mean μ and variance ξ to be defined. The second Base vector $\widehat{V}_2$ is generated according to the rule

$$\widehat{V}_2 = 2X_{best,G} - \widehat{V}_1 = (2 - \lambda_{M,G})X_{best,G} - \overline{\lambda}_{M,G} X_{r_1,G}$$

The two Trial vectors V_i for $i = 1, 2$ becomes

$$\begin{aligned} V_1 &= \widehat{V}_1 + F_b\big(X_{r_3,G} - X_{r_2,G}\big) \\ V_2 &= \widehat{V}_2 + F_b\big(X_{r_2,G} - X_{r_3,G}\big) = 2X_{best,G} - V_1 \end{aligned}$$

where F_b is a Gaussian variable with mean $\mu = 0$ and variance $\xi = F$. In the special case where $\mu = 2$ and $\xi = 0$ we get

$$\begin{aligned} \widehat{V}_1 &= 2X_{best,G} - X_{r_1,G} \\ \widehat{V}_2 &= X_{r_1,G} \end{aligned}$$

and the system simplifies to

$$\begin{aligned} V_1 &= 2X_{best,G} - X_{r_1,G} + F_b\big(X_{r_3,G} - X_{r_2,G}\big) \\ V_2 &= X_{r_1,G} + F_b\big(X_{r_2,G} - X_{r_3,G}\big) \end{aligned}$$

and V_2 recover a pseudo Secheme DE1 where $F_b \in [-F, F]$.

14.3.6 Improvements

The DE algorithm is found to be a powerful evolutionary algorithm for global optimisation in many real problems. As the DE algorithm performs mutation based on the distribution of the solutions in a given population, search directions and possible step sizes depend on the location of the individuals selected to calculate the mutation values. As a result, since the original article of Storn and Price [1995] many authors improved the DE model to increase the exploration and exploitation capabilities of the DE algorithm when solving optimisation problems. We are going to review a few changes to the DE algorithm which greatly improved the performances of our problem.

14.3.6.1 The tuning parameters

The tuning of the DE is mainly contolled by three variables: the number of vector (or individual) NP, the differential weight F, and the crossover factor CR. Finding bounds for their values has been a topic of intensive research. Zaharie [2002] studied theoretically the convergence of the DE by analysing the behavioural parameters of the DE models. She proved that the mutation scale factor F should never be smaller than F_{crit} defined as

$$F_{crit} = \sqrt{\frac{(1 - \frac{CR}{2})}{NP}}$$

Price et al. [2005] showed that only high values of CR guarantee the contour matching properties of DE. Further, only when $CR = 1$ is the mean number of function evaluation for an objective function and its rotated counterpart the same. In that setting, the DE is called rotationally invariant. As a rule of thumb (see Storn et al. [1997]), we get

- $F \in [0.5, 1.0]$
- $CR \in [0.8, 1.0]$
- $NP = 10D$

but they lack generality. Hence, the need to compute these parameters automatically.

14.3.6.2 Ageing

The DE selection is based on local competition only. The number of children that may be produced to compete against the parent $X_{i,G}$ should be chosen sufficiently high so that a sufficient number of child will enter the new population. Otherwise, it would lead to survival of too many old population vectors that may induce stagnation. To prevent the vector $X_{i,G}$ from surviving indefinitely, Storn [1996] used the concept of ageing. One can define how many generations a population vector may survive before it has to be replaced due to excessive age. If the vector $X_{i,G}$ is younger than Num generations it remains unaltered otherwise it is replaced by the vector $X_{r_3,G}$ with $r_3 \neq i$ being a randomly chosen integer in $[0, NP - 1]$.

14.3.6.3 Constraints on parameters

Commonly, we are searching for a solution to an optimisation problem between certain bounds. Given the parent vector $X_{i,G}$ for $i = 0, .., NP - 1$ we define upper and lower bounds for each initial parameters as

$$L(j) \leq X_{i,G_0}(j) \leq U(j) \, , j = 0, ..., D - 1$$

where G_0 is inception, and we randomly select the initial parameter values uniformly on the interval $[L(j), U(j)]$ as

$$X_{i,G_0}(j) = L(j) + U_j(0,1)\big(U(j) - L(j)\big) \, , j = 0, ..., D - 1$$

where $U_j(0,1)$ generates a random number in the range $[0,1]$ with a uniform distribution for each element j of the vector. Obviously, as the number of generation G increases, the DE algorithm will generate elements of the vector outside of the limits established (lower and upper) by an amount. Several alternatives exist for handling boundary constraints (see Onwubolu [2004]). Following Mezura-Montes et al. [2004a], this amount is substracted or added to the limit violated (reflecting barrier), in order to shift the value inside the limits. Should this cause the new value to violate the other bound, a random value will be generated, as when creating the initial parameter set. We can also set the element of the vector half way between the old position and the limit as follow

$$U_{i,G+1}(j) = \begin{cases} \frac{1}{2}(X_{i,G}(j) + U(j)) \text{ if } U_{i,G+1}(j) > U(j) \\ \frac{1}{2}(X_{i,G}(j) + L(j)) \text{ if } U_{i,G+1}(j) < L(j) \\ U_{i,G+1}(j) \text{ otherwise} \end{cases}$$

where $U_{i,G+1}(j)$ is the new Trial vector.

14.3.6.4 Convergence

In order to accelerate the convergence process, when a child replaces its parent, Mezura-Montes et al. [2004a] copied its value both into the new generation and into the current generation. It allows the new child, which is a new and better solution, to be selected among the r_l solutions and create better solutions. Therefore, a promising solution does not need to wait for the next generation to share its genetic code. Similarly, to improve performance and to accelerate the convergence process, Storn [1996] explored the idea of allowing a solution to generate more than one offspring. Once a child is better than its parent, the multiple offspring generation ends. Following the same idea, Coello Coello and Mezura-Montes [2003] and then Mezura-Montes et al. [2006] allowed for each parent at each generation to generate $k > 0$ offspring. Among these newly generated solutions, the best of them is selected to compete against its parent, increasing the chances to generate fitter offspring.

14.3.6.5 Self-adaptive parameters

It was proved that key control parameters in the DE algorithm, such as the crossover CR and the weight applied to random differential F, should be altered in the evolution process itself (see Liu et al. [2002]). These parameters can be adjusted by using heuristic rules, or they can be self-adapted (see Liu et al. [2005], Brest et al. [2006], Balamurugan et al. [2007]). That is, the control parameters are not required to be pre-defined and can change during the evolution process. These control parameters are applied at the individual levels in the population, such that better values should lead to better individuals producing better offspring and hence better values. However, in general, the random change to the parameters are applied irrespective of the quality of the current parameters. Noman et al. [2011] proposed to preserve better parameter choices, while changing the non-productive ones. We first describe the algorithm proposed by Brest et al. [2006] (jDE), and then introduce the adaptative algorithm of Noman et al. [2011] (aDE). The parameter F is a scaling factor controlling the amplification of the difference between two individuals to avoid search stagnation. At generation $G = 1$ the amplification factor $F_{i,G}$ for the ith individual ($i = 0, ..., NP - 1$) is generated randomly in the range $[0.1, 1.0]$. Then, at the next generations the control parameter is given by

$$F_{i,G+1} = \begin{cases} F_L + r_1 \times F_U \text{ if } r_2 < \tau_1 \\ F_{i,G} \text{ otherwise} \end{cases}$$

where r_j, for $j = 1, 2$ are uniform random values in $[0, 1]$, $F_L = 0.1$, $F_U = 0.9$ and τ_1 represent the probability to adjust the parameter F. Using the notation in Section (14.3.5.5), we can rewrite the scaling factor as

$$F_{i,G+1} = \begin{cases} U_i(F_l, F_u) \text{ if } r_2 < \tau_1 \\ F_{i,G} \text{ otherwise} \end{cases}$$

where $F_l = 0.1$ and $F_u = F_L + F_U = 1.0$. The only difference with the dither is the fact that the randomness of the differential weight depends on a probability of adjustment. According to Feoktistiv [2006], at the beginning of the evolution procedure, the mutation step length, and hence $F_{i,G}$, should be large, as individuals are far away from each other and the procedure could benefit from exploring beyond the current solution space. Since the individuals of a generation converge for subsequent generations, the step length, and hence $F_{i,G}$ should become smaller to allow a more concentrated search around the successful solution space. Thus, by applying custom parameters at individual levels, better values in the population lead to better individuals producing better Donor vectors and so on. Similarly, we can extend the crossover parameter CR as follow

$$CR_{i,G+1} = \begin{cases} r_3 \text{ if } r_4 < \tau_2 \\ CR_{i,G} \text{ otherwise} \end{cases}$$

where r_j, for $j = 3, 4$ are uniform random values in $[0, 1]$ and τ_2 represent the probability to adjust the parameter CR. The new parameter CR takes random values in the range $[0, 1]$. Since $F_L = 0.1$, $F_U = 0.9$, Brest et al. [2006] proposed to set $\tau_1 = \tau_2 = 0.1$ such that F takes random values in the range $[0.1, 1]$. Note, both $F_{k,G+1}$ and $CR_{k,G+1}$ are obtained before the mutation is performed in order to influence the mutation, crossover, and selection of the new vector $X_{i,G+1}$. Given that random adjustment can only be good when the parameter setting is not suitable, Noman

et al. [2011] proposed to compare the fitness of the offspring $f(U_G)$ with the average fitness value of the current generation, $f_{a,G}$. Then, the choice of the amplification factor F in offspring U_G for the parent $X_{i,G}$ is given by

$$F_G^{child} = \begin{cases} F_{i,G} \text{ if } f(U_G) < f_{a,G} \\ r_1(0.1, 1.0) \text{ otherwise} \end{cases}$$

and that of the crossover parameter is given by

$$CR_G^{child} = \begin{cases} CR_{i,G} \text{ if } f(U_G) < f_{a,G} \\ r_2(0.1, 1.0) \text{ otherwise} \end{cases}$$

where $r_j(a, b)$ are uniform random number in the range $[a, b]$. Initially, the parameters F and CR are created randomly for each individual. Note, the objective vector part of the offspring is created by using the original mutation and crossover values of the DE.

***Remark* 14.3.1** *Again, perturbing behavioural parameters introduce new parameters in the form of the boundaries for the stochastic sampling ranges and the adjustment probabilities.*

14.3.6.6 Selection

Santana-Quintero et al. [2005] maintained two different populations (primary and secondary) according to some criteria and considered two selection mechanisms that are activated based on the total number of generation G_{max} and the parameter $sel_2 \in [0.2, 1]$ which regulates the selection pressure. That is

$$\text{Type of selection } = \begin{cases} \text{Random if } G < (sel_2 * G_{max}) \\ \text{Elitist otherwise} \end{cases}$$

a random selection is first adopted followed by an elitist selection. In the random selection, three different parents are randomly selected from the primary population while in the elitist selection they are selected from the secondary one. In both selections, a single parent is selected as a reference so that all the parents of the main population will be reference parents only once during the generating process.

14.3.7 Convergence criteria revised

When applying a genetic algorithm to solve some NP-hard optimisation problem (the optimum is unkown), it is usually infeasible to compute the optimal solution of the problem. Still, we need to determine whether or not the algorithm has converged to some optimum. Since a genetic algorithm is said to converge when there is no significant improvement in the values of fitness of the population from one generation to the next, there is no defined difference between stopping criteria and convergence criteria. That is, the stopping criterion provides the user a guideline in stopping the algorithm with an acceptable solution close to the optimal solution. While, mathematically, that closeness may be judged in various ways, termination criteria should avoid needless computation and prevent premature termination. Thus, stopping criteria should account for

- Reliability guarantees termination within a finite time.
- Performace guarantees no premature termination and no needless computation.

Stopping criteria are generally based on time or fitness value. The simpliest termination criteria, called exhaustion-based criteria, is to select a maximum number of generations of evaluation functions, or, a maximal time budget (absolute time, CPU time). Alternatively, we can terminate the search when the best objective value $f^* = f_{min}$ reaches or surpasses a bound, $f^* \leq f_{lim}$. Usually, if there is no change in the best fitness value for K consecutive iterations, the algorithm is terminated. This method works well if the optimum or a lower bound is known. Some authors proposed tight bounds on the number of iterations required to achieve a level of confidence to guarantee that

a Genetic Algorithm has seen all strings (see Aytug et al. [1996]). Giggs et al. [2006] empirically studied a way to determine the maximum number of generations. However, determining the optimum time, or, finding a lower bound is a challenge. Thus, stopping criteria should contain the advantage of reacting adaptively to the state of the optimisation run.

The criteria based on objective function values use the underlying fitness function values to calculate auxiliary values as a measure of the state of the convergence of the GA. For instance,

- improvement-based criteria monitor the improvement of the best objective function value (ImpBest) (or its average ImpAvg) along the optimisation process, and stops when it falls below a user-defined threshold for a given number of generations.
- movement-based criteria monitor the distances between the population members in successive iterations (see Schwefel [1995]). The movement in the population can be calculated with respect to the average objective function value (MovObj), or with respect to positions (MovPar). Termination occurs when it is below a threshold for a given number of generations.
- distribution-based criteria monitor the distances of the population members at each iteration (see Zielinski et al. [2008]). It is assumed that all individuals converge to the optimum, such that convergence is reached when they are close to each other. MaxDist is when the maximum distance from every vector to the best population vector is below a threshold.
- combined criteria use several criteria in combination.

It is understood that the algorithm often focuses on global optimisation at the beginning, leading to large movements between population members in successive iterations, and that at the final stages of the optimisation process, the population generally converges to one point. There are different ways of measuring these distances, such as the standard deviation of positions to all, or part, of the population members, or the distance between the individuals with the best and worst objective function value. In any case, when the distance falls below a user-defined threshold for a given number of iterations, the algorithm terminates. For example, the Running Mean is the difference between the current best objective value $f_{min}(G)$, at generation G, and the average of the best objective value

$$f_{a,min}(n) = \frac{1}{G_{last}} \sum_{i=0}^{n} f_{min}(G_i) \, , G_i = G - i\Delta G$$

over a period of time $G_{last} = n\Delta G$, where ΔG is one generation. It must be less or equal to a given threshold, that is,

$$|f_{min}(G) - f_{a,min}(n)| \leq \epsilon$$

The Best-Worst is the difference between the best objective value $f_{min}(G)$ and the worst one $f_{max}(G)$ at generation G. At least $p\%$ of the individuals must be feasible, and it must be less or equal to a given threshold, $|f_{min}(G) - f_{max}(G)| \leq \epsilon$. We can also consider relative termination criterions such as $\frac{f_{min}}{f_{a,G}} \leq \epsilon$, or letting d_{ij} be the sum of all normalised distance between all individuals of the current generation, the ratio $\frac{d_{ij}}{K_{max}} \leq \epsilon$. The latter values the spacial speading of individuals of the current generation in the search space (normalised Euclidian distances $\frac{d_{ij}}{d}$ where d is the length of diagonal of the search space). Jain et al. [2001] proposed the Clus Term, combining information from the objective values and the distribution of individuals in the search space. They perform a cluster analysis (single linkage method) of the fittest individuals and determine the total amount $N(t)$ of of individuals in clusters. The search is terminated when the change of the average of $N(t)$ is equal or less than ϵ. Analysis of stopping criteria reacting adaptively to the state of an optimisation were perfomed (see Zielinski et al. [2005], Zielinski et al. [2007]).

Generally, EAs are stopped or terminated when the variance of fitness values of all the strings in the current population is less than a predefined threshold ϵ. It is assumed that after significantly many iterations the fitness values

of the strings present in the population are all close to each other (the population becomes homogeneous), thereby making the variance of the fitness values close to 0. Thus ϵ should be chosen close to zero. Further, one should select a significant number of iterations from which the fitness values will be considered in calculating the variance so that the algorithm gets enough opportunity to yield improved (better) solution. However, this is not correct since

- in elitist model, or other GAs, only the best string is preserved.
- any population containing an optimal string is sufficient for the convergence of the algorithm.
- there is a positive probability of obtaining a population after infinitely many iterations with exactly one optimal string and others are being not optimal.

A lot more iterations often occur after the global optimum has been reached, or nearly reached, to improve other inferior individuals. However, only the best objective function value is important, rather than the convergence of the whole population. Since at the beginning, global search dominates the optimisation algorithm, while at final stages, the algorithm focuses on local optimisation, Liu et al. [2009] proposed a combination of global and local methods. They monitor the average improvement of the whole population in the former, and they monitor the best objective function value in the latter. Bhandari et al. [2012] established theoretically that the variance of the best fitness values obtained in the iterations can be considered as a measure to decide the termination criterion of a GA with elitist model (EGA). The proposed criterion uses only the fitness function values and takes into account the inherent properties of the objective function. We let $f_{min}(i)$ be the best fitness function value obtained at the end of ith iteration, such that $f_{min}(1) \geq f_{min}(2) \geq ... \geq f_{min}(n) \geq ... \geq F_1$, where F_1 is the global optimal value of the fitness function (F_i denotes the ith lowest fitness function value). Then we get the statistical mean and variance of the best fitness values obtained up to the nth iteration as

$$\begin{aligned} f_{min,1}(n) &= \frac{1}{n}\sum_{i=1}^{n} f_{min}(i) \\ b(n) &= Var(f_{min}(n)) = \frac{1}{n}\sum_{i=1}^{n}\big(f_{min}(i) - f_{min,1}(n)\big)^2 = f_{min,2}(n) - f_{min,1}^2(n) \end{aligned}$$

where $f_{min,2}(n) = \frac{1}{n}\sum_{i=1}^{n} f_{min}^2(i)$. The variance can be iteratively calculated as follow

$$b_{n+1} = \frac{1}{n+1}\Big((nf_{min,2}(n) + f_{min}^2(n+1)) - \big(f_{min,1}(n) + f_{min}(n+1)\big)^2\Big)$$

such that only $f_{min,1}(n)$ and $f_{min,2}(n)$ at step n are required to keep in memory when computing the variance at step $(n+1)$. Alternatively, following Equation (B.10.9), we can write recursively the sample mean as

$$f_{min,1}(n+1) = f_{min,1}(n) + \frac{f_{min}(n+1) - f_{min,1}(n)}{n+1}$$

and following Equation (B.10.10), the sample variance becomes

$$b(n+1) = b(n) + f_{min,1}^2(n) - f_{min,1}^2(n+1) + \frac{f_{min}^2(n+1) - b(n) - f_{min,1}^2(n)}{n+1}$$

So far, the variance based criterion is not scale invariant, meaning it is sensitive to transformations of the fitness function [1]. One can easily avoid the impact of the scaling effect by a simple transformation of the fitness function, such as

[1] $g(x) = k \times f(x)$ where k is a constant.

$$g(x) = \frac{f(x)}{f^1_{min}}$$

where f^1_{min} is the minimum value of the fitness function obtained in the first iteration. If we let $b_n(g)$ be the variance of the best fitness values obtained up to the nth iteration for the function $g(x)$, then we get

$$b_n(g) = \frac{1}{n(f^1_{min})^2}\sum_{i=1}^{n}\big(f_{min}(i) - f_{min,1}(n)\big)^2 = \frac{1}{(f^1_{min})^2}b_n(f)$$

such that assuming the tolerence level ϵ_f as the value ϵ for the function f is equivalent to assuming $\epsilon_f \times (f^1_{min})^2$ as the value of ϵ for the function g. It implies that the user has to adjust the value of ϵ for the applied transformation. Note, we now need to use Equation (B.10.11) to obtain recursively the sample mean, and Equation (B.10.12) to obtain recursively the sample variance. In that setting, the GA is stopped or terminated after N iterations when the variance of the best fitness values obtained so far is bounded. That is, $b_N < \epsilon$, where $\epsilon > 0$ is a user defined small quantity corresponding to the difference between the fitness value of the best solution obtained so far and the global optimal solution.

14.4 Handling the constraints

14.4.1 Describing the problem

We saw above that EAs in general, and DE in particular, lacked a mechanism to deal with the constraints of the problems. Recently, various academics worked on solving that problem, and one of the most popular constraint handling mechanisms was proposed by Deb [2000] on genetic algorithms (GAs) who used the three feasibility rules. This algorithm generates feasible individuals, while maintaining a reasonable ratio between feasible and infeasible members in a population, allowing for a sparse feasible domain. Several studies assessed the performances of these rules on a large number of test functions, and very good results were found (see Zielinski et al. [2006], Zhang et al. [2012]). Before reviewing his method and showing how it was improved, we recall that Michalewicz [1995] discussed different constraint handling methods used in GAs and classified them in five categories

- methods based on preserving feasibility of solutions, that is, we use a search operator that maintains the feasibility of solutions
- methods based on penalty functions
- methods making distinction between feasible and infeasible solutions using different search operators for handling infeasible and feasible solutions
- methods based on decoders using an indirect representation scheme which carries instructions for constructing feasible solutions
- hybrid methods where evolutionary methods are combined with heuristic rules or classical constrained search methods

In a single-objective optimisation problem, the traditional approach for handling constraints is the penalty function method. The fitness of a candidate is based on a scale function F which is a weighted sum of the objective function value and the amount of design constraint violation

$$F(X) = f_1(X) + \Big(\sum_{k=1}^{p}\omega_k \max\left(g_k(X), 0\right) + \sum_{k=p+1}^{q}\omega_k|h_k(X)|\Big)$$

where ω_k are positive penalty function coefficients and such that the kth constraint $g_k(.)$ and $h_k(.)$ should be normalised. This method requires a careful tuning of the coefficients ω_k to obtain satisfactory design, that is a balance between the objective function and the constraints but also between the constraints themselves. Kusakci et al. [2012] presented a literature review summerising the constraint handling techniques for constrained optimisation problems (COPs).

14.4.2 Defining the feasibility rules

To overcome this problem, Deb [2000] proposed a penalty function approach based on the non-dominance concept, ranking candidates using the definition of domination between two candidates.

Definition 14.4.1 *A solution i is said to dominate a solution j if both of the following conditions are true*

1. *solution i is no worse than solution j in all objective*

$$\forall f_m(X_i) \leq f_m(X_j)$$

2. *solution i is strictly better than solution j in at least one objective*

$$\exists f_m(X_i) < f_m(X_j)$$

The constrained domination approach ranks candidates according to the following definition

Definition 14.4.2 *A solution i is said to constrained-dominate a solution j if any of the following conditions is true*

1. *solutions i and j are feasible and solution i dominates solution j.*
2. *solution i is feasible and solution j is not.*
3. *both solutions i and j are infeasible but solution i has a smaller constraint violation.*

He let the fitness function be

$$F(X) = \begin{cases} f(X) \text{ if } g_k(X) \leq 0 \; \forall k = 1, 2, .. \\ f_{max} + TACV \text{ otherwise} \end{cases}$$

where f_{max} is the objective value with the worst feasible solution in the population and (TACV) is the total amount of constraint violation

$$TACV = \sum_{k=1}^{p+q} \max\left(g_k(X), 0\right)$$

Therefore, solutions are never directly compared in terms of both objective function and constraint violation information. However, the high selection pressure generated by tournament selection will induce the use of additional procedure to preserve diversity in the population such as niching or sharing. Clearly, there is no tuning of the penalty function coefficients when the number of constraint is one. But, when multiple constraints are considered some considerations must be taken to relate constraints together. One way forward is to normalise the constraints such that every constraint has the same contribution to the comparing value as was done by Landa Becerra et al. [2006]. Letting $g_{max}(k)$ be the largest violation of the constraint $\max\left(g_k(X), 0\right)$ found so far, we define the new TACV as

$$NTACV = \sum_{k=1}^{p} \frac{\max\left(g_k(X), 0\right)}{g_{max}(k)}$$

14.4.3 Improving the feasibility rules

Again, many different approaches were proposed, for instance Coello Coello [2000] modified the definition of the constrained domination approach given in Definition (14.4.2) such that if the individuals are infeasible he compares the number of constraints violated first and only in the case of a tie would he use the total amount of constraint violation in the definition, getting

Definition 14.4.3 *A solution i is said to constrained-dominate a solution j if any of the following conditions is true*

1. *solutions i and j are feasible and solution i dominates solution j.*
2. *solution i is feasible and solution j is not.*
3. *both solutions i and j are infeasible but solution i violates less number of constraints than solution j.*
4. *both solutions i and j are infeasible and violating the same number of constraints but solution i has a smaller TACV than solution j.*

In that setting, the fitness of an infeasible solution not only depends the amount of constraint violation, but also on the population of solutions at hand. However, this technique may not be very efficient when the degrees of violation of constraints $g_k(X)$ are significantly different because the TACV is a single value. Alternatively, Coello Coello et al. [2002] handled constraints as additional objective functions and used the non-dominance concept (on objective functions) in Definition (14.4.1) to rank candidates. As a result, it required solving the objective function a large number of time. Going one step further, Oyama et al. [2005] introduced dominance in constraint space.

Definition 14.4.4 *A solution i is said to dominate a solution j in constraint space if both of the following conditions are true*

1. *solution i is no worse than solution j in all constraints*

$$\forall g_k(X_i) \leq g_k(X_j)$$

2. *solution i is strictly better than solution j in at least one constraint*

$$\exists g_k(X_i) < g_k(X_j)$$

Introducing the idea of non-dominance concept to the constraint function space, their proposed constraint-handling method is

Definition 14.4.5 *A solution i is said to constrained-dominate a solution j if any of the following conditions is true*

1. *solutions i and j are feasible and solution i dominates solution j in objective function space.*
2. *solution i is feasible and solution j is not.*
3. *both solutions i and j are infeasible but solution i dominates solution j in constraint space.*

In that setting, any non-dominance ranking can be applied to feasible designs and infeasible designs separately. As a result, in a single-objective constrained optimisation problem, Bloch [2010] modified the dominance-based tournament selection of Coello and Mezura with the non-dominance concept of Oyama et al., getting

Definition 14.4.6 *The new dominance-based tournament selection is*

1. *if solutions i and j are both feasible and solution i dominates in objective function solution j then solution i wins.*

2. *if solution i is feasible and solution j is not, solution i wins.*
3. *if solutions i and j are both infeasible and solution i dominates in constraint space solution j then solution i wins.*
4. *if solutions i and j are infeasible and non-dominated in constraint space, if solution i violates less number of constraints than solution j then solution i wins.*
5. *if solutions i and j are both infeasible, non-dominated in constraint space and violating the same number of constraints but solution i has a smaller TACV than solution j then solution i wins.*

14.4.4 Handling diversity

In order to explore new regions of the search space and to avoid premature convergence, a set of feasibility rules coupled with a diversity mechanism is proposed by Mezura-Montes et al. [2006]. It maintains besides competitive feasible solutions some solutions with a promising objective function value allowing for the DE to reach optimum solutions located in the boundary of the feasible region of the search space. Given a parameter S_r one can choose to select between parent and child based either only on the objective function value or on feasibility. As more exploration of the search space is required at the beginning of the process, the parameter S_r will be decreasing in a linear fashion with respect to the number of generation. Given an initial value $S_{r,0}$ and a terminal value $S_{r,\infty}$, the adjustment of the parameter is

$$S_{r,G+1} = \begin{cases} S_{r,G} - \Delta_{S_r} \text{ if } G > G_{max} \\ S_{r,\infty} \text{ otherwise} \end{cases}$$

where $\Delta_{S_r} = \frac{S_{r,0} - S_{r,\infty}}{G_{max}}$. After some generations, assuming that promising areas of the feasible region have been reached, we focus on keeping the feasible solutions found discarding the infeasible ones. In that setting, infeasible solutions with good objective function values will have a significant probability of being selected which will slowly decrease as the number of generation increases.

In some problems, it does not make sense to use the infeasible individual unless they are very close to the boundary between the feasible and infeasible region. This is the case of the calibration problem given in Section (14.2.2) where the constraints ensure that prices do not violate the AAO rules. To circumvent this issue, Mezura-Montes et al. [2004a] proposed that at each generation the best infeasible solution with lowest amount of constraint violation both in the parents (μ) and the children (λ) population will survive. Either of them being chosen with an appropriate probability, it will allows for infeasible solutions close to the boundary to recombine with feasible solutions. The pseudo code for introducing diversity in the DE algorithm is

```
if   flip(S_r)   then
   Select the best infeasible individual from the children population
else
   Select the best individual based on five selection criteria
end if
```

14.5 The proposed algorithm

Using the improvements in Section (14.3.6) and applying the Definition (14.4.6) on dominance-based tournament selection to handle the constraints, the pseudo code for the DE algorithm with constraints becomes

```
Begin
 G = 0   and   Age_i,G  =  0  ∀  i,   i = 1,..,NP
```

```
Create a random initial population  X_i,G ∀ i,  i = 1,..,NP
Evaluate  f(X_i,G) ∀ i,  i = 1,..,NP
while  niter < max_iter  and  G < Gmax  do
    fmin_old = fmin
   for  i = 1  to  NP  do
      for  k = 1  to  N_K  do
         Select randomly  r_1 ≠ r_2 ≠ r_3 ≠ i
          j_r = U(1,N)
         for  j = 1  to  N  do
            if  U(0,1) < CR  or  j = j_r  then
                U_i,G(j) = X_r_3,G(j) + F (X_r_1,G(j) − X_r_2,G(j))
            else
                U_i,G(j) = X_i,G(j)
            end if
         end for
         if  k > 1  then
            if  U_i,G(j)  is better than  U_i_best,G(j)  based on five selection
            criteria then
                U_i_best,G(j) = U_i,G(j)
         else
             U_i_best,G(j) = U_i,G(j)
      end of for
      Apply diversity :
      if  flip(S_r)  then
         if  f(U_i_best,G) ≤ f(X_i,G)  then
             X_i,G+1 = X_i,G = U_i_best,G
         else
             X_i,G+1 = X_i,G
         end if
      else
         if  U_i_best,G  is better than  X_i,G  based on five selection criteria then
             X_i,G+1 = X_i,G = U_i_best,G
         else
            if  Age_i,G < N_A  or  i = i_best  then
                X_i,G+1 = X_i,G
            else
               Select randomly  r_4 ≠ i
                X_i,G+1 = X_r_4,G
                Age_i,G = 0
            end if
         end if
      end if
      if  fmin > f(X_i,G)
          fmin = f(X_i,G)
         i_best = i
      end if
   end for
   Apply convergence criterions :
   if  fmin_old − f_min < precision
      niter = niter + 1
```

```
    else
       niter = 0
    end if
     G = G + 1
end while
End
```

14.6 Describing some benchmarks

According to the No Free Lunch theorem stating that any two algorithms A and B on average perform identially, it is very difficult (see Wolpert et al. [1997]), if not impossible, to devise a test suite identifying the best algorithm among different algorithms. Nonetheless, we can design a benchmark for assessing the performances of these algorithms based on some criteria, such as,

- the amount of iterations.
- the time taken by an algorithm to find a known optima, or how quickly it improves the best known solution of some optimisation problem.
- the number of evaluations of the fitness function needed to find the best solution.
- the rate of deterioration of the algorithm as the dimensions of the problem increase.
- whether the algorithms find the best solution.
- the distance to the known optimum.

We are now going to present a few test functions considered as benchmark for different optimisation methods. Some of these benchmark functions and set of engineering problems are detailed by Mezura-Montes et al. [2006b] and Witkowski [2011]. We are going to describe a few of them. To do so we let

$$f_n^N : \mathbb{R}^N \to \mathbb{R}$$

be the test function where N represents the dimensionality of the problem and n is a referrence number.

14.6.1 Minimisation of the sphere function

The problem consists in minimising the function

$$f_1^N(X) = \sum_{i=1}^{N} X_i^2$$

where

1. Search domain: $|X_i| < 5.12$, $i = 1, 2, ..., N$
2. Global minimum: $X^* = (0, ..., 0)$, $f(X^*) = 0$
3. No local minima besides the global minimum

14.6.2 Minimisation of the Rosenbrock function

The problem consists in minimising the function

$$f_2^N(X) = \sum_{i=1}^{N-1} \left[100(X_i^2 - X_{i+1})^2 + (X_i - 1)^2\right]$$

where

1. Search domain: $|X_i| < 5.12$, $i = 1, 2, ..., N$
2. Global minimum: $X^* = (1, ..., 1)$, $f(X^*) = 0$
3. Several local minima

14.6.3 Minimisation of the step function

The problem consists in minimising the function

$$f_3^N(X) = \sum_{i=1}^{N} \lfloor X_i \rfloor$$

where

1. Search domain: $|X_i| < 5.12$, $i = 1, 2, ..., N$
2. Global minimum: $X^* : X_i \leq -5$, $f(X^*) = -6N$, $i = 1, 2, ..., N$
3. No local minima besides the global minimum

14.6.4 Minimisation of the Rastrigin function

The problem consists in minimising the function

$$f_4^N(X) = 100N + \sum_{i=1}^{N} \left(X_i^2 - 10\cos(2\pi X_i)\right)$$

where

1. Search domain: $|X_i| < 5.12$, $i = 1, 2, ..., N$
2. Global minimum: $X^* = (0, ..., 0)$, $f(X^*) = 0$
3. Several local minima

14.6.5 Minimisation of the Griewank function

The problem consists in minimising the function

$$f_5^N(X) = \sum_{i=1}^{N} \frac{X_i^2}{4000} - \prod_{i=1}^{N} \cos(\frac{X_i}{\sqrt{i}}) + 1$$

where

1. Search domain: $|X_i| \leq 600$, $i = 1, 2, ..., N$
2. Global minimum: $X^* = (0, ..., 0)$, $f(X^*) = 0$
3. Several local minima

14.6.6 Minimisation of the Easom function

The problem consists in minimising the function

$$f_6^N(X_1, X_2) = -\cos(X_1)\cos(X_2)e^{-(X_1-\pi)^2-(X_2-\pi)^2}, X_i : |X_i| \leq 100$$

where

1. Number of variables: 2
2. Search domain: $|X_i| \leq 100$, $i = 1, 2$
3. Global minimum: $X^* = (\pi, \pi)$, $f(X^*) = -1$
4. No local minima besides the global minimum

14.6.7 Image from polygons

We consider the problem of finding the best combination of N semi-transparent coloured D-gons ($100 \leq N \leq 1000, 3 \leq D \leq 10$) that when rendered will produce an image I. We want to minimise the difference between the rendered image and the given one. The polygons are encoded so that each polygon is represented by $2D + 4$ real numbers $0 \leq X_i \leq 1$, and the first $2D$ values correspond to the positions of the points of the polygon. The last four values correspond the RGBA (Red, Green, Blue and Alpha). Given the encoding $\{0, 0, 0, 1, 1, 0, 1, 0, 0, 0.5\}$ and an input image I with the width of w_I and height h_I the encoded polygon will render to a fully red triangle with 50% alpha formed with by points $(0, 0)$, $(0.h_I)$, $(w_I, 0)$. The problem consists in minimising the function

$$f_7^{N,D,I}(X) = \sum_{i=1}^{w_I}\sum_{j=1}^{h_I}\big(I(i,j) - I^{'}(i,j)\big)^2$$

where

1. Large number of variables: $[1000, 14000]$
2. Search domain: $0 \leq X_i < 1$, $i = 1, 2, ..., N.(2D + 4)$
3. Global minimum: $X^* = (\pi, \pi)$, $f(X^*) = -1$
4. Unknown global minimum
5. Unknown quantity of local minima

14.6.8 Minimisation problem $g01$

The problem consists in minimising the function

$$f(X) = 5\sum_{i=1}^{4} X_i - 5\sum_{i=1}^{4} X_i^2 - \sum_{i=5}^{13} X_i$$

subject to the constraints

$$\begin{aligned}
g_1(X) &= 2X_1 + 2X_2 + X_{10} + X_{11} - 10 \leq 0 \\
g_2(X) &= 2X_1 + 2X_3 + X_{10} + X_{12} - 10 \leq 0 \\
g_3(X) &= 2X_2 + 2X_3 + X_{11} + X_{12} - 10 \leq 0 \\
g_4(X) &= -8X_1 + X_{10} \leq 0 \\
g_5(X) &= -8X_2 + X_{11} \leq 0 \\
g_6(X) &= -8X_3 + X_{12} \leq 0 \\
g_7(X) &= -2X_4 - X_5 + X_{10} \leq 0 \\
g_8(X) &= -2X_6 - X_7 + X_{11} \leq 0 \\
g_9(X) &= -2X_8 - X_9 + X_{12} \leq 0
\end{aligned}$$

where the bounds are $0 \leq X_i \leq 1$ for $i = 1, .., 9$, $0 \leq X_i \leq 100$ for $i = 10, 11, 12$, and $0 \leq X_{13} \leq 1$. The global optimum is located at $X^* = (1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 1)$ where $f(X^*) = -15$. Constraints g_i for $i = 1, .., 6$ are active.

14.6.9 Maximisation problem $g03$

The problem consists in maximising the function

$$f(X) = (\sqrt{N})^N \prod_{i=1}^{N} X_i$$

subject to the equality constraint

$$h(X) = \sum_{i=1}^{N} X_i^2 - 1 = 0$$

where $N = 10$ and $0 \leq X_i \leq 1$ for $i = 1, ..., N$. The global maximum is located at $X_i^* = \frac{1}{\sqrt{N}}$ for $i = 1, ..., N$ where $f(X^*) = 1$.

14.6.10 Maximisation problem $g08$

The problem consists in maximising the function

$$f(X) = \frac{\sin^3(2\pi X_1)\sin(2\pi X_2)}{X_1^3(X_1 + X_2)}$$

subject to

$$\begin{aligned}
g_1(X) &= X_1^2 - X_2 + 1 \leq 0 \\
g_2(X) &= 1 - X_1 + (X_2 - 4)^2 \leq 0
\end{aligned}$$

where $0 \leq X_1 \leq 10$ and $0 \leq X_2 \leq 10$. The global optimum is located at $X^* = (1.2279713, 4.2453733)$ where $f(X^*) = 0.095825$.

14.6.11 Minimisation problem $g11$

The problem consists in minimising the function

$$f(X) = X_1^2 + (X_2 - 1)^2$$

subject to the equality constraint

$$h(X) = X_2 - X_1^2 = 0$$

where $-1 \leq X_1 \leq 1$ and $-1 \leq X_2 \leq 1$. The global optimum is located at $X^* = (\pm\frac{1}{\sqrt{2}}, \frac{1}{2})$ where $f(X^*) = 0.75$.

14.6.12 Minimisation of the weight of a tension/compression spring

The problem consists in minimising the weight of a tension/compression spring subject to the constraints on minimum deflection, shear stress, surge frequency, limits on outside diameter and on design variables (see Arora [1989]). The design variables are the mean coil diameter D (X_2), the wire diameter d (X_1) and the number of active coils N (X_3). Formally, the problem is to minimise

$$(N+2)Dd^2$$

subject to

$$\begin{aligned} g_1(X) &= 1 - \frac{D^3 N}{71785 d^4} \leq 0 \\ g_2(X) &= \frac{4D^2 - dD}{12566(Dd^3 - d^4)} + \frac{1}{5108 d^2} - 1 \leq 0 \\ g_3(X) &= \frac{D+d}{1.5} - 1 \leq 0 \end{aligned}$$

where $0.05 \leq X_1 \leq 2$, $0.25 \leq X_2 \leq 1.3$, and $2 \leq X_3 \leq 15$.

Chapter 15

Introduction to CUDA Programming in Finance

This chapter has been written by Sebastien Gurrieri and we thank him for his time and effort.

15.1 Introduction

15.1.1 A birief overview

Parallel programming on Graphics Processing Units (GPUs) in Finance has gone from a curiosity in 2007 at the first release of CUDA by NVIDIA to a natural solution at the time of writing (2014). CUDA is now used in production environments for many applications in the Finance industry, including pricing of exotics or of large portfolios of vanillas, optimization and calibration problems, or risk calculations such as Value-at-Risk (VaR) or Potential Exposure (PFE) and Credit Valuation Adjustment (CVA).

On the consumer side, secrecy and competition lead to scarcity of details on how the technology is used. Bloomberg (see WST [2009]) and BNP Paribas [2009] were early adopters. JP Morgan reports deployment of GPU applications globally with large gains for risk calculations (see NVIDIA [2014a]). According to Davidson [2013], ING, Barclays, Société Générale, are known to use GPUs.

Information is more easily available on the software industry side, where companies readily report the introduction of GPU programming in their capabilities. A large number of solutions exist and we can unfortunately not cite all of them here - for a recent overview (see NVIDIA [2014a]). Several types of software are now available to banks and investment companies planning to use GPUs. Proprietary financial solutions such as Murex, Sungard offer fully integrated pricing and risk engines. Numerical libraries such as NAG can be used by the company's developers to link to their internal code, and applications such as MATLAB or SciFinance provide higher-level languages that benefit from GPU speed-ups without requiring to write explicit GPU code. A last type of solutions are "translators" such as Xcelerit, that wrap around the single-thread code of the developer (for instance C++) and translate it into CUDA.

When using a full proprietary software, numerical software to write high-level language or translators/wrappers around single-thread code, the company gives up some amount of control on the code. It may have no knowledge of the code at all, or have partial knowledge as in the case of the wrappers, but will not have full understanding of how the GPU code is translated from the higher-level language they have used. The company may not want to forgo this control, for instance because it may consider it can reach larger speed gains by having its developers have full control

over the syntax in order to reach the optimum speed. In such a case, quants with a knowledge of parallel programming and in particular of CUDA, will be required in order to write the translation by themselves in CUDA.

This document is meant to be a brief introduction for these quants who are interested in the subject of GPU programming, may want to use it for their own work by directly coding on the GPU, but have no or little experience of parallel programming. The learning curve can be quite steep, and we believe that there are few available resources to help this process: one can easily find very technical documents such as the CUDA manual (see NVIDIA [2014b]) or books that tend to assume a pre-existing familiarity with parallel programming and a background in computer science, in general without connection to Finance. The quant who wants to take the step and switch to CUDA without prior knowledge of parallel programming may then be left with few means of learning, most of which written in the language she will have difficulties understanding. The other type of learning resources is quant literature that is more easily understandable and directly related to the subject, but is unfortunately particularly scarce on the subject. In this article we will attempt to bridge this gap by providing the first few steps that can then put the reader on the right track to continue the learning with the other resources mentioned above.

This document is organised in the following manner:

- we will continue this introduction by a few words on parallel programming in general, GPU programming, CUDA, and why they can be used advantageously in Finance.
- we will then go into more details of the CUDA programming paradigm and syntax in section 2, keeping only what is necessary to start and gain a broad understanding of the main concepts.
- in section 3 we will describe two case studies: a Monte-Carlo simulation for exotic pricing with gains in the $\sim 100x$, and a multi-dimensional optimisation for implied volatility calibration using the Differential Evolution algorithm with gains of about $\sim 14x$ on cheap retail devices.
- we will then conclude with our views of future directions in this field.

15.1.2 Preliminary words on parallel programming

Parallel programming relies on the idea that independent calculations can be performed simultaneously on computing engines that have several "nodes" with the ability to execute actions concurrently. In an ideal situation, N nodes would lead to a speed-up factor of Nx, that is, the parallel calculation will be N times faster than if implemented on a single node. This is only an upper bound. In practice the speed gain will be lower than this due to several reasons among which:

- the algorithm may only be partially parallel. Most algorithms contain a sequential part that will not benefit from parallel hardware.
- parallel calculation speeds are often limited by data transfers to/from/within the nodes and read/write accesses to memory, which tend to be slow compared to arithmetic calculations on the nodes.
- the individual nodes on the parallel hardware may not be as powerful as the single node of the sequential calculation engine used for comparison.

This is why programming in parallel will usually involve the following actions:

- maximise the amount of parallel versus sequential tasks possibly by re-writing the algorithm.
- maximise the amount of active nodes (keep the hardware busy).
- minimise data transfers relatively to algebraic operations

15.1.3 Why GPUs?

Parallel hardware has existed since the 1950s, but remained for a long time only affordable to large institutions and research centers, constituting the so-called "super-computers". For several years now, multi-core CPUs have been available in retail computers at cheap starting prices. At the moment of writing (2014), typical retail computers contain 4 to 8 cores. Note that "cores" on CPUs and GPUs are usually able to calculate several streams of the parallel algorithm, also called "threads". Thanks to hyper-threading, a CPU core may run 2 threads in parallel, such that the upper bound on speed gain with an Intel i7 CPU (8 cores) using hyper-threading would be $16x$.

The comparison with GPUs is clear. A typical retail GPU (as of 2014) such as the NVIDIA GTX 750Ti contains 640 cores for a cost of about 150 \$. As GPU cores can run 32 threads simultaneously (a "warp"), this yields a maximum number of $20,480$ concurrent threads, several orders of magnitudes above what is available on CPUs at comparable levels of technology/cost.

The advantage of GPUs is not only in the present raw power, but also in their evolution through time. While the first GPU commercialised in 1999 by NVIDIA had 4 cores, 15 years later in 2014 a mid-level GPU will have hundreds of cores, while professional ones have thousands, not counting the fact that several GPUs can be added on the same machine. In the mean time, CPUs have made slower progress To illustrate this point, it is common to observe the evolution of the number of floating-point operations executable per second (GFLOP/s), a rough estimate of the speed, displayed in Figure (15.1).

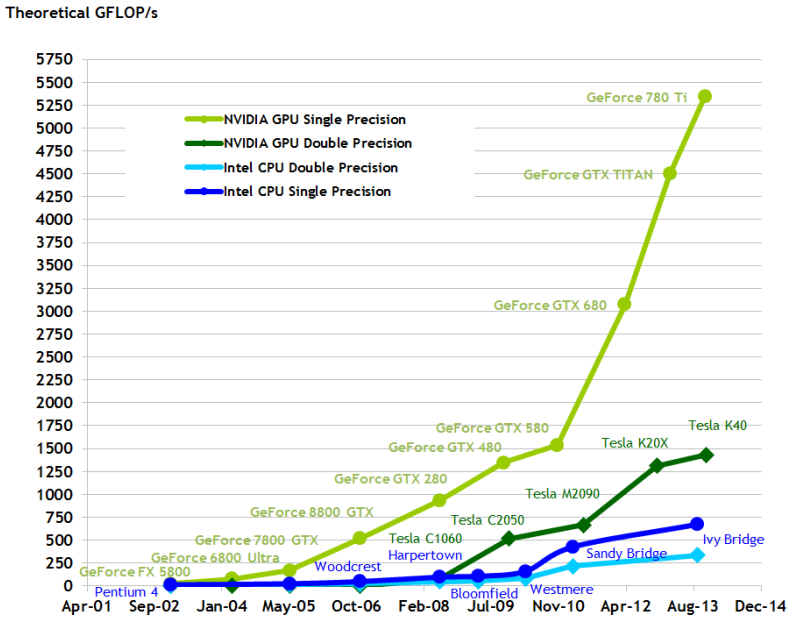


Figure 15.1: Comparison of theoretical GFLOP/s (extract from the CUDA programming guide, v6.5)

One can clearly see that GPUs have increased their computing power at a much larger pace than the CPUs over the last few years, especially for single precision. This means that, provided programming on GPUs is scalable, code written in the present will keep benefiting from significant speed gains in the coming years, which may not be so easily achievable on CPUs.

It should be kept in mind that, both on CPUs and GPUs, the maximum number of threads are usually poor indications of the actual speed-ups, in particular due to the limitations mentioned at the beginning of this section.

Furthermore, although the GPUs have a massive number of cores, these cores are usually less powerful than those of CPUs. Nevertheless, as we will show in this document, GPUs can bring speed gains up to $100x$ and more to financial applications, and can be considered as "game-changers".

15.1.4 Why CUDA?

This is all well but will not be very useful if there is no possibility to program the GPUs to perform the numerical calculations required by financial applications. This is where CUDA comes to help.

GPUs were originally used for video processing in video games and were optimised for image processing tasks, which were the main possible operations. Courageous pioneers realised that such operations are essentially multiplications and additions and that they could attempt to translate their code in video language in order to let the GPU calculate, thinking it was working on an image while it was actually calculating, say, a numerical integral.

CUDA is a C-based programming language introduced by NVIDIA in 2006 in order to shortcut this translation process. With CUDA, there is no longer any need to rewrite scientific algorithms in a video language. The CUDA compiler does it for the programmer. CUDA has additional features that make it very suitable for use in the financial industry. We will come back to these in more details later, but let us briefly mention them here:

- scalability: no need to rewrite the code when updating the hardware, code written on one device will benefit from speed improvements when run on the next generation of devices, without need for re-coding/re-building.
- compatibility with C: essentially a set of additional C-like commands, CUDA's syntax can easily be learnt by C programmers.
- free.
- integrates well in Microsoft Visual Studio.
- it has good debugging and profiling tools, in particular through the use of Nsight.

15.1.5 Applications in financial computing

Although in principle any set of independent operations can be accelerated by the use of GPUs and CUDA, in practice the application of these techniques to Finance tend to be in either of the three following groups:

1. Pricing: numerical pricing of individual complex products such as exotics can greatly benefit from GPU programming, especially when using the Monte-Carlo method that is "embarrassingly parallel", but also, to a lesser extent, with PDE methods. Significant gains can also be achieved when pricing very large portfolios of more simple products. This has been well illustrated in a number of references in the quant literature
   - Asian options in Black-Scholes, Monte-Carlo, $\sim 150x$.
   - Cancellable swaps in the BGM model, Monte-Carlo, $\sim 100x$.
   - PRDCs in Hybrid Local Volatility, Monte-Carlo, $\sim 100x$.
   - Structured equity options, Stochastic and Local Volatility, Monte-Carlo, $\sim 100x$.
   - Rainbow/basket options, Black-Scholes, PDE, $\sim 15x$.
   - PRDCs in Hybrid CEV, PDE, $\sim 40x$.
   - Options, Stochastic Volatility, PDE, $\sim 50x$.

2. Market/Credit Risk: significant gains can be obtained in the calculation of the Value-at-Risk and Credit/Counterparty risk measures such as the PFE and CVA. These methods tend to be based on Monte-Carlo simulations to generate future scenarios of the risk factors and repricing, which can be very efficiently implemented on GPUs. The gains depend a lot on the hardware and the particular algorithm, but speed-up of orders of $\sim 100x$ are consistently reported on GPUs (see Analytics Engines [2014], Dixon et al. [2009]). JP Morgan reported risk calculation speed-up leading to minutes rather than hours of calculations (see NVIDIA [2014a]), while HSBC experimented with intra-day CVA rather than overnight (see Woodie [2013]).

3. Calibration: much less represented than the other two applications above, and possibly included in the pricing, we believe that GPUs can significantly improve the calibration methods for complex models. For instance, it has been illustrated by Bernemann et al. [2011] for the piecewise time-dependent Heston model, and by Gurrieri [2012b] for the Hybrid Local Volatility (Dupire) model. Speed gains can be obtained from several sources, from the nature of the pricing formula (numerical integral, Monte-Carlo) to the fact that the models must usually be calibrated to a possibly large number of market instruments. Moreover, one may even consider parallelising the optimisation algorithm itself. We will describe one example of such calibration in this document.

15.2 Programming with CUDA

This section is largely inspired from the CUDA programming guide by NVIDIA [2014b]. We refer the reader to it for more details. Our goal here is to give more background and suitable application for quants in Finance. We introduce what we see as the most important points that can put the reader on the right track to start working, and leave aside deeper technical details that can be learnt later.

15.2.1 Hardware

For the purpose of this document, it is sufficient to understand a GPU as a collection of cores that have the ability to run simultaneous executions of operations. An indicator for the possible speed-ups of a GPU is this number of cores, but this can only give a rough estimate as numerous technical details have an impact on the effective number of simultaneous operations.

In reality, each core can execute a limited number of threads by construction of the hardware, but this number is also limited by the amount of memory required by the algorithm on each thread, with subtleties due to the thread hierarchy that we will discuss in the next section. As a result, it is quite difficult to predict the effective number of concurrent threads, but the number of cores on the device is certainly a decisive factor.

Another important hardware parameter is the amount of memory on the device as this can impact the choice of algorithm and/or the runtime configuration. Financial engineering algorithms may require storage of large amounts of data, for instance random numbers in Monte-Carlo simulations, and it is possible in realistic applications that the algorithm may ask for more memory than the device can provide (typically of the order of the GB). Joshi [2014] showed a good example of how to think about the algorithm in relation to the memory capacity of the GPU.

15.2.2 Thread hierarchy

Although there are certainly limitations to the possible algorithm configurations due to the hardware capabilities, the CUDA programming model is constructed in order to achieve scalability. One of the main goals is for parallel codes to run independently of the hardware. This is in opposition to past programming frameworks where knowledge of the hardware was often a requirement (for instance, knowing the number of cores), and this dependence was hard-coded in the algorithm.

This good level of independence is achieved in CUDA through a layer of abstraction, the "thread hierarchy". The parallel algorithm is written on a tree-like abstract structure made of a grid which contains blocks which contain threads, where the threads are running the parallel instructions. This is an abstract construction in the sense that the number of these structures is not tied to the hardware, and the user has no real control of how many blocks on the grid and threads within a block are concurrently running. Which threads and blocks are launched by the hardware is decided by the scheduler based on various considerations, an important one being available memory. Provided the programmer writes her code on this abstract structure, the same code will run on any CUDA-enabled GPU [1]. The thread hierarchy is illustrated in Figure (15.2). As in the C language, vector items are numbered from 0. In this example we have a 2-dimensional [2] grid of 6 blocks, and each block is 2-dimensional with 12 threads.

It is crucial for the programmer to keep in mind that she does not know how many of these blocks and threads are running simultaneously and in which order. For instance, one cannot be certain that $Block(0,1)$ will be launched after $Block(0,0)$, so the algorithm should not rely on such assumptions. In the same fashion, the programmer cannot know if $Thread(0,1)$ within $Block(1,1)$ will be run before or after $Thread(2,1)$ (or simultaneously), so again, such assumptions should never be made when coding an algorithm. How many block/threads are run concurrently and in which order is hardware/application/time dependent, and it is by attributing this control to the scheduler rather than the user that CUDA achieves good scalability. Contrary to what one may think, not having this control is a good thing.

One of the reasons for the existence of the concept of Blocks is for a more efficient use of memory. As we will see in the next section, there are several types of memories on the GPU, with different access patterns. It is often desirable for a subset of threads to share common information (such as input data) as it would be a waste of resources to have many threads perform the loading of the same data. Threads within blocks can share some parts of the memory. They also have the ability to wait so that each other has finished its tasks. Although this appears to be breaking the parallelism, this functionality can be useful and is common in parallel programming languages. As it obviously slows down the algorithm, it should be used only when strictly necessary.

15.2.3 Memory management

Several types of memory co-exist on the GPU, as is illustrated in Figure (15.3).

For the introductory purpose of this document, it is sufficient to mention three of them here:

- Global memory: similar to the RAM memory for a CPU. It is not built right in the cores ("off-chip"), and accesses to it are slow, but it comes in large quantities (2GB for the GTX 750Ti). It is accessible by all threads. It is the one used most of the time for communication between the CPU and the GPU. The CPU typically loads the inputs in it, and retrieves the outputs from it.
- Registers: local to each thread, and "on-chip", lying inside the core that calculates the thread. Their accesses are very fast but they come in small quantities (16KB per core). Since they are private to the thread, they tend to be used to define temporary local variables such as counters, accumulations for sums, etc...
- Shared memory: on-chip and very fast, it can be accessed by all threads within a block, but is limited to 48KB per core.

For optimum performance, the algorithm should be written so as to use the on-chip memory as much as possible. However, some interaction will appear with the number of simultaneous threads running on the device. Indeed, the cores will launch blocks and threads only so long as they do not use more than the available amount of memory in the core. Writing a code that uses more of the on-chip memory will result in less blocks/threads running simultaneously. This is a typical trade-off faced by the CUDA programmer. Profiling tools such as Nsight facilitate the task of optimising memory usage.

[1] Up to rather technical intricacies that are beyond the scope of this document.
[2] Grids can be 1D or 2D, blocks can be 1D, 2D or 3D.

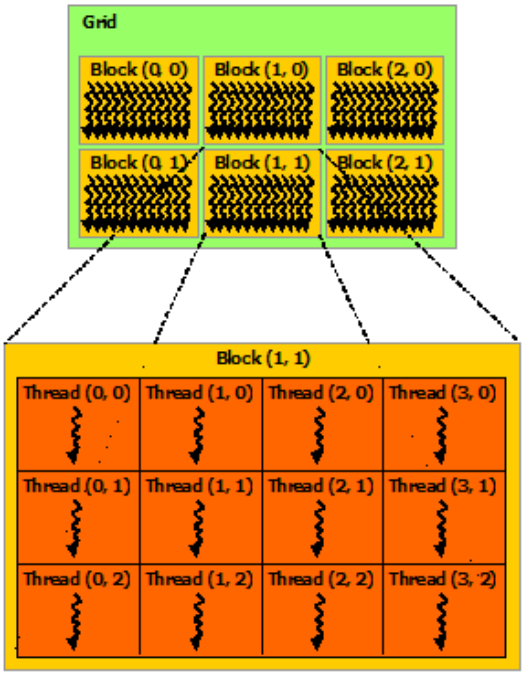


Figure 15.2: Thread hierarchy (extract from CUDA programming guide, v6.5)

15.2.4 Syntax and connetion to C/C++

The CUDA programming model is heterogeneous. Only some portions of code run on the GPU (also called "device"), while others will keep running on the CPU as before. It is up to the programmer to decide what is best left on the CPU or migrated to the GPU, knowing that not all algorithms can benefit from GPU acceleration [3].

The sequential part of the code is typically written in C/C++. This part is executed by the CPU (also called "host") and could be the reading of input information as well as pre-processing algorithms that are sequential in nature or would not benefit from acceleration on the GPU for other reasons (such as small numbers of operations). The input information is then sent to the GPU that executes the parallel algorithm. The outputs (usually in the global memory) are retrieved back on the CPU which can process them, then launch a new GPU algorithm, and so on and so forth, as illustrated in Figure (15.4).

There exist three types of functions in the C/CUDA language:

- Host functions: identical to standard C/C++ in that they are launched by the host and run on the host.
- Hybrid host-device functions: also called "kernels", launched by the host, run on the device. They are identified by the compiler thanks to the keyword

  `__global__`

- Device functions: launched by the device, run on the device. They are identified using the keyword

[3] Since each node of a GPU is slower than that of a CPU, it is even conceivable that a very sequential algorithm may be slower on the GPU than on the CPU.

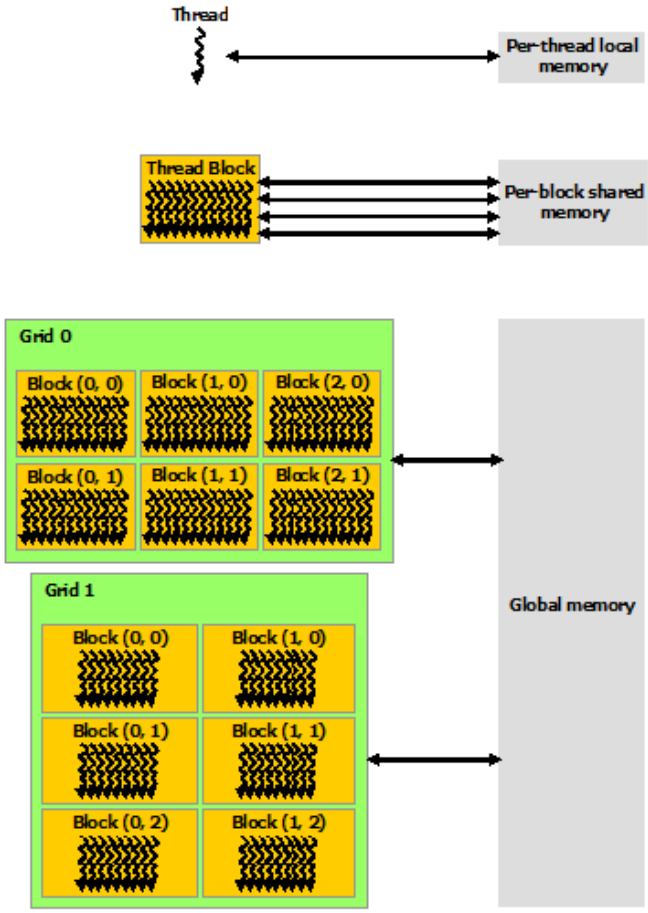


Figure 15.3: Memory types (extract from CUDA programming guide, v6.5)

```
__device__
```

We will be focusing on the kernels as these are the ones that involve most new concepts and syntax. Device functions can be used as helpers but are not mandatory, while kernels are crucial as the place where the CPU launches the GPU code.

A kernel is launched with the syntax in Figure (15.5). This piece of code is written inside a standard C $main()$ function. "dim3" is a new variable type that describes the grid and block configurations. The variable "threadsPerBlock" here is set so that each block is 2-dimensional, with 16 threads $\times$ 16 threads. The variable "numBlocks" contains the grid configuration. This grid is also 2-dimensional, with each dimension size determined as some integer N divided by the size of the block in that dimension. The sizes of the thread hierarchy are built-in member properties such as "threadsPerBlock.x" which represents the x-axis dimension of the variable threadsPerBlock.

The kernel itself is called "MatAdd" and is launched by the CPU with the thread hierarchy configuration within the signs $<<<$, $>>>$ as above, and with its arguments A, B, C as in any standard C function. The code for this kernel will be executed on the GPU and may look something like Figure (15.6) (schematically). This kernel performs a matrix addition. A and B are the inputs (loaded to the global memory of the device before launch) and C is the output (whose memory has been allocated before the kernel launch). This code describes what a single thread does. This is

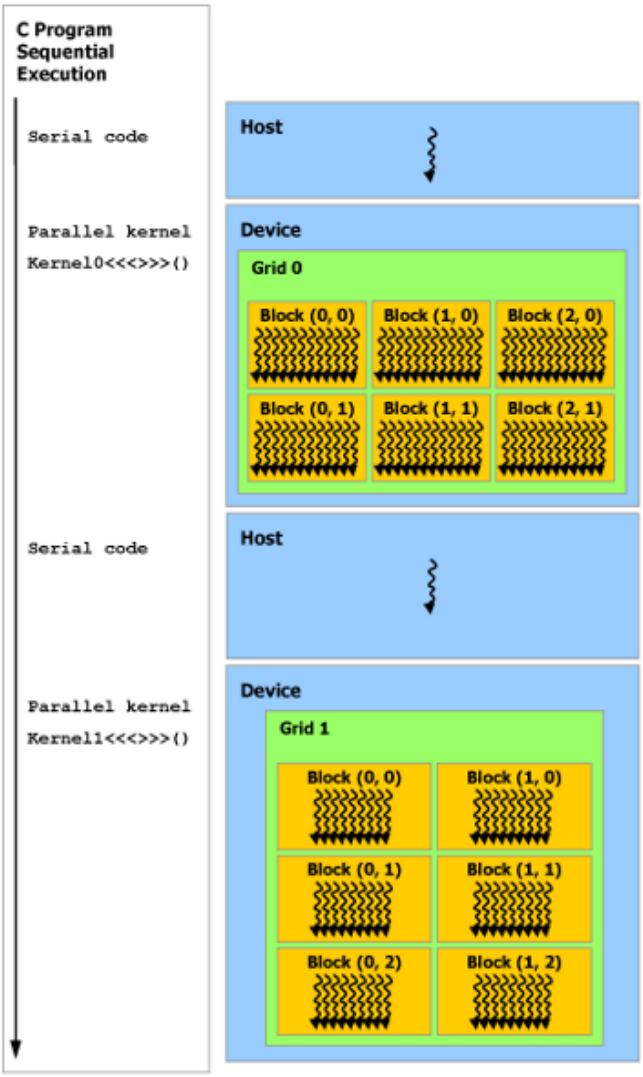


Figure 15.4: Heterogeneous programming (extract from CUDA programming guide, v6.5)

why, although the matrix is $N \times N$, there is no for loop over the rows and columns of C as one would have expected in a sequential code. Each thread represents one future entry of the matrix C and the role of the for loop is taken by the scheduler that launches the blocks on the cores, and the cores that launch the threads. As all these calculations are fully independent, there is no need to know in which order the blocks and threads are launched. A, B, C are in global memory, i and j are placed in registers (only usable by this thread).

What thread is this code running on can be identified thanks to a number of built-in variables:

- blockIdx represents the coordinates on the grid (x and y) of the block on which the thread resides.
- blockDim represents the dimensions of the blocks (that is, how many threads they have).
- threadIdx represents the coordinates on the block (x and y) of the thread. In particular, these are relative to the block (that is, they start at 0 on all blocks), not relative to the grid.

Memory allocation and transfers are performed by the CUDA functions $cudaMalloc()$ and $cudaMemcpy()$ called by the CPU. A sample syntax for the allocation and transfers before the kernel launch would look like Figure (15.7) (for pointers). The meaning of the lines in Figure (15.7) is the following:

- allocate memory to hold the host data (h_A, h_B). The filling in of data is not shown, this step is standard C language.
- allocate memory on the device for storage of the inputs (d_A, d_B) and future writing of the outputs (d_C). This is done with the CUDA function $cudaMalloc()$.
- transfer the data from the host (h_A, h_B) to the device (d_A, d_B), with the CUDA function $cudaMemcpy()$.

Once the kernel finished its tasks, the result of the matrix operation is held in the device variable d_C, which must be transferred to the CPU for further processing (such as display, or use in other algorithms). This step looks like Figure (15.8), where one recognises the CUDA function $cudaMemcpy()$ used in reverse direction (from device to host) and the CUDA de-allocation operator (equivalent of delete in C++) for the device (global) memory. At this point, the results of the algorithm are in the CPU pointer h_C and we are in a position to continue executing code on the CPU, or prepare for launching a new kernel.

```
int main()
{
    ...
    // Kernel invocation
    dim3 threadsPerBlock(16, 16);
    dim3 numBlocks(N / threadsPerBlock.x, N / threadsPerBlock.y);
    MatAdd<<<numBlocks, threadsPerBlock>>>(A, B, C);
    ...
}
```

Figure 15.5: Kernel call from the CPU (extract from CUDA programming guide, v6.5)

```
// Kernel definition
__global__ void MatAdd(float A[N][N], float B[N][N],
float C[N][N])
{
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
    if (i < N && j < N)
        C[i][j] = A[i][j] + B[i][j];
}
```

Figure 15.6: Example of Kernel (extract from CUDA programming guide, v6.5)

```
// Allocate input vectors h_A and h_B in host memory
float* h_A = (float*)malloc(size);
float* h_B = (float*)malloc(size);

// Initialize input vectors
...

// Allocate vectors in device memory
float* d_A;
cudaMalloc(&d_A, size);
float* d_B;
cudaMalloc(&d_B, size);
float* d_C;
cudaMalloc(&d_C, size);

// Copy vectors from host memory to device memory
cudaMemcpy(d_A, h_A, size, cudaMemcpyHostToDevice);
cudaMemcpy(d_B, h_B, size, cudaMemcpyHostToDevice);
```

Figure 15.7: Memory allocation and transfer to device (extract from CUDA programming guide, v6.5)

```
// Copy result from device memory to host memory
// h_C contains the result in host memory
cudaMemcpy(h_C, d_C, size, cudaMemcpyDeviceToHost);

// Free device memory
cudaFree(d_A);
cudaFree(d_B);
cudaFree(d_C);

// Free host memory
```

Figure 15.8: Copy to host and de-allocation (extract from CUDA programming guide, v6.5)

15.2.5 Random number generation

Before presenting our case studies, we would like to spend some time discussing the random number generation (RNG) process, which is the key to many financial applications that benefit from GPU acceleration.

Let us imagine a very simple Monte-Carlo simulation where we would calculate the expectation of N normally distributed random draws. Without entering into subtleties of thread hierarchy and speed optimisation at this point, let us say that we decide to launch N threads numbered from 0 to $N-1$.

Most RNG algorithms are highly sequential in nature. The i-th draw is usually derived from the $(i-1)$-th draw by some recurrence formula. If we want to run such an algorithm on our set of N threads, after initialising each thread to the first item in the random sequence, the thread i will then have to calculate all the draws in the sequence up to the i-th one before being able to proceed (that is, transform the uniform number into a Gaussian one). One can easily see that in principle there would be no significant speed gain from this, as the runtime of the simulation would be that of the single thread $(N-1)$ [4].

Research efforts have been spent on finding efficient "skip functions" that let us calculate a random draw farther in the sequence without having to calculate all the intermediate draws. Luckily for us, the Sobol sequence that we used in most of our Monte-Carlo simulations does have a particularly efficient one, based on the algorithm by Bromley [1995], provided that we skip 2^n draws for some integer n. This property is implemented in the sample code "SobolQRNG" in the CUDA SDK.

The skip functions are typically used in the following manner:

- choose an integer "stride" s, that is, the number of draws that are skipped each time a threads needs a random number (this should be a power of 2 for Sobol).
- choose the number of simulations N.
- thread i calculates draws $i, i+s, i+2s, ...$ by using the skip function, and stops before going above N.

The threads are then all independent and can be calculated in any order. In particular, not all of them need to be launched on the grid at the same time, which gives a lot of flexibility to configure the grid in an optimal way.

We now consider two strategies to design RNGs on the GPU, using memory storage or inline calculation.

[4] In more realistic applications one may still improve the speed because what happens after the generation of the uniform numbers will probably be truly parallel, and will often be the bottleneck of the simulation.

15.2.5.1 Memory storage

This strategy assumes that the random numbers have been generated before the launch of the kernel and are stored in the global memory. They could be read from text files, calculated by the CPU and sent to the GPU memory, or calculated on the GPU in a previous kernel.

From the point of view of the simulation kernel, they are inputs to be read from the global memory and to be dispatched to the threads. This has the advantage of being more elegant as a design in that the RNG code is separated from the Monte-Carlo code. Additionally, more types of RNGs may be implemented on the CPU that would then send the results to the GPU. But it has the disadvantage of requiring large numbers of global memory accesses, and this tends to be sub-optimal for speed improvement.

Furthermore, straight forward memory requirement estimates show that there may not be enough memory on the GPU to hold all the random numbers. Indeed, assume that we want to run $100,000$ simulations to calculate basket options with 10 underlyings and with 500 time steps to account for skew. The random numbers are typically float type so require $4B$ of memory (in single precision). The total memory required for storing them would be

$$100,000 \times 10 \times 500 \times 4 = 2GB$$

On the GTX 750Ti that we use as an example of mid-level device, this is already the limit of the available global memory.

15.2.5.2 Inline

The other strategy is to let the GPU Monte-Carlo engine calculate the random numbers at the point where it needs them. This eliminates the need for so many memory accesses, which will tend to be faster while not suffering from the potential lack of memory on the device. However, the design loses in elegance as the RNG algorithm sits in the middle of the simulation and cannot be decoupled from it.

15.3 Case studies

15.3.1 Exotic swaps in Monte-Carlo

15.3.1.1 Product and model

In this CUDA application, we consider Power Reverse Dual Currency swaps (PRDCs) with maturities up to $30Y$. One leg of the swap is essentially a series of call options on an FX rate (say USD/JPY), exchanged against floating payments of the domestic rate plus spread (say JPY Libor). We consider a Target Redemption exercise (TARN) which makes the product path-dependent. The swap is cancelled when the accumulated coupon on the structured leg goes above a pre-defined limit. As a model we choose the Dupire local volatility for FX and Hull-White for both the domestic and foreign interest rates (IR).

The long-dated nature of this product and its sensitivity to the FX skew imply that very complex hybrid models should be used, while the exotic exercise requires a numerical method. Although PDEs may be used with an additional intermediate variable to handle path-dependency (see Christara et al. ChristaraEtAl09), the most common method is Monte-Carlo, and this serves as a good example for our purpose of acceleration with CUDA.

The work described here has been presented by Gurrieri [2012a] for two factors (deterministic foreign rate) and extended to three factors by Gurrieri [2012b]. The latter reference contains a detailed account of the model definition and its calibration. Here we only recall the main characteristics of the model. It has three stochastic factors in total, one for FX and one for each IR, and the FX volatility is local in the sense that it is represented by a function of time t

and the FX underlying X_t, that is, $\sigma(t, X_t)$. The stochastic differential equation (SDE) of the FX underlying is given by

$$\frac{dX_t}{X_t} = (r_d - r_f)dt + \sigma(t, X_t)dW_X(t)$$

while the domestic short-rate has the typical Hull-White dynamics

$$dr_d = \big(\theta_d - a_d r_d\big)dt + \sigma_d dW_d(t)$$

with mean-reversion a_d, volatility σ_d, and deterministic curve shift θ_d. The foreign short-rate also follows Hull-White dynamics in the foreign measure, but is modified by a quanto term

$$dr_f = \big(\theta_f - a_f r_f - \sigma_f \rho_{fX} \sigma(t, X_t)\big)dt + \sigma_f dW_f(t)$$

The quanto term brings the foreign SDE to the domestic risk-neutral measure, and it is therefore expressed as a covariance between the foreign rate and the FX underlying. The three Brownian motions $dW_X(t)$, $dW_d(t)$, and $dW_f(t)$ are assumed correlated. See Gurrieri [2012b] for a full description of the model. In practice, the local volatility function is sampled on a set of discrete times t_n and FX values X_i, which results in the matrix

$$\sigma_{ni} = \sigma(t_n, X_i)$$

and is interpolated on the paths. Semi-analytical calibration methods exist for this model and were described by Bloch et al. [2008], while exact calibration through Monte-Carlo methods was proposed by Gurrieri [2012b] with implementation in CUDA and sample code.

15.3.1.2 Single-thread algorithm

Before looking at the algorithm in parallel, it helps to take a brief look at the single thread version, focusing on the most important part on the FX. On each Monte-Carlo path j, at each time t_n, we do the following:

1. Calculate next uniform random number
2. Transform to Gaussian, then Brownian motion increment $dW_j(n)$
3. Read previous spot $X_j(n)$ from memory
4. Calculate volatility σ by interpolating the volatility slice $\sigma(t_n, X_j(n))$ at t_n
5. Calculate the new spot
$$X_j(n+1) = X_j(n) e^{(r_d - r_f - \frac{1}{2}\sigma^2)(t_{n+1} - t_n) + \sigma dW_j(n)}$$
6. Calculate product(s)
7. Write new spot in memory

And then we loop on the paths, and the times.

15.3.1.3 Multi-thread algorithm

The most natural idea at first may be to let each thread calculate one path. This may be possible, but for memory management and performance reasons, it is not the best idea. Indeed, a general rule of thumb for achieving good performance is to maximise the amount of arithmetic operations compared to data transfers. Following this rule, it is often advantageous to let each thread calculate many paths to avoid multiple readings of the same inputs. We described this process in Section (15.2.5).

Let us fix the thread hierarchy. We take a 1D grid of N_b blocks, and 1D blocks of N_t threads (in our runs, $N_b = 128 = N_t$). We define the stride as $s = N_(b) \times N_t$, which represents the total number of threads. We run a number of simulations $N_m c$. Following the all threads from $j = 1$ to $j = N_m c$ are calculated once and only once. This is achieved by letting Thread a calculate the paths $a, a+s, a+2s...$ until reaching $N_m c$. Thread a first calculates the path a, then skips over the paths $a+1, a+2, ...$ using the RNG's skip function as mentioned in Section (15.2.5), and the next path is calculates is $a+s$, then skipping $a+s+1$, etc..

Let us take a simple example with a 1D grid of two 1D blocks with contain two threads, and let us run 14 simulations. The stride is 4 and we number the blocks and threads globally from 1 to 2 and 1 to 4 for simplicity. The Monte-Carlo algorithm follows this pattern

- Block 1, Thread 1, calculates paths $1, 5, 9, 13$
- Block 1, Thread 2, calculates paths $2, 6, 10, 14$
- Block 2, Thread 3, calculates paths $3, 7, 11$
- Block 2, Thread 4, calculates paths $4, 8, 12$

On each path j that it must calculate, Thread a calculates PV_j and accumulates it in a sum variable T_a. For example, in the simple case above $T_1 = PV_1 + PV_5 + PV_9 + PV_{13}$. Then, in a given block, we wait for all threads to finish their sum and once this is done, we accumulate the sums of the threads in a given block. The sum in Block 1 above will be $B_1 = T_1 + T_2$, and the sum in Block 2 is $B_2 = T_3 + T_4$.

Let us explain now how these partial sums can be performed in connection with memory types. Each thread can access its own registers, and this memory is accessible only by this thread. In the example above, PV_1 is written in a register of Thread 1, PV_6 is written in a register of Thread 2, PV_3 is written in a register of Thread 3, and Thread 1 can access neither PV_6 nor PV_7, even though Thread 1 and Thread 2 are located in the same Block 1.

Thread 1 sums all its PVs in the variable T_1 and Thread 2 in T_2. If we let T_1 and T_2 be on registers, we will not be able to sum them into the "block variable" $B_1 = T_1 + T_2$, because no thread will be able to access both. We thus need T_1 and T_2 to be in a memory type that is accessible by the thread that will sum them.

There are two possibilities to achieve this. Either we use the global memory, accessible by all threads on the device, which is comparatively slow, or we use the shared memory, accessible by all threads within a block, in small quantities but very fast. For performance, we should always try to use the shared memory as much as possible. In this situation, there is more than enough of it so it will be our choice.

Thread 1 and Thread 2 will write their own sums T_1 and T_2 in the shared memory of their block, Block 1. Once this is done, we use one of these threads to sum $B_1 = T_1 + T_2$ and similarly for B_2 on Block 2.

The next question is where to store B_1 and B_2. Ultimately we want to sum them and we will not be able to do so if we store them in the shared memories of their respective blocks as these cannot communicate. Therefore, we store them on the global memory.

In each block i, the thread that is responsible for calculating the block-sum B_i will write it in the global memory. At this point, we no longer have many variables. Indeed, although there may be tens or hundreds of thousands of PV_j or T_a, in our pricing configuration there are only 128 blocks so 128 B_i. A sum of 128 variables is not very advantageously performed on the GPU so we send the B_i from the global memory to the CPU memory where we do the final sum. This ends the algorithm. The performance was reported by Gurrieri [2012a], with speed-ups above $100x$ on realistic pricing configurations and on a GTX 460, a model that in 2014 is at the low end of the range.

Let us make a brief remark on the issue of single versus double precisions. Single precision can be a limitation in Monte-Carlo simulations when running in single thread because the more we add paths, the more we add small numbers to large numbers and the loss due to round-offs becomes more and more important with the number of simulations. Errors due to single precision can become significant at realistic numbers of simulations, while double precision prevents this problem. On the other hand, single precision can be significantly faster on GPUs, as is well illustrated by Figure (15.1).

The parallel algorithm illustrated above does not suffer from this issue because the Monte-Carlo sums are always partial. Threads sum a few paths, then blocks sum a few of these "thread sums", and the final Monte-Carlo result is the sum of the "block sums". In essence, this is a nested bucketed sum, and this is typically a way of avoiding the single precision issue. Note that since double precision is available on modern GPU devices, the programmer can always choose to run with it, although at the cost of some performance since double precision is not as optimised on GPUs as it is on CPUs.

Finally, note that this model can also benefit from GPU optimisation for its calibration. Gurrieri [2012a] took advantage of the fact that its FX calibration equation can be written explicitly as an expectation to propose a Monte-Carlo calibration algorithm that allows for a more accurate calibration than semi-analytical approximation formulas. Furthermore, this algorithm can straight forwardly be extended to other short-rate models such as CIR. However, the half-sequential half-parallel nature of this process implies that the speed gains are less dramatic than for a purely Monte-Carlo process. Nevertheless, we could obtain gains of about $\sim 30x$ on retail devices. Gurrieri [2012a] provided a detailed account of this algorithm together with sample CUDA code.

15.3.1.4 Using the texture memory

Finally, let us add a few words on the interpolation of the local volatility matrix. A typical size for it may be 200×200, which means about $40,000$ entries. When pricing in Monte-Carlo, it must be interpolated, mostly because $X(t)$ is a stochastic quantity not known in advance (by opposition to a lattice where it is known on the nodes) and thus has no reason to be one of the matrix pillars. Note that the time grid on the other hand, the t_n are known in advance, such that interpolation in the time direction is not necessary.

GPUs were originally used for video games and in particular for handling 2D surfaces for which interpolations were required. The hardware was thus optimised for very fast interpolation of matrices. For this purpose, there exists a special type of memory called "texture memory" that has fast cache and a built-in linear interpolation at the hardware level. The extrapolation is constant, and this is precisely how we interpolate the local volatility matrix. It is therefore natural to think of storing the LV entries in the texture memory of the GPU. This was originally done by Bernemann et al. [2011], and we refined the idea by noticing that only interpolation in the FX direction is needed such that one can actually use a slightly different type of textures with more flexibility, the layered textures. See Gurrieri [2012a] for more details and performance tests of the texture memory.

15.3.2 Volatility calibration by differential evolution

The results presented in this section have not been published yet, and we give only the main steps of the reasoning with preliminary performance reports.

15.3.2.1 Model and difficulties

Here we consider an entirely different CUDA application, that of the calibration of an implied volatility surface to the market of vanilla option prices. Broadly speaking, the problem is the following: given a continuous parametric function $\theta(K,T)$ where K is the strike and T is the maturity, what are the parameters defining $\theta(K,T)$ that lead to the best fit to a series of implied volatility quotes that are available at discrete market pillars (K_a,T_a)?

This is typically a problem of multi-dimensional constrained optimisation, as in general the model will have several parameters with potentially complex constraints. In this study we use the model described by Bloch [2010]. It is a mixture of normal densities whose parameters are time-dependent functions. For the three family model, there is a total of 19 parameters to optimise on. A common market configuration for equity indexes would lead to a fit to about 100 option prices.

The optimisation of implied volatility surfaces is known to be a delicate issue as the objective functions and their constraints often exhibit local minima which poses problems for standard (deterministic) optimisation algorithms such as the Simplex or gradient methods. The global minimum may be missed, and a high dependence on the starting point can be observed. Bloch [2010] proposed the use of the Differential Evolution (DE) algorithm, a member of the family of Genetic Algorithms. We will not go into details on what makes the Differential Evolution different from the other Genetic Algorithms, as this is not the main purpose of this study. The main steps of the algorithm for us are common to all Evolution algorithms:

1. Generate a population of N_p candidate parameter sets (parent population).
2. Cross-Over the parent population to produce an offspring population (stochastic).
3. Apply Mutations to the offspring population (stochastic).
4. Select the fittest elements between offspring and parents to generate the next parent population.
5. Iterate (that is, go to the next generation).

The "fittest" candidates are those for which the $\theta(K,T)$ function is the closest to the market quotes for some metric (we choose the L_2 distance), possibly incorporating a penalty for non-satisfied constraints (we use the so called "death penalty", that is, infinite distance).

While this algorithm is excellent at handling constraints and avoiding local minima, it does require large amounts of calculations as the objective function must be evaluated once for each member of the population and at each iteration. To avoid local minima, the algorithm will typically need more iterations than a standard deterministic algorithm, which tends to result in a much slower runtime. Happily, this algorithm can be naturally written in parallel (although not as "embarrassingly" as a Monte-Carlo simulation), and this is the subject of the remaining of this section.

15.3.2.2 Single-thread algorithm

As for the previous application, before looking at the parallel algorithm we briefly describe the single thread algorithm. Assuming a population of N_p candidate parameter sets (we will call them "individuals", 19-dimensional vectors here) already exist in memory with the value of their objective function calculated, the algorithm proceeds as

1. Choose N_p parent sets to generate N_p children (random process).
2. Let mutations modify these N_p children (random process).
3. Calculate the objective function for each child (use this child's parameters and calculate the option price by combining Black-Scholes functions).

4. Select the fittest individuals among parents and children.

5. Iterate.

As often with such optimisations, the bottleneck of the algorithm tends to be step 3) above, that is, the estimation of the modelling function on all the instruments we fit to. In our applications, there will be about 100 option prices to calculate on each individual, and each option calculation requires the evaluation of a few simple algebraic functions together with the cumulative normal density. The size of the population can be chosen by the user, but a good rule of thumb seems to be to use about $5D$ where D is the dimension of the problem, here 19 (for 19 parameters in the model). This means that at each iteration, about $10,000$ options must be calculated, and a few hundreds of generations may be needed for sufficient convergence. This results in a very large number of calculations that we would like to accelerate using GPUs and CUDA.

15.3.2.3 Multi-thread algorithm

An example of parallelisation of DE has been studied by Ramirez-Chavez et al. [2011] which adopted the strategy of parallelising on the population: since the evaluation of the objective function for each individual is independent, this step can be done in parallel with each thread calculating the objective for one of the population members.

This is certainly valid and a very generic principle for this algorithm, but we propose to take advantage of an other source of parallelism to further speed-up the calculation: the option direction. Indeed, while calculating the objective function is independent for each individual, within one such evaluation of the function, the calculation of each option is also independent of the others. We thus have a 2-dimensional source of parallelism here, which can very naturally be taken advantage of using the CUDA thread hierarchy.

We propose two algorithms with some trade-offs. The first algorithm leads to a minimum amount of CUDA code writing but less speed improvement. The second one is fully written on the GPU and faster by a factor of about $2x$ compared to the more simple algorithm.

Simple algorithm We keep all the purely DE steps on the CPU and parallelise only the calculation of the objective functions on the population. We choose a grid configuration with N_p blocks, that is, 1 block per individual. Each thread within a block calculates one option based on the parameters for the individual attached to this block. When all threads have calculated their assigned options, we reduce the calculation of the distance between market and model options, which is the objective function. This process is similar to that of the partial "block-sums" of our Monte-Carlo algorithm described in the previous section. The step-by-step description is as follow:

- generate the child population on the CPU
- configure a grid with N_p blocks where N_p is the size of the population
- send this population to the GPU, that is, send the 19 parameters of each child to each block (in the shared memory)
- each thread within a block calculates one option with the child parameters for this block, and loops until all options are calculated
- the objective function is calculated on this block and written to the global memory
- the CPU retrieves the objective functions on the population from the global memory, and uses these values to generate the next population according to the rules of DE
- iterate

This algorithm gave us only a moderate $6x$ gains on a GTX 460 over the single thread implementation on the CPU. The advantage of this algorithm is that, as DE is actually implemented on the CPU, only a small amount of CUDA code writing is required. The main disadvantage is in the number of memory transfers between the host and the device: at every generation, a population must be sent from host to device, and the objective functions must be sent from device to host. Another disadvantage, less significant in our opinion, is that we do not benefit from other possible gains on the DE algorithm itself.

Improved algorithm We suggest a second algorithm that addresses these issues. This time, all the DE steps are implemented on the GPU in a similar fashion to Ramirez-Chavez et al. [2011]: cross-over, mutation, and selection. This second version has the same thread hierarchy, but requires intermediate kernels for the DE steps. Although it requires more CUDA code, it benefits from no longer requiring host-device-host transfers, and can achieve up to 14x speed-ups on our GTX 460.

15.4 Conclusion

As of end of 2014, GPU programming and CUDA are more and more viewed as viable technologies for massive improvements of various calculations in financial applications, from pricing complex exotics and large portfolios of vanillas to risk calculations such as VaR and CVAs. Major banks and investment organisations report successful deployment of GPU solutions, and the number of software companies proposing GPU libraries has grown substantially since the launch of CUDA in 2007. These propose full solutions with closed code (such as Murex), libraries to link to or higher level languages (such as NAG, MATLAB), or wrappers around single thread code (Xcelerit).

Companies that prefer keeping full control of their code up until the GPU execution, though, will still need to hire quants with a knowledge of CUDA programming. This may not be very easy as single thread C/C++ programming is a standard, and the learning curve to switch to CUDA can be rather steep.

In our experience, most of the difficulty in using CUDA comes from the lack of experience in "thinking in parallel", and this is not specific to CUDA. The next difficulty in line is the understanding of the thread hierarchy and memory management, and how to use it to improve performance and scalability. We hope that this document will help other quants go through these steps with more ease. The results are worth the effort, with very large gains especially in Monte-Carlo implementations.

The level of gains one can obtain from this technology is not simply improving the speed of current calculations, it also allows for new models or possibilities to be explored. Such "game changers" are intra-day risk, or Monte-Carlo calibrations of models without closed-forms, a process that was prohibitively expensive before, but that can now be considered realistically for production.

Appendices

Part VI

Appendices

Appendix A

Review of some mathematical facts

A.1 Some facts on convex and concave analysis

Details can be found on text books on convex analysis (see Rockafellar [1995]) as well as on text books on mathematics for economics (see Wilson [2012]). We use relint (S) to refer to the relative interior of a convex set S, which is the set S minus all of the points on the relative boundary. We use closure (S) to refer to the closure of S, the smallest closed set containing all of the limit points of S. We use ∇ to denote a differential operator that indicates taking gradient in vector calculus. In the Cartesian coordinate system $\mathbb{R}^n$ with coordinates $(x_1, .., x_n)$ and standard basis $(\hat{e}_1, .., \hat{e}_n)$, del is defined in terms of partial derivative operators as

$$\nabla = (\frac{\partial}{\partial x_1}, .., \frac{\partial}{\partial x_n}) = \sum_{i=1}^{n} \hat{e}_i \frac{\partial}{\partial x_i}$$

The gradient product rule is given by

$$\nabla(fg) = f\nabla g + g\nabla f$$

and the rules for dot products satisfy

$$\nabla(u.v) = (u.\nabla)v + (v.\nabla)u + u \times (\nabla \times v) + v \times (\nabla \times u)$$

where u and v are vectors. The directional derivative of a scalar field $f(x, y, z)$ in the direction $a(x, y, z) = a_x\hat{x} + a_y\hat{y} + a_z\hat{z}$ is defined as

$$a.\ \text{grad}\ f = a_x \frac{\partial f}{\partial x} + a_y \frac{\partial f}{\partial y} + a_z \frac{\partial f}{\partial z} = (a.\nabla)f$$

which gives the change of a field f in the direction of a. The Laplace operator is a scalar operator that can be applied to either vector or scalar fields; for cartesian coordinate systems it is defined as

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = \nabla.\nabla = \nabla^2$$

The tensor derivative of a vector field v can be denoted simply as $\nabla \otimes v$ where $\otimes$ represents the dyadic product. This quantity is equivalent to the transpose of the Jacobian matrix of the vector field with respect to space.

A.1.1 Convex functions

If $x, y \in \mathbb{R}$ and $\alpha \in (0,1)$, then $(1-\alpha)x + \alpha y$ is a convex combination of x and y. Geometrically, a convex combination of x and y is a point somewhere between x and y. A set $X \subset \mathbb{R}$ is convex if $x, y \in X$ implies $(1-\alpha)x + \alpha y \in X$ for all $\alpha \in [0,1]$.

Definition A.1.1 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if any of the following conditions hold*

1. *for all $x, y \in \mathbb{R}^n$,*
$$\frac{1}{2}f(x) + \frac{1}{2}f(y) \geq f(\frac{x+y}{2})$$
2. *for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$,*
$$(1-\alpha)f(x) + \alpha f(y) \geq f((1-\alpha)x + \alpha y)$$
3. *for all random variables X, Jensen's inequality*
$$E[f(X)] \geq f(E[X]) \tag{A.1.1}$$
is satisfied. If f is strictly convex then the equality implies that $X = E[X]$ w.p. 1.

***Proposition* 3** *A sufficient condition for f to be convex is that*

$$\nabla^2 f \succeq 0$$

where $\succeq 0$ stands for positive semi-definitness, that is, $\nabla^2 f$ has non-negative eigenvalues.

***Proposition* 4** *If f is convex and differentiable, then for all $x_0, \delta \in \mathbb{R}^n$,*

$$f(x_0 + \delta) \geq f(x_0) + \delta.\nabla f(x_0)$$

Theorem A.1.1 *If f has a second derivative which is non-negative (positive) everywhere, then f is convex (strictly convex).*

A.1.2 Concave functions

Definition A.1.2 *A function $f : X \to \mathbb{R}$ is concave if for all $x, y \in X$ and $\alpha \in [0,1]$,*

$$(1-\alpha)f(x) + \alpha f(y) \leq f((1-\alpha)x + \alpha y)$$

Definition A.1.3 *A function $f : X \to \mathbb{R}$ is strictly concave if for all $x, y \in X$ with $x \neq y$ and $\alpha \in [0,1]$,*

$$(1-\alpha)f(x) + \alpha f(y) < f((1-\alpha)x + \alpha y)$$

Geometrically, a function f is concave if the cord between any two points on the function lies everywhere on or below the function itself.

Consider a list of functions $f_i : X \to \mathbb{R}$ for $i = 1, .., n$ and a list of numbers $\alpha_1, .., \alpha_n$. The function $f = \sum_{i=1}^n \alpha_i f_i$ is called a linear combination of $f_1, .., f_n$. If each of the weights $\alpha_i \geq 0$, then f is a non-negative linear combination of $f_1, .., f_n$.

Theorem A.1.2 *Suppose $f_1, .., f_n$ are concave functions and $(\alpha_1, .., \alpha_n) \geq 0$. Then, $f = \sum_{i=1}^n \alpha_i f_i$ is also a concave function. If at least one f_j is also strictly concave and $\alpha_j > 0$, then f is strictly concave.*

Even though a concave function need not be differentiable everywhere, the right and left hand derivatives always exist on the interior of the domain and $f^-(x) \geq f^+(x)$. As a result, f is both right and left continuous and therefore continuous. However, concave functions need not be continuous at the boundary.

For differentiable functions, the following theorem provides a simple necessary and sufficient conditions for concavity.

Theorem A.1.3 *Suppose $f : X \to \mathbb{R}$ is differentiable.*

1. *f is concave if and only if for each $x, y \in X$ we have*

$$f(y) - f(x) \leq f^{'}(x)(y - x)$$

2. *f is strictly concave if and only if the inequality is strict for each $x \neq y$.*

Even if a function is not differentiable everywhere, it is concave if and only if for each $x \in int(X)$, there is an $a \in \mathbb{R}$ such that $f(y) - f(x) \leq a(y - x)$ for all $y \in X$. This is an example of a supporting hyperplane for one dimension.

Theorem (A.1.3) implies that the first derivative function of a concave function is non-increasing.

Theorem A.1.4 *Suppose $f : X \to \mathbb{R}$ is differentiable.*

1. *f is concave if and only if $f^{'}$ is non-increasing.*
2. *f is strictly concave if and only if $f^{''}$ is strictly decreasing.*

Theorem A.1.5 *Suppose $f : X \to \mathbb{R}$ is twice differentiable.*

1. *f is concave if and only if $f^{''} \leq 0$.*
2. *if $f^{''} < 0$, then f is strictly concave.*

Note, f strictly concave does not imply that $f^{''}(x) < 0$ for all x. An example of strictly concave function $f : \mathbb{R}_{++} \to \mathbb{R}$ is

1. $f(x) = \frac{x^\alpha}{\alpha}$ for $\alpha \neq 0, \alpha < 1$
2. $f(x) = \log x$
3. $f(x) = bx - ax^2$ where $a > 0$

The following lemma is an immediate consequence of the definition of concave and convex functions.

Lemma A.1.1 *f is a (strictly) convex function if and only if $-f$ is a (strictly) concave function.*

A.1.3 Some approximations

1. Upper bound on the exponential function obtained by linearising e^x in 0

$$e^x \geq 1 + x$$

2. Lower bound on the logarithm function, derived by using the log operator on

$$\log(1 + x) \geq x$$

3. This inequality follows from the convexity of $f(z) = e^{\alpha z}$

$$e^{\alpha x} \leq 1 + (e^{\alpha} - 1)x \text{ for } x \in [0, 1]$$

4. Another inequality

$$-\log(1 - x) \leq x + x^2 \text{ for } x \in [0, \frac{1}{2}]$$

A.1.4 Conjugate duality

We are going to give a precise definition of the notion of duality as well as some useful results. We use the notation dom (f) to refer to the domain of a function f, that is, where it is defined and finite valued.

Definition A.1.4 *A function f is said to be proper if $f(x) > -\infty$ for all x and $f(x) < \infty$ for some x.*

Definition A.1.5 *A convex function $f : \mathbb{R}^K \to [-\infty, \infty]$ is said to be closed when the epigraph of f is a closed set, or equivalently, the set $\{x : f(x) \leq \alpha\}$ is closed for all $\alpha \in \mathbb{R}$.*

Definition A.1.6 *For any convex function $f : \mathbb{R}^K \to [-\infty, \infty]$, the convex conjugate f^* of f is defined as*

$$f^*(z) = \sup_{x \in \mathbb{R}^K} \left[z.x - f(x) \right]$$

For example, for $f(z) = \frac{1}{2}||z||_p^2$ we get $f^*(z) = \frac{1}{2}||z||q^2$ where $\frac{1}{p} + \frac{1}{p} = 1$. On of the property of the convex conjugate is $\nabla f^* = (\nabla f)^{-1}$. We are now going to cite two useful results.

Theorem A.1.6 *For any closed convex function $f : \mathbb{R}^K \to [-\infty, \infty]$, the conjugate f^* is also closed and convex, and $f^{**} = f$. Furthermore, we can write*

$$f^*(y) = \sup_{x \in \textit{ relint}(\textit{ dom } (f))} \left[y.x - f(x) \right]$$

It means that when taking the sup, we do not have to worry about what happens on the boundary. Further, it tells us that there is a one-to-one correspondence between every closed convex function and its dual. Hence, various properties of the function should translate when we go to the dual. For instance, the following result shows that differentiability is a dual property to strict convexity.

Theorem A.1.7 *Given a proper closed convex function $f : \mathbb{R}^K \to [-\infty, \infty]$, f is finite and differentiable everywhere on $\mathbb{R}^K$ if and only if its conjugate f^* is strictly convex on dom (f^*).*

A.1.5 A note on Legendre transformation

Given an arbitrary smooth convex function f, we can define the Legendre transformation which maps a point $x \in$ dom (f) via the rule $x \to \nabla f(x)$. Under certain circumstances, we get that this map is the inverse of the Legendre transformation of the conjugate f^*, that is, $\nabla f^*(\nabla f(x)) = x$ and $\nabla f(\nabla f^*(y)) = y$ for every $x \in$ dom (f) and $y \in$ dom (f^*). However, the latter only holds when f is strictly convex and the interior of dom (f) is non-empty (see Rockafellar [1995]).

A.1.6 A note on the Bregman divergence

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex with continuous first order partial derivatives.

Definition A.1.7 *The Bregman divergence between x and y with respect to a convex function f is given by*

$$D_f(x, y) = f(x) - f(y) - \nabla f(y)(x - y)$$

and due to convexity we get $D_f(x, y) \geq 0$ for all x and y.

The Bergman distance is in general not symmetric. For example, for $f(x) = \frac{1}{2}||x||^2$ we get $D_f(x, y) = \frac{1}{2}||x - y||^2$. Also, for $f(x) = \sum_{i=1}^n (x_i \log x_i - x_i)$ we get

$$D_f(x, y) = KL(x, y) - \sum_{i=1}^n x_i \log \frac{x_i}{y_i} + \sum_{i=1}^n (y_i - x_i)$$

We now states some properties of the Bregman divergence

Property A.1.1

1. $D_{f+g}(x, y) = D_f(x, y) + D_g(x, y)$
2. $D_f(x, v) + D_f(v, w) = D_f(u, v) + (u - v)(\Delta f(w) - \Delta f(v))$
3. *The Bergman projection into a convex set $\mathcal{K}$ exists and is unique. Let $w^{'}$ be the Bergman projection of the point w into the convex set $\mathcal{K}$. It follows*
$$w^{'} = \arg\min_{v \in \mathcal{K}} D_f(v, w)$$
4. *Generalised Pythagorean Theorem: for all $u \in \mathcal{K}$*
$$D_f(u, w) \geq D_f(u, w^{'}) + D_f(w^{'}, w)$$
5. $D_f(u, v) = D_{f^*}(\nabla f(x), \nabla f(u))$ *where f^* is the Legendre dual.*
6. $D_{f+g}(x, y) = D_f(x, y)$ *if $g(x)$ is linear.*
7. $\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y)$.
8. *If y minimise $f(\nabla f(y) = 0)$, then*
$$D_f(x, y) = f(x) - f(y)$$

A.2 The logistic function

The logistic function is used to study the relation to population growth and model the S-shaped curve of growth of some population P

$$P(t) = \frac{1}{1+e^{-t}}$$

with the property that $1 - P(t) = P(-t)$. The initial stage of growth is approximately exponential, then as saturation begins, the growth slows and stops at maturity. The derivative of the function is

$$\frac{d}{dt}P(t) = P(t)(1 - P(t))$$

which is a simple first-order non-linear differential equation. The logistic equation is an example of an autonomous ordinary differential equation (ODE) since the RHS is independent of t. Hence, if $P(t)$ solves the ODE, so does $P(t-c)$ for any constant c. The derivative is 0 at $P = 0$ or $P = 1$ and the derivative is positive for P in the range $[0, 1]$ and negative for P above 1 or less than 0. It yields an unstable equilibrium at 0, and a stable equilibrium at 1 and thus for any value of P greater than 0 and less than 1, P grows to 1. In general, the logistic equation is

$$\frac{d}{dt}P(t) = (b - aP(t))P(t) = b(1 - \frac{P(t)}{K_{ab}})P(t)$$

where b is the intrinsic growth rate and $K_{ab} = \frac{b}{a}$ is the environmental carrying capacity with critical points $P = 0$ and $P = K_{ab}$. The solution to the ODE is

$$P(t) = \frac{K_{ab}P(0)e^{bt}}{K_{ab} + P(0)(e^{bt} - 1)}$$

so, that if $0 < P(0) < K_{ab}$ then $P \to K_{ab}$ as t grows. As a result

$$\lim_{t\to\infty} P(t) = K_{ab}$$

and K_{ab} is the limiting value value of P, the highest value that the population can reach given infinite time. One should stress that the carrying capacity K_{ab} is asymptotically reached independently of the initial value $P(0) > 0$, as it happens also in the case $P(0) > K_{ab}$.

The logistic equation is separable

$$\frac{1}{(1 - \frac{P(t)}{K_{ab}})P(t)}dP(t) = bdt$$

Since

$$\frac{1}{P(t)} + \frac{\frac{1}{K_{ab}}}{(1 - \frac{P(t)}{K_{ab}})} = \frac{(1 - \frac{P(t)}{K_{ab}}) + \frac{P(t)}{K_{ab}}}{(1 - \frac{P(t)}{K_{ab}})P(t)} = \frac{1}{(1 - \frac{P(t)}{K_{ab}})P(t)}$$

the partial fractions gives

$$\big(\frac{1}{P(t)} + \frac{\frac{1}{K_{ab}}}{(1 - \frac{P(t)}{K_{ab}})}\big)dP(t) = bdt$$

which simplifies to

$$\Big(\frac{1}{P(t)} + \frac{1}{K_{ab} - P(t)}\Big)dP(t) = bdt$$

Integrating, we get

$$\ln|P| - \ln|K_{ab} - P| = bt + C^{'}$$

which gives

$$\ln\Big|\frac{P}{K_{ab} - P}\Big| = bt + C^{'}$$

Exponentiating, we get

$$\frac{P}{K_{ab} - P} = \pm e^{bt+C^{'}} = \pm e^{bt}e^{C^{'}}$$

which gives

$$\frac{P}{K_{ab} - P} = Ce^{bt} \text{ with } C = \pm e^{C^{'}}$$

Solving for P, we get

$$P(t) = C(K_{ab} - P(t))e^{bt} = CK_{ab}e^{bt} - CP(t)e^{bt}$$

which gives

$$P(t) + CP(t)e^{bt} = P(t)(1 + Ce^{bt}) = CK_{ab}e^{bt}$$

with solution

$$P(t) = \frac{CK_{ab}e^{bt}}{(1 + Ce^{bt})}$$

where we determine the constant C from initial condition. For example, if the initial condition is $P(0) = \frac{K_{ab}}{2}$ then

$$\frac{K_{ab}}{2} = \frac{CK_{ab}}{(1 + C)}$$

which implies

$$1 = 2\frac{C}{1 + C}$$

which gives $C = 1$, and the solution is

$$P(t) = K_{ab}\frac{e^{bt}}{(1 + e^{bt})} = K_{ab}\frac{1}{(e^{-bt} + 1)} \text{ with } \lim_{t\to\infty} P(t) = K_{ab}$$

In the general case, with initial condition $P(0) = x$ we get

$$x = \frac{CK_{ab}}{(1 + C)}$$

which gives $C = \frac{x}{K_{ab}-x}$. Hence, the solution becomes

$$P(t) = \frac{\frac{x}{K_{ab}-x}K_{ab}e^{bt}}{(1 + \frac{x}{K_{ab}-x}e^{bt})} = \frac{xK_{ab}e^{bt}}{((K_{ab} - x) + xe^{bt})}$$

and we recover

$$P(t) = \frac{xK_{ab}e^{bt}}{K_{ab} + x(e^{bt} - 1)} = \frac{K_{ab}P(0)e^{bt}}{K_{ab} + P(0)(e^{bt} - 1)}$$

Similarly, with terminal condition $P(T) = x$ we get

$$x = \frac{CK_{ab}e^{bT}}{(1 + Ce^{bT})}$$

which gives $C = \frac{x}{(K_{ab}-x)e^{bT}}$. Hence, the solution becomes

$$P(t) = \frac{\frac{x}{(K_{ab}-x)e^{bT}}K_{ab}e^{bt}}{(1 + \frac{x}{(K_{ab}-x)e^{bT}}e^{bt})} = \frac{xK_{ab}e^{-b(T-t)}}{((K_{ab} - x) + xe^{-b(T-t)})}$$

and we recover

$$P(t) = \frac{xK_{ab}e^{-b(T-t)}}{K_{ab} + x(e^{-b(T-t)} - 1)} = \frac{K_{ab}P(T)e^{-b(T-t)}}{K_{ab} + P(T)(e^{-b(T-t)} - 1)}$$

A.3 The convergence of series

For any sequence $\{a_n\}$ of numbers (real, complex), the associated series is defined as

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + ...$$

and the sequence of partial sums $\{S_k\}$ associated to that series is defined for each k as the sum of the sequence $\{a_n\}$ from a_0 to a_k

$$S_k = \sum_{n=0}^{k} a_n = a_0 + a_1 + ... + a_k$$

When summing a family $\{a_i\}, i \in I$ of non-negative numbers, one may define

$$\sum_{i \in I} a_i = \sup\{\sum_{i \in A} a_i | A \text{ finite }, A \subset I\} \in [0, +\infty]$$

When the sum is finite, the set of $i \in I$ such that $a_i > 0$ is countable. For every $n \geq 1$, the set $A_n = \{i \in I : a_i > \frac{1}{n}\}$ is finite since

$$\frac{1}{n} card(A_n) \leq \sum_{i \in A_n} a_i \leq \sum_{i \in I} a_i < \infty$$

Any sum over non-negative reals can be understood as the integral of a non-negative function with respect to the counting measure,which accounts for the many similarities between the two constructions.

Theorem A.3.1 *Convergence of series with positive terms*
An infinite series with positive terms either converges or diverges to ∞. The series converges if its partial sums are bounded and diverges if its partial sums are not bounded.

Theorem A.3.2 *The integral test for series with positive terms*
Suppose that the series

$$\sum_{n=n_0}^{\infty} a_n$$

with positive terms is such that $a_n = f(n)$ for integers $n \geq c$ with some c, where $y = f(x)$ is continuous on $[c, \infty)$ and decreasing for $X \geq c$. Then, the infinite series above converges if and only if the improper integral

$$\int_c^{\infty} f(x)dx$$

converges.

Theorem A.3.3 *Convergence of the p-series*
The infinite series

$$\sum_{n=1}^{\infty} \frac{1}{n^p} = 1 + \frac{1}{2^p} + \frac{1}{3^p} + ...$$

converges if $p > 1$ and diverges if $p \leq 1$.

Theorem A.3.4 *The Comparison Test with positive terms*
Suppose that $\sum_{n=n_0}^{\infty} a_n$ and $\sum_{n=n_0}^{\infty} b_n$ are series with positive terms.

1. *if $\sum_{n=n_0}^{\infty} b_n$ converges and there are constants M and N such that $a_n \leq Mb_n$ for $n \geq N$, then $\sum_{n=n_0}^{\infty} a_n$ also converges.*
2. *if $\sum_{n=n_0}^{\infty} b_n$ diverges and there are constants $M > 0$ and N such that $a_n \geq Mb_n$ for $n \geq N$, then $\sum_{n=n_0}^{\infty} a_n$ also diverges.*

In computer science, the prefix sum, scan, or cumulative sum of a sequence of numbers $x_0, x_1, x_2, ...$ is a second sequence of numbers $y_0, y_1, y_2, ...$, the sums of prefixes (running totals) of the input sequence

$$\begin{aligned} y_0 &= x_0 \\ y_1 &= x_1 + x_2 \\ y_2 &= x_1 + x_2 + x_3 \\ &... \end{aligned}$$

For instance, the prefix sum of the natural numbers are the triangular numbers

input numbers	1	2	3	4	5	6	...
prefix sums	1	3	6	10	15	21	...

Table A.1: prefix sums of the natural numbers

Prefix sums are trivial to compute in sequential models of computation, by using the formula

$$y_i = y_{i-1} + x_i$$

to compute each output value in sequence order.

A.4 The Heaviside function and the Dirac function

A.4.1 The Heaviside function

The Heaviside function $\mathcal{H}(t)$ is defined by the statement

$$\mathcal{H}(t) = \begin{cases} 0 \text{ for } t < 0 \\ 1 \text{ for } t > 0 \end{cases}$$

Note, $\mathcal{H}(t)$ is undefined when $t = 0$. Also, we can express the function

$$f(t) = \begin{cases} 0 \text{ for } t < T \\ 1 \text{ for } t > T \end{cases}$$

in terms of $\mathcal{H}(t)$ as

$$f(t) = \mathcal{H}(t - T)$$

We can define the Laplace transform of $\mathcal{H}(t - T)$ as follow

$$\begin{aligned} L[\mathcal{H}(t-T)] &= \int_0^\infty e^{-st}\mathcal{H}(t-T)dt = \int_0^T e^{-st}0dt + \int_T^\infty e^{-st}1dt \\ &= [\frac{e^{-st}}{-s}]_T^\infty = \frac{e^{-sT}}{s} \end{aligned}$$

In the special case where $T = 0$, we have $L[\mathcal{H}(t - T)] = \frac{1}{s}$. Given $a < b$, we let the rectangular pulse $P(t)$ of duration $b - a$ and magnitude k be defined by

$$P(t) = \begin{cases} k \text{ for } a < t < b \\ 0 \text{ for } t < a \text{ or } t > b \end{cases}$$

This pulse can be represented in terms of Heaviside function as

$$P(t) = k\big[\mathcal{H}(t-a) - \mathcal{H}(t-b)\big]$$

We can express the Laplace transform of the pulse $P(t)$ as

$$L[P(t)] = k\big[\frac{e^{-sa}}{s} - \frac{e^{-sb}}{s}\big] = k\frac{e^{-sa} - e^{-sb}}{s}$$

We now present a few remarks

1. The strength of the rectangular pulse is defined as the area of the rectangle with base $b-a$ and height k. That is,

$$\text{strength } = k(b-a)$$

2. In general, the expression

$$\big[\mathcal{H}(t-a) - \mathcal{H}(t-b)\big]f(t)$$

switch on the function $f(t)$ between $t = a$ and $t = b$, and switch off the function $f(t)$ when $t < a$ or $t > b$.

3. Similarly, the expression

$$\mathcal{H}(t-a)f(t) \tag{A.4.2}$$

switch on the function $f(t)$ when $t > a$, and switch off the function $f(t)$ when $t < a$.

Theorem A.4.1 *The second shifting theorem*

$$L[\mathcal{H}(t-T)f(t-T)] = e^{-sT}L[f(t)]$$

For example, given the function

$$f(t) = \begin{cases} (t-1)^2 \text{ for } t > 1 \\ 0 \text{ for } 0 < t < 1 \end{cases}$$

assuming $t > 0$, we can rewrite this function in terms of the Heaviside function as

$$f(t) = (t-1)^2\mathcal{H}(t-1)$$

Therefore, using $T = 1$ in the second shifting theorem, we get

$$L[f(t)] = e^{-s}L[t^2] = e^{-s}\frac{2}{s^3}$$

As another example, we consider the function

$$f(t) = \begin{cases} 5t - 11 \text{ for } 0 \leq t < 6 \\ \sin(3t) \text{ for } 6 \leq t < 7 \\ 4 \text{ for } 7 \leq t < 12 \\ t \text{ for } 12 \leq t \end{cases}$$

which we express in terms of the Heaviside function as

$$f(t) = (5t-11)\mathcal{H}(t) - (5t-11)\mathcal{H}(t-6) + \sin(3t)\mathcal{H}(t-6) - \sin(3t)\mathcal{H}(t-7) + 4\mathcal{H}(t-7) - 4\mathcal{H}(t-12) + t\mathcal{H}(t-12)$$

A.4.2 The Dirac function

Note, the Heaviside function can be expressed as a function of time t or space x. The derivative of the Heaviside function $\mathcal{H}(x)$ is zero for $x \neq 0$, and at $x = 0$ the derivative is undefined. We represent the derivative of the Heaviside function by the Dirac delta function $\delta(x)$. The delta function is zero for $x \neq 0$ and infinite at the point $x = 0$. Since the derivative of $\mathcal{H}(x)$ is undefined, $\delta(x)$ is not a function in the conventional sense of the word. While we can derive the properties of the delta function rigorously, we will focus on heuristic treatments. The Dirac delta function is defined by the following properties

$$\delta(x) = \begin{cases} 0 \text{ for } x \neq 0 \\ \infty \text{ for } x = 0 \end{cases}$$

and

$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

The second property comes from the fact that $\delta(x)$ represents the derivative of $\mathcal{H}(x)$. Even though the delta function is not a function, it is said to be a distribution, that is, it can only be used inside integrals. In fact, $\int \delta(x)dx$ can be regarded as an operator which pulls the value of a function at zero. Hence, as long as it is understood that the delta function is eventually integrated, we can use it as if it is a function. To see this, we let $f(x)$ be a continuous function vanishing at infinity, and consider the integral

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx$$

Using integration by parts to evaluate the integral, we get

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\delta(x)dx &= [f(x)\mathcal{H}(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f^{'}(x)\mathcal{H}(x)dx \\ &= -\int_{0}^{\infty} f^{'}(x)dx = [-f(x)]_{0}^{\infty} = f(0)\end{aligned}$$

In order to evaluate the integral by parts, we assumed that $f(x)$ vanishes at infinity. However, since the delta function is zero for $x \neq 0$, the integrand is nonzero only at $x = 0$. Thus, the behaviour of the function at infinity should not affect the value of the integral. Hence, it is reasonable $f(0) = \int_{-\infty}^{\infty} f(x)\delta(x)dx$ holds for all continuous functions. By changing variables and noting that $\delta(x)$ is symmetric, we can derive a more general formula

$$\begin{aligned}f(0) &= \int_{-\infty}^{\infty} f(y)\delta(y)dy \\ f(x) &= \int_{-\infty}^{\infty} f(y+x)\delta(y)dy \\ f(x) &= \int_{-\infty}^{\infty} f(y)\delta(y-x)dy \\ f(x) &= \int_{-\infty}^{\infty} f(y)\delta(x-y)dy\end{aligned}$$

This formula is very important in solving inhomogeneous differential equations. That is, if $x \in \mathbb{R}$, then the Dirac function δ_x represents a notional function with the properties

- $\delta_x(y) = 0$ if $y \neq x$
- $\int_{-\infty}^{\infty} g(y)\delta_x(y)dy = g(x)$ for all integrable $g : \mathbb{R} \to \mathbb{R}$

It satisfies the scaling property for non-zero scalar α

$$\int_{-\infty}^{\infty} \delta(\alpha y)dy = \int_{-\infty}^{\infty} \delta(u)\frac{du}{|\alpha|} = \frac{1}{|\alpha|}$$

so that $\delta(\alpha y) = \frac{\delta(y)}{|\alpha|}$. The composition with a function is

$$\delta_{x_0}(g(y)) = \frac{\delta(y - x_0)}{|g^{'}(x_0)|}$$

For a translation, the time-delayed Dirac function is

$$\int_{-\infty}^{\infty} f(t)\delta(t - T)dt = f(T)$$

If x is a set, $x_0 \in x$ is a marked point and Σ is any sigma algebra of subsets of x, then the measure defined on sets $A \in \Sigma$ is

$$\delta_{x_0}(A) = \begin{cases} 1 \text{ if } x_0 \in A \\ 0 \text{ otherwise} \end{cases}$$

We can view the delta function as a limit of the Gaussian density

$$\delta(x) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

or the Lorentzian

$$\delta(x) = \lim_{\epsilon \to 0} \frac{1}{\pi} \frac{\epsilon}{x^2 + \epsilon^2}$$

We can also consider the function $b(x, \epsilon)$ defined by

$$b(x, \epsilon) = \begin{cases} 0 \text{ for } |x| > \frac{\epsilon}{2} \\ \frac{1}{\epsilon} \text{ for } |x| < \frac{\epsilon}{2} \end{cases}$$

and define the delta function $\delta(x)$ as $b(x, \epsilon)$ in the limit as $\epsilon \to 0$. Further, when the delta function appears inside an integral, we can think of the delta function as a delayed limiting process

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx = \lim_{\epsilon \to 0} \int_{-\infty}^{\infty} f(x)b(x, \epsilon)dx$$

We can also relate the delta function to the Fourier transform. Given the function $f(t)$, the Fourier transform is given by

$$\tilde{f}(s) = \int_{-\infty}^{\infty} \frac{e^{its}}{\sqrt{2\pi}} f(t)dt$$

so that we go back to $f(t)$ from $\tilde{f}(s)$ by

$$f(t) = \int_{-\infty}^{\infty} \frac{e^{-its}}{\sqrt{2\pi}} \tilde{f}(s)ds$$

If we set $f(t) = \delta(t)$ in the above equations, we get

$$\tilde{\delta}(s) = \int_{-\infty}^{\infty} \frac{e^{its}}{\sqrt{2\pi}} \delta(t)dt = \frac{1}{\sqrt{2\pi}}$$

and

$$\delta(t) = \int_{-\infty}^{\infty} \frac{e^{-its}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} ds = \int_{-\infty}^{\infty} \frac{e^{-its}}{2\pi} ds$$

that is, the delta function and the constant $\frac{1}{\sqrt{2\pi}}$ are Fourier transform of each other. Another way to view the integral representation of the delta function is by using the limits. Given the limit of the Gaussian density above, we get

$$\begin{aligned} \delta(x) &= \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} \\ &= \lim_{\sigma \to 0} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{w^2\sigma^2}{2}} e^{-iwt} dw = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-iwt} dw \end{aligned}$$

A.5 Some linear algebra

A vector space is a set V equipped with two operations

- addition

$$V \times V \ni (x, y) \to x + y \in V$$

- and scalar multiplication

$$\mathbb{R} \times V \ni (r, x) \to rx \in V$$

having the following properties

Property A.5.1 *1.* $a + b = b + a$ *for all* $a, b \in V$

2. $(a + b) + c = a + (b + c)$ *for all* $a, b, c \in V$
3. *there exists an element of* V*, called the zero vector, and denoted* 0*, such that* $a + 0 = 0 + a = a$ *for all* $a \in V$
4. *for any* $a \in V$ *there exists an element of* V*, denoted* $-a$*, such that* $a + (-a) = (-a) + a = 0$
5. $r(a + b) = ra + rb$ *for all* $r \in \mathbb{R}$ *and* $a, b \in V$
6. $(r + s)a = ra + sa$ *for all* $r, s \in \mathbb{R}$ *and* $a \in V$
7. $(rs)a = r(sa)$ *for all* $r, s \in \mathbb{R}$ *and* $a \in V$
8. $1a = a$ *for all* $a \in V$

Any vector space V has a corresponding dual vector space consisting of all linear functionals on V together with a naturally induced linear structure. Given any vector space V over a field F, the dual space V^* is defined as the set of all linear maps $f : V \to F$. The dual space V^* itself becomes a vector space over F when equipped with an addition and scalar multiplication satisfying:

$$\begin{aligned} (f + g)(x) &= f(x) + g(x) \\ (af)(x) &= af(x) \end{aligned}$$

for all f and $g \in V^*$, $x \in V$ and $a \in F$.

Definition A.5.1 *A vector space* V_0 *is a subspace of a vector* V *if* $V_0 \subset V$ *and the linear operations on* V_0 *agree with the linear operations on* V.

Proposition 5 *A subset* S *of a vector space* V *is a subspace of* V *if and only* S *is nonempty and closed under linear operations, that is,*

$$\begin{aligned} &x, y \in S \Rightarrow x + y \in S \\ &x \in S \Rightarrow rx \in S \text{ for all } r \in \mathbb{R} \end{aligned}$$

Let V be a vector space and $v_1, v_2, ..., v_n \in V$. Consider the set L of all linear combinations $r_1v_1 + r_2v_2 + ... + r_nv_n$ where $r_1, r_2, ..., r_n \in \mathbb{R}$.

Theorem A.5.1 *L is a subspace of V.*

Let S be a subset of a vector space V.

Definition A.5.2 *The span of the subset S, denoted $Span(S)$, is the smallest subset of V that contains S. That is,*

- *$SpanS$ is a subset of V.*
- *for any subspace $W \subset V$ one has*

$$S \subset W \Rightarrow Span(S) \subset W$$

Let S be a subset of a vector space V.

- If $S = \{v_1, v_2, .., v_n\}$ then $Span(S)$ is the set of all linear combinations $r_1v_1 + r_2v_2 + ... + r_nv_n$ where $r_1, r_2, ..., r_n \in \mathbb{R}$.
- If S is an infinite set then $Span(S)$ is the set of all linear combinations $r_1u_1 + r_2u_2 + ... + r_ku_k$ where $u_1, u_2, ..., u_k \in S$ and $r_1, r_2, ..., r_k \in \mathbb{R}$ for $k \geq 1$.
- If S is the empty set then $Span(S) = \{0\}$.

Definition A.5.3 *A subset S of a vector space V is called a spanning set for V if $Span(S) = V$.*

We say that the set S spans the subspace W or that S is a spanning set for W. If S_1 is a spanning set for a vector space V and $S_1 \subset S_2 \subset V$, then S_2 is also a spanning set for V.

Definition A.5.4 *Let V be a vector space. Vectors $v_1, v_2, ..., v_k \in V$ are called linearly dependent if they satisfy a relation*

$$r_1v_1 + r_2v_2 + ... + r_kv_k = 0$$

where the coefficients $r_1, r_2, ..., r_k \in \mathbb{R}$ are not all equal to zero. Otherwise, vectors $v_1, v_2, ..., v_k$ are called linearly independent. That is, if

$$r_1v_1 + r_2v_2 + ... + r_kv_k = 0 \Rightarrow r_1 = .. = r_k = 0$$

An infinite set $S \subset V$ is linearly dependent if there are some linearly dependent vectors $v_1, v_2, ..., v_k \in S$. Otherwise S is linearly independent.

Theorem A.5.2 *The following conditions are equivalent:*

1. *vectors $v_1, v_2, ..., v_k$ are linearly dependent.*
2. *one of vectors $v_1, v_2, ..., v_k$ is a linear combination of the other $k - 1$ vectors.*

Theorem A.5.3 *Vectors $v_1, v_2, ..., v_m \in \mathbb{R}^n$ are linearly dependent whenever $m > n$ (the number of coordinates is less than the number of vectors).*

Definition A.5.5 *Let V be a vector space. A linearly independent spanning set for V is called a basis.*

Assuming that a set $S \subset V$ is a basis for V. Then, a spanning set means that any vector $v \in V$ can be represented as a linear combination

$$v = r_1v_1 + r_2v_2 + ... + r_kv_k$$

where $v_1, v_2, ..., v_k$ are distinct vectors from S and $r_1, r_2, ..., r_k \in \mathbb{R}$. Linearly independent implies that the above representation is unique

$$\begin{aligned} v &= r_1 v_1 + r_2 v_2 + ... + r_k v_k = r_1^{'} v_1 + r_2^{'} v_2 + ... + r_k^{'} v_k \\ &\Rightarrow (r_1 - r_1^{'}) v_1 + (r_2 - r_2^{'}) v_2 + ... + (r_k - r_k^{'}) v_k = 0 \\ &\Rightarrow (r_1 - r_1^{'}) = (r_2 - r_2^{'}) = ... = (r_k - r_k^{'}) = 0 \end{aligned}$$

Let $v_1, v_2, ..., v_k$ be vectors in $\mathbb{R}^n$.

Theorem A.5.4 *If $k < n$ then the vectors $v_1, v_2, ..., v_k$ do not span $\mathbb{R}^n$.*

Theorem A.5.5 *If $k > n$ then the vectors $v_1, v_2, ..., v_k$ are linearly dependent.*

Theorem A.5.6 *If $k = n$ then the following conditions are equivalent:*

1. *$\{v_1, v_2, ..., v_n\}$ is a basis for $\mathbb{R}^n$.*
2. *$\{v_1, v_2, ..., v_n\}$ is a spanning set for $\mathbb{R}^n$.*
3. *$\{v_1, v_2, ..., v_n\}$ is a linearly independent set.*

Theorem A.5.7 *Any vector space has a basis.*

Theorem A.5.8 *If a vector space V has a finite basis, then all bases for V are finite and have the same number of elements.*

Definition A.5.6 *The dimension of a vector space V, denoted $\dim V$, is the number of elements in any of its bases.*

Theorem A.5.9 *Let S be a subset of a vector space V. then the following conditions are equivalent:*

1. *S is a linearly independent spanning set for V, that is, a basis.*
2. *S is a minimal spanning set for V.*
3. *S is a maximal linearly independent subset of V.*

Definition A.5.7 *Given vector spaces V_1 and V_2, a mapping $L : V_1 \to V_2$ is linear if*

$$\begin{aligned} L(x + y) &= L(x) + L(y) \\ L(rx) &= rL(x) \end{aligned}$$

for any $x, y \in V_1$ and $r \in \mathbb{R}$.

A linear mapping $l : V \to \mathbb{R}$ is called a linear functional on V. If $V_1 = V_2$ (or V_1 and V_2 are functional spaces), then a linear mapping $L : V_1 \to V_2$ is called a linear operator. Some properties of linear mappings are

- If a linear mapping $L : V \to W$ is invertible, then the inverse mapping $L^{-1} : V \to W$ is also linear.
- If $L : V \to W$ and $M : W \to X$ are linear mappings, then the composition $MoL : V \to X$ is also linear.
- If $L_1 : V \to W$ and $L_2 : V \to W$ are linear mappings, then the sum $L_1 + L_2$ is also linear.

If $\{v_1, v_2, .., v_n\}$ is a basis for a vector space V, then any vector $v \in V$ has a unique representation

$$v = x_1 v_1 + x_2 v_2 + ... + x_n v_n$$

where $x_i \in \mathbb{R}$. The coefficients $x_1, x_2, .., x_n$ are called the coordinates of v with respect to the ordered basis $v_1, v_2, .., v_n$. This mapping is a linear transformation. Let V, W be vector spaces, and $L : V \to W$ be a linear mapping.

Definition A.5.8 *The range (or image) of L is the set of all vectors $w \in W$ such that $w = L(v)$ for some $v \in V$. The range of L is denoted $L(v)$.*

The kernel of L, denoted $Ker(L)$, is the set of all vectors $v \in V$ such that $L(v) = 0$.

Theorem A.5.10 *1. The range of L is a subspace of W.*

2. The kernel of L is a subspace of V.

Definition A.5.9 *Vectors $x, y \in \mathbb{R}^n$ are said to be orthogonal (denoted $x \perp y$) if $x.y = 0$.*

Definition A.5.10 *A vector $x \in \mathbb{R}^n$ is said to be orthogonal to a nonempty set $Y \subset \mathbb{R}^n$ (denoted $x \perp Y$) if $x.y = 0$ for any $y \in Y$.*

Definition A.5.11 *Nonempty sets $X, Y \in \mathbb{R}^n$ are said to be orthogonal (denoted $X \perp Y$) if $x.y = 0$ for any $x \in X$ and $y \in Y$.*

***Proposition* 6** *If $X, Y \in \mathbb{R}^n$ are orthogonal sets, then either they are disjoint or $X \cap Y = \{0\}$.*

***Proposition* 7** *Let V be a subspace of $\mathbb{R}^n$ and S be a spanning set for V. Then for any $x \in \mathbb{R}^n$*

$$x \perp S \Rightarrow x \perp V$$

Definition A.5.12 *Let $S \subset \mathbb{R}^n$. The orthogonal complement of S, denoted $S^\perp$, is the set of all vectors $x \in \mathbb{R}^n$ that are orthogonal to S. That is, $S^\perp$ is the largest subset of $\mathbb{R}^n$ orthogonal to S.*

Theorem A.5.11 *$S^\perp$ is a subspace of $\mathbb{R}^n$.*

Note that $S \subset (S^\perp)^\perp$, hence $Span(S) \subset (S^\perp)^\perp$.

Theorem A.5.12 *$(S^\perp)^\perp = Span(S)$. In particular, for any subspace V we have $(V^\perp)^\perp = V$.*

Theorem A.5.13 *Let V be a subspace of $\mathbb{R}^n$. Then any vector $x \in \mathbb{R}^n$ is uniquely represented as $x = p + o$, where $p \in V$ and $o \in V^\perp$.*

Note, p is called the ortogonal projection of the vector x onto the subspace V.

Theorem A.5.14 *$||x - v|| > ||x - p||$ for any $v \neq p$ in V. Thus,*

$$||o|| = ||x - p|| = \min_{v \in V} ||x - v||$$

is the distance from the vector x to the subspace V.

Theorem A.5.15 *$||x||_p$ is a norm on $\mathbb{R}^n$ for any $p \geq 1$.*

We let V be an inner product space with an inner product $< \bullet, \bullet >$ and the induced norm $|| \bullet ||$.

Definition A.5.13 *A nonempty set $S \subset V$ of nonzero vectors is called an orthogonal set if all vectors in S are mutually orthogonal. That is, $0 \notin S$ and $< x, y >= 0$ for any $x, y \in S$, $x \neq y$.*

Definition A.5.14 *An orthogonal set $S \subset V$ is called orthonormal if $||x|| = 1$ for any $x \in S$.*

For example, vectors $v_1, v_2, .., v_k \in V$ form an orthonormal set if and only if

$$< v_i, v_j >= \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Let V be a vector space with an inner product. Given the set $\{v_1, v_2, ..., v_n\}$ of orthogonal basis for V, an orthonormal set is formed by normalised vectors $w_i = \frac{v_i}{||v_i||}$ for $i = 1, ..., n$.

Theorem A.5.16 *Suppose $v_1, v_2, .., v_k$ are nonzero vectors that form an orthogonal set. Then, $v_1, v_2, .., v_k$ are linearly independent.*

Theorem A.5.17 *Let V be an inner product space and V_0 be a finite dimensional subspace of V. Then any vector $x \in V$ is uniquely represented as $x = p + o$, where $p \in V_0$ and $o \perp V_0$.*

Note, p is called the orthogonal projection of the vector x onto the subspace V_0. If V_0 is a one-dimensional subspace spanned by a vector v, then $p = \frac{<x,v>}{<v,v>} v$.

A.6 Some facts on matrices

We present a few facts on matrices that can be found in books, see for example Strang [1980]. We let $A = [a_{ij}]_{m \times n}$ and $C = [c_{ij}]_{p \times q}$ be two matrices with dimensions given in the subscript, and we let b be a real number. The scalar multiplication is defined as $bA = [ba_{ij}]_{m \times n}$ and the multiplication as $AC = [\sum_{k=1}^{n} a_{ik} c_{kj}]_{m \times q}$ provided that $n = p$. We let $A^\top = [a_{ij}^\top]$ be the transpose of A such that $a_{ij}^\top = a_{ji}$ and $(A^\top)^\top = A$. If $A^\top = A$, then A is a symmetric matrix. Further, $(AC)^\top = C^\top A^\top$ and $AC \neq CA$ in general. A square matrix $A_{m \times m}$ is non-singular or invertible if there exists a unique matrix $C_{m \times m}$ such that $AC = CA = I_m$, the $m \times m$ identity matrix. The matrix C is called the inverse matrix of A and is denoted by $C = A^{-1}$. The trace of $A_{m \times m}$ is the sum of the diagonal elements, that is, $tr(A) = \sum_{i=1}^{m} a_{ii}$. It has the following properties

- $tr(A + C) = tr(A) + tr(C)$
- $tr(A) = tr(A^\top)$
- $tr(AC) = tr(CA)$ provided that the two matrixes are conformable

Any linear map $f : V \to W$ between finite-dimensional vector spaces can be described by a matrix $A = (a_{ij})$ after choosing bases $v_1, ..., v_n$ of V and $w_1, ..., w_m$ of W which is such that

$$f(v_j) = \sum_{i=1}^{m} a_{ij} w_i \text{ for } j = 1, .., n$$

The transpose matrix $A^\top$ describes the transpose of the linear map given by A, with respect to the dual bases. More generally, the set of $m \times n$ matrices can be used to represent the $\mathbb{R}$-linear maps between the free modules $\mathbb{R}^m$ and $\mathbb{R}^n$ for an arbitrary ring $\mathbb{R}$ with unity. When $m = n$ composition of these maps is possible, and this gives rise to the matrix ring of $n \times n$ matrices representing the endomorphism ring of $\mathbb{R}^n$.

A number λ and a $m \times 1$ vector e, possibly complex-valued, are a right eigenvalue and eigenvector pair of the matrix A if $Ae = \lambda e$. There are m possible eigenvalues for the matrix A. For a real-valued matrix A, complex eigenvalues occur in conjugated pairs. The matrix A is non-singular if and only if all of its eigenvalues are non-zero. If we denote the eigenvalues by $\{\lambda_i\}_{i=1}^m$, we get $tr(A) = \sum_{i=1}^m \lambda_i$, and the determinant of the matrix A can be defined as $|A| = \prod_{i=1}^m \lambda_i$. Further, the rank of the matrix $A_{m\times n}$ is the number of non-zero eigenvalues of the symmetric matrix $AA^\top$. For a non-singular matrix A, then $(A^{-1})^\top = (A^\top)^{-1}$.
A square matrix $A_{m\times m}$ is a positive definite matrix if

1. A is symmetric, and
2. all eigenvalues of A are positive

Alternatively, A is a positive definite matrix if for any non-zero m-dimensional vector b, we have $b^\top Ab > 0$. Some useful properties of a positive definite matrix A include

1. all eigenvalues of A are real and positive
2. the matrix can be decomposed as

$$A = P\Lambda P^\top$$

where Λ is a diagonal matrix consisting of all eigenvalues of A, and P is a $m \times m$ matrix consisting of the m right eigenfactors of A.

We can write the eigenvalues as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ and the eigenvectors as $e_1, ..., e_m$ such that $Ae_i = \lambda_i e_i$ and $e_i^\top e_i = 1$. Further, these eigenvectors are orthogonal to each other, that is, $e_i^\top e_j = 0$ if $i \neq j$, if the eigenvalues are distinct. The matrix P is orthogonal, and the decomposition is called the spectral decomposition of the matrix A. For a symmetric matrix A, there exists a lower triangular matrix L with diagonal elements being 1 and a diagonal matrix G such that $A = LGL^\top$. If A is positive definite, then the diagonal elements of G are positive. In this case

$$A = L\sqrt{G}\sqrt{G}L^\top = (L\sqrt{G})(L\sqrt{G})^\top$$

where $L\sqrt{G}$ is again a lower triangular matrix and the square root is taking element by element. This decomposition, called the Cholesky decomposition, shows that a positive definite matrix A can be diagonalised as

$$L^{-1}A(L^\top)^{-1} = A(L^{-1})^\top = G$$

Since L is a lower triangular matrix with unit diagonal elements, L^{-1} is also lower triangular matrix with unit diagonal elements.

Writing a $m \times n$ matrix A in its columns as $A = [a_1, ..., a_n]$ we define the stacking operation as $vec(A) = (a_1^\top, ..., a_m^\top)^\top$, which is a $mn \times 1$ vector. For two matrices $A_{m\times n}$ and $C_{p\times q}$, the Kronecker product between A and C is

$$A \otimes C = \begin{bmatrix} a_{11}C & a_{12}C & ... & a_{1n}C \\ a_{21}C & a_{22}C & ... & a_{2n}C \\ . & . & ... & . \\ . & . & ... & . \\ a_{m1}C & a_{m2}C & ... & a_{mn}C \end{bmatrix}_{mp\times nq}$$

For example, asuume A is a 2×2 matrix and C is a 2×3 matrix

$$A = \begin{bmatrix} 2 & 1 \\ -1 & 3 \end{bmatrix}, \begin{bmatrix} 4 & -1 & 3 \\ -2 & 5 & 2 \end{bmatrix}$$

then $vec(A) = (2, -1, 1, 3)^\top$, $vec(C) = (4, -2, -1, 5, 3, 2)^\top$, and

$$A \otimes C = \begin{bmatrix} 8 & -2 & 6 & 4 & -1 & 1 \\ -4 & 10 & 4 & -2 & 5 & 2 \\ -4 & 1 & -3 & 12 & -3 & 9 \\ 2 & -5 & -2 & -6 & 15 & 6 \end{bmatrix}$$

Some useful proprties for the two operators are

- $A \otimes C \neq C \otimes A$ in general
- $(A \otimes C)^\top = A^\top \otimes C^\top$
- $A \otimes (C + D) = A \otimes C + A \otimes D$
- $(A \otimes C)(F \otimes G) = (AF) \otimes (CG)$
- if A and C are invertible, then $(A \otimes C)^{-1} = A^{-1} \otimes C^{-1}$
- for square matrixes A and C, $tr(A \otimes C) = tr(A)tr(C)$
- $vec(A + C) = vec(A) + vec(C)$
- $vec(ABC) = (C^\top \otimes A)vec(B)$
- $tr(AC) = vec(C^\top)^\top vec(A) = vec(A^\top)^\top vec(C)$

When dealing with symmetric matrices, one can generalise the stacking operation to the half-stacking operation, consisting of elements on or below the main diagonal. For a symmetric square matrix $A = [a_{ij}]_{k \times k}$, define

$$vech(A) = (a_{1.}^\top, a_{2*}^\top, ..., a_{k*}^\top)^\top$$

where $a_{1.}^\top$ is the first column of A, and $a_{i*} = (a_{ii}, a_{i+1,i}, ..., a_{k,i})^\top$ is a $(k-i+1)$-dimensional vector. The dimension of $vech(A)$ is $\frac{k(k+1)}{2}$. For example, for $k = 3$ we get $vech(A) = (a_{11}, a_{21}, a_{3,1}, a_{22}, a_{3,2}, a_{33})^\top$.

We consider the function $f(t, X_t) = X_t^\top A(t) X_t + B(t)^\top X_t + C(t)$ and recall that quadratic forms are additive. For a symmetric $n \times n$ matrix A, the vector $vec[A]$ [1] contains more information than is strictly necessary, since the matrix is completely determined by the symmetry together with the lower triangular portion, that is the $\frac{1}{2}n(n+1)$ entries on and below the main diagonal. For that symmetric matrix A, we let $v[A]$ denotes the $N \times 1$ vector with $N = \frac{1}{2}n(n+1)$ obtained from $vec[A]$ by eliminating the supradiagonal entries of A. Note, v is called the vector-half operator sometimes denoted by $vech$. For example, for a 2×2 matrix A we have

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and we get

$$v[A] = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$$

[1] $vec[A]$ is the vector containing all the entries of the matrix A

where $v[A]$ only collects the distinct elements of the matrix A. Note, we could have chosen to eliminate the subdiagonal entries of A (to get the upper triangular part of A) getting $v[A] = [a_{11}, a_{21}, a_{22}]^\top$. In fact, there exists a unique matrix D_n of size $n^2 \times N$ called the duplication matrix, such that

$$D_n v[A] = vec[A]$$

It turns out that all quadratic form $X^\top AX$ can be written in terms of $Z = v[XX^\top]$ as

$$\begin{aligned} X^\top AX &= vec[A]^\top vec[XX^\top] \\ &= vec[A]^\top D_n Z \end{aligned}$$

where Z only collects the distinct elements of the squared symmetric matrix $XX^\top$. Therefore, we can re-write the quadratic function $f(t, X_t)$ as

$$X_t^\top A(t)X_t + B^\top(t)X_t + C(t) = vec[A]^\top D_n Z + B^\top(t)X_t + C(t)$$

which is linear in the vectors X and Z.

A.7 Utility function

A.7.1 Definition

In economics and game theory, utility represents satisfaction experienced by the consumer of a good. In economics, utility is a representation of preferences over some set of goods and services. Preferences have a (continuous) utility representation so long as they are transitive, complete, and continuous. In finance, utility is applied to generate an individual's price for an asset called the indifference price (see Appendix (F.5.4)). Utility functions are also related to risk measures, with the most common example being the entropic risk measure (see Appendix (B.1)). Since desires can not be measured directly, economists inferred relative utility in people's willingness to pay.
Let X be the consumption set, the set of all mutually-exclusive baskets the consumer could conceivably consume. The consumer's utility function $u : X \to \mathbb{R}$ ranks each package in the consumption set. If the consumer strictly prefers x to y or is indifferent between them, then $u(x) \geq u(y)$. There are usually a finite set of L commodities, and a consumer may consume an arbitrary amount of each commodity. This gives a consumption set of $\mathbb{R}^L_+$, and each package $x \in \mathbb{R}^L_+$ is a vector containing the amounts of each commodity. A utility function $u : X \to \mathbb{R}$ represents a preference relation $\preceq$ on X if and only if for every $x, y \in X$, then $u(x) \leq u(y)$ implies $x \preceq y$. If u represents $\preceq$, then it implies that $\preceq$ is complete and transitive, and hence rational. In order to simplify calculations, various assumptions have been made on utility functions such as constant elasticity of substitution (CES) utility, exponential utility, quasilinear utility, homothetic preferences.

Utility functions can be defined either over the positive real line or over the whole real line. One can show that a utility function is only unique up to an increasing affine transformation. As we can rescale utility with an affine transformation, the actual number we see is not an intuitive measurement scale for investments and only the ordering of the utilities is meaningful. We can always translate the expected utility into the certain equivalent of an uncertain investment x as the monetary value $CE(x)$ that has the same utility as the expected utility as the investment.

$$u(CE(x)) = E[u(x)]$$

While both the certain equivalent and the expected utility are single numbers used to rank investments, the certain equivalent has a more intuitive interpretation than the expected utility. Note, in both cases the rankings they produce are preference ordering, and it coincides if the utility function is strictly monotonic increasing.

A.7.2 Some properties

Assuming more is better, the utility function is usually a strictly monotonic increasing function of wealth (W), that is, the marginal utility is always positive, $u^{'}(W) > 0$ for all X. Note, $u^{''}(X)$ determines the curvature of the utility function, and it can take either sign. We can characterise the risk preference as follow.

- if $u^{''}(W) > 0$, the utility is a convex function of W, with an increasing marginal utility. The investor is risk loving.
- if $u^{''}(W) < 0$, the utility is a concave function of W, with a diminishing marginal utility. The investor is risk averse.
- if $u^{''}(W) = 0$, the utility is a linear function of W. The investor is risk neutral.

An investment P may be described by a probability distribution over the utilities associated with all possible outcomes. We let μ_P and σ_P be the expectation and standard deviation of the distribution

$$\mu_P = E[u(P)] \text{ and } \sigma_P^2 = Var(u(P))$$

Using a second order Taylor series expansion of $u(P)$ around μ_P, we get

$$u(P) \approx u(\mu_P) + u^{'}(\mu_P)(P - \mu_P) + \frac{1}{2}u^{''}(\mu_P)(P - \mu_P)^2$$

Taking the expectation of the above equation and using the property

$$E[u(\mu_P)] = E[u(E[u(P)])] = \mu_P$$

and the property

$$E[P - \mu_P] = E[P] - \mu_P = 0$$

we get the expected utility of an investment approximated as

$$E[u(P)] \approx \mu_P + \frac{1}{2}\sigma_P^2 u^{''}(\mu_P) \tag{A.7.3}$$

and we can deduce that

- if $u^{''}(\mu_P) > 0$ then $CE(P) > \mu_P$ and the investor puts a greater certain equivalent on an uncertain investment (gamble) than its expected value.
- if $u^{''}(\mu_P) < 0$ then $CE(P) < \mu_P$ and the investor puts a lower certain equivalent on an uncertain investment (gamble) than its expected value.
- if $u^{''}(\mu_P) = 0$ then $CE(P) = \mu_P$ and the investor has a certain equivalent equal to the expected value.

As a diminishing marginal utility of wealth implies the investor is risk averse, the degree of risk aversion (the extent of the concavity of the utility function) is measured by the Coefficient of Absolute Risk Aversion (CARA) defined as

$$A(W) = \frac{-u^{''}(W)}{u^{'}(W)} \tag{A.7.4}$$

or by the Coefficient of Relative Risk Aversion (CRRA)

$$R(W) = \frac{-Wu^{''}(W)}{u^{'}(W)}$$

As an example, the logarithmic utility function

$$u(x) = \ln(x) \, , x > 0$$

has derivatives $u^{'}(x) = x^{-1}$ and $u^{''}(x) = -x^{-2}$ and the CARA is

$$A(W) = W^{-1}$$

so that $A^{'}(W) = -W^{-2} < 0$ with an ARA decreasing with wealth, meaning that the absolute value of the investment in risky assets will increase as the investor becomes more wealthy. Similarly, the logarithmic utility function has a constant CRRA given by

$$R(W) = 1$$

so that the investor will hold the same proportion of his wealth in risky assets no matter how rich he becomes. In general, investors with increasing ARA $(A^{'}(W) > 0)$ will hold less in risky assets in absolute terms as their wealth increases. On the other hand, investors with increasing RRA $(R^{'}(W) > 0)$ will hold proportionally less in risky assets as their wealth increases. Investors may have increasing, constant, or decreasing absolute or relative risk aversion depending on the functional form assumed for the utility function.

A.7.3 Some specific utility functions

Following Henderson et al. [2004], we define a utility function $u(x)$ as a twice continuously-differentiable function, strictly increasing to reflect that investors prefer more wealth to less, and strictly concave because investors are risk-averse. Considering the coefficient of absolute risk aversion (CARA) (see Appendix (E.2)), and defined above as $R_\alpha(x) = -\frac{u^{''}(x)}{u^{'}(x)}$, a utility function is of the Hara class if $R_\alpha(x)$ satisfies

$$R_\alpha(x) = \frac{1}{A + Bx} \, , x \in I_D \tag{A.7.5}$$

where I_D is the interval on which u is defined and B is a non-negative constant. The constant A is such that $A > 0$ if $B = 0$, whereas A can take any value if B is positive. If $B > 0$ then $u(x) = -\infty$ for $x < -\frac{A}{B}$ and $I_D = (-\frac{A}{B}, \infty)$. If $B > 0$ and $B \neq 1$, then integration leads to

$$u(x) = \frac{C}{B-1}(A + Bx)^{1-\frac{1}{B}} + D \, , C > 0, D \in \mathbb{R}, x > -\frac{A}{B}$$

where C and D are constants of integration. This is called the extended power utility function. If $A = 0$, it becomes the well known narrow power utility function

$$u(x) = \frac{CB^{-\frac{1}{B}}}{B-1} Bx^{1-\frac{1}{B}} + D \, , C > 0, D \in \mathbb{R}, x > 0$$

It is more usually written with $R = \frac{1}{B}$, $D = 0$, $C = B^{\frac{1}{B}}$, giving

$$u(x) = \frac{x^{1-R}}{1-R} \, , R \neq 1$$

The narrow power utility has constant relative risk aversion (RRA) of R, where relative risk aversion $R_r(x)$ is defined to be $R_r(x) = xR_\alpha(x)$. Setting $B = 1$ to a utility functions in the Hara class, we get

$$u(x) = C \ln(A + x) + E \, , C > 0, E \in \mathbb{R}, x > -A$$

called the logarithmic utility function. Taking $A = 0$, $E = 0$, $C = 1$ gives the standard or narrow form.

Investors with constant CRRA want the same percentage of their wealth in risky assets as their wealth increases. However, in general investment decisions will affect the wealth of the decision maker only marginally. In that case, many decision makers will adopt a constant CARA utility, where the absolute amount invested in risky assets is independent of their wealth. There are only two types of utility functions with the CARA property, the linear utility function

$$u(x) = A + Bx \, , B > 0$$

which has CARA equal to 0, and the exponential utility function. In our setting, for $B = 0$, we get the exponential utility function

$$u(x) = -\frac{F}{A}e^{-\frac{x}{A}} + G \, , F > 0, A > 0, G \in \mathbb{R}, x \in \mathbb{R}$$

It is usual to take $G = 0$, $A = \frac{1}{\gamma}$, $F = \frac{1}{\gamma^2}$ so that the coefficient of absolute risk aversion (CARA) becomes $R_\alpha(x) = \gamma$, a constant. The exponential utility is an appropriate choice for an investor who wants to hold the same dollar amount in risky assets as his wealth increases. As a result, the percentage of his wealth invested in risk assets will decrease as his wealth increases, and he will have decreasing RRA but constant ARA. Note, the CARA is not very intuitive as it is measured in $\$^{-1}$ units if the initial wealth is measured in $\$$. It is easier to express the exponential utility in terms of the Coefficient of Absolute Risk Tolerance (CART) measured in the same units as wealth. In addition, risk tolerance has an intuitive meaning not shared by risk aversion. In that case, the implicit assumption is that the initial wealth is 1 unit.

Among the utility functions that do not fit into the Hara class, is the quadratic utility function. If the percentage of wealth invested in risky assets increases with wealth, the investor can consider the quadratic utility function. Taking $B = -1$, $A > 0$ in Equation (A.7.5) we get

$$u(x) = x - \frac{1}{2A}x^2 \, , x \in \mathbb{R} \tag{A.7.6}$$

which only has increasing marginal utility when

$$u^{'}(x) = 1 - \frac{1}{A}x > 0 \text{ that is } x < A$$

so that the domain for a quadratic utility is restricted. Writing $a = \frac{1}{2A}$, the CRRA becomes

$$R(W) = \frac{2aW}{1 - 2aW}$$

and we get

$$R^{'}(W) = \frac{2a}{(1 - 2aW)^2} > 0$$

so that the quadratic utility function has increasing relative aversion, which implies the ARA must also be increasing. Therefore, a risk averse investor with a quadratic utility will increase the percentage of his wealth invested in risky assets as his wealth increases.

***Remark* A.7.1** *The quadratic utility function corresponds to the third strictly concave function in Appendix (A.1.2) with* $b = 1$ *and* $a = \frac{1}{2A}$.

This function decreases over part of the range, violating the assumption that investors desire more wealth (and so have an increasing utility function), but has excellent tractability properties.

A.7.4 Mean-variance criterion

A.7.4.1 Normal returns

We assume that an investor has an exponential utility function with a CARA γ given by

$$u(W) = -e^{-\gamma W}\ , \gamma > 0$$

and we further assume that the returns on a portfolio are normally distributed with expectation μ and standard deviation σ. Hence, the certain equialent of the portfolio can be approximated as

$$CE \approx \mu - \frac{1}{2}\gamma\sigma^2$$

The expected portfolio return is

$$\mu = w^\top E[r]$$

where w is the vector of portfolio weights and r is the vector of returns on the constituent assets, and the portfolio variance is

$$\sigma^2 = w^\top Q w$$

where Q is the covariance matrix of the asset returns. Since the best investment is the one giving the maximum certain equivalent, then for an investor with an exponential utility function investing in risky assets with normally distributed returns, the optimal allocation problem may be approximated by the simple optimisation

$$\max_w \left(\mu - \frac{1}{2}\gamma\sigma^2\right) = \max_w \left(w^\top E[r] - \frac{1}{2}\gamma w^\top Q w\right)$$

which is the mean-variance criterion. Note, when the utility is defined on the returns on the investment, we must multiply the utility function by the amount invested to find the utility of each investment. Similarly, to find the certain equivalent of a risky investment we multiply by the amount invested. However, we need to express the CRA γ as a proportion of the amount invested.

A.7.4.2 Non-normal returns

Assuming that investors borrow at zero interest rate, we consider the utility associated with an investment as being defined on the distribution of investment returns rather than on the distribution of the wealth arising from the investment. We further assume that the returns are non-normally distributed. Applying the expectation operator to a Taylor expansion of $u(R)$ about $u(\mu)$, the utility associated with the mean return, we get

$$E[u(R)] = u(\mu) + u^{'}(R)|_{R=\mu}E[R-\mu] + \frac{1}{2}u^{''}(R)|_{R=\mu}E[(R-\mu)^2] + \frac{1}{6}u^{'''}(R)|_{R=\mu}E[(R-\mu)^3] + ...$$

which is a simple approximation to the certain equivalent associated with any utility function since $E[u(R)] = E[u(X)] = u(CE(X))$. If we assume that the investor has an exponential utility function, then, the above equation to the fourth order becomes

$$e^{-\gamma CE} \approx e^{-\gamma\mu}\left(1 + \frac{1}{2}\gamma^2 E[(R-\mu)^2] - \frac{1}{6}\gamma^3 E[(R-\mu)^3] + \frac{1}{24}\gamma^4 E[(R-\mu)^4]\right)$$

Given the first four moments are

$$\sigma^2 = E[(R-\mu)^2]\ , S = E[(R-\mu)^3]\ , K = E[(R-\mu)^4]$$

where S is the skew and K is the kurtosis, we get the approximation

$$e^{-\gamma CE} \approx e^{-\gamma\mu}\big(1 + \frac{1}{2}(\gamma\sigma)^2 - \frac{S}{6}(\gamma\sigma)^3 + \frac{K-3}{24}(\gamma\sigma)^4\big)$$

where $K-3$ is the excess kurtosis (see Section (3.3.4.2)). Taking the logarithm and using the second order Taylor expansion, the approximated certain equivalent associated with the exponential utility function simplifies to

$$CE \approx \mu - \frac{1}{2}\gamma\sigma^2 + \frac{S}{6}\gamma^2\sigma^3 - \frac{K-3}{24}\gamma^3\sigma^4$$

The mean-variance criterion is a special case of the above equation with no skewnes and no kurtosis. In general, a risk averse investor having an exponential utility has aversion to risk associated with increasing variance, negative skewness, and increasing kurtosis.

A.8 Optimisation

Definition A.8.1 *We call numerical function of n real variables an application f of a set E of $\mathbb{R}^n$ in $\mathbb{R}$. The image of a point M, of coordinates $x_1, x_2, .., x_n$ is a real number, called $f(M)$ or $f(x_1, x_2, .., x_n)$.*

We call open ball of $\mathbb{R}^n$, of center $A = (a_1, a_2, .., a_n)$, $(A \in \mathbb{R}^n)$, and of radius r, $(r \in \mathbb{R}_+^*)$, the set

$$B_r(A) = \{M | M \in \mathbb{R}^n \text{ and } ||M - A|| < r\}$$

where $||M - A|| = \sqrt{(x_1 - a_1)^2 + ... + (x_n - a_n)^2}$ is the norm of $M - A$, also called the distance between M and A. We call neighbourhood of a point A all set $V(A)$ containing an open ball of centre A. $B_r(A)$ is a neighbourhood of each of his points, and in particular A.

We let f be a function of n real variables defined on the open set Ω of $\mathbb{R}^n$. One say that f has a local extremum at the point M_0 of Ω if there exists a neighbourhood $V(M_0)$ of M_0 such that for all $M \in V(M_0) \cap \Omega$, then $f(M) - f(M_0)$ keep a constant sign. If $f(M) - f(M_0) \geq 0$ (respectivly ≤ 0), it is a local minimum (respectivly maximum). If $f(M) - f(M_0)$ keeps a constant sign for all $M \in \Omega$, it is an absolute (or global) extremum.

Theorem A.8.1 *If f is continuously partially differentiable on Ω, for f to have an extremum at the point M_0 of Ω, it is necessary, but not sufficient that all the partial derivatives cancel on that point*

$$\forall i \in \{1, 2, .., n\}\ ,\ f'_{x_i}(M_0) = 0$$

A point M_0 where all the partial derivatives cancel is called a stationary point. For example, we consider a function f with $n = 2$ being continuously partially differentiable in the neibourhood of a stationary point M_0. From Taylor expansion, $f(x_0 + h, y_0 + k) - f(x_0, y_0)$ will be of the sign of

$$h^2 f''_{x^2}(x_0, y_0) + 2hk f''_{xy}(x_0, y_0) + k^2 f''_{y^2}(x_0, y_0)$$

when h and k will be in the neighborhoud of 0. Using Monge notation

$$r_0 = f''_{x^2}(x_0, y_0)\ ,\ s_0 = f''_{xy}(x_0, y_0)\ ,\ t_0 = f''_{y^2}(x_0, y_0)$$

we get the sufficient second order conditions

1. if $r_0 t_0 - s_0^2 > 0$, there is an extremum in M_0 which is a minimum if $r_0 > 0$ (or $t_0 > 0$), a maximum if $r_0 < 0$ (or $t_0 < 0$).
2. if $r_0 t_0 - s_0^2 < 0$, there is no extremum in M_0, and one says that M_0 is a col point.

3. if $r_0t_0 - s_0^2 = 0$, one can not conclude. One need to write the Taylor expension at a higher order, or directly study the sign of $f(x_0 + h, y_0 + k) - f(x_0, y_0)$ when h and k vary in the neighbourhood of 0.

Assuming a function f with n variables being continuously partially differentiable on the open set Ω of $mathbbR^n$, we now want to obtain the extremums of f, where the variables $x_1, x_2, .., x_n$ are linked by the constraint

$$g(x_1, x_2, .., x_n) = 0$$

such that g is also continuously partially differentiable on Ω. We distinguish two special cases

1. if the constraint g allows for one variable to be expressed in terms of the other $(n-1)$ variables, we recover the problem of obtaining the extremum of a function f with $(n-1)$ variables.

2. if the constraint g can be parametrised, that is, we can express $x_1, x_2, .., x_n$ in terms of the same real parameter t, we recover the problem of obtaining the extremum of a function f with a single variable

$$F(t) = f\big(x_1(t), x_2(t), .., x_n(t)\big)$$

In the general case, the previous problem is equivalent to that of finding the extremums of the function L, called the lagrangian

$$L(x_1, x_2, .., x_n, \lambda) = f(x_1, x_2, .., x_n) + \lambda g(x_1, x_2, .., x_n)\ , \lambda \in \mathbb{R}$$

where λ is the Lagrange multiplicator. The necessary conditions of the first order will allow us to determine the stationary points $\hat{M}(\hat{x}_1, \hat{x}_2, .., \hat{x}_n)$ which will be the only points where there can be an extremum. In order to get the nature of these points we need to study the sign of

$$f(\hat{x}_1 + h_1, \hat{x}_2 + h_2, .., \hat{x}_n + h_n) - f(\hat{x}_1, \hat{x}_2, .., \hat{x}_n)$$

where the variables $(h_1, .., h_n)$ varies in the neighbourhood of 0, and are linked by the constraint

$$g(\hat{x}_1 + h_1, \hat{x}_2 + h_2, .., \hat{x}_n + h_n) = 0$$

Note, if we use the Taylor expension to study the sign of the previous difference, then the terms of the first order in $(h_1, .., h_n)$ will not be null, and the variables $(h_1, .., h_n)$ are not independent. In the previous example of the function f with $n = 2$ we get the determinant

$$\Delta_3 = \begin{bmatrix} L''_{x^2} & L''_{xy} & g'_x \\ L''_{xy} & L''_{y^2} & g'_y \\ g'_x & g'_y & 0 \end{bmatrix}$$

If M_0 is a stationary point, then

1. $\Delta_3 < 0$, M_0 is a minimum

2. $\Delta_3 > 0$, M_0 is a maximum

In the general case, if there are p, $(p \geq 2)$ constraints $g_j(x_1, x_2, .., x_n) = 0$, $j = 1, .., p$, we introduce p Lagrange multiplicators $\lambda_1, .., \lambda_p$ and we find the extremums of the Lagrangian

$$L(x_1, x_2, .., x_n, \lambda_1, \lambda_2, .., \lambda_p) = f(x_1, x_2, .., x_n) + \sum_{j=1}^{p} \lambda_j g_j(x_1, x_2, .., x_n)$$

A.9 Conjugate gradient method

Given a single function f depending on one or more independent variables, we want to find the value of those variables where f takes on a maximum or a minimum. Since an extremun can be either global or local, finding a global extremum is a very difficult problem (see Press et al. [1992]).

- One approach consits in finding local extrema starting from a widely varying starting values of the independent variables, and then picking the most extreme of these.
- Another approach consists in perturbating a local extremum by taking a finite amplitute step away from it, and then see if the routine returns a better point, or always the same one.

In multidimensions, starting at a point P in N-dimensional space and proceeding from there in some vector direction n, the line methods consist in minimising $f(P)$ along the line n by a one-dimensional methods, obtaining sequences of such line minimisations. At each stage, different methods will only differe by the way they choose the next direction n to try. One can can construct a black-box subalgorithm as follow:

```
Begin
Given as input the vectors P and n, and the function f,
find the scalar λ minimising f(P + λ n)
Replace P by P + λ n
Replace n by λ n
End
```

Taking the unit vector $e_0, e_1, .., e_{N-1}$ as a set of directions, and using the above subalgorithm, we move along the first direction to its minimum, then from there along the second direction to its minimum, and so on, cycling through the whole set of directions as many times as necessary, until the function stops decreasing. However, if the second derivatives of the function are much larger in magnitude in some directions than in others, then many cycles through all N basis vectors will be necessary. Hence, we need a better set of directions than the e_i's that either

1. includes some very good directions taking us far along narrow valleys, or else
2. includes some number of noninterfering directions (called conjugate directions) with the property that minimisation along one is not spoiled by subsequent minimisation along another.

Note, if we minimise a function along some direction u, then the gradient of the function must be perpendicular to u at the line minimum, otherwise there would still be a nonzero directional derivative along u. Given a point P of dimension N as the origin of the coordinate system with coordinates X, then any function f can be approximated by its Taylor series

$$\begin{aligned} f(X) &= f(P) + \sum_i \frac{\partial f}{\partial x_i} x_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} x_i x_j + ... \\ &\approx c - b.X + \frac{1}{2} X.A.X \end{aligned}$$

where

$$c = f(P) \, , b = -\nabla f|_P \, , [A]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}\Big|_P$$

and ∇ denotes a differential operator that indicates taking gradient in vector calculus. The matrix A, whose components are the second partial derivative matrix of the function, is called the Hessian matrix of the function at point P. In the above approximation, the gradient of f is easily calculated as

$$\nabla f = A.X - b$$

implying that the function is at an extremum at a value of X obtained by solving

$$A.x = b$$

The change of the gradient as we move along some direction is given by

$$\delta(\nabla f) = A.(\delta X)$$

If we have moved along some direction u to a minimum and want to move along a new direction v without spoiling the minimisation along u, then the change in the gradient must stay perpendicular to u, that is,

$$0 = u.\delta(\nabla f) = u.A.v$$

In that case, the vectors u and v are said to be conjugate. Doing successive line minimisations of a function along a conjugate set of directions, then we do not need to redo any of those directions. Powell first discovered a direction set method producing N mutually conjugate directions.

```
Begin
Initialise the set of directions u_i to the basis vectors,
u_i = e_i  i = 0,..,N-1
Repeat the following sequence of steps until the function stop decreasing:
Save the starting position as P_0 .
For i = 0,...,N-1 , move P_i to the minimum along direction
u_i and call this point P_i+1 .
For i=0, ..., N-2 , set u_i ← u_i+1 .
Set u_N-1 ← P_N - P_0 .
Move P_N to the minimum along direction u_N
and call this point P_0 .
End
```

Powell showed that for a quadratic form, k iterations of the above algorithm produce a set of directions u_i whose last k members are mutually conjugate. Hence, N iterations of thisalgorithm, amounting to $N(N+1)$ line minimisations, will exactly minimise a quadratic form. However, the procedure of throwing away, at each stage, u_0 in favour of $P_N - P_0$ tends to produce sets of directions that fold up on each other and become linearly dependent. Fortunately, there are a number of ways to fix up this problem of linear dependence.

We are now going to describe the conjugate gradient method (see Press et al. [1992]). Considering the approximation of the function f with the Taylor series described above, the number of unknown parameters in f is equal to the number of free parameters in A and B, that is, $\frac{1}{2}N(N+1)$, which is of order N^2. Changing any one of these parameters can move the location of the minimum. A simple algorithm can be described as follow

```
Begin
Start at point P_0
Move from point P_i to the point P_i+1 by minimising
along the line from P_i in the direction of the local downhill
gradient -∇ f(P_i) .
Repeat until convergence.
End
```

This methods will perform many small steps in going down a long, narrow valley, even if the valley is a perfect quadratic form. Since the new gradient at the minimum point of any line minimisation is perpendicular to the direction just traversed, then with the steepest descent method we must make a right angle turn, which does not lead to the minimum. Hence, we want to follow a direction that is constructed to be conjugate to the old gradient, and if possible to all previous directions traversed. This method is called the conjugate gradient methods (CGM). We now introduce the Fletcher-Reeves version of the CGM. Starting with an arbitrary initial vector g_0 and letting $h_0 = g_0$, the CGM constructs two sequences of vectors from the recurrence

$$g_{i+1} = g_i - \lambda_i A.h_i \ , \ h_{i+1} = g_{i+1} + \gamma_i h_i \ , \ i = 0, 1, ...$$

and the vectors satisfy the orthogonality and conjugacy conditions

$$g_i.g_j = 0 \ , \ h_i.A.h_j = 0 \ , \ g_i.h_j = 0 \ , \ j < i$$

The scalars λ_i and γ_i are given by

$$\begin{aligned} \lambda_i &= \frac{g_i.g_i}{h_i.A.h_i} \\ \gamma_i &= \frac{g_{i+1}.g_{i+1}}{g_i.g_i} \end{aligned}$$

If we knew the Hessian matrix A we could find conjugate directions h_i along which to line-minimise, and after N computations, arrive at the minimum of the quadratic form. But in practice we do not know the matrix A. A way around is to compute

$$g_i = -\nabla f(P_i)$$

for some point P_i. Then proceed from P_i along the direction h_i to the local minimum of f located at some point P_{i+1}, and then set $g_{i+1} = -\nabla f(P_{i+1})$. Then this g_{i+1} is the same vector as would have been constructed from the above equation if the matrix A was known. Hence, a sequence of directions h_i is constructed by using line minimisations, evaluations of the gradient vector, and an auxiliary vector to store the latest in sequence of g's. Polack and Ribiere introduced a significant change by modifying γ_i as follow

$$\gamma_i = \frac{(g_{i+1} - g_i).g_{i+1}}{g_i.g_i}$$

Appendix B

Some probabilities

For details see text books by Grimmett et al. [1992], Oksendal [1998] and Jacod et al. [2004].

B.1 Some definitions

Definition B.1.1 *The set of all possible outcomes of an experiment is called the sample space and is denoted* Ω.

Definition B.1.2 *An event is a property which can be observed either to hold or not to hold after the experiment is done. In mathematical terms, an event is a subset of* Ω.

We think of the collection of events as a subcollection $\mathcal{F}$ of the set of all subsets of Ω such that

1. if $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
3. the empty set $\oslash$ belongs to $\mathcal{F}$

Any collection $\mathcal{F}$ of subsets of Ω which satisfies these three conditions is called a field. It follows from the properties of a field $\mathcal{F}$ that

$$\text{if } A_1, A_2, .., A_n \in \mathcal{F} \text{ then } \bigcup_{i=1}^{n} A_i \in \mathcal{F}$$

so that $\mathcal{F}$ is closed under finite unions and hence under finite intersections also.

Definition B.1.3 *A collection* $\mathcal{F}$ *of subsets of* Ω *is called a* σ*-field if it satisfies the following conditions*

1. *the empty set* $\oslash$ *belongs to* $\mathcal{F}$
2. *if* $A_1, A_2, ... \in \mathcal{F}$ *then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
3. *if* $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$

We consider a set E in $\mathbb{R}^d$ and let the measure μ on E associate to some measurable subsets $A \subset E$ be a positive number $\mu(A) \in [0, \infty]$ called the measure of A. The domain of definition of a measure on E is a collection of subsets of E called a σ -algebra which contains the empty set, is stable under unions and contains the complementary of every element. We define the counting measure $\mu_X = \sum_i \delta_{x_i}$ on a countable set of points $X = \{x_i, i = 0, 1, ...\} \subset E$

where $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \neq A$ is a Dirac measure such that for any $A \subset E$, $\mu_X(A)$ counts the number of points x_i in A

$$\mu(A) = \#\{i\ ,\ x_i \in A\} = \sum_{i \geq 1} I_{x_i \in A}$$

It is an integer valued measure. A finite measure with mass 1 is called a probability measure.

Definition B.1.4 *Let $E \subset \mathbb{R}^d$. A Radon measure on $(E, \mathcal{B})$ is a measure μ such that for every compact measurable set $B \in \mathcal{B}$, $\mu(B) < \infty$.*

A measure μ_0 which gives zero mass to any point is said to be diffusive or atomless, that is $\forall x \in E, \mu_0(\{x\}) = 0$. Any Radon measure can be decomposed into a diffusive part and a sum of Dirac measures

***Proposition* 8** *Any Radon measure μ can be decomposed into a diffusive part μ_0 and a linear combination of Dirac measures*

$$\mu = \mu_0 + \sum_{j \geq 1} b_j \delta_{x_j}\ x_j \in E, b_j > 0$$

We can now look at measurable functions

Definition B.1.5 *We consider two measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, then a function $f : E \to F$ is called measurable if for any measurable set $A \in \mathcal{F}$, the set*

$$f^{-1}(A) = \{x \in E, f(x) \in A\}$$

is a measurable subset of E.

If the measure μ can be decomposed as in Proposition (8) then the integral of μ with respect to f denoted by $\mu(f)$ is

$$\mu(f) = \int f(x)\mu_0(dx) + \sum_{j \geq 1} b_j f(x_j)$$

We let Ω be the set of scenarios equipped with a σ-algebra $\mathcal{F}$ and consider a probability measure on $(\Omega, \mathcal{F})$ which is a positive finite measure $\mathbb{P}$ with total mass 1. Therefore, $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space and any measurable set $A \in \mathcal{F}$ called an event is a set of scenarios to which a probability can be assigned. The probability measure assigns value in $[0, 1]$ to each event such that

$$\begin{aligned} \mathbb{P} : \quad \mathcal{F} &\to [0, 1] \\ A &\to \mathbb{P}(A) \end{aligned}$$

An event A with probability $\mathbb{P}(A) = 1$ is said to occur almost surely and if $\mathbb{P}(A) = 0$ the event is impossible. We will say that a property holds $\mathbb{P}$-almost surely if the set of $\omega \in \Omega$ for which the property does not hold is a null set (subset of an impossible event). Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ are equivalent if they define the same impossible events

$$\mathbb{P} \sim \mathbb{Q} \Longleftrightarrow [\forall A \in \mathcal{F}, \mathbb{P}(A) = 0 \Longleftrightarrow \mathbb{Q}(A) = 0]$$

A random variable X taking values in E is a measurable map $X : \Omega \to E$ where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. $X(\omega)$ represents the outcome of the random variable if the scenario ω happens and is called the realisation of X in the scenario ω. The law of X is the probability measure on E defined by $\mu_X(A) = \mathbb{P}(X \in A)$.

B.2 Random variables

Definition B.2.1 *A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.*

Definition B.2.2 *The distribution function of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by*

$$F(x) = P(X \leq x)$$

B.2.1 Discrete random variables

Definition B.2.3 *The random variable X is called discrete if it takes values in some countable subset $\{x_1, x_1, ..\}$, only, of $\mathbb{R}$.*

Its distribution function $F(x) = P(X \leq x)$ is a jump function.

Definition B.2.4 *The probability mass function of a discrete random variable X is the function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = P(X = x)$.*

The distribution and mass functions are related by

$$F(x) = \sum_{i:x_i \leq x} f(x_i)$$

Lemma B.2.1 *The probability mass function $f : \mathbb{R} \to [0, 1]$ satisfies*

1. *$f(x) \neq 0$ if and only if x belongs to some countable set $\{x_1, x_2, ..\}$*
2. *$\sum_i f(x_i) = 1$*

Let $x_1, x_2, .., x_N$ be the numerical outcomes of N repetitions of some experiment. The average of these outcomes is

$$m = \frac{1}{N} \sum_i x_i$$

In advance of performing these experiments we can represent their outcomes by a sequence $X_1, X_2, .., X_N$ of random variables, and assume that these variables are discrete with a common mass function f. Then, roughly speaking, for each possible value x, about $Nf(x)$ of the X_i will take that value x. So, the average m is about

$$m \approx \frac{1}{N} \sum_x xNf(x) = \sum_x xf(x)$$

where the summation is over all possible values of the X_i. This average is the expectation or mean value of the underlying distribution with mass function f.

Definition B.2.5 *The mean value or expectation of X with mass function f is defined to be*

$$E[X] = \sum_{x:f(x)>0} xf(x)$$

whenever this sum is absolutely convergent.

Definition B.2.6 *If k is a positive integer, then the kth moment m_k of X is*

$$m_k = E[X^k]$$

The kth central moment σ_k is

$$\sigma_k = E[(X - m_1)^k]$$

The two moments of most use are $m_1 = E[X]$ and $\sigma_2 = E[(X - E(X))^2]$ called the mean and variance of X.

B.2.2 Continuous random variables

Definition B.2.7 *The random variable X is called continuous if its distribution function can be expressed as*

$$F(x) = \int_{-\infty}^{x} f(u)du \ , \ x \in \mathbb{R}$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$.

The expectation of a discrete variable X is $E[X] = \sum_x xP(X = x)$ which is an average of the possible values of X, each value being weighted by its probability. For continuous variables, expectations are defined as integrals.

Definition B.2.8 *The expectation of a continuous random variable X with density function f is*

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

whenever this integral exists.

We shall allow the existence of $\int g(x)dx$ only if $\int |g(x)|dx < \infty$. Note, the definition of the kth moment m_k in Appendix (B.2.1) applies to continuous random variables, but the moments of X may not exist since the integral

$$E[X^k] = \int x^k f(x)dx$$

may not converge.

B.3 Introducing stochastic processes

A martingale process is defined on the basis of semi-martingales.

Definition B.3.1 *A random process $(X_t)_{t>0}$ is called a submartingale if*

$$E[|X_t|] < \infty$$

and

$$E[X_t|\mathcal{F}_s] \geq X_s \ , \ s < t$$

a supermartingale if, instead

$$E[X_t|\mathcal{F}_s] \leq X_s \ , \ s < t$$

and it is a martingale if the process is both a submartingale and a supermartingale.

We now provide the definitions of a Markov process and that of an independent process.

Definition B.3.2 *A random process $(X_t)_{t>0}$ is called a a Markov process if, for each n and every $i_0, .., i_n$, then*

$$P(X_{n+1} = j|X_0 = i_0, ..., X_n = i_n) = P(X_{n+1} = j|X_n = i_n)$$

where $P(\bullet|\bullet)$ denotes conditional probability.

Definition B.3.3 *We let $\{X_t; t = 1, 2, ...\}$ be a sequence of random variables on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $E[X_t = 0]$ and $\{\mathcal{F}_t; t = 1, 2, ...\}$ a current of σ-algebras on the measurable space $(\Omega, \mathcal{F})$, where Ω is the complete universe of all possible events. Then $\{X_t\}$ is a sequence of independent random variables with respect to $\{\mathcal{F}_t\}$ if X_t is measurable with respect to $\mathcal{F}_t$ and is independent of $\mathcal{F}_{t-1}$ for all $t = 1, 2, ...$*

We can now define a random walk (RW) process and geometric Brownian motion (GBM).

Definition B.3.4 *A random walk is a Markov process with independent innovations*

$$X_t - X_{t-1} = \epsilon_t$$

where $\epsilon_t \approx IID$, standing for independent and identically distributed process.

Definition B.3.5 *A geometric Brownian motion is a random walk of natural logarithm of the original process X_t, where $L_t = \ln(X_t)$, so that*

$$\Delta L_t = L_t - L_{t-1} = \epsilon_t$$

where $\epsilon_t \approx IID$.

Note, martingale is more general than random walk since semi-martingales allow for dependence in the process. Thus, random walk implies martingale but martingale does not imply random walk in the process.

B.4 The characteristic function, moments and cumulants

B.4.1 Definitions

We start by recalling some definitions together with the properties of the characteristic functions. The characteristic function of a random variable is the Fourier transform of its distribution

Definition B.4.1 *The characteristic function of a $\mathbb{R}^d$ random variable X is the function $\Phi_X : \mathbb{R}^d \to \mathbb{C}$ defined by*

$$\Phi_X(z) = E[e^{iz.X}] = \int_{\mathbb{R}^d} e^{iz.x} d\mu_X(x) \text{ for } z \in \mathbb{R}^d$$

where μ_X is the measure of X.

The characteristic function of a random variable completely characterises its law. Smoothness properties of Φ_X depend on the existence of moments of the random variable X which is related on how fast the distribution μ_X decays at infinity. If it exists, the n-th moment m_n of a random variable X on $\mathbb{R}$ is

$$m_n = E[X^n]$$

The first moment of X called the mean or expectation measures the central location of the distribution. Denoting the mean of X by μ_X, the nth central moment of X, if it exists, is defined as

$$m_n^c = E[(X - \mu_X)^n]$$

The second central moment σ_X^2 called the variance of X measures the variability of X. The third central moment measures the symmetry of X with respect to its mean, and the fourth central moment measures the tail behaviour of X. In statistics, skewness and kurtosis, respectively the normalised third and fourth central moments of X are used to summarise the extent of asymmetry and tail thickness. The are defined as

$$S = \frac{m_3^c}{(m_2^c)^{\frac{3}{2}}} = E[\frac{(X - \mu_X)^3}{\sigma_X^3}] \, , K = \frac{m_4^c}{(m_2^c)^2} = E[\frac{(X - \mu_X)^4}{\sigma_X^4}]$$

Since $K = 3$ for a normal distribution, the quantity $K - 3$ is called the excess kurtosis. The moments of a random variable are related to the derivatives at 0 of its characteristic function.

***Proposition* 9** *If $E[|X|^n] < \infty$ then Φ_X has n continuous derivatives at $z = 0$ and*

$$m_k = E[X^k] = \frac{1}{i^k} \frac{\partial^k \Phi_X}{\partial z^k}(0)$$

***Proposition* 10** *X possesses finite moments of all orders iff $z \to \Phi_X(z)$ is C^∞ at $z = 0$. Then the moments of X are related to the derivatives of Φ_X by*

$$m_n = E[X^n] = \frac{1}{i^n} \frac{\partial^n \Phi_X}{\partial z^n}(0)$$

If X_i with $i = 1, .., n$ are independent random variables, the characteristic function of $S_n = X_1 + ... + X_n$ is the product of characteristic functions of individual variables X_i

$$\Phi_{S_n}(z) = \prod_{i=1}^{n} \Phi_{X_i}(z) \tag{B.4.1}$$

We see that $\Phi_X(0) = 1$ and that the characteristic function Φ_X is continuous at $z = 0$ and $\Phi_X(z) \neq 0$ in the neighborhood of $z = 0$. It leads to the definition of the cumulant generating function or log characteristic function of X.

Definition B.4.2 *There exists a unique continuous function Ψ_X called the cumulant generating function defined around zero such that*

$$\Psi_X(0) = 0 \text{ and } \Phi_X(z) = e^{\Psi_X(z)}$$

The cumulants k_n of a probability distribution are a set of quantities providing an alternative to the moments of the distribution. It is defined via the cumulant-generating function $\Psi_X(z)$, which is the natural logarithm of the moment generating function

$$\Psi_X(z) = \ln \Phi_X(z)$$

The cumulants k_n are obtained from the power series expansion of the cumulant generating function

$$\Psi_X(z) = \sum_{n=1}^{\infty} k_n \frac{z^n}{n!}$$

so that the nth cumulant can be obtained by differentiating the above equation n-times and evaluating the result at zero

$$k_n = \Psi_X^{(n)}(z)\big|_{z=0}$$

B.4.2 The first two moments

We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let X and Y be two discrete random variables. Given the definition of the kth moment m_k and the kth central moment σ_k in Appendix (B.2.1) we get

$$\begin{aligned} Cov(X,Y) &= E[XY] - E[X]E[Y] = E\big[(X - E[X])(Y - E[Y])\big] \\ E[XY] &= \sum_{x,y} xy f_{XY}(x,y) \\ Var(X) &= Cov(X,X) = E[X^2] - (E[X])^2 = E\big[(X - E[X])^2\big] \\ \rho(X,Y) &= \frac{Cov(X,Y)}{\big(Var(X)Var(Y)\big)^{\frac{1}{2}}} \end{aligned}$$

For $\rho(X,Y) = 0$ we must have $Cov(X,Y) = 0$ which leads to

$$E[XY] = E[X]E[Y]$$

Moreover, if X and Y are independent, we get

$$Var(X+Y) = Var(X) + Var(Y)$$

Otherwise, if they are correlated, that is $\rho(X,Y) \neq 0$ we set $Z = X + Y$ and plug it back into the variance equation

$$\begin{aligned} Var(Z) &= E[Z^2] - (E[Z])^2 = E[(X+Y)^2] - (E[X+Y])^2 \\ &= E[X^2 + Y^2 + 2XY] - \big(E[X] + E[Y]\big)^2 \\ &= E[X^2] + E[Y^2] + E[2XY] - (E[X])^2 - (E[Y])^2 - 2E[X]E[Y] \\ &= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 + 2\big(E[XY] - E[X]E[Y]\big) \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

More generally, for n random variables $X_1, ..., X_n$ the variance becomes

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j}^{n} Cov(X_i, X_j)$$

B.4.3 Trading correlation

We are going to briefly explain the difference between the correlation obtained from the dispersion trades and the correlation obtained from the correlation swap. We consider a basket made of n-underlying X_i for $i = 1, .., n$ with volatility given by σ_i. In the dispersion trades, the correlation comes from the composition of the basket, that is

$$\begin{aligned} Var(Indice) &= \sum_{i=1}^{n} w_i^2 Var(X_i) + 2\sum_j \sum_{i<j} w_i w_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} w_i^2 \sigma_i^2 + 2\sum_j \sum_{i<j} w_i w_j \rho_{ij} \sigma_i \sigma_j \end{aligned}$$

where w_i is the weight of the ith-stock in the basket and ρ_{ij} is the correlation between the ith-stock and the jth-stock in the basket. The problem is that the correlations between the stocks vary with the market volatility. One way to look at it is to study the derivative with respect to the correlation, that is,

$$\frac{\partial Var(Indice)}{\partial \rho_{ij}} = 2\sum_j \sum_{i<j} w_i w_j \sigma_i \sigma_j$$

which depends on the volatility of the i-th stock and the j-th stock. Hence, it is not a pure correlation product. On the contrary, a correlation swap with the payoff at maturity being $(\rho - K)$ plays directly with the realised correlation. It is a pure correlation product since its derivative with respect to the correlation is 1, that is, $\frac{\partial RS(t,T)}{\partial \rho} = 1$.

B.5 Introduction to subordinated stochastic processes

Rather than indexing discrete stochastic processes with integers $X(0), X(1), .., X_t, X_{t+1}$, the process can be indexed by a set of numbers $t_1, t_2, ...$ where these numbers are themselves a realisation of a stochastic process with positive increments, so that $t_1 \leq t_2 \leq ...$ (see Bochner [1960], Feller [1971]). Hence, if $T(t)$ is a positive stochastic process, then a new process $X(T(t))$ may be formed which is said to be subordinated to $X(t)$. $T(t)$ is called the directed process, and the distribution of $\Delta X(T(t))$ is said to be subordinate to the distribution of $\Delta X(t)$. The following theorem holds for very general classes of subordinated stochastic process with independent increments. It provides a simple formula for calculating the variance of the increments, and shows that the variance is finite for processes having increments with finite variance.

Theorem B.5.1 *Let $X(t)$ and $T(t)$ be processes with stationary independent increments. Let the increments of $X(t)$ be drawn from a distribution with mean 0 and finite variance σ^2. Let the increments of $T(t)$ be drawn from a positive distribution with mean α, independent of the increments of $X(t)$. Then, the subordinated stochastic process $X(T(t))$ has stationary independent increments with mean 0 and variance $\alpha\sigma^2$.*

It says that if the directing process has a finite mean, then $\Delta X(T)$ will have a finite variance unless ΔT does not. Following Clark [1973], we show some results on the limit distribution of a random sum of random variables

Theorem B.5.2 *Central Limit Theorem*
Let $\{Y_i\}$ be a sequence of i.i.d. random variables with mean 0 and variance 1. Let $S_n = \sum_{i=1}^n Y_i$. Then the distribution of $\frac{S_n}{\sqrt{n}}$ tends to the unit normal distribution.

It can be generalised to the case where the number of terms, n, in the sum S_n is itself a random variable.

Theorem B.5.3 *We let $\{Y_i\}$ be distributed as in the above Theorem, and we let*

$$S_{N_n} = \sum_{i=1}^{N_n} Y_i$$

Let $N_n = [Z_n]$ for large n, where Z is a random variable with mean 1, again independent of $\{Y_i\}$. Then, $\frac{S_{N_n}}{\sqrt{n}}$ has

$$f(u) = \frac{1}{\sqrt{2\pi Z}} e^{-\frac{u^2}{2Z}}$$

as its density.

Corollary 5 *If $X(t)$ is normal with stationary independent increments, and $T(t)$ has stationary independent positive increments with finite second moment which are independent of X, then the kurtosis, k, of the increments of $X(T(t))$ is an increasing function of the variance of the increments of $T(t)$.*

This Corollary is directly applicable to the limit distributions described in the above Theorem since the limit distribution of a random sum of random variables obeying the Central Limit Theorem is asymptotically normal with random variance, or subordinate to the normal distribution.

As a special case, of the subordinate distributions, consider a process $X(t)$ whose independent increments $\Delta X(t)$ are normally distributed, directed by a process $T(t)$, whose independent increments are lognormally distributed. We let $f(x;\mu,sigam_1^2)$ be the density, and define the mean of x as $\mu_x = e^{\mu+\frac{\sigma_1^2}{2}}$ and the variance of x as $\sigma_x^2 = e^{2\mu+\sigma_1^2}(e^{\sigma_1^2} - 1)$. The previous Theorem tells us that if $\Delta X(t)$ is normally distributed with mean 0 and variance σ_2^2, then the increments $\Delta X(T(t))$ of the lognormal-normal process have mean 0 and variance

$$\sigma^2_{\Delta X(T(t))} = \sigma_2^2\mu_x$$

Further, given $\mu + \frac{\sigma_1^2}{2}$ and increasing σ_1^2, the variance of the distribution stays constant while its kurtosis increases as much as desired.

Theorem B.5.4 *A random process subordinated to a normal process with independent increments distributed $N(0,\sigma_2^2)$ and directed by a lognormal with independent increments (and parameters μ and σ_1^2) has the following lognormal-rormal increments:*

$$f_{LNN}(y) = \frac{1}{2\pi\sigma_1^2\sigma_2^2}\int_0^\infty v^{-\frac{3}{2}} e^{-\frac{(\log v-\mu)^2}{2\sigma_1^2}} e^{-\frac{y^2}{2v\sigma_2^2}} dv$$

which may be approximated by numerical integration techniques.

B.6 Conditional moments

The related concept of conditional probability dates back from Laplace who calculated conditional distributions. Kolmogorov [1933] formalised it by using the Radon-Nikodym theorem. Halmos and Doob [1953] generalised the conditional expectation by using sub-sigma-algebras.

B.6.1 Conditional expectation

Let A and B be two events defined on a probability space. Let $f_n(A)$ denote the number of times A occurs divided by n. As n gets large $f_n(A)$ should be close to $P(A)$, that is, $\lim_{n\to\infty} f_n(A) = P(A)$. When computing $P(A|B)$ we do not want to count the occurrences of $A \cap B^c$ since we know B has occurred. Hence, counting only the occurrences of A where B also occurs, this is $nf_n(A\cap B)$. Now the number of trials is the number of occurrences of B (all other trials are discarded as impossible since B has occurred). Therefore, the number of relevant trials is $nf_n(B)$. Consequently we should have

$$P(A|B) \approx \frac{nf_n(A\cap B)}{nf_n(B)} = \frac{f_n(A\cap B)}{f_n(B)}$$

and taking limits in n motivates the definition of the conditional probability which is

Definition B.6.1 *Given A and B two events, if $P(B) > 0$ then the conditional probability that A occurs given that B occurs is*

$$P(A/B) = \frac{P(A\cap B)}{P(B)}$$

A family $B_1, B_2, ..., B_n$ of events is called a partition of Ω if

$$B_i \cap B_j = \oslash \text{ when } i \neq j \text{ and } \bigcup_{i=1}^{n} B_i = \Omega$$

Lemma B.6.1 *(Partition Equation) For any events A and B*

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

More generally, let $B_1, B_2, ..., B_n$ be a (finite or countable) partition of Ω. Then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

This may be set in the more general context of the conditional distribution of one variable Y given the value of another variable X (Discrete Case). Let X and Y be two random variables with Y taking values in $\mathbb{R}$ and with X taking only countably many values.

***Remark* B.6.1** *Suppose we know that the event $\{X = j\}$ for some value j has occurred. The expectation of Y may change given this knowledge.*

Indeed, if $Q(\Lambda) = P(\Lambda|X = j)$, it makes more sense to calculate $E^Q[Y]$ than it does to calculate $E^P[Y]$.

Definition B.6.2 *Let X have values in $(x_1, x_2, ..., x_n, ..)$ and Y be a random variable. Then if $P(X = x_j) > 0$ the conditional expectation of Y given $\{X = x_j\}$ is defined to be*

$$E[Y|X = x_j] = E^Q[Y]$$

where Q is the probability given by $Q(\Lambda) = P(\Lambda|X = j)$, provided $E^Q[|Y|] < \infty$.

Theorem B.6.1 *In the previous setting, and if further Y is countably valued with values $(y_1, y_2, ..., y_n, ..)$ and if $P(X = x_j) > 0$, then*

$$E[Y|X = x_j] = \sum_{k=1}^{\infty} y_k P(Y = y_k|X = x_j)$$

provided the series is absolutely convergent.

Proof:

$$E[Y|X = x_j] = E^Q[Y] = \sum_{k=1}^{\infty} y_k Q(Y = y_k) = \sum_{k=1}^{\infty} y_k P(Y = y_k|X = x_j)$$

Recall, $P(Y = y_k|X = x_j) = \frac{P(Y=y_k, X=x_j)}{P(X=x_j)}$.

Definition B.6.3 *The conditional distribution function of Y given $X = x$, written $F_{Y|X}(.|x)$ is defined by*

$$F_{Y|X}(y|x) = P(Y \leq y|X = x)$$

for any x such that $P(X = x) > 0$. The conditional mass function of Y given $X = x$, written $f_{Y|X}(.|x)$ is defined by

$$f_{Y|X}(y|x) = P(Y = y|X = x)$$

Next, still with X having at most a countable number of values, we wish to define the conditional expectation of any real valued r.v. Y given knowledge of the random variable X, rather than given only the event $\{X = x_j\}$. To this end we consider the function

$$f(x) = \begin{cases} E[Y|X = x] \text{ if } P(X = x) > 0 \\ \text{any arbitrary value if } P(X = x) = 0 \end{cases}$$

Definition B.6.4 *Let X be countably valued and let Y be a real valued random variable. The conditional expectation of Y given X is defined to be*

$$E[Y|X] = f(X)$$

where f is given above and provided that it is well defined.

Definition B.6.5 *Let $\psi(x) = E[Y|X = x]$. Then $\psi(x)$ is called the conditional expectation of Y given X, written $E[Y|X]$.*

***Remark* B.6.2** *Note, a conditional expectation is actually a random variable.*

Theorem B.6.2 *The conditional expectation $\psi(X) = E[Y|X]$ satisfies*

$$E[\psi(X)] = E[Y]$$

Using this theorem, we can compute the marginal mean $E[Y]$ as

$$E[Y] = \sum_x E[Y|X = x]P(X = x) \tag{B.6.2}$$

Thus, the marginal mean is the weighted average of the conditional means, with weights equal to the probability of being in the subgroup determined by the corresponding value of the conditioning variable. We can express the variance of the random variable Y in terms of conditional variance as

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$$

In the continuous case, we can compute $E[Y]$ as

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x]f_X(x)dx$$

Theorem B.6.3 *Let X and Y be random variables, and suppose that $E[Y^2] < \infty$. The best predictor of Y given X is the conditional expectation $E[Y|X]$.*

B.6.2 Conditional variance

Consider $E[Y|X]$ as a new random variable U as follows: Randomly pick an x from the distribution X so that the new r.v. U has the value $E[Y|X = x]$. For example, $Y = height$ and $X = sex$. Randomly pick a person from the population in question

$$U = \begin{cases} E[Y|X = female] \text{ if the person is female} \\ E[Y|X = male] \text{ if the person is male} \end{cases}$$

If U is a discrete r.v. the expected value of this new random variable is

$$E[U] = \sum_u P(u)u$$

and we get

$$E[E[Y|X]] = \text{ weighted average of conditional means } = E[Y]$$

The definition of the (population) (marginal) variance of a random variable Y is

$$Var(Y) = E[(Y - E[Y])^2]$$

Letting $\overline{Y} = E[Y]$ and expending the above term, we can rewrite the variance as

$$Var(Y) = E[Y^2 + (\overline{Y})^2 - 2Y\overline{Y}] = E[Y^2] + (\overline{Y})^2 - 2\overline{Y}E[Y] = E[Y^2] - (\overline{Y})^2$$

Similarly, if we are considering a conditional distribution $Y|X$, we define the conditional variance

$$Var(Y|X) = E\big[(Y - E[Y|X])^2|X\big]$$

Alternatively, we can write the conditional variance as

$$Var(Y|X) = E[Y^2|X] - (E[Y|X])^2 \tag{B.6.3}$$

Proof:
Letting $\psi(X) = E[Y|X]$ we can rewrite the conditional variance as

$$Var(Y|X) = E[Y^2|X] + (\psi(X))^2 - 2\psi(X)E[Y|X] = E[Y^2|X] - (\psi(X))^2$$

As with $E[Y|X]$, we can consider $Var(Y|X)$ as a random variable. Hence, we can compute the expected value of the conditional variance as

$$E[Var(Y|X)] = E[E[Y^2|X]] - E[(E[Y|X])^2]$$

Since the expected value of the conditional expectation of a random variable is the expected value of the original random variable the above equation simplifies to

$$E[Var(Y|X)] = E[Y^2] - E[(E[Y|X])^2] \tag{B.6.4}$$

We can also compute the variance of the conditional expectation as

$$Var(E[Y|X]) = E[(E[Y|X])^2] - (E[E[Y|X]])^2$$

and since $E[E[Y|X]] = E[Y]$ it simplifies to

$$Var(E[Y|X]) = E[(E[Y|X])^2] - (E[Y])^2 \tag{B.6.5}$$

Combining Equation (B.6.4) with Equation (B.6.5) we get

$$E[Var(Y|X)] + Var(E[Y|X]) = E[Y^2] - (E[Y])^2$$

which is the marginal variance $Var(Y)$. That is, the marginal variance is

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X]) \tag{B.6.6}$$

The marginal (overall) variance is the sum of the expected value of the conditional variance and the variance of the conditional means. Since variances are always non-negative, we get

$$Var(Y) \geq E[Var(Y|X)]$$

Further, since $Var(Y|X) \geq 0$ and $E[Var(Y|X)]$ must also be positive then

$$Var(Y) \geq Var(E[Y|X])$$

Note, $E[Var(Y|X)]$ is a weighted average of $Var(Y|X)$. We can also write the variance of the conditional expectation as

$$Var(E[Y|X]) = E\big[(E[Y|X] - E[Y])^2\big]$$

which is a weighted average of $(E[Y|X] - E[Y])^2$.

B.6.3 More details on conditional expectation

B.6.3.1 Some discrete results

We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let X be a discrete random variable taking values in the countable set $\{x_1, x_2, ..\}$. Its distribution function $F(x) = P(X \leq x)$ is a jump function with pdf $f(x) = P(X = x)$. We have

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

The expected value of X is

$$E[X] = \sum_{x:f(x)>0} xf(x)$$

A discrete random variable X_n can be defined as

$$X_n(\omega) = \sum_{i=1}^{n} x_i I_{A_i(\omega)}$$

where A_i is a partition of Ω. The conditional expectation of X_n given event B with $\mathbb{P}(B) > 0$ is

$$\begin{aligned} E[X_n|B] &= \int_\Omega X_n(\omega) d\mathbb{P}(\omega|B) = \sum_{i=1}^{n} x_i \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \frac{1}{\mathbb{P}(B)} \int_\Omega X_n(\omega) I_B(\omega) d\mathbb{P}(\omega) = \frac{E[X_n I_B]}{\mathbb{P}(B)} \end{aligned}$$

Letting n goes to infinity, that is partitioning finer and finer we get the continuous case

$$E[X|B] = \lim_{n\to\infty} E[X_n|B] = \frac{E[XI_B]}{\mathbb{P}(B)}$$

The Cauchy-Schwarz inequality states that for any X and Y we have

$$E[(XY)]^2 \leq E[X^2]E[Y^2] \tag{B.6.7}$$

with equality if and only if $P(aX = bY) = 1$. We let $\psi(x) = E[Y|X = x]$ and call $\psi(X) = E[Y|X]$ the conditional expectation of Y given X. It satisfies

$$E[\psi(X)] = E[Y]$$

As a result, another way of computing $E[Y]$ is

$$E[Y] = \sum_x E[Y|X = x]P(X = x) \tag{B.6.8}$$

B.6.3.2 Some continuous results

Generally, if B is a sufficiently nice subset of $\mathbb{R}$ then, for the variable X we get

$$P(X \in B) = \int_B f_X(x)dx$$

where $f_X(x)$ is the probability density function of X. We can then think of the pdf $f_X(x)dx$ as the element of the probability $P(X \in dx)$ since

$$P(X \in dx) \approx f_X(x)dx$$

A random variable X is continuous if its distribution function $F(x) = P(X \leq x)$ is

$$F(x) = \int_{-\infty}^{x} f_X(u)du$$

where $f_X(.)$ is the probability density function of X. If X and $g(X)$ are continuous random variables, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

Therefore, if we let $h(x) = g(x)f_X(x)$ we can use the properties of the Dirac function in Appendix (A.4) to get

$$\begin{aligned} E[g(X)\delta_x(X)] &= \int_{-\infty}^{\infty} h(y)\delta_x(y)dy = h(x) = g(x)f_X(x) \\ &= g(x)P(X = x) \end{aligned}$$

In the special case where $g(X) = 1$ we get $E[\delta_x(X)] = P(X = x)$. We can relate it to the definition of the conditional expectation in Equation (B.6.1) when $G(X_T) = X_T - x$

$$E[F(X_T)\delta_0(X_T - x)] = \int_{-\infty}^{\infty} F(y)\delta_0(y - x)f_X(y)dy = F(x)f_X(x)$$

and the conditional expectation becomes

$$E[F(X_T)|(X_T - x) = 0] = \frac{F(x)P(X_T = x)}{P(X_T = x)} = F(x)$$

B.7 About fractal analysis

B.7.1 The fractional Brownian motion

We define the line-to-line Brownian function $B(t)$ as a random function such that for all t and Δt

$$P\Big(\frac{[B(t + \Delta t) - B(t)]}{|\Delta t|^H} < x\Big) = erf(x)$$

where x is a real number, $erf(.)$ is the error function, and $H = \frac{1}{2}$. The function $B(t)$ is continuous but it is not differentiable. Each assumption can be generalised and every process obtained is significantly different from $B(t)$. The variation of the Brown line-to-line function $B(t)$ between $t = 0$ and $t = 2\pi$ decomposes into

1. the trend defined by

$$B^*(t) = B(0) + \frac{t}{2\pi}[B(2\pi) - B(0)]$$

2. and an oscillatory remainder $B_B(t)$.

To define the fractional Brown line-to-line function $B_H(t)$ we consider $B(t)$ and change the exponent from $H = \frac{1}{2}$ to any real number satisfying $0 < H < 1$. Cases where $H \neq \frac{1}{2}$ are properly fractional. All the $B_H(t)$ are continuous and nondifferentiable. Clearly

$$< [B_H(t + \Delta t) - B_H(t)]^2 >= |\Delta t|^{2H}$$

and the spectral density of $B_H(t)$ is f^{-2H-1}. The discrete fractional Gaussian noise is the sequence of increments of $B_H(t)$ over successive unit time spans. Its correlation is

$$2^{-1}\left[|d+1|^{2H} - 2|d|^{2H} + |d-1|^{2H}\right]$$

We now set $B_H(0) = 0$ and define the past increment as $-B_H(-t)$ and the future increment as $B_H(t)$. We get the correlation of past and future as

$$< -B_H(-t)B_H(t) >= 2^{-1}\Big(< [B_H(t) - B_H(-t)]^2 > -2 < [B_H(t)]^2 >\Big) = 2^{-1}(2t)^{2H} - t^{2H}$$

Dividing by $< [B_H(t)]^2 >= t^{2H}$ we obtain the correlation which is independent of t and given by $2^{2H-1} - 1$. In the classical case where $H = \frac{1}{2}$ the correlation vanishes as expected. For $H > \frac{1}{2}$ the correlation is positive, expressing persistence, and it becomes 1 when $H = 1$. On the other hand, for $H < \frac{1}{2}$ the correlation is negative, expressing anti-persistence, and it becomes $-\frac{1}{2}$ when $H = 0$. While classical algorithm for generating random function between $t = 0$ and $t = T$ is independent of T, it is no-longer the case when generating fractional Brownian functions.

The levy stable line-to-line functions are random functions having stationary independent increments and such that the incremental random variable $X(t) - X(0)$ is Levy stable. The scaling factor $a(t)$ making $[X(t) - X(0)]a(t)$ independent of t must take the form $a(t) = t^{-\frac{1}{D}}$. This process generalises the ordinary Brownian motion to $D \neq 2$. The process $X(t)$ is discontinuous and includes jumps. When $D < 1$, the process $X(t)$ includes only jumps. The number of jumps occurring between t and $t + \Delta t$ and having an absolute value exceeding u is a Poisson random variable of expectation equal to $|\Delta t|u^{-D}$. The relative numbers of positive and negative jumps are $\frac{1}{2}(1 + \beta)$ and $\frac{1}{2}(1 - \beta)$. The case $\beta = 1$ involves only positive jumps, which is called stable subordinator and serves to define the Levy staircases. Note, since $u^{-D} \to \infty$ as $u \to 0$, the total expected number of jumps is infinite no matter how small is Δt. However, the jumps for which $u < 1$ add to a finite cumulative total. That is, the expected length of small jumps is finite. It is proportional to

$$\int_0^1 Du^{-D-1}udu = D\int_0^1 u^{-D}du < \infty$$

In the case $1 < D < 2$ the above integral diverges, hence the total contribution of small jump's expected length is infinite. As a result, $X(t)$ includes a continuous term and a jump term. Both are infinite, but they have a finite sum.

Mandelbrot defined a fractal set as a set in a metric space for which

$$\text{Hausdorff Besicovitch dimension } D > \text{ Topological dimension } D_T$$

A fractal can be defined alternatively as a set for which

$$\text{Frostman capacitary dimension } > \text{ Topological dimension}$$

While standard assumptions in time series analysis state

1. that $< X^2 >< \infty$, and
2. that X is weakly (short-run) depenent

B.7.2 The R/S analysis

Mandelbrot showed that long-tailed records are often best interpreted by accepting $< X^2 >= \infty$. To tackle the question of weather a record is weakly or strongly dependent he disregarded the distribution of $X(t)$ by considering rescaled range analysis (R/S analysis) which is a statistical technique concerned with the distinction between the short and the very long run. Introducing the Hurst Coefficient J, such that $0 \leq J \leq 1$, $J = \frac{1}{2}$ is characteristic of independent, Markov and other short-run dependent random functions. The intensity of very long-run dependence is measured by $J - \frac{1}{2}$, and can be estimated from the data. The Hurst Coefficient J is robust with respect to the marginal distribution and continues to be effective even when $X(t)$ is so far from Gaussian that $< X^2(t) >$ diverges and all second order techniques are invalid. In continuous time t, we let $X^*(t) = \int_0^t X(u)du$ be the cumulative value of the process X, and we further define $X^{2*}(t) = \int_0^t X^2(u)du$ and $X^{*2} = (X^*)$. In discrete time i we get $X^*(t) = \sum_{i=1}^{[t]} X(i)$ with $X^*(0) = 0$ and where $[t]$ is the integer part of t. For every lag $d > 0$ the adjusted range of $X^*(t)$ in the time interval 0 to d is given by

$$R(d) = \max_{0 \leq u \leq d} \{X^*(u) - \frac{u}{d} X^*(d)\} - \min_{0 \leq u \leq d} \{X^*(u) - \frac{u}{d} X^*(d)\}$$

We then estimate the sample standard deviation of $X(t)$ as

$$S^2(d) = \frac{1}{d} X^{2*}(d) - \frac{1}{d^2} X^{*2}(d)$$

The expression

$$Q(d) = \frac{R(d)}{S(d)}$$

is the R/S statistics, or self-rescaled self-adjusted range of $X^*(t)$. Assuming that there exists a real number J such that, as $d \to \infty$, $\frac{1}{d^J} \frac{R(d)}{S(d)}$ converges in distribution to a nondegenerate limit random variable, then Mandelbrot proved that $0 \leq J \leq 1$. The function X is then said to have the R/S exponent H with a constant R/S prefactor. More generally, the ratio $\frac{1}{d^J L(d)} \frac{R(d)}{S(d)}$ converges in distribution to a nondegenerate limit random variable, where $L(d)$ denotes a slowly varying function at infinity [1]. For $L(d) = \log d$, the function X is said to have the R/S exponent J and the prefactor $L(d)$. Note, $J = \frac{1}{2}$ whenever $S(d) \to < X^2 >$ and the rescaled $a^{-\frac{1}{2}} X^*(at)$ converges weakly to $B(t)$ as $a \to \infty$. In order to obtain $J = H \neq \frac{1}{2}$ with a constant prefactor, it suffices that $S(d) \to < X^2 >$ and that $X^*(t)$ be attracted by $B_H(t)$ with $< X^*(t) > \tilde{t}^{2H}$. More generally, $J = H \neq \frac{1}{2}$ with the prefactor $L(d)$ prevails if $S(d) \to < X^2 >$, and $X^*(t)$ is attracted by $B_H(t)$ and satisfies $< (X^*(t))^2 > \tilde{t}^{2H} L(t)$. Also, $J \neq \frac{1}{2}$ when $S(d) \to < X^2 >$, and $X^*(t)$ is attracted by a non-Gaussian scaling random function of exponent $H = J$. When X is a white Levy stable noise, then $< X^2 >= \infty$, and we get $J = \frac{1}{2}$.

B.8 Some continuous variables and their distributions

For details see text book by Grimmett et al. [1992].

[1] a function satisfying

$$\frac{L(td)}{L(d)} \to 1 \text{ as } d \to \infty \,, \forall t > 0$$

B.8.1 Some popular distributions

B.8.1.1 Uniform distribution

The continuous uniform distribution is a family of symmetric probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by two parameters a and b, which are its minimum and maximum values. The distribution $U(a,b)$ is the maximum entropy probability distribution for a random variate X under the constraint that it is in the distribution's support. X is uniform on $[a,b]$ with $-\infty < a < b < \infty$ if

$$F(x) = \begin{cases} 0 \text{ if } x \leq a \\ \frac{(x-a)}{(b-a)} \text{ if } a < x \leq b \\ 1 \text{ if } x > b \end{cases}$$

Roughly speaking, X takes any value between a and b with equal probability. The probability density function is given by

$$f(x) = \begin{cases} \frac{1}{(b-a)} \text{ for } x \in [a,b] \\ 0 \text{ otherwise} \end{cases}$$

The first moment of the distribution is $E[X] = \frac{1}{2}(a+b)$ and the second centralised moment is $Var(X) = \frac{1}{12}(b-a)^2$.

B.8.1.2 Exponential distribution

The exponential distribution is the probability distribution describing the time between events in a Poisson process, that is, a process where events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless. X is exponential with parameter $\lambda(>0)$ if

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, x \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

which is the continuous limit of the waiting time distribution. The probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ for } x \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

The mean of X is given by

$$E[X] = \int_0^\infty [1 - F(x)]dx = \frac{1}{\lambda}$$

and the variance of X is given by $Var(X) = \frac{1}{\lambda^2}$ so that the standard deviation of X is equal to its mean. The moments of X for $n = 1, 2, ..$ are given by

$$E[X^n] = \frac{n!}{\lambda^n}$$

An exponentially distributed random variable T follows the relation

$$P(T > s + t | T > s) = P(T > t), \forall s, t \geq 0$$

B.8.1.3 Normal distribution

The normal (or Gaussian) distribution is very useful due to the central limit theorem which states that averages of rondom variables independently drawn from independent distributions are normally distributed. Gauss [1809] introduced the distribution as a way of rationalising the method of least squares. Given the two parameters μ and σ^2, the probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty$$

where μ is the mean or expectation of the distribution and σ is its standard deviation. A random variable with a Gaussian distribution is said to be normally distributed and it is denoted by $N(\mu, \sigma^2)$. Let X be $N(\mu, \sigma^2)$, where $\sigma > 0$ and let

$$Y = \frac{(X-\mu)}{\sigma}$$

For the cumulative distribution of Y we get

$$P(Y \leq y) = P(X \leq y\sigma + \mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}v^2} dv$$

Thus Y is a standard normal deviate with $N(0,1)$. The normal distribution is the only absolutely continuous distribution whose cumulants beyond the first two (mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a specified mean and variance. The value of its distribution is practically zero when the value x lies more than a few standard deviations away from the mean. The error function is defined as the probability of a random variable with normal distribution of mean 0 and variance $\frac{1}{2}$ falling in the range $[-x, x]$

$$erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-v^2} dv$$

The CDF and error function are related by

$$N(x) = \frac{1}{2}\left[1 + erf(\frac{x}{\sqrt{2}})\right]$$

Hence, for a normal distribution f with mean μ and standard deviation σ, the CDF is

$$F(x) = N(\frac{x-\mu}{\sigma}) = \frac{1}{2}\left[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})\right]$$

The graph of the standard normal CDF has 2-fold rotational symmetry at the point $(0, \frac{1}{2})$ where $N(-x) = 1 - N(x)$. Its indefinite integral is given by

$$\int N(x)dx = xN(x) + f(x)$$

The values $N(x)$ may be approximated by a variety of methods such as numerical integration, Taylor series, asymptotic series and continued fractions. For instance, the CDF of the standard normal distribution can be expanded by integration by parts into a series. Marsaglia [2004] proposed a simple algorithm with arbitrary precision based on the Taylor series expansion

$$N(x) = \frac{1}{2} + f(x)\left[x + \frac{x^3}{3} + \frac{x^5}{3.5} + ... + \frac{x^{2n+1}}{(2n+1)!!} + ...\right]$$

where $n!!$ is the double factorial [2]. The first derivative of the density is $f^{'} = -xf(x)$, the second derivative is $f^{''}(x) = (x^2 - 1)f(x)$ so that the nth derivative satisfies

$$f^n(x) = (-1)^n He_n(x)f(x)$$

where $He_n(x)$ is the Hermite polynomial of order n. If X has a normal distribution, the moments exist and are finite for any p whose real part is greater than -1. For any non-negative integer p, the plain central moments are

$$E[X^p] = \begin{cases} 0 \text{ if p is odd} \\ \sigma^p(p-1)!! \text{ if p is even} \end{cases}$$

The Fourier transform of a normal distribution f with mean μ and standard deviation σ is

$$\hat{f}(t) = \int_{-\infty}^{\infty} f(x)e^{itx} = e^{i\mu t}e^{-\frac{1}{2}(\sigma t)^2}$$

where i is the imaginary number. The moment generating function of X is given by

$$M(t) = \hat{f}(-it) = e^{\mu t}e^{\frac{1}{2}\sigma^2 t^2}$$

and the cumulant generting function is the logarithm of the moment generating function

$$g(t) = \ln M(t) = \mu t + \frac{1}{2}\sigma^2 t^2$$

Hence, the first and second derivatives of the cumulant generating function are $g^{'}(t) = \mu + \sigma^2 t$ and $g^{''}(t) = \sigma^2$, respectively. The cumulants are therefore

$$k_1 = \mu \text{ , } k_2 = \sigma^2 \text{ , } k_3 = k_4 = ... = 0$$

B.8.1.4 Gamma distribution

The gamma distribution is a two-parameter family of continuous probability distributions where the exponential distribution and the chi-squared distribution are special cases. It is the maximum entropy probability distribution for a random variable X for which $E[X] = k\theta = \frac{\alpha}{\beta}$ is fixed and greater than zero, and $E[\ln X] = \psi(k) + \ln\theta = \psi(\alpha) - \ln\beta$ is fixed and $\psi(\bullet)$ is the digamma function. In the following we denote the pair (k, θ) by $(t, \frac{1}{\lambda})$ corresponding to $\alpha = t$ and $\beta = \lambda$. The variable X has the gamma distribution with parameters λ, $t > 0$, denoted by $\Gamma(\lambda, t)$ if it has the probability density function

$$f(x) = \frac{1}{\Gamma(t)}\lambda^t x^{t-1}e^{-\lambda x} \text{ , } x \geq 0$$

where $\Gamma(t)$ is the gamma function

$$\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}dx$$

If $t = 1$ then X is Exponentially distributed with parameter λ. The cumulative distribution function is the regularised gamma function

[2] The product of all the integer from 1 up to some non-negative integer n having the same parity as n is called the double factorial of n, given by

$$n!! = \prod_{k=0}^{m}(n - 2k) = n(n-2)(n-4)...$$

where $m = \lceil \frac{n}{2} \rceil - 1$.

$$F(x) = \int_0^x f(u)du = \frac{\gamma(t, \lambda x)}{\Gamma(t)}$$

where $\gamma(t, \lambda x)$ is the lower incomplete gamma function. If t is a positive integer, the CDF follows the series expansion

$$F(x) = e^{-\lambda x} \sum_{i=t}^{\infty} \frac{(\lambda x)^i}{i!}$$

The skewness $\frac{2}{\sqrt{t}}$ depends only on the shape parameter t and approaches a normal distribution when t is large (when $t > 10$).

B.8.1.5 Beta distribution

The beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ with two shape parameters α and β. It is applied to model the behaviour of random variables limited to interval of finite length. The probability density function of the beta distribution for $0 \leq x \leq 1$ is a power function of the variable X and of its reflection $(1 - X)$

$$\begin{aligned} f(x) &= cst \times x^{\alpha-1}(1-x)^{\beta-1} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

where $\Gamma(\bullet)$ is the gamma function, and $B(\bullet, \bullet)$ is a normalisation constant ensuring that the total probability integrates to 1. Note, this definition includes both ends $x = 0$ and $x = 1$, but some authors choose to exclude them and consider $0 < x < 1$ instead. Beta densities are symmetric (for $\alpha = \beta = 1$), unimodal ($\alpha, \beta > 1$), uniantimodal, increasing, decreasing or constant depending on the values α and β relative to 1, and have many more attractive properties. The beta density is U-shaped when $\alpha, \beta < 1$, it has positive skew when $\alpha < \beta$ and negative skew when $\alpha > q$. Since the beta distribution approaches the Bernoulli distribution in the limit when both shape parameters α and β approach zero, some authors denote the pair (α, β) by (p, q). The slope of the pdf is given by

$$\begin{aligned} f^{'}(x) &= f(x)\frac{(\alpha+\beta-2)x-(\alpha-1)}{(x-1)x} \\ &= -\frac{x^{\alpha-2}(1-x)^{\beta-2}}{B(\alpha,\beta)}\big((\alpha+\beta-2)x-(\alpha-1)\big) \end{aligned}$$

and at $x = \frac{1}{2}$, for $\alpha = \beta$, the slope of the pdf is zero. Further, we get the differential equation

$$(x-1)xf^{'}(x) + \big(\alpha - 1 - (\alpha+\beta-2)x\big)f(x) = 0$$

The cumulative distribution function is given by

$$F(x) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} = I_x(\alpha,\beta)$$

where $B(\bullet; \alpha, \beta)$ is the incomplete beta function and $I_\bullet(\alpha, \beta)$ is the regularised incomplete beta function. The mode of a beta distributed random variable X with $\alpha, \beta > 1$ (corresponding to the peak in the PDF) is given by

$$\frac{\alpha-1}{\alpha+\beta-2}$$

The mean of X is given by

$$\begin{aligned}\mu = E[X] = \int_0^1 xf(x)dx &= \int_0^1 x\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\ &= \frac{\alpha}{\alpha+\beta} = \frac{1}{1+\frac{\beta}{\alpha}}\end{aligned}$$

which only depends on the ratio $\frac{\beta}{\alpha}$. For $\alpha = \beta$, we get the mean $\mu = \frac{1}{2}$, which is at the center of the (symmetric) distribution. We also get the following limits

$$\lim_{\frac{\beta}{\alpha}\to 0}\mu = 1 \text{ and } \lim_{\frac{\beta}{\alpha}\to\infty}\mu = 0$$

For the former limit ratio, the beta distribution becomes a one-point degenerate distribution with a Dirac delta function spike at $x = 1$ with probability 1 and zero probability elsewhere else. Similarly, for the latter limit ratio the spike is at $x = 0$. Next, we consider the limit cases where one parameter is finite (non-zero) and the other one approaches the limits

$$\lim_{\beta\to 0}\mu = \lim_{\alpha\to\infty}\mu = 1 \text{ and } \lim_{\beta\to\infty}\mu = \lim_{\alpha\to 0}\mu = 0$$

The variance of a beta distributed random variable X is given by

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

and for $\alpha = \beta$ the variance simplifies to

$$Var(X) = \frac{1}{4(2\beta+1)}$$

Setting $\alpha = \beta = 0$ in the above equation, we obtain the maximum variance $Var(X) = \frac{1}{4}$. The skewness of the beta distribution is

$$\gamma_1 = \frac{E[(X-\mu)^3]}{(Var(X))^{\frac{3}{2}}} = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$$

For $\alpha = \beta$ in the above equation, we obtain $\gamma_1 = 0$ showing that in that setting the distribution is symmetric. We get positive skew (right-tailed) for $\alpha < \beta$ and negative skew (left-tailed) for $\alpha > \beta$.

B.8.1.6 Kumaraswamy distribution

The Kumaraswamy's double bounded distribution (also known as minmax distribution) is a family of continuous probability distributions defined on the interval $[0, 1]$ with two shape parameters a and b (see Kumaraswamy [1980]). While similar to the beta distribution, the Kumaraswamy distribution is simpler to use due to the simple closed form of both its probability density function and cumulative distribution function. The pdf is given by

$$f(x) = abx^{a-1}(1-x^a)^{b-1}\,,\, 0 < x < 1\,,\, a, b > 0$$

and the cdf is given by

$$F(x) = 1 - (1 - x^a)^b$$

In a more general form, using linear transformation, the normalised variable x is replaced with the unshifted and unscaled variable z where

$$x = \frac{z - z_{min}}{z_{max-z_{min}}}, z_{min} \leq z \leq z_{max}$$

In a discrete setting where the density satisfies $P(X \in dx) \approx f_X(x)dx$, then when considering the unshifted variable z we get $dx = \frac{1}{z_{max-z_{min}}}$. We can invert the distribution function to obtain the quantile function

$$Q(y) = F^{-1}(y) = \left[1 - (1-y)^{\frac{1}{b}}\right]^{\frac{1}{a}}, 0 < y < 1$$

We can then trivially generate random variable as for $U \sim U(0,1)$, then $X \sim f$ if

$$X = (1 - U^{\frac{1}{b}})^{\frac{1}{a}}$$

The Kumaraswamy distribution has the same basic shape properties as the beta distribution (see Jones [2009])

- unimodal: $a > 1, b > 1$.
- uniantimodal: $a < 1, b < 1$.
- increasing: $a > 1, b \leq 1$.
- decreasing: $a \leq 1, b > 1$.
- constant: $a = b = 1$.

and the Kumaraswamy density also matches that of the beta density at the boundaries of their support

- $f(x) \sim x^{a-1}$ as $x \to 0$.
- $f(x) \sim (1-x)^{b-1}$ as $x \to 1$

The raw moments of the Kumaraswamy distribution are given by

$$m_n = \frac{b\Gamma(1 + \frac{n}{a})\Gamma(b)}{\Gamma(1 + b + \frac{n}{a})} = bB(1 + \frac{n}{a}, b)$$

where $B(\bullet, \bullet)$ is the Beta function. Similarly to the beta distribution, they exist for all $n > -a$. The variance, skewness, and excess kurtosis can be computed from these raw moments. For instance, the variance is given by $\sigma^2 = m_2 - m_1^2$

$$Var(X) = bB(1 + \frac{2}{a}, b) - \left(bB(1 + \frac{1}{a}, b)\right)^2$$

Assuming that $X_{a,b}$ is a Kumaraswamy distributed random variable, then it is the a-th root of a suitably defined $Beta$ distributed random variable (see Jones [2009]). Let $Y_{1,b}$ denote a Beta r.v. with parameters $\alpha = 1$ and $\beta = b$, then $X_{a,b} = Y_{1,b}^{\frac{1}{a}}$ with equality in distribution

$$P(X_{a,b} \leq x) = \int_0^x abu^{a-1}(1-u^a)^{b-1}du = \int_0^{x^a} b(1-u)^{b-1}du = P(Y_{1,b} \leq x^a) = P(Y_{1,b}^{\frac{1}{a}} \leq x)$$

Considering a generalised distribution with Beta distributed r.v. $Y_{\alpha,\beta}^{\frac{1}{\gamma}}$ with $\gamma > 0$, the raw moments are given by

$$m_n = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\frac{n}{\gamma})}{\Gamma(\alpha)\Gamma(\alpha+\beta+\frac{n}{\gamma})}$$

where the original moments are recovered by setting $\alpha = 1$,$\beta = b$ and $\gamma = a$. However, in general the CDF does not have a closed form solution. Note,

- if $X \sim K(1,1)$ then $X \sim U(0,1)$.
- if $X \sim B(1,b)$ then $X \sim K(1,b)$.
- if $X \sim B(a,1)$ then $X \sim K(a,1)$.
- if $X \sim K(a,b)$ then $X \sim GB1(a,1,1,b)$.

where $GB1(a,1,1,b)$ is the generalised beta distribution of the first kind.

B.8.1.7 Generalised beta distribution

McDonald [1984] introduced the generalised beta distribution of the first kind characterised by its density function

$$f_{GB1}(x) = B^{-1}(p,q)\left[ax^{ap-1}(1-x^a)^{q-1}\right], 0 < x < 1$$

where we recover the classical beta distribution for $a = 1$ $(1,p,q)$, and the Kumaraswamy distribution for $p = 1$ $(a,1,q)$. It is the distribution of the $\frac{1}{a}$ power of a $B(p,q)$ random variable or of the p-th order statistic of a sample of size $p+q-1$ from the power function distribution $B(a,1)$. Jones [2004] introduced the general class of beta-generated distibutions characterised by their density function

$$f_{BG}(x) = B^{-1}(\alpha,\beta)f(x)[F(x)]^{\alpha-1}[1-F(x)]^{\beta-1}, x \in \mathcal{I}$$

where $F(x)$ is the parent distribution function and $f(x)$ is its density. Jones concentrated on the cases where F is symmetric about zero with no free parameter other than location and scale, and where $\mathcal{I}$ is the whole real line. Since then, Beta-generated distributions with more general parents have been studied extensively. Even though the shapes of the BG-distributions are more flexible than the beta-normal, they depend on two parameters adding only a limited structure to the generated distribution. Alexander et al. [2010] proposed to use a more flexible generator distribution such as a generalised beta distribution of the first kind $GB1$(a,p,q). They recover theclassical BG and Kumaraswamy generated distributions as special cases. Given a parent distribution $F(x)$ with density $f(x)$, the generalised beta-generated (GBG) density is given by

$$f(GBG)(x) = B^{-1}(p,q)f(x)\left[aF^{ap-1}(x)(1-F^a(x))^{q-1}\right], x \in \mathcal{I}$$

where we recover the beta-generated distribution for $a = 1$ and the Kumaraswamy distribution for $p = 1$. Further, the GBG with parent $F(x)$ is a standard beta-generated distribution with parent $F^a(x)$.

B.8.1.8 Chi-square distribution

Given the Gamma distribution, when $\lambda = \frac{1}{2}$ and $t = \frac{1}{2}d$ for some integer d, then X is said to have the Chi-squared distribution $\chi^2(d)$ with d degrees of freedom with density

$$f(x) = \frac{1}{\Gamma(\frac{1}{2}d)}(\frac{1}{2})^{\frac{1}{2}d}x^{\frac{1}{2}d-1}e^{-\frac{1}{2}x}, x \geq 0$$

Definition B.8.1 *If Z is a standard normal random variable, the distribution of $U = Z^2$ is called the chi-square distribution with 1 degree of freedom.*

Note, if $X \sim N(\mu, \sigma^2)$ then $U = \frac{(X-\mu)}{\sigma} \sim N(0,1)$ and therefore $[\frac{(X-\mu)}{\sigma}]^2 \sim \chi_1^2$.

Definition B.8.2 *If $U_1, U_2, .., U_n$ are independent chi-square random variables with 1 degree of freedom, the distribution of $V = U_1 + U_2 + .. + U_n$ is called the chi-square distribution with n degrees of freedom denoted by χ_n^2.*

Note, if U and V are independent and $U \sim \chi_n^2$ and $V \sim \chi_m^2$ then $U + V \sim \chi_{m+n}^2$.

B.8.1.9 Weibull distribution

Introduced by Frechet [1927], the Weibull distribution is a continuous probability distribution described by Weibull [1951]. The random variable X is Weibull with parameters $\alpha, \beta > 0$ if the cdf satisfies

$$F(x) = 1 - e^{-\alpha x^\beta}, x \geq 0$$

which is a stretched exponential function. Note, β is the shape parameter and α is the scale parameter of the distribution. If the r.v. X is a time to failure, then the Weibull distribution gives a distribution for which the failure rate is proportional to a power time where β is that power plus one. Differentiating the cdf, we get the density function

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}, x \geq 0$$

The form of the density function changes drastically with the value of β. For example, setting $\beta = 1$, we recover the Exponential distribution with $\alpha = \lambda$.

B.8.2 Normal and Lognormal distributions

Given a random variable X normally distributed with mean μ_X and variance σ_X^2, its probability distribution function is

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}}$$

Assuming the random variables X and Y are jointly Gaussian with $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and correlation ρ, the bivariate normal distribution function is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

If X is lognormally distributed variable $X \sim LN(\mu_X, \sigma_X^2)$ then the variable $Y = \log X$ is normally distributed with mean μ_Y and standard deviation σ_Y, and the pdf of X satisfies $\phi_X(x) = \frac{dP(X \leq x)}{dx} = \frac{dP(Y \leq y)}{dy}\frac{dy}{dx}$, that is

$$\phi_X(x; \mu_Y, \sigma_Y) = \frac{1}{x} f_Y(\log(x)), x \in (0, \infty)$$

The expected value and variance of the variable X are

$$\mu_X = E[X] = e^{\mu_Y + \frac{1}{2}\sigma_Y^2}$$
$$\sigma_X^2 = Var(X) = (e^{\sigma_Y^2} - 1)e^{2\mu_Y + \sigma_Y^2}$$

Alternatively, if we know the expected value and variance of X we can recover the parameter μ_Y and σ_Y with

$$\mu_Y = \log E[X] - \frac{1}{2}\log(1 + \frac{Var(X)}{(E[X])^2})$$
$$\sigma_Y^2 = \log(1 + \frac{Var(X)}{(E[X])^2})$$

We consider the variables X_i for $i = 1, 2$ such that $Y_i = \log X_i$ are normally distributed $Y_i \sim N(\mu_{Y_i}, \sigma_{Y_i})$ and assume their joint distribution is bivariate normal with correlation coefficient ρ_Y. Hence, X_1 and X_2 are bivariate lognormally distributed $LN(\mu_{Y_1}, \mu_{Y_2}, \sigma_{Y_1}, \sigma_{Y_2}, \rho_Y)$ with joint distribution

$$\phi_{X_1,X_2}(x_1, x_2) = \frac{1}{x_1 x_2} f_{Y_1,Y_2}(\log(x_1), \log(x_2))$$

Johnson et al give the correlation coefficient

$$\rho_X = \frac{e^{\rho_Y \sigma_{Y_1} \sigma_{Y_2}} - 1}{\sqrt{(e^{\sigma_{Y_1}^2} - 1)(e^{\sigma_{Y_2}^2} - 1)}}$$

B.8.3 Multivariate Normal distributions

Given $X_1, X_2, .., X_n$, the multivariate normal distribution is obtained by rescaling the exponential of a quadratic form. A quadratic form is a function $Q : \mathbb{R}^n \to \mathbb{R}$ of the form

$$Q(x) = \sum_{1 \leq i,j \leq n} a_{ij} x_i x_j = x A x^\top$$

where $x = (x_1, ..., x_n)$, $x^\top$ is the transpose of x, and $A = (a_{ij})$ is a real symmetric matrix with non-zero determinant. A well known theorem about diagonalizing matrices states that there exists an orthogonal matrix B such that

$$A = B \Lambda B^\top$$

where Λ is the diagonal matrix with eigenvalues $\lambda_1, .., \lambda_n$ of A on its diagonal. So, the quadratic form becomes

$$Q(x) = y \Lambda y^\top = \sum_i \lambda_i y^2$$

where $y = xB$. Q is called a positive definite quadratic form if $Q(x) > 0$ for all vectors x with some non-zero coordinate. From matrix theory, $Q > 0$ if and only if $\lambda_i > 0$ for all i.

Definition B.8.3 *$X = (X_1, .., X_n)$ has the multivariate normal distribution written $N(\mu, V)$, if its joint density function is*

$$f(x) = [(2\pi)^n |V|]^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)V^{-1}(x-\mu)^\top}$$

where V is a positive definite symmetric matrix.

Theorem B.8.1 *If X is $N(\mu, V)$ then*

1. *$E[X] = \mu$, which is to say that $E[X_i] = \mu_i$ for all i*
2. *$V = (v_{ij})$ is called the covariance matrix because $v_{ij} = Cov(X_i, X_j)$*

We often write

$$V = E[(X-\mu)^\top (X-\mu)]$$

where $(X-\mu)^\top(X-\mu)$ is a matrix with (i,j)th entry $(X_i-\mu_i)(X_j-\mu_j)$. A very important property of this distribution is its invariance of type under linear changes of variables.

Theorem B.8.2 *If $X = (X_1,..,X_n)$ is $N(0,V)$ and $Y = (Y_1,..,Y_m)$ is given by $Y = XD$ for some matrix D of rank $m \leq n$, then Y is $N(0, D^\top V D)$.*

A similar result holds for linear transformations of $N(\mu, V)$ variables. We now present another way of defining the multivariate normal distribution.

Definition B.8.4 *The vector $X = (X_1,..,X_n)$ of random variables is said to have the multivariate normal distribution whenever, for all $a \in \mathbb{R}^n$, $xa^\top = a_1X_1 + a_2X_2 + ... + a_nX_n$ has a normal distribution.*

That is, X is multivariate normal if and only if every linear combination of the X_i is univariate normal.

B.8.4 Distributions arising from the Normal distribution

B.8.4.1 Presenting the problem

Statisticians are frequently faced with a collection $X_1, X_2, .., X_n$ of random variables arising from a sequence of experiments. They might assume that they are independent $N(\mu, \sigma^2)$ variables for some fixed but unknown values for μ and σ^2. This assumption is often a very close approximation to reality. They proceed to estimate the values of μ and σ^2 by using functions of $X_1, .., X_n$. They commonly use the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as a guess at the value of μ and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

as a guess at the value of σ^2. From the Definition (B.8.2) we see that $(n-1)\frac{S^2}{\sigma^2}$ is a sum of independent chi-square random variables, and we obtain the following Definition:

Theorem B.8.3 *If $X_1, X_2, ..$ are independent $N(\mu, \sigma^2)$ variables then $\overline{X}$ and S^2 are independent. Further*

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \text{ and } (n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Note, σ is only a scaling factor for $\overline{X}$ and S. As a preliminary to showing that $\overline{X}$ and S^2 are independently distributed, we get the theorem

Theorem B.8.4 *The random variable $\overline{X}$ and the vector of random variables $\left(X_1 - \overline{X}, ..., X_n - \overline{X}\right)$ are independent.*

***Corollary* 6** *Given $\overline{X}$ and S^2 defined as above, then*

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

B.8.4.2 The t-distribution

Given the remark in the previous Appendix, we consider two random variables

$$U = \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$$

which does not depend on σ, and

$$V = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \sim N(0,1)$$

which does not depend on σ. Hence the random variable

$$T = \frac{V}{\sqrt{\frac{U}{n-1}}} = \frac{\sqrt{n}}{S}(\overline{X} - \mu)$$

has a distribution which does not depend on σ. T is the ratio of two independent random variables, and it is said to have a t-distribution with $(n-1)$ degrees of freedom written $t(n-1)$. It is also called Student's t distribution.

Definition B.8.5 *If $V \sim N(0,1)$ and $U \sim \chi^2(n)$ and U and V are independent, then the distribution of $\frac{V}{\sqrt{\frac{U}{n}}}$ is called the t distribution with n degrees of freedom.*

The joint density of U and V is

$$f(u,v) = \frac{(\frac{1}{2})^r e^{-\frac{1}{2}u} u^{\frac{1}{2}r-1}}{\Gamma(\frac{1}{2}r)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}$$

where $r = n-1$. See Appendix (B.8.1.4) for a description of the $\chi^2(d)$ density. Then map (u,v) onto (s,t) by

$$s = u \, , t = v(\frac{u}{r})^{-\frac{1}{2}}$$

and use the Corollary

Corollary 7 *If X_1 and X_2 have joint density function f, then the pair Y_1, Y_2 given by $(Y_1, Y_2) = T(X_1, X_2)$ has joint density function*

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} f(x_1(y_1,y_2), x_2(y_1,y_2))|J(y_1,y_2)| \text{ if } (y_1,y_2) \text{ is in the range of } T \\ 0 \text{ otherwise} \end{cases}$$

to get

$$f_{U,T}(s,t) = (\frac{s}{r})^{\frac{1}{2}} f(s, t(\frac{s}{r})^{\frac{1}{2}})$$

Integrating over s we obtain

$$f_T(t) = \frac{\Gamma(\frac{1}{2}(r+1))}{\sqrt{\pi r}\Gamma(\frac{1}{2}r)}(1 + \frac{t^2}{r})^{-\frac{1}{2}(r+1)} \, , \; -\infty < t < \infty$$

as the density function of the $t(r)$ distribution. Note, the t-distribution is symmetric about zero.

***Remark* B.8.1** *As the number of degrees of freedom approaches infinity, the t-distribution tends to the standard normal distribution. For more than 20 or 30 degrees of freedom, the distributions are very close.*

Also, the tails become lighter as the degrees of freedom increase.

B.8.4.3 The F-distribution

Another important distribution in statistics is the F-distribution. Let U and V be independent variables with the $\chi^2(r)$ and $\chi^2(s)$ distributions respectively. Then

$$F = \frac{\frac{U}{r}}{\frac{V}{s}}$$

is said to have the F-distribution with r and s degrees of freedom, written $F(r,s)$. Note, the following two properties

- F^{-1} is $F(s,r)$
- T^2 is $F(1,r)$ if T is $t(r)$

The density function of the $F(r,s)$ distribution is

$$f(x) = \frac{r\Gamma(\frac{1}{2}(r+s))}{s\Gamma(\frac{1}{2}r)\Gamma(\frac{1}{2}s)} \frac{(\frac{rx}{s})^{\frac{1}{2}r-1}}{(1+(\frac{rx}{s}))^{\frac{1}{2}(r+s)}}, x > 0$$

B.8.5 Approximating the probability distribution

The Gram-Charlier A series and the Edgeworth series are series approximating a probability distribution in terms of its cumulants. The series are the same, but the arrangement of terms differ.

B.8.5.1 The Gram-Charlier A series

The idea of the Gram-Charlier A series is to approximate the characteristic function of the distribution, whose probability density function is f, in terms of the characteristic function of a distribution with known and suitable properties. We can then recover f with the inverse Fourier transform. Given a continuous random variable X, we let Ψ be the CF of its distribution with pdf f, and k_r its cumulants (see details in Appendix (B.4)). We expand in terms of a known distribution with pdf ϕ, CF Φ, and cumulants γ_r. The density function ϕ is generally chosen to be that of the normal distribution. By definition of the cumulants, we have

$$\Psi(t) = e^{\sum_{r=1}^{\infty}(k_r-\gamma_r)\frac{(it)^r}{r!}}\Phi(t)$$

By the properties of the Fourier transform, $(it)^r\Phi(t)$ is the Fourier transform of $(-1)^r[D^r\phi](-x)$, where D is the differential operator with respect to x. Thus, setting $x = -x$ in the above equation, we ger the formal expansion

$$f(x) = e^{\sum_{r=1}^{\infty}(k_r-\gamma_r)\frac{(-D)^r}{r!}}\phi(x)$$

If ϕ is the normal density with mean $\mu = k_1$ and variance $\sigma^2 = k_2$, then the expansion of the density becomes

$$f(x) = e^{\sum_{r=3}^{\infty}k_r\frac{(-D)^r}{r!}}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Expanding the exponential and collecting terms according to the order of the derivatives we obtain the Gram-Charlier A series. Focussing on the first two correction terms to the normal distribution, we get

$$f(x) \approx \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\left[1 + \frac{k_3}{3!\sigma^3}He_3(\frac{x-\mu}{\sigma}) + \frac{k_4}{4!\sigma^4}He_4(\frac{x-\mu}{\sigma})\right]$$

where $He_3(x) = x^3 - 3x$ and $He_4(x) = x^4 - 6x^2 + 3$ are Hermite polynomials. Note, this expression is not guaranteed to be positive, so that it is not a valid probability distribution. There are many cases of interest where the Gram-Charlier A series diverges. It converges only if the density $f(x)$ falls at a faster rate than $e^{-\frac{x^2}{4}}$ at infinity (see Cramer [1957]). Hence, when it does not converge the series is not a true asymptotic expansion because it is not possible to estimate the error of the expansion.

B.8.5.2 The Edgeworth series

Edgeworth developed a similar expansion as an improvement of the central limit theorem. The main advantage of Edgeworth series is that the error is controlled, leading to a true asymptotic expansion. Given $\{X_i\}$ a sequence of i.i.d. random variables with mean μ and variance σ^2, we let Y_n be their standardised sum

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

We let F_n be the cumulative distrbution functions of the variables Y_n. Then, by the central limit theorem, we get

$$\lim_{n\to\infty} F_n(x) = N(x) = \int_{-\infty}^{x} \phi(u)du$$

with $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ for every x, as long as the mean and variance are finite. If we now assume that the random variables X_i with $i = 1, .., n$ have higher cumulants $k_r = \sigma^r \lambda_r$, and if we expand in terms of the standardised normal distribution $\phi(x)$, then the cumulant differences in the formal expansion of the CF $\Psi_n(t)$ of F_n are

$$\begin{aligned} k_1^{F_n} - \gamma_1 &= 0 \\ k_2^{F_n} - \gamma_2 &= 0 \\ k_r^{F_n} - \gamma_r &= \frac{k_r}{\sigma^r n^{\frac{r}{2}-1}} = \frac{\lambda_r}{n^{\frac{r}{2}-1}} \text{ , } r \geq 3 \end{aligned}$$

Following the same approach as the Gram-Charlier A series except that terms are collected according to powers of n, we get

$$\Psi_n(t) = \big[1 + \sum_{j=1}^{\infty} \frac{P_j(it)}{n^{\frac{j}{2}}}\big] e^{-\frac{t^2}{2}}$$

where $P_j(x)$ is a polynomial of degree $3j$. After applying the inverse Fourier transform, the distribution function is given by

$$F_n(x) = N(x) + \sum_{j=1}^{\infty} \frac{P_j(-D)}{n^{\frac{j}{2}}} N(x)$$

Letting $N^{(j)}(x)$ be the jth derivative of $N(\bullet)$ at point x, we can recover the first few terms of the expansion. Further, since the derivatives of the density of the normal distribution are related to the normal density by $\phi^{(n)}(x)$ is $(-1)^n He_n(x)\phi(x)$, then we obtain an alternative representation in terms of the density function. However, Edgeworth expansions are not guaranteed to be a proper probability distribution since the integral of the density needs not to integrate to one, and the probabilities can be negative. Further, they can be inaccurate, especially in the tails, because they are obtained under a Taylor series around the mean, and they guarantee (asymptotically) an absolute error but not a relative one.

B.9 Some results on Normal sampling

B.9.1 Estimating the mean and variance

Given $[\mathbb{R}^n, N(\mu, \sigma^2)^{\otimes n}]$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ unknown, we set $\theta = (\mu, \sigma^2)^\top$. The likelihood of the model is

$$l_n(\theta) = \frac{1}{(2\pi\sigma^2)^{-\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

Setting $L_n(\theta) = \log l_n(\theta)$, the score vector is

$$\partial_\theta L_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \mu)^2 \end{bmatrix}$$

We can then obtain the maximum likelihood estimator for μ and σ^2 as

$$\hat{\mu} = \overline{Y}\ ,\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Note, $\hat{\mu}$ is a non-biased estimator but $\hat{\sigma}^2$ is a biased estimator. We must consider

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$$

to get a non-biased estimator since $E[s^2] = \sigma^2$.

Theorem B.9.1 *(Theorem of Fisher)*
$\overline{Y}$ and s^2 are two independent statistics and we get

$$\overline{Y} \sim N(\mu, \frac{\sigma^2}{n}) \text{ and } (n-1)\frac{s^2}{\sigma^2} \sim \chi^2(n-1)$$

B.9.2 Estimating the mean with known variance

Given $[\mathbb{R}^n, N(\theta, \sigma_0^2)^{\otimes n}]$ with $\theta \in \mathbb{R}$ and σ_0 known, we look for the confidence interval of θ. Note, the function $\frac{\sqrt{n}(\overline{Y}-\theta)}{\sigma_0}$ is pivotal since its law $N(0,1)$ is fixed. We let $Z_p = Z_{1-\frac{\alpha}{2}}$ be the $(1-\frac{\alpha}{2})$ percentile point of the $N(0,1)$ distribution, we get

$$\forall\theta\ ,\ P_\theta(-Z_p \leq \frac{\sqrt{n}(\overline{Y}-\theta)}{\sigma_0} \leq Z_p) = 1-\alpha$$

which we rearrange as

$$\forall\theta\ ,\ P_\theta(\overline{Y} - Z_p\frac{\sigma_0}{\sqrt{n}} \leq \theta \leq \overline{Y} + Z_p\frac{\sigma_0}{\sqrt{n}}) = 1-\alpha$$

The interval $\overline{Y} \pm Z_p\frac{\sigma_0}{\sqrt{n}}$ is the confidence interval of level $1-\alpha$.

B.9.3 Estimating the mean with unknown variance

Given $[\mathbb{R}^n, N(\mu, \sigma^2)^{\otimes n}]$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ unknown, we look for the confidence interval of μ. We have n independent observations of the same law $N(\mu, \sigma^2)$. We let $\theta = (\mu, \sigma^2)^\top$, then the function g is the first coordinate $g(\theta) = \mu$. The function $\frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma}$ is no-longer pivotal for μ as it does not depend on θ only through μ. However, the function $\frac{\sqrt{n}(\overline{Y}-\mu)}{s}$ with

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$$

also has a fixed law, which is the Student law with $(n-1)$ degrees of freedom. This is therefore a pivotal function. We can therefore deduce a confidence interval of level $(1-\alpha)$ symmetric around $\overline{Y}$ with bounds given by

$$\overline{Y} \pm \frac{s}{\sqrt{n}}t_p$$

with t_p being the quantile of order $(1-\frac{\alpha}{2})$ of the Student law with $(n-1)$ degrees of freedom. Note, in this case the length of the interval $2\frac{s}{\sqrt{n}}t_p$ is random.

B.9.4 Estimating the parameters of a linear model

Given the linear model

$$Y = \underline{X}b + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$ and $\underline{X}$ is an $(n \times K)$ matrix of rank K, we consider b_i the ith element of b. We let $\hat{b}_i$ be the ith component of the least square estimate $\hat{b}$ of b. Hence, $\hat{b}_i$ follow the law $N(b_i, \sigma^2 a_{ii})$ where a_{ii} is the ith diagonal element of $\frac{1}{\overline{X}'\overline{X}}$. Moreover

$$(n-K)\tilde{\sigma}^2 = ||Y - \overline{X}\hat{b}||^2 \sim \chi^2(n-K)$$

is independent from $\hat{b}_i$. As a result $\frac{\hat{b}_i - b_i}{\tilde{\sigma}\sqrt{a_{ii}}}$ follows the Student law with $(n-K)$ degrees of freedom. Therefore, it is a pivotal function for b_i. We can deduce a confidence interval of level $(1-\alpha)$ for b_i symmetric around $\hat{b}_i$ with bounds

$$\hat{b}_i + \hat{\sigma}_i t_p$$

where t_p is the quantile of order $(1-\frac{\alpha}{2})$ of the Student law with $(n-K)$ degrees of freedom and $\hat{\sigma}_i = \tilde{\sigma}\sqrt{a_{ii}}$.

B.9.5 Asymptotic confidence interval

We consider a semi-parametric sample model and do not make any assumption on the law of the variables Y_i for $i = 1, .., n, ..$ We only assume that the Y_i are independent from the same law, and that the mean m and the variance σ^2 of Y exist. We are looking for an asymptotic confidence interval of level α for the mean m. The least square estimate of m is $\overline{Y}$, and using the law of the large numbers $\overline{Y}$ is asymptotically normal (central limit theorem)

$$\sqrt{n}(\overline{Y} - m)\frac{L}{n \to \infty} > N(0, \sigma^2)$$

As a result, an asymptotic pivotal function is

$$\frac{\sqrt{n}(\overline{Y} - m)}{\hat{\sigma}} \text{ with } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

because

$$\frac{\sqrt{n}(\overline{Y} - m)}{\hat{\sigma}} > \frac{L}{n \to \infty} > N(0, 1)$$

We can then deduce the asymptotic confidence interval of level α for the mean m as

$$\{\overline{Y} - \frac{\hat{\sigma}}{\sqrt{n}}Z_{1-\frac{\alpha}{2}} \leq m \leq \overline{Y} + \frac{\hat{\sigma}}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}\}$$

where $Z_{1-\frac{\alpha}{2}}$ is the quantile of order $(1-\frac{\alpha}{2})$ of the distribution $N(0,1)$.

B.9.6 The setup of the Monte Carlo engine

Given a continuous process $(X_t)_{t\geq 0}$, we want to estimate $\theta = E[f(X_T)]$ for a fixed maturity T. We take a discrete process $\hat{X}_t$ given by $\{\hat{X}_h, \hat{X}_{2h}, ..., \hat{X}_{mh}\}$ such that $mh = T$. To simplify notation we define $f_j = f(\hat{X})$ and consider the estimated value of $\hat{\theta} = E[f(\hat{X}_T)]$ to be

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n} f_j$$

From the strong law of large numbers we get

$$\hat{\theta}_n \to \theta \text{ as } n \to \infty$$

Also, if we consider $Z \sim N(0,1)$ and let $Z_{1-\frac{\alpha}{2}}$ be the $(1-\frac{\alpha}{2})$ percentile point of the $N(0,1)$ distribution so that $P(-Z_{1-\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1-\alpha$ we can recover a confidence interval.

Setting $\alpha = 5\%$ we get $Z_{97.5} = 1.96$ which is the approximate value of the $(1-\frac{\alpha}{2}) = 97.5$ percentile point of the normal distribution used in probability and statistics. That is, $1-\alpha = 95\%$ of the area under a normal curve lies within roughly 1.96 standard deviations of the mean, and due to the central limit theorem, this number is therefore used in the construction of approximate 95% confidence intervals. Hence, $P(Z > 1.96) = 0.025$ and $P(Z < 1.96) = 0.975$ and as the normal distribution is symmetric we get P(-1.96 < Z < 1.96) = 0.95.

The approximate $100(1-\frac{\alpha}{2})\%$ confidence interval for θ when n is large is

$$[L(Y), U(Y)] = [\hat{\theta}_n - Z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\theta}_n + Z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}}]$$

where $\hat{\sigma}_n$ is an estimated standard deviation of the process $\hat{X}_t$.

***Remark* B.9.1** *$\hat{\theta}_n$ is an unbiased estimate of $\hat{\theta}$ but it can be a very biased estimate of θ.*

In order to measure that bias we consider the discretisation error D given by

$$D = |E[f(X_T)] - E[f(\hat{X}_T)]|$$

It leads to two types of error, the discretisation error and the statistical error.

- small value of $m \to$ greater discretisation error
- small value of $n \to$ greater statistical error

The values m and n must be chosen judiciously to control the convergence of the estimated $\hat{\theta}_n$ to the true value θ at the minimum computational cost. From the confidence interval, we see that the accuracy of the Monte Carlo pricer is governed by the relation

$$\frac{\hat{\sigma}_n}{\sqrt{n}}$$

so that we can

- multiply n by 4 to decrease the confidence interval by a factor of 2
- reduce the variance

B.10 Some random sampling

For details see text book by Rice [1995]. We assume that the population is of size N and that associated with each member of the population is a numerical value of interest denoted by $x_1, x_2, .., x_N$. The variable x_i may be a numerical variable such as age or weight, or it may take on the value 1 or 0 to denote the presence or absence of some characteristic. The latter is called the dichotomous case.

B.10.1 The sample moments

In general, the population variance of a finite population of size N is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \text{ with } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where μ is the population mean. In the dichotomous case μ equals the proportion p of individuals in the population having the particular characteristic. The population total is

$$\tau = \sum_{i=1}^{N} x_i = N\mu$$

In many practical situations, the true variance of a population is not known a priori and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population and one must estimate the variance of a population from a sample. We take a sample with replacement of n values $X_1, ..., X_n$ from the population, where $n < N$ and such that X_i is a random variable. X_i is the value of the ith member of the sample, and x_i is that of the ith member of the population. The joint distribution of the X_i is determined by that of the x_i. We let $\xi_1, .., \xi_m$ be the elements of a vector corresponding to the possible values of x_i. Since each member of the population is equally likely to be in the sample, we get

$$P(X_i = \xi_j) = \frac{n_j}{N}$$

The sample mean is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Theorem B.10.1 *With simple random sampling* $E[\overline{X}] = \mu$

We say that an estimate is unbiased if its expectation equals the quantity we wish to estimate.

$$\text{Mean squared error } = \text{ variance } + (biased)^2$$

Since $\overline{X}$ is unbiased, its mean square error is equal to its variance.

Lemma B.10.1 *With simple random sampling*

$$Cov(X_i, X_j) = \begin{cases} \sigma^2 \text{ if } i = j \\ -\frac{\sigma^2}{(N-1)} \text{ if } i \neq j \end{cases}$$

It shows that X_i and X_j are not independent of each other for $i \neq j$, but that the covariance is very small for large values of N.

Theorem B.10.2 *With simple random sampling*

$$Var(\overline{X}) = \frac{\sigma^2}{n}\big(1 - \frac{n-1}{N-1}\big)$$

Note, the ratio $\frac{n}{N}$ is called the sampling fraction and

$$p_c = \big(1 - \frac{n-1}{N-1}\big)$$

is the finite population correction. If the sampling fraction is very small we get the approximation

$$\sigma_{\overline{X}} \approx \frac{\sigma}{\sqrt{n}}$$

We estimate the variance on the basis of this sample as

$$\hat{\sigma}^2 = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ with } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is the biased sample variance. The unbiased sample variance is

$$\overline{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ with } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

While the first one may be seen as the variance of the sample considered as a population (over n), the second one is the unbiased estimator of the population variance (over N) where $(n-1)$ is the Bessel's correction. Put another way, we get

$$E[\sigma_n^2] = \frac{n-1}{n}\frac{N}{N-1}\sigma^2 \text{ and } E[\overline{\sigma}_n^2] = \sigma^2$$

That is, the unbiased estimate of σ^2 may be obtained by multiplying σ_n^2 by the factor $\frac{n}{n-1}\frac{N-1}{N}$. If the population is large relative to n, the dominant bias is due to the term $\frac{n-1}{n}$. In general one set n between 50 and 180 days. Being a function of random variables, the sample variance is itself a random variable, and it is natural to study its distribution. In the case that y_i are independent observations from a normal distribution, Cochran's theorem shows that $\overline{\sigma}_n^2$ follows a scaled chi-squared distribution

$$(n-1)\frac{\overline{\sigma}_n^2}{\sigma^2} \approx \chi_{n-1}^2$$

If the conditions of the law of large numbers hold for the squared observations, $\overline{\sigma}_n^2$ is a consistent estimator of σ^2 and the variance of the estimator tends asymptotically to zero. The obtained standard deviation $\tilde{\sigma}_N$ is an estimator of σ called the historical volatility.

***Corollary* 8** *An unbiased estimate of* $Var(\overline{X})$ *is*

$$s_{\overline{X}}^2 = \frac{\hat{\sigma}^2}{n}\frac{n}{n-1}\frac{N-1}{N}\frac{N-n}{N-1} = \frac{s^2}{n}(1 - \frac{n}{N})$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

One can compute the sample mean and sample variance recursively as follow. Starting from the identities

$$(n+1)\overline{X}_{n+1} = n\overline{X}_n + X_{n+1}$$

and

$$(n+1)\big(\sigma_{n+1}^2 + (\overline{X}_{n+1})^2\big) = n\big(\sigma_n^2 + (\overline{X}_n)^2\big) + X_{n+1}^2$$

we get the recursive sample mean

$$\overline{X}_{n+1} = \overline{X}_n + \frac{X_{n+1} - \overline{X}_n}{n+1} \tag{B.10.9}$$

and the recursive sample variance

$$\sigma_{n+1}^2 = \sigma_n^2 + (\overline{X}_n)^2 - (\overline{X}_{n+1})^2 + \frac{X_{n+1}^2 - \sigma_n^2 - (\overline{X}_n)^2}{n+1} \tag{B.10.10}$$

such that σ_{n+1}^2 only depends on σ_n^2, $\overline{X}_n$, $\overline{X}_{n+1}$ and X_{n+1}. If we normalise the data with a constant K such that the ith value is $\frac{X_i}{K}$, we let $\tilde{X} = \frac{1}{n}\sum_{i=1}^n \frac{X_i}{K}$ be the sample mean and $\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (\frac{X_i}{K} - \tilde{X})^2$ be the sample variance. Then, the recursive sample mean becomes

$$\tilde{X}_{n+1} = \tilde{X}_n + \frac{1}{n+1}\big(\frac{X_{n+1}}{K} - \tilde{X}_n\big) \tag{B.10.11}$$

and the recursive sample variance becomes

$$\tilde{\sigma}_{n+1}^2 = \tilde{\sigma}_n^2 + (\tilde{X}_n)^2 - (\tilde{X}_{n+1})^2 + \frac{1}{n+1}\Big(\frac{X_{n+1}^2}{K^2} - \tilde{\sigma}_n^2 - (\tilde{X}_n)^2\Big) \tag{B.10.12}$$

B.10.2 Estimation of a ratio

We consider the estimation of a ratio, and assume that for each member of a population, two values, x and y, may be measured. The ratio of interest is

$$r = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \frac{\mu_y}{\mu_x}$$

Assuming that a sample is drawn consisting of the pairs (X_i, Y_i), the natural estimate of r is $R = \frac{\overline{Y}}{\overline{X}}$. Since R is a nonlinear function of the random variables $\overline{X}$ and $\overline{Y}$, there is no closed form for $E[R]$ and $Var(R)$ and we must approximate them by using $Var(\overline{X})$, $Var(\overline{Y})$, and $Cov(\overline{X}, \overline{Y})$. We define the the population covariance of x and y as

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$$

One can show that

$$Cov(\overline{X}, \overline{Y}) = \frac{\sigma_{xy}}{n} p_c$$

Theorem B.10.3 *With simple random sampling, the approximate variance of $R = \frac{\overline{Y}}{\overline{X}}$ is*

$$Var(R) \approx \frac{1}{\mu_x^2}(r^2\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 - 2r\sigma_{\overline{XY}}) = \frac{1}{n}p_c\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})$$

The population correlation coefficient given by

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

is a measure of the strength of the linear relationship between the x and the y values in the population. We can express the variance in the above theorem in terms of ρ as

$$Var(R) \approx \frac{1}{n} p_c \frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y)$$

so that strong correlation of the same sign as r decreases the variance. Further, for small μ_x we get large variance, since small values of $\overline{X}$ in the ratio $R = \frac{\overline{Y}}{\overline{X}}$ cause R to fluctuate wildly.

Theorem B.10.4 *With simple random sampling, the expectation of R is given approximately by*

$$E[R] \approx r + \frac{1}{n} p_c \frac{1}{\mu_x^2}(r^2\sigma_x^2 - \rho\sigma_x\sigma_y)$$

so that strong correlation of the same sign as r decreases the bias, and the bias is large if μ_x is small. In addition, the bias is of the order $\frac{1}{n}$, so its contribution to the mean squared error is of the order $\frac{1}{n^2}$ while the contribution of the variance is of order $\frac{1}{n}$. Hence, for large sample, the bias is negligible compared to the standard error of the estimate. For large samples, truncating the Taylor Series after the linear term provides a good approximation, since the deviations $\overline{X} - \mu_x$ and $\overline{Y} - \mu_y$ are likely to be small. To this order of approximation, R is expressed as a linear combination of $\overline{X}$ and $\overline{Y}$, and an argument based on the central limit theorem can be used to show that R is approximately normally distributed, and confidence intervals can be formed for r by using the normal distribution. In order to estimate the standard error of R, we substitute R for r in the formula of the above theorem where the x and y population variances are estimated by s_x^2 and s_y^2. The population covariance is estimated by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - n\overline{XY})$$

and the population correlation is estimated by

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}$$

The estimated variance of R is thus

$$s_R^2 = \frac{1}{n} p_c \frac{1}{\overline{X}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$$

and the approximate $100(1-\alpha)\%$ confidence interval for r is $R \pm z(\frac{\alpha}{2})s_R$.

B.10.3 Stratified random sampling

The population is partitioned into subpopulations, or strata, which are then independently sampled. We are interested in obtaining information about each of a number of natural subpopulations in addition to information about the population as a whole. It guarantees a prescribed number of observations from each subpopulation, whereas the use of a simple random sample can result in underrepresentation of some subpopulations. Further, the stratified sample mean can be considerably more precise than the mean of a simple random sample, especially if there is considerable variation between strata.

We will denote by N_l, where $l = 1, .., L$ the population sizes in the L strata such that $N_1 + N_2 + .. + N_L = N$ the total population size. The population mean and variance of the lth stratum are denoted by μ_l and σ_l^2 (unknown). The overall population mean can be expressed in terms of the μ_l as follows

$$\mu = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l} x_{il} = \frac{1}{N}\sum_{l=1}^{L} N_l\mu_l = \sum_{l=1}^{L} W_l\mu_l$$

where x_{il} denotes the ith population value in the lth stratum and $W_l = \frac{N_l}{N}$ is the fraction of the population contained in the lth stratum.

Within each stratum a simple random sample of size n_l is taken. The sample mean in stratum l is denoted by

$$\overline{X}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il}$$

where X_{il} denotes the ith observation in the lth stratum. By analogy with the previous calculation we get

$$\overline{X}_s = \sum_{l=1}^{L} \frac{N_l\overline{X}_l}{N} = \sum_{l=1}^{L} W_l\overline{X}_l$$

Theorem B.10.5 *The stratisfied estimate, $\overline{X}_s$ of the population mean is unbiased. That is, $E[\overline{X}_s] = \mu$.*

Since we assume that the samples from different strata are independent of one another and that within each stratum a simple random sample is taken, the variance of $\overline{X}_s$ can easily be calculate.

Theorem B.10.6 *The variance of the stratisfied sample mean is given by*

$$Var(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \frac{1}{n_l}(1 - \frac{n_l - 1}{N_l - 1})\sigma_l^2 \tag{B.10.13}$$

Note, $\frac{n_l-1}{N_l-1}$ represents the finite population corrections. If the sampling fractions within all strata are small ($\frac{n_l-1}{N_l-1} << 1$), we get the approximation

$$Var(\overline{X}_s) \approx \sum_{l=1}^{L} \frac{W_l^2}{n_l}\sigma_l^2 \tag{B.10.14}$$

The estimate of σ_l^2 is given by

$$S_l^2 = \frac{1}{n_l - 1}\sum_{i=1}^{n_l}(X_{il} - \overline{X}_l)^2$$

and $Var(\overline{X}_s)$ is estimated by

$$S_{\overline{X}_s}^2 = \sum_{l=1}^{L} W_l^2 \frac{1}{n_l}(1 - \frac{n_l - 1}{N_l - 1})S_l^2$$

The question that naturally arises is how to choose $n_1, .., n_L$ to minimise $Var(\overline{X}_s)$ subject to the constraint $n_1 + .. + n_L = n$. Ignoring the finite population correction within each stratum we get the Neyman allocation.

Theorem B.10.7 *The sample sizes $n_1, .., n_L$ that minimise $Var(\overline{X}_s)$ subject to the constraint $n_1 + .. + n_L = n$ are given by*

$$n_l = n\frac{W_l\sigma_l}{\sum_{k=1}^{L} W_k\sigma_k}$$

where $l = 1, .., L$.

This theorem shows that those strata for which $W_l\sigma_l$ is large should be sampled heavily. If W_l is large, the stratum contains a large fraction of the population. If σ_l is large, the population values in the stratum are quite variable, and a relatively large sample size must be used. Substituting the optimal values of n_l into the Equation (B.10.13) we get the following corollary.

Corollary 9 *Denoting by $\overline{X}_{so}$ the stratified estimate using the optimal Neyman allocation and neglecting the finite population correction, we get*

$$Var(\overline{X}_{so}) = \frac{1}{n}\left(\sum_{l=1}^{L} W_l\sigma_l\right)^2$$

The optimal allocations depend on the individual variances of the strata, which generally will not be known. A simple alternative method of allocation is to use the same sampling fraction in each stratum

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = ... = \frac{n_L}{N_L}$$

which holds if

$$n_l = n\frac{N_L}{N} = nW_l \tag{B.10.15}$$

$l = 1, .., L$. This method is called the Proportional allocation. The estimate of the population mean based on proportional allocation is

$$\overline{X}_{sp} = \sum_{l=1}^{L} W_l\overline{X}_l = \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n_l} X_{il}$$

since $\frac{W_l}{n_l} = \frac{1}{n}$. This estimate is simply the unweighted mean of the sample values.

Theorem B.10.8 *With stratified sampling based on proportional allocation, ignoring the finite population correction, we get*

$$Var(\overline{X}_{sp}) = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2$$

We can compare $Var(\overline{X}_{so})$ and $Var(\overline{X}_{sp})$ to define when optimal allocation is substantially better than proportional allocation.

Theorem B.10.9 *With stratified random sampling, the difference between the variance of the estimate of the population mean based on proportional allocation and the variance of that estimate based on optimal allocation is, ignoring the finite population correction,*

$$Var(\overline{X}_{sp}) - Var(\overline{X}_{so}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \overline{\sigma})^2$$

where

$$\overline{\sigma} = \sum_{l=1}^{L} W_l \sigma_l$$

As a result, if the variances of the strata are all the same, proportional allocation yields the same results as optimal allocation. The more variable these variances are, the better it is to use optimal allocation.

We can also compare the variance under simple random sampling with the variance under proportional allocation. Neglecting the finite population correction, the variance under simple random sampling is $Var(\overline{X}) = \frac{\sigma^2}{n}$. We first need a relationship between the overall population variance σ^2 and the strata variances σ_l^2. The overall population variance may be expressed as

$$\sigma^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_L} (x_{il} - \mu)^2$$

Further, using the relation

$$(x_{il} - \mu)^2 = (x_{il} - \mu_l)^2 + 2(x_{il} - \mu_l)(\mu_l - \mu) + (\mu_l - \mu)^2$$

and realising that when both sides are summed over l, the middle term on the right-hand side becomes zero, we have

$$\sum_{i=1}^{N_L} (x_{il} - \mu)^2 = \sum_{i=1}^{N_L} (x_{il} - \mu_l)^2 + N_L(\mu_l - \mu)^2 = N_L \sigma_l^2 + N_L(\mu_l - \mu)^2$$

Dividing both sides by N and summing over l, we have

$$\sigma^2 = \sum_{l=1}^{L} W_l \sigma_l^2 + \sum_{l=1}^{L} W_l (\mu_l - \mu)^2$$

Substituting this expression for σ^2 into $Var(\overline{X})$ and using the formula for $Var(\overline{X}_{sp})$ completes the proof of the following theorem.

Theorem B.10.10 *The difference between the variance of the mean of a simple random sample and the variance of the mean of a stratified random sample based on proportional allocation is, neglecting the finite population correction,*

$$Var(\overline{X}) - Var(\overline{X}_{sp}) = \frac{1}{n} \sum_{l=1}^{L} W_l (\mu_l - \mu)^2$$

Thus, stratified random sampling with proportional allocation always gives a smaller variance than does simple random sampling, providing that the finite population correction is ignored. Typically, stratified random sampling can result in substantial increases in precision for populations containing values that vary greatly in size.

In order to construct the optimal number of strata, the population values themselves (which are unknown) would have to be used. Stratification must therefore be done on the basis of some related variable that is known or on the results of earlier samples.

According to the Neyman-Pearson Paradigm, a decision as to weather or not to reject H_0 in favour of H_A is made on the basis of $T(X)$, where X denotes the sample values and $T(X)$ is a statistic.

The statistical properties of the methods are relevant if it is reasonable to model the data stochastically. There exists methods that are sample analogues of the cumulative distribution function of a random variable. It is useful in displaying the distribution of sample values.

Given $x_1, .., x_n$ a batch of numbers, the empirical cumulative distribution function (ecdf) is defined as

$$F_n(x) = \frac{1}{n}(nox_i \leq x)$$

where $F_n(x)$ gives the proportion of the data less than or equal to x. It is a step function with a jump of height $\frac{1}{n}$ at each point x_i. The ecdf is to a sample what the cumulative distribution is to a random variable. We now consider some of the elementary statistical properties of the ecdf when $X_1, .., X_n$ is a random sample from a continuous distribution function F. We choose to express F_n as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

The random variables $I_{(-\infty,x]}(X_i)$ are independent Bernoulli random variables

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1 \text{ with probability } F(x) \\ 0 \text{ with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable with the first two moments being

$$\begin{aligned} E[F_n(x)] &= F(x) \\ Var(F_n(x)) &= \frac{1}{n}F(x)[1 - F(x)] \end{aligned}$$

B.10.4 Geometric mean

The use of a geometric mean normalises the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. The geometric mean applies only to positive numbers. It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment. The geometric mean of a data set $\{a_1, a_2, .., a_n\}$ is given by

$$(\prod_{i=1}^{n} a_i)^{\frac{1}{n}}$$

The geometric mean of a data set is less than the data set's arithmetic mean unless all members of the data set are equal, in which case the geometric and arithmetic means are equal. By using logarithmic identities to transform the formula, the multiplications can be expressed as a sum and the power as a multiplication

$$(\prod_{i=1}^{n} a_i)^{\frac{1}{n}} = e^{\frac{1}{n}\sum_{i=1}^{n} \ln a_i}$$

This is sometimes called the log-average. It is simply computing the arithmetic mean of the logarithm-transformed values of a_i (i.e., the arithmetic mean on the log scale) and then using the exponentiation to return the computation to the original scale.

Appendix C

Introducing random number generators

C.1 The random number generators

C.1.1 The need to generate independent uniform random variables

When simulating stochastic models in finance, we must assume random variables from different probability distributions. To do so we need to generate a sequence of independent uniform variates and transform them appropriately. A simple way of obtaining independent random variables $X_1, X_2, ...$ with distribution function F from a sequence of i.i.d. $U(0,1)$ random variables $U_1, U_2, ...$ is to define

$$X_j = F^{-1}(U_j) = \min\{x|F(x) \geq U_j\}$$

For instance, each replication in a Monte Carlo simulation can be interpreted as the result of applying a series of transformations to an input sequence of uniformly distributed random variables U_i with $i = 1, .., d$ producing the output $f(U_1, .., U_d)$. In option pricing theory, f is the result of a transformation converting the U_i to normal random variables, which are transformed to paths of underlying assets, which are transformed to the discounted payoff of an option. The objective being to compute

$$E[f(U_1, .., U_d)] = \int_{[0,1)^d} f(x)dx$$

which can be approximated by

$$\int_{[0,1)^d} f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i) \tag{C.1.1}$$

where n is the number of simulation. Even though the random variables U_i with $i = 1, .., d$ can not be realised, they can be approximated. That is, random number generator (RNGs) are small computer programs producing sequences of numbers behaving as if they were generated randomly from a specified probability distribution. Since they are not truly random, they are called pseudorandom numbers. Other generators designed to produce numbers with random properties are the quasirandom numbers.

C.1.2 Defining random number generators

C.1.2.1 The pseudorandom numbers

A pseudorandom number generator (PRNG) is an algorithm for generating a sequence of numbers that approximates the properties of random numbers. The sequence is not truly random in that it is completely determined by a relatively

small set of initial vales called the state. That is, the numbers are generated quickly and easily by simple and deterministic computer program, except for its initial state which can be randomly selected. Some parameters of the generator may also be selected randomly, and become part of the state. For simulation purposes, we need speed, small memory requirement, and good statistical properties. Common classes of these algorithms are Linear Congruential Generators (LCGs), Lagged Fibonacci Generators (LFGs) etc ... Fast uniform random number generators with extremely long periods have been defined on linear recurrences modulo 2. Recent instances of pseudorandom algorithms include Blum Blum Shub (see Blum et al. [1986]), Fortuna, the twisted GFSR and the Mersenne Twister. We let the reader refer to L'Ecuyer [2007], where a review of the basic principles underlying the design of a uniform random number generators is performed. For simplicity of use, we recommand the combined multiple recursive random number generators (CMRGs) introduced by L'Ecuyer [1998]. Alternatively, one can use a Linear Feedback Shift (LFSR), or Tausworthe random number generators, which are based on linear recurrences modulo two with primitive characteristic polynomials. For details, see L'Ecuyer [1991] and Panneton et al. [2006]. A PRNG can be started from any arbitrary starting state using a seed state. It will always produce the same sequence thereafter when initialised with that state. The maximum length of the sequence before it repeats itself is determined by the size of the state measued in bits. In general, the length of the maximum period doubles with each bit of state added. So, if the internal state contains n bits, its period is no longer than 2^n results. In practice, the output from many PRNGs exhibit artifacts which cause them to fail statistical pattern detection tests. The Mersenne Twister (see Matsumoto et al. [1998]) algorithm avoids many of the problems with earlier generators. It has a period of $2^{19937}-1$ and is proven to be equidistributed in up to 623 dimensions (for 32-bit) and run faster than other statistically reasonable generators. However, this type of generators have equidistributions far from optimal in large dimensions. The WELL (see Panneton et al. [2006]) approximate random values evolving in a more chaotic way than the Mersenne Twister, reducing the impact of persistent dependencies among successive output values. Saito et al. [2006] proposed the single instruction, multiple data-oriented fast Mersenne Twister (SFMT) as a variant of MT designed to be fast on 128-bit SIMD. It is about twice as fast as the MT with better equidistribution of v-bit accuracy, but worse than WELL.

C.1.2.2 The quasirandom numbers

A low-discrepancy sequence is a sequence with the property that for all values of N, its sequence $X_1, .., X_N$ has a low discrepancy. That is, if the number of points in the sequence falling into an arbitrary set B (hyperspheres, hypercubes etc ..) is close to proportional to the measure of B, as would happen on average in the case of a uniform distribution. Put another way, the goal of low-discrepancy methods is to construct points x_i such that the error in Equation (C.1.1) is small for a large class of integrands f. It is equivalent to choosing the points x_i to fill the hypercube uniformly. Hence, the low-discrepancy sequences are also called quasi-random or sub-random sequences due to their common use as a replacement of uniformly distributed random numbers. The quasi modifier is used to denote that the values are neither random nor pseudorandom but they share some properties of random variables. Low-discrepancy methods have the potential to accelerate convergence from $\mathcal{O}(\frac{1}{\sqrt{n}})$ with n points generated in the classical Monte Carlo to nearly $\mathcal{O}(\frac{1}{n})$ convergence. More precisely, under the appropriate conditions, the error is $\mathcal{O}(\frac{1}{n^{1-\epsilon}})$ for all $\epsilon > 0$. However, as opposed to ordinary Monte Carlo simulation where the vectors $(U_1, .., U_d), (U_{d+1}, .., U_{2d}), ..$ drawn from a sequence of uniforms $U_1, U_2, ..$ produce an i.i.d. sequence of points from the d-dimensional hypercube, in QMC, the construction of the points x_i depends explicitly on the dimension of the problem. That is, the vectors x_i in $[0, d)^d$ can not be constructed by taking sets of d consecutive elements from a scalar sequence. The dependence of QMC methods on problem dimension is one of the features that differentiate it most from ordinary Monte Carlo.

C.2 Presenting PRNGs

C.2.1 Introducing the problem

We want successive output values of an PRNG, $u_0, u_1, ...$, to imitate independent random variables from the uniform distribution over the interval $[0, 1]$ (i.i.d. $U(0, 1)$), or over the two-element set $\{0, 1\}$ (independent random bits).

In both cases, we denote the hypothesis of perfect behaviour by $\mathcal{H}_0$. Under the i.i.d. $U(0,1)$ hypothesis, any pre-specified sequence of bits must be a sequence of independent random bit, such that statistical tests for bit sequences can be used as well for testing the null hypothesis. In the $U(0,1)$ case, $\mathcal{H}_0$ states that for each integer $n > 0$, the vector $(u_0, ..., u_{n-1})$ is uniformly distributed over the n-dimensional unit cube $[0,1]^n$, which is not formally true since the vectors always take their values only from the finite set

$$\Psi_n = \{(u_0, ..., u_{n-1}) : s_0 \in S\}$$

of all n-dimensional vectors of n successive values produced by the generator from all its possible initial states s_0 (or seeds). Thus, the set Ψ_n is a finite subset of the unit cube $[0,1]^n$, which can be viewed as the sample space from which the n-dimensional points are drawn, rather than drawing them uniformaly from $[0,1]^n$. The cardinality of this set can not exceed the number of admissible seeds for the PRNG. Assuming the seed is randomly chosen, the vectors are actually generated over Ψ_n to approximate the uniform distribution over $[0,1]^n$, suggesting that Ψ_n (the sample space) should be very evenly distributed over the unit cube. For Ψ_n to provide a dense and uniform coverage of the hypercube, for small and moderate values of n, S must have large cardinality. When we consider bits, the null hypothesis $\mathcal{H}_0$ is not formally true as soon as the length n of the sequence exceeds the number b of bits in the generator's state, since the number of distinct sequences of bits produced can not exceed 2^b. For $b < n$, the fraction of all sequences that can be visited is at most 2^{b-n}. We must therefore make sure that the sequences that can be visited are uniformly scattered in the set of all 2^n possible sequences. Even though the number of tests detecting deficiencies of the null hypothesis is infinite, there is no universal test that can guarantee, when passes, that a given generator is fully reliable for all kinds of simulations. Testing RNGs is an heuristic process, and none of them can pass every conceivable statistical test (see Knuth [1981], L'Ecuyer et al. [2007b]). In general, PRNGs with very long periods, good structure of their set Ψ_n, repeatable, portable, and based on recurrences not too simplistic, pass most reasonable tests, whereas PRNGs with short periods, or bad structures, are usually easy to crack by standard statistical tests. When a generator starts failing a test decisively, the p-value of the test tends to converge exponentially fast to 0 or 1 as a function of the sample size when the latter increases further.

C.2.2 Describing a few generators

C.2.2.1 Defining generators

We follow L'Ecuyer [2006] and define a generator as

Definition C.2.1 *A generator is a structure $\mathcal{G} = (S, s_0, T, U, G)$ where S is a finite set of states, $s_0 \in S$ is the initial state, $T : S \to S$ is the transition function, U is a finite set of output symbols, and $G : S \to U$ is the output function.*

A generator starts from the initial state s_0 (called the seed), let $u_0 = G(s_0)$, then for $n = 1, 2, ...$, we let the state evolves according to the recurrence $s_n = T(s_{n-1})$ and the output at step n satisfies $u_n = G(s_n) \in U$. The sequence $\{u_n\}$ is the output of the generator and the elements are called the observations. The set U is either a set of integers of the form $\{0, ..., m-1\}$, or a finite set of values between 0 and 1 to approximate the $U(0,1)$ distribution. Most of the RNGs can be expressed by linear recurrences in modular arithmetic, over a finite set S. Thus, the RNG must eventually return to a previously visited state, that is, $s_{l+j} = s_l$ for some $l \geq 0$ and $j > 0$, so that $s_{n+j} = s_n$ and $u_{n+j} = u_n$ for all $n \geq l$. The smallest $j > 0$ for which this happens is the period length ρ. If the state can be represented with b bits of memory, then $\rho \leq 2^b$. In general, the transition function has the form $T(s) = \alpha s$, where α, $s \in S$. In that case, S has the form $S = \mathcal{F}_{m^k}$, where $m = p^e$, p is prime and e, k are positive integers. For $k = 1$, we get $S = \mathcal{F}_m$, the finite field with m elements. Note, $\mathcal{F}_m$ exists if and only if m is a power of a prime. When m is prime, we can identify $\mathcal{F}_m$ with the set $\mathbb{Z}_m = \{0, 1, ..., m-1\}$ on which arithmetic operations are performed modulo m (see Lidl et al. [1986]). Assuming m to be prime and $\alpha \in S = \mathcal{F}_{m^k}$, then the state s_n of a generator evolves in $\mathcal{F}_{m^k}$ as

$$s_n = \alpha s_{n-1}$$

We let

$$P(z) = z^k - a_1 z^{k-1} - ... - a_k \in \mathcal{F}_m(z)$$

be the minimal polynomial of α over $\mathcal{F}_m$. Then, in $\mathcal{F}_{m^k}$, one has $P(\alpha) = 0$, that is,

$$\alpha^n = a_1 \alpha^{n-1} + ... + a_k \alpha^{n-k}$$

and k is called the order of the recurrence. We define the output function as a composition of the form $G = G_1 \circ G_2$ where $G_1 : \mathcal{F}_{m^k} \to \mathcal{F}_m$ is a linear form over $\mathcal{F}_{m^k}$, and $G_2 : \mathcal{F}_m \to [0, 1]$. If $x_n = G_1(s_n) \in \mathcal{F}_m$, one has

$$x_n = a_1 x_{n-1} + ... + a_k x_{n-k}$$

in $\mathcal{F}_m$. The transformation G_2 is sometimes defined by $G_2(x) = \frac{x}{m}$, where $x \in \mathcal{F}_m$ is identified with its representative in $\mathbb{Z}_m$. The sequence $\{x_n\}$ is called a linear recurring sequence with characteristic polynomial $P(z)$.

C.2.2.2 Linear generators

The Linear Congruential Generator (LCG) is one of the best known PRNG where $k = 1$, m prime, and it is defined by the recurrence relation

$$x_n = (ax_{n-1} + c) \mod m$$

where $u_n = \frac{x_n}{m}$ is the sequence of pseudorandom values at the nth step with $0 < a < m$ the multiplier, $0 \leq c < m$ the increment and $m > 0$ is the modulus. The most efficient LCG have an m equal to a power of 2, most often $m = 2^{32}$ or $m = 2^{64}$ as it allows for the modulus operation to be computed by merly truncating the rightmost 32 or 64 bits. LCGs are fast and require minimal memory (32 or 64 bits) to retain state. If higher quality random numbers are needed and sufficient memory is available ($\approx$ 2 kilobytes) then the Mersenne Twister algorithm is a preferred choice. The Multiple Recursive Generators (MRGs) based on a linear recurrence of order $k \geq 1$, modulo m satisfies

$$x_n = \left(a_1 x_{n-1} + ... + a_k x_{n-k}\right) \mod m \tag{C.2.2}$$

The generator's state at step n is the vector $s_n = (x_n, ..., x_{n+k-1}) \in \mathbb{Z}_m^k$, which can be transformed into the output value $u_n \in [0, 1]$ by $u_n = G(s_n) = \frac{x_n}{m}$. A key issue for implementation is the way

$$a_i x \mod m$$

is computed when m is large. As a first approach one can use approximate factoring, or alternatively, one can compute the product and the division by m directly in floating-point arithmetic. In 64-bit floating point standard, all integers up to 2^{53} are represented exactly in floating point, so that the latter works if $am < 2^{53}$. A third approach called the powers-of-2 decomposition, assumes that a is a sum or a difference of a small number of powers of 2, such as $a = \pm 2^q \pm 2^r$ which we will detail later on. The point set Ψ_n produced by an MRG has a lattice structure, and it is measured via a figure of merit for the quality of that lattice. This is the spectral test. The period is at most m and for some choice of a, much less than that. For properly chosen a_i's, the sequence has maximal period length $\rho = m^k - 1$, which can be achieved with only two non-zero a_i's (see Knuth [1981]), where the primitive polynomial becomes $P(z) = z^k - a_r z^{k-r} - a_k$ and the recurrence satisfies

$$x_n = \left(a_r x_{n-r} + a_k x_{n-k}\right) \mod m \tag{C.2.3}$$

Further, one can take $m = 2^e$ for $e > 1$ and obtain a long period by considering linear recurrence with a carry

$$\begin{aligned} x_n &= \left(a_1 x_{n-1} + ... + a_k x_{n-k} + c_{n-1}\right) \mod b \\ c_n &= \left(a_1 x_{n-1} + ... + a_k x_{n-k} + c_{n-1}\right) \text{ div } b \\ u_n &= \frac{x_n}{b} \end{aligned}$$

where div is the integer division, b can be a power of two, and c_n is the carry at step n. This is the Multiply-with-Carry (MWC) generator which is approximately equivalent to an LCG with modulus $m = \sum_{i=0}^{k} a_i b^i$ where $a_0 = -1$ and multiplier a equal to the inverse of b modulo m (see Couture et al. [1997]). Lagged-Fibonacci generators are MRGs when the selected operation is addition or subtraction, but multiplicative LF generators are not. An effective way of implementing high quality MRGs is to combine two (or more) of them by adding their outputs modulo 1. Further, we can use the above recurrence, define $s_n = (x_{ns}, ..., x_{ns+k-1})$ and let the output be a digital expansion in base m given by

$$u_n = G(s_n) = \sum_{j=1}^{L} x_{ns+j-1} m^{-j}$$

where the step size s and the length $L \leq k$ are positive integers. Computing s_n from s_{n-1} involves performing s steps of the recurrence in Equation (C.2.2). Such a generator is called a digital multistep sequence. If the recurrence in Equation (C.2.2) has a full period $\rho = m^k - 1$ and s is coprime to ρ, then the digital multistep sequence is also periodic with a period $\rho = m^k - 1$. The MRG is a special case with $s = L = 1$. For $m = p = 2$ we get the Tausworthe generator where the output values $\{u_n, n \geq 0\}$ are constructed by taking blocks of L successive bits from the binary sequence with spacings of $s - L \geq 0$ bits between the blocks (see Tausworthe [1965], Knuth [1981]). This generator is also called a linear feedback shift register (LFSR).

C.2.3 Equidistribution and measures of quality

The set Ψ_t of t-dimensional vectors of successive output values produced by a generator, from all its possible initial states, is given by

$$\Psi_t = \{u_n = (u_n, ..., u_{n+t-1}) | n \geq 0, (x_0, ..., x_{k-1}) \in \mathbb{Z}_m^k\}$$

which is the intersection of a lattice L_t with the unit hypercube $[0,1)^t$. Hence, the points of Ψ_t lie in successive parallel hyperplanes at a distance d_t of each other, where $\frac{1}{d_t}$ is the Euclidean length of the shortest nonzero vector in the dual lattice of L_t. Computing d_t is called the spectral test. One can use the figure of merit

$$M_T = \min_{2 \leq t \leq T} S_t$$

for some integer T, where $S_t = (\rho_t m^{\frac{k}{t}} d_t)^{-1}$, and for $t \leq 8$, ρ_t is the γ_t defined in Knuth [1981], while for $t > 8$, $\rho_t = e^{\frac{R(t)}{t}}$ where $R(t)$ is Roger's bound on the density of sphere packings. Thus, S_t and M_T are always in the range $[0,1]$ and we seek generators with M_T close to 1, meaning that Ψ_t is evenly distributed over the unit hypercube, for all $t \leq T$. Further, dividing the interval $[0,1)$ into 2^l equal segments determine a partition of the unit hypercube $[0,1)^t$ into 2^{tl} cubic cells of equal size, called a (t,l)-equidissection in base 2. The set Ψ_t is (t,l)-equidistributed if each cell contains the same number of points of Ψ_t, 2^{k-tl}, which is only possible if $l \leq L$ and $tk \leq k$. If Ψ_t is (t, l_t^*)-equidistributed for $0 \leq t \leq k$, where $l_t^* = \min(L, \lfloor \frac{k}{t} \rfloor)$, then the output sequence is called maximally-equidistributed (ME). An ME sequence having more cells than points is called collision-free (CF), and ME-CF sequences have point sets very evenly distributed in all dimensions (in terms of equidissections). For non-ME generators, we let t_l be the largest dimension t for which Ψ_t is (t,l)-equidistributed, and define the dimension gap for l bits of resolution as

$$\delta_l = t_l^* - t_l$$

where $t_l^* = \lfloor \frac{k}{l} \rfloor$ is an upper bound on the best possible value of t_l. Panneton and L'Ecuyer defined two measures of uniformity, the worst-case dimension gap

$$\Delta_\infty = \max_{1 \leq l \leq w} \delta_l$$

and the sum of dimension gaps

$$\Delta_1 = \sum_{l=1}^{w} \delta_l$$

Further, good linear generators over $\mathcal{F}_2$ must have characteristic polynomials $P(z)$ with number of nonzero coefficients in the vicinity of $\frac{k}{2}$. Thus, as another quality criterion, we consider N_1 the number of nonzero coefficients in $P(z)$.

C.2.4 Combining linear generators

Combining parallel multiple recursive sequences provides an efficient way of implementing random number generators with long periods and good structural properties. However, combining elements with distinct prime moduli leads to another MRG with non-prime modulus m equal to the product of the moduli of the components, and the period can be up to half the product of the component's periods. One approach combines J copies of Equation (C.2.2)

$$x_{j,n} = \left(a_{j,1}x_{j,n-1} + ... + a_{j,k}x_{j,n-k}\right) \mod m_j \ , j = 1, .., J$$

where $\{x_{j,n}, n \geq 0\}$ is the jth copy, the m_j are distinct primes and the jth recurrence has order k and period length $\rho_j = m_j^k - 1$.

C.2.4.1 The combined MRGs

Letting $\delta_1, ..., \delta_j$ be arbitrary integers such that δ_j is relatively prime to m_j for each j, L'Ecuyer [1996] proposed

$$w_n = \left(\sum_{j=1}^{J} \delta_j \frac{x_{j,n}}{m_j}\right) \mod 1$$

and

$$\begin{aligned} z_n &= \left(\sum_{j=1}^{J} \delta_j x_{j,n}\right) \mod m_1 \\ \tilde{u}_n &= \frac{z_n}{m_1} \end{aligned}$$

where the sequences $\{w_n, n \geq 0\}$ and $\{\tilde{u}_n, n \geq 0\}$ define two diffent CMRGs. The first CMRG is exactly equivalent to an MRG with modulus $m = \prod_{j=1}^{J} m_j$, and the second CMRG is approximately the same as the first one.

C.2.4.2 The combined LFSRs

As an alternative to LFSR, we can have J copies of Equation (C.2.2) running in parallel, with different initial values, and use one copy for each digit of the fractional expansion of u_n. Letting $\{x_{j,n}, n \geq 0\}$ be the jth LFSR copy, if $x_{j,n} = x_{n+d_j}$, then

$$u_n = G(s_n) = \sum_{j=1}^{L} x_{j,n} m^{-j} = \sum_{j=1}^{L} x_{n+d_j} m^{-j}$$

In the special case where $m = 2$ and Equation (C.2.3) is used, then the generator is called a generalised feedback shift refister (GFSR) generator. L'Ecuyer [1996b] combined several trinomial-based LFSR generators of relatively prime period lengths, by bitwise xor, giving another LFSR. For $x_n = (x_{1,n} + ... + x_{J,n}) \mod 2$, if $\{u_{j,n}, n \geq 0\}$ is the output sequence from the jth LFSR, then

$$u_n = u_{1,n} \oplus ... \oplus u_{J,n}$$

where $\oplus$ is the bitwise exclusive-or in the binary expansion. We call $\{x_n\}$ the combined LFSR sequence with reducible characteristic polynomial $P(z) = P_1(z)...P_J(z)$, and $\{u_n\}$ a combined LFSR generator with period length $\rho = (2^{k_1} - 1) \times ... \times (2^{k_J} - 1)$. This way, the polynomial $P(z)$ contains many more nonzero coefficients. For instance, combining J trinomial-based LFSRs, and $3^J < k$, we get up to 3^J nonzero coefficients. We can therefore efficiently build generators with values of N_1 up to a few hundreds on 32-bit computers.

C.2.4.3 Results

Since the CMRGs above are special implementations of an MRG, they can be analysed with the spectral test. For $J = 2, 3$, $k = 3, 5, 7$ and prime moduli slightly smaller than 2^e for $e = 31, 32, 63, 64, 127, 128$, L'Ecuyer [1998] searched for good values of M_T for $T = 8, 16, 32$. Some coefficients $a_{j,i}$ must be constrained. For instance, some coefficients should be set to zero and

- (B) the product $a_{j,i}(m_j - 1)$ is less than 2^{53}.
- (C) the coefficient $a_{j,i}$ satisfies $a_{j,i}(m_j \mod a_{j,i}) < m_j$.

Doing so, the combination can reach good figures of merit M_T. After comparing different implementation, L'Ecuyer proposed the following models

- MRG32ka, it has two components of order three with period length $\approx 2^{191}$
- MRG32k5a, it has two components of order five with period length $\approx 2^{319}$
- MRG63k3a, it is a 64-bit integer arithmetic with two components of order three with period length $\approx 2^{377}$

The implementation is available from the author's website

`http://www.iro.umontreal.ca/~lecuyer`

L'Ecuyer [1996b] proposed ME-CF combined LFSR generators with length $L = 32$ and $L = 64$, whose components have recurrences with primitive trinomials of the form $P_j(z) = z^{k_j} - z^{q_j} - 1$ with $0 < 2_{q_j} < k_j$, and with step size s_j satisfying $0 < s_j \leq k_j - q_j < k_j \leq L$ and $gcd(s_j, 2^{k_j} - 1)$. Since a large number of generators had good properties, L'Ecuyer [1999b] performed extensive computer searches and introduced specific instances of such generators. For $L = 32$, the generators have period lengths $(2^{31} - 1)(2^{29} - 1)(2^{28} - 1)(2^{25} - 1) \approx 2^{113}$ and characteristic polynomials of degree 113. The procedure LFSR113 has a period $\rho \approx 2^{113}$ and the procedure LFSR258 has a period $\rho \approx 2^{258}$.

Using the powers-of-2 decomposition method, L'Ecuyer et al. [2000b] proposed CMRGs that are faster for an equivalent statistical quality. Considering $a = \pm 2^q \pm 2^r$, the product of x by each power of 2 can be implemented by a left shift of the binary representation of x, and the product ax is computed by adding /or subtracting. Combining MRG with $J = 2$ components of order $k = 3$ with parameters defined such that each component has only two nonzero coefficients, one of the form $a_{ij} = 2^q$ and the other one of the form $a_{ij} = 2^q + 1$, they obtained the MRG31k3p. It has two distinct cycles of length $\rho = \frac{m_1 m_2}{2} \approx 2^{185}$ and provided very good results in terms of speed compared to the other MRGs having similar periods.

C.2.5 Matrix notation

More generally, we can rewrite the recurrence in Equation (C.2.2) in matrix form. We let A be a $k \times k$ transition matrix with elements in $\mathcal{F}_m$ and B be a $w \times k$ output matrix. Given the characteristic polynomial of the matrix A

$$P(z) = det(A - zI) = z^k - a_1 z^{k-1} - ... - a_k$$

where I is the identity matrix, we consider the recurrence

$$\begin{aligned} X_n &= AX_{n-1} \\ Y_n &= BX_n \\ u_n &= \sum_{l=1}^{w} y_{i,l-1} m^{-l} = y_{i,0} y_{i,1} y_{i,2} .... \end{aligned}$$

where each X_n is a k-dimensional column vector of elements of $\mathcal{F}_m$. One can show that $\{X_n\}$ follows the recurrence

$$X_n = a_1 X_{n-1} + ... + a_k X_{n-k}$$

Thus, we have k copies of the same linear recurring sequence evolving in parallel, with possibly different lags between themselves. This is the parallel MRG implementation of the matrix generator, and the state is redefined as $s_n = (X_n, ..., X_{n+k-1})$. Setting $X_n = (x_{n,1}, ..., x_{n,k})^\top$ to be the k-bit state vector at step n and $Y_n = (y_{n,1}, ..., y_{n,w})^\top$ to be the w-bit output vector at step n, all operations are performed in $\mathcal{F}_m$ and we have

$$\Psi_t = \{(u_0, u_1, ..., u_{t-1}) : s_0 \in \mathcal{F}_m^k\}$$

By carefully choosing the matrices A and B we can recover well known generators. In the special case where B is the identity matrix, we define $y_n = x_{n,1}$, for $j > 1$ we let d_j be the lag associated with the component j of X_n. That is, $x_{n,j} = x_{n+d,1} = y_{n+d_j}$ for $2 \leq j \leq k$ and all $n \geq 0$. A special case is when A is the companion matrix of $P(z)$

$$A_c = \begin{bmatrix} 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \\ a_k & a_{k-1} & ... & a_1 \end{bmatrix}$$

then the recurrences associated with the successive components of X_n are shifted one unit apart, $d_j = j - 1$ for all j. In general, if $P(z)$ is the characteristic polynomial of A, we can write $A = PA_cP^{-1}$ for some regular matrix P. Then, $X_n = A^n X_0 = PA_c^n P^{-1} X_0$ for all n. If the output of the matrix generator is produced by a composition of the form $G = G_1 \circ G_2$ as above, then the matrix generator is no more general than the MRG. That is, if we define $x_n = G_1(X_n)$, then the sequence $\{x_n\}$ obeys again the linear recurence in Equation (C.2.2). There are other ways of combining the elements of X_n leading to different recurrence. For instance,

$$X_n = X_{n-r} \oplus X_{n-k} \tag{C.2.4}$$

where $\oplus$ is the bitwise exclusive-or, provides a very fast way of implementing the GFSR. The lagged-Fibonacci generator is a modified GFSR where $\oplus$ can be replaced by any arithmetic or logical operation, such as $+$, $-$, etc. In the case of the additive generator, we get

$$X_n = \left(a_r X_{n-r} + a_k X_{n-k}\right) \mod m$$

where we can set $m = 2^L$. The add-with-carry (AWC) and subtract-with-borrow (SWB) were proposed by Marsaglia et al. [1991]. However, all these generators have gross structural defects. Better modifications of Equation (C.2.4) maintaining the speed and increasing the period length from $2^k - 1$ to $2^{kL} - 1$ are the twisted GFSR proposed by

Matsumoto et al. [1994] and the Mersenne twister introduced by Matsumoto et al. [1998]. These generators provide a very efficient implementation of linear generators over $\mathcal{F}_2$ with polynomial $P(z)$ of very large degree k, but their values of N_1 are much smaller than $\frac{k}{2}$. In addition, they have large value of Δ_1 since their equidistribution is far from optimal in large dimensions. As an example, the MT19937 has $k = 19937$, $N_1 = 135$, and $\Delta_1 = 6750$. Niederreiter [1995] introduced the multiple recursive matrix method as a generalisation framework encompassing many of these modifications and variants. For $m = 2$, the sequences $\{x_{i,j}, i \geq 0\}$ and $\{y_{i,j}, i \geq 0\}$ both obey the linear recurrence

$$x_{i,j} = \left(a_1 x_{i-1,j} + ... + a_k x_{i-k,j}\right) \mod 2$$

Setting $a_k = 1$, the recurrence has order k, is purely periodic, and has the upper bound $2^k - 1$ if $P(z)$ is a primitive polynomial over $\mathcal{F}_2$. One can jump directly from X_n to $X_{n+\nu}$ for an arbitrary ν by precomputing $A^\nu \mod 2$ and then multiply X_n by this matrix. It allows to split a sequence into streams and substreams. In order to obtain robust generators, the matrices A and B must perform a large enough amount of bit transformations, which can be obtained by performing operations such as bit shifts, xors, and bit masks spread out in the matrix. Doing so, Panneton et al. [2006] proposed generators with better equidistribution and bit-mixing properties than other generators for equivalent period length and speed. The Well Equidistributed Long-period Linear (WELL) are $\mathcal{F}_2$-linear generators with primitive characteristic polynomials, speed and period length comparable to the Mersenne twister, and such that $N_1 \approx \frac{k}{2}$ and $\Delta_1 = 0$.

C.2.6 Initialisation

Initialising PRNGs is very sensitive to the choice of the seed and few information is provided on the best seed to use. One can not simply use the system time or process identity. This is particularly true for TFSR, MT, WELL etc. For instance, the MT uses 624 32-bit integers to represent its internal state plus a few more for housekeeping (19937 bits of internal state). Thus, there is no bias-free way to use a single full-range 32-bit value to initialise this type of PRNGs. Further, one problem when initialising MT, or WELLs, is that if two initial states (x_i and x_j) are too near with respect to the Hamming distance, then the corresponding output sequences are close to each other. For example, Jeff Szuhay (Psychology Software Tools) described the problem he faced trying to generate repeatable, random sequences within a trial block regardless of when it occurs in an experiment. He used a tt800 TFSR represented internally as an array of 25 unsigned integer keys (32 bits each). The initialisation routine applied a simple XOR operator on the seed and each of the 25 keys, producing a starting key state that was not significantly different from the default key state. An initialisation routine was needed to generate a sequence of very wide numbers from a given seed. He chose Marsaglia's Super Duper LCG. Further, Martin Kretschmar initialised the state vector with many zeroes, or some bit-pattern, and observed a tendency for non-randomness to remain for a long time. Thus, when initialising MT, or WELL, we want the most significant bit of the seed to be well reflected to the state vector. One way forward is to use the CMRGs, such as the MRG32ka or the MRG63k3a discussed in Appendix (C.2.4.3). T. Nishimura and M. Matsumoto proposed in 2002 an improvement to the initialisation of the Mersenne Twister (MT). Since the state needed for a MT implementation is an array of n values of w bits each, a w-bit seed value is used to supply x_0 through x_{n-1} by setting x_0 to the seed value and thereafter setting

$$x_i = f \times \left(x_{i-1} \oplus (x_{i-1} >> (w-2))\right) + i$$

for i from 1 to $n-1$. The constant f forms another parameter to the generator, which is 1812433253 for MT19937 and 6364136223846793005 for MT19937-64. See Knuth [1997] for details on the multiplier f.

C.3 The Sobol' sequence

Given that point sets and sequences with low discrepancy is a fruitful approach for numerical integration, one can use the notion of a (t, m, d)-net and a (t, d)-sequence to contruct and describe such points. A (t, m, d)-net is a finite set of points in the hypercube $[0, 1)^d$ possessing a degree of uniformity quantified by t, and a (t, d)-sequence is a sequence of

points whith certain segments forming a (t, m, d)-net. For integers $0 \leq t \leq m$, a (t, m, d)-net in base b is a set of b^m points in $[0, 1)^d$ with the property that exactly b^t points fall in each b-ary box of volume b^{t-m}. One can refer to () for proper definitions. Note, smaller values of t are associated to with greater uniformity. In addition, other things being equal, a smaller base b is better because the uniformity properties of (t, m, d)-nets and (t, d)-sequences are exhibited in sets of b^m points. With larger b, more points are required for these properties to hold. The simplest constructions of low-discrepancy sequences, producing Halton sequences and Hammersley points yield neither (t, d)-sequences nor (t, m, d)-nets (see Halton [1960]). Faure sequences are $(0, d)$-sequences optimising the uniformity parameter t, but requiering a base at least as large as the smallest prime greater than or equal to the dimension of d. Sobol' sequences use base 2 regardless of the dimension, but their t parameter grows with the dimension d (see Sobol' [1967]). Working in base 2 has computational advantages through bit-level operations.

C.3.1 Some theory

The Sobol' generator produces a sequence of $\widehat{N}$ random numbers, usually a power of 2 ranging from 2^{10} to 2^{20}, that is, $\widehat{N} = 2^N$ per dimension. When pricing an derivative options with a Monte Carlo Engine, we need to generate a fixed number of random values called the dimension D, corresponding to the number of time steps multiplied by the number of underlyings. Altogether, to converge to the true price we need many simulations for each dimension, giving the total number of random number generated. Hence, we have to generate D sequences of $\widehat{N}$ random numbers, that is, a total of $\widehat{N} \times D$ points or coordinates of the form

$$x_{n,\widehat{d}} = \frac{m_{n,\widehat{d}}}{2^N}$$

where N is the number of bits (e.g. $N = 32$ bits), n is the point index, $\widehat{d}$ is the dimension index and D can range from 1 to several thousand. The points $x_{n,\widehat{d}}$ are in the range $[0, 1)$ and are stored in a matrix where $n \in [0, \widehat{N} - 1]$ and $\widehat{d} \in [0, D - 1]$. The various coordinates of a d-dimensional Sobol' sequence result from permutations of segments of the Van der Corput sequence, and these permutations result from multiplying expansions of consecutive integers by a set of generator matrices, one for each dimension. All coordinates of a Sobol' sequence follow the same construction, but each with its own generator.

C.3.1.1 Generating a Sobol' sequence

For simplicity, we are going to assume that $D = 1$ for the moment, and we are going to show how to generate a Sobol' sequence of $\widehat{N}$ points. Given N-bit, an integer n can be represented on mode two as

$$n = (b_N b_{N-1} ... b_2 b_1)_2$$

in binary (see Table (C.1)). To generate one sequence of N-bit low-discrepency Sobol' numbers, we choose odd integers M_i for $0 \leq i < N$ and define N direction numbers v_i

$$v_i = \frac{M_i}{2^i} = 0.v_{i1}v_{i2}.. \tag{C.3.5}$$

where v_{ij} denote the binary expansion of v_i. Then, we choose a primitive polynomial $P(X)$ of degree d with coefficients a_i from the two-element finite field $GF(2)$

$$P(x) = x^d + a_1 x^{d-1} + ... + a_{d-1}x + 1$$

where a_i for $i \in [1, d-1]$ are either 0 or 1. The number of coefficient is $coeff = (d-1)$. We identify the coefficients of a primitive polynomial of degree d with the integer I_d

$$I_d = (a_1 a_2 ... a_{d-1})_2$$

Hence, each primitive polynomial is uniquely specified by its degree d together with the number I_d. For example, given $d = 7$ and $I_d = 28 = (011100)_2$ we obtain the polynomial

$$P(x) = x^7 + x^5 + x^4 + x^3 + 1$$

Binary addition of integers modulo two, which amounts to bitewise addition without carry, is a very fast operation on modern computers known as Exclusive Or (XOR) (see Marsaglia [2003]). As a result, we can define a sequence of positive integers $(M_1, M_2, ..)$ by the recurrence relation

$$M_k = 2a_1 M_{k-1} \oplus ... \oplus 2^{d-1} a_{d-1} M_{k-d+1} \oplus 2^d M_{k-d} \oplus M_{k-d} \tag{C.3.6}$$

where $\oplus$ is an Exclusive-Or (XOR). The initial values M_k for $k \in [1, d]$ can be chosen freely provided that each of them is odd and less than 2^k. These coefficients a_i are used to calculate each direction vector v_i as

$$v_i = a_1 v_{i-1} \oplus ... \oplus a_{d-1} v_{i-d+1} \oplus v_{i-d} \oplus [v_{i-d} >> d] \tag{C.3.7}$$

where the last term is v_{i-d} right-shifted by d bits. Note, another way of writing that term is the following

$$v_{i-d} >> d \to \frac{v_{i-d}}{2^d}$$

so that Equation (C.3.5) rewrite $v_i = M_i >> i$. But the direction number implemented is

$$v_i = \frac{M_i}{2^i} 2^N = M_i 2^{N-i}$$

which gives $v_i = M_i << (N - i)$.

***Remark* C.3.1** *The direction vector v_i is implemented with Equation (C.3.7) for all $i > d$. For $i \in [1, d]$ it is implemented with Equation (C.3.5).*

A one-dimensional N-bit wide low-discrepancy Sobol' sequence $x_1, x_2, ..$ can be generated based on this set of direction vectors. Take the n-th term of this sequence x_n with $n = (b_N b_{N-1} ... b_2 b_1)_2$ in binary, then

$$x_n = b_1 v_1 \oplus ... \oplus b_N v_N$$

requiring at most N lookups and $(N - 1)$ XORs. Considering the gray-coded representation of Antonov et al. [1979] given in Appendix (C.3.5.2), this effort can be drastically reduced. Hence, we can generate the Sobol points using

$$x_n = g_{n,1} v_1 \oplus g_{n,2} v_2 \oplus ... \oplus g_{n,N} v_N$$

where $g_{n,k}$ is the kth digit from the right of the Gray code of n in binary, that is $gray(n) = (..g_{n,3} g_{n,2} g_{n,1})_2$ (see Table (C.3)). Similarly, since $gray(n)$ and $gray(n - 1)$ differ in one known position, we can generate the point x_n recursively based on x_{n-1} with only one lookup and one XOR as

$$x_n = x_{n-1} \oplus v_{c_{n-1}} \text{ with } x_0 = 0$$

where c_n is the index of the first 0 digit from the right in the binary representation $n = (b_3 b_2 b_1)_2$. With the Gray code implementation, we simply obtain the points in a different order, while still preserving their uniformity properties.

C.3.1.2 Generating sequences of random numbers

In general $D > 1$ and one sequence per dimension needs to be generated. For instance, in finance D is the product of the number of timesteps and the dimension of the SDE. For each dimension $\hat{d}$, there is a set of, for example, $N = 32$-bit unsigned integer direction vectors $v_{i,\hat{d}}$ such that

$$m_{n,\hat{d}} = g_1 v_{1,\hat{d}} \oplus g_2 v_{2,\hat{d}} \oplus ... \oplus g_N v_{N,\hat{d}}$$

where g_k are the bits of the Gray code g with g_1 being the least significant bit. If the total number of random values is $\widehat{N} = 2^N$, then only the first N vectors are used for $n < \hat{N}$. For instance, given $N = 32$, we can handle $\widehat{N} = 2^{32}$ random numbers which is more than enough.

C.3.1.3 Initialisation

To construct a Sobol' sequence, a set of initial direction numbers v_i must be selected. But there is some freedom in that selection. A bad selection of initial numbers can considerably reduce the efficiency of Sobol' sequences when used for computation. Good initialisation numbers for different dimensions are provided by several authors. The key reference is the paper by Bratley and Fox [1988]. Recently, Joe and Kuo [2008] expanded the maximum dimension from 64 to 1111.

C.3.1.4 Randomization

By definition, a randomized sequence no-longer has the homogeneity properties of the original sequence, that is, the means and the skews are not exactly 0 while the variances and the kurtosis are not equal at all dimensions. Glasserman [2004] provides two good reasons for randomizing QMC, one, it offers the possibility of measuring error through a confidence interval while preserving much of the accuracy of pure QMC, two, there are settings in which randomization improves accuracy. For instance, Owen [1997] showed that for smooth integrand, the root mean square error of integration using a class of randomized nets is $\mathcal{O}(\frac{1}{n^{1.5-\epsilon}})$ as opposed to $\mathcal{O}(\frac{1}{n^{1-\epsilon}})$ without it. Recently, another reason was proposed by Gurrieri [2012c] who showed that a simple randomized version of Sobol sequence could remove most of the auto-correlation bias of the original sequence while retaining most of its good convergence properties. Several methods exist to randomize sequence such as Random Shift, Random Permutation of Digits, Scrambled Nets or Linear Permutation of Digits and we refer the readers to L'Ecuyer et al. [2000] for a more extensive treatment of the topic.

C.3.2 Examples: direction numbers

Using a primitive polynomial we derive a sequence of positive integers which are used to derive the direction numbers. We first consider the case

- $d = 1$ with 0 coefficient

For the integer $I_1 = 0 = (000)_2$ the primitive polynomial becomes

$$P(x) = x + 1$$

Since $d = 1$, only M_1 can be chosen freely and we assume $M_1 = 1$. Using Equation (C.3.6) we get

$$M_2 = 2M_1 \oplus M_1 = 2 \oplus 1 = (010)_2 \oplus (001)_2 = (011)_2 = 3$$

then M_3 is

$$M_3 = 2M_2 \oplus M_2 = 6 \oplus 3 = (110)_2 \oplus (011)_2 = (101)_2 = 5$$

and M_4 is

$$M_4 = 2M_3 \oplus M_3 = 10 \oplus 5 = (1010)_2 \oplus (101)_2 = (1111)_2 = 15$$

The elements of the direction vector are $v_1 = \frac{1}{2} = (0.1)_2$ or equivalently, setting $v_0 = 0$ and using Equation (C.3.7), we get $v_1 = v_0 \oplus [v_0 >> 1] = (0.1)_2$. Then $v_2 = \frac{3}{4} = (0.11)_2$ or equivalently $v_2 = v_1 \oplus [v_1 >> 1] = (0.1) \oplus 0.01 = (0.11)_2$. We then consider the case

- $d = 2$ with $d - 1 = 1$ coefficient

For the integer $I_2 = 1 = (001)_2$ we get $a_1 = 1$ and the polynomial becomes

$$P(x) = x^2 + x + 1$$

For $d = 2$, we assume $M_1 = 1$ and $M_2 = 1$. Using Equation (C.3.6) we get

$$M_3 = 2M_2 \oplus 4M_1 \oplus M_1 = 2 \oplus 4 \oplus 1 = (010)_2 \oplus (100)_2 \oplus (001)_2 = (111)_2 = 7$$

and M_4 is

$$M_4 = 2M_3 \oplus 4M_2 \oplus M_2 = 14 \oplus 4 \oplus 1 = (1110)_2 \oplus (0100)_2 \oplus (0001)_2 = (1011)_2 = 11$$

The direction numbers are $v_1 = \frac{1}{2} = (0.1)_2$ or equivalently, setting $v_0 = 0$ and using Equation (C.3.7), we get $v_1 = v_0 \oplus [v_0 >> 1] = (0.1)_2$. Then $v_2 = \frac{1}{4} = (0.01)_2$ or equivalently $v_2 = v_1 \oplus v_0 \oplus [v_0 >> 2] = (0.1)_2 \oplus (000)_2 \oplus (0.01)_2 = (0.11)_2$. We consider the case

- $d = 3$ with $d - 1 = 2$ coefficient

For the integer $I_3 = 1 = (001)_2$ we get $a_1 = 0$, $a_2 = 1$ and the polynomial becomes

$$P(x) = x^3 + x + 1$$

For $d = 3$ and $I_3 = 1$, we assume $M_1 = 1$, $M_2 = 3$ and $M_3 = 7$. Using Equation (C.3.6) we get

$$M_4 = 4M_2 \oplus 8M_1 \oplus M_1 = 12 \oplus 8 \oplus 1 = (1100)_2 \oplus (1000)_2 \oplus (0001)_2 = (0101)_2 = 5$$

and M_5 is

$$M_5 = 4M_3 \oplus 8M_2 \oplus M_2 = 28 \oplus 24 \oplus 3 = (11100)_2 \oplus (11000)_2 \oplus (00011)_2 = (0111)_2 = 7$$

The direction numbers are $v_1 = \frac{1}{2} = (0.1)_2$ or equivalently, setting $v_0 = 0$ and using Equation (C.3.7), we get $v_1 = v_0 \oplus [v_0 >> 1] = (0.1)_2$. Then $v_2 = \frac{3}{4} = (0.11)_2$ or equivalently $v_2 = v_1 \oplus v_0 \oplus [v_0 >> 3] = (0.1)_2 \oplus (000)_2 \oplus (0.0001)_2 = (0.11)_2$. Then $v_3 = \frac{7}{8} = (0.111)_2$. For the integer $I_3 = 2 = (010)_2$ we get $a_1 = 1$, $a_2 = 0$ and the polynomial becomes

$$P(x) = x^3 + x^2 + 1$$

We consider the case

- $d = 4$ with $d - 1 = 3$ coefficient

For the integer $I_4 = 1 = (001)_2$ we get $a_1 = 0$, $a_2 = 0$, $a_3 = 1$ and the polynomial becomes

$$P(x) = x^4 + x + 1$$

For the integer $I_4 = 4 = (100)_2$ we get $a_1 = 1$, $a_2 = 0$, $a_3 = 0$ and the polynomial becomes

$$P(x) = x^4 + x^3 + 1$$

We consider the case

- $d = 5$ with $d - 1 = 4$ coefficients

For the integer $I_5 = 2 = (0010)_2$ we get $a_1 = 0$, $a_2 = 0$, $a_3 = 1$, $a_4 = 0$ and the polynomial becomes

$$P(x) = x^5 + x^2 + 1$$

For the integer $I_4 = 4 = (0100)_2$ we get $a_1 = 0$, $a_2 = 1$, $a_3 = 0$, $a_4 = 0$ and the polynomial becomes

$$P(x) = x^5 + x^3 + 1$$

C.3.3 Some rules and considerations

We saw in Section (C.3.1) that the generation of Sobol' numbers is initially carried out on a set of integers in the interval from 1 to a power of two minus one, $[1, 2^N - 1]$, where N represents the number of bits in an unsigned integer on a given computer (typically 32 bits). For pseudo-random number generators, by the central limit theorem, the number of iterations only affect the expected variance of the result. However, for low-discrepancy numbers the situation is different. For number generators based on integer arithmetics modulo two, by construction, they provide additional equidistribution properties whenever the number of iteration is $\widehat{N} = 2^N - 1$ for some positive integer N. For instance, on a unit interval in one dimension, such choice of draws always results in a perfectly regular distribution of points. In fact, careful construction from number theoretical principles already tries to match all the moments in a well-balanced way and interfering with them can have unexpected effects. Hence, using $\widehat{N} = 2^N - 1$ vector draws ensure that the first moment is exactly met when using Sobol' numbers. Moreover, the second moments are almost exactly met especially when compared to pseudo-random numbers. Consequently, when simulating a Sobol' sequence the test to be passed or failed is that the mean of the standard normal random variables should be 0.

In addition, unlike pseudo-random numbers, low-discrepancy numbers have the antithetic feature build into them, but only approximately. Hence, whenever we use a recommended number of draws such as $\widehat{N} = 2^N - 1$ for some N, the first moment of a Gaussian variate vector is correct within the numerical accuracy of the conversion from uniform $(0, 1)$ to Gaussian variates. As low-discrepancy numbers are very carefully designed, adding the antithetic method is unlikely to improve the accuracy and can lead to erroneous results. Further, one should include an option to randomize the Sobol' sequence. This is a trick to avoid a deficiency in Sobol sequence for large dimensional-paths, when we do not use the Brownian bridge. For high-dimensional problems, every set of initialisation numbers will work as well as any other, but one should use initialisation numbers providing properties A and $A^{'}$ for the lowest dimensions and for the higher dimensions at least ensure that any regularity in the initialisation set is broken up. For instance, one can use a pseudo-random number generator to draw uniform variates $\in (0, 1)$ from and initialised in a special way.

It is well known that if there is an advantage in ordering the dimensions according to their importance, as there is for low-discrepancy numbers, the Brownian bridge method offers the benefit of almost optimal ordering (in the sense of maximal variability explained) while only requiering three multiplication per dimension. Alternatively, another empirical trick is to skip the first N draws. For instance, with $N = 16$ skipping about 65K draws made the performance just as good as the Brownian bridge. As a result, it is better to skip 2^N draws due to the structure of Sobol' numbers. One can use the fast skipping function as it is faster and safer to skip 2^N numbers in Sobol'. Safer because the Sobol' sequence is filling the interval $[0, 1[$ in a specific order and one should be careful when playing with the sequence, to avoid disrupting the filling. As the Sobol' filling is uniform by layers of 2^N, to be on the safe side we should skip the whole layers, otherwise we might end up with bad fillings. It is also faster, but only if you use the special shortcut when skipping 2^N, with a special function.

C.3.4 Improving the Sobol' sequence in high dimensions

It is well known that Sobol' sequence suffers from a well known loss of efficiency at high dimensions. Even though the Brownian bridge construction is used to circumvent this drawback, the mechanisms by which it is achieved are not yet

fully understood. Nonetheless, the combination of Sobol' Sequence with the Brownian bridge has been empirically demonstrated by many authors. One of the main reason put forward is that the high dimensions coordinates have worse uniformity than those at low dimensions. By scrambling these coordinates, the Brownian bridge would mix coordinates of high dimensions with low ones (the good and the bad points), resulting in less apparent errors. Gurrieri [2012c] illustrated the loss of convergence efficiency occuring in Sobol' sequence at high dimensions, and highlited empirically the improvement obtainened with the use of a Brownian bridge. Proceeding by elimination, he isolated the causes of the phenomenon showing the problem was not related to a lower quality of the coordinates at high dimensions, and he then performed statistical analysis of Sobol' sequence at high dimensions. He observed an auto-correlation bias appearing in the incremental construction but not in the Brownian bridge construction, concluding that this bias was actually causing the loss of efficiency at high dimensions and that the Brownian bridge reduces the number of Gaussian deviates required to reach the events of a product that matter most. More precisely, the auto-correlations are small but biased on the negative side. The individual Gaussian deviates being not exactly independent, the total variance involves their correlations. As the correlations have a bias on the negative side, he showed that the extra error in the incremental side was

$$C = \frac{2}{m} \sum_{i<j} \rho_{ij} \, , T = mh$$

which is negative for Sobol sequence. Therefore, having less Sobol' Gaussian deviates to sum, we reduce the biased auto-correlation and improve convergence. Further, Gurrieri showed that a simple randomised version of Sobol' sequence could remove most of the bias of the original sequence while retaining most of its good convergence properties. That is, even though the randomised sequence no-longer holds the homogeneity properties of the original sequence, he observed that the bias on the low variances had disapeared leading to a sequence well distributed around 1 and that the bias related to the auto-correlation was largely removed.

C.3.5 A few points on XOR and Gray code

C.3.5.1 The tables

Decimal	Bit
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111
16	10000
17	10001
18	10010
19	10011
20	10100
21	10101
22	10110
23	10111
24	11000
25	11001
26	11010
27	11011
28	11100

Table C.1: From decimals to bits

P+Q	P	Q
0	0	0
1	0	1
1	1	0
0	1	1

Table C.2: XORing bit

Decimal	Bit	Gray
0	$(000)_2$	$(000)_2 = 0$
1	$(001)_2$	$(001)_2 = 1$
2	$(010)_2$	$(011)_2 = 3$
3	$(011)_2$	$(010)_2 = 2$
4	$(100)_2$	$(110)_2 = 6$
5	$(101)_2$	$(111)_2 = 7$
6	$(110)_2$	$(101)_2 = 5$
7	$(111)_2$	$(100)_2 = 4$
8	$(1000)_2$	$(1100)_2 = 12$
9	$(1001)_2$	$(1101)_2 = 13$
10	$(1010)_2$	$(1111)_2 = 15$
11	$(1011)_2$	$(1110)_2 = 14$
12	$(1100)_2$	$(1010)_2 = 10$
13	$(1101)_2$	$(1011)_2 = 11$
14	$(1110)_2$	$(1001)_2 = 9$
15	$(1111)_2$	$(1000)_2 = 8$

Table C.3: From decimals to Gray code

C.3.5.2 XOR and Gray code

As shown in Table (C.2), XOR is only true when either of the two bits is true, else it is false. Note, if $v_1 = 0.1$ then $v_1 >> 3 = 0.0001$. That is, v_1 is right-shifted by 3 bits. The (binary-reflected) Gray code of an integer i is defined as

$$gray(i) = i \oplus [\frac{i}{2}] = (..b_3b_2b_1)_2 \oplus (..b_4b_3b_2)$$

where the last bit is removed in the second binary representation. For example, given the decimal

- $i = 1 = (001)_2$

then we get the Gray code

$$gray(1) = (001)_2 \oplus (000)_2 = (001)_2 = 1$$

Similarly, given the decimal

- $i = 2 = (010)_2$

then we get

$$gray(2) = (010)_2 \oplus (001)_2 = (011)_2 = 3$$

Again, given the decimal

- $i = 3 = (011)_2$

then we get

$$gray(3) = (011)_2 \oplus (001)_2 = (010)_2 = 2$$

If we carry on, we get the Gray code in Table (C.3). We see that the Gray code is simply a reordering of the nonnegative integers within every block of 2^m numbers from $m = 0, 1, ...$ We let c_n be the index of the first 0 digit from the right in the binary representation of $n = (b_3b_2b_1)_2$. Then we get $c_0 = 1$, $c_1 = 2$, $c_2 = 1$, $c_3 = 3$, $c_4 = 1$, $c_5 = 2$ etc..

Appendix D

Stochastic processes and Time Series

D.1 Introducing time series

D.1.1 Definitions

Following Brockwell et al. [1991], a time series is a set of observations x_t, each one being recorded at a specified time t. A discrete time series is one in which the set T_0 of times at which observations are made is a discrete set, as it is the case when observations are made at fixed time intervals. Continuous time series are obtained when observations are recorded continuously over some time interval. We assume that each observation x_t is a realised value of a certain random variable X_t. The time series $\{x_t\ , t \in T_0\}$ is a realisation of the family of random variables $\{X_t\ , t \in T_0\}$. Hence, we can model the data as a realisation of a stochastic process $\{X_t\ , t \in T\}$ where $T \supseteq T_0$.

Definition D.1.1 *A stochastic process is a family of random variables* $\{X_t\ , t \in T\}$ *defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$.

When dealing with a finite number of random variables, we need to compute the covariance matrix to gain insight into the dependence between them. For a time series $\{X_t\ , t \in T\}$ we extend that concept to deal with an infinite collections of random variables which is called the Autocovariance function.

Definition D.1.2 *If* $\{X_t\ , t \in T\}$ *is a process such that* $Var(X_t) < \infty$ *for each* $t \in T$*, then the Autocovariance function* $\gamma_X(.,.)$ *of* X_t *is defined by*

$$\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - EX_r)(X_s - EX_s)]\ , r, s \in T$$

Definition D.1.3 *(Weak Stationarity) The time series* $\{X_t\ , t \in \mathbb{Z}\}$ *with index set* $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$ *is said to be Stationary if*

1. $E|X_t^2| < \infty$ *for all* $t \in \mathbb{Z}$
2. $E[X_t] = m$ *for all* $t \in \mathbb{Z}$
3. $\gamma_X(r,s) = \gamma_X(r+t, s+t)$ *for all* $r, s, t \in \mathbb{Z}$

If $\{X_t\ , t \in \mathbb{Z}\}$ is stationary then $\gamma_X(r,s) = \gamma_X(r-s, 0)$ for all $r, s, \in \mathbb{Z}$. Hence, we can redefine the Autocovariance function of a stationary process as the function of just one variable

$$\gamma_X(h) = \gamma_X(h,0) = Cov(X_{t+h}, X_t) \text{ for all } t, h \in \mathbb{Z} \tag{D.1.1}$$

The Autocorrelation function of X_t is defined analogously as the function whose value at lag h is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_{t+h}, X_t) \text{ for all } t, h \in \mathbb{Z}$$

Definition D.1.4 *(Gaussian Time Series) The process X_t is a Gaussian time series if and only if the distribution functions of X_t are all multivariate normal.*

For example, given an independent and identically distributed (iid) sequence of zero-mean random variables Z_t with finite variance σ_Z^2, we let $X_t = Z_t + \theta Z_{t-1}$. Then the Autocovariance function of X_t is

$$\gamma_X(t+h,t) = Cov(Z_{t+h} + Z_{t+h-1}, Z_t + Z_{t-1}) = \begin{cases} (1+\theta^2)\sigma_Z^2 \text{ if } h = 0 \\ \theta\sigma_Z^2 \text{ if } h = \pm 1 \\ 0 \text{ if } |h| > 1 \end{cases}$$

and X_t is stationary.

D.1.2 Estimation of trend and seasonality

In general, when analysing time series we check if the data is a realisation of the process

$$X_t = m_t + s_t + Y_t \tag{D.1.2}$$

where m_t is a slowly changing function called the trend, s_t is a function with known period d called the seasonal component, and Y_t is a random noise component which is stationary. If the seasonal and noise fluctuations appear to increase with the level of the process, then a preliminary transformation of the data is often used to make the transformed data compatible with the model in Equation (D.1.2).

One can obtain smoothing by means of a Moving Average. Assuming no seasonality term s_t in Equation (D.1.2) and discrete time $t = 1, .., n$, we let q be a non-negative integer and consider the two-sided moving average

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}$$

of the process $\{X_t\}$ in Equation (D.1.2). Then for $q + 1 \leq t \leq n - q$ we get

$$\begin{aligned} W_t &= \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \\ &\approx m_t \end{aligned}$$

assuming that m_t is approximately linear over the interval $[t-q, t+q]$ and that the average of the error terms over this interval is close to zero. The moving average provides us with the estimates

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j} \text{ for } q+1 \leq t \leq n-q$$

Note, as X_t is not observed for $t \leq 0$ or $t > n$, we can not use this equation for $t \leq q$ or $t > n - q$ and one can define $X_t = X_1$ for $t < 1$ and $X_t = X_n$ for $t > n$. It is useful to think of $\{\hat{m}_t\}$ in the above equation as a process obtained from $\{X_t\}$ by application of a linear operator or linear filter

$$\hat{m}_t = \sum_{j=-\infty}^{\infty} a_j X_{t+j}$$

with weights $a_j = \frac{1}{2q+1}$ for $-q \leq j \leq q$ and $a_j = 0$ for $|j| > q$. This filter is a low-pass filter since it takes the data $\{x_t\}$ and removes from it the rapidly fluctuating (or high frequency) component $\{\hat{Y}_t\}$ to leave the slowly varying estimated trend term $\{\hat{m}_t\}$. For q large enough, provided $\frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t+j} \approx 0$, it will not only attenuate noise, but it will allow linear trend functions $m_t = at + b$ to pass without distortion. But q can not be too large since if m_t is not linear, the filtered process will not be a good estimate of m_t. Clever choice of the weights $\{a_j\}$ will allow for larger class of trend functions.
Suppose the mean level of a series drifts slowly over time, then a naive one-step-ahead forecast is $X_t(1) = X_t$. Letting all past observations play a part in the forecast, but giving greater weights to those that are more recent we choose weights to decrease exponentially

$$X_t(1) = \frac{1-w}{1-w^t}\left(X_t + wX_{t-1} + w^2X_{t-2} + ... + +w^{t-1}X_1\right)$$

where $0 < w < 1$. Defining S_t as the right hand side of the above as $t \to \infty$

$$S_t = (1-w)\sum_{s=0}^{\infty} w^s X_{t-s}$$

S_t can serve as a one-step-ahead forecast $X_t(1)$. For any fixed $a \in [0,1]$, the one sided moving averages $\hat{m}_t$ with $t = 1, .., n$ defined by the recursions

$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1}\ , t = 2, ..., n$$

with $\hat{m}_1 = X_1$ can also be used to smooth data. This equation is referred to as Exponential smoothing since it follows from these recursions that, for $t \geq 2$ then

$$\hat{m}_t = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1}X_1$$

is a weighted moving average of $X_t, X_{t-1}, ..$, with weights decreasing exponentially (apart from the last one). Simple algebra gives

$$\begin{aligned} S_t &= aX_t + (1-a)S_{t-1} \\ X_t(1) &= X_{t-1}(1) + a(X_t - X_{t-1}(1)) \end{aligned}$$

To get things started we might set S_0 equal to the average of the first few data points. We can play around with α choosing it to minimise the mean square forecasting error. In practice, α is in the range $[0.25, 0.5]$.

D.1.3 Some sample statistics

From the observations $\{x_1, x_2, .., x_n\}$ of a stationary time series X_t we wish to estimate the Autocovariance function $\gamma(.)$ of the underlying process X_t in order to gain information on its dependence structure. To do so we use the sample Autocovariance function.

Definition D.1.5 *The sample Autocovariance function of* $\{x_1, x_2, .., x_n\}$ *is defined by*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(x_{j+h} - \overline{x})(x_j - \overline{x})\ , 0 \leq h < n$$

where $\hat{\gamma}(h) = \hat{\gamma}(-h)$, $-n < h \leq 0$, *where* $\overline{x}$ *is the sample mean* $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$.

Notice that in defining $\hat{\gamma}(h)$ we divide by n rather than by $(n-h)$. When n is large relative to h it does not much matter which divisor we use. However, for mathematical simplicity and other reasons there are advantages in dividing by n. The sample Autocorrelation function is defined in terms of the sample Autocovariance function as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, |h| < n$$

The sample Autocovariance and Autocorrelation functions can be computed from any data set $\{x_1, x_2, .., x_n\}$ and are not restricted to realisations of a stationary process. The plot of $\hat{\rho}(h)$ against h is known as the correlogram. For data containing a trend, $|\hat{\rho}(h)|$ will exhibit slow decay as h increases, and for data with a substantial deterministic periodic component, $\hat{\rho}(h)$ will exhibit similar behaviour with the same periodicity. Thus $\hat{\rho}(.)$ can be useful as an indicator of non-stationarity.

D.2 The ARMA model

The simplest kind of time series $\{X_t\}$ is one in which the random variables X_t for $t = 0, \pm 1, \pm 2, ..$ are independently and identically distributed with zero mean and variance σ^2. From a second order point of view, that is ignoring all properties of the joint distribution of $\{X_t\}$ except those which can be deduced from the moments $E[X_t]$ and $E[X_s X_t]$, such processes are identified with the class of all stationary process having mean zero and autocovariance function (see Equation (D.1.1))

$$\gamma(h) = \begin{cases} \sigma^2 \text{ if } h = 0 \\ 0 \text{ if } h \neq 0 \end{cases} \tag{D.2.3}$$

Definition D.2.1 *The process $\{Z_t\}$ is said to be a white noise with mean 0 and variance σ^2, written*

$$\{Z_t\} \sim WN(0, \sigma^2)$$

if and only if $\{Z_t\}$ has zero mean and covariance function in Equation (D.2.3)

If the random variable Z_t are independently and identically distributed with mean 0 and variance σ^2 then we shall write

$$\{Z_t\} \sim IDD(0, \sigma^2)$$

***Remark* D.2.1** *The difference between a white noise and an i.i.d. variable lies in the fact that the white noise is a process.*

Increasing in complexity, we consider the class of time series $\{X_t, t = 0, \pm 1, \pm 2, ..\}$ defined in terms of linear difference equations with constant coefficients called the autoregressive moving average or ARMA processes. For any autocovariance function $\gamma(.)$ such that $\lim_{h\to\infty} \gamma(h) = 0$, and for any integer $k > 0$, it is possible to find an ARMA process with autocovariance function $\gamma_X(.)$ such that $\gamma_X(h) = \gamma(h)$ for $h = 0, 1, .., k$. The linear structure of ARMA processes leads to a simple theory of linear prediction.

Definition D.2.2 *(The $ARMA(p,q)$ process)*
The process $\{X_t, t = 0, \pm 1, \pm 2, ..\}$ is said to be an $ARMA(p,q)$ process if $\{X_t\}$ is stationary and if for every t

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q} \tag{D.2.4}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$. We say that $\{X_t\}$ is an $ARMA(p,q)$ process with mean μ if $\{X_t - \mu\}$ is an $ARMA(p,q)$ process.

The Equation (D.2.4) can be written symbolically in the more compact form

$$\phi(B)X_t = \theta(B)Z_t \ , t = 0, \pm 1, \pm 2, .. \tag{D.2.5}$$

where ϕ and θ are the pth and qth degree polynomials

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$$

and

$$\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q$$

and B is the backward shift operator defined by

$$B^j X_t = X_{t-j} \ , t = 0, \pm 1, \pm 2, ..$$

- If $\phi(z) = 1$, then

$$X_t = \theta(B)Z_t$$

and the process is said to be a moving average process of order q (or $MA(q)$). The difference equation have the unique solution $\{X_t\}$ which is a stationary process since for $\theta_0 = 1$ and $\theta_j = 0$ for $j > q$ we have

$$E[X_t] = \sum_{j=0}^{q} \theta_j E[Z_{t-j}] = 0$$

and

$$\gamma_X(t+h,t) = Cov(X_{t+h}, X_t) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} \text{ if } |h| \leq q \\ 0 \text{ if } |h| > q \end{cases}$$

- If $\theta(z) = 1$ then

$$\phi(B)X_t = Z_t$$

and the process is said to be an autoregressive process of order p (or $AR(p)$) and the existence and uniqueness of a stationary solution to the above equation needs closer investigation. For example, in the case $\phi(z) = 1 - \phi_1 z$ we get

$$X_t = Z_t + \phi_1 X_{t-1}$$

and by successive recursion we get

$$X_t = Z_t + \phi_1 Z_{t-1} + ... + \phi_1^k Z_{t-k} + \phi_1^{k+1} X_{t-k-1}$$

If $|\phi_1| < 1$ and $\{X_t\}$ is stationary then $||X_t||^2 = E[X_t^2]$ is constant so that

$$||X_t - \sum_{j=0}^{k} \phi_1^j Z_{t-j}||^2 = \phi_1^{2k+2} ||X_{t-k-1}||^2 \to 0 \text{ as } k \to \infty$$

and since $\sum_{j=0}^{\infty} \phi_1^j Z_{t-j}$ is mean-square convergent (by the Cauchy criterion), we conclude that

$$X_t = \sum_{j=0}^{\infty} \phi_1^j Z_{t-j}$$

which is only valid in the mean-square sense. Further, $\{X_t\}$ is stationary since

$$E[X_t] = \sum_{j=0}^{\infty} \phi_1^j E[Z_{t-j}] = 0$$

and

$$\begin{aligned} Cov(X_{t+h}, X_t) &= \lim_{n\to\infty} E[(\sum_{j=0}^{n} \phi_1^j Z_{t+h-j})(\sum_{k=0}^{n} \phi_1^k Z_{t-k})] \\ &= \sigma^2 \phi_1^{|h|} \sum_{j=0}^{\infty} \phi_1^{2j} = \frac{\sigma^2 \phi_1^{|h|}}{(1-\phi_1^2)} \end{aligned}$$

and it is the unique stationary solution. An easier way to obtain these results is to multiply the $AR(1)$ equation above by X_{t-h} and take the expected value, getting

$$E[X_t X_{t-h}] = E[Z_t X_{t-h}] + \phi_1 E[X_{t-1} X_{t-h}]$$

thus

$$\gamma_h = \phi_1 \gamma_{h-1} \text{ , } h = 1, 2, ..$$

Similarly, squaring the $AR(1)$ equation and taking the expected value, we get

$$E[X_t^2] = E[Z_t^2] + \phi_1^2 E[X_{t-1}^2] + 2E[Z_t \phi_1 X_{t-1}] = \sigma^2 + \phi_1^2 E[X_{t-1}^2]$$

and so $\gamma_0 = \frac{\sigma^2}{(1-\phi_1^2)}$. In the case when $|\phi_1| > 1$ the series does not converge in L^2, but we can rewrite it in the form

$$X_t = -\phi_1^{-1} Z_{t+1} + \phi_1^{-1} X_{t+1}$$

which becomes

$$X_t = -\phi_1^{-1} Z_{t+1} - ... - \phi_1^{-k-1} Z_{t+k+1} + \phi_1^{-k-1} X_{t+k+1}$$

and we get

$$X_t = -\sum_{j=1}^{\infty} \phi_i^{-j} Z_{t+j}$$

which is the unique stationary solution. However, it is regarded as unnatural since X_t is correlated with $\{Z_s, s > t\}$ which is not the case when $|\phi_1| < 1$. If $|\phi_1| = 1$ there is no stationary solution. Restricting attention to $AR(1)$ processes with $|\phi_1| < 1$, such processes are called causal or future-independent autoregressive processes.

Given the causal $AR(p)$ process defined as

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + Z_t$$

the autocorrelation function can be found by multiplying the above equation by X_{t-h} taking the expected value and dividing by γ_0 thus producing the Yule-Walker equations

$$\rho_h = \phi_1 X_{h-1} + ... + \phi_p X_{h-p} + Z_t \text{ , } h = 1, 2, ..$$

are linear recurrence relations, with general solution of the form

$$\rho_h = C_1 w_1^{|h|} + ... + C_p w_p^{|h|}$$

where $w_1, ..., w_p$ are the roots of

$$w_p - \phi_1 w^{p-1} - \phi_2 w^{p-2} - .. - \phi_p = 0$$

and $C_1, .., C_p$ are determined by $\rho_0 = 1$ and the equations for $h = 1, .., p-1$. It is natural to require $\gamma_h \to 0$ as $h \to \infty$, in which case the roots must lie inside the unit circle, that is $|w_i| < 1$ restricting the chosen values $\phi_1, .., \phi_p$.

Definition D.2.3 *An $ARMA(p,q)$ process defined by the equations $\phi(B)X_t = \theta(B)Z_t$ is said to be causal if there exists a sequence of constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \ , \ t = 0, \pm 1, ... \tag{D.2.6}$$

***Proposition* 11** *If $\{X_t\}$ is any sequence of random variables such that $\sup_t E[|X_t|] < \infty$ and if $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ then the series*

$$\psi(B)X_t = \sum_{j=-\infty}^{\infty} \psi_j B^j X_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} \tag{D.2.7}$$

converges absolutely with probability one. If in addition $\sup_t E[|X_t|^2] < \infty$ then the series converges in mean-square to the same limit.

***Proposition* 12** *If $\{X_t\}$ is a stationary process with autocovariance function $\gamma(.)$ and if $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then for each $t \in \mathbb{Z}$ the series in Equation (D.2.7) converges absolutely with probability one and in mean-square to the same limit. If*

$$Y_t = \psi(B)X_t$$

then the process $\{Y_t\}$ is stationary with autocovariance function

$$\gamma_Y(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma(h - j + k)$$

Note, operators such as $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$ with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, when applied to stationary processes inherit the algebriac properties of power series. In particular if $\sum_{j=-\infty}^{\infty} |\alpha_j| < \infty$, $\sum_{j=-\infty}^{\infty} |\beta_j| < \infty$, $\alpha(z) = \sum_{j=-\infty}^{\infty} \alpha_j z^j < \infty$, $\beta(z) = \sum_{j=-\infty}^{\infty} \beta_j z^j < \infty$ and

$$\alpha(z)\beta(z) = \psi(z) \ , \ |z| \leq 1$$

then $\alpha(z)\beta(z)X_t$ is well defined and

$$\alpha(z)\beta(z)X_t = \beta(B)\alpha(B)X_t = \psi(B)X_t$$

Definition D.2.4 *Let $\{X_t\}$ be an $ARMA(p,q)$ process for which the polynomials $\phi(.)$ and $\theta(.)$ have no common zeroes. Then $\{X_t\}$ is causal if and only if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\psi_j\}$ in Equation (D.2.6) are determined by the relation*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)} \ , \ |z| < 1$$

Note, $ARMA$ process for which $\phi(.)$ and $\theta(.)$ have common zeroes are rarely considered. Further, if $\phi(.)$ and $\theta(.)$ have no common zeroes and if $\phi(z) = 0$ for some $z \in \mathbb{C}$ with $|z| = 1$, then there is no stationary solution of $\phi(B)X_t = \theta(B)Z_t$.

Definition D.2.5 *An $ARMA(p,q)$ process defined by the equation $\phi(B)X_t = \theta(B)Z_t$ is said to be invertible if there exists a sequence of constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and*

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \ , \ t = 0, \pm 1, ..$$

Theorem D.2.1 *Let $\{X_t\}$ be an $ARMA(p,q)$ process for which the polynomials $\phi(.)$ and $\theta(.)$ have no common zeroes. Then $\{X_t\}$ is invertible if and only if $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\pi_j\}$ are determined by the relation*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)} \ , \ |z| \leq 1$$

Put simply, if $\{X_t\}$ is a stationary solution of the equations

$$\phi(B)X_t = \theta(B)Z_t \ , \ \{Z_t\} \sim WN(0, \sigma^2)$$

and if $\phi(z)\theta(z) \neq 0$ for $|z| \leq 1$, then the power series coefficients of $C(z) = \frac{\theta(z)}{\phi(z)} = \psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ for $|z| \leq 1$ give an expression for X_t as

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

But also, $Z_t = D(B)X_t$ where $D(z) = \frac{\phi(z)}{\theta(z)} = \sum_{j=0}^{\infty} \pi_j z^j$ for $|z| \leq 1$ as long as the zeros of θ lie strictly outside the unit circle and thus

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

Hence, we will concentrate on causal invertible $ARMA$ processes. The advantage of the representation above is that given $(..., X_{t-1}, X_t)$ we can calculate values for $(..., Z_{t-1}, Z_t)$ and so can forecast X_{t+1}. In general, if we want to forecast X_{t+h} from $(..., X_{t-1}, X_t)$ we use

$$\hat{X}_{t,h} = \sum_{j=0}^{\infty} \psi_{h+j} Z_{t-j}$$

which has the least mean squared error over all linear combinations of $(..., Z_{t-1}, Z_t)$. In fact,

$$E[(\hat{X}_{t,h} - X_{t+h})^2] = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2$$

In practice, there is an alternative recursive approach. Define

$$\hat{X}_{t,h} = \begin{cases} X_{t+h} \text{ if } -(t-1) \leq h \leq 0 \\ \text{optimal predictor of } X_{t+h} \text{ given } X_1, .., X_t \text{ for } 1 \leq h \end{cases}$$

then we have the recursive relation

$$\hat{X}_{t,h} = \sum_{i=1}^{p} \phi_i \hat{X}_{t,h-i} + \hat{Z}_{t+h} + \sum_{j=1}^{q} \theta_j \hat{Z}_{t+h-j}$$

For $h = -(t-1), -(t-2), .., 0$ it gives estimates of $\hat{Z}_t$ for $t = 1, .., n$. For $h > 0$ it gives a forecast $\hat{X}_{t,h}$ for X_{t+h}, and we take $\hat{Z}_t = 0$ for $t > n$. To start the recursion process, we need to know $(X_t, t \leq 0)$ and Z_t for $t \leq 0$. There are two standard approaches

1. Conditional approach: take $X_t = Z_t = 0$ for $t \leq 0$
2. Backcasting: we forecast the series in the reverse direction to determine estimators of $X_0, X_{-1}, ..$ and $Z_0, Z_{-1}, ..$

There are several ways to compute the autocovariance function of an ARMA process. The autocovariance function γ of the causal $ARMA(p,q)$ process $\phi(B)X_t = \theta(B)Z_t$ satisfy

$$\gamma(k) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|k|}$$

where

$$\phi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$$

and we want to determine the coefficients ψ_j. One way is based on the difference equations for $\gamma(k)$ for $k = 0, 1, 2, ..$ which are obtained by multiplying equation (D.2.5) by X_{t-k} and taking expectations

$$\gamma(k) - \phi_1 \gamma(k-1) - ... - \phi_p \gamma(k-p) = \sigma^2 \sum_{k \leq j \leq q} \theta_j \psi_{j-k} \, , 0 \leq k < \max(p, q+1) \tag{D.2.8}$$

and

$$\gamma(k) - \phi_1 \gamma(k-1) - ... - \phi_p \gamma(k-p) = 0 \, , k \geq \max(p, q+1)$$

with general solution

$$\gamma(h) = \sum_{i=1}^{k} \sum_{j=0}^{r_i - 1} \beta_{ij} h^j \xi_i^{-h} \, , h \geq \max(p, q+1) - p$$

where the p constants β_{ij} and the covariance $\gamma(j)$ for $0 \leq j < \max(p, q+1) - p$ are uniquely determined from the boundary conditions above after first computing $\psi_0, \psi_1, .., \psi_q$.

The autocovariance function of an $MA(q)$ process

$$X_t = \sum_{j=0}^{q} \theta_j Z_{t-j} \, , \{Z_t\} \sim WN(0, \sigma^2)$$

has the extremely simple form

$$\gamma(k) = \begin{cases} \sigma^2 \sum_{j=0}^{q} \theta_j \theta_{j+|k|} \text{ if } |k| \leq q \\ 0 \text{ if } |k| > q \end{cases}$$

where θ_0 is defined to be 1 and θ_j for $j > q$ is defined to be zero. The autocovariance function of an $AR(p)$ process

$$\phi(B)X_t = Z_t$$

has an autocovariance function of the form

$$\gamma(h) = \sum_{i=1}^{k} \sum_{j=0}^{r_i - 1} \beta_{ij} h^j \xi_i^{-h} \ , h \geq 0$$

where ξ_i for $i = 1, .., k$ are the zeroes (possibly complex) of $\phi(z)$, and r_i is the multiplicity of ξ_i. The constants β_{ij} are found from Equation (D.2.8). Note, the numerical determination of the autocovariance function $\gamma(.)$ from Equation (D.2.8) can be carried out by first finding $\gamma(0), .., \gamma(p)$ from the equations with $k = 0, 1, .., p$ and then using the subsequent equations to determine $\gamma(p+1), \gamma(p+2), ..$ recursively.

The partial autocorrelation function, like the autocorrelation function, conveys vital information regarding the dependence structure of a stationary process. Both functions depends only on the second order properties of the process. The partial autocorrelation $\alpha(k)$ at lag k may be regarded as the correlation between X_1 and X_{k+1} adjusted for the intervening observations $X_2, .., X_k$.

Definition D.2.6 *The partial autocorrelation function (pacf) $\alpha(.)$ of a stationary time series is defined by*

$$\alpha(1) = Corr(X_2, X_1) = \rho(1)$$

and

$$\alpha(k) = Corr\big(X_{k+1} - P_{\overline{sp}\{1,X_2,..,X_k\}} X_{k+1}, X_1 - P_{\overline{sp}\{1,X_2,..,X_k\}} X_1\big) \ , \ k \geq 2$$

where $P_{\overline{sp}\{1,X_2,..,X_k\}} X_{k+1}$ and $P_{\overline{sp}\{1,X_2,..,X_k\}} X_1$ are projections. The value $\alpha(k)$ is known as the partial autocorrelation at lag k.

Note, the projection satisfies

$$\hat{X}_k = E[X_k | X_1, ..., X_{k-1}] = P_{\overline{sp}\{X_1,..,X_{k-1}\}} X_k \ , \ k \geq 2$$

It is thus the correlation of the two residuals obtained after regressing X_{k+1} and X_1 on the intermediate observations $X_2, .., X_k$.
One can give an equivalent definition of partial autocorrelation function. Let $\{X_t\}$ be a zero-mean stationary process with autocovariance function $\gamma(.)$ such that $\gamma(h) \to 0$ as $h \to \infty$ and suppose that $\phi_{j,k}$ for $j = 1, .., k$ and $k = 1, 2, ..$ are the coefficients in the representation

$$P_{\overline{sp}\{X_1,..,X_k\}} X_{k+1} = \sum_{j=1}^{k} \phi_{j,k} X_{k+1-j}$$

together with the equations

$$< X_{k+1} - P_{\overline{sp}\{X_1,..,X_k\}} X_{k+1}, X_j >= 0 \ , j = k, .., 1$$

Identifying an $AR(p)$ process
Since the $AR(p)$ process has $\rho(h)$ decaying exponentially, it can be difficult to recognise in the correlogram. Suppose we have a process X_t which we believe is $AR(k)$ with

$$X_t = \sum_{j=1}^{k} \phi_{j,k} X_{t-j} + Z_t$$

with Z_t independent from $X_1, .., X_{t-1}$. Given the data $X_1, .., X_n$, the least squares estimates of $(\phi_{1,k}, .., \phi_{1,k})$ are obtained by minimising

$$\frac{1}{n}\sum_{t=k+1}^{n}\left(X_t - \sum_{j=1}^{k}\phi_{j,k}X_{t-j}\right)^2$$

which is approximately equivalent to solving equations similar to the Yule-Walker equations

$$\hat{\gamma}_j = \sum_{l=1}^{k}\hat{\phi}_{l,k}\hat{\gamma}_{|j-l|}\ , j = 1, .., k$$

It can be solved by the Levinson-Durbin recursion:

1. $\sigma_0^2 = \hat{\gamma}_0, \hat{\phi}_{1,1} = \frac{\hat{\gamma}_1}{\hat{\gamma}_0}, k = 0$
2. Repeat until $\hat{\phi}_{k,k}$ near 0

$$\begin{aligned}
&k = k + 1\\
&\hat{\phi}_{k,k} = \frac{1}{\sigma_{k-1}^2}\left(\hat{\gamma}_k - \sum_{j=1}^{k-1}\hat{\phi}_{j,k-1}\hat{\gamma}_{k-j}\right)\\
&\hat{\phi}_{j,k} = \hat{\phi}_{j,k-1} - \hat{\phi}_{k,k}\hat{\phi}_{k-j,k-1} \text{ for } j = 1, .., k-1\\
&\sigma_k^2 = \sigma_{k-1}^2(1 - \hat{\phi}_{k,k}^2)
\end{aligned}$$

The statistic $\hat{\phi}_{k,k}$ is called the kth sample partial autocorrelation coefficient (PACF). If the process X_t is genuinely $AR(p)$ then the population PACF $\hat{\phi}_{k,k}$ is exactly zero for all $k > p$. Thus a diagnostic for $AR(p)$ is that the sample PACFs are close to zero for $k > p$.

Both the sample ACF and PACF are approximately normally distributed about their population values, and have standard deviation of about $\frac{1}{\sqrt{n}}$ where n is the length of the series. A rule of thumb is that $\rho(h)$ (and similarly $\phi_{k,k}$) is negligible if $\hat{\rho}(h)$ (similarly $\hat{\phi}_{k,k}$) lies between $\pm\frac{2}{\sqrt{n}}$ (2 is an approximation to 1.96). Care is needed in applying this rule of thumb as it is important to realise that the sample autocorrelations $\hat{\rho}(1), \hat{\rho}(2), ..$ (and sample partial autocorrelations $\hat{\phi}_{1,1}, \hat{\phi}_{1,1}, ..$) are not independently distributed. The probability that any one $\hat{\rho}(h)$ should lie outside $\pm\frac{2}{\sqrt{n}}$ depends on the values of the other $\hat{\rho}(h)$.

If $\{X_t\}$ is a stationary process with autocovariance function $\gamma(.)$, then its autocovariance generating function is defined by

$$G(z) = \sum_{k=-\infty}^{\infty}\gamma(k)z^k$$

provided the series converges for all z in some annulus $r^{-1} < |z| < r$ with $r > 1$. When the generating function is easy to calculate, the autocovariance at lag k may be determined by identifying the coefficient of either z^k or z^{-k}. Clearly $\{X_t\}$ is white noise if and only if the autocovariance generating function $G(z)$ is constant for all z. If

$$X_t = \sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}\ , \{Z_t\} \sim WN(0, \sigma^2) \tag{D.2.9}$$

and there exists $r > 1$ such that

$$\sum_{j=-\infty}^{\infty} |\psi_j| z^j < \infty \, , r^{-1} < |z| < r$$

the generating function $G(.)$ takes a simple form. We get

$$\gamma(k) = Cov(X_{t+k}, X_t) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|k|}$$

so that the generating function becomes

$$\begin{aligned} G(z) &= \sigma^2 \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|k|} z^k \\ &= \sigma^2 \big[\sum_{j=-\infty}^{\infty} \psi_j^2 + \sum_{k=1}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+k} (z^k + z^{-k}) \big] \\ &= \sigma^2 \big(\sum_{j=-\infty}^{\infty} \psi_j z^j \big) \big(\sum_{k=-\infty}^{\infty} \psi_k z^{-k} \big) \end{aligned}$$

Defining

$$\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j \, , |z| < r$$

we can rewrite the generating function as

$$G(z) = \sigma^2 \psi(z) \psi(z^{-1}) \, , r^{-1} < |z| < r$$

For example, given an $ARMA(p,q)$ process $\phi(B)X_t = \theta(B)Z_t$ for which $\phi(z) \neq 0$ when $|z| = 1$ can be written in the form in Equation (D.2.9) with

$$\psi(z) = \frac{\theta(z)}{\phi(z)} \, , r^{-1} < |z| < r$$

for some $r > 1$. Hence, we get

$$G(z) = \sigma^2 \frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})} \, , r^{-1} < |z| < r$$

When determining the appropriate $ARMA(p,q)$ model to represent an observed stationary time series one must consider the choice of p and q, and estimate the remaining parameters like the mean, the coefficients $\{\phi_i, \theta_j : i = 1, .., p; j = 1, .., q\}$ and the white noise variance σ^2 for given values of p and q. Goodness of fit of the model must also be checked and the estimation procedure repeated with different values of p and q. Final selection of the most appropriate model depends on a variety of goodness of fit tests such as the AICC statistic. We now assume that the data has been adjusted by subtraction of the mean, so that the problem becomes that of fitting a zero-mean ARMA model to the adjusted data $x_1, .., x_n$. If the model fitted to the adjusted data is

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + .. + \theta_q Z_{t-q} \, , \{Z_t\} \sim WN(0, \sigma^2)$$

then the corresponding model for the original stationary series $\{Y_t\}$ is found by substituting $Y_j - \overline{y}$ for X_j with $j = t, ..., t-p$ where $\overline{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$ is the sample mean of the original data.

In the case $q = 0$ a good estimate of ϕ can be obtained by the simple device of equating the sample and theoretical autocovariances at lags $0, 1, .., p$ called the Yule-Walker estimator. When $q > 0$ the corresponding procedure is neither simple nor efficient. One can use least squares or maximum likelihood estimators for solving non-linear optimisation problems.

We assume that $\{X_t\}$ is a Gaussian process with mean zero and covariance function $\kappa(i, j) = E[X_i X_j]$. We consider $X_n = (X_1, .., X_n)^\top$ and $\hat{X}_n = (\hat{X}_1, .., \hat{X}_n)^\top$ where $\hat{X}_1 = 0$ and

$$\hat{X}_j = E[X_j | X_1, ..., X_{j-1}] = P_{\overline{sp}\{X_1,..,X_{j-1}\}} X_j \ , j \geq 2$$

We let Γ_n be the covariance matrix $\Gamma_n = E[X_n X_n^\top]$ and assume that it is non-singular. The likelihood of X_n is

$$L(\Gamma_n) = (2\pi)^{-\frac{n}{2}} (det\Gamma_n)^{-\frac{1}{2}} e^{-\frac{1}{2} X_n^\top \Gamma_n^{-1} X_n}$$

Note, the direct calculation of $(det\Gamma_n)$ and Γ_n^{-1} can be avoided by re-expressing them in terms of the one-step predictors $\hat{X}_j$ and their mean-squared errors v_{j-1} for $j = 1, ..., n$ which are computed recursively from the innovations algorithm. We let θ_{ij} for $j = 1, .., i$ and $i = 1, 2, ..$ denote the coefficients obtained when applying the innovations algorithm to the covariance function κ of $\{X_t\}$ with $\theta_{i0} = 1$, $\theta_{ij} = 0$ for $j < 0$ and $i = 0, 1, 2, ...$ We define the $(n \times n)$ lower triangular matrix $C = [\theta_{i,i-j}]_{i,j=0}^{n-1}$ and the $(n \times n)$ diagonal matrix

$$D = diag(v_0, v_1, .., v_{n-1})$$

so that the innovations representation of $\hat{X}_j$ for $j = 1, ..., n$ can be written in the form

$$\hat{X}_n = (C - I)(X_n - \hat{X}_n)$$

where I is the $(n \times n)$ identity matrix. Hence, we get

$$X_n = X_n - \hat{X}_n + \hat{X}_n = C(X_n - \hat{X}_n)$$

Since D is the covariance matrix of $(X_n - \hat{X}_n)$ we get

$$\Gamma_n = CDC^\top$$

from which the Cholesky factorisation $\Gamma_n = UU^\top$ with U lower triangular can be deduced. Hence, combining the above terms we get

$$X_n^\top \Gamma_n^{-1} X_n = (X_n - \hat{X}_n)^\top D^{-1} (X_n - \hat{X}_n) = \sum_{j=1}^{n} \frac{1}{v_{j-1}} (X_j - \hat{X}_j)^2$$

and

$$det\Gamma_n = (detC)^2 (detD) = v_0 v_1 .. v_{n-1}$$

so that the likelihood of the vector X_n simplifies to

$$L(\Gamma_n) = (2\pi)^{-\frac{n}{2}} (v_0 v_1 .. v_{n-1})^{-\frac{1}{2}} e^{-\frac{1}{2} \sum_{j=1}^{n} \frac{1}{v_{j-1}} (X_j - \hat{X}_j)^2} \tag{D.2.10}$$

where from the covariance κ we get $\hat{X}_1, \hat{X}_2, .., v_0, v_1, ...$ and hence $L(\Gamma_n)$. If Γ_n can be expressed in terms of a finite number of unknown parameters $\beta_1, .., \beta_r$ as it is the case when $\{X_t\}$ is an $ARMA(p, q)$ process and $r = p+q+1$, it is necessary to estimate the parameters from the data X_n. In this situation, one maximise the likelihood $L(\beta_1, .., \beta_r)$ with respect to $\beta_1, .., \beta_r$. Hence, a natural estimation procedure for Gaussian processes is to maximise the above likelihood with respect to $\beta_1, .., \beta_r$. Even if $\{X_t\}$ is not Gaussian we can still regard the above likelihood as a measure of the

goodness of fit of the covariance matrix $\Gamma_n(\beta_1, .., \beta_r)$ to the data and choose $\beta_1, .., \beta_r$ to maximise the likelihood. As a result, the estimators $\hat{\beta}_1, .., \hat{\beta}_r$ are called the maximum likelihood estimators.

We consider that $\{X_t\}$ is a causal $ARMA(p,q)$ process with Equation (D.2.4) where $\theta_0 = 1$ and assume that the coefficients θ_i and white noise variance σ^2 have been adjusted to ensure that $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q \neq 0$ for $|z| < 1$. We know that the one-step predictors $\hat{X}_{i+1}$ and their mean-squared errors are given by

$$\begin{aligned}\hat{X}_{i+1} &= \sum_{j=1}^{i} \theta_{ij}(X_{i+1-j} - \hat{X}_{i+1-j}) \text{ , } 1 \leq i < m = \max(p,q) \\ \hat{X}_{i+1} &= \phi_1 X_i + .. + \phi_p X_{i+1-p} + \sum_{j=1}^{q} \theta_{ij}(X_{i+1-j} - \hat{X}_{i+1-j}) \text{ , } i \geq m\end{aligned}$$

and

$$E[(X_{i+1} - \hat{X}_{i+1})^2] = \sigma^2 r_i$$

where θ_{ij} and r_i are estimated from the covariance function and are independent of σ^2. Plugging back in Equation (D.2.10) the Gaussian likelihood of the vector X_n is

$$L(\phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}}(r_0 r_1 .. r_{n-1})^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{1}{r_{j-1}}(X_j - \hat{X}_j)^2} \tag{D.2.11}$$

Differentiating $\ln L(\phi, \theta, \sigma^2)$ partially with respect to σ^2 and noting that $\hat{X}_j$ and r_j are independent of σ^2 we deduce that the maximum likelihood estimators $\hat{\phi}$, $\hat{\theta}$ and $\hat{\sigma}^2$ satisfy

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\phi}, \hat{\theta})$$

where

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n} \frac{1}{r_{j-1}}(X_j - \hat{X}_j)^2$$

and $\hat{\phi}$, $\hat{\theta}$ are the values of ϕ, θ minimising

$$l(\phi, \theta) = \ln \frac{1}{n} S(\phi, \theta) + \frac{1}{n}\sum_{j=1}^{n} \ln r_{j-1}$$

where $l(\phi, \theta)$ is the reduced likelihood. One can use a non-linear minimisation program in conjunction with the innovations algorithm to search for the value of ϕ and θ minimising $l(\phi, \theta)$. The search procedure can be greatly accelerated by choosing initial values ϕ_0 and θ_0 close to the minimum of l. Further, it is essential to start the search with a causal parameter ϕ_0 as the causality is assumed in the computation of $l(\phi, \theta)$.

An alternative estimation procedure is to minimise the weighted sum of squares

$$S(\phi, \theta) = \sum_{j=1}^{n} \frac{1}{r_{j-1}}(X_j - \hat{X}_j)^2$$

with respect to ϕ and θ. The estimators $\tilde{\phi}$ and $\tilde{\theta}$ of ϕ and θ are called the least-squares estimators. For the minimisation of $S(\phi, \theta)$ it is necessary to restrict ϕ to be causal, but also to restrict θ to be invertible as otherwise there will be no finite (ϕ, θ) at which S achieves its minimum value. The least-squares estimator $\tilde{\sigma}^2_{LS}$ is given by

$$\tilde{\sigma}^2_{LS} = \frac{1}{n-p-q} S(\tilde{\phi}, \tilde{\theta})$$

where $(n-p-q)$ is used since $\frac{1}{\sigma^2} S(\tilde{\phi}, \tilde{\theta})$ is distributed approximately as chi-square with $(n-p-q)$ degrees of freedom.

D.3 Fitting ARIMA models

When selecting an appropriate model for a given set of observations $\{X_t, t = 1, .., n\}$ if the data

1. exhibits no apparent deviations from stationarity
2. has a rapidly decreasing autocorrelation function

we shall seek a suitable ARMA process to represent the mean-corrected data. If not, we shall first look for a transformation of the data which generates a new series with the above properties. This can be achieved by differencing and hence considering the class of ARIMA (autoregressive integrated moving average) processes. Once the data has been suitably transformed, the problem becomes one of finding a satisfactory $ARMA(p,q)$ model and choosing p and q. Among the various criterions for model selection, a general criterion for model selection is the information criterion of Akaike [1973] known as the AIC. It was designed to be an approximately unbiased estimate of the Kullback-Leibler index of the fitted model relative to the true model. Later, Hurvich and Tsai [1989] proposed a bias-corrected version of the AIC called AICC. According to this criterion we compute maximum likelihood estimators of ϕ, θ and σ^2 for a variety of competing p and q values and choose the fitted model with smallest AICC values. If the fitted model is satisfactory, the residuals should resemble white noise.

The ARIMA models incorporate a wide range of non-stationary series which after differencing finitely many times reduce to ARMA processes. For instance, if the original process $\{X_t\}$ is not stationary, we can look at the first order difference process

$$Y_t = \nabla X_t = X_t - X_{t-1}$$

or the second order differences

$$Y_t = \nabla^2 X_t = \nabla(\nabla X)_t = X_t - 2X_{t-1} + X_{t-2}$$

and so on. When the differenced process is a stationary process we can look for a ARMA model. The process $\{X_t\}$ is said to be an autoregressive integrated moving average process $ARIMA(p,d,q)$ if $Y_t = \nabla^d X_t$ is $ARMA(p,q)$ process.

Definition D.3.1 *(The $ARIMA(p,d,q)$ process) If d is a non-negative integer, then $\{X_t\}$ is said to be an $ARIMA(p,d,q)$ process if $Y_t = (1-B)^d X_t$ is a causal $ARMA(p,q)$ process.*

This means that $\{X_t\}$ satisfies a difference equation of the form

$$\phi(B)\nabla(B)^d X_t = \theta(B) Z_t \, , \{Z_t\} \sim WN(0, \sigma^2)$$

where $\nabla B = I - B$, or alternatively

$$\phi^*(B) X_t = \phi(B)(1-B)^d X_t = \theta(B) Z_t \, , \{Z_t\} \sim WN(0, \sigma^2)$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q respectively and $\phi(z) \neq 0$ for $|z| \leq 1$. The polynomial $\phi^*(z)$ has a zero of order d at $z = 1$. The process $\{X_t\}$ is stationary if and only if $d = 0$. Note, if $d \geq 1$ we can add an

arbitrary polynomial trend of degree $(d-1)$ to $\{X_t\}$ without violating the above equation so that ARIMA can be used to represent data with trend.
For example, $\{X_t\}$ is an $ARIMA(1,1,0)$ process if for some $\phi \in (-1,1)$

$$(1-\phi B)(1-B)X_t = Z_t \ , \{Z_t\} \sim WN(0,\sigma^2)$$

we can write

$$X_t = X_0 + \sum_{j=1}^{t} Y_j \ , t \geq 1$$

where

$$Y_t = (1-B)X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

The Box-Jenkins procedure is concerned with fitting an ARIMA model to data. It has three parts: identification, estimation, and verification. As part of the identification process, the data may require pre-processing to make it stationary. To achieve stationarity we may do any of the following

- Re-scale it (for instance, by a logarithmic or exponential transform.)
- Remove deterministic components.
- Difference it, by taking $\nabla(B)^d X$ until stationary. In practice $d = 1, 2$ should suffice.

We recognise stationarity by the observation that the autocorrelations decay to zero exponentially fast. Once the series is stationary, we can try to fit an $ARMA(p,q)$ model. An $ARMA(p,q)$ process has kth order sample ACF and PACF decaying geometrically for $k > \max(p,q)$.

We assume that the ARIMA models have been differenced finitely many times and reduce to ARMA processes. Let $\{X_t\}$ denote the mean-corrected transformed series, we want to find the most satisfactory $ARMA(p,q)$ model to represent $\{X_t\}$ by identifying appropriate values for p and q. Even though it might appear that the higher the values of p and q chosen, the better the fitted model will be, we must be aware of the danger of overfitting (tailoring the fit too closely to the particular numbers observed). Akaike's AIC criterion and Parzen's CAT criterion attempt to prevent overfitting by effectively assigning a cost to the introduction of each additional parameter. We consider a bias-corrected form of the AIC defined for an $ARMA(p,q)$ model with coefficient vectors ϕ and θ, by

$$AICC(\phi,\theta) = -2\ln L(\phi,\theta,\frac{1}{n}S(\phi,\theta)) + \frac{2(p+q+1)n}{(n-p-q-2)}$$

and the model selected is the one which minimises the value of the AICC. One can think of $\frac{2(p+q+1)n}{(n-p-q-2)}$ as a penalty term to discourage over-parametrisation. Once a model has been found which minimises the AICC value, it must then be checked for goodness of fit by checking that the residuals are like white noise. Note, the search for a model minimising the AICC can be very lengthy without some idea of the class of models to be explored. A variety of techniques can be used to accelerate the search by considering preliminary estimates of p and q based on sample autocorrelation and partial autocorrelation functions.

The identification of a pure autoregressive or moving average process is reasonably straightforward using the sample autocorrelation and partial autocorrelation functions and the AICC. However, for $ARMA(p,q)$ processes with p and q both non-zero, the sample ACF and PACF are much more difficult to interpret. We can directly search for values p and q such that the AICC is minimum. The search can be carried out in a variety of ways: by trying all (p,q) values such that $p+q=1$, then $p+q=2$ etc., or by using the following steps

1. use maximum likelihood estimation to fit ARMA processes of orders $(1,1),(2,2),..$, to the data, selecting the model which gives smallest value of the AICC.
2. starting from the minimum AICC $ARMA(p,q)$ model, eliminate one or more coefficients (guided by the standard errors of the estimated coefficients), maximise the likelihood for each reduced model and compute the AICC value.
3. select the model with smallest AICC value (subject to its passing the goodness of fit tests)

When an ARMA model is fitted to a given series, an essential part of the procedure is to examine the residuals, which should resemble white noise for the model to be satisfactory. If the autocorrelations and partial autocorrelations of the residuals suggest that they come from some other identifiable process, then this more complicated model for the residuals can be used to suggest a more appropriate model for the original data. For instance, if the residuals appear to come from an ARMA process with coefficient vectors ϕ_Z and θ_Z it indicates that $\{Z_t\}$ in our fitted model should satisfy

$$\phi_Z(B)Z_t = \theta_Z(B)W_t$$

where $\{W_t\}$ is white noise. Applying the operator $\phi_Z(B)$ to each side of the equation defining $\{X_t\}$ we obtain

$$\phi_Z(B)\phi(B)X_t = \phi_Z(B)\theta(B)Z_t = \theta_Z(B)\theta(B)W_t$$

where $\{W_t\}$ is white noise. The modified model for $\{X_t\}$ is thus an ARMA process with autoregressive and moving operators $\phi_Z(B)\phi(B)$ and $\theta_Z(B)\theta(B)$ respectively.

The Akaike Information Criterion, AIC (Akaike [1973]), and a bias-corrected version, AICC (Sugiura, 1978; Hurvich and Tsai [1989]) are two methods for selection of regression and autoregressive models. Both criteria may be viewed as estimators of the expected Kullback-Leibler information. The main idea is that we want to fit a model with parametrised likelihood function $f(X|\theta)$ for $\theta \in \Theta$, and this includes the true model for some $\theta_0 \in \Theta$. Let $X = (X_1,..,X_n)$ be a vector of n independent samples and let $\hat{\theta}(X)$ be the maximum likelihood estimator of θ. Suppose Y is a further independent sample. Then

$$-2nE_Y E_X[\log f(Y|\hat{\theta}(X))] = -2E_X[\log f(X|\hat{\theta}(X))] + 2k + o(\frac{1}{\sqrt{n}})$$

where $k = |\Theta|$. The left hand side is $2n$ times the conditional entropy of Y given $\hat{\theta}(X)$, that is, the average number of bits required to specify Y given $\hat{\theta}(X)$. The right hand side is approximately the AIC and this is to be minimised over a set of models, say $(f_1,\Theta_1),..,(f_m,\Theta_m)$. Generally, we use the maximum likelihood estimators, or least squares numerical approximations to the MLEs. The essential idea is prediction error decomposition. We can factorise the joint density of $(X_1,..,X_n)$ as

$$f(X_1,..,X_n) = f(X_1)\prod_{t=2}^{n} f(X_t|X_1,..,X_{t-1})$$

Suppose $f(X_t|X_1,..,X_{t-1})$ the conditional distribution of X_t given $(X_1,..,X_{t-1})$ is normal with mean $\hat{X}_t$ and variance P_{t-1}, and suppose also that X_1 is normal $N(\hat{X}_1,P_0)$. Here $\hat{X}_t$ and P_{t-1} are functions of the unknown parameters $\phi_1,..,\phi_p,\theta_1,..,\theta_q$ and the data. The log likelihood is

$$-2\log L = -2\log f = \sum_{t=1}^{n}\big(\log 2\pi + \log P_{t-1} + \frac{(X_t-\hat{X}_t)^2}{P_{t-1}}\big)$$

We can minimise this equation with respect to $\phi_1, .., \phi_p, \theta_1, .., \theta_q$ to fit the $ARMA(p,q)$ model. Additionally, the second derivative matrix of $-\log L$ (at the MLE) is the observed information matrix, whose inverse is an approximation to the variance-covariance matrix of the estimators.

More formally, if X is an n-dimensional random vector whose probability density belongs to the family $\{f(.;\psi), \psi \in \Psi\}$, the Kullback-Leibler discrepancy between $f(.;\psi)$ and $f(.;\theta)$ is defined as

$$d(\psi|\theta) = \Delta(\psi|\theta) - \Delta(\theta|\theta)$$

where

$$\Delta(\psi|\theta) = E^{\theta}[-2\ln f(X;\psi)] = \int_{\mathbb{R}^n} -2\ln\left(f(x;\psi)\right) f(x;\theta)dx$$

is the Kullback-Leibler index of $f(.;\psi)$ relative to $f(.;\theta)$. Applying Jensen's inequality, we get

$$\begin{aligned} d(\psi|\theta) &= \int_{\mathbb{R}^n} -2\ln\left(\frac{f(x;\psi)}{f(x;\theta)}\right) f(x;\theta)dx \\ &\geq -2\ln\int_{\mathbb{R}^n} \frac{f(x;\psi)}{f(x;\theta)} f(x;\theta)dx \\ &= -2\ln\int_{\mathbb{R}^n} f(x;\psi)dx = 0 \end{aligned}$$

with equality holding if and only if $f(x;\psi) = f(x;\theta)$ a.e.. Given observations $X_1, .., X_n$ of an ARMA process with unknown parameters $\theta = (\beta, \sigma^2)$ one could identify the true model if it were possible to compute the Kullback-Leibler discrepancy between all candidate models and the true model. Instead, one must estimate the Kullback-Leibler discrepancies and choose the model whose estimated discrepancy is minimum. To do so, we can assume that the true model and the alternatives are all Gaussian. Then, for any given $\theta = (\beta, \sigma^2)$, $f(x;\theta)$ is the probability density of $(Y_1, .., Y_n)^{\top}$ where $\{Y_t\}$ is a Gaussian $ARMA(p,q)$ process with coefficient vector β and white noise variance σ^2.

Assume that our observations $X_1, .., X_n$ are from a Gaussian ARMA process with parameter vector $\theta = (\beta, \sigma^2)$ and assume that the true order is (p,q). Let $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ be the maximum likelihood estimator of θ based on $X_1, .., X_n$ and let $Y_1, .., Y_n$ be an independent realisation of the true process (with parameter θ), then

$$-2\ln L_Y(\hat{\beta}, \hat{\sigma}^2) = -2\ln L_X(\hat{\beta}, \hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} S_Y(\hat{\beta}) - n$$

so that

$$\begin{aligned} E^{\theta}[\Delta(\hat{\theta}|\theta)] &= E^{\beta,\sigma^2}[-2\ln L_Y(\hat{\beta}, \hat{\sigma}^2)] \\ &= E^{\beta,\sigma^2}[-2\ln L_X(\hat{\beta}, \hat{\sigma}^2)] + E^{\beta,\sigma^2}[\frac{1}{\hat{\sigma}^2} S_Y(\hat{\beta})] - n \end{aligned}$$

Making some local linearity approximation, we get

$$E^{\beta,\sigma^2}[\frac{1}{\hat{\sigma}^2} S_Y(\hat{\beta})] - n \approx \frac{2(p+q+1)n}{(n-p-q-2)}$$

Thus the quantity

$$-2\ln L_X(\hat{\beta}, \hat{\sigma}^2) + \frac{2(p+q+1)n}{(n-p-q-2)}$$

is an approximately unbiased estimate of the expected Kullback-Leibler index $E^{\theta}[\Delta(\hat{\theta}|\theta)]$. Since these calculations are based on the assumption that the true order is (p,q), we select the values of p and q for the fitted model to be those minimising $AICC(\hat{\beta})$ where

$$AICC(\beta) = -2\ln L_X(\beta, \frac{1}{n}S_X(\beta)) + \frac{2(p+q+1)n}{(n-p-q-2)}$$

The AIC statistic, defined as

$$AIC(\beta) = -2\ln L_X(\beta, \frac{1}{n}S_X(\beta)) + 2(p+q+1)$$

can be used in the same way. Note, both statistics are minimised for any given β by setting $\sigma^2 = \frac{1}{n}S_X(\beta)$. Further, the penalty factors $\frac{2(p+q+1)n}{(n-p-q-2)}$ and $2(p+q+1)$ are asymptotically equivalent as $n \to \infty$. However, the AICC statistic has a more extreme penalty for large-order models which counteracts the overfitting tendency of the AIC. Hurvich and Tsai [1991] considered both normal linear regression and autoregressive candidate models. They showed that the bias of AICC is typically smaller, often dramatically smaller, than that of AIC. A simulation study in which the true model is an infinite-order autoregression shows that, even in moderate sample sizes, AICC provides substantially better model selections than AIC.

The third stage in the Box-Jenkins algorithm is to check whether the model fits the data. There are several tools we may use

- Overfitting. Add extra parameters to the model and use likelihood ratio test or t-test to check that they are not significant.
- Residuals analysis. Calculate the residuals from the model and plot them. The autocorrelation functions, ACFs, PACFs, spectral densities, estimates, etc., and confirm that they are consistent with white noise.

The goodness of fit of a statistical model to a set of data is judged by comparing the observed values with the corresponding predicted values obtained from the fitted model. If the fitted model is appropriate, then the residuals should behave in a consistent manner with the model. In the case of an $ARMA(p,q)$ model with estimators $\hat{\phi}$, $\hat{\theta}$, and $\hat{\sigma}^2$ we let the predicted values $\hat{X}_t(\hat{\phi},\hat{\theta})$ of X_t based on $X_1,..,X_{t-1}$ be computed for the fitted model. The residuals are given by

$$\hat{W}_t = \frac{(X_t - \hat{X}_t(\hat{\phi},\hat{\theta}))}{\sqrt{r_{t-1}(\hat{\phi},\hat{\theta})}}, t = 1,..,n$$

In the case where the maximum likelihood $ARMA(p,q)$ model is the true process generating $\{X_t\}$, then $\{\hat{W}_t\} \sim WN(0,\hat{\sigma}^2)$. But, since we assume that $X_1,..,X_n$ is generated by an $ARMA(p,q)$ process with unknown parameters ϕ, θ, and σ^2 with maximum likelihood estimators $\hat{\phi}$, $\hat{\theta}$, and $\hat{\sigma}^2$ then $\{\hat{W}_t\}$ is not a true white noise process. Nonetheless, $\hat{W}_t$ for $t = 1,..,n$ should have properties similar to those of the white noise sequence

$$W_t(\phi,\theta) = \frac{(X_t - \hat{X}_t(\phi,\theta))}{\sqrt{r_{t-1}(\phi,\theta)}}, t = 1,..,n$$

As $E[(W_t(\phi,\theta) - Z_t)^2]$ is small for large t, so that the properties of the residuals $\{\hat{W}_t\}$ should reflect those of the white noise sequence $\{Z_t\}$ generating the underlying $ARMA(p,q)$ process.

The next step is to check that the sample autocorrelation function of $\hat{W}_1,..,\hat{W}_n$ behaves as it should under the assumption that the fitted model is appropriate. The sample autocorrelation function of an iid sequence $Z_1,..,Z_n$ with $E[Z_t^2] < \infty$ are for large n approximately iid with distribution $N(0,\frac{1}{n})$. Therefore, assuming that we have fitted an

appropriate ARMA model to our data and that the ARMA model is generated by an iid white noise sequence, the same approximation should be valid for the sample autocorrelation function of $\hat{W}_t$ for $t = 1, .., n$ defined by

$$\hat{\rho}_W(h) = \frac{\sum_{t=1}^{n-h}(\hat{W}_t - \overline{W})(\hat{W}_{t+h} - \overline{W})}{\sum_{t=1}^{n}(\hat{W}_t - \overline{W})^2}, h = 1, 2, ..$$

where $\overline{W} = \frac{1}{n}\sum_{t=1}^{n}\hat{W}_t$. However, since each $\hat{W}_t$ is a function of the maximum likelihood estimator $(hat\phi, \hat{\theta})$ then $\hat{W}_1, .., \hat{W}_n$ is not an iid sequence and the distribution of $\hat{\rho}_W(h)$ is not the same as in the iid case. In fact $\hat{\rho}_W(h)$ has an asymptotic variance which for small lags is less than $\frac{1}{n}$ and which for large lags is close to $\frac{1}{n}$.

If $\{X_t\}$ is a causal invertible ARMA process, assuming $h \geq p + q$ we set

$$T_h = [a_{i-j}]_{1 \leq i \leq h, 1 \leq j \leq p+q}$$

and

$$\tilde{\Gamma}_{p+q} = [\sum_{k=0}^{\infty} a_k a_{k+|i-j|}]_{i,j=1}^{p+q}$$

and

$$Q = T_h \frac{1}{\tilde{\Gamma}_{p+q}} T_h^{\top} = [q_{ij}]_{i,j=1}^{h}$$

The matrix $\tilde{\Gamma}_{p+q}$ is the covariance matrix of $(Y_1, .., Y_{p+q})$ where $\{Y_t\}$ is an $AR(p + q)$ process. It can be shown that

$$\hat{\rho}_W \text{ is } AN(0, \frac{1}{n}(I_h - Q))$$

where I_h is the $h \times h$ identity matrix. The asymptotic variance of $\hat{\rho}_W(i)$ is thus

$$\frac{1}{n}(1 - q_{ii})$$

Instead of checking to see if each $\hat{\rho}_W(i)$ falls within the confidence bounds

$$\pm 1.96 \frac{1}{\sqrt{n}} \sqrt{1 - q_{ii}}$$

it is possible to consider a single statistic which depends on $\hat{\rho}_W(i)$ for $1 \leq i \leq h$. To do so, we assume that h depends on the sample size n in such way that

- $h_n \to \infty$ as $n \to \infty$
- $\psi_j = o(\frac{1}{\sqrt{n}})$ for $j \geq h_n$ where ψ_j for $j = 0, 1, ..$ are the coefficients in the expansion $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$
- $h_n = o(\sqrt{n})$

Then as $h_n \to \infty$ the matrix $\tilde{\Gamma}_{p+q}$ may be approximated by $T_h^{\top} T_h$ and the matrix Q may be approximated by the projection matrix

$$T_h \frac{1}{T_h^{\top} T_h} T_h^{\top}$$

which has rank $p + q$. Hence the distribution of

$$Q_W = n\hat{\rho}_W^\top \hat{\rho}_W = n\sum_{j=1}^{h} \hat{\rho}_W^2(j)$$

is approximately chi-squared with $h-(p+q)$ degrees of freedom. The adequacy of the model is therefore rejected at level α if

$$Q_W > \chi_{1-\alpha}^2(h-p-q)$$

Examination of the squared residuals may often suggest departures of the data from the fitted model which could not otherwise be detected from the residuals themselves. We can test the squared residuals for correlation by letting

$$\hat{\rho}_{WW}(h) = \frac{\sum_{t=1}^{n-h}(\hat{W}_t^2 - \overline{W}^2)(\hat{W}_{t+h}^2 - \overline{W}^2)}{\sum_{t=1}^{n}(\hat{W}_t^2 - \overline{W}^2)}, h \geq 1$$

be the sample autocorrelation function of the squared residuals where $\overline{W}^2 = \frac{1}{n}\sum_{t=1}^{n}\hat{W}_t^2$. Then McLeod and Li (1983) showed that

$$\tilde{Q}_{WW} = n(n+2)\sum_{j=1}^{h}\frac{1}{n-j}\hat{\rho}_{WW}^2(j)$$

has an approximate $\chi^2(h)$ distribution under the assumption of model adequacy. As a result, the adequacy of the model is rejected at level α if

$$\tilde{Q}_{WW} > \chi_{1-\alpha}^2(h)$$

In practice, portmanteau tests are more useful for disqualifying unsatisfactory models from consideration than for selecting the best-fitting model among closely competing candidates. There are a number of other tests available for checking the hypothesis of randomness of $\{\hat{W}_t\}$, that is, the hypothesis that it is an iid sequence. One can consider a test based on turning points, the difference-sign test, or the rank test (see Kendall and Stuart [1976]).

Checking for normality

If it can be assumed that the white noise process $\{Z_t\}$ generating an $ARMA(p,q)$ process is Gaussian, then stronger conclusion can be drawn from the fitted model. One can specify an estimated mean squared error for predicted values, and asymptotic prediction confidence bounds can also be computed. So, let $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$ be the order statistics of a random sample $Y_1, .., Y_n$ from the distribution $N(\mu, \sigma^2)$. If $X_{(1)} < X_{(2)} < ... < X_{(n)}$ are the order statistics from a $N(0,1)$ sample size n, then

$$E[Y_{(j)}] = \mu + \sigma m_j$$

where $m_j = E[X_{(j)}]$ for $j = 1, .., n$. Thus, a plot of the points $(m_1, Y_{(1)}), .., (m_n, Y_{(n)})$ should be approximately linear. This is not the case if the sample values Y_i are not normally distributed. As a result, the squared correlation of the points $(m_i, Y_{(i)})$ for $i = 1, .., n$ should be near one if the normal assumption is correct. The assumption of normality is therefore rejected if the squared correlation R^2 is sufficiently small. If we approximate m_i by $\Phi^{-1}(\frac{i-0.5}{n})$ then R^2 reduces to

$$R^2 = \frac{\left(\sum_{i=1}^{n}(Y_{(i)} - \overline{Y})\Phi^{-1}(\frac{i-0.5}{n})\right)^2}{\sum_{i=1}^{n}(Y_{(i)} - \overline{Y})^2 \sum_{i=1}^{n}(\Phi^{-1}(\frac{i-0.5}{n}))^2}$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

D.4 State space models

A state space model is defined by a measurement equation postulating the relationship between an observable vector and a state vector, and a transition equation describing the generating process of the state variables. State space models are an alternative formulation of time series with a number of advantages for forecasting.

1. All ARMA models can be written as state space models.
2. Nonstationary models (e.g., ARMA with time varying coefficients) are also state space models.
3. Multivariate time series can be handled more easily.
4. State space models are consistent with Bayesian methods.

In general, the model consists of

$$\begin{aligned} &X_t = F_t S_t + v_t \text{ observed data} \\ &S_t = G_t S_{t-1} + w_t \text{ unobserved state} \\ &v_t \sim N(0, V_t) \text{ observation noise} \\ &w_t \sim N(0, W_t) \text{ state noise} \end{aligned} \tag{D.4.12}$$

where X_t is a $(n, 1)$ vector of time series, and the state vector S_t is made of $(m, 1)$ state variables. Further, v_t, w_t are independent and F_t, G_t are known matrices, often time dependent because of seasonality, of dimension (n, m) and (m, m).
Example 1

$$\begin{aligned} X_t &= S_t + v_t \\ S_t &= \phi S_{t-1} + w_t \end{aligned}$$

Define

$$Y_t = X_t - \phi X_{t-1} = (S_t + v_t) - \phi(S_{t-1} + v_{t-1}) = w_t + v_t - \phi v_{t-1}$$

The autocorrelations of $\{y_t\}$ are zero at all lags greater than 1. So, $\{Y_t\}$ is $MA(1)$ and thus $\{X_t\}$ is $ARMA(1, 1)$.
Example 2
The general $ARMA(p, q)$ model

$$X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \sum_{s=1}^{q} \theta_s \epsilon_{t-s}$$

is a state space model. We write $X_t = F_t S_t$ where

$$F_t = (\phi_1, .., \phi_p, 1, \theta_1, .., \theta_q)$$

and

$$S_t = (X_{t-1}, .., X_{t-p}, \epsilon_t, ..., \epsilon_{t-q})^\top \in \mathbb{R}^{p+q+1}$$

with $v_t = 0$, $V_t = 0$.

The Kalman filter (see Kalman [1960])
Given observed data $X_1, ..., X_t$ we want to find the conditional distribution of S_t and a forecast of X_{t+1}. Recall the following multivariate normal fact: If

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right)$$

then

$$(Y_1|Y_2) = N\big(\mu_1 + A_{12}A_{22}^{-1}(Y_2 - \mu_2), A_{11} - A_{12}A_{22}^{-1}A_{21}\big)$$

Conversely, if $(Y_1|Y_2)$ satisfies the above equation, and $Y_2 \sim N(\mu_2, A_{22})$ then the joint distribution is as above. Now let $\mathcal{F}_{t-1} = (X_1, .., X_{t-1})$ and suppose we know that $(S_{t-1}|\mathcal{F}_{t-1}) \sim N(\hat{S}_{t-1}, P_{t-1})$. That is, $\hat{S}_{t-1} = E_{t-1}[S_{t-1}]$ and $P_{t-1} = E_{t-1}[(\hat{S}_{t-1} - S_{t-1})(\hat{S}_{t-1} - S_{t-1})^\top]$ is the covariance matrix. Then, given the dynamics of the state vector

$$S_t = G_t S_{t-1} + w_t$$

we get

$$(S_t|\mathcal{F}_{t-1}) \sim N(G_t\hat{S}_{t-1}, G_tP_{t-1}G_t^\top + W_t)$$

where $\hat{S}_{t|t-1} = E_{t-1}[S_t] = G_t\hat{S}_{t-1}$ and $P_{t|t-1} = E_{t-1}[(\hat{S}_{t|t-1} - S_t)(\hat{S}_{t|t-1} - S_t)^\top] = G_tP_{t-1}G_t^\top + W_t$ is the covariance matrix. See the invariance of the covariance of multivariate normal distribution under linear changes of variables in Theorem (B.8.2). Note, we also have $(X_t|S_t, \mathcal{F}_{t-1}) \sim N(F_tS_t, V_t)$. We put $Y_1 = X_t$ and $Y_2 = S_t$ and let $R_t = P_{t|t-1} = G_tP_{t-1}G_t^\top + W_t$ be the covariance matrix of $(S_t|\mathcal{F}_{t-1})$. Taking all variables conditional on $\mathcal{F}_{t-1}$ we can use the converse of the multivariate normal fact and identify

$$\mu_2 = G_t\hat{S}_{t-1} \text{ and } A_{22} = R_t$$

Since S_t is a random variable, we get

$$\mu_1 + A_{12}A_{22}^{-1}(S_t - \mu_2) = F_tS_t \to A_{12} = F_tR_t \text{ and } \mu_1 = F_t\mu_2$$

Also,

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = V_t \to A_{11} = V_t + F_tR_tR_t^{-1}R_t^\top F_t^\top = V_t + F_tR_tF_t^\top$$

which says that

$$\begin{bmatrix} X_t \\ S_t \end{bmatrix}|_{\mathcal{F}_{t-1}} = N\left(\begin{bmatrix} F_tG_t\hat{S}_{t-1} \\ G_t\hat{S}_{t-1} \end{bmatrix}, \begin{bmatrix} V_t + F_tR_tF_t^\top & F_tR_t \\ R_t^\top F_t^\top & R_t \end{bmatrix} \right)$$

Since $X_{t|t-1} = E_{t-1}[X_t] = F_tG_t\hat{S}_{t-1}$, we can define $I_t = X_t - X_{t|t-1}$ to be the innovation process representing the observed error in forecasting X_t with covariance matrix $I_t^c = (V_t + F_tR_tF_t^\top)$. Now we can apply the multivariate normal fact directly to get $(S_t|X_t, \mathcal{F}_{t-1}) = (S_t|\mathcal{F}_t) \sim N(\hat{S}_t, P_t)$ where

$$\begin{aligned} \hat{S}_t &= G_t\hat{S}_{t-1} + R_tF_t^\top(V_t + F_tR_tF_t^\top)^{-1}(X_t - F_tG_t\hat{S}_{t-1}) = \hat{S}_{t|t-1} + R_tF_t^\top(I_t^c)^{-1}I_t \\ P_t &= R_t - R_tF_t^\top(V_t + F_tR_tF_t^\top)^{-1}F_tR_t = \big(I_m - R_tF_t^\top(I_t^c)^{-1}F_t\big)R_t \end{aligned}$$

with $P_t = E_t[(\hat{S}_t - S_t)(\hat{S}_t - S_t)^\top]$ is the covariance matrix of $(S_t|\mathcal{F}_t)$. This system of equation is the Kalman filter updating equations or the innovation representation (see Harvey [1989]). The form of the right hand side of the expression for $\hat{S}_t$ contains the term $G_t\hat{S}_{t-1}$, which is simply what we would predict if it were known that $S_{t-1} = \hat{S}_{t-1}$

plus a term that depends on the observed error in forecasting X_t, that is, the innovation process I_t. This is similar to the forecast updating expression for simple exponential smoothing (). All we need to start updating the estimates are the initial values $\hat{S}_0$ and P_0. Three ways are commonly used

1. Use a Bayesian prior distribution.
2. If F,G, V,W are independent of t the process is stationary. We could use the stationary distribution of S to start.
3. Choosing $S_0 = 0$, $P_0 = kI$ (k large) reflects prior ignorance.

Prediction

Suppose we want to predict the term $X_{T+k} = X_{T+k|T}$ given $(X_1, .., X_T)$. We already have

$$(X_{T+1}|X_1, .., X_T) \sim N(F_{T+1}G_{T+1}S_t, V_{T+1} + F_{T+1}R_{T+1}F_{T+1}^\top)$$

which solves the problem for the case $k = 1$. By induction, Harvey [1989] showed that $S_{T+k} = S_{T+k|T}$ satisfies

$$(S_{T+k}|X_1, .., X_T) \sim N(\hat{S}_{T+k}, P_{T+k})$$

where

$$\begin{aligned}
\hat{S}_{T,0} &= \hat{S}_T \\
P_{T,0} &= P_T \\
\hat{S}_{T,k} &= G_{T+k}\hat{S}_{T,k-1} \\
P_{T,k} &= G_{T+k}P_{T,k-1}G_{T+k}^\top + W_{T+k}
\end{aligned}$$

and hence we obtain

$$(X_{T+k}|X_1, .., X_T) \sim N(F_{T+k}\hat{S}_{T,k}, V_{T+k} + F_{T+k}P_{T,k}F_{T+k}^\top)$$

D.5 ARCH and GARCH models

D.5.1 The ARCH process

Volatility clustering has long been a salient feature of series generated from financial data. However, only recently did researchers in finance recognise the importance of explicitly modelling time varying second-order moments. One of the most important and widely used approach is that of the Autoregressive Conditional Heteroskedastic (ARCH) model introduced by Engle [1982]. Following his seminal paper we define a zero mean $ARCH(p)$ process $\{X_t\}$ by

$$X_t = \epsilon_t h_t^{\frac{1}{2}}, \epsilon_t \sim N(0, 1)$$

and

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2$$

The above equations imply that $X_t \sim N(0, h_t)$ and that

$$X_t = \left(\alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2\right)^{\frac{1}{2}} \epsilon_t$$

A logic extension is to replace X_t by $X_t - \mu$ for all time t to get an $ARCH(p)$ process having a constant mean $E[X_t] = \mu$. That is

$$X_t - \mu = \left(\alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2\right)^{\frac{1}{2}} \epsilon_t$$

The attractiveness of the ARCH model for financial data is that it captures the tendency for volatility clustering. That is, large (small) changes tend to be followed by large (small) changes but of unpredictable sign.

For instance, consider the non-zero mean $ARCH(1)$ process

$$X_t - \mu = \left(\alpha_0 + \alpha_1 (X_{t-1} - \mu)^2\right)^{\frac{1}{2}} \epsilon_t$$

with $\epsilon_t \sim N(0,1)$. Then

$$Var(X_t | X_{t-1}) = \alpha_0 + \alpha_1 (X_{t-1} - \mu)^2$$

Hence, large deviation of X_{t-1} from the mean μ cause a large variance for the next time period. Conditioning on the immediate past, the process becomes heteroskedastic. Providing $\alpha_1 < 1$, the process is stationary with finite unconditional variance given by $\frac{\alpha_0}{1-\alpha_1}$. If $3\alpha_1^2 < 1$ then the fourth moment is finite and the kurtosis k_x is given by

$$k_x = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2}$$

Hence, we see that the kurtosis for an $ARCH(1)$ process exceed 3 for all $\alpha_1 > 0$ which is consistent with financial data.

D.5.2 The GARCH process

In empirical applications of the ARCH model, a relatively long lag in the conditional variance equation is often called for. In light of this, Bollerslev extended the ARCH class of models to allow for a longer memory and flexible lag structure. The Generalised Autoregressive Conditional Heteroskedastic (GARCH) model is defined as

$$X_t = h_t^{\frac{1}{2}} \epsilon_t$$

where

$$h_t = \alpha_0 + \sum_{i=1}^{p}$$

and

$$p \geq 0\,,\, q > 0\,,\, \alpha_0 > 0\,,\, \alpha_i \geq 0\,,\, \beta_j \geq 0$$

In the $ARCH(p)$ process the conditional variance is specified as a linear function of past sample variances only, whereas the $GARCH(p,q)$ process allows lagged conditional variances to enter as well. Bollerslev [1986] states that it corresponds to some sort of adaptive learning mechanism.

For the $GARCH(1,1)$ model, under second order stationarity, the unconditional variance is equal to $\frac{\alpha_0}{1-(\alpha_1+\beta_1)}$. It can be shown that for the $GARCH(1,1)$ model, the $\{X_t\}$ are uncorrelated and the squared process is correlated, making it particularly attractive to the closing price series.

D.5.3 Estimating model parameters

We are now considering the method used to estimate the model parameters and limit our analysis to the $ARCH(p)$ process. The development of estimation methods for the $GARCH(p,q)$ model being similar. In order to estimate the parameters of the $ARCH(p)$ model we employ maximum likelihood estimation seeking the parameter values which maximise the log-likelihood function.

For the zero-mean $ARCH(p)$ process we let L be the overall likelihood, L_i be the conditional log-likelihood of the i-th observation and N be the sample size. Then we get

$$L = \sum_{i=1}^{N} L_{t_i}$$

and

$$L_{t_i} = -\frac{1}{2}\log h_{t_i} - \frac{X_{t_i}^2}{2h_{t_i}}$$

ignoring constants. The log-likelihood function is nonlinear and Fisher scoring is used to obtain the maximum likelihood estimates (see Engle [1982]). In general the maximum likelihood estimation is performed with several constraints imposed upon the parameters to protect against numerical problems from negative, zero, or infinite variances. For instance, in the $GARCH(1,1)$ model, under second order stationarity, it has unconditional variance equal to $\frac{\alpha_0}{1-(\alpha_1+\beta_1)}$ so that we require $\alpha_1 + \beta_1 < 1$. Estimating parameters by maximising the Gaussian likelihood yields consistent estimates when the errors are not Gaussian distributed, provided that they are i.i.d..

D.6 The linear equation

D.6.1 Solving linear equation

Following Karatzas & Shreve [1997], we consider the linear SDE

$$dX_t = \big(K_0(t) + K_1(t)X_t\big)dt + \Sigma_X(t)dW_t \text{ with } X_0 = \xi$$

where $X_t \in \mathbb{R}^n$, $K_0 \in \mathbb{R}^n$ and $K_1 \in \mathbb{R}^{n\times n}$ with the corresponding deterministic differential equation

$$\dot{Q}(t) = K_0(t) + K_1(t)Q(t) \text{ with } Q_0 = Q$$

where $Q_t \in \mathbb{R}^n$. It has the associated homogeneous equation

$$\dot{Q}(t) = K_1(t)Q(t)$$

where standard calculus guarantee a unique, absolutely continuous solution to this initial value problem. In the special case where the matrix $K_1(t)$ is constant, the solution to the linear system is

$$Q(t) = e^{K_1 t}Q(0)$$

Moler et al. [2003] showed that there are many different ways to compute the exponential of a matrix and tried to classify them according to some criterions such as accuracy and efficiency. We are now going to use Q_t as an integrating factor to solve for X_t. We first take its inverse $Q^{-1}(t)$ and differentiate it with respect to time t, getting

$$\frac{dQ^{-1}(t)}{dt} = -Q^{-1}(t)\frac{dQ(t)}{dt}Q^{-1}(t)$$

We then apply Ito's lemma to get the dynamic of the product of the multi-factor Ornstein-Uhlenbeck process with the matrix $Q^{-1}(t)$

$$dQ^{-1}(t)X_t = Q^{-1}(t)K_0(t)dt + Q^{-1}(t)\Sigma_X dW_X^T(t)$$

Integrating between $[t,T]$, we get the solution

$$X_T = Q^{-1}(t)X_t + \int_t^T Q^{-1}(s)K_0(s,T)ds + \int_t^T Q^{-1}(s)\Sigma_X(s)dW_X^T(s) \tag{D.6.13}$$

where $\int_t^T Q^{-1}(s)\Sigma_X(s)dW_X(s)$ is a $(n,1)$ stochastic integral. In that case, the conditional mean vector $M(t,T)$ and variance matrix $V(t,T)$ of the process X_t are

$$\begin{aligned} M(t,T) &= Q^{-1}(t)X_t - \int_t^T Q^{-1}(s)K_0(s)ds \\ V(t,T) &= \int_t^T Q^{-1}(s)\Sigma_X\Sigma_X^\top(s)(Q^{-1}(s))^\top ds \end{aligned}$$

D.6.2 A simple example

We consider a multivariate Ornstein-Uhlenbeck process $X = (X_t \in \mathbb{R}^2)_{t\in[0,T]}$ with initial value $X_0 \in \mathbb{R}^2$ defined by the SDE

$$dX_t = (K_0(t) - K_1X_t)dt + \Sigma(t)dW_t$$

with $K_0(t) \in \mathbb{R}^{2\times 2}$, $\Sigma(t) \in \mathbb{R}^{2\times 2}$ and $W_t \in \mathbb{R}^2$. The solution given in Equation (D.6.13) can be rewritten as

$$X_t = e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du + \int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u$$

D.6.2.1 Covariance matrix

We have

$$Var(X_t|X_s) = Var(e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du + \int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u)$$

and then

$$Var(X_t|X_s) = Var(e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du) + Var(\int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u)$$

which simplifies to

$$Var(X_t|X_s) = Var(\int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u)$$

From the definition of the variance we get

$$Var(X_t|X_s) = E[(\int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u)(\int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u)^\top]$$

From Itô's isometry we get

$$Var(X_t|X_s) = E[\int_s^t (e^{-(t-u)K_1}\Sigma(u))(e^{-(t-u)K_1}\Sigma(u))^\top du]$$

which becomes

$$Var(X_t|X_s) = \int_s^t (e^{-(t-u)K_1}\Sigma(u))(e^{-(t-u)K_1}\Sigma(u))^\top du$$

D.6.2.2 Expectation

From the Equation (D.6.13) we have the conditional expectation

$$E[X_t|X_s] = E[e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du + \int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u]$$

which we split as

$$E[X_t|X_s] = E[e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du] + E[\int_s^t e^{-(t-u)K_1}\Sigma(u)dW_u]$$

The conditional expectation becomes

$$E[X_t|X_s] = e^{-(t-s)K_1}X_s + \int_s^t e^{-(t-u)K_1}K_0(u)du$$

D.6.2.3 Distribution and probability

We have proven that X_t is a normal multivariate random variable with mean $M_t = E[X_t|X_0] \in \mathbb{R}^2$ and variance $V_t = Var(X_t|X_0) \in \mathbb{R}^2$ (We can write $X_t \sim \mathcal{N}(M_t, V_t)$). In order to compute the probability of the first element, we assume that V_t is diagonalisable. We can then decompose it as $V_t = U\Lambda U^{-1}$, with $U \in \mathbb{R}^{2\times 2}$ being the unit eigenvectors and $\Lambda \in \mathbb{R}^{2\times 2}$ the diagonal matrix of the eigenvalues. Since $X_t \sim \mathcal{N}(M_t, V_t)$, we have:

$$X_t = M_t + U\Lambda^{\frac{1}{2}}Y$$

with $Y \sim \mathcal{N}(0, I)$. Let's have the following matrix representation:

$X_t = \begin{pmatrix} X \\ \bar{X} \end{pmatrix}$, $A = U\Lambda^{\frac{1}{2}} = \{A_{ij}\}_{i=1,2,j=1,2}$, $M_t = \{M_i\}_{i=1,2}$ and $Y = \{y_i\}_{i=1,2}$, then:

$$\begin{pmatrix} X \\ \bar{X} \end{pmatrix} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

Then we have the first element of our process corresponding to:

$$X = M_1 + A_{11}Y_1 + A_{12}Y_2$$

Let us have $\bar{Y} \sim \mathcal{N}(0, 1)$, then from the independence of Y_1 and Y_2, we get

$$X = M_1 + \sqrt{A_{11}^2 + A_{12}^2}\bar{Y}$$

Let us define $m_{X_t} = M_1$ and $\sigma_X = \sqrt{A_{11}^2 + A_{12}^2}$, then we have $X_t \sim \mathcal{N}(m_{X_t}, \sigma_X^2)$ and we can then compute the probability:

$$P(X_t < a) = P(m_{X_t} + \sigma_X\bar{Y} < a) = P\left(\bar{Y} < \frac{a - m_{X_t}}{\sigma_X}\right)$$

$$P(X_t < a) = \Phi\left(\frac{a - m_{X_t}}{\sigma_X}\right)$$

where Φ being the cumulative distribution function of a standard normal random variable. As a result, the probability becomes

$$P(X_t \geq a) = 1 - P(X_t < a) = 1 - \Phi\left(\frac{a - m_{X_t}}{\sigma_X}\right) = \Phi\left(\frac{m_{X_t} - a}{\sigma_X}\right)$$

D.6.3 From OU to AR(1) process

ARMA models are mathematical models of the persistence or autocorrelation in a time series. One subset of ARMA models are the autoregressive (AR) models and express a time series as a linear function of its past values. The order of the AR model tells how many lagged past values are included. Discretising an Ornstein-Uhlenbeck process, we recover the $AR(1)$ model.

D.6.3.1 The Ornstein-Uhlenbeck process

Assuming $(X_t)_{t\geq 0}$ is an Ornstein-Uhlenbeck process with positive constant parameters ρ and σ, the dynamics of the model are

$$dX_t = \rho(\overline{X} - X_t)dt + \sigma dW_t \tag{D.6.14}$$

In that model, the distribution of future values depends on the current value. The solution of the SDE is

$$\begin{aligned} X_t &= X_0 e^{-\rho t} + \rho \int_0^t \overline{X} e^{-\rho(t-s)} ds + Z_X(0,t) \\ &= X_0 e^{-\rho t} + \overline{X}(1 - e^{-\rho t}) + Z_X(0,t) \end{aligned}$$

where $Z_X(0,t) = \int_0^t \sigma e^{-\rho(t-s)} dW(s)$ is normally distributed with mean equal to zero and variance $V_{Z_X}(0,t) = \int_0^t \sigma^2 (e^{-\rho(t-s)})^2 ds = \frac{\sigma^2}{2\rho}(1 - e^{-2\rho t})$. The stationary (or unconditional) mean and variance are computed in the limit by letting the time to infinity

$$\begin{aligned} E[X_t] &= \overline{X} \\ Var(X_t) &= \frac{\sigma^2}{2\rho} \end{aligned}$$

Without loss of generality, we set $\overline{X} = 0$ since for any $\alpha \in \mathbb{R}$ the process $(X_t - \alpha)$ is also an OU process. Hence, if $X_0 = x$ with probability 1, then X_t has the distribution

$$(X_t | X_0 = x) \sim N\big(xe^{-\rho t}, \frac{\sigma^2}{2\rho}(1 - e^{-2\rho t})\big)$$

and we recover our Gaussian distribution with $\alpha(0,t) = xe^{-\rho t}$ and variance $V_{Z_X}(0,t)$. The covariance functions are given by

$$Cov(X_t, X_s | X_0 = x) = \frac{\sigma^2}{2\rho}\big(e^{-\rho|t-s|} - e^{-\rho(t+s)}\big)$$

D.6.3.2 Deriving the discrete model

Given the dynamics of the OU process in Equation (D.6.14), we let $T = N\Delta t$ and assume that historical data satisfies the discretised dynamics

$$X_n = \rho_0 \overline{X} \Delta t + (1 - \rho_0 \Delta t) X_{n-1} + \sigma_0 \sqrt{\Delta t} y \ , n = 1, .., N$$

where $y \sim N(0, 1)$, that is, the distribution of y is time independent. Given X_0, the parameters ρ_0 and σ_0 are unknown and must be estimated. We consider a parametric model where the conditional mean and variance of X_n given $X_{n-1} = x_{n-1}$ belong to the families

$$\{\rho \overline{X} \Delta t + (1 - \rho \Delta t) x_{n-1}, \rho \in \mathbb{R}\}$$
$$\{\sigma^2 \Delta t, \sigma^2 \in \mathbb{R}^+\}$$

Setting $K_1 = e^{-\rho \Delta t} \approx (1 - \rho \Delta t)$ and $\overline{\sigma} = \sigma \sqrt{\Delta t}$, we rewrite the process as

$$X_n = C + K_1 X_{n-1} + \overline{\sigma} y_n$$

for $C = (1 - K_1)\overline{X}$. The sequence $X_0, ..., X_N$ is a first order autoregressive sequence with lag-one correlation coefficient K_1. Positive autocorrelation might be considered a specific form of persistence, a tendency for a system to remain in the same state from one observation to the next. Interpolating linearly the values $X_n = X(n\Delta t)$ for $n \in [1, N]$ we recover the desired path. The equation for X_n is the recursive representation of the $AR(1)$ with conditional mean and variance

$$E[X_n | X_{n-1}] = C + K_1 X_{n-1}$$
$$Var(X_n | X_{n-1}) = \overline{\sigma}^2$$

If we lag that equation by p periods where $p = 1$ is the original equation and substitute each time the result back in the first equation, we get

$$X_n = C \sum_{j=0}^{p-1} K_1^j + K_1^p X_{n-p} + \overline{\sigma} \sum_{j=0}^{p-1} K_1^j y_{n-j}$$

where

$$C \sum_{j=0}^{p-1} K_1^j = (1 - K_1)\overline{X} \sum_{j=0}^{p-1} K_1^j = \overline{X} \sum_{j=0}^{p-1} K_1^j - K_1 \overline{X} \sum_{j=0}^{p-1} K_1^j = \overline{X} \sum_{j=0}^{p-1} K_1^j - \overline{X} \sum_{j=0}^{p-1} K_1^{j+1}$$

Assuming stationarity, that is $|K_1| < 1$, and taking the limit $p \to \infty$ then K_1^p will approach zero and from the infinite geometric series [1], we get

$$X_n = \overline{X} + \overline{\sigma} \sum_{j=0}^{\infty} K_1^j y_{n-j}$$

since $\frac{C}{1-K_1} = \overline{X}$, which is an infinite order moving average. Therefore, the variable X_n can be written as an infinite sum of past shocks where most distant shocks get smaller and smaller weights. In fact, the coefficients are geometrically declining since $|K_1| < 1$. Hence, when $K_1 = 1$ we get the unit root case and X_n has infinite memory,

[1] $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$ for $|x| < 1$

that is past shocks never dies out. As a result, the distant past matters more the closer K_1 is from unity. Since each y_{n-j} is an i.d.d standard normal and since $|K_1| < 1$ we compute the unconditional mean and variance as

$$E[X_n] = \overline{X}$$
$$Var(X_n) = \overline{\sigma}^2 \sum_{j=0}^{\infty} K_1^{2j} Var(y_{n-j}) = \frac{\overline{\sigma}^2}{1 - K_1^2}$$

where the unconditional variance is larger than the conditional one when $K_1 \neq 0$. The further we look into the future the larger the number of shocks which come into play so that the set of possibilities or variance grows as a function of K_1.

D.6.4 Some facts about AR series

D.6.4.1 Persistence

Autocorrelation refers to the correlation of a time series with its own past and future values. Positively autocorrelated series are referred to as persistent implying a positive dependence on the past and show up in a time series plot as unusually long runs or stretches of several consecutive observations above or below the mean. In the $AR(1)$ model, the higher K_1, the more persistence and the longer the series stay above or below the mean. When $K_1 = 1$ we get a random walk and the speed of mean reversion ρ in the continuous counterpart is zero. But as K_1 decreases to zero, the time series mean revert faster and ρ increases. Because the departures for computing autocorrelation are computed relative to the mean, a horizontal line plotted at the sample mean is useful in evaluating autocorrelation with the time series plot. Hence, the equations on the process X_t are somewhat simpler if the time series is first reduced to zero-mean by subtracting the sample mean, that is $V_n = X_n - \overline{X}$ for $n = 1, .., N$ where X_n is the original time series, $\overline{X}$ is its sample mean and V_n is the mean-adjusted series. In that case, the $AR(1)$ model for the mean-adjusted series in year n becomes

$$V_n = K_1 V_{n-1} + \overline{\sigma} y_n \tag{D.6.15}$$

We can deduce the e-folding time constant τ of the $AR(1)$ model at lag Δt from the relation

$$K_1 = e^{-\rho \Delta t} = e^{-\frac{\Delta t}{\tau}}$$

such that $\rho = \frac{1}{\tau}$.

D.6.4.2 Prewhitening and detrending

Prewhitening refers to the removal of autocorrelation from a time series prior to using the time series in some application. For a series with positive autocorrelation, prewhitening acts to damp those time series features that are characteristic of persistence. In general, a trend is a long term change in the mean but can also refer to change in other statistical properties. Detrending is the statistical operation of removing trend from the series. It can be applied to remove a feature thought to distort the relationships of interest or can be used as a preprocessing step to prepare time series for analysis by methods that assume stationarity. The variance at low frequencies is diminished relative to the variance of high frequencies resulting in lower spectrum (at lowest frequencies) of the detrended series compared with the spectrum of the original data. Simple linear trend in mean can be removed by subtracting a least-square fit straight line.

D.6.4.3 Simulation and prediction

Simulation is the generation of synthetic time series with the same persistence properties as the observed series. In our example, it effectively mimics the low-frequency behaviour of the observed series. That is, given Equation (D.6.15) we estimate the autoregressive parameter by modelling the time series as an $AR(1)$ process. Then we generate a time series of random noise ω by sampling from an appropriate distribution. Assuming some starting values for V_{n-1}, we recursively generate a time series of V_n. Usually, we assume that the noise is normally distributed with mean zero and variance equal to the variance of the residuals from fitting the $AR(1)$ model to the data. We can use simulation to generate empirical confidence intervals of the relationship between the observed time series and the climate variable.

Prediction is the extension of the observed series into the future based on past and present values. It differs from simulation in that the objective of prediction is to estimate the future value of the time series as accurately as possible from the current and past values. A prediction form of the $AR(1)$ model in Equation (D.6.15) is

$$\hat{V}_n = \hat{K}_1 V_{n-1}$$

where the hat indicates an estimate. The equation can be applied one step ahead to get estimate $\hat{V}_n$ from the observed V_{n-1} while k-step ahead prediction can be made by applying recursively the above equation. Note, because the modelling of V_n assume a departure from the mean, the convergence in terms of the original time series is a convergence toward the mean. Therefore, in the limit for large enough k, the predictions will eventually converge to zero. Besides the randomness of the residuals, we are concerned with the statistical significance of the model coefficients. Significance of the $AR(p)$ coefficients can be evaluated by comparing the estimated parameters with the standard deviation. In the $AR(1)$ model, the estimated first order autoregressive coefficient $\hat{K}_1$, is normally distributed with variance

$$Var(\hat{K}_1) = \frac{(1-\hat{K}_1^2)}{N} \tag{D.6.16}$$

Therefore, the 95% confidence interval for $\hat{K}_1$ is two standard deviations around $\hat{K}_1$, that is

$$95\%\,\text{CI} = \hat{K}_1 \pm 2\sqrt{Var(\hat{K}_1)}$$

D.6.5 Estimating the model parameters

For simplicity of exposition we change notations and denote the regression model with autocorrelated disturbances as follows

$$X_t = a_x + b_x X_{t-1} + \varepsilon_t^x$$

where ε_t^x is an error process. At this point we would need to estimate the model parameters. A first and common method would consist in computing the ordinary least squares (OLS) *i.e.* minimising the $\mathcal{L}^2(\mathbb{R})$ norm of residuals. But, using the OLS method implicitly assume that the process ε_t^x is a white noise and then that an hexogen perturbation of the temperature has no consequence on the future sea-level values. To test the hypothesis for white noise, we perform the *generalized Durbin-Watson tests* on the data. If the test is rejected, it is not desirable to use ordinary regression analysis for the data we are dealing with since the assumptions on which the classical linear regression model is based will be obviously violated.

Violation of the independent errors assumption has three important consequences for ordinary regression. First, statistical tests of the significance of the parameters and the confidence limits for the predicted values are not correct. Second, the estimates of the regression coefficients are not as efficient as they would be if the autocorrelation were taken into account. Third, since the ordinary regression residuals are not independent, they contain information that can be used to improve the prediction of future values. In that case, we need to introduce a dynamics on the errors in order to capture this effect. For instance, we can consider the model

$$\left\{\begin{array}{lcl} X_t & = & a_x + b_x X_{t-1} + \varepsilon_t^x, \\ \varepsilon_t^x & = & \rho_1^x \varepsilon_{t-1}^x + \cdots + \rho_p^x \varepsilon_{t-p_x}^x + u_t^x \\ u_t^x & \rightsquigarrow & \mathcal{WN}(0, \xi_x^2) \end{array}\right. \tag{D.6.17}$$

where the notation $u_t \rightsquigarrow \mathcal{WN}(0, \xi^2)$ indicates that u_t are uncorrelated with mean 0 and variance ξ^2. To estimate the parameters of the model we initially fit a high-order model with many autoregressive lags and then sequentially remove autoregressive parameters until all remaining autoregressive parameters have significant *t-tests*. To fit the model an *exact maximum likelihood method* is used. This method is based on the hypothesis that the white noises are normally distributed, which is in accordance with the Kolmogoroff test for normality [2].

[2] if tests for normality are *ex post* rejected the Yule-Walker estimation or the unconditional least squares can be used

Appendix E

Defining market equilibrirum and asset prices

The theory of general equilibrium started with Walras [1874-7] who considered demand and supply to explain the prices of economical goods, and was formalised by Arrow-Debreu [1954] and McKenzie [1959]. In parallel Arrow [1953] and then Debreu [1953] generalised the theory, which was static and deterministic, to the case of uncertain future by introducing contingent prices. Arrow [1953] proposed to create financial markets and was at the origin of the modern theory of financial markets equilibrium. Radner [1976] improved Arrow's model by considering more general assets, and introduced the concept of rational anticipation. In view of presenting some well known approches to value asset prices and define market equilibrium, we follow Dana and Jeanblanc-Picque [1994] and consider models in discrete time with one or two time periods with a finite number of states of the world.

E.1 Introducing the theory of general equilibrium

E.1.1 1 period, $(d+1)$ assets, k states of the world

We consider a market with one period of time, $(d+1)$ assets and k states of the world. We let S^i be the price at time 0 of the ith asset $(i = 0, 1, .., d)$ with value at time 1 and in the state of the world j being v_j^i. We let the portfolio $(\theta_0, \theta_1, .., \theta_d)$ have value $\sum_{i=0}^d \theta_i S^i$ at time 0, and value $\sum_{i=0}^d \theta_i v_j^i$ at time 1 in the jth state of the world. Hence, S is a column vector with element S^i, θ is a column vector with element θ_i, and V is the $(k \times (d+1))$ gain matrix with element $(v_j^i, 1 \leq j \leq k)$. In matricial notation $\theta.S = \sum_{i=0}^d \theta_i S^i$ is the scalar product of vectors θ and S, and $V\theta$ is a vector in $\mathbb{R}^k$ with element $(V\theta)_j = \sum_{i=0}^d \theta_i v_j^i$. Since a riskless asset has a value of 1 in all states of the world, we let $S = \frac{1}{1+r}$ be its value at time 0 where r is the risk-free rate. Given $\mathbb{R}_+^k$ the set of vectors in $\mathbb{R}^k$ with positive elements and $\mathbb{R}_{++}^k$ the set of vectors in $\mathbb{R}^k$ with strictly positive elements, we let Δ^{k-1} be the unit simplex of $\mathbb{R}^k$

$$\Delta^{k-1} = \{\lambda \in \mathbb{R}_+^k | \sum_{i=1}^k \lambda_i = 1\}$$

Further, given z and $z^{'}$ two vectors in $\mathbb{R}^k$, we let $z \geq z^{'}$ denote $z_i \geq z_i^{'}$ for all i. Following Ross [1976] [1978], we are going to define the notion of arbitrage opportunity.

Definition E.1.1 *There is an arbitrage opportunity if one of the following conditions is satisfied*

- *there exists a portfolio $\theta = (\theta_0, \theta_1, .., \theta_d)$ such that the initial value $\theta.S = \sum_{i=0}^d \theta_i S^i$ is strictly negative and the value at time 1 is positive in all the states of the world, that is, $\sum_{i=0}^d \theta_i v_j^i \geq 0$ for $j \in \{1, .., k\}$.*

- *there exists a portfolio* $\theta = (\theta_0, \theta_1, .., \theta_d)$ *such that the initial value* $\theta.S$ *is negative or null, and the value at time* 1 *is positive in all the states of the world and strictly positive in at least one state, that is,* $\sum_{i=0}^{d} \theta_i v_j^i \geq 0$ *for* $j \in \{1, .., k\}$ *and there exists* j_0 *such that* $\sum_{i=0}^{d} \theta_i v_{j_0}^i > 0$.

***Hypotheses* 1** *No-Arbitrage Opportunity (NAO)*
There is no-arbitrage opportunity.

Theorem E.1.1 *The hypothesis of NAO is equivalent to the existence of a series* $(\beta_j)_{j=1}^k$ *of strictly positive numbers called state prices, such that*

$$S^i = \sum_{j=1}^{k} v_j^i \beta_j \ , \ i \in \{0, .., d\} \tag{E.1.1}$$

Note, β is a vector of state prices with elements β_j corresponding to the price at time 0 of an asset with value being 1 at the time 1 in the state of the world j and 0 in the other states. For the riskless asset we have

$$v_j^0 = 1 \ , \ j \in \{1, .., k\}$$

such that using Equation (E.1.1) we get

$$S^0 = \frac{1}{1+r} = \sum_{j=1}^{k} \beta_j \tag{E.1.2}$$

Hence, setting $\pi_j = (1+r)\beta_j$ we get positive numbers such that $\sum_{j=1}^{k} \pi_j = 1$, and we can consider them as probabilities on the state of the world. As a result, the price at time 0 of the ith asset becomes

$$S^i = \frac{1}{1+r} \sum_{j=1}^{k} \pi_j v_j^i$$

where the price S^i of the ith asset is the expected value of its price at time 1 discounted by the risk-free rate. Building the portfolio $\theta = (\theta_0, \theta_1, .., \theta_d)$, we get

$$(1+r) \sum_{i=0}^{d} \theta_i S^i = \sum_{j=1}^{k} \pi_j \sum_{i=0}^{d} \theta_i v_i^i$$

where π is a risk-neutral probability. By definition, the return of the ith asset in the state of the world j is $\frac{v_j^i}{S^i}$, and its expected return under the probability π is

$$\sum_{j=1}^{k} \pi_j \frac{v_j^i}{S^i} = (1+r)$$

which is the return of the riskless asset.

Property E.1.1 *Given the hypothesis of NAO and a riskless asset* 0*, there exists a probability* π *on the states of the world such that the price of the ith asset at time* 0 *equal the expected value of its price at time* 1 *discounted by the risk-free rate*

$$S^i = \frac{1}{1+r} \sum_{j=1}^{k} \pi_j v_j^i \tag{E.1.3}$$

E.1.2 Complete market

Definition E.1.2 *A market is complete if, for all vector w of $\mathbb{R}^k$, we can find a portfolio θ such that $V\theta = w$, that is, θ such that*

$$(V\theta)_j = \sum_{i=0}^{d} \theta_i v_j^i = w_j \ , \ j \in \{1, .., k\}$$

***Proposition* 13** *A market is complete if and only if the matrix V has rank k.*

In a complete market, for all $j \in \{1, .., k\}$, there exists a portfolio θ_j such that $V\theta_j = (\delta_{1,j}, .., \delta_{k,j})^\top$ with $\delta_{i,j} = 0$ if $i \neq j$ and $\delta_{j,j} = 1$ and we get an Arrow-Debreu asset. Further, if there is no-arbitrage, the initial value of the portfolio is $S.\theta_j = \beta^\top V\theta_j = \beta_j$. To conclude, if there exists β such that $V^\top\beta = S$, then β is unique. If there exists a probability π satisfying $V^\top\pi = (1 + r)S$ then it is unique and we call it the risk-neutral probability.

We can then use the NAO to value assets in a complete market. Given z a vector in $\mathbb{R}^k$ and assuming NAO, if there exists a portfolio $\theta = (\theta_0, \theta_1, .., \theta_d)$ taking the value z at time 1

$$\sum_{i=0}^{d} \theta_i v_j^i = z_j$$

then z is replicable. The value of the portfolio at time 0 is $z_0 = \sum_{i-0}^{d} \theta_i^{S^i}$, and it does not depend on the chosen solution. Such a portfolio is called a hedging portfolio.

***Proposition* 14** *In a complete market with no-arbitrage, the value of $z \in \mathbb{R}^k$ is given by*

$$\frac{1}{1+r} \sum_{j=1}^{k} \pi_j z_j = \sum_{j=1}^{k} \beta_j z_j$$

Note, the value of z is linear in z. Hence, the price at time 0 of the replicating portfolio $(z_j, j = 1, .., k)$ is the expected value under π of its discounted value at time 1. For instance, for a call option on the ith asset, we have $z_j = (v_j^i - K)^+$ and the arbitrage price becomes

$$\frac{1}{1+r} \sum_{j=1}^{k} \pi_j (v_j^i - K)^+$$

E.1.3 Optimisation with consumption

We consider a simple economy with one consumer good taken as numeraire and a single economical agent. This agent has a known wealth W^0 at time 0 and a wealth W^j at time 1 in the state of the world j. The objective being to maximise wealth over time. In view of modifying his future income, he can buy at time 0 a portfolio of assets, but he can not have debts. He can further consume c^0 at time 0 and get the amount of consumption c^j at time 1 in the jth state of the world. Given the portfolio θ, the agent must satisfies the constraints

1. $W^0 \geq c^0 + \sum_{i=0}^{d} \theta_i S^i$
2. $W^j \geq c^j - \sum_{i=0}^{d} \theta_i v_j^i$, $j \in \{1, .., k\}$ income from portfolio at time 1

The second constraint states that consumption at time 1 comes from wealth together with the value of the portfolio. The set of consumption compatible with the agent's income is

$$B(S) = \{c \in \mathbb{R}_+^{k+1}; \exists\, \theta \in \mathbb{R}^{d+1} \text{ satisfying the above constraints } \}$$

For details see text book by Demange et al. [1992]. The agent has preferences on $\mathbb{R}_+^{k+1}$, or complete pre-order notted $\succeq$.

More generally, given the finite dimensional space vector C made with $(c^j, j = 1, .., k)$, we provide C with the scalar product

$$< c, c^\top >= E[cc^\top]$$

with associated norm $||c||_2$. In particular we get $< 1, c >= E[c]$ which we also denote $< c >$. Agents or investors have some preferences over the elements of C characterised by utility functions $U_i : C \to \mathbb{R}$ for $(i = 1, .., m)$ with the property of aversion for variance or risk. That is, for all pair $(c, c^\top) \in C^2$ the inequality $Var(c) < Var(c^\top)$ implies $U_i(c) > U_i(c^\top)$ which is equivalent to $c \succeq c^\top$. We assume that the utility functions U_i are strictly increasing with respect to each variables, strictly concave and differentiable, and that agents or investors maximise their utilities under budget constraints. In general, we make the assumption that the utility functions only depends on the first two moments of the random variables, that is, it can be written $\tilde{U}_i(E[c], Var(c))$ where $\tilde{U}_i$ is increasing with respect to the first coordinate and decreasing with respect to the second coordinate. Put another way, there is an equilibrium if the utility functions are concave functions of the mean and variance of the variables, such that they are increasing with respect to the first coordinate and decreasing with respect to the second coordinate.

Back to our settings, our objective being to maximise wealth over time, we say that $c^* \in B(S)$ is an optimal consumption if

$$u(c^*) = \max\{u(c); c \in B(S)\} \tag{E.1.4}$$

***Proposition* 15** *There is an optimal solution if and only if S satisfies the hypothesis of NAO. The optimal solution is strictly positive.*

Since c^* is strictly positive, using the Lagrange's multiplier, one can deduce that one sufficient and necessary condition for c^* to be optimal is that $\theta^* \in \mathbb{R}^{d+1}$ and $\lambda^* \in \mathbb{R}_+^{k+1}$ such that some constraints are satisfied. Given

$$\beta_j = \frac{\lambda_j^*}{\lambda_0^*} = \frac{\frac{\partial u}{\partial c^j}}{\frac{\partial u}{\partial c^0}}(c^*) \tag{E.1.5}$$

the β_j are strictly positive and we get

$$S^i = \sum_{j=1}^k \beta_j v_j^i$$

and one get a formula to value the price of assets. Simplification to the optimisation problem can be made in a complete market with no-arbitrage. In such a market there exists a unique β such that $S = V^\top \beta$, and one get the budget's constraint (agent can not have debts)

$$c^0 + \sum_{j=1}^k \beta_j c^j \leq W^0 + \sum_{j=1}^k \beta_j W^j \tag{E.1.6}$$

where an agent buys consumption goods c^j at price β_j. Similarly, if the market is complete, we get

$$B(S) = \{c \in \mathbb{R}_+^{k+1} | \text{ satisfying the above constraint } \}$$

and there exists θ such that

$$c^j - \sum_{i=0}^{d} \theta_i v_j^i - W^j = 0 \text{ for all } j \in \{1, .., k\}$$

Hence the problem of optimisation has now the single budget's constraint given in Equation (E.1.6). If there is a riskless asset, we get the risk-neutral probabilities $\beta_j = \frac{\pi_j}{(1+r)}$ so that letting asset 0 be the riskless asset, Equation (E.1.6) rewrite

$$c^0 + \sum_{j=1}^{k} \pi_j \frac{c^j}{(1+r)} \leq W^0 + \sum_{j=1}^{k} \pi_j \frac{W^j}{(1+r)}$$

Note, we can use this equation in continuous time to transform a trajectorial constraint into a constraint on average.

E.2 An introduction to the model of Von Neumann Morgenstern

We are now going to specialise the utility function in the optimisation with consumption. For details see text book by Kreps [1990].

E.2.1 Part I

We consider one period of time with a single good of consumption. We let $\mathcal{P}$ be the set of probabilities on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, and assume a finite number of states of the world such that the state j happens with the probability μ_j. We let the consumption C at time 1 be a random variable taking the values c^j and assume that the law μ_C of C with $\mu_C = \sum_{j=1}^{k} \mu_j \delta_{c^j}$ is an element of $\mathcal{P}$. Further, assuming the agent has a complete pre-order $\succeq$ on $\mathcal{P}$, we say that $u : \mathcal{P} \to \mathbb{R}$ is a utility function with pre-order of preference if $u(\mu) \geq u(\mu^{'})$ is equivalent to $\mu \succeq \mu^{'}$. To get a Von-Neumann Morgenstern (VNM) utility (see Von Neumann et al. [1944]), there must exist $v : \mathbb{R}_+ \to \mathbb{R}$ such that

$$u(\mu) = \int_0^{\infty} v(x) d\mu(x)$$

In the special case where μ_C is the discrete sum defined above, the VNM utility simplifies to

$$u(\mu) = \sum_{j=1}^{k} \mu_j v(c^j)$$

Our objective of maximising wealth over time is realised via the optimal consumption in Equation (E.1.4). Note, μ_j is a probability mass function (see Definition (B.2.4)) such that $\sum_{j=1}^{k} \mu_j = 1$ (see Lemma (B.2.1)), and from Definition (B.2.5) we get the first two moments

$$\begin{aligned} E[C] &= \sum_{j=1}^{k} \mu_j c^j \, , Var(C) = \sum_{j=1}^{k} \mu_j (c^j - E[C])^2 \\ E[v(C)] &= \sum_{j=1}^{k} \mu_j v(c^j) \, , Var(v(C)) = \sum_{j=1}^{k} \mu_j (v(c^j) - E[v(C)])^2 \end{aligned} \tag{E.2.7}$$

Hence, the criterion becomes that of maximising the expected value of the utility of consumption where

$$u(\mu) = E[v(C)] =< v(C) >$$

Given Jensen's inequality, when v is concave, the agent is risk-averse since

$$v(E[C]) \geq E[v(C)]$$

and the investor prefers the certain future consumption $E[C]$ to the consumption c^j with probability μ_j for all j. The investor is risk-neutral if v is affine, and we get $v(E[C]) = E[v(C)]$. Therefore, we can let λ be the market price of risk linked to the random consumption C and defined by

$$v(E[C] - \lambda) = E[v(C)] \tag{E.2.8}$$

such that $E[C] - \lambda$ is the certain equivalent to C. Assuming v to be C^2 and $(c^j - E[C])$ to be small, we use the Taylor expansion around $E[C]$ to get

$$v(c^j) \approx v(E[C]) + (c^j - E[C])v^{'}(E[C]) + \frac{1}{2}(c^j - E[C])^2 v^{''}(E[C])$$

Taking expectation (see Equation (E.2.7)) we get

$$E[v(C)] = \sum_{j=1}^{k} \mu_j v(c^j) \approx v(E[C]) + \frac{1}{2} Var(C) v^{''}(E[C])$$

Again, doing a Taylor expansion to $v(E[C] - \lambda)$ in Equation (E.2.8), we get

$$v(E[C] - \lambda) \approx v(E[C]) - \lambda v^{'}(E[C])$$

so that putting terms together, the market price of risk becomes

$$\lambda \approx -\frac{1}{2} \frac{v^{''}(E[C])}{v^{'}(E[C])} Var(C) = \frac{1}{2} \alpha Var(C)$$

where the coefficient $\alpha = -\frac{v^{''}(E[C])}{v^{'}(E[C])}$ is called the absolute aversion index for the risk in $E[C]$. It is also called the coefficient of absolute risk aversion (see Pratt [1964] and Arrow [1971]).

E.2.2 Part II

Following from the previous Appendix, we now consider two periods of time and assume that at time 1 the state of the world happens with probability $\mu = (\mu_j)_{j=1}^k$. We also assume complete market and the existence of a riskless asset. We further assume that the utility functions are additively separable with respect to time

$$u(c^0, C) = v^0(c^0) + \frac{1}{1+r} E[v(C)] = v^0(c^0) + \frac{1}{1+r} \sum_{j=1}^{k} \mu_j v(c^j)$$

where v^0 and v are strictly concave functions, strictly increasing functions in C^2 satisfying $\frac{\partial v^0}{\partial c}(c) \to \infty$ and $\frac{\partial v}{\partial c}(c) \to \infty$ as $c \to 0$. Assuming strictly positive optimal consumption, from Equation (E.1.5) we get the state prices

$$\beta_j = \frac{\mu_j}{(1+r)} \frac{v^{'}(c^{j*})}{v^{0'}(c^{0*})}$$

and given the riskless asset, we have

$$1 = \sum_{j=1}^{k} \beta_j (1+r)$$

so that

$$v^{0'}(c^{0*}) = E[v^{'}(C^*)]$$

We can then express the risk-neutral probability as

$$\pi_j = (1+r)\beta_j = \mu_j \frac{v^{'}(c^{j*})}{E[v^{'}(C^*)]}$$

As in Appendix (E.2.1), we use Taylor expansion around $E[C^*]$, to get

$$v^{'}(c^{j*}) \approx v^{'}(E[C^*]) + (c^{j*} - E[C^*])v^{''}(E[C^*])$$

and taking the expectation (see Equation (E.2.7)) we get

$$E[v^{'}(C^*)] \approx v^{'}(E[C^*])$$

so that

$$\frac{\pi_j}{\mu_j} \approx 1 + \frac{(c^{j*} - E[C^*])v^{''}(E[C^*])}{v^{'}(E[C^*])} = 1 + \alpha(E[C^*] - c^{j*})$$

The probability gets larger as the absolute aversion index α gets larger or the spread between the mean consumption and the consumption in state j gets larger. From the price formula in Equation (E.1.3), the ith asset at time 0 becomes

$$S^i = \frac{1}{1+r}\sum_{j=1}^{k} \mu_j \frac{v^{'}(c^{j*})}{E[v^{'}(C^*)]} v_j^i$$

***Remark* E.2.1** *For the investor to be risk-neutral* $(\lambda = 0)$ *then* $v^{''}(.) = 0$ *and* $v^{'}(.) = cst$ *so that he would pay* $\frac{1}{1+r}\mu_j$ *at time* 0 *to get* \$1 *at time* 1 *in the state of the world* j*. If he is risk averse, he would pay* $\frac{1}{1+r}\mu_j \frac{v^{'}(c^{j*})}{E[v^{'}(C^*)]}$ *at time* 0 *to get* \$1 *at time* 1 *in the state of the world* j*.*

E.3 Simple equilibrium model

E.3.1 m agents, $(d+1)$ assets

We assume complete market and the existence of a riskless asset. We consider an economy with a single good of consumption and m economical agents with $(d+1)$ assets. The agent h has e_h^0 unit of wealth at time 0 and e_h^j unit of wealth at time 1 in the jth state of the world. He can buy at time 0 a portfolio $\theta_h = (\theta_h^0, .., \theta_h^d)$ without incurring debts. Given a vector of price S, the set of consumption compatible with the agent's income is

$$B_h(S) = \{c \in \mathbb{R}_+^{k+1} | \exists\, \theta \in \mathbb{R}^{d+1}, e_h^0 \geq c^0 + \theta.S; e_h^j \geq c^j - (V\theta)_j, j \in \{1, .., k\}\}$$

The agent h has some preferences represented by the VNM utility function

$$u(c^0, C) = v_h^0(c^0) + \frac{1}{(1+r)}\sum_{j=1}^{k} \mu_j v_h(c^j)$$

Definition E.3.1 *We say that* $\{\overline{S}, (\overline{c}_h, \overline{\theta}_h); h = 1, .., m\}$ *is a market equilibrium if, given* $\overline{S}$

1. $\overline{c}_h$ *maximise* $u_h(c_h^0, C_h)$ *under the constraint* $c_h(c_h^0, C_h) \in B_h(\overline{S})$

2. $\sum_{h=1}^m \bar{c}_h^j = \sum_{h=1}^m e_h^j = e^j$, $j \in \{1,..,k\}$
3. $\sum_{h=1}^m \bar{\theta}_h = 0$

Assuming that there exists an equilibrium, and given the results in Appendix (E.2.2), the price at time 0 for all h is

$$S^i = \frac{1}{1+r}\sum_{j=1}^k \mu_j \frac{v_h^{'}(\bar{c}_h^j)}{v_h^{0'}(\bar{c}_h^0)} v_j^i = \frac{1}{1+r}\sum_{j=1}^k \mu_j \frac{v_h^{'}(\bar{c}_h^j)}{E[v_h^{'}(\overline{C}_h)]} v_j^i$$

Since in a complete market the equation $S = V^T\beta$ has a unique solution, the quantity $\frac{v_h^{'}(\bar{c}_h^j)}{v_h^{0'}(\bar{c}_h^0)}$ is independent from h, and one can consider a dummy agent with VNM utility function

$$u(c^0, C) = v^0(c^0) + \frac{1}{(1+r)}\sum_{j=1}^k \mu_j v(c^j)$$

where

$$\begin{aligned} v^0(c) &= \max\{\sum_{h=1}^m \frac{v_h^0(c_h)}{v_h^{0'}(\bar{c}_h^0)}; \sum_{h=1}^m c_h = c\} \\ v(c) &= \max\{\sum_{h=1}^m \frac{v_h(c_h)}{v_h^{0'}(\bar{c}_h^0)}; \sum_{h=1}^m c_h = c\} \end{aligned}$$

One can check that we get

$$u(e^0, e) = \sum_{h=1}^m \frac{v_h^0(\bar{c}_h)}{v_h^{0'}(\bar{c}_h^0)} + \frac{1}{(1+r)}\sum_{j=1}^k\sum_{h=1}^m \frac{v_h(c_h^j)}{v_h^{0'}(\bar{c}_h^0)}\mu_j$$

where e is a random variable taking the value e^j with probability μ_j. One can show that v^0 and v are differentiable, and that

$$v^{'}(e^j) = \frac{v_h^{'}(\bar{c}_h^j)}{v_h^{0'}(\bar{c}_h^0)} = \frac{v_h^{'}(\bar{c}_h^j)}{E[v_h^{'}(\overline{C}_h)]} \text{ for all } h = 1,..,m \text{ and } j = 1,..,k$$

and

$$v^{0'}(e^0) = E[v^{'}(e)] = 1$$

As a result, the price becomes

$$S^i = \frac{1}{1+r}\sum_{j=1}^k \mu_j \frac{v^{'}(e^j)}{E[v^{'}(e)]} v_j^i = \frac{1}{1+r}\sum_{j=1}^k \mu_j v^{'}(e^j) v_j^i \tag{E.3.9}$$

E.3.2 The consumption based asset pricing model

In the special case where $v^{'}$ is linearly decreasing

$$v^{'}(c) = -ac + b\,,\, a > 0$$

corresponding to the quadratic utility function in Equation (A.7.6) or the third concave funvtion v described in Appendix (A.1.2), we get

$$v^{'}(e^j) = E[v^{'}(e)] + a\big(E[e] - e^j\big) = 1 + a\big(E[e] - e^j\big)$$

In that setting, the equilibrium price in Equation (E.3.9) becomes

$$\begin{aligned} S^i &= \frac{1}{(1+r)}\sum_{j=1}^{k}\mu_j\big(1 - a(e^j - E[e])\big)v_j^i \\ &= \frac{1}{(1+r)}\big\{E[V^i] - aCov(e, V^i)\big\} \end{aligned} \tag{E.3.10}$$

where V^i is a random variable taking values v_j^i with probability μ_j with the first two moments being

$$E[V^i] = \sum_{j=1}^{k}\mu_j v_j^i \, , Var(V^i) = \sum_{j=1}^{k}\mu_j(v_j^i - E[V^i])^2$$

This equilibrium price correspond to the Consumption based Capital Asset Pricing Model (CCAPM) (see Dana [1993]). Let $R^i = \frac{V^i}{S^i}$ be the return of the ith asset, and let the vector M be the market portfolio such that $e = VM$. Dividing Equation (E.3.10) by S^i, we get

$$E[R^i] - (1+r) = aCov(e, R^i)$$

and setting $R_M = \frac{e}{S.M}$ we get

$$E[R_M] - (1+r) = aCov(e, R_M) = (aS.M)Var(R_M)$$

since $(S.M)R_M = e$. Combining the two equations, we get

$$E[R^i] - (1+r) = \frac{Cov(R^i, R_M)}{Var(R_M)}\big\{E[R_M] - (1+r)\big\} \tag{E.3.11}$$

which is related to the beta formula in the Capital Asset Pricing Model (CAPM) where

$$\beta_i = \frac{Cov(R^i, R_M)}{Var(R_M)}$$

In that model, the risk premium of the ith asset $E[R^i] - (1+r)$ is a linear function of its covariance with the return of the market portfolio called the Security Market Line (SML) (see Huang et al. [1988]). More generally, relaxing the constraint on $v^{'}$, but assuming that e^j is close to $E[e]$, one can approximate the CCAPM formula since

$$\frac{v^{'}(e^j)}{E[v^{'}(e)]} \approx 1 + \alpha(E[e] - e^j)$$

so that the equilibrium price in Equation (E.3.10) becomes

$$S^i \approx \frac{1}{(1+r)}E[V^i] - \frac{1}{(1+r)}\alpha Cov(e, V^i)$$

E.4 The n-dates model

We consider a model with one state of the world at time 0, and k_n states of the world at time n where $M(n)$ is the set of states of the world at time n. Given $j \in M(n)$, we let $\Delta^n(j)$ be the states of the world at time $(n+1)$ coming from j. Further, we let $S_n^i(j)$ be the price of the ith asset at time n in the jth state of the world, and we let S_n be a vector with elements S_n^i, $i \in \{0, .., d\}$. Letting the asset 0 be the riskless element, we get

$$\forall n, \forall j \in M(n), \forall l \in \Delta^n(j) \ , \ S_{n+1}(l) = (1 + r(j))S_n^0(j)$$

At last, θ_n is the vector $(\theta_n^0, .., \theta_n^d)$ being the quantity of assets hold at time n and the value of the portfolio $\theta_n.S_n$ depends on the state of the world where we stand.

Definition E.4.1 *There is NAO if, for all n and for all $j \in M(n)$, there is NAO between j and $\Delta^n(j)$.*

As a result, for all n, for all $j \in M(n)$, there exists a family of reals $\pi_n(j, l)$, $l \in \Delta^n(j)$ such that

1. $\pi_n(j, l) \geq 0$
2. $\sum_{l \in \Delta^n(j)} \pi_n(j, l) = 1$
3. $S_n^i(j) = \frac{1}{1+r(j)} \sum_{l \in \Delta^n(j)} \pi_n(j, l) S_{n+1}^i(l)$

corresponding to some probabilities. It is the probability of being at time $(n + 1)$ in the lth state of the world, when at time n we were in the jth state. Put another way, they are transition probabilities between time periods n and $(n + 1)$, and from one state of the world to another. We can normalise the third equation with the riskless asset, getting

$$\frac{S_n^i(j)}{S_n^0(j)} = \sum_{l \in \Delta^n(j)} \pi_n(j, l) \frac{S_{n+1}^i(l)}{S_{n+1}^0(l)}$$

and calling $P_n^i(.)$ the price at time n of the ith asset discounted by the riskless asset

$$P_n^i(j) = \frac{S_n^i(j)}{S_n^0(j)}$$

the equation rewrites

$$P_n^i(j) = \sum_{l \in \Delta^n(j)} \pi_n(j, l) P_{n+1}^i(l)$$

Hence, the price $P_n^i(j)$ is the expected value of the price $P_{n+1}^i(.)$ under the probability $\pi_n(j, .)$ noted

$$P_n^i(j) = E^{\pi_n(j,.)}[P_{n+1}^i(.)] \tag{E.4.12}$$

As a result, $P_\cdot^i$ is a martingale for the probability π.

***Proposition* 16** *Assuming NAO, there exists a probability π on the tree of events such that the price of assets discounted by the riskless asset are martingales.*

Note, if there exists a probability such that discounted prices are martingales, there is no-arbitrage. In a complete market, this probability is unique (see Harrison et al. [1981]).

E.5 Discrete option valuation

We consider a simple financial market with two dates, the date 0 and the date 1, and two states of the world at the date 1. The market contains one risky asset and one riskless asset (for details see Cox et al. [1985b]). The value of the risky asset at date 0 is S, and it becomes S_u or S_d if its price goes up or down. We consider a call option with strike K such that $S_d \leq K \leq S_u$. We can build a portfolio with weights (α, β) where α is invested in the riskless asset and β is invested in the risky asset. At date 0 the value of the portfolio is $\alpha + \beta S$, and at date 1 the value becomes $\alpha(1+r) + \beta S_u$ if the asset price goes up and $\alpha(1+r) + \beta S_d$ if it is down. The portfolio duplicates the option when its value at date 1 is equal to the gain realised by the option whatever the state of the world

$$\begin{aligned}\alpha(1+r) + \beta S_u &= S_u - K \\ \alpha(1+r) + \beta S_d &= 0\end{aligned}$$

Solving this system we obtain the weights (α^*, β^*)

$$\alpha^* = -\frac{S_d(S_u - K)}{(S_u - S_d)(1+r)}, \beta^* = \frac{(S_u - K)}{(S_u - S_d)}$$

The option price being the value of this portfolio at time 0

$$q = \alpha^* + \beta^* S = \frac{(S_u - K)}{(S_u - S_d)}(S - \frac{S_d}{1+r})$$

If we relax the constraint on the strike, we get

$$\begin{aligned}\alpha^*(1+r) + \beta^* S_u &= (S_u - K)^+ = C_u \\ \alpha^*(1+r) + \beta^* S_d &= (S_d - K)^+ = C_d\end{aligned}$$

and the price becomes

$$q = \alpha^* + \beta^* S = \frac{1}{1+r}(\pi C_u + (1-\pi)C_d)$$

where

$$\pi = \frac{1}{(S_u - S_d)}((1+r)S - S_b)$$

If $S_d \leq (1+r)S \leq S_u$, then $\pi \in [0,1]$ and the pricing equation above can be interpreted in terms of risk-neutral probabilities where

$$(1+r)S = \pi S_u + (1-\pi)S_d$$

Property E.5.1 *The option price is the expectation of the discounted gain under the risk-neutral probability.*

In that setting, if the asset can take more than two values at date 1, we can no-longer replicate the option. Further, there exists several risk-neutral probability measures. We therefore get a range of prices. Let's assume that the asset price at date 1 is the random variable S_1 taking values in $[S_d, S_u]$, and that $S_d \leq (1+r)S \leq S_u$. Let $\mathcal{P}$ be the set of risk-neutral probabilities, that is, the probabilities P such that

$$E^P[\frac{S_1}{(1+r)}] = S$$

Property E.5.2 *For all convex function h (for example $h(x) = (x-K)^+$), we get*

$$\sup_{P\in\mathcal{P}} E^P[\frac{h(S_1)}{(1+r)}] = \frac{h(S_u)}{(1+r)}\frac{S(1+r)-S_d}{(S_u-S_d)} + \frac{h(S_d)}{(1+r)}\frac{S_u-S(1+r)}{(S_u-S_d)}$$

The supremum is reached when S_1 only takes the values S_u and S_d. If h is of class C^1, we get

$$\inf_{P\in\mathcal{P}} E^P[\frac{h(S_1)}{(1+r)}] = \frac{h((1+r)S)}{(1+r)}$$

The inf is reached when S_1 is equal to $(1+r)S$.

We define the price to sell an option as the smallest price allowing for the seller to hedge himself. It is the smallest wealth to invest into the portfolio (α,β) such that its final value gets bigger than the option value $h(S_1)$. Therefore, the selling price is

$$\inf_{(\alpha,\beta)\in\mathcal{A}}(\alpha+\beta S)$$

where $\mathcal{A} = \{(\alpha,\beta)|\alpha(1+r)+\beta x \geq h(x)\ ,\forall x\in[S_d,S_u]\}$. We get

$$\inf_{(\alpha,\beta)\in\mathcal{A}}(\alpha+\beta S) = \sup_{P\in\mathcal{P}} E^P[\frac{h(S_1)}{(1+r)}]$$

One say that the two problems

$$\sup_{P\in\mathcal{P}} E^P[\frac{h(S_1)}{(1+r)}] \text{ and } \inf_{(\alpha,\beta)\in\mathcal{A}}(\alpha+\beta S)$$

are dual problems. The buying price of the option is

$$\sup_{(\alpha,\beta)\in\mathcal{C}}(\alpha+\beta S)$$

where $\mathcal{C} = \{(\alpha,\beta)|\alpha(1+r)+\beta x \leq h(x)\ ,\forall x\in[S_d,S_u]\}$, and we get

$$\sup_{(\alpha,\beta)\in\mathcal{C}}(\alpha+\beta S) = \inf_{P\in\mathcal{P}} E^P[\frac{h(S_1)}{(1+r)}]$$

E.6 Valuation in financial markets

E.6.1 Pricing securities

Following Dybvig et al. [2003], we extend the results on optimisation with consumption, presented in Appendix (E.1.3) in the case of the simple models, to financial markets price securities with payoffs extending out in time. In a discrete time world with asset payoffs $h(X)$ at time T, contingent on the realisation of a state of nature $X\in\Omega$, absence of arbitrage opportunity (AAO) (from the Fundamental Theorem of asset pricing (see Dybvig et al.)) implies the existence of positive state space prices, that is, the Arrow-Debreu contingent claims prices $p(X)$ paying \$1 in state X and nothing in any other states (see Theorem (E.1.1)). If the market is complete, then these state prices are unique. The current value C_h of an asset paying $h(X)$ in one period is given by

$$C_h = \int h(X)dP(X)$$

where $P(X)$ is a price distribution function ($dP(X) = p(X)dX$). Letting $r(X^0)$ be the riskless rate as a function of the current state X^0, such that $\int p(X)dX = e^{-r(X^0)T}$ (see Equation (E.1.2)), we can rewrite the price as

$$\begin{aligned} C_h &= \int h(X)dP(X) = (\int dP(X)) \int h(X) \frac{dP(X)}{\int dP(X)} = e^{-r(X^0)T} \int h(X)d\pi^*(X) \\ &= e^{-r(X^0)T} E^*[h(x)] = E[h(X)\xi(X)] \end{aligned}$$

where the asterisk denotes the expectation in the martingale measure and where the pricing kernel, that is, the state-price density $\xi(X)$ is the Radon-Nikodym derivative of $P(X)$ with respect to the natural (historical) measure denoted $F(X)$. With continuous distribution, we get $\xi(X) = \frac{p(X)}{f(X)}$ where $f(X)$ is the natural probability, that is, the relevant subjective probability distribution, and the risk-neutral probabilities are given by

$$\pi^*(X) = \frac{p(X)}{\int p(X)dX} = e^{r(X^0)T} p(X)$$

(see Appendix (E.1.1) for details). We let X_i denote the current state, and X_j be a state one period ahead, and we assume that it fully describes the state of nature so that the stock price can be written $S(X_i)$. From the forward equation for the martingale probabilities

$$Q(X_i, X_j, T) = \int_X Q(X_i, X, t)Q(dX, X_j, T-t)$$

where $Q(X_i, X_j, T)$ is the forward martingale probability transition function for going from state X_i to state X_j in T periods, and where the integration is over the intermediate state X at time t. This is a very general framework allowing for many interpretations. To avoid dealing with interest rates, one can use state prices rather than martingales, defined as

$$P(X_i, X_j, t, T) = e^{-r(X_i)(T-t)} Q(X_i, X_j, T-t)$$

and assuming a time homogeneous process where calendar time is irrelevant, for the transition from any time t to $t+1$, we have

$$P(X_i, X_j) = e^{-r(X_i)} Q(X_i, X_j)$$

We let f be the natural (time homogeneous) transition density, and define the kernel as the price per unit of probability in continuous state spaces

$$\xi(X_i, X_j) = \frac{p(X_i, X_j)}{f(X_i, X_j)} \tag{E.6.13}$$

An equivalent statement of no arbitrage is that a positive kernel exists. As an example, Dybvig et al. [2003] considered an intertemporal model with a representative agent having additively time separable preferences and a constant discount factor δ. It correspond to the Von Neumann Morgenstern utility function described in Appendix (E.2). Letting $c(X)$ be the consumption at time t as a function of the state, over any two periods the agent want to maximise

$$\max_{c(X_i), \{c(X)\}_{X \in \Omega}} \left\{ u(c(X_i)) + \delta \int u(c(X)) f(X_i, X) dX \right\}$$

subject to the constraint

$$c(X_i) + \int c(X) p(X_i, X) dX = w$$

where w is the wealth of the agent. Note, the simplified version of this optimisation with consumption was discussed in Appendix (E.1.3). One can use the Lagrange's multiplier to solve for the optimum and obtain the Arrow-Debreu

price given in Equation (E.1.5). Following the same approach, considering the first order condition for the optimum, one can interpret the kernel as

$$\xi(X_i, X_j) = \frac{p(X_i, X_j)}{f(X_i, X_j)} = \frac{\delta u^{'}(c(X_j))}{u^{'}(c(X_i))} \tag{E.6.14}$$

This equation is the equilibrium solution for an economy with complete market, in which consumption is exogenous and prices are defined by the first order condition for the optimum. In a multiperiod mode with complete markets and state independent, intertemporally additive separable utility, there is a unique representative agent utility function satisfying the above optimum condition. The kernel is the agent's marginal rate of substitution (MRS) as a function of aggregate consumption. The marginal rate of substitution (MRS) across time for an agent with an intertemporally additively separable utility function is a function of the final state and depend only on the current state as a normalisation. Note the path independence, since the pricing kernel only depends on the MRS between future and current consumption, is a key element in the recovery theorem.

Definition E.6.1 *A kernel is transition independent if there is a positive function of the states g, and a positive constant δ such that for any transition from X_i to X_j, the kernel has the form*

$$\xi(X_i, X_j) = \delta \frac{g(X_j)}{g(X_i)}$$

Note, the intertemporally additive utility function is one among others generating a transition independent kernel. Using transition independence we can rewrite Equation (E.6.13) as

$$p(X_i, X_j) = \xi(X_i, X_j) f(X_i, X_j) = \delta \frac{g(X_j)}{g(X_i)} f(X_i, X_j) \tag{E.6.15}$$

where $g(X) = u^{'}(c(X))$ in the representative model.

E.6.2 Introducing the recovery theorem

In a recent article, Ross [2013] assumed that the state-price transition function $p(X_i, X_j)$ is observable and solved the system in Equation (E.6.15) to recover the three unknowns, the natural probability transition function $f(X_i, X_j)$, the pricing kernel $\xi(X_i, X_j)$, and the discount rate δ. Note, the notion of transition independence was necessary to separately determine the kernel and the natural probability distribution. This is because in the Equation (E.6.13) there are more unknowns than equations. Other approaches used the historical distribution of returns to estimate the unknown kernel and thereby link the historical estimate of the natural distribution to the risk-neutral distribution. Alternatively, one can assume a functional form for the kernel. Ross [2013] showed that the equilibrium system in Equation (E.6.15) could be solved without the need of historical data or any further assumptions than a transition independent kernel.

From the definition of the kernel in Equation (E.6.14) Ross induced some properties to the market and natural densities. Since ξ is decreasing with respect to $c(X_j)$, fixing X_i, since both densities integrate to one and since ξ exceeds δ for $c(X_j) < c(X_i)$, then defining v by $\delta u^{'}(v) = u^{'}(c(X_i))$, it follows that $p > f$ for $c < v$ and $p < f$ for $c > v$. This is the single crossing property and verifies that f stochastically dominates p. In a single period model, terminal wealth and consumption are the same. Hence the following theorem.

Theorem E.6.1 *The risk-neutral density for consumption and the natural density for consumption have the single crossing property, and the natural density stochastically dominates the risk-neutral density. Equivalently, in a one period world, the natural density stochastically dominates the risk-neutral density.*

The following proof of the existence of a risk premium is interesting for our understanding of market's return. Since in a one period world, consumption coincides with the value of the market, from stochastic dominance at any future date T, the return in the risk-neutral measure satisfies

$$R^* \sim R - Z + \epsilon$$

where R is the natural return, Z is strictly non-negative, and ϵ is mean zero conditional on $R - Z$. Taking expectation, we get

$$E[R] = r + E[Z] > r$$

We now introduce the recovery theorem on discrete state spaces (bounded state space).

Theorem E.6.2 *Recovery Theorem*
If there is no arbitrage, if the pricing matrix is irreducible, and if it is generated by a transition independent kernel, then there exists a unique (positive) solution to the problem of finding the natural probability transition matrix F, the discount rate δ, and the pricing kernel ξ. For any given set of state prices, there is a unique compatible natural measure and a unique pricing kernel.

Hence, under some assumptions, using the recovery theorem, one can disentangled the kernel from the natural probabilities.

E.6.3 Using implied volatilities

Implied volatilities are a function of the risk-neutral probabilities, the product of the natural probabilities and the pricing kernel (risk aversion and time discounting), and as such, they embody all of the information needed to determine state prices. From the volatility surface and the formula for the value of a call option, one can derive the state-price distribution $p(S,T)$ at any tenor T

$$C(K,T) = \int_0^\infty (S-K)^+ p(S,T) dS$$

where $C(K,T)$ is the price of the call option. Differentiating twice with respect to the strike, we obtain the Breeden and Litzenberger result (see Breeden et al. [1978])

$$p(S,T) = \partial_{KK} C(K,T)$$

To apply the recovery theorem, we need the $m \times m$ state-price transition matrix $P = [p(i,j)]$, but it is not directly observed in the market. One can estimate it from the state price distributions at different tenors. Assuming that we are currently in state c and observe prices of options across strikes and tenors, then we can extract the state prices at each future date T

$$p^T(c) =< p(1,T), .., p(m,T) >$$

The vector of one period ahead state prices with $T = 1$ corresponds to the row of the state price transition matrix P. To solve for the remaining elements of P we can apply the forward equation recursively to create the sequence of $(m-1)$ equations

$$p^{t+1} = p^t P \, , t = 1, .., m$$

where m is the number of states. In each of these equations, the current state-price for a security paying off in state j at time $T+1$ is the state-price for a payment at time T in some intermediate state k multiplied by the transition price of going from state k to state j, that is, $p(k,j)$ and then added up over all the possible intermediate states k. Hence,

by looking at m time periods, we have m^2 equations to solve for the m^2 unknown transition prices $p(i,j)$. We can then use the recovery theorem to the transition pricing matrix P to recover the pricing kernel and the resulting natural probability transition matrix.

E.6.4 Bounding the pricing kernel

In view of testing efficient market hypothesis, Ross [2005] proposed to find an upper bound to the volatility of the pricing kernel corresponding to a simple byproduct of recovery. Assuming that μ is an unobservable stochastic process, Hansen et al. [1991] found the lower bound

$$\sigma(\xi) \geq e^{-rT}\frac{\mu}{\sigma}$$

where μ is the absolute value of the excess return and σ is the standard deviation of any asset. It implies that $\sigma(\xi)$ is bounded from below by the largest observed discounted Sharpe ratio. Equivalently, $\sigma(\xi)$ is also an upper bound on the Sharpe ratio for any investment (strategy) to be consistent with efficient markets. That is, if the Sharpe ratio is above $\sigma(\xi)$, then the deal is too good (see Cochrane [2000] and Bernardo et al. [2000]). Alvarez et al. [2005] derived and estimated a lower bound for the volatility of the permanent component of asset pricing kernels, based on return properties of long-term zero-coupon bonds, risk-free bonds, and other risky securities. They found that the permanent component of pricing kernels was very volatile, at least as large as the volatility of the stochastic discount factor. Alternatively, Ross proposed to decompose excess returns R_t on an asset or portfolio strategy as

$$R_t = \mu(I_t) + \epsilon_t$$

where the mean depends on the information set I, and the residual term is not correlated with I. The variance satisfies

$$\sigma^2(R_t) = \sigma^2(\mu(I_t)) + \sigma^2(\epsilon_t) \leq E[\mu^2(I_t)] + \sigma^2(\epsilon_t)$$

Rearranging, we get an upper bound on the R^2 of the regression

$$R^2 = \frac{\sigma^2(\mu(I_t))}{\sigma^2(R_t)} \leq \frac{E[\mu^2(I_t)]}{\sigma^2(R_t)} \leq e^{2rT}\sigma^2(\xi)$$

and R^2 is bounded above by the volatility of the pricing kernel. Note, to avoid arbitrarily high volatility from the kernel, the proper maximum to be used is the volatility of the projection of the kernel on the stock market. Hence, the above tests must be done on strategies based on stock returns and the filtration they generate. Note also that the results of the tests are subject to the assumptions of the recovery theorem, and that any strategy must overcome transaction costs to be an implementable violation.

Appendix F

Pricing and hedging options

F.1 Valuing options on multi-underlyings

In Appendix (E), following Dana and Jeanblanc-Picque [1994], we presented simple models in discrete time with one or two time periods and with a finite number of states of the world to value asset prices and define market equilibrium. We are now going to consider more complex models in continuous time. To do so, one can either take the results obtained in discrete time and consider the limit case, or one can consider a probabilist approach and define a probability under which the price of assets discounted by the riskless asset are martingales. We consider the latter, and follow the text book by Dana and Jeanblanc-Picque [1994] and the lecture notes by El Karoui [1997].

F.1.1 Self-financing portfolios

***Assumption* F.1.1** *A simplified world*

- *there is no transaction costs when buying or selling assets*
- *illimited short-selling is allowed*
- *assets do not pay dividends*
- *one can buy and sell assets at all time*

We let T_H be the market horizon, and assume $(d+1)$ assets $S = (S^0, S^1, .., S^d)$ traded between 0 and T_H. We let S_t^i be the price of the ith asset at time t with price process continuous in time. The uncertainties in the economy $(\Omega, \mathcal{F}, \mathbb{P})$ are given by a k-dimensional Brownian motion $(\hat{W}_t)_{t\in\mathbb{R}^+}$ with components $\hat{W}_t^j$ being independent Brownian motions. We let S^0 be the riskless asset with return rdt over the time interval $[t, t+dt]$. The risky-assets $S^i; 1 \leq i \leq d$ are Ito's random functions satisfying the dynamics

$$dS_t^i = S_t^i\big(b_t^i dt + \sum_{j=1}^{k} \sigma_j^i(t)\hat{W}_t^j\big)$$

where b_t is an adapted vector in $\mathbb{R}^d$ with components b_t^i corresponding to the rate of return of the assets. The matrix σ_t of dimension $d \times k$ is the adapted matrix of volatility of the assets. The risk-free asset S^0 satisfies the dynamics

$$dS_t^0 = r_t S_t^0 dt$$

Given an investor with initial capital x, at time t, his portfolio is made of $\delta^i(t)$ parts of the asset i for $i = 0, 1, .., d$ which can be positive or negative. A portfolio strategy is given by the process $(\delta_i(t))_{0\leq i\leq d}$ corresponding to the quantity invested in each asset. We consider simple strategies where the composition of the portfolio changes on a finite set of dates called trading dates. In discrete time, any strategy is a simple strategy.

Definition F.1.1 *A simple strategy written on assets is given by a finite set of trading dates*

$$\Theta = \{(t_i)_{0\leq i\leq n}; 0 = t_0 < t_1 < t_2 < ... < t_n = T\}$$

and $(d+1)$ *process* $(\delta_i(t))_{0\leq i\leq d}$ *giving the allocation of the assets in the portfolio over time*

$$\delta^i(t) = n_0^i I_{[0,t_1]}(t) + n_k^i I_{]t_k,t_{k+1}]}(t) + ... + +n_{N-1}^i I_{]t_{N-1},t_N]}(t)$$

where the variables n_k^i *are* $\mathcal{F}_{t_k}$*-mesurable.*

The financial value of the portfolio δ is given by $V(\delta)$. At time t, that value is given by

$$V_t(\delta) =< \delta(t), S_t >= \sum_{i=0}^{d} \delta^i(t) S_t^i$$

Not, for all t in the interval $]t_k, t_{k+1}]$ then $\delta^i(t) = \delta^i(t_{k+1}) = n^i(k)$. That is, the part invested in the ith asset is $\mathcal{F}_{t_k}$-measurable and only depends on the information available at the previous trading date. Hence, the process $\delta^i(t)$ is predictable and the process $V_t(\delta)$ is adapted. In continuous time, as the simple strategies are processes left-continuous so is the value of a simple portfolio.

Between the dates t_k and t_{k+1} an investor following the strategy δ put n_k^i units in asset S^i. Just before the trading date t_{k+1} the portfolio value is $< n_k, S_{t_{k+1}} >$. At time t_{k+1}, the investor build a new portfolio, that is, reallocate the weights of the portfolio from the information at time t_{k+1}. Assuming no cash is added or removed from the portfolio, then at time t_{k+1} the self-financing condition can be expressed as

$$< n_k, S_{t_{k+1}} >=< n_{k+1}, S_{t_{k+1}} >$$

or when considering the assets' variation between two dates

$$V_{t_k}(\delta) + < n_k, S_{t_{k+1}} - S_{t_k} >=< n_k, S_{t_{k+1}} >= V_{t_{k+1}}(\delta)$$

and the variations of a self-financing portfolio are only due to the variations of the assets.

A self-financing portfolio is a strategy to buy or sell assets whose value is not modified by adding or removing cash. Further, the self-financing condition implies that the value of the portfolio does not jump at trading dates. The self-financing condition is a necessary and sufficient condition for the continuity process of the value of the portfolio.

Definition F.1.2 *Given* (Θ, δ) *a simple self-financing trading strategy, the value of the portfolio is characterised by*

$$V_t(\delta) =< \delta(t), S_t >$$
$$V_t(\delta) - V_0(\delta) = \int_0^t < \delta(u), dS_u >$$

If the portfolio is not self-financing, Follmer-Schweitzer defined the cost process

$$C_t(\delta) = V_t(\delta) - V_0(\delta) - \int_0^t < \delta(u), dS_u >$$

We are going to express the dynamics of the self-financing portfolio in terms of the rate of return and volatility of each asset.

$$\begin{aligned} dV_t(\delta) &= \sum_{i=0}^{d} \delta^i(t) dS_t^i = \delta^0(t) dS_t^0 + \sum_{i=1}^{d} \delta^i(t) dS_t^i \\ &= \delta^0(t) S_t^0 r_t dt + \sum_{i=1}^{d} \delta^i(t) S_t^i b_t^i dt + \sum_{i=1}^{d} \sum_{j=1}^{k} \delta^i(t) S_t^i \sigma_j^i(t) d\hat{W}_t^j \\ &= \delta^0(t) S_t^0 r_t dt + < (\delta S)_t, b_t > dt + < (\delta S)_t, \sigma_t d\hat{W}_t > \end{aligned}$$

We can get rid off $\delta^0(t)$ by using the self-financing condition, to get the dynamics

$$dV_t(\delta) = r_t V_t dt + < (\delta S)_t, b_t - r_t I > dt + < (\delta S)_t, \sigma_t d\hat{W}_t > \tag{F.1.1}$$

where $(\delta S)_t = \pi_t$ correspond to the vector with component $(\delta^i(t) S_t^i)_{1 \leq i \leq d}$ describing the amount to be invested in each stock. The linear Equation (F.1.1) having a unique solution, knowing the initial investment and the weights of the portfolio is enough to characterise the value of the portfolio.

***Remark* F.1.1** *A process V_t solution to the Equation (F.1.1) is the financial value of a self-financing portfolio corresponding to investing the quantity $(\delta^i(t))_{1 \leq i \leq d}$ in risky assets and the quantity $\frac{1}{S_t^0}(V_t(\delta) - \sum_{i=1}^{d} \delta^i(t) S_t^i)$ in the risk-free asset.*

$$V_t(\delta) = \delta^0(t) S_t^0 + \sum_{i=1}^{d} \delta^i(t) S_t^i \text{ where } \delta^0(t) = \frac{1}{S_t^0}(V_t(\delta) - \sum_{i=1}^{d} \delta^i(t) S_t^i)$$

Some examples:

- Fixed time strategy
  The trader decides to maintain the amount invested in the risky asset at 50% of the portfolio value, that is $\delta_t S_t = \frac{1}{2} V_t$. Given Equation (F.1.1), the portfolio value is solution to the SDE

  $$dV_t(\delta) = r_t V_t dt + \frac{1}{2} V_t (\frac{dS_t}{S_t} - r dt)$$

  for $\frac{dS_t}{S_t} = b_t dt + \sigma_t d\hat{W}_t$. In the special case where the trader maintains 100% of the amount invested in the risky asset, we get $\delta_t S_t = V_t$ and the equation becomes

  $$dV_t(\delta) = r_t V_t dt + V_t (\frac{dS_t}{S_t} - r dt)$$

- Random time strategy
  The trader decides to rebalance his portfolio as soon as the stock price varies more than 2%. Regarding the choice of the weights, he can keep the same rule as in the previous example. The trading dates are therefore random dates being stopping time characterised by the first time the asset crosses a 2% barrier level.

Note, one can add options in his portfolio provided that they are very liquid and easily tradable on the market. In that case, letting $(C_t^i)_{i=1}^d$ be option prices, and δ_t^i be the quantity of these claims, then the self-financing equation becomes

$$dV_t = \delta_t dS_t + \sum_{i=1}^d \delta^i(t) dC_t^i + \big(V_t - \delta_t S_t - \sum_{i=1}^d \delta^i(t) C_t^i\big) r_t dt$$

where the portfolio is

$$V_t = \delta_t S_t + \sum_{i=1}^d \delta^i(t) C_t^i + \frac{1}{S_t^0}\big(V_t - \delta_t S_t - \sum_{i=1}^d \delta^i(t) C_t^i\big) S_t^0$$

We can rewrite the SDE as

$$dV_t = r_t V_t dt + \delta_t(dS_t - r_t S_t dt) + \sum_{i=1}^d \delta^i(t)(dC_t^i - r_t C_t^i dt)$$

F.1.2 Absence of arbitrage opportunity and rate of returns

***Assumption* F.1.2** *We assume that there is no arbitrage opportunity between admissible portfolio strategies, in which case the market is viable.*

As a result of the absence of arbitrage opportunity, there is a constraint on the rate of return of financial assets. The riskier the asset, the higher the return, to justify its presence in the portfolio.

Theorem F.1.1 *Given a viable Ito market*

1. *two admissible and non-risky portfolios have the same rate of return r_t*
2. *there exists an adapted random vector λ_t taking values in $\mathbb{R}^k$, called the market price of risk, such that*

$$dS_t^i = S_t^i\big[r_t dt + \sum_{j=1}^k \sigma_j^i(t)(d\hat{W}_t^j + \lambda_t^j dt)\big]$$

The instantaneous rate of returns of the risky assets satisfy

$$b_t = r_t I + \sigma_t \lambda_t \text{ , } d\mathbb{P} \times dt \text{ a.s.} \tag{F.1.2}$$

A process V_t is the financial value of an adapted strategy δ if and only if it satisfies

$$dV_t = r_t V_t dt + < (\delta S)_t, \sigma_t(d\hat{W}_t + \lambda_t dt) > \tag{F.1.3}$$

together with the integrability condition $E[\int_0^T V_t^2 dt + |(\delta S)_t|^2 dt] < +\infty$.

Note, one assumes that the rate of returns of risky assets are greater than the one from the risk-free assets. The equilibrium prices of risky assets exhibit a risk premium which is the spread between the instantaneous rate of return b_t of the risky assets and the rate of return r_t of the risk-free asset. However, in a risk-neutral economy, all assets have the same rate of return equal to the market interest rate.

F.1.3 Numeraire

Definition F.1.3 *A numeraire is a monetary reference with value in Euros which is an adapted and strictly positive random function of Ito.*

***Proposition* 17** *1. a simple self-financing strategy is invariant under change of numeraire*

2. *a portfolio strategy which is an arbitrage in a given numeraire is an arbitrage in all numeraire*
3. *if the asset S^0 is chosen as numeraire, all stochastic integral $\frac{x}{S_0^0} + \int_0^t < \delta(u), d\frac{S_u}{S_u^0} >$ with simple process δ is the financial value of a self-financing portfolio*

We let X_t be the value in Euros of a numeraire. The price at time t of the ith asset expressed in that numeraire is $\frac{S_t^i}{X_t}$ and the financial value of a portfolio is $\frac{V_t(\delta)}{X_t}$. Given a self-financing portfolio $V_t(\delta) =< \delta_t, S_t >$ with variation $dV_t(\delta) =< \delta_t, dS_t >$, we get

$$\frac{V_t(\delta)}{X_t} =< \delta_t, \frac{S_t}{X_t} > \text{ and } d\frac{V_t(\delta)}{X_t} =< \delta_t, d\frac{S_t}{X_t} >$$

Given (X_t) a numeraire with dynamics

$$\frac{dX_t}{X_t} = r_t dt - r_t^X dt + < \gamma_t^X, d\hat{W}_t + \lambda_t dt >$$

we let γ_t^X be a volatility vector belonging to the image of $(\sigma_t)^\top$, that is, there exists a vector π_t^X such that $\gamma_t^X = (\sigma_t)^\top \pi_t^X$. In the X-market, we let $S_t^X = \frac{S_t}{X_t}$ be the price process under the numeraire X_t.

Theorem F.1.2 *Given a viable initial market*

1. *the parameters in the X-market are*

$$\lambda_t^X = \lambda_t - \gamma_t^X \text{ , } r_t^X = \mu_t^X$$

2. *we let (Z_t) be an admissible price process with volatility vector σ_t^Z and dynamics*

$$\frac{dZ_t}{Z_t} = r_t dt + < \sigma_t^Z, d\hat{W}_t + \lambda_t dt >$$

The volatility vector of Z^X is given by $\sigma^{Z^X} = \sigma_t^Z - \gamma_t^X$ and the dynamics are

$$\frac{dZ_t^X}{Z_t^X} = r_t^X dt + < \sigma_t^Z - \gamma_t^X, d\hat{W}_t + (\lambda_t - \gamma_t^X) dt >$$

We consider the M-numeraire called the market numeraire and introduced by Long [1990] in reference to the market portfolio described by Markowitz [1952]. Long showed that if any set of assets is arbitrage-free, then there always exists a numeraire portfolio comprised of just these assets. The prices expressed in that numeraire have no specific return over time. They are white noise under the historical probability as the interest rates and risk premiums are null. As a result, the M-market prices are local martingales and risk-neutral under the historical probability.

We let λ_t be a volatility vector belonging to the image of $(\sigma_t)^\top$, that is, there exists a vector α_t such that $\lambda_t = (\sigma_t)^\top \alpha_t$. This condition is not restrictive and we can always decompose λ_t as $(\lambda_t^1, \lambda_t^2)$ where λ_t^1 belongs to the seed of σ_t and λ_t^2 belongs to the orthogonal space $Ker(\sigma_t) = Image(\sigma_t^\top)$. The market price of risk (λ_t) and (λ_t^2) have

the same impact on the dynamics of the price as they are linked to the volatility through $\sigma_t\lambda_t = \sigma_t\lambda_t^2$. Hence, we consider a self-financing strategy where the weights in the risky assets are $\phi_t^M = (\frac{\alpha_t^i}{S_t^i})_{i=1}^d$ corresponding to an initial investment of \$1, and we assume that this strategy is admissible. The value of this strategy noted M_t is called market numeraire.

Theorem F.1.3 *1. Given the M-market numeraire, a process with initial value equal to* 1 *and volatility* (λ_t)*, the dynamics are*

$$\frac{dM_t}{M_t} = r_t dt + (\lambda_t)^\top (d\hat{W}_t + \lambda_t dt) = r_t dt + |\lambda_t|^2 dt + \lambda_t^\top d\hat{W}_t \, , \, M_0 = 1$$

- *in the M-market the investors are risk-neutral*
- *the M-price* $Z_t^M = \frac{Z_t}{M_t}$ *of an asset or a portfolio is a local martingale*

2. Arbitrage Pricing Theory
In the initial market, the expected return of an asset Z *is given by the risk-free rate plus the infinitesimal covariance between the return of the risky asset and that of the market numeraire*

$$\mu_t^Z = r_t + \sigma_{Z,M}(t) \, , \, \sigma_{Z,M}(t) = Cov_t(\frac{dM_t}{M_t}, \frac{dZ_t}{Z_t})$$

In the APT, the excess return with respect to cash is measured by the beta of the portfolio return with respect to the market numeraire

$$\mu_t^Z - r_t = \frac{\sigma_{Z,M}(t)}{\sigma_{M,M}(t)}(\mu_t^M - r_t)$$

where $\sigma_{M,M}(t) = Var_t(\frac{dM_t}{M_t})$.

Proof
Using no-arbitrage arguments we can show

$$\mu_t^Z - r_t =< \sigma_t^Z, \lambda_t >$$

where σ_t^Z is the volatility of Z and λ_t is the vector of market price of risk. As λ_t is also the volatility of the market numeraire, we get

$$Cov_t(\frac{dM_t}{M_t}, \frac{dZ_t}{Z_t}) =< \sigma_t^Z, \lambda_t > dt$$
$$Cov_t(\frac{dM_t}{M_t}, \frac{dM_t}{M_t}) = |\lambda_t|^2 dt = (\mu_t^M - r_t)dt$$

F.1.4 Evaluation and hedging

We can then compute the price of contingent claims as they are replicable with an admissible portfolio.

***Proposition* 18** *We let* $\mathcal{B}_T = \{\Phi_T = V_T(\delta); \delta$ *admissible self-financing strategy* $\}$ *be the set of simulable flux which are square integrable.*

- *assuming absence of arbitrage opportunity, two strategies replicating Φ_T have the same value at all intermediary dates, which is the price $C_t(\Phi_T)$ of Φ_T satisfying the equation*

$$\begin{aligned} dC_t(\Phi_T) &= C_t(\Phi_T)r_t dt + < (\delta S)_t, \sigma_t(d\hat{W}_t + \lambda_t dt) > \\ C_T(\Phi_T) &= \Phi_T \end{aligned}$$

  where δ is a hedging portfolio of the contingent claim Φ.

- *assuming absence of arbitrage opportunity, the application which at $\Phi_T = V_T(\delta) \in \mathcal{B}_T$ associates its price at time t given by $V_t(\delta) = C_t(\Phi_T)$ is a positive linear function.*

The price of a simulable asset is the unique solution to a linear SDE with known terminal value. The hedging portfolio δ is unknown, just like is unknown the price $C_t(\Phi_T)$. Hence, the pair $(C_t(\Phi_T), \delta_t)$ is solution to a backward SDE (see Peng and Pardoux (1987)).

***Remark* F.1.2** *When the interest rates and the market price of risk are null, one can easily compute the price of a contingent claim as the expected value of the terminal flux. These conditions are satisfied in the M-market.*

***Proposition* 19** *Evaluation in the M-market*
We assume that the numeraire M is regular enough for $\frac{1}{M}$ to belongs to $\mathbb{H}^{2+\epsilon}$ which is the case when r_t, σ_t and λ_t are bounded. Given $\Phi_T \in \mathcal{B}_t$ the terminal flux of a simulable contingent claim which is square integrable, its price $C_t(\Phi_T)$ is given by $C_t^M(\Phi_T^M) = E[\Phi_T^M|\mathcal{F}_t]$ which is written in the usual numeraire as

$$C_t(\Phi) = E[\Phi_T \frac{M_t}{M_T}|\mathcal{F}_t]$$

In the M-market the prices of contingent claims are the expected value of their terminal flux. One is then left to compute the weights of the hedging portfolio when the terminal flux is unknown. In the simple Markowian case of the Black-Scholes formula the weights are expressed in terms of the gradients of the price which is the expectation of the derivative of the random variable Φ^M with respect to the Markowian variables. In the general case, we need to compute the weights representing the random variable X_T as a stochastic integral with respect to the Brownian motions. We need to characterise the set of replicable derivatives. Hence, we need to characterise the complete markets in which the derivatives are replicable. We use the probability result that says that all random variable X_T in $\mathbb{L}^1(\mathbb{P})$ and measurable with respect to the sigma-algebra generated by the vectorial Brownian motion $\hat{W}$ can be expressed in terms of a stochastic integral

$$X_T = E[X_T] + \int_0^T < z_s, d\hat{W}_s > \ , \ \int_0^T |z_s|^2 ds < \infty$$

In a financial market, one need to express the prices on the basis of the asset variations and not on the basis of Brownian motions. One must therefore assume that there is enough assets to cover all the noises.

***Assumption* F.1.3** *The matrix inferred from the volatility matrix $\sigma_t\sigma_t^\top$ is invertible and bounded as well as the inverse matrix.*

***Proposition* 20** *Assuming a financial market with no interest rates, and no risk premium such as the M-market. All random variable Φ_T which is measurable with respect to the Brownian filtration $\hat{W}$ belonging to $\mathbb{L}^{1+\epsilon}$, is replicable with a portfolio which is a uniformly integrable martingale*

$$\begin{aligned} \Phi_T &= E[\Phi_T] + \int_0^T < \alpha_t, d\hat{W}_t > = E[\Phi_T] + \int_0^T < \delta_t S_t, \frac{dS_t}{S_t} > \\ \delta_t S_t &= (\sigma_t\sigma_t^\top)^{-1}\sigma_t\alpha_t \end{aligned}$$

Theorem F.1.4 *Complete market*
We let the matrix σ_t satisfies the assumptions above, and assume that the M-numeraire is regular enough such that $\sup_{0\leq t\leq T}(M_t)$ and $\sup_{0\leq t\leq T}(M_t^{-1})$ belongs to $\mathbb{L}^{1+\epsilon}$. Given $\Phi_T \in \mathbb{L}^2(\mathcal{F}_T^{\hat{W}}, \mathbb{P})$, the terminal flux of a simulable contingent claim, square integrable, and measurable with respect to the Brownian filtration. The market is complete in the sense where Φ_T is replicable with an admissible portfolio.

In practice, we can not observe the M-numeraire, and we need to find an observable numeraire with similar properties to value and hedge contingent claims. For the price to be the expected value of the terminal flux expressed in the new numeraire, we need to make the market risk-neutral which is possible after a change of probability.

Theorem F.1.5 *Risk-neutral probability*
We assume that the vectors λ_t and r_t are bounded [1], and we choose the cash S^0 as numeraire.

1. *there exists a probability $\mathbb{Q}$ equivalent to $\mathbb{P}$ such that*
   - *$W_t = \int_0^t d\hat{W}_s + \lambda_s ds$ is a $\mathbb{Q}$-Brownian motion*
   - *the prices processes written in that numeraire (the discounted values of self-financing portfolios) are local $\mathbb{Q}$-martingale satisfying for $Z_t^a = \frac{Z_t}{S_t^0}$ the dynamics*

     $$\frac{dZ_t^a}{Z_t^a} = \sigma_t^Z dW_t$$

     and $\mathbb{Q}$ is called the risk-neutral probability.
2. *we further assume that the negative part of the interest rate $(r_t)^-$ is bounded. If $\Phi_T \in \mathcal{B}_T$ is the terminal flux of a contingent claim, we get*

   $$\frac{C_t(\Phi_T)}{S_t^0} = E^Q[\frac{\Phi_T}{S_T^0}|\mathcal{F}_t]$$

   or

   $$C_t(\Phi_T) = E^Q[e^{-\int_t^T r_s ds}\Phi_T|\mathcal{F}_t]$$

3. *when the market is complete, there exists a unique risk-neutral probability, and the risk-neutral rule of valuation can be applied to all contingent claims which are square integrable.*

The same argument can be applied to all numeraire provided that we consider the appropriate risk-neutral probability. In the M-market, the M-prices are local martingales with respect to the historical probability $\mathbb{P}$, which is called a M-martingale measure.

***Remark* F.1.3** *The historical probability is the risk-neutral probability when the market numeraire is taken as a reference.*

Theorem F.1.6 *Given (X_t) a portfolio numeraire such that $X_t^M = \frac{X_t}{M_t}$ is a uniformly integrable $\mathbb{P}$-martingale, there exists a probability $\mathbb{Q}^X$ defined by its Radon-Nikodym derivative with respect to $\mathbb{P} = \mathbb{Q}^M$ such that*

$$\frac{d\mathbb{Q}^X}{d\mathbb{P}} = \frac{X_T}{M_T}\frac{M_0}{X_0} = \frac{X_T^M}{X_0^M}$$

The X-prices are local $\mathbb{Q}^X$-martingales, that is, $\mathbb{Q}^X$ are X-martingale measures.

[1] it is enough that the Novikov condition $E[e^{\frac{1}{2}\int_0^T |\lambda_s|^2 ds}] < +\infty$ be satisfied.

To conclude, we introduce the Ross [2013] recovery theorem which gives sufficient conditions under which the probability measure $\mathbb{P}$ can be obtained from the measure $\mathbb{Q}$. The theorem states that in a complete market, if the utility function of the representative investor is state independent and intertemporally additively separable, and if the state variable is a time homogeneous Markov process X with a finite discrete state space, then one can recover the real-world transition probability matrix from the assumed known matrix of Arrow-Debreu state prices. Considering the restrictions on preferences made by Ross impossible to test, Carr et al. replaced that concept with restrictions on beliefs.

F.2 The dynamics of financial assets

We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}_t$ is a right continuous filtration including all $\mathbb{P}$ negligible sets in $\mathcal{F}$. For simplicity, we let the market be complete and assume that there exist an equivalent martingales measure as defined in a mixed diffusion model by Bellamy and Jeanblanc in [2000].

F.2.1 The Black-Scholes world

In the special case where we assume that the stock prices are lognormally distributed we can derive most of the contract market values in closed form solution. In the risk-neutral pricing theory we introduce a new probability measure $\mathbb{Q}$ such that the discounted value process $\overline{V}_t = e^{-rt}V_t$ of the replicating portfolio is a martingale under $\mathbb{Q}$. By the fundamental theorem of asset pricing (see Harrison et al. [1981]), such a measure exists if and only if the market is arbitrage-free. We let the stock price S_t under the historical measure $\mathbb{P}$ takes values in $\mathbb{R}$ with dynamics

$$\frac{dS_t}{S_t} = \mu dt + \sigma d\tilde{W}_S(t)$$

where μ is the drift, σ is the volatility and $\tilde{W}_S(t)$ is a standard Brownian motion. Note, μ is the annualised expected rate of return of the stock by unit time. The parameter of reference is $\mu - r$. Given $S_0 = x$ we get

$$S_t = f(t, \tilde{W}_S(t)) = xe^{\mu t - \frac{1}{2}\sigma^2 t + \sigma \tilde{W}_S(t)}$$

The first two moments are

$$E[S_t] = xe^{\mu t}\ ,\ E[S_t^2] = x^2 e^{(2\mu+\sigma^2)t}$$
$$Var(S_t) = x^2 e^{2\mu t}(e^{\sigma^2 t} - 1)$$

and the Sharpe ratio of the stock price is

$$M = \frac{E[S_t] - x}{\sqrt{Var(S_t)}} = \frac{(e^{\mu t} - 1)}{e^{\mu t}\sqrt{(e^{\sigma^2 t} - 1)}}$$

which is independent from the initial value x. We can also compute the Sharpe ratio by unit time of excess returns with respect to cash which is

$$\lambda = \frac{\frac{1}{dt}E[\frac{dS_t}{S_t}] - r}{\sqrt{\frac{1}{dt}Var(\frac{dS_t}{S_t})}} = \frac{\mu - r}{\sigma}$$

It corresponds to the market price of risk λ associated to the Brownian motion $\tilde{W}$. Hence, the dynamics of the stock price becomes

$$\frac{dS_t}{S_t} = rdt + \sigma\big(d\tilde{W}_S(t) + \lambda dt\big)$$

In the risk-neutral measure the portfolio replicates the derivative such that at maturity we have $V_T = h(S_T)$ where $h(.)$ is a sufficiently smooth payoff function. As a result, the price of a European call option $C(t,x)$ on $[0,T] \times [0,+\infty[$ is

$$C(t,S_t) = e^{-r(T-t)} E^Q[h(S_T)|\mathcal{F}_t]$$

and the price of the derivative is a linear function. The change of measure is achieved by using the Girsanov's Theorem which states that there exists a measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ such that the discounted stock price is a martingale under $\mathbb{Q}$. By Girsanov's Theorem the new measure is computed via the Radon-Nikodym derivative

$$\frac{dQ}{dP}|_{\mathcal{F}_T} = e^{-\frac{1}{2}\int_0^T \lambda_s^2 ds + \int_0^T \lambda_s d\tilde{W}(s)}$$

where the market price of risk λ_t follows

$$\lambda_t = \frac{\mu - r}{\sigma} \tag{F.2.4}$$

and such that $E[e^{-\frac{1}{2}\int_0^T \lambda_s^2 ds}] < \infty$. Since the Brownian motion

$$W(t) = \tilde{W}(t) + \int_0^t \lambda_s ds$$

is a Brownian motion on the space $(\Omega, \mathcal{F}, \mathbb{Q})$ the SDE of the asset price becomes

$$\frac{dS_t}{S_t} = rdt + \sigma dW_S(t)$$

F.2.2 The dynamics of the bond price

We consider the underlying process to be the zero-coupon bond price of maturity T where $P(t,T)$ is the price at time t of 1\$ paid at maturity T. However, practitioners would rather work with the forward instantaneous rate which is related to the bond price by

$$f_P(t,T) = -\partial_T \ln P(t,T)$$

We assume that zero-coupon bonds of all maturities are continuously traded in the market and follow Ito processes. Then, each asset has an instantaneous return and a vector of volatility under the historical probability measure $\mathbb{P}$. We let $\hat{W}_P(t) = \left(W_{P,1}(t), ..., W_{P,n}(t)\right)^\top$ be a column vector with dimension $(n,1)$ and $\Gamma(t,T)$ be a matrix of dimension $(1,n)$ with element $\Gamma_j(t,T)$. The absence of arbitrage opportunities implies that there exist a vector of market price of risk λ_t [2] such that the dynamic of the bond price is

$$\frac{dP(t,T)}{P(t,T)} = r_t dt + < \Gamma(t,T), \lambda_t dt > \pm < \Gamma(t,T), d\hat{W}_P(t) > \ , P(T,T) = 1$$

where we use the notation $\pm$ to indicate when we consider the bond price or the instantaneous forward rate as the underlying process. Given K_t an adapted process, the volatility must satisfies

$$|\partial_T \Gamma(t,T) = \gamma(t,T)| \leq K_t$$

Moreover, the AAO implies that all asset prices depend on the Brownian motion W_t such that

$$dW_t = d\hat{W}_t \pm \lambda_t dt$$

[2] $< x, y >$ is the scalar product of two vectors.

The market price of risk λ_t is cancelled out by introducing the probability $\mathbb{Q}$ where W_t is a centred Brownian motion, that is, there exists a risk-neutral probability $\mathbb{Q}$ equivalent to $\mathbb{P}$ such that

$$P(t,T) = E^Q[e^{-\int_t^T r_s ds}|\mathcal{F}_t]$$

As a result, to model the bond price we can characterise a dynamic of the short rate r_t. However, the AAO allows us to describe the dynamic of the bond price from its initial value and the knowledge of its volatility function. Therefore, assuming further hypothesis, the shape taken by the volatility function fully characterise the dynamic of the bond price and some specific functions gave their names to popular models commonly used in practice. Hence, the dynamic of the zero-coupon bond price is

$$\frac{dP(t,T)}{P(t,T)} = r_t dt \pm \Gamma_P(t,T) dW_P(t) \text{ with } P(T,T) = 1 \tag{F.2.5}$$

where $(W_P(t))_{t\geq 0}$ is valued in $\mathbb{R}^n$ and $\Gamma_P(t,T)$ [3] is a family of local volatilities parameterised by their maturities T. Market uncertainty is expressed via the presence of noises and in general the market introduces as many noises as there are traded maturities. We therefore need to assume something about how all these noises are correlated. So far, that structure of correlation is modelled with a finite number of independent Brownian motions and a vector of volatility $\Gamma(t,T)$. Using the properties of the stochastic integral and the Brownian motion, we can reduce the diffusion term to the volatility $vol(t,T)$ and a unidimentional Brownian motion $Z_T(t)$ as

$$\Gamma(t,T)dW(t) = vol(t,T)dZ_T() \tag{F.2.6}$$

In that setting, the link between the vector of local volatilities and the zero-coupon bond volatility is

$$vol^2(t,T) = \sum_{j=1}^{k} \Gamma_j^2(t,T)$$

Using vectorial notation, the instantaneous volatility is $vol(t,T) = |\Gamma(t,T)|$. We can therefore calculate the instantaneous correlations between zero-coupon bond of different maturities

$$Cov(\frac{dP(t,T+\theta)}{P(t,T+\theta)}, \frac{dP(t,T)}{P(t,T)}) = \Gamma(t,T+\theta)\Gamma(t,T)^\top dt \tag{F.2.7}$$

The relationship between the bond price and the rates in general was found by Heath et al. [1992],and following their approach the forward instantaneous rate is

$$f_P(t,T) = f_P(0,T) \mp \int_0^t \gamma_P(s,T)dW_P(s) + \int_0^t \gamma_P(s,T)\Gamma_P(s,T)^T ds$$

where $\gamma_P(s,T) = \partial_T \Gamma_P(t,T)$. The spot rate $r_t = f_P(t,t)$ is therefore

$$r_t = f_P(0,t) \mp \int_0^t \gamma_P(s,t)dW_P(s) + \int_0^t \gamma_P(s,t)\Gamma_P(s,t)^T ds$$

Similarly to the bond price, the short rate is characterised by the initial yield curve and a family of bond price volatility functions.

[3] $\Gamma_P(t,T)dW(t) = \sum_{j=1}^n \Gamma_{P,j}(t,T)dW_j(t)$

F.3 From market prices to implied volatility

F.3.1 The Black-Scholes formula

In the special case where we assume that the stock prices $(S_t^x)_{t\geq 0}$ are lognormally distributed and pay a continuous dividend we can derive most of the contract market values in closed form solution. For example, Black and Scholes [1973] derived the price of a call option seen at time t with strike K and maturity T as

$$C_{BS}(t,x,K,T) = xe^{-q(T-t)}N(d_1(T-t,F(t,T),K)) - Ke^{-r(T-t)}N(d_2(T-t,F(t,T),K)) \tag{F.3.8}$$

where $F(t,T) = xe^{(r-q)(T-t)}$ and

$$d_2(t,x,y) = \frac{1}{\sigma\sqrt{t}}\log\frac{x}{y} - \frac{1}{2}\sigma\sqrt{t} \text{ and } d_1(t,x,y) = d_2(t,x,y) + \sigma\sqrt{t}$$

For notational purpose, we let $P(t,T)$ be the discount factor, $Re(t,T)$ be the repo factor and we define $\eta = \frac{K}{F(t,T)} = \frac{KP(t,T)}{xRe(t,T)}$ to be the forward moneyness of the option. It leads to the limit case $\lim_{\eta\to 1} d_2(.) = -\frac{1}{2}\sigma\sqrt{t}$ and $\lim_{\eta\to 1} d_1(.) = \frac{1}{2}\sigma\sqrt{t}$. It is well known that when the spot rate, repo rate and volatility are time-dependent, we can still use the Black-Scholes formula (F.3.8) with the model parameters expressed as

$$r = \frac{1}{T-t}\int_t^T r(s)ds\ ,\ q = \frac{1}{T-t}\int_t^T q(s)ds\ ,\ \sigma^2 = \frac{1}{T-t}\int_t^T \sigma^2(s)ds$$

Further, we let the Black-Scholes total variance be given by $\omega(t) = \sigma^2 t$, and rewrite the BS-formula in terms of the total variance, denoted $C_{TV}(t,x,K,T)$, where

$$d_2(t,x,y) = \frac{1}{\sqrt{\omega(t)}}\log\frac{x}{y} - \frac{1}{2}\sqrt{\omega(t)} \text{ and } d_1(t,x,y) = d_2(t,x,y) + \sqrt{\omega(t)} \tag{F.3.9}$$

Expressing the strike in terms of the forward price $K = \eta F(t,T)$, the call price in Equation (F.3.8) becomes

$$C_{TV}(t,x,K,T)\big|_{K=\eta F(t,T)} = xe^{-q(T-t)}\big(N(d_1(\eta,\omega(T-t)) - \eta N(d_2(\eta,\omega(T-t)))\big)$$

where

$$d_2(\eta,\omega(t)) = -\frac{1}{\sqrt{\omega(t)}}\log\eta - \frac{1}{2}\sqrt{\omega(t)} \text{ and } d_1(\eta,\omega(t)) = d_2(\eta,\omega(t)) + \sqrt{\omega(t)} \tag{F.3.10}$$

which only depends on the forward moneyness and the total variance. This is the scaled Black-Scholes function discussed by Durrleman [2003]. In the special case where the spot price x is exactly at-the-money forward $K = F(t,T)$ with $\eta = 1$, the call price can be approximated with

$$C_{TV}(t,x,K,T)\big|_{K=F(t,T)} \approx xe^{-q(T-t)}\big(N(\frac{1}{2}\sqrt{\omega(T-t)}) - N(-\frac{1}{2}\sqrt{\omega(T-t)})\big) = 0.4xe^{-q(T-t)}\sqrt{\omega(T-t)} \tag{F.3.11}$$

which is linear in the spot price and the square root of the total variance.

F.3.2 The implied volatility in the Black-Scholes formula

A call price surface parametrised by s is a function

$$C : [0,\infty) \times [0,\infty) \to \mathbb{R}$$
$$(K,T) \to C(K,T)$$

along with a real number $s > 0$. However, market practise is to use the implied volatility when calculating the Greeks of European options. We let the implied volatility (IV) be a mapping from time, spot prices, strike prices and expiry days to $\mathbb{R}^+$

$$\Sigma : (t, S_t, K, T) \to \Sigma(t, S_t; K, T)$$

Hence, given the option price $C(t, S_t, K, T)$ at time t for a strike K and a maturity T, the market implied volatility $\Sigma(t, S_t; K, T)$ satisfies

$$C(t, S_t, K, T) = C_{BS}(t, S_t, K, T; \Sigma(K, T)) \tag{F.3.12}$$

where $C_{BS}(t, S_t, K, T; \sigma)$ is the Black-Scholes formula for a call option with volatility σ. Consequently, we refer to the two-dimensional map

$$(K, T) \to \Sigma(K, T)$$

as the implied volatility surface. Note, we will some time denote $\Sigma_{BS}(K, T)$ the Black-Scholes implied volatility. When visualising the IVS, it makes more sense to consider the two-dimensional map $(\eta, T) \to \Sigma(\eta, T)$ where η is the forward moneyness (corresponding to the strike $K = \eta F(t, T)$). Further, we let the total variance be given by $\omega(\eta, T) = \nu^2(\eta, T) = \Sigma^2(\eta, T)T$ and let the implied total variance satisfies

$$C(t, S_t, K, T) = C_{TV}(t, S_t, \eta F(t, T), T; \omega(\eta, T)) \tag{F.3.13}$$

where the two-dimensional map $(\eta, T) \to \omega(\eta, T)$ is the total variance surface. Note, $\sqrt{\omega(\eta, T)}$ corresponds to the time-scaled implied volatility in log-moneyness form discussed by Roper [2010].

F.3.3 The robustness of the Black-Scholes formula

El Karoui et al. [1998] provided conditions under which the Black-Scholes formula was robust with respect to a misspecification of volatility, that is, when the underlying assets do not satisfy the Black-Scholes hypothesis of deterministic volatility. To do so, they assumed that the contingent claims have convex payoffs, and the only source of randomness in misspecified volatility is a dependence on the current price of stock. Note, these assumptions ensure that they are working in a complete market. They found that when the misspecified volatility dominates (or is dominated by) the true volatility, then the contingent claim price corresponding to the misspecified volatility dominates (is dominated by) the true contingent claim. More formally, for both European and American contingent claims with convex payoffs, when interest rates are deterministic and the volatility depends only on time and the current stock price, then the price of the contingent claim is a convex function of the price of the stock. While Bergman et al. [1996] obtained this result by analysing the parabolic partial differential equation satisfied by the contingent claim price, El Karoui et al. used the theory of stochastic flows and the Girsanov theorem. They also showed that if the misspecified volatility dominates (is dominated by) the true volatility, then the self-financing value of the misspecified hedging portfolio exceeds (is exceeded by) the payoff of the contingent claim at expiration. Further, when the volatility of the underlying stock is allowed to be random in a path-dependent way, the price of a European can can fail to be convex in the stock price. Similarly, Bergman et al. considered an example where the volatility is decreasing with increasing initial stock price, and showed that dependence of the volatility on a second Brownian motion or jumps in the stock price can lead to nonincreasing, non-convex European call prices. This is to relate to the long-range dependence observed on volatility and the fact that financial markets are not complete. The convexity of market is strongly related to market completeness. Lyons [1995] considered non-convex options in a model with uncertain volatility.

F.4 Some properties satisfied by market prices

F.4.1 The no-arbitrage conditions

Market prices must satisfy strong constraints for the no-arbitrage conditions to apply. We refer the readers to Harrison and Pliska [1981] for detailed proof. Moreover, in the presence of discrete dividends the stock price can not fall bellow a floor value, imposing stronger constraints of no-arbitrage. For instance, call prices may not always be increasing in time. In fact, the classical no-arbitrage conditions still hold but for the pure diffusion process. For simplicity, we only assume dividend yield $D(t,T)$ from time t to maturity T. Further, we let $C(t,T) = \frac{Re(t,T)}{P(t,T)}$ be the capitalisation factor from time t to maturity T. The necessary no-arbitrage conditions are

Theorem F.4.1 *Under the diffusion model framework, a market has no arbitrage if the prices of the calls at time* $t_0 = 0$ *satisfy*

1. *the prices are decreasing and convex in the strike price* K
2. $\forall K \geq 0,\ T \geq 0$
   $P(t_0,T)\big(C(t_0,T)S_{t_0} - D(t_0,T)\big) \geq C(K,T) \geq P(t_0,T)\big(C(t_0,T)S_{t_0} - (D(t_0,T)+K)\big)^+$
3. $\forall \theta \geq 0$
   $C(K,T+\theta) \geq C(\frac{K+D(T,T+\theta)}{C(T,T+\theta)}, T)$

while the sufficient no-arbitrage conditions are

Theorem F.4.2 *If there exists a system of option prices twice differentiable in the strike that satisfies the conditions of Theorem (F.4.1) then the corresponding prices are arbitrage-free.*

F.4.2 Pricing two special market products

F.4.2.1 The digital option

Given the stock price $(S_t)_{t\geq 0}$, a Digital option $D(K,T)$ for strike K and maturity T pays \$1 when the stock price S_T is greater than the strike K, and zero otherwise. The price of the Digital option is

$$D(K,T) = \lim_{\Delta K \to 0} \frac{C(K,T) - C(K+\Delta K,T)}{\Delta K} = -\frac{\partial}{\partial K} C(K,T)$$

Given $C(K,T) = C_{TV}(K,T;\omega_{BS}(K,T))$ where $\omega_{BS}(K,T) = \Sigma^2_{BS}(K,T)T$ is the BS implied total variance for strike K and maturity T, and using the chain rule, the Digital option becomes

$$D(K,T) = -\frac{\partial}{\partial K} C_{TV}(K,T;\omega_{BS}(K,T)) = -\frac{\partial}{\partial K} C_{TV}(K,T;\omega_{BS}) - \frac{\partial}{\partial \omega} C_{TV}(K,T;\omega(K,T)) \frac{\partial}{\partial K}\omega(K,T)$$

We can express the Digital option in terms of the total variance Vega and the total variance Skew as

$$D(K,T) = -\frac{\partial}{\partial K} C_{TV}(K,T;\omega_{BS}) - Vega_{TV}(K,T) Skew_{TV}(K,T) \tag{F.4.14}$$

where $Vega_{TV}(K,T;\omega_{BS}(K,T))$ is the Black-Scholes total variance vega for the strike K and maturity T, and $\partial_K C_{TV}(K,T;\omega_{BS})$ is the BS digital price for the total variance $\omega_{BS}(K,T)$.

F.4.2.2 The butterfly option

Assuming that the volatility surface has been constructed from European option prices, we consider a butterfly strategy centred at K where we are long a call option with strike $K - \Delta K$, long a call option with strike $K + \Delta K$, and short two call options with strike K. The value of the butterfly for strike K and maturity T is

$$B(t_0, K, T) = C(K - \Delta K, T) - 2C(K, T) + C(K + \Delta K, T) \approx P(t_0, T)\phi(t_0; K, T)(\Delta K)^2$$

where $\phi(t_0; K, T)$ is the probability density function (PDF) of S_T evaluated at strike K. As a result, we have

$$\phi(t_0; K, T) \approx \frac{1}{P(t_0, T)} \frac{C(K - \Delta K, T) - 2C(K, T) + C(K + \Delta K, T)}{(\Delta K)^2}$$

and letting $\Delta K \to 0$, the density becomes

$$\phi(t_0; T, K) = \frac{1}{P(t_0, T)} \frac{\partial^2}{\partial K^2} C(K, T) \tag{F.4.15}$$

Hence, for any time T one can recover the marginal risk-neutral distribution of the stock price from the volatility surface.

F.5 Introduction to indifference pricing theory

F.5.1 Martingale measures and state-price densities

We saw in Appendix (F.1) that option pricing in complete markets consists in fixing a measure $\mathbb{Q}$ under which the discounted traded assets are martingales, and to calculate option prices via expectation under this measure. This is related to the notion of a state-price-density from economics. The advantage of using a state-price-density ξ_T is that prices π can be calculated as expectations under the physical measure, that is, $\pi = E[\xi_T C_T]$. In an incomplete market there is more than one martingale measure, or equivalently, there are infinitely many state-price-densities. For example, we consider a model on a stochastic basis $(\Omega, (\mathcal{F}_t)_{0 \le t \le T}, \mathbb{P})$ with a single traded asset with price process S_t and a second auxiliary process Y_t, which may correspond to a related but non-traded stock. The dynamics of the system is given by

$$\begin{aligned} \frac{dS_t}{S_t} &= (r_t + \sigma_t \lambda_t)dt + \sigma_t dW_S(t) \\ dY_t &= b_t dt + a_t d\hat{W}_Y(t) \end{aligned}$$

where $< dW_S, d\hat{W}_Y >_t = \rho_t dt$ and λ_t is the Sharpe ratio (market price of risk) of the traded asset S. We assume that even though the asset Y_t is not directly traded, its value is an observable quantity. We can write the Brownian motion $\hat{W}_Y$ as a composition of two independent Brownian motions W_S and W_Y such that $d\hat{W}_Y(t) = \rho_t dW_S(t) + \overline{\rho}_t dW_Y(t)$ where $\overline{\rho}_t = \sqrt{1 - \rho_t^2}$. A natural choice for Y to be an asset is to take $a_t = \eta Y_t$ and $b_t = Y_t(r + \eta\chi)$ where χ is the Sharpe ratio of the non-traded asset. In that setting, the equivalent martingale measures are given by

$$\frac{dQ}{dP}\Big|_{\mathcal{F}_T} = Z_T$$

where Z is a uniformly integrable martingale of the form

$$Z_t = e^{-\frac{1}{2}\int_0^t \lambda_s^2 ds - \int_0^t \lambda_s dW_S(s) - \frac{1}{2}\int_0^t \chi_s^2 ds - \int_0^t \chi_s dW_Y(s)}$$

with χ_t undetermined. For $\mathbb{Q}$ to be a true probability measure it is necessary to have $E[Z_T] = 1$. Note, Z_t can be seen as the value of S_t^0 in the market numeraire, that is, $Z_t = \frac{S_t^0}{M_t}$. An example of a candidate martingale Z_t is given by the

choice $\chi_t = 0$ leading to the minimal martingale measure $\mathbb{Q}^0$ of Follmer et al. [1990]. The state-price-densities ξ_T take the form

$$\xi_t = e^{-\int_0^t r_s ds} Z_t$$

with the property that $\xi_t S_t$ is a $\mathbb{P}$-local martingale. In the case where interest rates are deterministic, then the state-price density and the density of the martingale measure differ only by a positive constant, otherwise they differ by a stochastic discount factor. The martingale measures $\mathbb{Q}^\chi$, the associated martingales Z^χ, and state-price densities ξ_T^χ can all be parametrised by the process χ_t governing the change of drift on the non-traded Brownian motion W_Y. In a complete market, the fair price does not depend on the choice of numeraire, and the definition of martingale measure does depend on the choice of numeraire

$$\xi_T = \frac{1}{N_T} \frac{dQ^N}{dP}\Big|_{\mathcal{F}_T}$$

where N_t is the numeraire, and $\mathbb{Q}^N$ is a martingale measure for this numeraire. In an incomplete market there is risk, and an agent needs to specify the units in which these risks are to be measured, as well as the concave utility function to be used. Some authors consider cash at time T as the units in which utility is measured.

F.5.2 An overview

F.5.2.1 Describing the optimisation problem

We let π_b be the price at which the investor is indifferent (in the sense that his expected utility under optimal trading is unchanged) between paying nothing and not having the claim C_T, and paying π_b now to receive the claim (payoff) C_T at time T. We consider the problem with $k > 0$ units of the claim, and assume the investor initially has wealth x and zero endowment of the claim. We define

$$V(x,k) = \sup_{X_T \in \mathcal{A}(x)} E[u(X_T + kC_T)] \tag{F.5.16}$$

where the supremum is taken over all wealth X_T which can be generated from initial fortune x. When the wealth is build over time by investing in the financial market, we let $X_T^{x,\theta}$ denotes the terminal fortune of an investor with initial wealth x who follows a trading strategy consisting of holding θ_t units of the traded asset. In that case, the primal approach of the value function becomes

$$V(x,k) = \sup_\theta E[u(X_T^{x,\theta} + kC_T)] \tag{F.5.17}$$

The utility indifference buy (or bid) price $\pi_b(k)$ is the solution to

$$V(x - \pi_b(k), k) = V(x,0) \tag{F.5.18}$$

with rate $r = \frac{X_T + kC_T - \pi_b(k)}{X_T}$, where the investor is willing to pay at most the amount $\pi_b(k)$ today for k units of the claim C_T at time T. Similarly, the utility indifference sell (or ask) price $\pi_s(k)$ is the smallest amount the investor is willing to accept in order to sell k units of C_T satisfying

$$V(x + \pi_s(k), -k) = V(x,0)$$

Note,the proposed price is for an individual with particular risk preferences (see Detemple et al. [1999]). In contrast to the Black and Scholes price, utility indifference prices are non-linear in the number k of options. The investor is not willing to pay twice as much for twice as many options, but requires a reduction in this price to take on the additional risk. If the market is complete or if the claim C_T is replicable, the utility indifference price $\pi(k)$ is equivalent to the complete market price for k units.

Let π_i be the utility indifference price for one unit of payoff C_T^i and let $C_T^1 \leq C_T^2$, then $\pi_1 \leq \pi_2$ (Monotonicity). Let π_λ be the utility indifference price for the claim $\lambda C_T^1 + (1-\lambda)C_T^2$ where $\lambda \in [0,1]$, then $\pi_\lambda \geq \lambda\pi_1 + (1-\lambda)\pi_2$ (Concavity). Note, if we consider sell prices rather than buy prices then π_λ is convex rather than concave. In order to compute the utility indifference price of a claim from Equation (F.5.18), two stochastic control problems must be solved. The first is the optimal investment problem when the investor has a zero position in the claim (see Merton [1969] [1971]). Merton used dynamic programming to solve for an investor's optimal portfolio in a complete market where asset prices follow Markovian diffusions, leading to Hamilton Jacobi Bellman (HJB) equations and a PDE for the value function representing the investor's maximum utility. The second is the optimal investment problem when the investor has bought or sold the claim (LHS of Equation (F.5.18)). This optimisation involves the option payoff, and problems are usually formulated as one of stochastic optimal control and again solved in the Markovian case using HJB equations. An alternative solution approach is to convert this primal problem into the dual problem which involves minimising over state-price densities or martingale measures (see Karatzas et al. [1991] and Cvitanic et al. [2001]). Under this approach, the problems are no longer restricted to be Markovian in nature.

F.5.2.2 The dual problem

While in a complete market it is possible to write the set of attainable terminal wealth generated from an initial fortune x and a self-financing strategy as the set of random variables satisfying $E[\xi_T X_T] \leq x$, in an incomplete market this condition becomes that $E[\xi_T X_T] \leq x$ for all state-price-densities. One can therefore take a Lagrangian approach for solving Equation (F.5.17). For all state-price-densities ξ_T, terminal wealth X_T satisfying the budget constraint and non-negative Lagrange multipliers μ, then

$$E[u(X_T + kC_T) - \mu(\xi_T X_T - x)] \leq \mu x + \mu k E[\xi_T C_T] + E[\tilde{u}(\mu\xi_T)]$$

where $\tilde{u}$ is the Legendre-Frenchel transform of $-u$. Optimising over wealth on one hand, and Lagrange multipliers and state-price-densities on the other hand, we get

$$\sup_{X_T} E[u(X_T + kC_T)] \leq \inf_{\mu} \inf_{\xi_T} \left\{ \mu x + \mu k E[\xi_T C_T] + E[\tilde{u}(\mu\xi_T)] \right\} \tag{F.5.19}$$

corresponding to the dual problem. If we can find suitable random variables $X_T^{k,*}$ and $\xi_T^{k,*}$, and a constant $\mu^{k,*}$ such that $u'(X_T^{k,*} + kC_T) = \mu^{k,*}\xi_T^{k,*}$, then there should be equality in the above equation, and $X_T^{k,*}$ should be the optimal primal variable, and $\mu^{k,*}$ and $\xi_T^{k,*}$ should be the optimal dual variables. In the special case of deterministic interest rates and exponential utility function, the minimisation over ξ_T, or equivalently Z_T, in Equation (F.5.19), simplifies to

$$\inf_{Z_T} \left\{ E[Z_T \ln Z_T] + \gamma k E[Z_T C_T] \right\}$$

In addition, given the simple dependence of exponential utility function on initial wealth, one can deduce an expression for the form of the UIP given by

$$\pi(k) = \frac{e^{-rT}}{\gamma} \left(\inf_{Z_T} \left\{ E[Z_T \ln Z_T] + \gamma k E[Z_T C_T] \right\} - \inf_{Z_T} E[Z_T \ln Z_T] \right)$$

(see Frittelli [2000] and Rouge et al. [2000]).

***Remark* F.5.1** *A consequence of the dual problem is that the Sharpe ratio (market price of risk) plays a fundamental role in the characterisation of the solution to the utility indifference pricing problem. Moreover, it is the Sharpe ratio which determines whether an investment is a good deal.*

As a special case of the UIP theory, one can interpret super-replication pricing via a degenerate utility function. El Karoui et al. [1995] characterised the super-replication price as

$$\sup_{\xi_T} E[\xi_T C_T]$$

where the supremum is taken over the set of state-price densities, so that the price is the sell price under the worst case state-price density.

F.5.3 The non-traded assets model

F.5.3.1 Discrete time

As an example, we consider the pricing of European options on non-traded assets in a simple one-period binomial model where current time is denoted 0 and the terminal date is time 1 (see Smith et al. [1995]). The market consists of a riskless asset, a traded asset with price P_0 today and a non-traded asset with price Y_0 today. The traded price P_0 may move up to $P_1 = P_0\psi^u$ or down to $P_1 = P_0\psi^d$ where $0 < \psi^d < 1 < \psi^u$, and the non-traded price satisfies $Y_1 = Y_0\phi$ where $\phi = \phi^d, \phi^d$ and $\phi^d < \phi^u$. Wealth X_1 at time 1 is given by $X_1 = \beta + \alpha P_1 = x + \alpha(P_1 - P_0)$ where α is the number of shares of stock held, β is the money in the riskless asset, and x is the initial wealth. The investor is pricing k units of a claim with payoff C_1 and has exponential utility described in Appendix (A.7). The value function in Equation (F.5.17) becomes

$$V(x,k) = \sup_{\alpha} E[-\frac{1}{\gamma}e^{-\gamma(X_1+kC_1)}]$$

The utility indifference buy price satisfies Equation (F.5.18) and becomes

$$\pi_b(k) = E^{Q^0}[\frac{1}{\gamma}\log E^{Q^0}[e^{\gamma kC_1}|P_1]]$$

where $\mathbb{Q}^0$ is the measure under which the traded asset P is a martingale, and the conditional distribution of the non-traded asset given the traded one is preserved with respect to the real world measure $\mathbb{P}$. It corresponds to the minimal martingale measure of Follmer et al. [1990]. In this setting, the price can be written as a new non-linear, risk adjusted payoff, and then expectations are taken with respect to $\mathbb{Q}^0$ of this new payoff.

***Remark* F.5.2** *Exponential utility and the non-traded assets model is one of the few examples for which an explicit form for the utility indifference price is known.*

F.5.3.2 Continuous time

The theory of UIP has been extended to continuous time, and the canonical situation of a security which is not traded has been studied by Davis [1999]. We now consider the value function in Equation (F.5.17) with a trading strategy consisting of holding θ_t units of the traded asset, and assume the self-financing condition

$$dX_t^{x,\theta} = \theta_t dS_t + r_t(X_t^{x,\theta} - \theta_t S_t)dt$$

and sufficient regularity conditions to exclude doubling strategies. In the primal approach, we consider the dynamic version of the optimisation problem at an intermediate time t where

$$V(x,0) = V(x,s,y,t) = \sup_{\theta} E_t[u(X_T^{x,\theta})|X_t = x, S_t = s, Y_t = y]$$

Using the observation that $V(x,s,y,t)$ is a martingale under the optimal strategy θ, and a supermartingale otherwise, we have that V solves an equation of the form

$$\sup_{\theta} \mathcal{L}^{\theta}V = 0\,,\, V(x,s,y,T) = u(x)$$

where $\mathcal{L}^\theta$ is an operator. Given that $\mathcal{L}^\theta$ is quadratic in θ, the minimisation in θ is trivial and the problem can be reduced to solving a non-linear Hamilton-Jacobi-Bellman equation in four variables. In the special case of exponential utility function, wealth factors out of the problem and it is possible to consider $V = -\frac{1}{\gamma}e^{-\gamma x}\overline{V}(s,y,t)$ where $\overline{V}(s,y,T) = 1$. Assuming constant interest rates, and letting Y_t follows an exponential Brownian motion with dynamics

$$\frac{dY_t}{Y_t} = (r + \eta\chi)dt + \eta dW_Y(t)$$

it follows that with exponential utility we get

$$V(x,s,y,t) = -\frac{1}{\gamma}e^{-\gamma x e^{r(T-t)} - \frac{1}{2}\lambda^2(T-t)}$$

We are left with evaluating the left-hand side of Equation (F.5.18) under the assumption that $C_T = C(Y_T)$. The only change from the previous analysis is that the boundary condition becomes $V(x,s,y,T) = u(x + kC(y))$, but it is not possible to remove the dependence on y. Assuming exponential utility function, then the non-linear HJB equation can be linearised using the Hopf-Cole transformation, and the value function at $t = 0$ becomes

$$V(x,k) = -\frac{1}{\gamma}e^{-\gamma x e^{rT} - \frac{1}{2}\lambda^2 T}\left(E^{Q^0}[e^{-k\gamma(1-\rho^2)C(Y_T)}]\right)^{\frac{1}{1-\rho^2}}$$

where $\mathbb{Q}^0$ is the minimal martingale measure (see Henderson et al. [2002]). It follows that the price can be expressed as

$$\pi(k) = -\frac{e^{-rT}}{\gamma(1-\rho^2)}E^{Q^0}[e^{-k\gamma(1-\rho^2)C(Y_T)}]$$

where the price is independent of the initial wealth of the agent, and such that it is a non-linear concave function of k. Note, when $k > 0$, and C is non-negative, the bid price is well defined, but for $k < 0$ it may be that the price is infinite. Thus one of the disadvantages of exponential utility is that the ask price for many important examples of contingent claims is infinite.

F.5.4 The pricing method

We let $(X_t)_{t\geq 0}$ be the underlying asset under the historical measure $\mathbb{P}$ taking values in $\mathbb{R}$. When it is not possible to hold the underlying asset, we have to price the contingent claim in an incomplete market. Defining the risk premium λ, we can then price the option under the risk-neutral measure $\mathbb{Q}^\lambda$. Hence, the price at time t of the contingent claim under the probability measure $\mathbb{Q}^\lambda$ is

$$C(t,T) = P(t,T)E_t^\lambda[h(X_T)]$$

where $h(\bullet)$ is a sufficiently smooth payoff function, and $P(t,T)$ is a zero-coupon bond. However, to compute this price, we need to know the values of existing prices to infer the risk premium. When existing prices are seldom, or when there is not yet a developed market, one can not use this approach to price options on the underlying asset.

Nonetheless, we can use the indifference pricing theory (IPT) to find a range a possible agreement prices for the contingent claim corresponding to the minimum price for the seller and the maximum price for the buyer (see Davis et al. [2010]). The idea being that the buyer of the option wants the net present value (NPV) of his investment to be greater or equal to his initial wealth utility. Doing so, he will not loose money by entering into the transaction (similarly for the seller with the NPV being less or equal to his initial wealth utility). We identify the increasing concave functions $u_b : \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ and $u_s : \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ as respectively the buyer and the seller utility functions. We also define $w_b \geq 0$ and $w_s \geq 0$ as their respective initial wealth. Moreover, given the payoff $C(X)$ and assuming $k = 1$ unit of claim, we define $\pi_b(C) \geq 0$ and $\pi_s(C) \geq 0$ as indifference buyer or seller price that we will try to estimate. We can then solve the equations

$$E_P[u_b(w_b)] = E_P[u_b\big(w_b + C(X) - \pi_b(X)\big)]$$
$$E_P[u_s(w_s)] = E_P[u_s\big(w_s + C(X) - \pi_s(X)\big)]$$

where $w_l + C(X) - \pi_l(X)$ for $l = b, s$ is the wealth after the event. Note, we can define the rate r_l as

$$r_l = \frac{w_l + C(X) - \pi_l(X)}{w_l}, l = b, s$$

We can then find analytically, or numerically, the two prices $\pi_b(c)$ and $\pi_s(C)$ giving us the range of prices that can be settled between the two counterparts (see Carmona et al. [2009]). Further, given utility functions for the two parties, we apply Marginal Utility theory to estimate the optimum quantity traded.

F.5.4.1 Computing indifference prices

We define the processes X_s and X_b corresponding to the view of the seller and buyer, respectively, and get the indifference prices

$$E_P[u_b(w_b)] = E_P[u_b\big(w_b + C(X_b) - \pi_b(X_b)\big)]$$
$$E_P[u_s(w_s)] = E_P[u_s\big(w_s + C(X_s) - \pi_s(X_b)\big)]$$

We consider the exponential utility:

$$u(x) = 1 - e^{-\lambda x}, \forall x \in \mathbb{R}$$

where $\lambda > 0$ is the coefficient of risk aversion representing the counterpart vision of risk. This utility function is widely used in the literature and has no constraint on the sign of the cash-flow. Further, the indifference prices over this utility function are independent of the initial wealth, simplifying the number of parameters to include in the computation of the prices.

Buyer indifference price Using the exponential utility function with the coefficient of risk aversion λ_b, we have:

$$E_P[1 - e^{-\lambda_b w_b}] = E_P[1 - e^{-\lambda_b(w_b + C(X_b) - \pi_b(X_b))}]$$

which becomes

$$E_P[e^{-\lambda_b w_b}] = E_P[e^{-\lambda_b w_b}]E_P[e^{-\lambda_b(C(X_b) - \pi_b(X_b))}]$$

$$1 = E_P[e^{-\lambda_b(C(X_b) - \pi_b(X_b))}]$$

Expending, we get

$$e^{-\lambda_b \pi_b(X_b)} = E_P[e^{-\lambda_b C(X_b)}]$$

The buyer indifference price becomes

$$\pi_b(X_b) = -\frac{1}{\lambda_b} \log E_P[e^{-\lambda_b C(X_b)}]$$

Seller indifference price Similarly, for the seller with an exponential utility function with a coefficient of risk aversion λ_s we have:

$$E_P[1 - e^{-\lambda_s w_s}] = E_P[1 - e^{-\lambda_s(w_s - C(X_s) + \pi_s(X_s))}]$$

which becomes

$$E_P[e^{-\lambda_s w_s}] = E_P[e^{-\lambda_s w_s}]E_P[e^{-\lambda_s(\pi_s(X_s) - C(X_s))}]$$

$$1 = E_P[e^{-\lambda_s(\pi_s(X_s) - C(X_s))}]$$

Expending, we get

$$e^{\lambda_s \pi_s(X_s)} = E_P[e^{\lambda_s C(X_s)}]$$

The seller indifference price becomes

$$\pi_s(X_s) = \frac{1}{\lambda_s} \log E_P[e^{\lambda_s C(X_s)}]$$

F.5.4.2 Computing option prices

We consider a call option price with maturity T, strike K, and discounted payoff

$$C = P_T(X_T - K)^+$$

where $P_T = P(0,T)$ is the discount factor and X_t is the underlying price at time t.

Buyer indifference price Given the indifference price of the buyer $\pi_b(X_b)$ defined in the previous section, we need to compute the expectation of the right hand side. To simplify notations, we use $X = X_b$ and $E[X] = E_P[X]$. We can then write

$$E[e^{-\lambda_b C(X)}] = E[e^{-\lambda_b P_T(X_T - K)^+}]$$

$$E[e^{-\lambda_b C(X)}] = E[e^{-\lambda_b P_T(X_T - K)\mathbb{1}_{\{X_T \geq K\}}}]$$

Expending we get,

$$E[e^{-\lambda_b C(X)}] = E[\mathbf{I}_{\{X_T < K\}} + \mathbf{I}_{\{X_T \geq K\}} e^{-\lambda_b P_T(X_T - K)}]$$

$$E[e^{-\lambda_b C(X)}] = E[\mathbf{I}_{\{X_T < K\}}] + \mathbb{E}[\mathbf{I}_{\{X_T \geq K\}} e^{-\lambda_b P_T(X_T - K)}]$$

which we can write as

$$E[e^{-\lambda_b C(X)}] = P(X_T < K) + E[\mathbf{I}_{\{X_T \geq K\}} e^{-\lambda_b P_T(X_T - K)}]$$

Assuming the underlying process X_t follows a multivariate Ornstein-Uhlenbeck process, we get

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + E[\mathbf{I}_{\{X_T \geq K\}} e^{-\lambda_b P_T(X_T - K)}]$$

which we rewrite as

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + \int_K^{+\infty} e^{-\lambda_b P_T(x-K)} f_{X_T}(x)dx$$

where $f_{X_T} : \mathbb{R} \to [0;1]$ is the density function of X_T. Further, we have $X_t \sim \mathcal{N}(m_{X_t}, \sigma_X^2)$ and so $f_{X_T}(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-m_{X_T})^2}{2\sigma_X^2}}$. Hence,

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + \int_K^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\lambda_b P_T(x-K) - \frac{(x-m_{X_T})^2}{2\sigma_X^2}} dx$$

We now want to complete the square in the exponential term. We get

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + \int_K^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{2\sigma_X^2\lambda_b P_T(x-K) + x^2 - 2xm_{X_T} + m_{X_T}^2}{2\sigma_X^2}} dx$$

which gives

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + \int_K^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{x^2 - 2x(m_{X_T} - \sigma_X^2\lambda_b P_T) - 2\sigma_X^2\lambda_b P_T K + m_{X_T}^2}{2\sigma_X^2}} dx$$

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + \int_K^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-(m_{X_T} - \sigma_X^2\lambda_b P_T))^2 - (m_{X_T} - \sigma_X^2\lambda_b P_T)^2 - 2\sigma_X^2\lambda_b P_T K + m_{X_T}^2}{2\sigma_X^2}} dx$$

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{(m_{X_T} - \sigma_X^2\lambda_b P_T)^2 + 2\sigma_X^2\lambda_b P_T K - m_{X_T}^2}{2\sigma_X^2}} \int_K^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-(m_{X_T} - \sigma_X^2\lambda_b P_T))^2}{2\sigma_X^2}} dx$$

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^4(\lambda_b P_T)^2 + 2\sigma_X^2\lambda_b P_T K - 2m_{X_T}\sigma_X^2\lambda_b P_T}{2\sigma_X^2}} \int_K^{+\infty} f_G(x)dx$$

Where $f_G(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-(m_{X_T} - \sigma_X^2\lambda_b P_T))^2}{2\sigma_X^2}}$ is the density function of a normal random variable G such that $G \sim \mathcal{N}(m_{X_T} - \sigma_X^2\lambda_b P_T, \sigma_X^2)$ and so,

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_b P_T)^2}{2} + \lambda_b P_T(K - m_{X_T})} P(G \geq K)$$

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_b P_T)^2}{2} + \lambda_b P_T(K - m_{X_T})} P(m_{X_T} - \sigma_X^2\lambda_b P_T + \sigma_X Y \geq K)$$

Where $Y \sim \mathcal{N}(0,1)$,

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_b P_T)^2}{2} + \lambda_b P_T(K - m_{X_T})} P\left(Y \geq \frac{K - m_{X_T} + \sigma_X^2\lambda_b P_T}{\sigma_X}\right)$$

$$E[e^{-\lambda_b C(X)}] = \Phi\left(\frac{K - m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2\lambda_b^2}{2} + \lambda_b(K - m_{X_T})} \Phi\left(\frac{m_{X_T} - \sigma_X^2\lambda_b - K}{\sigma_X}\right)$$

We can then deduce the buyer indifference price:

$$\pi_b(X) = -\frac{1}{\lambda_b}\log\left[\Phi\left(\frac{K-m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_b P_T)^2}{2}+\lambda_b P_T(K-m_{X_T})}\Phi\left(\frac{m_{X_T}-\sigma_X^2\lambda_b P_T - K}{\sigma_X}\right)\right]$$

Seller indifference price Given the indifference price of the seller $\pi_s(X_s)$, the computation of $\mathbb{E}[e^{-\lambda' C(X)}]$ does not depend on the sign of λ_b so that we can have $\lambda_b = -\lambda_s \ \forall \lambda_s$ and $X = X_s$. Then,

$$E[e^{\lambda_s C(X)}] = E[e^{-\lambda_b C(X)}]$$

$$E[e^{\lambda_s C(X)}] = \Phi\left(\frac{K-m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_b P_T)^2}{2}+\lambda_b' P_T(K-m_{X_T})}\Phi\left(\frac{m_{X_T}-\sigma_X^2\lambda_b P_T - K}{\sigma_X}\right)$$

$$E[e^{\lambda_s C(X)}] = \Phi\left(\frac{K-m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_s P_T)^2}{2}-\lambda_s P_T(K-m_{X_T})}\Phi\left(\frac{m_{X_T}+\sigma_X^2\lambda_s P_T - K}{\sigma_X}\right)$$

We can then conclude that the seller's indifference price $\pi_s(X)$ is given by

$$\pi_s(X) = \frac{1}{\lambda}\log\left[\Phi\left(\frac{K-m_{X_T}}{\sigma_X}\right) + e^{\frac{\sigma_X^2(\lambda_s P_T)^2}{2}-\lambda_s P_T(K-m_{X_T})}\Phi\left(\frac{m_{X_T}+\sigma_X^2\lambda_s P_T - K}{\sigma_X}\right)\right]$$

Appendix G

Some results on signal processing

As discussed by Press et al. [1992], the Fourier methods have revolutionised fields of science and engineering with the help of the fast Fourier transform (FFT) having applications to the convolution or deconvolution of data, correlation and autocorrelation, optimal filtering, power spectrum estimation, and the computation of Fourier integrals. However, in the spectral domain, one limitation of the FFT is that it represents a function as a polynomial in $z = e^{2\pi f \Delta}$, and some processes may have spectra with shapes not well represented by this form. Another limitation of all FFT methods is that they require the input data to be sampled at evenly spaced intervals. The wavelet methods inhabit a representation of function space that is neither in the temporal, nor in the spectral domain, but rather something in between. Like the FFT, the discrete wavelet transform (DWT) is a fast, linear operation operating on a data vector whose length is an integer power of two, and transforming it into a numerically different vector of the same length. The transforms being invertible and orthogonal, they can be viewed as a rotation in function space, from the input space (or time) domain, where the basis functions are the unit vector e_i, or Dirac delta functions in the continuum limit, to a different domain. While the new domain in the FFT has basis functions that are sines and cosines, in the wavelet domain they are more complicated, called wavelets. Unlike sines and cosines, individual wavelet functions are quite localised in space, but like sines and cosines they are also quite localised in frequency or characteristic scale, making large classes of functions and operators sparse when transformed into the wavelet domain.

G.1 A short introduction to Fourier transform methods

G.1.1 Some analytical formalism

A physical process can be described either in the time domain, by the values of some quantity h as a function of time $h(t)$, or in the frequency domain where the process is specified by giving its amplitude H as a function of frequency f, that is, $H(f)$ with $-\infty < f < \infty$. One can think of $h(t)$ and $H(f)$ as being two different representations of the same function represented by the Fourier transform equations

$$\begin{aligned} H(f) &= \int_{-\infty}^{\infty} h(t) e^{2\pi i f t} dt \\ h(t) &= \int_{-\infty}^{\infty} H(f) e^{-2\pi i f t} df \end{aligned}$$

where, if t is measured in seconds, then f is in cycles per second or Hertz. Note, these equations work for different units. For instance, h can be a function of position x (in meters), in which case H will be a function of inverse wavelength (cycles per meter). Considering the angular frequency w given in radians per sec, the relation between w and f as well as $H(w)$ and $H(f)$ is

$$w = 2\pi f \ , H(w) = [H(f)]_{f=\frac{w}{2\pi}}$$

so that the above equations become

$$\begin{aligned} H(w) &= \hat{h}(w) = \int_{-\infty}^{\infty} h(t)e^{iwt}dt \\ h(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} H(w)e^{-iwt}dw \end{aligned}$$

While in the time domain the function $h(t)$ may have special symmetries, be purely real or purely imaginary, even or odd, in the frequency domain these symmetries lead to relationships between $H(f)$ and $H(-f)$. Further, using the symbol $\Longleftrightarrow$ to indicate transform pairs, the Fourier transform has the following elementary properties. If

$$h(t) \Longleftrightarrow H(f)$$

is a transform pair, then the other ones are

- time scaling
$$h(at) \Longleftrightarrow \frac{1}{|a|}H(\frac{f}{a})$$
- frequency scaling
$$\frac{1}{|b|}h(\frac{t}{b}) \Longleftrightarrow H(bf)$$
- time shifting
$$h(t-t_0) \Longleftrightarrow H(f)e^{2\pi i f t_0}$$
- frequency shifting
$$h(t)e^{-2\pi i f_0 t} \Longleftrightarrow H(f-f_0)$$

Given two functions $h(t)$ and $g(t)$ and their corresponding Fourier transforms $H(f)$ and $G(f)$ we can look at two special combinations. The convolution of the two functions defined by

$$g * h = \int_{-\infty}^{\infty} g(\tau)h(t-\tau)d\tau$$

is a member of the simple transform pair

$$g * h \Longleftrightarrow G(f)H(f) \text{ convolution theorem}$$

stating that the Fourier transform of the convolution is just the product of the individual Fourier transforms. The correlation of two functions defined by

$$Corr(g,h) = \int_{-\infty}^{\infty} g(\tau+t)h(\tau)d\tau$$

is a member of the transform pair

$$Corr(g,h) \Longleftrightarrow G(f)H^*(f) \text{ correlation theorem}$$

in the case where g and h are real functions. That is, multiplying the Fourier transform of one function by the complex conjugate of the Fourier transform of the other gives the Fourier transform of their correlation. The autocorrelation being the correlation of a function with itself, the transform pair becomes

$$Corr(g,g) \Longleftrightarrow |G(f)|^2 \text{ Wiener-Khinchin theorem}$$

The total power in a signal is the same if we compute it in the time domain or in the frequency domain, leading to the Parseval's theorem

$$\text{Total Power } = \int_{-\infty}^{\infty} |h(t)|^2 dt = \int_{-\infty}^{\infty} |H(f)|^2 df$$

When interested in the amount of power contained in the frequency interval $[f, f+df]$, the one-sided power spectral density (PSD) of the function h is

$$P_h(f) = |H(f)|^2 + |H(-f)|^2 \, , 0 \leq f < \infty \tag{G.1.1}$$

so that the total power is just the integral of $P_h(f)$ from $f = 0$ to $f = \infty$. In the case where the function $h(t)$ is real, the two terms above are equal, and we get

$$P_h(f) = 2|H(f)|^2$$

Note, in some cases, PSDs are defined without this factor two, and are called two-sided power spectral densities.

***Remark* G.1.1** *When the function $h(t)$ goes endlessly from $-\infty < t < \infty$, then its total power and power spectral density will, in general, be infinite. Of interest is then the power spectral density per unit time, which is computed by taking a long, but finite, stretch of the function $h(t)$, and then dividing the resulting PSD by the length of the stretch used.*

Parseval's theorem in this case states that the integral of the one-sided PSD per-unit-time, which is a function of frequency f, converges as one evaluates it using longer and longer stretches of data. To conclude, we are rarely given a continuous function $h(t)$ to work with, but rather a list of measurements of $h(t_i)$ for a discrete set of t_i.

G.1.2 The Fourier integral

Around 1807, the french mathematician Joseph Fourier asserted that any 2π periodic function could be represented by superposition of sines and cosines. We let h be a function of position x and assume that $h(x)$ is a periodic function of period 2π that can be represented by a trigonometric series

$$h(x) = a_0 + \sum_{n=1}^{\infty} \big(a_n \cos nx + b_n \sin nx\big)$$

that is, we assume that this series converges and has $h(x)$ as its sum. A classical problem is to determine the coefficients a_n and b_n (called the Fourier coefficients of $h(x)$) of that series called the Fourier series of $h(x)$. Periodic functions in applications rarely have period 2π but some other period $p = 2L$. If such a function $h(x)$ has a Fourier series, it is of the form

$$h(x) = a_0 + \sum_{n=1}^{\infty} \big(a_n \cos \frac{n\pi}{L} x + b_n \sin \frac{n\pi}{L} x\big)$$

with the Fourier coefficients of $h(x)$ given by the Euler formulas

$$\begin{aligned} a_0 &= \frac{1}{2L}\int_{-L}^{L} h(x)dx \\ a_n &= \frac{1}{L}\int_{-L}^{L} h(x)\cos\frac{n\pi x}{L}dx \text{ for } n = 1, 2, .. \\ b_n &= \frac{1}{L}\int_{-L}^{L} h(x)\sin\frac{n\pi x}{L}dx \text{ for } n = 1, 2, .. \end{aligned}$$

Fourier series are powerful tools for solving various problems involving periodic functions. However, many practical problems do not involve periodic functions and it becomes natural to generalise the method of Fourier series to include nonperiodic functions. We consider any periodic function $f_L(x)$ of period $2L$ which can be represented with Fourier series as

$$h_L(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\omega_n x + b_n \sin\omega_n x$$

where $\omega_n = \frac{n\pi}{L}$. We let the Fourier coefficients be given by the Euler formulas above and re-write the Fourier series as

$$\begin{aligned} h_L(x) &= \frac{1}{2L}\int_{-L}^{L} h_L(v)dv + \frac{1}{L}\sum_{n=1}^{\infty}\big(\cos\omega_n x\int_{-L}^{L} h_L(v)\cos\omega_n v dv \\ &+ \sin\omega_n x\int_{-L}^{L} h_L(v)\sin\omega_n v dv\big) \end{aligned}$$

We then let $\Delta\omega = \omega_{n+1} - \omega_n = \frac{\pi}{L}$ such that $\frac{1}{L} = \frac{\Delta\omega}{\pi}$ and the Fourier series become

$$\begin{aligned} h_L(x) &= \frac{1}{2L}\int_{-L}^{L} h_L(v)dv + \frac{1}{\pi}\sum_{n=1}^{\infty}\big(\cos\omega_n x\Delta\omega\int_{-L}^{L} h_L(v)\cos\omega_n v dv \\ &+ \sin\omega_n x\Delta\omega\int_{-L}^{L} h_L(v)\sin\omega_n v dv\big) \end{aligned}$$

which is valid for any finite value of L. We now let $L \to \infty$ and assume that the nonperiodic function

$$h(x) = \lim_{L\to\infty} h_L(x)$$

is absolutely integrable on the x-axis, that is, the integral

$$\int_{-\infty}^{\infty} |h(x)|dx \tag{G.1.2}$$

exists. Therefore, $\frac{1}{L} \to 0$ and $\Delta\omega \to 0$ so that the infinite series become the integral

$$h(x) = \frac{1}{\pi}\int_0^{\infty}\big(\cos\omega x\int_{-\infty}^{\infty} h(v)\cos\omega v dv + \sin\omega x\int_{-\infty}^{\infty} h(v)\sin\omega v dv\big)d\omega$$

Introducing the notation

$$\begin{aligned} A(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} h(v)\cos\omega v dv \\ B(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} h(v)\sin\omega v dv \end{aligned}$$

we can then represent the function $h(x)$ in terms of the Fourier integral by

$$h(x) = \int_0^{\infty} \big(A(\omega)\cos\omega x + B(\omega)\sin\omega x\big)d\omega$$

Sufficient conditions for the validity of the Fourier integral are

Theorem G.1.1 *If $h(x)$ is piecewise continuous in every finite interval and has a right-hand derivative and a left-hand derivative at every point and if the integral (G.1.2) exists, then $h(x)$ can be represented by a Fourier integral. At a point where $h(x)$ is discontinuous the value of the Fourier integral equals the average of the left- and right-hand limits of $h(x)$ at that point.*

G.1.3 The Fourier transformation

The Fourier transformation is obtained from the Fourier integral in complex form. So, we first consider the complex form of the Fourier integral. We defined in Section (G.1.2) the real Fourier integral and use the property of even function on the cosine and the property of odd function on the sine, getting

$$h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\big(\int_{-\infty}^{\infty} h(v)\cos(\omega x - \omega v)dv\big)d\omega$$

We now use the Euler formula $e^{it} = \cos t + i\sin t$ for the complex exponential function. We set $t = (\omega x - \omega v)$ obtaining

$$h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(v)e^{i\omega(x-v)}dvd\omega$$

which is the complex Fourier integral. We can now express this integral as a product of exponential functions, getting

$$h(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\big(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(v)e^{-i\omega v}dv\big)e^{i\omega x}d\omega$$

We call the Fourier transform of h the expression of ω in brackets and denote it by $\hat{h}(\omega)$. Writing $v = x$, we get

$$\hat{h}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} h(x)e^{-i\omega x}dx$$

so that $h(x)$ is the inverse Fourier transform of $\hat{h}(\omega)$ given by

$$h(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{h}(\omega)e^{i\omega x}d\omega$$

if $\hat{h}(\omega) \in L^1(\mathbb{R})$ [1]. Another notation is $\mathcal{H}(h) = \hat{h}(\omega)$ and $\mathcal{H}^{-1}$ for the inverse. We are now going to mention without proof the sufficient conditions for the existence of the Fourier transform $\mathcal{H}(f)$ of a function $h(x)$.

[1] Recall, $h \in L^p(I)$ if $\int_I |h(t)|^p dt$ exists and is bounded.

***Proposition* 21** *The following two conditions are sufficient for the existence of the Fourier transform of a function $h(x)$ defined on the x-axis.*

- *$h(x)$ is piecewise continuous on every finite interval*
- *$h(x)$ is absolutely integrable on the x-axis [2] or put another way the integral in Equation (G.1.2) exists and is bounded.*

For the Fourier transform of $h(x)$ to exist and have an inverse it is enough that the function $h(x)$ be square integrable, that is, $h(x) \in L^2(\mathbb{R})$. We get $\mathcal{H}^{-1}\mathcal{H}(h) = h$ but this inversion formula holds as well in other cases.

G.1.4 The discrete Fourier transform

Assuming $h(t)$ to be a function sampled at evenly spaced intervals Δ in time, we get the sequence

$$h_n = h(n\Delta) \, , n = .., -1, 0, 1, ..,$$

For any sampling interval Δ, there is a special frequency f_c, called the Nyquist critical frequency, given by

$$f_c = \frac{1}{2\Delta}$$

One can measure time in units of the sampling interval Δ to get $f_c = \frac{1}{2}$. The sampling theorem states that if $H(f) = 0$ for all $|f| \geq f_c$, then the function $h(t)$ is completely determined by its sample h_n. However, all of the power spectral density lying outside of the frequency range $-f_c < f < f_c$ is spuriously moved into that range, which is called aliasing. One can overcome aliasing by knowing the natural bandwidth limit of the signal and then sample at a rate sufficiently rapid to give at least two points per cycle of the highest frequency present. Given N consecutive sampled values

$$h_k = h(t_k) \, , t_k = k\Delta \, , k = 0, 1, ..., N-1$$

with sampling interval Δ, we assume N is even. Rather than estimating the Fourier transform $H(f)$ at all values of f in the range $-f_c$ to f_c we do it only at the discrete values

$$f_n = \frac{n}{N\Delta} \, , n = -\frac{N}{2}, .., \frac{N}{2} \tag{G.1.3}$$

where $n = -\frac{N}{2}$ and $n = \frac{N}{2}$ correspond to the lower and upper limits of the Nyquisit critical frequency range $-f_c$ and f_c. While we have $N+1$ values (including 0), the two extreme values of n are not independent, and we are left with N independent values. The integral can now be approximated by the discrete sum

$$H(f_n) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f_n t} dt \approx \sum_{k=0}^{N-1} h_k e^{2\pi i f_n t_k} \Delta = \Delta \sum_{k=0}^{N-1} h_k e^{2\pi i k \frac{n}{N}}$$

which is the discrete Fourier transform of the N points h_k. Hence, the DFT H_n given by

$$H_n = \sum_{k=0}^{N-1} h_k e^{2\pi i k \frac{n}{N}}$$

maps N complex numbers (the h_k) into N complex numbers (the H_n), so that we get

[2] that is the following limits exist

$$\lim_{a \to -\infty} \int_a^0 |h(x)| dx + \lim_{b \to \infty} \int_0^b |h(x)| dx$$

$$H(f_n) \approx \Delta H_n$$

Since H_n is periodic in n with period N, we get $H_{-n} = H_{N-n}$ for $n = 1, 2, ...$ and we can let the index n varies from 0 to $N-1$ (one complete period) so that n and k (in h_k) vary exactly over the same range. Using the discrete frequency in Equation (G.1.3), the zero frequency corresponds to $n = 0$, the value $n = \frac{N}{2}$ corresponds to both $f = f_c$ and $f = -f_c$, positive frequencies $0 < f < f_c$ correspond to values $1 \leq n \leq \frac{N}{2} - 1$, and negative frequencies $-f_c < f < 0$ correspond to $\frac{N}{2} + 1 \leq n \leq N - 1$. Note, the DFT has symmetry properties almost exactly the same as the continuous Fourier transform. Hence, the formula for the discrete inverse Fourier transform is

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-2\pi i k \frac{n}{N}}$$

The only differences between the DFT and its discrete inverse are

- changing the sign in the exponential
- dividing the answer by N

so that computing the DFT we recover its inverse transform.

G.1.5 The Fast Fourier Transform algorithm

We let $h(.)$ be an appropriate function with enough regularities. The Fourier transform of that function is

$$H(u) = \int_{-\infty}^{\infty} e^{iux} h(x) dx$$

where $u = 2\pi f$ is the angular frequency given in radians per sec, and x is a location. As explained in Appendix (G.1.4), we can estimate that transform from a finite number of sampled points N, but to avoid classifying frequency from $f = 0$, $0 < f < f_c$, and $-f_c < f < 0$ we are considering another discretisation. We let $h_p = h(x_p)$ with $x_p = (\frac{N}{2} - p)\Delta$ for $p = 0, .., N-1$, where Δ is the sampling interval. Note, in this notation x_p goes from $(-\frac{N}{2} + 1)\Delta$ to $\frac{N}{2}\Delta$ giving a total of N elements. We assume that the number of iteration, N, is a power of 2, and choose Δ and N so that $\Delta << 1$ and $N\Delta >> 1$. We can then re-write the Fourier transform as

$$H(u) \approx \int_{-\alpha\frac{N}{2}}^{\alpha\frac{N}{2}} e^{iux} h(x) dx \approx \sum_{p=-\frac{N}{2}}^{\frac{N}{2}-1} e^{iux_p} h(x_p)\Delta$$

so that x_p goes from μ to $N\Delta$. Since the function $h(.)$ is sampled over N points, the output transform $H(.)$ must also be sampled over N points. Therefore, we will discretise the Fourier transform in $u_q = \frac{2\pi(\frac{N}{2} - q)}{N\Delta}$ for $q = 0, .., N-1$ such that u_q goes from $2\pi\frac{1}{2\Delta}$ to $2\pi(-\frac{1}{2\Delta} + \frac{1}{N\Delta})$ where $\frac{1}{N\Delta}$ is the step of the frequency. Note, we could as well have considered the discretisation $x_p = (p - \frac{N}{2})\Delta$ and $u_q = \frac{2\pi(q - \frac{N}{2})}{N\Delta}$ for $p, q = 0, .., N-1$. Then, given

$$iu_q x_p = \pi i(\frac{N}{2} - p) - i\pi q + \frac{2\pi i}{N} pq$$

the Fourier transform becomes

$$H(u_q) = \sum_{p=-\frac{N}{2}}^{\frac{N}{2}-1} e^{iu_q x_p} h(x_p)\Delta = e^{-i\pi q} \sum_{p=-\frac{N}{2}}^{\frac{N}{2}-1} e^{\frac{2\pi i}{N} pq} h(x_p) e^{\pi i(\frac{N}{2} - p)}\Delta$$

So, if we define $A_p = h(x_p)e^{i\pi(\frac{N}{2}-p)}\Delta$, $H_q = e^{i\pi q}H(u_q)$ and $W_N^{pq} = e^{\frac{2\pi i}{N}pq}$ we get the system of equations

$$\begin{bmatrix} H_0 \\ ... \\ ... \\ H_{N-1} \end{bmatrix} = \begin{bmatrix} W_N^{0,0} & ... & ... & W_N^{0,N-1} \\ ... & ... & ... & 0 \\ ... & ... & ... & ... \\ W_N^{N-1,0} & ... & ... & W_N^{N-1,N-1} \end{bmatrix} \begin{bmatrix} A_0 \\ ... \\ ... \\ A_{N-1} \end{bmatrix}$$

and we see that the calculation with this method is of order $0(N^2)$. However, we can reduce the order of computation by using the FFT algorithm, getting the Fourier transform as

$$(H_q)_{(0\leq q\leq N-1)} = \left[FFT(N, (A_p)_{(0\leq p\leq N-1)})\right]_{(0\leq q\leq N-1)}$$

The main idea behind the FFT is to write this sum of length N as the sum of two Fourier transforms each of them being of length $\frac{N}{2}$ (see Danielson and Lanczos). So we get $W_N^{2pq} = e^{\frac{2\pi i}{N}pq} = e^{\frac{2\pi i}{N/2}pq} = W_{\frac{N}{2}}^{pq}$ and also $W_N^{p(q+N)} = e^{\frac{2\pi i}{N}p(q+N)} = e^{\frac{2\pi i}{N}pq+2\pi ip} = W_N^{pq}$ so that $W_N^{p(q+\frac{N}{2})} = -W_N^{pq}$. Therefore, the Fourier transform becomes

$$\begin{aligned} H_q &= \sum_{p=0}^{N-1} W_N^{pq}A_p = \sum_{p=0}^{N-1} W_N^{2pq}A_{2p} + \sum_{p=0}^{N-1} W_N^{(2p+1)q}A_{2p+1} \\ &= \sum_{p=0}^{N-1} W_{\frac{N}{2}}^{pq}A_{2p} + W_N^q \sum_{p=0}^{N-1} W_N^{2pq}A_{2p+1} \\ &= \sum_{p=0}^{N-1} W_{\frac{N}{2}}^{pq}A_{2p} + W_N^q \sum_{p=0}^{N-1} W_{\frac{N}{2}}^{pq}A_{2p+1} \end{aligned}$$

Let's now define $(B_p)_{(0\leq p\leq N-1)}$ and $(C_p)_{(0\leq p\leq N-1)}$ such that $B_p = A_{2p}$ and $C_p = A_{2p+1}$. Then, if $0 \leq q \leq \frac{N}{2} - 1$, the Fourier transform re-write

$$\begin{aligned} H_q &= \sum_{p=0}^{\frac{N}{2}-1} W_{\frac{N}{2}}^{pq}B_p + W_N^q \sum_{p=0}^{\frac{N}{2}-1} W_{\frac{N}{2}}^{pq}C_p \\ &= \left[FFT(\frac{N}{2}, (B_p)\right]_q + W_N^q\left[FFT(\frac{N}{2}, (C_p)\right]_q \end{aligned}$$

while if $\frac{N}{2} \leq q \leq N-1$

$$\begin{aligned} H_q &= H_{k+\frac{N}{2}} = \sum_{p=0}^{\frac{N}{2}-1} W_{\frac{N}{2}}^{pk}B_p + W_N^q \sum_{p=0}^{\frac{N}{2}-1} W_{\frac{N}{2}}^{pk}C_p \\ &= \left[FFT(\frac{N}{2}, (B_p)\right]_{q-\frac{N}{2}} + W_N^q\left[FFT(\frac{N}{2}, (C_p)\right]_{q-\frac{N}{2}} \end{aligned}$$

which can then be computed recursively using the symmetry in the calculation of the factors W_N^{pq}. Note, $FFT(\frac{N}{2}, (B_p)$ is the Fourier transform of length $\frac{N}{2}$ from the even components, and $FFT(\frac{N}{2}, (C_p)$ is the corresponding transform of length $\frac{N}{2}$ formed from the odd components. This solution is recursive, as we can reduce $FFT(\frac{N}{2}, (B_p)$ by computing the transform of its $\frac{N}{4}$ even-numbered input data and $\frac{N}{4}$ odd-numbered data, hence the necessity for N to be an integer power of 2. This process is repeated until we have subdivided the data all the way down to transforms of length one. Hence, there is a one-point transform that is just one of the input numbers u_q, and using bit reversal one can find out the q corresponding to the right equation.

G.2 From spline analysis to wavelet analysis

The need for a continuous signal representation comes up every time we need to implement numerically an operator initially defined in the continuous domain. The sampling theory introduced by Shannon [1949] provide a solution to this problem. It describes an equivalence between a band-limited function and its equidistant samples taken at a frequency that is superior or equal to the Nyquist rate. However, this approaches faces a lot of problems. An alternative approach is to use splines introduced by Schoenberg [1946], which offer many practical advantages. One of the main advantage of using splines is that we can always obtain a continuous representation of a discrete signal by fitting it with a spline in one or more dimensions. The fit may be exact (interpolation) or approximate (least-squares or smoothing splines). While splines developed in various field (computer science), there was little crossover to signal processing until the development of wavelet theory (see Mallat [1989]). Since we have extensively discussed wavelet analysis to filter and forecast time series, we are now going to introduce splines in view of relating the two methods. We will follow Unser [1999] who provided a clear tutorial on splines.

G.2.1 An introduction to splines

Splines are piecewise polynomials with pieces that are smoothly connected together, where joining points are called knots. For a spline of degree n, each segment is a polynomial of degree n, but imposing the continuity of the spline and its derivatives up to the order $(n-1)$ at the knots, there is only one degree of freedom per segment. Considering only splines with uniform knots and unit spacing, they are uniquely characterised in terms of B-spline expansion

$$s(x) = \sum_{k\in\mathbb{Z}} c(k)\beta^n(x-k) \tag{G.2.4}$$

which involves the integer shifts of the central B-spline of degree n denoted by $\beta^n(x)$. We can construct symmetrical, bell-shaped functions B-spline from the $(n+1)$-fold convolution of a rectangular pulse β^0 given by

$$\beta^0(x) = \begin{cases} 1 \text{ if } -\frac{1}{2} < x < \frac{1}{2} \\ \frac{1}{2} \text{ if } |x| = \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

with

$$\beta^n = \beta^0 * \beta^0 * ... * \beta^0(x) \text{ , } (n+1) \text{ times}$$

Since the B-spline in Equation (G.2.4) is linear, we can easily study the properties of the basic atoms. Further, each spline is characterised by its sequence of coefficients $c(k)$, giving the spline the convenient structure of a discrete signal, even though the underlying model is continuous. Using β^0 we can construct the cubic B-spline as

$$\beta^3(x) = \begin{cases} \frac{2}{3} - |x|^2 + \frac{|x|^3}{2} \text{ if } 0 < |x| < 1 \\ \frac{(2-|x|)^3}{6} \text{ if } 1 \leq |x| < 2 \\ 0 \text{ if } 2 \leq |x| \end{cases}$$

which is used for performing high-quality interpolation. The interpolation problem consists in determining the B-spline model of a given input signal $s(k)$, where the coefficients are determined such that the function goes through the data points exactly. For degrees $n = 0$ and $n = 1$ the coefficients are identical to the signal samples, $c(k) = s(k)$, but it is not the case for higher degrees. One can solve the problem by using digital filtering techniques. We consider the B-spline kernel b_m^n obtained by sampling the B-spline of degree n expanded by a factor m, that is,

$$b_m^n = \beta^n(\frac{x}{m})\big|_{x=k} \leftrightarrow B_m^n(z) = \sum_{k\in\mathbb{Z}} b_m^n(k)z^{-k}$$

Then, given the signal samples $s(k)$, we want to estimate the coefficients $c(k)$ from Equation (G.2.4) such that we have a perfect fit at the integers. That is, we want to solve

$$\sum_{l\in\mathbb{Z}} c(l)\beta^n(x-l)\big|_{x=k} = s(k)$$

Using the discrete B-splines, we can rewrite this constraint in terms of a convolution

$$s(k) = (b_1^n * c)(k)$$

and defining the inverse convolution operator

$$(b_1^n)^{-1}(k) \leftrightarrow \frac{1}{B_1^n(z)}$$

the solution is found by inverse filtering

$$c(k) = (b_1^n)^{-1} * s(k)$$

We are now going to define the cardinal spline basis functions that are the spline analogs of the $sinc$ function in the approach for band-limited functions. We have

$$s(x) = \sum_{k\in\mathbb{Z}} s(k)\eta^n(x-k)$$

where the cardinal spline of degree n is

$$\eta^n(x) = \sum_{k\in\mathbb{Z}} (b_1^n)^{-1}(k)\beta^n(x-k)$$

which provides a spline interpolation formula using the signal values coefficients.

In order to introduce spline sampling theory, we first define a general spline generating function

$$\phi(x) = \sum_{k\in\mathbb{Z}} p(k)\beta^n(x-k) \tag{G.2.5}$$

with the restriction that the sequence p is such that the integer translations of ϕ form a basis of the basic spline space. Two important special cases are the B-spline with $p(k) = \delta(k)$, and the cardinal spline with $p = (b_1^n)^{-1}$. Since we want to vary the sampling step, we define the spline space of degree n with step size T by rescaling the model in Equation (G.2.4) as

$$S_T^n = \{s_T(x) = \sum_{k\in\mathbb{Z}} c(k)\phi(\frac{x}{T} - k); c(k) \in l_2\}$$

That is, we take linear combinations (with finite energy) of the spline basis functions ϕ rescaled by a factor T and spaced accordingly. We now want to approximate an arbitrary signal $s(x)$ with a spline $s_T \in S_T^n$ using the L_2-norm $||s - S_T||_2$ induced by the L_2 inner product [3]. According to this criterion, the minimum error approximation of $s(x) \in L_2$ in S_T^n is given by its orthogonal projection onto S_T^n. As a result, the coefficients of the best approximation (least-squares solution) are given by

$$c_T(k) = \frac{1}{T} < s(x), \dot{\phi}(\frac{x}{T} - k) > \tag{G.2.6}$$

where $\dot{\phi}(x) \in S_1^n$ is the dual of $\phi(x)$, in the sense that

[3] $< f, g >= \int_{-\infty}^{\infty} f^*(x)g(x)dx$ and $||f||_2 =< f, f >^{\frac{1}{2}}$

$$< \dot{\phi}(x-k), \phi(x-l) >= \delta(k-l)$$

(bi-orthogonality condition). Note, the process of performing a least-square spline approximation of a signal is linked to that of obtaining its band-limited representation using the standard sampling procedure. The only difference being in the choice of the appropriate analog pre-filter. Since we are performing an orthogonal projection, the approximation error will be non-zero unless the signal is already included in the approximation space. The error can be controlled by choosing a sufficiently small sampling step T. In approximation theory, a fundamental result states that the rate of decay L of the error as a function of T depends on the ability of the representation to produce polynomials of degree $n = L - 1$. The approximation error also depends on the bandwidth of the signal. The relevant measure is

$$||s^{(L)}|| = (\frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^{2L}|\hat{s}(\omega)|^2 d\omega)^{\frac{1}{2}}$$

where $\hat{s}(\omega)$ is the Fourier transform of s. Hence, it is the norm of the L-th derivative of s. The key result from the Strang-Fix theory of approximation is the following error bound (see Strang et al. [1971])

$$\forall s \in W_2^L \ , ||s - P_T s|| \leq C_L T^L ||s^{(L)}||$$

where $P_T s$ is the least-squares spline approximation of s at sampling step T, C_L is a known constant, and W_2^L denotes the space of functions being L-times differentiable in the L_2 space. Hence, the error will decay like $\mathcal{O}(T^L)$, where the order $L = n + 1$ is one more than the degree n. That is, spline interpolation gives the same rate of decay as the least-squares approximation in Equation (G.2.6), but with a larger leading constant.

G.2.2 Multiresolution spline processing

If we dilate a spline by a factor m with knots at the integers, the resulting function is also a spline with respect to the initial integer grid. That is, for scale-invariance to hold, we need the spline knots to be positioned on the integers. This observation is the key to the multiresolution properties of splines, making them perfect candidates for the construction of wavelets and pyramids. Hence, we consider the shifted causal B-splines

$$\phi^n(x) = \beta^n(x - \frac{n+1}{2})$$

having the required property. Similarly to the centred B-spline $\beta^n(x)$, we can construct $\phi^n(x)$ from the $(n+1)$-fold convolution of ϕ^0, the indicator function in the unit interval. That is, $\phi^0(\frac{x}{m})$ which is 1 for $x \in [0, m)$ and 0 otherwise, can be written as

$$\phi^0(\frac{x}{m}) = \sum_{k=0}^{m-1} \phi^0(x-k) = \sum_{k\in\mathbb{Z}} h_m^0(k)\phi^0(x-k)$$

where $h_m^0(k)$ is the filter whose z-transform is $H_m^0(z) = \sum_{k=0}^{m-1} z^{-k}$ (discrete pulse of size m). By convolving this equation with itself $(n+1)$-times and performing the appropriate normalisation, we get

$$\phi^n(\frac{x}{m}) = \sum_{k\in\mathbb{Z}} h_m^n(k)\phi^n(x-k) \tag{G.2.7}$$

where

$$H_m^n(z) = \frac{1}{m^n}\big(H_m^0(0)\big)^{n+1} = \frac{1}{m^n}\big(\sum_{k=0}^{m-1} z^{-k}\big)^{n+1}$$

This is a two-scale equation, indicating that a B-spline of degree n dilated by m can be expressed as a linear combination of B-splines. Note, the two-scale Equation (G.2.7) holds for any integer m, and the refinement filter is the

$(n+1)$-fold convolution of the discrete rectangular impulse of width m. In the standard case where $m = 2$, $H_2^n(z)$ is the binomial filter playing a central role in the wavelet transform theory (see Strang et al. [1996]). The filter coefficients appear in the Pascal triangle, and in the case of the B-spline of degree 1 we get the third line of Pascal's triangle. When constructing multiscale representations of signals, or pyramids, we consider scaling factors that are powers of two. The implication of the two-scale relation for $m = 2$ is that the spline subspaces S_m^n, with $m = 2^i$, are nested $S_1^n \supset S_2^n \supset ... \supset S_{2^i}^n ...$. If we let $P_{2^i} s = s_i$ be the minimum error approximation of some continuously defined signal $s(x) \in L_2$ at the scale $m = 2^i$, we get

$$P_{2^i} s = \sum_{k \in \mathbb{Z}} c_{2^i}(k) \phi(\frac{x}{2^i} - k)$$

where $\phi(\frac{x}{2^i} - k)$ are the spline basis functions at the scale $m = 2^i$. Note, one implication of the nested property is that the coefficients $c_{2^i}(k)$, formally computed from Equation (G.2.6), can be computed iteratively using a combination of discrete pre-filtering and down-sampling operations.

***Remark* G.2.1** *The key observation is that we can obtain $P_{2^i} s = s_i$ if we simply re-approximate s_{i-1} at the next finer scale, $P_{2^i} s = P_{2^i} s_{i-1}$.*

Thus, we may compute the expansion coefficients as

$$c_{2^i}(k) = \frac{1}{2^i} < \sum_{l \in \mathbb{Z}} c_{2^{i-1}} \phi(\frac{x}{2^{i-1}} - l), \dot{\phi}(\frac{x}{2^i} - k) >$$

Using the two-scale relation to precompute the sequence of inner products

$$\dot{h}(k) = \frac{1}{2^i} < \phi(\frac{x}{2^{i-1}} + k), \dot{\phi}(\frac{x}{2^i} - k) > = < \phi(x + k), \dot{\phi}(\frac{x}{2}) >$$

one can show that $c_{2^i}(k)$ are evaluated by pre-filtering with $\dot{h}$ and down-sampling by a factor of 2

$$c_{2^i}(k) = (\dot{h} * c_{2^{i-1}})(2k)$$

Note, rather than minimising the continuous L_2 error, we can also construct spline pyramids that are optimal in the discrete l_2 norm by performing a small modification of the reduction filter $\dot{h}$. This technique is known as spline regression in statistics. Most of the spline pyramids use symmetric filters centred on the origin.

The L_2 spline pyramid described above has all the required properties for a multiresolution analysis of L_2 in the sense defined by Mallat [1989]. In the wavelet theory, the multiresolution analysis is dense in $L2$, so that we can construct the associated wavelet bases of L_2 with no difficulty, and obtain an efficient, non-redundant, way of representing the difference images. Since image reduction is achieved by repeated projection, the difference between two successive signal approximations $P_{2^{i-1}} f$ and $P_{2^i} f$ belong to the subspace $W_{2^i}^n$. It is the complement of $S_{2^i}^n$ with respect to $S_{2^{i-1}}^n$, that is,

$$S_{2^{i-1}}^n = S_{2^i}^n \oplus W_{2^i}^n$$

with $S_{2^i}^n \cap W_{2^i}^n = \{0\}$. Then, the wavelet $\psi(x)$ generate the basis functions of the residual spaces

$$W_{2^i}^n = Span(\psi(\frac{x}{2^i} - k))_{k \in \mathbb{Z}}$$

For some applications, it is more concise to express the residues

$$P_{2^{i-1}} f - P_{2^i} f \in W_{2^i}^n$$

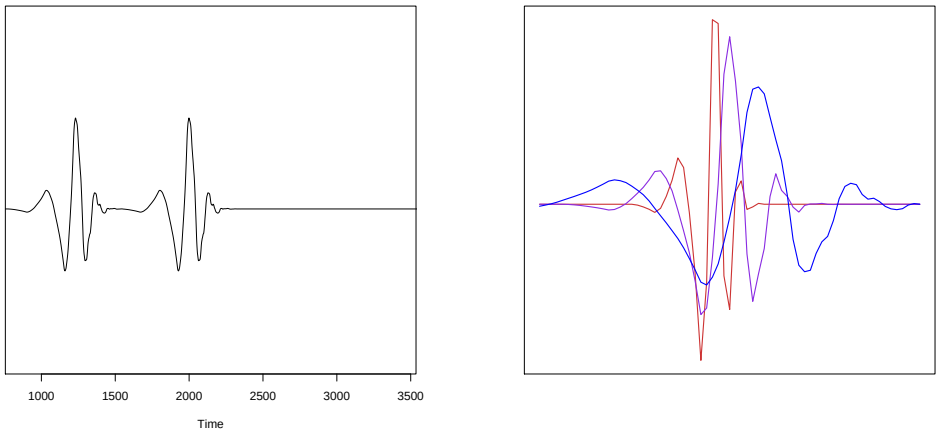
using wavelets rather than the basis functions of $V_{2^{i-1}}^n$. In wavelet theory, splines constitute a case apart because they give rise to the only wavelets having a closed-form formula (piecewise polynomial). All other wavelet bases are defined indirectly by an infinite recursion. This is why most of the earlier wavelet constructions were based on splines. For instance, the Haar wavelet transform $(n = 0)$ (see Haar [1910]), the Franklin system $(n = 1)$, Stromberg's one-sided orthogonal splines, the Battle-Lemarie wavelets (see Battle [1987]). There are now several other sub-classes of spline wavelets available differing in the type of projection used and in their orthogonality properties. For instance, the class of semi-orthogonal wavelets, which are orthogonal with respect to dilation (see Unser et al. [1993]). They span the same space as the Battle-Lemarie splines, but they are not constrained to be orthogonal. Of particular interest are the B-spline wavelets, which are compactly supported and optimally localised in time and frequency (see Unser et al. [1992]). The only downside of semi-orthogonal wavelets is that some of the corresponding wavelet filters are IIR. While it is not a serious problem in practice due to fast recursive algorithms, researchers have also designed spline wavelets such that the corresponding wavelet filters are FIR. These bi-orthogonal wavelets are constructed using two multiresolutions instead of one, with the spline spaces on the synthesis side. The major difference with the semi-orthogonal case being that the underlying projection operators are oblique rather than orthogonal. Bi-orthogonal spline wavelets are short, symmetrical, easy to implement (FIR filter bank), and very regular. Further, in that subclass, we can still orthogonalise the wavelets with respect to shifts, leading to the shift-orthogonal wavelets.

G.3 A short introduction to wavelet transform methods

G.3.1 The continuous wavelet transform

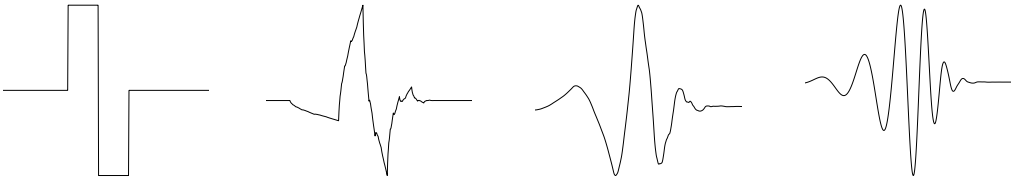
Following Addison [2002], we give an overview of the theory of wavelet transform methods focusing on continuous and discrete wavelet transforms of continuous signal. For further details and applications the readers should refer to the book. The wavelet transform is a method for converting a function (or signal) into another form either making certain features of the original signal more amenable to study, or enabling the original data set to be described more succinctly. In any case, we need a wavelet, which is a function $\psi(t)$ with certain properties, undergoing translation (movements along the time axis) (see Figure (G.1(a))) and dilation (spreading out of the wavelet) (see Figure (G.1(b))) to transform the signal into another form which unfolds it in time and scale. There are a large number of wavelets to choose from depending on both the nature of the signal and what we require from the analysis (see Figure (G.2)). For instance, all derivatives of the Gaussian function may be employed as a wavelet. As an example, the second derivative of the Gaussian distribution function $e^{-\frac{t^2}{2}}$ with unit variance but without the normalisation factor $\frac{1}{\sqrt{2\pi}}$ is called the Mexican hat wavelet defined as

$$\psi(t) = (1 - t^2)e^{-\frac{t^2}{2}}$$



(a) Wavelet translation (b) Wavelet rescaling

Figure G.1: The two properties that allow wavelet to capture signal features



(a) Haar (b) Daubechies 4 (c) Daubechies 8 (d) Daubechies 20

Figure G.2: Some examples of wavelets

A function must satisfy certain mathematical criteria to be classified as a wavelet

1. it must have finite energy
$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty$$
where E is the energy.
2. if $\hat{\psi}(f)$ is the Fourier transform of $\psi(t)$, then the following condition must hold
$$C_g = \int_0^{\infty} \frac{|\hat{\psi}(f)|^2}{f} df < \infty$$
implying that the wavelet $\psi(t)$ must have a zero mean. This is the admissibility condition and C_g is the admissibility constant.
3. in case of complex wavelets, the Fourier transform must both be real and vanish for negative frequencies.

Wavelets satisfying the admissibility condition are bandpass filters, letting through only signal components within a finite range of frequencies (passband) and in proportion characterised by the energy spectrum of the wavelet. A plot of the squared magnitude of the Fourier transform against frequency for the wavelet gives its energy spectrum. The peak of the energy spectrum occurs at a dominant frequency of $f_p = \pm\frac{\sqrt{2}}{2\pi}$. The second moment of area of the energy spectrum is used to define the passband centre of the energy spectrum

$$f_c = \sqrt{\frac{\int_0^\infty f^2|\hat{\psi}(f)|^2 df}{\int_0^\infty |\hat{\psi}(f)|^2 df}}$$

where f_c is the standard deviation of the energy spectrum about the vertical axis. The energy of a function is also given by the area under its energy spectrum

$$E = \int_{-\infty}^{\infty} |\hat{\psi}(f)|^2 df$$

so that

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{\psi}(f)|^2 df$$

In practice, the wavelet function is normalised to get unit energy. To make the wavelet more flexible we can either stretch and squeeze it (dilation) with a parameter a, or we can move it (translation) with a parameter b. The shifted and dilated versions of the mother wavelet are denoted $\psi(\frac{t-b}{a})$. The wavelet transform of a continuous signal with respect to the wavelet function is defined as

$$T(a,b) = w(a)\int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-b}{a})dt$$

where $w(a)$ is a weighting function. The asterisk indicates that the complex conjugate of the wavelet function is used in the transform. The wavelet transform can be considered as the cross-correlation of a signal with a set of wavelets of various widths. For instance, for reasons of energy conservation, one set $w(a) = \frac{1}{\sqrt{a}}$ leading to the continuous wavelet transform (CWT). The normalised wavelet function is often written as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{G.3.8}$$

so that the transform integral becomes

$$T(a,b) = \int_{-\infty}^{\infty} x(t)\psi^*_{a,b}(t)dt \tag{G.3.9}$$

corresponding to the convolution of the wavelet and the signal. Using the inner product, we get

$$T(a,b) =< x, \psi_{a,b} >$$

The wavelet transform is sometime called a mathematical microscope, where b is the location on the time series and a is associated with the magnification at location b. Continuous wavelet transforms are computed over a continuous range of a and b leading to a wavelet transform plot presented in a contour plot or as a surface plot. Another useful property of the wavelet transform is its ability to identify abrupt discontinuities (edges) in the signal. As the wavelet traverses the discontinuity there are first positive then negative values returned by the transform integral (located in the vicinity of the discontinuity). The inverse wavelet transform is defined as

$$x(t) = \frac{1}{C_g}\int_{-\infty}^{\infty}\int_0^{\infty} T(a,b)\psi_{a,b}(t)\frac{1}{a^2}dadb \tag{G.3.10}$$

to recover the original signal from its wavelet transform. Note, the original wavelet function is used rather than its conjugate which is used in the forward transformation. By limiting the integration range over a range of a scales, we can perform a basic filtering of the original signal. We are reconstructing the signal using

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_{a^*}^{\infty} T(a,b)\psi_{a,b}(t) \frac{1}{a^2} dadb$$

over the range of scales $a^* < a < \infty$ where a^* is the cutt-off scale. Since high frequency corresponds to small a scale, we are in effect reducing the high frequency noise, known as scale-dependent thesholding. This way of reconstructing the signal allows for denoising applied locally. A better way to separate pertinent signal features from unwanted noise, using the continuous wavelet transform, is by using a wavelet transform modulus maxima method. The modulus maxima lines are the loci of the local maxima and minima of the transform plot with respect to b, traced over wavelet scales. Following maxima lines down from large to small a scales allows the high frequency information corresponding to large features within the signal to be differentiated from high frequency noise components. It can be used as methods for filtering out noise from coherent signal features.

We are now presenting some signal energy such as wavelet-based energy and power spectra. The total energy contained in a signal $x(t)$ is defined as its integrated squared magnitude

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt = ||x(t)||^2$$

so that the signal must contain finite energy. The relative contribution of the signal energy contained at a specific a scale and b location is given by the two dimensional wavelet energy density function

$$E(a,b) = |T(a,b)|^2$$

and the plot of $E(a,b)$ is the scalogram. It can be integrated across a and b to recover the total energy in the signal using the admissibility constant

$$E = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_{0}^{\infty} |T(a,b)|^2 \frac{1}{a^2} dadb$$

The scalogram surface highlights the location and scale of dominant energetic features within the signal. The relative contribution to the total energy contained within the signal at a specific a scale is given by the scale dependent energy distribution

$$E(a) = \frac{1}{C_g} \int_{-\infty}^{\infty} |T(a,b)|^2 db$$

and peaks in $E(a)$ highlight the dominant energetic scales within the signal. We can convert the scale dependent wavelet energy spectrum of the signal, $E(a)$, to a frequency dependent wavelet energy spectrum $E_W(f)$ to compare with the Fourier energy spectrum of the signal $E_F(f)$. We must convert from the wavelet a scale to a characteristic frequency of the wavelet. One can use the passband centre of the wavelet's power spectrum or another representative frequency of the mother wavelet such as the spectral peak frequency f_p or the central frequency f_0. Since the spectral components are inversely proportional to the dilation $f \propto \frac{1}{a}$ and for $a = 1$ we get f_c, using this passband frequency, the characteristic frequency associated with a wavelet of arbitrary a scale is

$$f = \frac{f_c}{a}$$

where f_c becomes a scaling constant, and f is the representative or characteristic frequency for the wavelet at scale a. We can now associate the scale dependent energy $E(a)$ to the passband frequency of the wavelet. The total energy in the signal is given by

$$E = \int_0^\infty E(a)\frac{1}{a^2}da$$

Making the change of variable $f = \frac{f_c}{a}$, we can rewrite the equation in terms of the passband frequency. The derivative in the integral becomes $\frac{da}{a^2} = -\frac{df}{f_c}$, and swapping the integral limits to get rid of the negative sign, we get

$$E = \int_0^\infty E_W(f)df$$

where $E_W(f) = \frac{E(a)}{f_c}$ and the subscripts W corresponds to wavelet to differentiate it from its Fourier counterpart. The wavelet energy spectrum, which is the plot of the wavelet energy $E_W(f)$ against f, has an area underneath it equal to the total signal energy and may be compared with the Fourier energy spectrum $E_F(f)$ of the signal. The total energy in the signal becomes

$$E = \frac{1}{C_g f_c}\int_{-\infty}^\infty \int_0^\infty |T(f,b)|^2 df db$$

where $T(f,b) = T(a,b)$ for $f = \frac{f_c}{a}$. Further, the energy density surface in the time-frequency plane, defined by $E(f,b) = \frac{|T(f,b)|^2}{C_g f_c}$, contains a volume equal to the total energy of the signal

$$E = \int_{-\infty}^\infty \int_0^\infty E(f,b) df db$$

which can be compared to the energy density surface of the short time Fourier transform (the spectrogram). Since the peaks in $E(a,b)$ and $E(a)$ correspond to the most energetic parts of the signal as do the peaks in $E(f,b)$ and $E(f)$, we can use both the scalogram and the scale dependent energy distribution to determine the energy distribution relative to the wavelet scale. Scalograms are normally plotted with a logarithmic a scale axis. The power spectrum is the energy spectrum divided by the time period of the signal, and the area under the power spectrum gives the average energy per unit time (the power) of the signal. For a signal of length τ, the Fourier and wavelet power spectra are, respectively,

$$\begin{aligned} P_F(f) &= \frac{1}{\tau}E_F(f) \\ P_W(f) &= \frac{1}{\tau}E_W(f) = \frac{1}{\tau f_c C_g}\int_0^\tau |T(f,b)|^2 db \end{aligned}$$

The wavelet spectrum is more than simply a smeared version of the Fourier spectrum as the shape of the wavelet itself is an important parameter in the analysis of the signal. Also, the resulting wavelet power spectrum of the signal is dependent on the characteristic frequency of the wavelet (f_c in this study). At last, the wavelet variance, defined in the continuous time as

$$\sigma^2(a) = \frac{1}{\tau}\int_0^\tau |T(a,b)|^2 db$$

is used in practice to determine dominant scales in the signal (τ must be of sufficient length). Note, it differs from the scale dependent energy distribution and the power spectral density function only by constant multiplicative factors.

We can use the convolution theorem to express the wavelet transform in terms of products of the Fourier transform of the signal $\hat{x}(f)$ and wavelet $\hat{\psi}_{a,b}(f)$

$$T(a,b) = \int_{-\infty}^\infty \hat{x}(f)\hat{\psi}^*_{a,b}(f)df$$

where the conjugate of the wavelet function is used. The Fourier transform of the dilated and translated wavelet is

$$\hat{\psi}_{a,b}(f) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{a}} \psi(\frac{t-b}{a}) e^{-i(2\pi f)t} dt$$

Making the substitution $t^{'} = \frac{t-b}{a}$ ($dt = adt^{'}$) we get

$$\hat{\psi}_{a,b}(f) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{a}} \psi(t^{'}) e^{-i(2\pi f)(at^{'}+b)} adt^{'}$$

Separating out the constant part of the exponential function and dropping the prime from $t^{'}$ we get

$$\hat{\psi}_{a,b}(f) = \sqrt{a} e^{-i(2\pi f)b} \int_{-\infty}^{\infty} \psi(t) e^{-i(2\pi af)t} dt$$

which is the Fourier transform of the wavelet at rescaled frequency af. Hence, we get the relation

$$\hat{\psi}_{a,b}(f) = \sqrt{a} \hat{\psi}(af) e^{-i(2\pi f)b}$$

Adding the asterisk to the wavelet in the above equation and changing the sign in front of the imaginary term, we obtain the Fourier transform of the wavelet conjugate. At last, the wavelet transform can be written in expanded form as

$$T(a,b) = \sqrt{a} \int_{-\infty}^{\infty} \hat{x}(f) \hat{\psi}^{*}(af) e^{i(2\pi f)b} df$$

which has the form of an inverse Fourier transform. It is very useful when using the discretised approximation of the continuous wavelet transform with large signal data set as we can use the fast Fourier transform (FFT) algorithm to speed up computation time. Further, the Fourier transform of the wavelet function $\hat{\psi}_{a,b}(f)$ is usually known in analytic form, and need not be computed using an FFT. Only an FFT of the original signal $\hat{x}(f)$ is required, then, to get $T(a,b)$ we take the inverse FFT of the product of the signal Fourier transform and the wavelet Fourier transform for each required a scale and multiply the result by $\sqrt{a}$.

One property of wavelet is that it can localise itself in time for short duration (high frequency) fluctuations. But, in that time frame, there is an associated spreading of the frequency distribution associated with wavelets. Conversely, there is a spreading in temporal resolution at low frequencies. The spread of $|\psi_{a,b}(t)|^2$ and $|\hat{\psi}_{a,b}(f)|^2$ can be quantified using σ_t and σ_f respectively (standard deviations around their respective means). The spread of the wavelets in the time-frequency plane can be represented by drawing boxes of side lengths $2\sigma_t$ by $2\sigma_f$, called the Heisenberg boxes after the Heisenberg uncertainty principle which address the problem of the simultaneous resolution in time and frequency that can be attained when measuring a signal. The more accurate the temporal measurement (smaller σ_t) the less accurate the spectral measurement (larger σ_f) and vice versa. One solution is to consider the short-time Fourier transform (STFT) which employs a window function to localise the complex sinusoid

$$F(f,b) = \int_{-\infty}^{\infty} x(t) h(t-b) e^{-i2\pi ft} dt$$

where $h(t-b)$ is the window function which confines the complex sinusoid $e^{-i2\pi ft}$. There are many shapes of window available such as Hanning, Hamming, cosine, Kaiser and Gaussian. The combined window plus the complex sinusoid is known as the window Fourier atom or time-frequency atom given by

$$h_{f,b}(t) = h(t-b) e^{i2\pi ft}$$

We obtain the time-frequency decomposition by convolving the complex conjugate of this atom with the signal $x(t)$. Assuming a Gaussian window, we get the Gabor STFT having a very similar form to the Morlet wavelet transform.

While the Morlet wavelet has a form very similar to the analysing function used for the STFT within a Gaussian window, in the former we scale the window and enclosed sinusoid together, whereas in the later we keep the window length constant and scale only the enclosed sinusoid.

G.3.2 The discrete wavelet transform

G.3.2.1 An infinite summations of discrete wavelet coefficients

When certain criteria are met it is possible to completely reconstruct the original signal using infinite summations of discrete wavelet coefficients rather than continuous integrals. One can sample the parameters a and b by using a logarithmic discretisation of the a scale and link it to the size of steps taken between b locations. To do so, we move in discrete steps to each location b which are proportional to the scale a, getting the wavelet form

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}}\psi(\frac{t - nb_0a_0^m}{a_0^m})$$

where the integers m and n control the wavelet dilation and translation respectively. Also, $a_0 > 1$ is a fixed dilation step parameter, and $b_0 > 0$ is the location parameter. Hence, the step of the translation steps $\Delta b = b_0a_0^m$ is directly proportional to the wavelet scale a_0^m. In that setting, the wavelet transformation becomes

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\frac{1}{a_0^{\frac{1}{2}}}\psi(a_0^{-m}t - nb_0)dt$$

written with the inner product as $T_{m,n} =< x, \psi_{m,n} >$ where $T_{m,n}$ is the discrete wavelet transform with values known as wavelet coefficients (detail coefficients). The wavelet frames, providing a general framework, are constructed by discretely sampling the time and scale parameters of a continuous wavelet transform as detailed above. Within that framework, the energy of the resulting wavelet coefficients must lies with a certain bounded range of the energy, E, of the original signal

$$AE \leq \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}|T_{m,n}|^2 \leq BE$$

where the bounds A and B depend on the parameters a_0 and b_0 chosen. In the special case where $A = B$ the frame is tight, and the formula becomes

$$x(t) = \frac{1}{A}\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}T_{m,n}\psi_{m,n}(t)$$

and

- when $A = B = 1$ the frame forms an orthonormal basis,
- if $A = B > 1$ the frame is redundant,
- and for $A \neq B$ the constant in the above equation is modified to $\frac{2}{A+B}$.

For $a_0 = 2$ and $b_0 = 1$ we get the dyadic grid arrangement (power of two logarithmic scaling of the dilation and translation steps), and the wavelet becomes

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}}\psi(\frac{t - n2^m}{2^m}) = \frac{1}{\sqrt{2^m}}\psi(\frac{t}{2^m} - n) \tag{G.3.11}$$

Assuming $A = B = 1$, the wavelets are both orthogonal to each other and normalised to have unit energy. That is,

$$\int_{-\infty}^{\infty} \psi_{m,n}(t)\psi_{m',n'}(t)dt = \begin{cases} 1 \text{ if } m = m^{'} \text{ and } n = n^{'} \\ 0 \text{ otherwise} \end{cases}$$

so that the information stored in the wavelet coefficient $T_{m,n}$ is not repeated elsewhere, avoiding redundancy. Given the discrete wavelet transform (DWT)

$$T_{m,n} =< x, \psi_{m,n} >= \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt$$

if $\psi_{m,n}(t)$ is an orthonormal basis ($A = B = 1$), we can reconstruct the original signal in terms of $T_{m,n}$ using the inverse discrete wavelet transform (IDWT)

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t)$$

So, on one hand we have the DWT where the transform integral remains continuous but determined only on a discretised grid of a and b, and on the other hand the discretised approximations of the CWT (discrete approximation of the transform integral computed on a discrete grid of a and b).

G.3.2.2 The scaling function

The scaling function is associated with the smoothing of the signal and has the form of the wavelet

$$\phi_{m,n}(t) = 2^{-\frac{m}{2}}\phi(2^{-m}t - n) \tag{G.3.12}$$

with the property

$$\int_{-\infty}^{\infty} \phi_{0,0}(t)dt = 1$$

where $\phi_{0,0}(t) = \phi(t)$ is the father scaling function or father wavelet. The scaling function is orthogonal to translation of itself, but not to dilations of itself, and it can be convolved with the signal to produce approximation coefficients

$$S_{m,n} =< x, \phi_{m,n} >= \int_{-\infty}^{\infty} x(t)\phi_{m,n}(t)dt \tag{G.3.13}$$

which are simply weighted averages of the continuous signal factored by $2^{\frac{m}{2}}$. These approximations at a specific scale m are the discrete approximation of the signal at that scale and can be combined to generate the continuous approximation at scale m

$$x_m(t) = \sum_{n=-\infty}^{\infty} S_{m,n}\phi_{m,n}(t)$$

where $x_m(t)$ is a smooth, scaling function dependent, version of the signal $x(t)$ at scale m. It approaches $x(t)$ at small scales, as $m \to -\infty$. We can represent a signal $x(t)$ using a combined series expansion using both the approximation coefficients and the wavelet coefficients

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m_{0,n}}\phi_{m_{0,n}}(t) + \sum_{m=-\infty}^{m_0} \sum_{n=-\infty}^{\infty} T_{m,n}\phi_{m,n}(t)$$

The signal detail at scale m is defined as

$$d_m(t) = \sum_{n=-\infty}^{\infty} T_{m,n}\phi_{m,n}(t) \tag{G.3.14}$$

so that the signal becomes

$$x(t) = x_{m_0}(t) + \sum_{m=-\infty}^{m_0} d_m(t) \tag{G.3.15}$$

which is a linear combination of the details. Hence, one can show that

$$x_{m-1}(t) = x_m(t) + d_m(t) \tag{G.3.16}$$

stating that adding the signal detail at an arbitrary scale (index m) to the approximation at that scale, we get the signal approximation at an increased resolution (index $m-1$), which is called the multiresolution representation.

The scaling equation (or dilation equation) describes the scaling function $\phi(t)$ in terms of contracted and shifted versions of itself

$$\phi(t) = \sum_k c_k \phi(2t-k) \tag{G.3.17}$$

where $\phi(2t-k)$ is a contracted version of $\phi(t)$ shifted along the time axis by an integer step k and factored by an associated scaling coefficient c_k. That is, we can build a scaling function at one scale from a number of scaling equations at the previous scale. Note, wavelets of compact support have sequences of nonzero scaling coefficients of finite length. The scaling coefficients must satisfy the constraint

$$\sum_k c_k = 2$$

so that to create an orthogonal system we must have

$$\sum_k c_k c_{k+2k'} = \begin{cases} 2 \text{ if } k' = 0 \\ 0 \text{ otherwise} \end{cases}$$

and the sum of the squares of the scaling coefficients is equal to 2. Similarly, the differencing of the associated wavelet equation satisfies

$$\psi(t) = \sum_k (-1)^k c_{1-k}\phi(2t-k) \tag{G.3.18}$$

ensuring orthogonality between the wavelets and the scaling functions. In the case of wavelets of compact support the equation becomes

$$\psi(t) = \sum_k (-1)^k c_{N_k-1-k}\phi(2t-k)$$

over the interval $[0, N_k - 1]$. Given the coefficient

$$b_k = (-1)^k c_{N_k-1-k}$$

where the sum of all b_k is zero, the equation becomes

$$\psi(t) = \sum_k b_k \phi(2t-k)$$

Plugging the scaling equation (G.3.17) into the scaling function in Equation (G.3.12), we get

$$\phi_{m,n}(t) = \sum_k c_k 2^{-\frac{m}{2}} \phi(2^{-m+1}t - (2n+k))$$

As a result, for a wavelet at scale index $m+1$, we get

$$\phi_{m+1,n}(t) = \frac{1}{\sqrt{2}} \sum_k c_k \phi_{m,2n+k}(t) = \sum_k h_k \phi_{m,2n+k}(t)$$

where $h_k = \frac{c_k}{\sqrt{2}}$, stating that the scaling function at an arbitrary scale is made of a sequence of shifted scaling functions at the next smaller scale each factored by their respective scaling coefficients. Similarly, for the wavelet function we get

$$\psi_{m+1,n}(t) = \frac{1}{\sqrt{2}} \sum_k b_k \phi_{m,2n+k}(t) = \sum_k g_k \phi_{m,2n+k}(t)$$

The vectors containing the sequences $\frac{1}{\sqrt{2}}c_k$ and $\frac{1}{\sqrt{2}}b_k$ represent the filters h_k and g_k, respectively. That is, $h_k = \frac{1}{\sqrt{2}}c_k$ is a low-pass filter, and $g_k = \frac{1}{\sqrt{2}}b_k$ is a high-pass filter in the associated analysis filter bank. The former is the lowpass filter letting through low signal frequencies (a smoothed version of the signal), and the latter is the highpass filter letting through the high frequencies corresponding to the signal details. These filters defined the Quadrature Mirror Filters (QMF) (see Vaidyanathan [1992]).

G.3.2.3 The FWT algorithm

We can now express the approximation coefficients and the wavelet coefficients in term of the scaling function at a specific scale. The approximation coefficients at scale index $m+1$ are given by

$$S_{m+1,n} = < x, \phi_{m+1,n} > = \int_{-\infty}^{\infty} x(t)\phi_{m+1,n}(t)dt$$

Replacing $\phi_{m+1,n}(t)$ with its expression above, we get

$$S_{m+1,n} = \int_{-\infty}^{\infty} x(t)\big(\frac{1}{\sqrt{2}} \sum_k c_k \phi_{m,2n+k}(t)\big)dt$$

which we rewrite as

$$S_{m+1,n} = \frac{1}{\sqrt{2}} \sum_k c_k \big(\int_{-\infty}^{\infty} x(t)\phi_{m,2n+k}(t)dt\big)$$

where the integral gives the approximation coefficients $S_{m,2n+k}$ for each k. Hence, the approximation coefficients become

$$S_{m+1,n} = \frac{1}{\sqrt{2}} \sum_k c_k S_{m,2n+k} = \frac{1}{\sqrt{2}} \sum_k c_{k-2n} S_{m,k} = \sum_k h_{k-2n} S_{m,k} \tag{G.3.19}$$

We can therefore generate the approximation coefficients $S_{\bullet,n}$, at scale index $m+1$, by using the scaling coefficients at the previous scale. Similarly, the wavelet coefficients, $T_{\bullet,n}$, can be found from the approximation coefficients at the previous scale by using the reordered scaling coefficients b_k

$$T_{m+1,n} = \frac{1}{\sqrt{2}} \sum_k b_k S_{m,2n+k} = \frac{1}{\sqrt{2}} \sum_k b_{k-2n} S_{m,k} = \sum_k g_{k-2n} S_{m,k} \tag{G.3.20}$$

Hence, knowing the approximation coefficients $S_{m_0,n}$ at a specific scale m_0, and then using the above equations repeatedly, we can generate the approximation and detail wavelet coefficients at all scales larger than m_0. Note, we do not need to know exactly the continuous signal $x(t)$ but only $S_{m_0,n}$. Equations (G.3.19) and (G.3.20) represent the multiresolution decomposition algorithm, which is the first half of the fast wavelet transform (FWT). These iterating equations perform respectively a highpass and lowpass filtering of the input where the vectors containing the sequences h_k and g_k represent the filters. Expanding Equation (G.3.16), we get

$$x_{m-1}(t) = \sum_n S_{m,n}\phi_{m,n}(t) + \sum_n T_{m,n}\psi_{m,n}(t)$$

which we can expand in terms of the scaling function at the previous scale as

$$x_{m-1}(t) = \sum_n S_{m,n}\frac{1}{\sqrt{2}}\sum_k c_k\phi_{m-1,2n+k}(t) + \sum_n T_{m,n}\frac{1}{\sqrt{2}}\sum_k b_k\phi_{m-1,2n+k}(t)$$

Rearranging the summation indices, we get

$$x_{m-1}(t) = \sum_n S_{m,n}\frac{1}{\sqrt{2}}\sum_k c_{k-2n}\phi_{m-1,k}(t) + \sum_n T_{m,n}\frac{1}{\sqrt{2}}\sum_k b_{k-2n}\phi_{m-1,k}(t)$$

or equivalently

$$x_{m-1}(t) = \sum_n S_{m,n}\sum_k h_{k-2n}\phi_{m-1,k}(t) + \sum_n T_{m,n}\sum_k g_{k-2n}\phi_{m-1,k}(t)$$

We can also expand $x_{m-1}(t)$ in terms of the approximation coefficients at scale $m-1$

$$x_{m-1}(t) = \sum_n S_{m-1,n}\phi_{m-1,n}(t)$$

Equating the coefficients in these two equations we see that the index k at scale index m relates to the location index n at scale index $m-1$. Further, the location index n in the first equation above corresponds to scale index m with associated location spacings 2^m, while the index n in the second equation corresponds to scale index $m-1$ with discrete location spacings 2^{m-1} making this n indices twice as dense as the first one. We can therefore swap the indices k and n in the first equation before equating the two expressions, getting the reconstruction algorithm

$$S_{m-1,n} = \frac{1}{\sqrt{2}}\sum_k c_{n-2k}S_{m,k} + \frac{1}{\sqrt{2}}\sum_k b_{n-2k}T_{m,k} = \sum_k h_{n-2k}S_{m,k} + \sum_k g_{n-2k}T_{m,k} \tag{G.3.21}$$

where k is reused as the location index of the transform coefficients at scale index m to differentiate it from n, the location index at scale $m-1$. The reconstruction algorithm is the second half of the fast wavelet transform corresponding to the synthesis filter bank.

***Remark* G.3.1** *In general, when discretising the continuous wavelet transform, the FWT, DWT, decomposition/reconstruction algorithms, fast orthogonal wavelet transform, multiresolution algorithm, are all used to mean the same thing.*

G.3.3 Discrete input signals of finite length

G.3.3.1 Discribing the algorithm

We now consider the wavelet multiresolution framework in the case of discrete input signals specified at integer spacings. The signal approximation coefficients at scale index $m = 0$ is defined by

$$S_{0,n} = \int_{-\infty}^{\infty} x(t)\phi(t-n)dt$$

and will allow us to generate all subsequent approximation and detail coefficients, $S_{m,n}$ and $T_{m,n}$, at scale indices greater than $m = 0$. We assume that the given discrete input signal $S_{0,n}$ is of finite length N, which is an integer power of 2, that is, $N = 2^M$ so that the range of scale we can investigate is $0 < m < M$. Substituting $m = 0$ and $m = M$ in Equation (G.3.15), and noting that we have a finite range of n halving at each scale, the signal approximation scale $m = 0$ (input signal) can be written as the smooth signal at scale M plus a combination of detailed signals

$$\sum_{n=0}^{2^{M-m}-1} S_{0,n}\phi_{0,n}(t) = S_{M,n}\phi_{M,n}(t) + \sum_{m=1}^{M}\sum_{n=0}^{2^{M-m}-1} T_{m,n}\psi_{m,n}(t)$$

We can rewrite this equation as

$$x_0(t) = x_M(t) + \sum_{m=1}^{M} d_m(t) \tag{G.3.22}$$

where the mean signal approximation at scale M is

$$x_M(t) = S_{M,n}\phi_{M,n}$$

The indexing is such that $m = 1$ corresponds to the finest scale (high frequencies). The detail signal approximation corresponding to scale index m is defined for a finite length signal as

$$d_m(t) = \sum_{n=0}^{2^{M-m}-1} T_{m,n}\psi_{m,n}(t) \tag{G.3.23}$$

Hence, adding the approximation of the signal at scale index M to the sum of all detail signal components across scales $0 < m < M$ gives the approximation of the original signal at scale index 0. We can rewrite Equation (G.3.16) as

$$x_m(t) = x_{m-1}(t) - d_m(t)$$

and starting with the input signal at scale $m - 1 = 0$,we see that at scale index $m = 1$, the signal approximation is given by

$$x_1(t) = x_0(t) - d_1(t)$$

At the next scale $(m = 2)$ the signal approximation is given by

$$x_2(t) = x_0(t) - d_1(t) - d_2(t)$$

and so on, corresponding to the successive stripping of high frequency information from the original signal. Once we have the discrete input signal $S_{0,n}$, we can compute $S_{m,n}$ and $T_{m,n}$ using the decomposition algorithm given by Equations (G.3.19) and (G.3.20). At scale index M we have performed a full decomposition of the finite length, discrete input signal. We are left with an array of coefficients: a single approximation coefficient value, $S_{M,0}$, plus the detail coefficients $T_{m,n}$ corresponding to discrete wavelets of scale $a = 2^m$ and location $b = 2^m n$. The finite time series is of length $N = 2^M$ giving the ranges of m and n for the detail coefficients as respectively $1 < m < M$ and $0 < n < 2^{M-m} - 1$. At the smallest wavelet scale, index $m = 1$, $\frac{2^M}{2^1} = \frac{N}{2}$ coefficients are computed, at the next scale, index $m = 2$, $\frac{2^M}{2^2} = \frac{N}{4}$ coefficients are computed, and so on until the largest scale $m = M$ where one coefficient $\frac{2^M}{2^M}$ is computed. The total number of detail coefficients, $T_{m,n}$, for a discrete time series of length $N = 2^M$ is then $1 + 2 + 4 + ... + 2^{M-1}$, or $\sum_{m=1}^{M-1} 2^m = 2^M - 1 = N - 1$. Note, the single approximation coefficient $S_{M,0}$ remains to fully represent the discrete signal. Thus, a discrete input signal of length N can be broken down into exactly N components without any loss of information using discrete orthogonal wavelets. Further, no signal information is

repeated in the coefficient representation, which is known as zero redundancy. After a full decomposition, the energy contained within the coefficients at each scale is

$$E_m = \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^2$$

The total energy of the discrete input signal $E = \sum_{n=0}^{N-1}(S_{0,n})^2$ is equal to the sum of the squared detail coefficients over all scales plus the square of the remaining coefficient $S_{M,0}$, that is,

$$E = (S_{M,0})^2 + \sum_{m=1}^{M} \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^2$$

Since the energy contained within the transform vector at all stages of the multiresolution decomposition remains constant, we can write the conservation of energy more generally as

$$E = \sum_{i=0}^{N-1} (W_i^m)^2$$

where W_i^m are the individual components of the transform vector W^m. The wavelet transform vector after the full decomposition has the form

$$W^M = \left(S_M, T_M, T_{M-1}, .., T_m, ..., T_2, T_1\right) \tag{G.3.24}$$

where $T_m = \{T_{m,n} : n \in \mathbb{Z}\}$ represents the sub-vector containing the coefficients $T_{m,n}$ at scale index m (with n in the range 0 to $2^{M-m} - 1$). If we halt the transformation process before the full decomposition at arbitrary level m_0, the transform vector has the form

$$W^{m_0} = \left(S_{m_0}, T_{m_0}, T_{m_0-1}, ..., T_2, T_1\right)$$

where m_0 can take the range $1 \leq m_0 \leq M - 1$. The transform vector always contains $N = 2^M$ components. Also, we can express the original input signal as the transform vector at scale index zero, that is, W^0.

G.3.3.2 Presenting thresholding

Once we have performed the full decomposition, we can manipulate the coefficients in the transform vector in a variety of ways by setting groups of components to zero, setting selected individuals components to zero, or reducing the magnitudes of some components. Since the transform vector contains both small and large values, we can throw away the smallest valued coefficients in turn and perform the inverse transforms. The least significant components have been first smoothed out, leaving intact the more significant fluctuating parts of the signal. We have threshold the wavelet coefficients at increasing magnitudes. We can define the scale-dependent smoothing of the wavelet coefficients as

$$T_{m,n}^{scale} = \begin{cases} 0 \text{ if } m \geq m^* \\ T_{m,n} \text{ if } m < m^* \end{cases}$$

where m^* is the index of the threshold scale, or the transform vector. Considering sequentially indexed coefficients W_i, and assuming a full decomposition, the thresholding criterion becomes

$$W_i^{scale} = \begin{cases} 0 \text{ if } i \geq 2^{M-m^*} \\ W_i \text{ if } i < 2^{M-m^*} \end{cases}$$

where the range of the sequential index i is from 0 to $N - 1$, and N is the length of the original signal. Hence, $i = 2^{M-m^*}$ is the first location index within the transform vector where the coefficients are set to zero. Magnitude

thresholding is carried out to remove noise from a signal, to partition signals into two or more components, or simply to smooth the data. The two most popular methods for selecting and modifying the coefficients are hard and soft thresholding. While scale-dependent smoothing removes all small scale coefficients below the scale index m^* regardless of amplitude, hard and soft thresholding remove or reduce the smallest amplitude coefficients regardless of scale. To hard threshold the coefficients, one must define the threshold λ in relation with some mean value of the wavelet coefficients at each scale such as the standard deviation or the mean absolute deviation. The coefficients above the threshold correspond to the coherent part of the signal, and those below the threshold correspond to the random or noisy part of the signal. It is defined as

$$W_i^{hard} = \begin{cases} 0 \text{ if } |W_i| < \lambda \\ W_i \text{ if } |W_i| \geq \lambda \end{cases}$$

where one makes the decision to keep or remove the coefficients. The soft version recognises that the coefficients contain both signal and noise, and attempts to isolate the signal by removing the noisy part from all coefficients. It is defined as

$$W_i^{soft} = \begin{cases} 0 \text{ if } |W_i| < \lambda \\ sign(W_i)(|W_i| - \lambda) \text{ if } |W_i| \geq \lambda \end{cases}$$

where all the coefficients below λ are set to zero and the ones above are shrunk towards zero by an amount λ. One commonly used measure of the optimum reconstruction is the mean square error between the reconstructed signal and the original one. In the case where the exact form of either the underlying signal or the corruption noise is not known, we can use the universal threshold defined as

$$\lambda_U = (2\ln N)^{\frac{1}{2}}\sigma$$

where $(2\ln N)^{\frac{1}{2}}$ is the expected maximum value of a white noise sequence of length N and unit standard deviation, and σ is the standard deviation of the noise in the signal. However, in practice the universal threshold tends to oversmooth. Further, since we do not know σ for our signal we need to use robust estimate $\hat{\sigma}$, typically set to the median of absolute deviation (MAD) of the wavelet coefficients at the smallest scale divided by 0.6745 to calibrate with the standard deviation of a Gaussian distribution. Hence, the universal threshold becomes

$$\lambda_U = (2\ln N)^{\frac{1}{2}}\frac{MAD}{0.6745} = (2\ln N)^{\frac{1}{2}}\hat{\sigma}$$

Other thresholding methods exist such as the minimax method, the SURE method, the HYBRID method, cross-validation methods, the Lorentz method and various Bayesian approaches.

G.3.4 Wavelet-based statistical measures

While turbulent statistical measures have been calculated in the Fourier space, the non-local nature of the Fourier modes lead to important temporal information losses. On the other hand, the local properties of wavelets make it ideal to quantify the temporal and spectral distribution of the energy in statistical terms such as wavelet variance, skewness, flatness, etc. These statistics enable both scale and location dependent behaviour to be quantified. For simplicity, we consider only discrete transform coefficients $T_{m,n}$ generated from full decompositions using real-valued, discrete orthonormal wavelet transforms. We further assume that the mean has been removed from the signal, and that it contains $N = 2^M$ data points.

The pth order statistical moment of the wavelet coefficients $T_{m,n}$ at scale index m is given by

$$< T_{m,n}^p >_m = \frac{1}{2^{M-m}} \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^p$$

where $<>_m$ denotes the average taken over the number of coefficients at scale m. The variance at scale index m is obtained by setting $p = 2$ in the above formula. It corresponds to the average energy wrapped up per coefficient at each scale m. A general dimensionless moment function can be defined as

$$F_m^p = \frac{< T_{m,n}^p >_m}{(< T_{m,n}^p >_m)^{\frac{p}{2}}}$$

where the pth order moment is normalised by dividing it by the rescaled variance. For example, the scale-dependent coefficient skewness factor is defined as the normalised third moment with $p = 3$, and the scale-dependent coefficient flatness factor is obtained with $p = 4$. The later gives a measure of the peakedness (or flatness) of the probability distribution of the coefficients at each level. In the case of a Gaussian distribution the flatness factor is 3. The wavelet-based scale-dependent energy is defined as

$$E_m = \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^2 \Delta t$$

where Δt is the sampling time. The scale-dependent energy per unit time, or scale-dependent power, is $P_m = \frac{E_m}{\tau}$ where τ is the total time period of the signal. In the case where $\tau = 2^M \Delta t$ we get

$$P_m = \frac{1}{2^M} \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^2$$

so that as long as the signal has zero mean, both the total energy and total power of the signal can be found by summing E_m and P_m respectively over all scale indices m. We can then construct a wavelet power spectrum for direct comparison with the Fourier spectrum

$$P_W(f_m) = \frac{1}{\tau} \frac{2^m \Delta t}{\ln 2} \sum_{n=0}^{2^{M-m}-1} (T_{m,n})^2 \Delta t = \frac{1}{\tau} \frac{2^m \Delta t}{\ln 2} E_m$$

where $\frac{2^m \Delta t}{\ln 2}$ stems from the dyadic spacing of the grid. The temporal scale of the wavelet at scale index m is equal to $2^m \Delta t$. Another common statistical measure of the energy distribution across scales is the normalised variance of the wavelet energy. It is called the fluctuation intensity (FI) (or the coefficient of variation (CV)), given by

$$FI_m = \frac{\left(< T_{m,n}^4 >_m - (< T_{m,n}^2 >_m)^2\right)^{\frac{1}{2}}}{< T_{m,n}^2 >_m}$$

and measures the standard deviation of the variance in coefficient energies at scale index m. The intermittency at each scale can be viewed directly by using the intermittency index $I_{m,n}$ given by

$$I_{m,n} = \frac{(T_{m,n})^2}{< T_{m,n}^2 >_m}$$

The index $I_{m,n}$ is the ratio of local energy to the mean energy at temporal scale $2^m \Delta t$. It allows the investigator to visualise the uneven distribution of energy through time at a given wavelet scale. The correlation between the scales can be measured using the pth moment scale correlation R_m^p given by

$$R_m^p = 2^{M-m} \sum_{n=0}^{2^{M-(m-1)}-1} B_{m,[\frac{n}{2}]}^p B_{m-1,n}^p$$

where $B_{m,n}^p$ is the pth order moment function, and $[\frac{n}{2}]$ requires that the integer part only be used. $B_{m,n}^p$ is the pth order moment function defined as

$$B^p_{m,n} = \frac{(T_{m,n})^p}{\sum_{n=0}^{2^{M-m}-1}(T_{m,n})^p}$$

It has a similar form to the intermittency index when $p = 2$, except that $B^p_{m,n}$ has a normalised sum at each scale, that is, $\sum_n B^p_{m,n} = 1$, whereas the sum of the intermittency indices at scale m is equal to the number of coefficients at that scale, that is, $\sum_n I_{m,n} = 2^{M-m}$.

G.4 The problem of shift-invariance

There is a large literature on redundant transforms, using different notations. We are going to review some of these transforms, trying to be consistent with the notation introduced in the continuous and discrete wavelet transform.

G.4.1 A brief overview

G.4.1.1 Describing the problem

The DWT is an (bi-) orthogonal transform and provides a sparse time-frequency representation of the original signal, making it computationally efficient. However, the use of critical sub-sampling in the DWT, where every second wavelet coefficient at each decomposition level is discarded, forces it to be shift variant. This critical sub-sampling results in wavelet coefficients that are highly dependent on their location in the sub-sampling lattice, leading to small shifts in the input waveform which causes large changes in the wavelet coefficients, large variations in the distribution of energy at different scales, and possibly large changes in reconstructed waveforms. Alternatively, considering the frequency response of the mother wavelets, when the WT sub-bands are sub-sampled by a factor of two, the Nyquist criteria is strictly violated and frequency components above (or below) the cut-off frequency of the filter will be aliased into the wrong sub-band. Note, the aliasing introduced by the DWT cancels out when the inverse DWT (IDWT) is performed using all of the wavelet coefficients (when the original signal is reconstructed). As soon as the coefficients are not included in the IDWT, or they are quantised, the aliasing no-longer cancels out and the output is no-longer shift-invariant.

All the techniques attempting to eliminate or minimise the amount of aliasing consider relaxing the critical sub-sampling criteria and/or reducing the transition bandwidth of the mother wavelets. One way forward is to use the a trous algorithm which do not perform any sub-sampling at all. Since in that algorithm the mother wavelet has to be dilated (by inserting zeros) at each level of the transform, it requires additional computation and memory. Note, it is only strictly shift-invariant under circular convolution (periodic boundary extension). Alternatively, the power shiftable discrete wavelet transform (PSDWT) achieves approximate shift-invariance by limiting the sub-band sub-sampling . One can also build two wavelet decomposition trees (with alternate phase sub-sampling), one for a mother wavelet with even symmetry and the other for the same mother wavelet, but with odd symmetry. In this way, the dual tree complex wavelet transform (DTCWT) measures both the real (even) and the imaginary (odd) components of the input signal. However, since we must perform two decompositions, the computation and memory requirements are twice that of the Mallat DWT. Another method is to use the wavelet transform modulus maxima (WTMM), which is a fully sampled dyadic WT, using a mother wavelet with one or two vanishing moments, applied to estimate the multi-resolution gradient of the signal. Since the dyadic WT has the same properties as the CWT, it is shift-invariant. Further, if the coefficients can be adaptively sub-sampled to keep only the coefficients being locally maximum or locally minimum (the modulus maxima) at each scale, this sub-sampled representation is also shift-invariant. However, using a pseudo inverse transform, exact reconstruction from the wavelet modulus maxima is not possible.

Since Shensa [1992] showed that the Mallat and a trous algorithms are special cases of the same filter bank structure, it is possible to combine them to benefit from both approaches. Bradley [2003] proposed a generalisation of the Mallat and a trous discrete wavelet transform called the over complete discrete wavelet transform (OCDWT)

which achieves various levels of shift-invariance by controlling the amount of sub-sampling applied at each level of the transform. He applied the Mallat algorithm to the first M levels of an L-level decomposition and then applied the a trous algorithm to the remaining $(L-M)$ levels. This method can be seen as an initial down-sampling of the signal prior to a fully sampled a trous decomposition.

G.4.1.2 The a trous algorithm

We saw in Appendix (G.3.2.1) that the DWT is still computationally intensive, and that one way forward is to use the fast wavelet transform (FWT). Alternatively, we can apply the a trous algorithm proposed by Holschneider et al. [1989] which can be described as a modification of the FWT algorithm. Given $(x(t))_{t\in\mathbb{Z}}$ a discrete time process and $(g_k, h_k)_{k\in\mathbb{Z}}$ the filter banks with $T_{m,n} =< x, \psi_{m,n} >$ and $S_{m,n} =< x, \phi_{m,n} >$, then Equations (G.3.19) and (G.3.20) are modified in the following way

$$S_{m,n} = \sum_{k\in\mathbb{Z}} h_k S_{m-1,k} \text{ and } T_{m,n} = \sum_{k\in\mathbb{Z}} g_k S_{m-1,k}$$

where h_{k-2n} and g_{k-2n} are replaced with h_k and g_k. We set $T_m = \{T_{m,n} : n \in \mathbb{Z}\} \in \mathcal{L}^2(\mathbb{Z})$ and $S_m = \{S_{m,n} : n \in \mathbb{Z}\} \in \mathcal{L}^2(\mathbb{Z})$, and let h^r, g^r be recursive filters with $g^0 = h$ and $h^0 = g$. The h^r, g^r are computed by introducing zeros between each component of g^{r-1} and h^{r-1}. Two operators G^r, H^r are defined as

$$G^r : \mathcal{L}^2(\mathbb{Z}) \to \mathcal{L}^2(\mathbb{Z}) \text{ with } c \to \{(G^r c)_n = \sum_{k\in\mathbb{Z}} g^r_{k-n} c_k\}$$

and

$$H^r : \mathcal{L}^2(\mathbb{Z}) \to \mathcal{L}^2(\mathbb{Z}) \text{ with } c \to \{(H^r c)_n = \sum_{k\in\mathbb{Z}} h^r_{k-n} c_k\}$$

The adjoint functions G^{r^*}, H^{r^*} are defined analogously to invert this mapping. The a trous decomposition algorithm is performed as follow:
As input we need $S_0 = \{S_{0,n} : n \in \mathbb{Z}\}$ and $M \in \mathbb{N}$, where 2^M is the maximal scale. We then gradually compute

$$\forall m = 1, .., M \ , T_m = G^{m-1} S_{m-1} \ , S_m = H^{m-1} S_{m-1}$$

and yield S_M, T_m for $m = 1, .., M$ to fill the vector in Equation (G.3.24). It is a multiscale decomposition of the time series with S_M containing the information about the highest scale (the long-term component). For the reconstruction of the time series, we start with M, S_M, T_m for $m = 1, .., M$ and gradually compute

$$\forall m = M, M-1, .., 1 \ , S_{m-1} = H^{m^*} S_m + G^{m^*} T_m$$

The result is S_0, and from that we yeld the time series via inversion of the respective convolution.

G.4.1.3 Relating the a trous and Mallat algorithms

Shensa [1992] showed that the a trous and Mallat algorithms are both filter bank structures where the only distinguishing feature is the choice of two finite length filters, a lowpass filter h and a bandpass filter g. The lowpass condition given by $\sum_k h_k = \sqrt{2}$ is necessary to the construction of the corresponding continuous wavelet function, and the bandpass requirement ensures that finite energy signals lead to finite energy transforms. Under these conditions, the filter bank output is referred to as the discrete wavelet transform (DWT). We discussed in Appendix (G.3.2) the Mallat algorithm and described the lowpass and bandpass filters in Equations (G.3.19) and (G.3.20). We will try to keep the same notation. In the a trous algorithm, the lowpass filter satisfies the condition $h_{2k} = \frac{\delta(k)}{\sqrt{2}}$, where $\delta(\bullet)$ is the Dirac delta function, which corresponds to a non-orthogonal wavelet decomposition. If the filter h is a trous, then the DWT coincides with a CWT by wavelet $g(t)$ (ψ in our notation) whose samples $g(n)$ form the filter g. Given the CWT

defined in Equation (G.3.9), we consider discrete values for a and b, and assume that $a = 2^i$ where i is the octave of the transform. We let $T_{i,n} = T(2^i, n)$ be the wavelet series and take b to be a multiple of a, getting $b = n2^i$. We recover the dyadic wavelet transform $\psi_{m,n}(t)$ in Equation (G.3.11) with scale m replaced by i. Discretising the integral, we get

$$T(2^i, n2^i) = T_{i,n} = \sum_k x(k)\psi_{i,n}(k)$$

which are decimated wavelet transforms since the octave i is only output every 2^i samples. Hence, the resulting algorithms will not be translation invariant. To restore the invariance we can either filter separately the even and odd sequences, or we shall use the following algorithm

$$\begin{aligned} S_{i,n} &= \sum_j h_{2n-j} S_{i-1,j} \\ T_{i,n} &= \sum_j g_{n-j} S_{i-1,j} \end{aligned}$$

where S_0 is the original signal. Decimation is represented by the matrix $\Delta_{kj} = \delta(2k - j)$ with transpose $D_{kj} = \delta(k - 2j)$, and $(h^*)_k = h_{-k}$ is the adjoint filter. We now approximate the values at nonintegral points through interpolation via a finite filter h^*, where

$$\sum_k h^*_{n-2k} g(k) = \begin{cases} g(\frac{n}{2}) \text{ if } n \text{ is even} \\ \frac{1}{2}(g(\frac{n-1}{2}) + g(\frac{n+1}{2})) \text{ if } n \text{ is odd} \end{cases}$$

approximates a sampling of $g(\frac{t}{2})$. With the help of the dilation operator D, this method can be generalised, leading to

$$[h^* * (Dg^*)]_n = \sum_k h^*_{n-2k} g(k) \approx \frac{1}{\sqrt{2}} g(\frac{n}{2})$$

Inserting this approximation into the discrete sum, we get

$$T_{1,n} \approx \sum_{k,m} h^*_{k-2n-2m} g^*_m x(k) = [g * (\Delta(h * S))]_n$$

for $i = 1$. By induction, replacing S above with S_{i-1}, we obtain $T(2^i, n2^i) \approx T_{i,n}$ for all i. Note, we can rewrite the a trous algorithm as

$$\begin{aligned} S_{i+1} &= \Delta(h * S_i) \\ T_i &= g * S_i \end{aligned}$$

which is a DWT for which the filter h satisfies

$$h_{2k} = \frac{\delta(k)}{\sqrt{2}}$$

and the filter g is obtained by sampling an apriori wavelet function $g(t)$. When $T_{i,n}$ is replaced by $T_{i,2n}$ in the above system of equation, we recover the Mallat algorithm. Under certain regularity conditions, and some properties of the filters h and g, we get

$$T_{i,2n} = \int x(t) g(\frac{t}{2^i} - n) dt$$

provided that

$$S_{0,k} = \int x(t)\phi(t-k)dt \tag{G.4.25}$$

(approximation coefficients) where ϕ is related to $g(t)$ and the filter g by

$$g(t) = \sum_k \sqrt{2} g_{-k}\phi(2t-k)$$

Note, in Mallat algorithm, the sampled signal must lie in an appropriate subspace given by $S_{0,k}$. Further, it is necessary and sufficient for h to be a trous, for the condition in Equation (G.4.25) to be dropped.

G.4.2 Describing some redundant transforms

While wavelet transforms can be classified as either redundant or non-redundant (orthogonal), the major drawback of the latter is their non-invariance in time (or space). That is, the coefficients of a delayed signal are not a time shifted version of those of the original signal. The time invariance property is of importance in statistical signal processing applications, such as detection or parameter estimation of signals with unknown arrival time. The non-invariance implies that if a detector is designed in the wavelet coefficient domain, its performances will then depend on the arrival time of the signal. This is very problematic in finance where we try to make forecast directly on the smooth time series obtained by denoising the original signal. As a way around, practitioners used redundant transforms in detection/estimation problems. Pesquet et al. [1996] showed that it was possible to build different orthogonal wavelet representations of a signal while keeping the same analysing wavelet. These decompositions differ in the way the time-scale plane is sampled. Optimising the decomposition to best-fit the time localisation of the signal, they obtained an improved representation which is time invariant. We first review their approach and then follow Nason et al. [1995] to review the basics of discrete wavelet transform (DWT) using a filter notation which we use to describe extensions of the DWT, namely the ϵ-decimated DWT and the stationary wavelet transform (SWT). The main idea of these extensions is to fill the gaps caused by the decimation step in the standard wavelet transform, leading to an over-determined, or redundant, representation of the original data, but with considerable statistical potential.

G.4.2.1 The multiresolution analysis

We briefly review the multiresolution analysis detailed in Appendix (G.3.2). We can decompose a signal $x(t) \in L^2(\mathbb{R})$ with wavelet coefficients $\{T_a^b(x)\}_{(a,b)\in\mathcal{Z}^2}$ as in Equation (G.3.9) with normalised wavelet function given in Eqation (G.3.8). The orthonormal wavelet basis $\psi_{a,b}(t)$ can be built from a multiresolution analysis of $L^2(\mathbb{R})$. In this case, we need to sample the parameters a and b and consider the discrete step wavelet form $\psi_{j,k}(t)$ given in Equation (G.3.11) with scale a_0^j and translation step $\Delta b = b_0 a_0(j)$, for $a_0 = 2$ and $b_0 = 1$. In this setting, the wavelet coefficients $\{T_j^k(x)\}_{k\in\mathcal{Z}}$ become

$$T_{j,k} =< x, \psi_{j,k} >= \int_{-\infty}^{\infty} x(t)\psi_{j,k}^*(t)dt$$

where $\psi_{j,k}(t) = \frac{1}{2^{\frac{j}{2}}}\psi(\frac{t-k2^j}{2^j})$. Then, considering the scaling function $\phi_{j,k}(t)$ given in Equation (G.3.12), it is convolved with the signal to produce the approximation coefficients $\{S_{j,k}\}_{(j,k)\in\mathcal{Z}^2}$ in Equation (G.3.13). For well chosen filters $\{h_k\}_{k\in\mathbb{Z}}$ and $\{g_k\}_{k\in\mathbb{Z}}$, the mother wavelet ψ and the scaling function ϕ satisfy the two-scale equations

$$\begin{aligned} 2^{-\frac{1}{2}}\phi(\frac{t}{2}-k) &= \sum_{l=-\infty}^{\infty} h_{l-2k}\phi(t-l) \\ 2^{-\frac{1}{2}}\psi(\frac{t}{2}-k) &= \sum_{l=-\infty}^{\infty} g_{l-2k}\phi(t-l) \end{aligned}$$

and the approximation coefficients and wavelet coefficients satisfiy the Equations (G.3.19) and (G.3.20). Given the vector spaces $V_j = Span\{\phi_{j,k}(t), k \in \mathbb{Z}\}$ and $O_j = Span\{\psi_{j,k}(t), k \in \mathbb{Z}\}$, it results that

$$V_{j+1} = V_j \oplus^{\perp} O_j \text{[4]}$$

such that for every $j^* \in \mathbb{Z}$, $\{\psi_{j,k}(t), k \in \mathbb{Z}, j \leq j^*\} \cup \{\phi_{j^*,k}(t), k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

This decomposition, which is non-invariant in time (or space), implies that the wavelet coefficients of $\mathcal{T}_\tau[x(t)] = x(t-\tau)$ for $\tau \in \mathbb{R}$ are generally not delayed versions of $\{T_j^k(x)\}_{k\in\mathcal{Z}}$. One way forward is to consider the redundant decomposition of the signal $x(t)$ given by

$$\begin{aligned} \tilde{T}_{2^j}^{\theta}(x) &= T_j^{\theta 2^{-j}} = < x, \psi_{j,\theta 2^{-j}} > \\ \tilde{S}_{2^j}^{\theta}(x) &= S_j^{\theta 2^{-j}} = < x, \phi_{j,\theta 2^{-j}} > \ , \theta \in \mathbb{R}, j \in \mathbb{Z} \end{aligned} \tag{G.4.26}$$

where $\psi_{j,\theta 2^{-j}}(t) = \frac{1}{2^{\frac{j}{2}}}\psi(\frac{t-\theta}{2^j})$. This representation is time-invariant since the redundant wavelet and approximation coefficients of $\mathcal{T}_\tau[x(t)]$ are $\mathcal{T}_\tau[\tilde{T}_{2^j}^{\theta}(x)]$ and $\mathcal{T}_\tau[\tilde{S}_{2^j}^{\theta}(x)]$, respectively. One obvious way to construct/retrieve a signal from its wavelet representation is to select the subset of coefficients $\{T_j^k(x)\}_{(k,j)\in\mathcal{Z}^2}$ from the set of redundant wavelet coefficients and reconstruct the signal from its orthonormal wavelet representation. However, there exists many different ways of achieving this reconstruction as one can extract different orthonormal bases from the wavelet family $\{\psi_{j,\theta 2^{-j}}\ , \theta \in \mathbb{R}, j \in \mathbb{Z}\}$.

The wavelet packets are a generalisation of wavelets allowing for optimising the representation the signal (see Coifman et al. [1992]). We let $\omega_m(t)$, $m \in \mathbb{N}$ be a function of $L^2(\mathbb{R})$ such that $\int_{-\infty}^{\infty} \omega_0(t)dt = 1$ and for all $k \in \mathbb{Z}$

$$\begin{aligned} 2^{-\frac{1}{2}}\omega_{2m}(\frac{t}{2} - k) &= \sum_{l=-\infty}^{\infty} h_{l-2k}\omega_m(t-k) \\ 2^{-\frac{1}{2}}\omega_{2m+1}(\frac{t}{2} - k) &= \sum_{l=-\infty}^{\infty} g_{l-2k}\omega_m(t-k) \end{aligned}$$

where $\{h_k\}_{k\in\mathbb{Z}}$ and $\{g_k\}_{k\in\mathbb{Z}}$ satisfy the QMF filters. For every $j \in \mathbb{Z}$, given the vector space $\Omega = Span\{\omega_{j,m}^k, k \in \mathbb{Z}\}$ with $\omega_{j,m}^k = \omega_m(\frac{t-k2^j}{2^j})$, we can show that

$$\Omega_{j,m} = \Omega_{j+1,2m} \oplus^{\perp} \Omega_{j+1,2m+1}$$

Hence, if we let $\mathcal{P}$ be a partition [5] of $\mathbb{R}^+$ into the intervals $I_{j,m} = [2^{-j}m, .., 2^{-j}(m+1)[$, $j \in \mathbb{Z}$ and $m \in \{0, .., 2^j - 1\}$, then $\{2^{-\frac{j}{2}}\omega_{j,m}^k, k \in \mathbb{Z}, (j,m)/I_{j,m} \in \mathcal{P}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. Such basis is called wavelet packet, where the coefficients resulting from the decomposition of a signal $x(t)$ are

$$C_{j,m}^k =< x, \omega_{j,m}^k >$$

such that by varying the partition $\mathcal{P}$, differentchoices of wavelet packets are possible. For instance, we can set $\phi(t) = \omega_0(t)$ and $\psi(t) = \omega_1(t)$, getting $V_j = \Omega_{j,0}$ and $O_j = \Omega_{j,1}$. This structure can be described by a binary tree (called frequency tree) with nodes indexed by (j,m) and leaves on that node satisfying $I_{j,m} \in \mathcal{P}$. One should select the partition $\mathcal{P}$ for which an optimised representation of the analysed signal is obtained. For instance, one can choose the entropy criteria

[4] The symbol $\oplus^{\perp}$ means the orthogonal sum of vector spaces.
[5] A partition $\mathcal{P}$ of a set $\mathcal{B}$ is a set of nonempty disjoints subsets whose union is $\mathcal{B}$.

$$\mathcal{H}(\{\alpha_k\}_{k\in\mathbb{Z}}) = -\sum_k P_k \ln P_k$$

where $P_k = \frac{|\alpha_k|^2}{\sum_l |\alpha_l|^2}$ and $\{\alpha_k\}_{k\in\mathbb{Z}}$ is the sequence of coefficients of the decomposition in a given basis. Coifman et al. proposed a binary tree search method to find the wavelet packet minimising a given criterion $\mathcal{H}(\bullet)$.

Pesquet et al. [1996] proposed an extension of the wavelet packet decomposition and showed how to achieve time invariance while preserving the orthonormality.

Proposition 22 *Let two vector spaces be defined as*

$$\begin{aligned} V_{j,p} &= span\{\phi_{j,k}(t-p)\,,\, k\in\mathbb{Z}\} \\ O_{j,p} &= span\{\psi_{j,k}(t-p)\,,\, k\in\mathbb{Z}\} \end{aligned}$$

for $j\in\mathbb{N}$ *and* $p\in\{0,..,2^j-1\}$. *It follows that*

$$V_{j,p} = V_{j+1,p} \oplus^{\perp} O_{j+1,p} = V_{j+1,p+2^j} \oplus^{\perp} O_{j+1,p+2^j} \tag{G.4.27}$$

$\{\phi_{j,k}(t-p), k\in\mathbb{Z}\}$ *and* $\{\psi_{j,k}(t-p), k\in\mathbb{Z}\}$ *being respectively orthonormal bases of* $V_{j,p}$ *and* $O_{j,p}$.

While two different orthonormal bases are possible for decomposing the space $V_{j,p}$ at the next lower resolution 2^{-j-1}, they differ in the time localisation of the basis functions. A binary tree can be used to describe the different possible choices at each resolution level j where each node is indexed by parameters (j,p). Note, the redundant wavelet coefficients $\{\tilde{T}^k_{2^j}\}_{k\in\mathbb{Z}}$, for $j\geq 1$, may be structured according to this tree by associating the set $\{\tilde{T}^{k2^j+p}_{2^j}\}_{k\in\mathbb{Z}}$ to the node (j,p). That is, given $\tilde{T}^{\theta}_{2^j}$ in Equation (G.4.26), with $\theta = k2^j+p$, we obtain $T^{k+p2^{-j}}_j$ with

$$\psi_{j,k+p2^{-j}}(t) = \frac{1}{2^{\frac{j}{2}}}\psi(\frac{t-p-k2^j}{2^j}) = \frac{1}{2^{\frac{j}{2}}}\psi(\frac{t-p}{2^j}-k) = \psi_{j,k}(t-p)$$

where $\{\psi_{j,k}(t-p), k\in\mathbb{Z}\}$ corresponds to $\{\psi_{j,k+p2^{-j}}(t), k\in\mathbb{Z}\}$. Assuming that the multiscale decomposition is performed on j^* levels, there exists (at least) 2^{j^*} different ways of reconstructing a given signal. From Proposition (22), the reconstruction may be recursively obtained by using the relation

$$\tilde{S}^{2^jk+p}_{2^j} = \sum_{l=-\infty}^{\infty} h_{k-2l}\tilde{S}^{2^{j+1}l+p}_{2^{j+1}} + \sum_{l=-\infty}^{\infty} g_{k-2l}\tilde{T}^{2^{j+1}l+p}_{2^{j+1}}$$

This filter bank is generally visualised in a graph where the operator $2\uparrow$ is an interpolator by a factor 2, that is, its inputs $\{e_k\}_{k\in\mathbb{Z}}$ and its output $\{s_k\}_{k\in\mathbb{Z}}$ satisfy

$$s_k = \begin{cases} e_{\frac{k}{2}} \text{ if k is even} \\ 0 \text{ if k is odd} \end{cases}$$

Similarly, the reconstruction may also be recursively obtained by using the relation

$$\tilde{S}^{2^jk+p}_{2^j} = \sum_{l=-\infty}^{\infty} h'_{k-2l}\tilde{S}^{2^{j+1}l+p+2^j}_{2^{j+1}} + \sum_{l=-\infty}^{\infty} g'_{k-2l}\tilde{T}^{2^{j+1}l+p+2^j}_{2^{j+1}}$$

The filter bank is now visualised with the operator $2\uparrow'$ whose inputs $\{e_k\}_{k\in\mathbb{Z}}$ and output $\{s_k\}_{k\in\mathbb{Z}}$ satisfy

$$s_k = \begin{cases} e_{\frac{k-1}{2}} \text{ if k is odd} \\ 0 \text{ if k is even} \end{cases}$$

These filter banks may be associated to the dual filter banks corresponding to the decimator by a factor 2, $2\downarrow$ (resp. $2\downarrow'$) is such that the output is obtained from its input by

$$s_k = e_{2k} \text{ resp. } s_k = e_{2k+1}$$

and retaining the even samples amounts to the decimation commonly used in any orthonormal wavelet decomposition. Hence, we can give a simple digital filtering interpretation to the relation in Equation (G.4.27). This show that the original signal can be reconstructed from its redundant wavelet decomposition by correctly selecting different sets of orthonormal coefficients. We are now left with choosing a criterion, such as the entropy, to select the best set of orthonormal coefficients. Considering the entropy criterion with extra additivity property, Pesquet et al. proposed an efficient optimisation algorithm.

G.4.2.2 The standard DWT

The discrete wavelet transform (DWT) for the sequence $x_0, .., x_{N-1}$, with $N = 2^J$ for some integer J is based on filters $\mathcal{H}$ and $\mathcal{G}$, and on a binary operator $\mathcal{D}_0$. The filter $\mathcal{H}$ is a lowpass filter defined by the sequence $\{h_k\}$ where only few values are non-zero. Its action is defined by

$$(\mathcal{H}x)_k = \sum_l h_{l-k} x_l$$

The definitions for sequences of finite length depending on the choice of treatment at the boundaries, we assume periodic boundary conditions. The filter satisfies the internal orthogonality relation

$$\sum_k h_k h_{k+2j} = 0$$

for all integers $j \neq 0$, and have the sum of squares $\sum_k h_k^2 = 1$. The highpass filter $\mathcal{G}$ is defined by the sequence

$$g_k = (-1)^k h_{1-k} \, , \forall k$$

It satisfies the same internal orthogonality relations as $\mathcal{H}$, and it also obeys the mutual orthogonality relation

$$\sum_k h_k g_{k+2j} = 0 \, , \forall j$$

These filters are called quadratic mirror filters. The binary decimation operator $\mathcal{D}_0$ chooses every even member of a sequence, and is defined as

$$(\mathcal{D}_0 x)_j = x_{2j} \, , \forall j$$

From the properties of the quadratic mirror filters, the mapping of a sequence x to the pair of sequences $(\mathcal{D}_0\mathcal{G}x, \mathcal{D}_0\mathcal{H}x)$ is an orthogonal transformation. Hence, given the finite sequence x of length 2^J, with periodic boundary conditions, each of $\mathcal{D}_0\mathcal{G}x$ and $\mathcal{D}_0\mathcal{H}x$ is a sequence of length 2^{J-1}. In the multiresolution analysis, we define the smooth (approximation) at level J, written c_J, to be the original data

$$c_{J,k} = x_k \text{ for } k = 0, ..., 2^J - 1$$

For $j = J-1, ..., 0$ we define recursively the smooth c_j at level j and the detail d_j at level j by

$$c_j = \mathcal{D}_0\mathcal{H}c_{j+1} \text{ and } d_j = \mathcal{D}_0\mathcal{G}c_{j+1}$$

where c_j and d_j are sequences of length 2^j. We see that the smooth at each level is fed down to the next level, giving the new smooth and detail at that level. Since the mapping $(\mathcal{D}_0\mathcal{G}x, \mathcal{D}_0\mathcal{H}x)$ is an orthogonal transform, it can easily be

inverted to find c_{j+1} in terms of c_j and d_j. To do so, we write the transform as a matrix and transpose it. Writing $\mathcal{R}_0$ for the inverse transform, we get

$$c_{j+1} = \mathcal{R}_0(c_j, d_j) \text{ for each } j$$

Continuing this process, we obtain the detail at each level with the smooth at the zero level, so that the original sequence is orthogonally transformed to the sequence of sequences

$$d_{J-1}, d_{J-2}, .., d_0, c_0$$

of total length 2^J. The process can be reversed by reconstructing c_1 from d_0 and c_0, then c_2 from d_1 and c_1, and so on. Given that $\{h_k\}$ has a finite number of non-zero elements, the overall number of arithmetic operations for both the transform and its inverse is $\mathcal{O}(2^J)$. Stopping the process at any level R will give the sequence of sequences $d_{J-1}, d_{J-2}, .., d_R, c_R$ called DWT curtailed at level R. Note, Nason et al. constructed bases by dilating and translating ϕ and ψ according to

$$\phi_j(t) = 2^{\frac{j}{2}}\phi(2^j t) \text{ and } \psi_j(t) = 2^{\frac{j}{2}}\psi(2^j t)$$

so that for a function f and sequence c_J we get

$$f(t) = \sum_k c_{J,k}\phi_J(t - 2^{-J}k)$$

and the expansion for $R < J$ of a function f in orthonormal functions is given by

$$f(t) = \sum_k c_{R,k}\phi_R(t - 2^{-R}k) + \sum_{j=R}^{J-1}\sum d_{j,k}\psi_j(t - 2^{-j}k) = \sum_k c_{R,k}2^{\frac{R}{2}}\phi(2^R t - k) + \sum_{j=R}^{J-1}\sum d_{j,k}2^{\frac{j}{2}}\psi(2^j t - k)$$

and

$$d_{j,k} = \int \psi_j(t - 2^{-j}k)f(t)dt = \int 2^{\frac{j}{2}}\psi(2^j t - k)f(t)dt$$

so that the detail coefficient $d_{j,k}$ gives information about f at scale 2^{-j} near position $t = 2^{-j}k$. In terms of the original sequence c_J, it corresponds to scale 2^{J-j} near position $2^{J-j}k$.

G.4.2.3 The ϵ-decimated DWT

The DWT being an orthogonal transform, it corresponds to a particular choice of basis for the space $\mathbb{R}^N$ where the original sequence lies. As a result, depending on the choice of the basis, we can get modifications of the DWT. We could also select every odd number of each sequence with the operator $(\mathcal{D}_1 x)_j = x_{2j+1}$, getting the mapping $(\mathcal{D}_1\mathcal{G}x, \mathcal{D}_1\mathcal{H}x)$ which is an orthogonal transformation. Reconstruction would be obtained by successive application of the corresponding inverse operator, $\mathcal{R}_1$. Further, if we let $\epsilon_{J-1}, .., \epsilon_0$ be a sequence of 0 and 1, we can then use the operator $\mathcal{D}_{\epsilon_j}$ at level j, and perform the reconstruction by using the corresponding sequence of operators $\mathcal{R}_{\epsilon_j}$, giving a different orthogonal transformation from the original sequence for each choice of the sequence ϵ. This transformation is called the ϵ-decimated DWT. To understand the mechanism, we consider the shift operator $\mathcal{S}$ defined by

$$(\mathcal{S}x)_j = x_{j+1}$$

such that $\mathcal{D}_1 = \mathcal{D}_0\mathcal{S}$ and, thus, $\mathcal{R}_1\mathcal{S}^{-1}\mathcal{R}_0$. We can also see that $\mathcal{S}\mathcal{D}_0 = \mathcal{D}_0\mathcal{S}^2$ and that the operator $\mathcal{S}$ commutes with $\mathcal{H}$ and $\mathcal{G}$. We now let S be the integer with binary representation $\epsilon_0\epsilon_1...\epsilon_{J-1}$, one can show that the coefficient sequences c_j and d_j yielded by the ϵ-decimated DWT are all shifted versions of the original DWT applied to the shifted sequence $\mathcal{S}^S x$. For example, fixing j, we let s_1 and s_2 be the integers with binary representations $\epsilon_0\epsilon_1...\epsilon_{j-1}$

and $\epsilon_j\epsilon_{j+1}...\epsilon_{J-1}$. In DWT the sequence $d_j = \mathcal{D}_0\mathcal{G}(\mathcal{D}_0\mathcal{H})^{J-j-1}c_J$, while in the ϵ-decimated case we get $d_j = \mathcal{D}_0\mathcal{G}(\mathcal{D}_0\mathcal{H})^{J-j-1}\mathcal{S}^{s_2}c_J$. Applying the operator $\mathcal{S}^{s_1}$, we get $\mathcal{S}^{s_1}d_j = \mathcal{D}_0\mathcal{G}(\mathcal{D}_0\mathcal{H})^{J-j-1}\mathcal{S}^S c_J$ since $S = 2^{J-j}s_1 + s_2$. Thus, d_j shifted by s_1 is the jth detail sequence of the DWT applied to the original data shifted by an amount S. The result for c_j can be similarly derived. Hence, the basis vectors of the ϵ-decimated DWT can be obtained from those of the DWT by applying the shift operator $\mathcal{S}^S$, and the choice of ϵ corresponds to a choice of origin with respect to which the basis functions are defined. If we let $t_0 = 2^{-J}S$, the coefficient sequences obtained will give an expansion of f in terms of $\phi_R(t - t_0 - 2^{-R}k)$ and $\psi_j(t - t_0 - 2^{-j}k)$ for integers k and for $j = R, R+1, .., J-1$. In terms of the original sequence, it corresponds to scale 2^{J-j} near position $2^{J-j}k + S$, a grid of integers of gauge 2^{J-j} shifted to have origin S.

G.4.2.4 The stationary wavelet transform

In the stationary wavelet transform (SWT) we do not decimate but simply apply appropriate high and low pass filters to the data at each level to produce two sequences at the next level each having the same length as the original one. At each level, the filters are modified by padding them out with zeros. That is, the operator $\mathcal{Z}$ alternates a given sequence with zeros, so that for all integers j, we get $(\mathcal{Z}x)_{2j} = x_j$ and $(\mathcal{Z}x)_{2j+1} = 0$. We define the filters $\mathcal{H}^{[r]}$ and $\mathcal{G}^{[r]}$ to have weights $\mathcal{Z}^r h$ and $\mathcal{Z}^r g$, respectively, such that $\mathcal{H}^{[r]}$ has weights

$$\begin{aligned} \mathcal{H}^{[r]}_{2^r} &= h_j \\ \mathcal{H}^{[r]}_k &= 0 \text{ if k is not a multiple of } 2^r \end{aligned}$$

Thus, the filter $\mathcal{H}^{[r]}$ is obtained by inserting a zero between every adjacent pair of elements of the filter $\mathcal{H}^{[r-1]}$, and similarly for $\mathcal{G}^{[r]}$. Hence, both $\mathcal{H}^{[r]}$ and $\mathcal{G}^{[r]}$ commute with $\mathcal{S}$, and we get

$$\mathcal{D}_0^r\mathcal{H}^{[r]} = \mathcal{H}\mathcal{D}_0^r \text{ and } \mathcal{D}_0^r\mathcal{G}^{[r]} = \mathcal{G}\mathcal{D}_0^r$$

As a result, setting a_J to be the original sequence, and b_j be the detail, for $j = J-1, J-2, .., , 1$, we obtain recursively

$$a_{j-1} = \mathcal{H}^{[J-1]}a_j \text{ and } b_{j-1} = \mathcal{G}^{[J-1]}a_j$$

Given the vector a_J of length 2^J, then all the vectors a_j and b_j are of the same length, rather than decreasing as in the DWT. Thus, to find

$$b_{J-1}, b_{J-1}, .., b_0, a_0$$

takes $\mathcal{O}(J2^J)$ operations compared with $\mathcal{O}(2^J)$ with the DWT. One can show that SWT contains the coefficients of the ϵ-decimated DWT for every choice of ϵ. That is, for any given ϵ and corresponding origin S, the details at level j are a shifted version of $\mathcal{D}_0^{J-1}\mathcal{S}^S b_j$, and similarly for the data $\mathcal{D}_0^{J-1}\mathcal{S}^S a^j$. Given the same example as in Section (G.4.2.3), we let $d_j(\epsilon)$ be the j detail sequence obtained from the ϵ-decimated DWT. We then have for each j, $\mathcal{S}^{-s_1}\mathcal{D}_0^{J-1}\mathcal{S}^S b_j = d_j$, and the same link exists for a_j and c_j. Associating x_J with the function f, for any j and k, considering a decimated DWT with $S = k$ and $t_0 = 2^{-J}k$, we get

$$b_{j,k} = \int \psi_j(t - 2^{-J}k)f(t)dt$$

which gives information at scale 2^{J-1} localised at position k. There is no-longer any restriction of the localisation position to a grid of integers.

G.4.3 The autocorrelation functions of compactly supported wavelets

Even though the coefficients of the orthonormal wavelet expansions are of finite size (allowing for exact computer implementation), they are not shift invariant and they also have asymmetric shape. Hence, symmetric basis functions are preferred since their use simplifies finding zero-crossings (or extrema) corresponding to the locations of edges in images at later stages of processings. One way forward is to construct approximately symmetric orthonormal wavelets giving rise to approximate QMF (see Mallat [1989]), or to use biorthogonal bases so that the basis functions may be chosen to be exactly symmetric (see Cohen et al. [1992]). An alternative solution proposed by Beylkin et al. [1992] is to use a redundant representation using dilations and translations of the auto-correlation functions of compactly supported wavelets. In that setting, the decomposition filters are exactly symmetric. Hence, rather than using the wavelets, the auto-correlation shell are used for signal analysis. The recursive definition of the auto-correlation functions of compactly supported wavelets leads to fast recursive algorithms to generate the multiresolution representation (see Saito et al. [1993]).

We let $\Phi(x)$ be the auto-correlation function

$$\Phi(x) = \int_{-\infty}^{\infty} \phi(y)\phi(y-x)dy \tag{G.4.28}$$

where $\phi(\bullet)$ is a scaling function (used in wavelet analysis), and corresponds to the fundamental function of the symmetric iterative interpolation scheme (see Dubuc [1986]). Thus, there is a one-to-one correspondence between the iterative interpolation schemes and compactly supported wavelets. In general, the scaling functions corresponding to Daubechies' wavelets with M vanishing moments lead to an iterative interpolation schemes using the Lagrange polynomials of degree $L = 2M$ (see Deslauriers et al. [1989]). We are now going to derive the two-scale difference equation for the function $\Phi(x)$. Let $m_0(\xi)$ and $m_1(\xi)$ be the 2π-periodic functions

$$m_0(\xi) = \frac{1}{\sqrt{2}} \sum_{k=0}^{L-1} h_k e^{ik\xi}$$

and

$$m_1(\xi) = \frac{1}{\sqrt{2}} \sum_{k=0}^{L-1} g_k e^{ik\xi} = e^{i(\xi+\pi)} m_0^*(\xi+\pi)$$

satisfying the quadrature mirror filter (QMF) condition

$$|m_0(\xi)|^2 + |m_1(\xi)|^2 = 1$$

Given the trigonometric polynomial solutions to $m_i(\xi)$ for $i = 0, 1$, we get

$$|m_0(\xi)|^2 = \frac{1}{2} + \frac{1}{2} \sum_{k=1}^{\frac{L}{2}} a_{2k-1} \cos(2k-1)\xi$$

where $\{a_k\}$ are the auto-correlation coefficients of the filter $H = \{h_k\}_{0 \leq k \leq L-1}$

$$a_k = 2 \sum_{l=0}^{L-1-k} h_l h_{l+k} \text{ for } k = 1, .., L-1$$

and

$$a_{2k} = 0 \text{ for } k = 1, .., \frac{L}{2} - 1$$

Using the scaling equation (G.3.17), we obtain

$$\Phi(x) = \Phi(2x) + \frac{1}{2}\sum_{l=1}^{\frac{L}{2}} a_{2l-1}\big(\Phi(2x-2l+1) + \Phi(2x+2l-1)\big) \tag{G.4.29}$$

We can then introduce the autocorrelation function of the wavelet

$$\Psi(x) = \int_{-\infty}^{\infty} \psi(y)\psi(y-x)dy$$

and repeat the same process. Note, both $\Phi(x)$ and $\Psi(x)$ are supported within the interval $[-L+1, L-1]$, and they have vanishing moments

$$\begin{aligned}\mathcal{M}_{\Psi}^{m} &= \int_{-\infty}^{\infty} x^m \Psi(x)dx = 0 \text{ for } 0 \leq m \leq L-1 \\ \mathcal{M}_{\Phi}^{m} &= \int_{-\infty}^{\infty} x^m \Phi(x)dx = 0 \text{ for } 1 \leq m \leq L-1\end{aligned}$$

and $\int_{-\infty}^{\infty} \Phi(x)dx = 1$. One can also show that even moments of the coefficients a_{2k-1} vanish, that is,

$$\sum_{k=1}^{\frac{L}{2}} a_{2k-1}(2k-1)^{2m} = 0 \text{ for } 1 \leq m \leq M-1$$

where $L = 2M$. Since L consecutive moments of the autocorrelation function $\Psi(x)$ vanish, we have

$$\hat{\Psi}(\xi) = O(\xi^l)$$

where $\hat{\Psi}(\xi)$ is the Fourier transform of $\Psi(\xi)$. As a result, $\hat{\Psi}(\xi)$ can be seen as an approximation of the derivative operator $(\frac{d}{dx})^L$, such that convolution with $\Psi(x)$ behaves like a differential operator in detecting changes of spatial intensity, and it is designed to act at any desired scale.

We can now relate the autocorrelation function in Equation (G.4.28) to iterative interpolation scheme. Given values of $f(x)$ on set B_0, where B_n is the set of dyadic rationals $\frac{m}{2^n}$ for $m = 0, 1, ..,$ Dubuc [1986] proposed to extend f to $B_1, B_2, ..$ in an iterative manner. He suggested computing

$$f(x) = \frac{9}{16}\big(f(x-h) + f(x+h)\big) - \frac{1}{16}\big(f(x-3h) + f(x+3h)\big)\,, h = \frac{1}{2^{n+1}}$$

It was further generalised by Deslauriers et al. [1989] to

$$f(x) = \sum_{k\in\mathbb{Z}} F(\frac{k}{2}) f(x+kh) \text{ for } x \in B_{n+1}/B_n \text{ and } h = \frac{1}{2^{n+1}}$$

where the function $F(\frac{k}{2})$ satisfy

$$F(\frac{x}{2}) = \sum_{k\in\mathbb{Z}} F(\frac{k}{2})F(x-k)$$

Using the Lagrange polynomials with $L = 2M$ nodes, we get

$$f(x) = \sum_{k=1}^{M} \mathcal{P}_{2k-1}^{L}(0)\big(f(x-(2k-1)h) + f(x+(2k-1)h)\big)$$

where $\{\mathcal{P}^L_{2k-1}(x)\}_{-M+1\leq k\leq M}$ is a set of the Lagrange polynomials of degree $L-1$ with nodes $\{-L+1,-L+3,...,L-3,L-1\}$ given by

$$\mathcal{P}^L_{2k-1}(x) = \prod_{l=-M+1,l\neq k}^{M} \frac{x-(2l-1)}{(2k-1)-(2l-1)}$$

and we get the fundamental function F_L

$$F_L(x) = F_L(2x) + \sum_{k=1}^{M} \mathcal{P}^L_{2k-1}(0)\big(F_L(2x-2k+1) + F_L(2x+2k-1)\big)$$

which is a special case of $f(x)$ called the Lagrange iterative interpolation. Note, setting $L=4$ we recover $f(x)$. Thus, we have

$$F(x) = \Phi(x)$$

and using the two-scale difference equation, we get

$$\Phi(\frac{k}{2}) = \Phi(k) + \frac{1}{2}\sum_{l\in\mathbb{N}}^{M} a_{2l-1}\big(\Phi(k-2l+1) + \Phi(k+2l+1)\big)$$

and therefore

$$\Phi(\frac{k}{2}) = \frac{a_k}{2}$$

As a result, the two-scale difference equation for the function Φ becomes

$$\Phi(\frac{x}{2}) = \sum_{k\in\mathbb{Z}} \Phi(\frac{k}{2})\Phi(x-k) \tag{G.4.30}$$

It may beshown that the QMF relation for the periodic function $m_0(\xi)$ can be rewritten as

$$|m_0(\xi)|^2 = \frac{1}{2} + \frac{1}{2}\big[\frac{(2M-1)!}{(M-1)!4^{M-1}}\big]^2 \sum_{k=1}^{M} \frac{(-1)^{k-1}\cos(2k-1)\xi}{(2k-1)(M-m)!(M+k-1)!}$$

and if $M\to\infty$, then

$$|m_0(\xi)|^2 = \frac{1}{2} + \frac{1}{2}\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{2k-1}\cos(2k-1)\xi$$

which is the Fourier series of the characteristic function of $[-\frac{\pi}{2},\frac{\pi}{2}]$. Hence, the corresponding autocorrelation function is

$$\Phi_\infty(x) = sinc(x) = \frac{\sin\pi x}{\pi x}$$

so that if the number M of vanishing moments of the compactly supported wavelets approaches infinity, then $\phi_\infty(x) = sinc(x)$. As a result, we get the relation

$$\phi_\infty(x) = \Phi_\infty(x)$$

and

$$\sqrt{2}h_k = \frac{a_k}{2} = \frac{\sin\frac{\pi k}{2}}{\frac{\pi k}{2}} \text{ for } k \in \mathbb{Z}$$

We then obtain a family of the symmetric iterative interpolation schemes parametrised by the number of vanishing moments $1 \leq M < \infty$. Further, the derivative of the function $f(x)$ is computed via

$$f^{'}(x) = \sum_{k=1}^{L-2} r_k\big(f(x+kh) - f(x-kh)\big)$$

where $h = 2^{-n}$, $x \in B_m$, where $m \leq n$, and

$$r_k = \int_{-\infty}^{\infty} \phi(x-k)\frac{d}{dx}\phi(x)dx$$

The coefficient r_k may be computed by solving

$$r_k = 2\big[r_{2k} + \frac{1}{2}\sum_{l=1}^{\frac{L}{2}} a_{2l-1}\big(r_{2k-2l+1} + r_{2k+2l-1}\big)\big]$$

and

$$\sum_{k\in\mathbb{Z}} kr_k = -1$$

where a_{2l-1} are given above. If the number of vanishing moments of the wavelet are such that $M \geq 2$, then the above equations have a unique solution with a finite number of non-zero r_k. That is, $r_k \neq 0$ for $-L+2 \leq k \leq L-2$ and

$$r_k = -r_{-k}$$

We now present the recursive algorithm proposed by Saito et al. [1993] to generate the multiresolution representation. We assume that the finest scale of interest is described by the $N = 2^M$ dimensional subspace $V_0 \subset L^2(\mathbb{R})$, and we only consider circulant shifts on V_0. We call the set of functions $\{\Psi_{j,k}(x)\}_{1\leq j\leq m_0, 0\leq k\leq N-1}$ and $\{\Phi_{m_0,k}(x)\}_{0\leq k\leq N-1}$ the shell of the auto-correlation functions of orthonormal wavelets, where m_0 ($\leq M$) is the coarsest scale of interest and

$$\begin{aligned} \Phi_{j,k}(x) &= 2^{-\frac{j}{2}}\Phi(2^{-j}(x-k)) \\ \Psi_{j,k}(x) &= 2^{-\frac{j}{2}}\Psi(2^{-j}(x-k)) \end{aligned} \tag{G.4.31}$$

We now need a fast algorithm to expand the function $f \in V_0 = Span(\phi^*_{0,k} : k \in \mathbb{Z})$, $f = \sum_{k=0}^{N-1} S^0_k\phi_{0,k}$. We let the coefficients $\{p_k\}$ and $\{q_k\}$ be those of the two-scale difference equations (G.4.29) with solution in Equation (G.4.30), which we rewrite as

$$p_k = \begin{cases} 2^{-\frac{1}{2}} \text{ for } k = 0 \\ 2^{-\frac{3}{2}}a_{|k|} \text{ otherwise} \end{cases}$$

and

$$q_k = \begin{cases} 2^{-\frac{1}{2}} \text{ for } k = 0 \\ -p_k \text{ otherwise} \end{cases}$$

which we use as symmetric filters $P = \{p_k\}_{-L+1 \leq k \leq L-1}$ and $Q = \{q_k\}_{-L+1 \leq k \leq L-1}$ with only $\frac{L}{2} + 1$ distinct non-zero coefficients. As an example of coefficients $\{p_k\}$, for Daubechies' wavelets with two vanishing moments and $L = 4$, the coefficients are $2^{-\frac{1}{2}}\left(-\frac{1}{16}, 0, \frac{9}{16}, 1, \frac{9}{16}, 0, -\frac{1}{16}\right)$. Note, these filters do not form a QMF pair, but their role and use is similar in the algorithm. For the shift-invariance, we apply P and Q without subsampling at each scale, getting

$$\begin{aligned} S_k^j &= \sum_{l=-L+1}^{L-1} p_l S_{k+2^{j-1}l}^{j-1} \\ D_k^j &= \sum_{l=-L+1}^{L-1} q_l S_{k+2^{j-1}l}^{j-1} \end{aligned} \tag{G.4.32}$$

where S_k^j are the residuals and D_k^j the details. Starting from the original discrete signal $\{S_k^0\}_{0 \leq k \leq N-1}$ we apply Equations (G.4.32) recursively and we obtain the auto-correlation shell coefficients $\{D_k^j\}_{1 \leq j \leq m_0}$ and $\{S_k^{m_0}\}_{0 \leq k \leq N-1}$. Saito et al. established a relation between the original discrete signal and the auto-correlation shell.

Proposition 23 *For any function $f \in V_0$, $f(x) = \sum_{k=0}^{N-1} S_k^0 \phi(x-k)$, the coefficients $\{S_k^j\}$ and $\{D_k^j\}$ computed with Equations (G.4.32) satisfy the following identities*

$$\begin{aligned} \sum_{k=0}^{N-1} S_k^j \Phi_{0,k} &= \sum_{k=0}^{N-1} S_k^0 \Phi_{j,k} \\ \sum_{k=0}^{N-1} D_k^j \Phi_{0,k} &= \sum_{k=0}^{N-1} S_k^0 \Psi_{j,k} \end{aligned}$$

where $\Phi_{j,k}$ and $\Psi_{j,k}$ are defined in Equations (G.4.31).

Since $p_k = -q_k$ for $k \neq 0$, adding Equations (G.4.32) yields a simple reconstruction formula

$$S_k^{j-1} = \frac{1}{\sqrt{2}}(S_k^j + D_k^j) \text{ for } j = 1, .., m_0 \text{ , } k = 0, ..., N-1$$

That is, given a smoothed signal at two consecutive resolution levels, the detailed signal can be derived as

$$D_k^j = \sqrt{2} S_k^{j-1} - S_k^j$$

Hence, given the auto-correlation shell coefficients $\{D_k^j\}_{1 \leq j \leq m_0, 0 \leq k \leq N-1}$ and $\{S_k^{m_0}\}_{0 \leq k \leq N-1}$, the above equation leads to

$$S_k^0 = 2^{-\frac{m_0}{2}} S_k^{m_0} + \sum_{j=1}^{m_0} 2^{-\frac{j}{2}} D_k^j \text{ for } k = 0, ..., N-1$$

At each scale j, we obtain the set of detail coefficients $\{D^j\}$ having the same number of samples as the original signal, and the set $\{S^{m_0}\}$ which provides the residual. Adding D^j for $j = m_0, m_0 - 1, ..$ gives an increasingly more accurate approximation of the original signal. The additive form of the reconstruction allows one to combine the predictions in a simple additive manner.

Note, representations using the auto-correlation functions of compactly supported wavelets can also be viewed as a way of obtaining a continuous multiresolution analysis. Further, since they can also be viewed as pseudo-differential

operators of even order, the zero-crossings in that setting corresponds to the location of edges at different scales in the signal. At last, this approach can be modified to produce the maxima-based representation of Mallat et al. [1992b] by considering $\int_{-\infty}^{x} \Psi(y)dy$ instead of $\Psi(x)$ and the corresponding two-scale difference equation.

Bibliography

[2012] Abernethy J., Chen Y., Vaughan J.W., Efficient market making via convex optimization, and a connection to online learning. Working Paper, University of California, Berkeley.

[1993] Abry P., Goncalves P., Flandrin P., Wavelet-based spectral analysis of $1/f$ processes. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **3**, pp 237–240.

[1995] Abry P., Goncalves P., Flandrin P., Wavelets, spectrum estimation and $1/f$ processes. in A. Antoniadis and G. Oppenheim, eds, *Lecture Notes in Statistics: Wavelets and Statistics*, Springer-Verlag, pp 15–30.

[1996] Abry P., Sellan F., The wavelet-based synthesis for fractional Brownian motion. proposed by F. Sellan and Y. Meyer, Remarks and Fast Implementation, *Applied and Computational Harmonic Analysis*, **3**, (4), pp 377–383.

[1998] Abry P., Veitch D., Wavelet analysis of long-range-dependent traffic. *IEEE Transaction on Information Theory*, **44**, (1), pp 2–15.

[1999] Abry P., Sellan F., A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Transaction on Information Theory*, **45**, (3), pp 878–897.

[2000] Abry P., Flandrin P., Taqqu M.S., Veitch D., Wavelets for the analysis, estimation, and synthesis of scaling data. in *Self-similar Network Traffic and Performance Evaluation*, ed. K. Park, W. Willinger, Wiley, pp 39–87.

[2002] Abry P., Baraniuk R., Flandrin P., Riedi R., Veitch D., Multiscale nature of network traffic. *IEEE Signal Processing Magazine*, **19**, (3), pp 28–46.

[2001] Abu-Mostafa Y.S., Atiya A.F., Magdon-Ismail M., White H., Neural networks in financial engineering. *IEEE Transactions on Neural Networks*, **12**, (4), pp 653–656.

[2002] Addison P.S., The illustrated wavelet transform handbook. Taylor & Francis Group, New York.

[1998] Adya M., Collopy F., How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, **17**, pp 481–495.

[2004] Agarwal V., Naik N.Y., Risk and portfolio decisions involving hedge funds. *The Review of Financial Studies*, **17**, (1), pp 63–98.

[2011] Agarwal S., Delage E., Peters M., Wang Z., Ye Y., A unified framework for dynamic prediction market design. *Operations Research*, **59**, (3), pp 550–568.

[2010] Ahmed N.K., Atiya A.F., El Gayar N., El-Shishiny H., An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, **29**, (5), pp 594–621.

[2003] Ahn D.H., Conrad J., Dittmar R.F., Risk adjustment and trading strategies. *The Review of Financial Studies*, **16**, pp 459–485.

[1998] Ait-Sahalia Y., Lo A.W., Nonparametric estimation of state price densities implicit in financial asset prices. *Journal of Finance*, **53** pp 499–547.

[2000] Ait-Sahalia Y., Lo A.W., Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, **94** pp 9–51.

[1973] Akaike H., Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.). Second international symposium on information theory (pp. 267–281). Budapest: Academiai Kiado

[2009] Aldridge I., High-frequency trading: A practical guide to algorithmic strategies and trading systems. John Wiley & Sons, Inc., New Jersey.

[2002] Alessio E., Carbone A., Castelli G., Frappietro V., Second-order moving average and scaling of stochastic time series. *European Physical Journal*, **27**, B, pp 197–200.

[2002] Alexander C., Dimitriu A., The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies. Working Paper, SSRN eLibrary.

[2008] Alexander C., Market risk analysis: Practical financial econometrics. John Wiley & Sons, Ltd., Chichester.

[2010] Alexander C., Sarabia J-M., Generalized beta-generated distributions. ICMA Centre Discussion Papers in Finance, Henley Business School at Reading, UK.

[1988] Algoet P., Cover T., Asymptotic optimality asymptotic equipartition properties of log-optimum investments. *Annals of Probability*, **16**, pp 876–898.

[1992] Algoet P., Universal schemes for prediction, gambling, and portfolio selection. *Annals of Probability*, **20**, pp 901–941.

[1994] Algoet P., The strong law of large numbers for sequential decisions under uncertainty. IEEE *Transactions on Information Theory*, **40**, pp 609–634.

[1953] Allais M., Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'ecole americaine. *Econometrica*, **21**, 503–546.

[1999] Allen F., Karjalainen R., Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, **51**, pp 245–271.

[2002] Alrumaih R.M., Al-Fawzan M.A., Time series forecasting using wavelet denoising: An application to Saudi Stock Index. *Journal of King Saud University*, Engineering Sciences, **2**, (14), pp 221–234.

[2005] Alvarez F., Jermann U.J., Using asset prices to measure the persistence of the marginal utility of wealth. Working Paper.

[2003] Amenc N., Malaise P., Martellini L., Sfeir D., Tactical style allocation: A new form of market neutral strategy. Working Paper, EDHEC Risk and Asset Management Research centre.

[2014] Analytics Engines, Value at risk calculations using Monte Carlo methods.

[1979] Antonov A., Saleev V.M., An economic method of computing lp-sequences, *USSR Computational Mathematics and Mathematical Physics*, **19** (1), pp 252–256.

[1998] Andersen T.G., Bollerslev T., Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, **39**, (4), pp 885–905.

[1976] Annis A.A., Lloyd E.H., The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika*, **63**, (1), pp 111–116.

[2001] Antoniadis A., Bigot J., Sapatinas T., Wavelet estimators in non-parametric regression: A comparative simulation study. *Journal of Statistical Software*, **6**, (6), pp 1–83.

[2004] Aoki M., Modeling aggregate behavior and fluctuations in economics: Stochastic views of interacting agents. Cambridge, Cambridge University Press.

[1999] Appel G., Technical analysis power tools for active investors. Financial Times Prentice Hall.

[2008] Appel G., Appel M., A quick tutorial in MACD: Basic concepts. Working Paper.

[1990] Archibald B.C., Parameter space of the Holt-Winters' model. *International Journal of Forecasting*, **6**, pp 199–209.

[1995] Arino M.A., Morettin P., Vidakovic B., Wavelet scalograms and their applications in economic time series. Discussion Paper No. 94-13, Institute of Statistics and Decision Sciences, Duke University.

[2006] Arisoy Y.E., Altay-Salih A. and Akdeniz L., Is volatility risk priced in the securities market? Evidence from S&P 500 index options. Working Paper, Bilkent University, Faculty of Business Administration.

[1991] Arneodo A., Bacry E., Muzy J-F., Wavelets and multifractal formalism for singular signals: Application to turbulence data. *Phys. Rev. Lett.*, **67**, pp 3515–3518.

[1995] Arneodo A., Bacry E., Graves P.V., Muzy J-F., Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.*, **74**, 3293.

[1998] Arneodo A., Muzy J-F., Dornette D., Discrete causal cascade in the stock market. *Eur. Phys. J.*, **2**, pp 277–282.

[1989] Arora J.S., Introduction to optimum design. McGraw-Hill, New York.

[1951] Arrow K.J., Alternative approaches to the theory of choice in risk-taking situations. *Econometrica*, **19**, (4), pp 404–437.

[1952] Arrow K.J., Le role des valeurs boursieres pour la repartition la meilleure des risques. International Colloqium on Econometrics.

[1953] Arrow K.J., Le role des valeurs boursieres pour la repartition la meilleure des risques. *Econometrie*, **40**, Cahier du CNRS, pp 41–47.

[1954] Arrow K.J., Debreu G., Existence of an equilibrium for a competitive economy. *Econometrica*, **22**, pp 265–290.

[1971] Arrow K.J., Essays in the theory of risk-bearing. Chicago: Markham.

[1974] Arrow K.J., The use of unbounded utility functions in expected-utility maximisation: Response. *Quarterly Journal of Econometrics*, **88**, (1), pp 136–138.

[1999] Artzner P., Delbaen F., Eber J-M., Heath D., Coherent measures of risk. *Mathematical Finance*, **9**, (3), pp 203–228.

[1999] Ausloos M., Vandewalle N., Boveroux P., Minhuet A., Ivanova K., Applications of statistical physics to economic and financial topics. *Physica A: Statistical Mechanics and its Applications*, **274**, pp 229–240.

[2000] Ausloos M., Statistical physics in foreign exchange currency and stock markets. *Physica A*, **285**, pp 48–65.

[2007] Ausloos M., Lambiotte R., Clusters of networks of economies? A macroeconomy study through gross domestic product. *Physica A: Statistical Mechanics and its applications*, **382**, pp 16–21.

[2010] Ausloos M., Econophysics in Belgium. The first 15 years. *Science and Culture*, **76**, (9-10), pp 380–385.

[1998] Aussem A., Campbell J., Murtagh F., Wavelet based feature extraction and decomposition strategies for financial forecasting. *J. Computational Intelligence Finance*, **6**, (2), pp 5–12.

[2008] Avellaneda M., Lee JH., Statistical arbitrage in the U.S. equities market. Working Paper, Courant Institute of Mathematical Sciences, New York.

[1996] Aytug H., Koehler G.J., New stopping criterion for genetic algorithms. Working Paper, University City Blvd and University of Florida.

[1900] Bachelier L., Theorie de la speculation. Thesis for the doctorate in Mathematical Sciences. *Annales Scientifique de l'Ecole Normale Superieure*, **3**, (17), pp 21–86.

[2008] Bacon C.R., Practical risk-adjusted performance measurement. The Wiley Finance Series, John Wiley & Sons.

[2000] Bacry E., Delour J., Muzy J-F., A multivariate multifractal model for return fluctuations. *Quantitative Finance Papers*.

[2001] Bacry E., Delour J., Muzy J-F., Multifractal random walk. *Physical Review E*, **64**, 026103–026106.

[2003] Bacry E., Muzy J-F., Log-infinitely divisible multifractal processes. *Communications in Mathematical Physics*, **236**, pp 449–475.

[2008] Bacry E., Kozhemyak A., Muzy J-F., Continuous cascade model for asset returns. *Journal of Economic Dynamics and Control*, **32**, pp 156–199.

[2012] Bacry E., Duvernet L., Muzy J-F., Continuous-time skewed multifractal processes as a model for financial returns. *Journal of Applied Probability*, **49**, pp 482–502.

[2013] Bacry E., Kozhemyak A., Muzy J-F., Lognormal continuous cascades: Aggregation properties and estimation. *Quantitative Finance*, **13**, pp 795–818.

[1996] Baillie R.T., Bollerslev T., Mikkelsen H.O., Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **74**, pp 3–30.

[2003] Bakshi G., Kapadia N., Delta-hedged gains and the negative market volatility risk premium. *Review of Financial Studies*, **16**, pp 527–566.

[2007] Balamurugan R., Subramanian S., Self-adaptive differential evolution based power economic dispatch of generators with valve-point effects and multiple fuel options. *International Journal of Computer Science and Engineering*, Winter.

[2012a] Baltas A.N., Kosowski R., Momentum strategies in futures markets and trend-following funds. Working Paper, Imperial College London.

[2012b] Baltas A.N., Kosowski R., Improving time-series momentum strategies: The role of trading signals and volatility estimators. Working Paper, Imperial College London.

[2012] Bhandari D., Murthy C.A., Pal S.K., Variance as a stopping criterion for genetic algorithms with elitist model. *Fundamenta Informaticae*, **120**, pp 145–164.

[1993] Bansal R., Viswanathan S., No arbitrage and arbitrage pricing. *Journal of Finance*, **48**, pp 1231–1262.

[1991] Barabasi A-L., Vicsek T., Multifractality of self-affine fractals. *Physical Review A*, **44**, (4), pp 2730–2733.

[1998] Barberis N., Shleifer A., Vishny R., A model of investor sentiment. *Journal of Financial Economics*, **49**, pp 307–343.

[2002] Barberis N., Thaler T., A survey of behavioral finance. NBER Working Paper, 9222.

[2001] Barndorff-Nielsen O.E., Prause K., Apparent scaling. *Finance and Stochastics*, **5**, pp 103–113.

[2002] Barndorff-Nielsen O.E., Shephard N., Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, **64**, (2), pp 253–280.

[2008] Barone-Adesi G., Engle R., Mancini L., A GARCH option pricing model in incomplete markets. *Review of Financial Studies*, **21**, pp 1223–1258.

[2012] Barroso P., Santa-Clara P., Managing the risk of momentum. Working Paper, SSRN eLibrary.

[2012] Barunik J., Understanding the source of multifractality in financial markets. *Physica A*, **391**, (17), pp 4234–4251.

[1987] Battle G., A block spin construction of ondelettes, Part I: Lemarie functions. *Commun. Math. Phys.*, **110**, pp 601–615.

[2013] Battula B.P., Satya Prasad R., An overview of recent machine learning strategies in data mining. *International Journal of Avanced Computer Science and Applications*, **4**, (3), pp 50–54.

[2001] Becherer D., The numeraire portfolio for unbounded semi-martingales. *Finance and Stochastics*, **5**, (3), pp 327–341.

[2000] Bekaert G., Wu G., Asymmetric volatility and risk in equity markets. *The Review of Financial Studies*, **13**, (1), pp 1–42.

[1980] Bell R.M., Cover T.M., Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, **5**, pp 161–166.

[2000] Bellamy N., Jeanblanc M., Incompleteness of markets driven by a mixed diffusion. *Finance and Stochastics*, **4**, Number 2.

[2010] Bender J., Briand R., Nielsen F., Stefek D., Portolio of risk premia: a new approach to diversification. *Journal of Portfolio Management*, **36**, (2), pp 17–25.

[1977] Benedetti J.K., On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society*, **39**, Series B, pp 248–253.

[1994] Bengio Y., Simard P., Frasconi P., Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**, (2), pp 157–166.

[2012] Bennett C., Gil M., Measuring historical volatility. Equity Derivatives, Santander.

[2003] Bentz Y., Quantitative equity investment management with time-varying factor sensitivities. In Dunis, C., Laws, J. and Naim, P, eds., Applied Quantitative Methods for Trading and Investment, John Wiley & Sons, Chichester, pp 213–237.

[1994] Beran J., Statistics for long-memory processes. Monographs on Statistics and Applied Probability, Chapman & Hall, New York.

[1996] Bergman Y.Z., Grundy B.D., Wiener Z., General properties of option prices. Preprint, Wharton School of the Univertity of Pennsylvania.

[1964] Berman S.M., Limiting theorems for the maximum term in stationary sequences. *Annals of Mathematical Statistics*, **35**, pp 502–516.

[1989] Bernard V.L., Thomas J.K., Post-earnings annoucement drift: Delayed price response or risk premium? *Journal of Accounting Research*, **27**, pp 1–36.

[1996] Bernardo A.E., Ledoit O., Gain, loss and asset pricing. Working Paper.

[2000] Bernardo A.E., Ledoit O., Gain, loss, and asset pricing. *Journal of Political Economy*, **108**, pp 144–172.

[2011] Bernemann A., Shreyer R., Spanderen K., Accelerating exotic option pricing and model calibration using GPUs. Working Paper, SSRN.

[1738-1954] Bernoulli D., Specimen theoriae novae de mensura sortis, in Commentarii Academiae Scientiarum Imperialis Petropolitannae. Exposition of a new theory on the measurement of risk. translated by Dr. Louise Sommer, *Econometrica*, **22**, (1), pp 22–36.

[1713] Bernoulli J., Ars conjectandi. Thurnisorium, Basil.

[2013] Bertrand P., Fhima M., Guillin A., Local estimation of the Hurst index of multifractional Brownian motion by increment ratio statistic method. ESAIM *Probability and Statistics*, **17**, (1), pp 307–327.

[1992] Beylkin G., Saito N., Wavelets, their autocorrelation functions, and multiresolution representation of signals. D.P. Casasent, eds., *Intelligent Robots and Computer Vision XI: Biological, Neural Net, and 3D Methods*, **39**, doi:10.1117/12.131585.

[2005] Billah B., Hyndman R.J., Koehler A.B., Empirical information criteria for time series forecasting model selection. *Journal of Statistical Computation and Simulation*.

[2006] Bishop C., Pattern recognition and machine learning. Springer Verlag.

[1972] Black F., Capital market equilibrium with restricted borrowing. *Journal of Business*, **45**, (3), pp 444–454.

[1972] Black F., Jensen M.C., Scholes M., The capital asset pricing model: Some empirical tests. *Studies in the Theory of Capital Markets*, M.C. Jensen, ed. New York: Praeger, pp 79–121.

[1973] Black F., Scholes M., The pricing of options and corporate liabilities. *Journal of Political Economics*, **81**, pp 637–659.

[1990] Black F., Litterman R., Asset allocation: Combining investor views with market equilibrium. Working Paper, Goldman Sachs.

[2008] Bloch D.A., Nakashima Y., Multi-currency local volatility model. Working Paper, SSRN-.

[2010] Bloch D.A., A practical guide to implied and local volatility Working Paper, SSRN-id1538808.

[2011] Bloch D.A., Coello Coello C.A., Smiling at evolution. *Applied Soft Computing*, **11**, (8), pp 5724–5734.

[1986] Blum L., Blum M., Schub M., A simple unpredictable pseudo-random number generator. *SIAM Journal on Computing*, **15**, (2), pp 364–383.

[2009] BNP Paribas, BNP Paribas CIB reduces supercomputer environmental impact.

[1992] Boashash B., Estimating and interpreting the instantaneous frequency of a signal part 2: Algorithms and applications. *Proceedings of the IEEE*, **80**, pp 540–568.

[1960] Bochner S., Harmonic analysis and the theory of probability. University of California Press, Berkeley.

[1986] Bollerslev T., Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, pp 307–327.

[1999] Bollerslev T., Jubinski D., Equality trading volume and volatility: Latent information arrivals and common long-run dependencies. *Journal of Business & Economic Statistics*, **17**, pp 9–21.

[2011] Bollerslev T., Todorov V., Tails, fears, and risk premia. *Journal of Finance*.

[2007] Bookstaber R., A demon of our own design: Markets, hedge funds, and the perils of financial innovation. Wiley.

[1987] Boothe P., Glassman D., The statistical distribution of exchange rates, empirical evidence and economic implications. *Journal of International Economics*, **22**, pp 297–319.

[2010] Bordalo P., Gennaioli N., Shleifer A., Salience theory of choice under risk. Harvard University working paper.

[1966] Bossons J., The effects of parameter misspecification and non-stationary on the applicability of adaptive forecasts. *Management Science*, **12**, pp. 659–669.

[2000] Bouchaud J-P., Potters M., Theory of financial risks: from statistical physics to risk management. Cambridge University Press, Cambridge.

[1970] Box G.E.P., Pierce D., Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, **65**, pp 1509–1526.

[1994] Box G.E.P., Jenkins G.M., Reinsel G.C., Time series analysis: Forecasting and control. 3rd edition, Prentice Hall: Englewood Cliffs, New Jersey.

[1997] Brace A., Gatarek D., Musiela M., The market model of interest rate dynamics. *Mathematical Finance*, **7** (2), pp 127–155.

[2003] Bradley A.P., Shift-invariance in the discrete wavelet transform. in C. Sun, H. Talbot, S. Ourselin, T. Adriaansen, eds., *Proc. VIIth Digital Image Computing: Techniques and Applications*, pp 29–38.

[2005] Brandt M.W., Kinlay J., Estimating historical volatility. Research Article, Investment Analytics.

[1988] Bratley P., Fox B.L, Algorithm 659: Implementing Sobol's quasirandom sequence generator, *ACM Transactions on Mathematical Software*, **14**, pp 88–100.

[1978] Breeden D.T., Litzenberger R.H., Prices of state-contingent claims implicit in option prices. *Journal of Business*, **51**, pp 621–651.

[1989] Breeden D.T., Gibbons M.R., Litzenberger R.H., Empirical tests of the consumption oriented CAPM. *Journal of Finance*, **44**, (2), pp 221–262.

[1998] Breidt F.J., Crato N., de Lima P., On the detection and estimation of long memory in stochastic volatility. *Journal of Econometrics*, **83**, pp 325–348.

[1961] Breiman L., Optimal gambling systems for favorable games. in *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley University, pp 65–78.

[2006] Brest J., Greiner S., Boskovic B., Mernik M., Zumer V., Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark. *IEEE Transactions on Evolutionary Computation*, **10**, pp 646–657.

[1950] Brier G., Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, (1), pp 1–3.

[2000] Britten-Jones M., Neuberger A., Option prices, implied price processes, and stochastic volatility. *Journal of Finance*, **55**, pp 839–866.

[1987] Brock W.A., Dechert W.D., Scheinkman J.A., LeBaron B., A test for independence based on the correlation dimension. Working Paper, Classifications C10, C52.

[1991] Brockwell P.J., Davis R.A., Time series: Theory and methods. Springer Series in Statistics.

[1996] Brockwell P.J., Davis R.A., Introduction to time series and Forecasting. Springer Series in Statistics, New York.

[1995] Bromley B.C., Quasirandom number generators for parallel Monte Carlo algorithms. *Journal of Parallel and Distributed Computing*, **38**, (1), pp 101–104.

[1959] Brown R.G., Statistical Forecasting for Inventory Control. McGraw-Hill, New York, NY.

[1963] Brown R.G., Smoothing, forecasting and prediction of discrete time series. Englewood Cliffs, NJ, Prentice-Hall.

[2011] Bruder B., Dao TL., Richard JC., Roncalli T., Trend filtering methods for momentum strategies. White Paper, Quant Research by LYXOR.

[1997] Brush J.S., Comparisons and combinations of long and long-short strategies. *Financial Analysts Journal*, **53**, pp 81–89.

[2011] Bryhn A.C., Dimberg P.H., An operational definition of a statistically meaningful trend. *PLoS ONE*, **6**, (4), e19241.

[1994] Burke G., A sharper Sharpe ratio. *The Computerized Trader*, March.

[2005] Bush K., Tsendjav B., Improving the richness of echo state features using next ascent local search. in *Proceedings of the Artificial Neural Networks in Engineering Conference*, pp 227–232, St. Louis.

[2006] Bush K., Anderson C., Exploiting iso-error pathways in the N, k-plane to improve echo state network performance.

[1994] Cai J., A Markov model of switching-regime ARCH. *Journal of Business*, **12**, pp 309–316.

[2001] Cai T.T., Silverman B.W., Incorporating information on neighbouring coefficients into wavelet estimation. Sankhya, Series A, 63.

[2004] Cajueiro D.O., Tabak B.M., The Hurst exponent over time: Testing the assertion that emerging markets are becoming more efficient. *Physica A*, **336**, pp 521–537.

[2007] Cajueiro D.O., Tabak B.M., Long-range dependence and multifractality in the term structure of LIBOR interest rates. *Physica A*, **373**, pp 603–614.

[1995] Caldwell R.B., Performances metrics for neural network-based trading system development. *NeuroVet Journal*, **3**, (2), pp 22–26.

[2001] Calvet L., Fisher A., Forecasting multifractal volatility. *Journal of Econometrics*, **105**, pp 27–58.

[2002] Calvet L., Fisher A., Multifractality in asset returns: Theory and evidence. *The Review of Economics and Statistics*, **84**, (3), pp 381–406.

[2004] Calvet L., Fisher A., Regime-switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, **2**, pp 44–83.

[2006] Calvet L., Fisher A., Thompson S., Volatility comovement: A multi-frequency approach. *Journal of Econometrics*, **31**, pp 179–215.

[2013] Calvet L., Fisher A., Wu L., Staying on top of the curve: A cascade model of term structure dynamics. Working Paper.

[1988] Campbell J.Y., Shiller R.J., Stock prices, earnings and expected dividends. *Journal of Finance*, **43**, (3), pp 661–676.

[1997] Campbell J.Y., Lo A.W., MacKinlay A.C., The econometrics of financial markets. Princeton University Press, New jersey.

[2004] Carbone A., Castelli G., Stanley H., Time-dependent Hurst exponents in financial time series. *Physica A*, **344**, pp 267–271.

[2007] Carbone A., Algorithm to estimate the Hurst exponent of high-dimensional fractals. Working Paper, Physics Department, Politecnico di Torino.

[1997] Carhart M.M., On persistence in mutual fund performance. *Journal of Finance*, **52**, (1), pp 57–82.

[2009] Carmona R., Cinlar E., Indifference pricing, theory and application. Princeton.

[2012] Carr P., Yu J., Risk, return, and Ross recovery. *Journal of Derivatives*, **20**, pp 38–59.

[1994] Carter C.K., Kohn R., On Gibbs sampling for state space models. *Biometrica*, **81**, pp 541–553.

[2001] Castiglione F., Bernaschi M., Market fluctuations: Simulation and forecasting. Working Paper

[2006] Cesa-Bianchi N., Lugosi G., Prediction, learning, and games. Cambridge University Press.

[1988] Chan N.H., Wei C.Z., Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics*, **16**, pp 367–401.

[2009] Chan E., Quantitative trading: How to build your own algorithmic trading business. John Wiley & Sons, Hoboken, New Jersey.

[1994] Chande T.S., Kroll S., The new technical trader. John Wiley, New York.

[1988] Chatfield C., Yar M., Holt-Winters forecasting: Some practical issues. *Journal of the Royal Statistical Society*, Series D, **37**, pp 129–140.

[1996] Chatfield C., Model uncertainty and forecast accuracy. *Journal of Forecasting*, **15**, pp 495–508.

[2002] Chatfield C., Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society*, Series D, **51**, pp 1–20.

[2008] Chauhan A., Automated stock trading and portfolio optimization using XCS trader and technical analysis. Master of Science, Artificial Intelligence, School of Informatics, University of Edinburgh.

[2007] Chen Y., Pennock D.M., A utility framework for bounded-loss market makers. in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*.

[2008] Chen Y., Fortnow L., Lambert N., Pennock D.M., Wortman J., Complexity of combinatorial market makers. in *Proceedings of the 9th ACM Conference on Electronic Commerce*.

[1989] Chhabra A., Jensen R.V., Direct determination of the $f(\alpha)$ singularity spectrum. *Phys. Rev. Lett.*, **62**, (12), pp 1327–1330.

[2005] Chianca C., Ticona A., Penna T., Fourier-detrended fluctuation analysis. *Physica A*, **357**, (447), pp –.

[2008] Chin W., Spurious long-range dependence: Evidence from malaysian equity markets. MPRA Paper No. 7914.

[2008b] Chin W.C., A sectoral efficiency analysis of Malaysian stock exchange under structural break. *American Journal of Applied Sciences*, **5**, pp 1291–1295.

[1993] Chopra V.K., Ziemba W.T., The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, **19**, pp 6–11.

[1960] Chow G., Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, pp 591–605.

[2009] Christara C., Dang D.M., Jackson K., Lakhany A., A PDE pricing framework for cross-currency interest rate derivatives with Target Redemption features. Working Paper, SSRN.

[2005] Christensen M.M., Platen E., A general benchmark model for stochastic jumps. *Stochastic Analysis and Applications*, **23** (5), pp 1017–1044.

[2011] Christensen M.M., On the history of the growth optimal portfolio. Chapter 1, *World Scientific Review*, **9** (6), pp 1–70.

[2005] Chui A.C., Titman S., Wei K.C.J., Individualism and momentum around the world. Working Paper, Hong Kong Polytechnic University, University of Texas at Austin and NBER.

[1998] Chung H., Lee B.S., Fundamental and non-fundamental components in stock prices of Pacific-Rim countries. *Pacific-Basin Journal of Finance*, **6**, 321–346.

[1973] Clark P.K., A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, **41**, (1), pp 135–155.

[1980] Clark R.M., Calibration, cross-validation and carbon 14 ii.. *Journal of the Royal Statistical Society*, **143**, Series A, pp 177–194.

[2002] Clarke R.G., de Silva H., Thorley S., Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, **58**, (5), pp 48–66.

[2004] Clarke R.G., de Silva H., Sapra S.G., Towards more information-efficient portfolios'. *Journal of Portfolio Management*, **31**, (1), pp 54–63.

[2005] Clarke R.G., de Silva H., Murdock R., A factor approach to asset allocation exposure to global market factors. *Journal of Portfolio Management*, **32**, pp 10–21.

[2006] Clarke R.G., de Silva H., Thorley S., The fundamental law of active portfolio management with full covariance matrix. *Journal of Investment Management*, **4**, (3), pp 54–72.

[2008] Clarke R.G., de Silva H., Sapra S.G., Thorley S., Long-short extensions: How much is enough? *Financial Analysts Journal*, **64**, (1), pp 16–30.

[2009] Clauset A., Cosma R.S., Newman M.E.J., Power-law distributions in empirical data. *SIAM Review*, **51**, pp 661–703.

[2006] Clegg R.G., A practical guide to measuring the Hurst parameter. *International Journal of Simulation: Systems, Science and Technology*, **7**, (2), pp 3–14.

[2000] Cochrane J.H., Beyond arbitrage: Good deal asset price bounds in incomplete markets. *Journal of Political Economy*, **108**, pp 79–119.

[2001] Cochrane J.H., Asset pricing. Princeton University Press.

[2000] Coello Coello C.A., Constraint-handling using an evolutionary multiobjective optimization technique. *Civil Engineering and Environmental Systems*, Vol. **17**, pp 319–346.

[2002] Coello Coello C.A., Mezura-Montes E., Constraint-handling in genetic algorithms through the use of dominance-based tournament selection. *Advanced Engineering Informatic*, Vol. **16**, pp 193–203.

[2003] Coello Coello C.A., Mezura-Montes E., Increasing successful offspring and diversity in differential evolution for engineering design. *Advanced Engineering Informatic*, Vol. **16**, pp 193–203.

[2000] Coeurjolly J.F., Simulation and identification of the fractional Brownian motion: A bibliographical and comparative study. *Journal of Statistical Software*, **5**, (7).

[1992] Cohen A., Daubechies I., Feauveau J-C., Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, **45**, (5), pp 485–560.

[2012] Cohen R., Signal denoising using wavelets. Project Report, Department of Electrical Engineering, Technion, Israel Institute of Technology.

[1992] Coifman R.R., Wickerhauser M.V., Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, **38**, pp 713–718.

[1992] Collopy F., Armstrong J.S., Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, **38**, pp 1394–1414.

[2005] Conejo A.J., Plazas M.A. Espinola R. Molina A.B., Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. *IEEE Transactions on Power Systems*, **20**, (2), pp 1035–1042.

[2002] Cont R., Tankov P., Calibration of jump-diffusion option-pricing models : a robust non-parametric approach. CMAP, **42**, September.

[2003] Cont R., Tankov P., Financial modelling with jump processes. Chapman & Hall, CRC Press.

[1994] Connor J.T., Martin R.D., Atlas L.E., Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, **5**, (2), pp 240–254.

[1964] Cootner P., The random character of stock market prices. Cambridge: MIT Press.

[2003] Costa R.L., Vasconcelos G.L., Long-range correlations and nonstationarity in the Brazilian stock market. *Physica A*, **329**, pp 231–248.

[2005] Couillard M., Davison M., A comment on measuring the Hurst exponent of financial time series. *Physica A*, **348**, pp 404–418.

[1997] Couture R., L'Ecuyer P., Distribution properties of multiply-with-carry random number generators. *Mathematics of Computation*, **66**, (218), pp 591–607.

[1984] Cover T.M., An algorithm for maximizing expected log investment return. IEEE *Transactions on Information Theory*, **30**, pp 369–373.

[1998] Cover T.M., Ordentlich E., Universal portfolios with short sales and margin. in Proceedings of IEEE International Symposium on Information Theory.

[1974] Cox D.R., Hinkley D.V., Theoretical statistics. Chapman and Hall, London.

[1985] Cox J.C., Ingersoll J.E. and Ross S.A., A theory of the term structure of interest rates. *Econometrica*, **53**, pp 373–384.

[1985b] Cox J.C., Rubinstein M., Options markets. Prentice-Hall, Englewood Cliffs.

[1957] Cramer H., Mathematical methods of statistics. Princeton University Press, Princeton.

[2000] Cristi R., Tummula M., Multirate, multiresolution, recursive Kalman filter. *Signal Processing*, **80**, pp 1945–1958.

[2001] Cvitanic J., Schachermayer W., Wang H., Utility maximisation in incomplete markets with random endowment. *Finance and Stochastics*, **5**, (2), pp 259–272.

[1988] Cybenko G., Continuous valued neural networks with two hidden layers are sufficient. Technical Report, Department of Computer Science, Tufts University, Medford, MA.

[1989] Cybenko G., Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2**, pp 303–314.

[2008] Czarnecki L., Grech D., Pamula G., Comparison study of global and local approches describing critical phenomena on the Polish stock exchange market. *Physica A: Statistical Mechanics and its Applications*, **387**, (29), pp 6601–6811.

[1993] Dacorogna M.M., Muller U.A., Nagler R.J., Olsen R.B., Pictet O.V., A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance*, **12**, (4), pp 413–438.

[1998] Dacorogna M.M., Muller U.A., Olsen R.B., Pictet O.V., Modelling short-term volatility with GARCH and HARCH models. in C. Dunis and B. Zhou, (eds.), *Nonlinear Modelling of High Frequency Financial Time Series*, John Wiley & Sons, pp 161–176.

[2001] Dacorogna M.M., Gencay R., Muller U.A., Olsen R.B., Pictet O.V., An introduction to high frequency finance. Academic Press, San Diego, CA.

[1973] D'Agostino R., Pearson E., Tests for departures from normality. Empirical results for the distribution of $\sqrt{b_1}$ and b_2. *Biometrika*, **60**, pp 613–622.

[1993] Dana R.A., Existence and uniqueness of equilibria when preferences are additively separable. *Econometrica*, **61**, (4), pp 953–957.

[1994] Dana R.A., Jeanblanc-Picque M., Marche financiers en temps continu: valorisation et equilibre. Recherche en Gestion, Economica, Paris.

[1998] Darbellay G.A., Predictability: An information-theoretic perspective. in *Signal Analysis and Prediction*, edt, A. Prochazka, J. Uhlir, P.W.J. Rayner and N.G. Kingsbury, pp 249–262.

[2000] Darbellay G.A., Slama M., Forecasting the short-term demand for electricity? Do neural networks stand a better chance? *International Journal of Forecasting*, **16**, pp 71–83.

[2003] Darst D.M., The art of asset allocation: Asset allocation principles and investment strategies for any market. McGraw-Hill, New York.

[1882] Darwin C.R., The variation of animals and plants under domestication. Murray, London, second edition.

[2005] Das S., Konar A., Chakraborty U.K., Two improved differential evolution schemes for faster global search. in *Proceedings of the 2005 conference on Genetic and Evolutionary Computing*, (GECCO 2005), pp 991–998.

[2005] Da Silva S., Matsushita R., Gleria I., Figueiredo A., Rathie P., International finance, Levy distributions, and the econophysics of exchange rates. *Communications in Nonlinear Science and Numerical Simulation*, **10**, (4), pp 365–393.

[1988] Daubechies I., Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, **41**, pp 909–996.

[1992] Daubechies I., Ten lectures on wavelets. *Regional Conference Series in Applied Mathematics*, CSIAM

[2004] Daubechies I., Defrise M., De Mol C., An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, **57**, (11), pp 1413–1457.

[2003] Dash G.H., Kajiji N., Forecasting hedge fund index returns by level and classification: A comparative analysis of RBF neural network topologies. Working Paper.

[1998] Davidson R., Labys W.C., Lesourd J-B., Wavelet analysis of commodity price behaviour. *Comput. Econ.*, **11**, pp 103–128.

[2013] Davidson C., Chip and win: Banks expand use of GPUs. *Risk magazine*, 25th of March.

[1999] Davis M.H.A., Option valuation and basis risk. In T.E. Djaferis and L.C. Shick, editors, *System Theory: Modelling, Analysis and Control*, Academic Publishers.

[2010] Davis M.H.A. and Yoshikawa D., An equilibrium approach to indifference pricing. Working Paper, Imperial College London.

[1988] Deaton A., Agricultural pricing policies and demand patterns in thailand. Unpublished Manuscript.

[2000] Deb K., An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering*, **186**, (2/4), pp 311-338.

[1985] De Bondt W.F.M., Thaler R., Does the stock market overreact? *Journal of Finance*, **40**, (3), pp 793–805.

[1987] De Bondt W.F.M., Thaler R., Further evidence of investor overreaction and stock market seasonality. *Journal of Finance*, **42**, pp 557–581.

[1953] Debreu G., Une economie de l'incertain. Electricte de France.

[1959] Debreu G., Theorie de la valeur. Dunod, Paris.

[1994] Delbaen F., Schachermayer W., A general version of the fundamental theorem of asset pricing. *Math. Ann.*, **300**, pp 463–520.

[1990] De Long J.B., Shleifer A., Summers L.H., Waldmann R.J., Positive feedback investment strategies and destabilizing rational speculation. *Journal of Finance*, **45**, pp 375–395.

[1992] Demange G., Rochet J.C., Methode mathematiques de la finance. Economica, Paris.

[1998] Dempster A., Logicist statistics I. Models and modelling. *Statistical Science*, **13**, pp 248–276.

[1999] Dempster M.A.H., Jones C.M., Can technical pattern trading be profitably automated? Working Paper, Judge Institutes of management Studies, Cambridge, UK.

[2001] Dempster M.A.H., Jones C.M., A real-time adaptive trading system using genetic programming. *Quantitative Finance*, **1**, Institute of Physics Publishing, pp 397–413.

[2001b] Dempster M.A.H., Payne T.W., Romahi Y., Thompson G.W.P., Computational learning techniques for intraday FX trading using popular technical indicators. IEEE Transactions on Neural Networks, **12**, (4).

[2006] Der R., Lee D., Beyond Gaussian processes: On the distributions of infinite networks. *Advances in Neural Information Processing Systems*, **18**, pp –.

[1989] Deslauriers G., Dubuc S., Symmetric iterative interpolation processes. *Constructive Approximation*, **5**, pp 49–68.

[1991] Detemple J., and Selden L., A general equilibrium analysis of option and stock market interactions. *International Economic Review*, **32**, pp 279–303.

[1999] Detemple J., Sundaresan S., Nontraded asset valuation with portfolio constraints: A binomial approach. *Review of Financial Studies*, **12**, pp. 835–872.

[2008] Devlin K., The unfinished game. New York, NY: Basic Books.

[1979] Dickey D.A., Fuller W.A., Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association*, pp 427–431.

[1989] Diebold F.X., Nerlove M., The dynamics of exchange rate volatility: A multivariate latent factor ARCH model. *Journal of Applied Econometrics*, **4**, pp 1–21.

[2003] Dieker A.B., Simulation of fractional Brownian motion. Master Thesis, Vrije Universiteit Amsterdam.

[1993] Ding Z., Granger C.W.J., Engle R.F., A long memory property of stock market returns and a new model. *Journal of the Empirical Finance*, **1**, pp 83–106.

[2009] Dixon M., Chong D., Accelerating market value-at-risk estimation on GPUs.

[2006] Do B., Faff R., Hamza K., A new approach to modeling and estimation for pairs trading. Working Papaer.

[1995] Donoho D.L., Johnstone I.M., Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, pp 1200–1224.

[1995b] Donoho D.L., Johnstone I.M., Kerkyacharian G., Picard D., Wavelet shrinkage: Asymptopia? *J. R. Statistical Society*, **57**, (2), pp 301–369.

[1953] Doob J.L., Stochastic processes. John Wiley & Sons.

[2000] Dowd K., Adjusting for risk: An improved Sharpe ratio. *International Review of Economics and Finance*, **9**, (3), pp 209–222.

[1992] Doya K., Bifurcations in the learning of recurrent neural networks. in *Proceedings of IEEE International Symposium on Circuits and Systems*, **6**, pp 2777–2780.

[1999] Drummond C., Hearne T., The lessons. A series of 30 multi-media lessons, Drummond and Hearne Publications, Chicago.

[2007] Duarte J., Jones C.S., The price of market volatility risk. Working Paper, University of Washington, University of Southern California.

[1986] Dubuc S., Interpolation through an iterative scheme. *J. Math. Anal. and Appl.*, **114**, pp 185–204.

[1993] Duffie D. Kan R., A yield factor model of interest rates. *Journal of Finance*, **1**.

[1997] Duffie D., Pan J., An overview of value at risk. *Journal of Derivatives*, Spring, pp 7–48.

[2000] Duffie D., Pan J., Singleton K., Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, **68**, (6), pp 1343–1376.

[2003] D. Duffie, D. Filipovic and W. Schachermayer, Affine processes and applications in finance. *Annals of Applied Probability* **13**, pp 984–1053.

[1998] Dunis C.L., Zhou B., Nonlinear modeling of high frequency financial time series. Wiley.

[2005] Dunis C.L., Shannon G., Emerging markets of south-east and central Asia: Do they still offer a diversification benefit? *Journal of Asset Management* **6**, (3), pp 168–190.

[2010] Dunis C.L., Giorgioni G., Laws J., Rudy J., Statistical arbitrage and high-frequency data with an application to Eurostoxx 50 equities. Working Paper, Liverpool Business School.

[1994] Dupire B., Pricing with a smile. *Risk*, **7**, pp. 18–20.

[2003] Durrleman V., A note on initial volatility surface. Working Paper.

[1989] Dutilleux P., An implementation of the algorithme a trous to compute the wavelet transform. in *Wavelets, time-frequency methods and phase space*, *Proceedings of the International Conference*, Marseille, J.M. Combes, A. Grossman, Ph. Tchamitchian, eds., Springer, Berlin, pp 298–304.

[2003] Dybvig P.H., Ross S.A. Arbitrage, state prices and portfolio theory. in G.M. Constantinedes, M. Harris, and R.M. Stultz, eds. *Handbook of the Economics if Finance*, (Elsevier).

[1908] Einstein A., Uber die von der molekularkinetischen theorie der warme geforderte bewegung von in ruhenden flussigkeiten suspendierten teilchen. *Annals of Physics*, **322**.

[2001] Einstein A., Wu H.S., Gil J., Detrended fluctuation analysis of chromatin texture for diagnosis in breast cytology. *Fractals*, **9**, (4).

[2004] Eisler Z., Kertesz J., Multifractal model of asset returns with leverage effect. *Physica A*, **343**, pp 603–622.

[2002] Eke A., Hermann P., Kocsis L., Kozak L.R., Fractal characterization of complexity in temporal physiological signals. *Physiol. Meas.*, **23**, R1–R38.

[2006] Eling M., Autocorrelation, bias and fat tails: Are Hedge Funds really attractive investments? *Derivatives Use, Trading and Regulation*, **12**, pp 1–20.

[2006] Eling M., Schuhmacher F., Does the choice of performance measure influence the evaluation of hedge funds? Working Papers on Risk Management and Insurance No. 29, September.

[1992] El Karoui N., Myneni R. Viswanathan R., Arbitrage pricing and hedging of interest rate claims with state variables : I theory. Working Paper, University of Paris.

[1995] El Karoui N., Quenez M.C., Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal of Control and Optimization*, **33**, pp. 29–66.

[1997] El Karoui N., Modèles stochastiques de taux d'intéret. Laboratoire de Probabilités, Universite de Paris 6.

[1998] El Karoui N., Jeanblanc M., Shreve S., Robustness of the BS formula. *Mathematical Finance*, **8**, pp 93-126.

[2002] Embrechts P., Maejima M., Selfsimilar processes. Princeton University Press.

[1982] Engle R.F., Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, pp 987–1008.

[1986] Engle R.F., Granger C.W.j., Rice j., Weiss A., Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, pp 310–320.

[1987] Engle R.F., Granger C.W.j., Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, **55**, 2, pp 251–276.

[1992] Evertsz C.J.G., Mandelbrot B.B., Multifractal measures. in H-O. Pleitgen, H. Jurgens and D. Saupe, eds., *Chaos and Fractals: New Frontiers of Science*, Berlin, Springer.

[1963] Fama E.F., Mandelbrot and the stable Paretian hypothesis. *Journal of Business*, **36**, pp 420–429.

[1965a] Fama E.F., The behavior of stock-market prices. *Journal of Business*, **38**, (1), pp 34–105.

[1965] Fama E.F., Portfolio analysis in a stable Paretian market. *Management Science*, **11**, pp 404–419.

[1970] Fama E.F., Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, **25**, (2), 383–417.

[1971] Fama E.F., Roll R., Parameter estimates for symmetric stable distributions. *Journal of the American Association*, **66**, pp 331–338.

[1972] Fama E.F., Miller M.H., The theory of finance. New York: Holt, Rinehart and Winston.

[1973] Fama E.F., MacBeth J.D., Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, **81**, (3) pp 607–636.

[1989] Fama E.F., French K., Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, **25**, pp 23–49.

[1992] Fama E.F., French K., The cross-section of expected stock returns. *Journal of Finance*, **47**, (2), 427–465.

[1993] Fama E.F., French K., Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, (1) pp 3–56.

[1996] Fama E.F., Multifactor portfolio efficiency and multifactor asset pricing. *Journal of Financial and Quantitative Analysis*, **31**, (4), pp 441–465.

[2004] Fama E.F., French K., The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, **18**, (3), pp 25–46.

[1997] Fang H., Lai T., Cokurtosis and capital asset pricing. *Financial Review*, **32**, pp 293–307.

[2002] Favre L., Galeano J.A., Mean-modified value-at-risk optimization with hedge funds. *Journal of Alternative Investments*, **5** (Fall), pp 21–25.

[1988] Feder J., Fractals. Plenum Press, New York.

[1951] Feller W., The asymptotic distribution of the range of sums of independent random variables. *Ann. Math. Statist.*, **22**, (3), pp 427–432.

[1971] Feller W., An introduction to probability theory and its applications. Wiley, Vol. 2, New York.

[2006] Feoktistov V., Differential evolution, in search of solutions. Springer, Optimisation and its Applications, **5**.

[2006] Fergussen K., Platen E., On the distributional characterisation of daily log-returns of a world stock index. *Applied Mathematical Finance*, **13**, pp 19–38.

[2005] Fernandez V., Time-scale decompositions of price transmissions in international markets. *Emerging markets Finance and Trade*, **41**, (4), pp 57–90.

[2008] Fernandez-Blanco P., Technical market indicators optimization using evolutionary algorithms. In proceedings of the 2008 GECCO conference companion on genetic and evolutionary computation. Atlanta, USA.

[1992] Fildes R., The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, **8**, pp 81–98.

[1998] Fildes R., Hibon M., Makridakis S., Meade N., Generalising about univariate forecasting methods: Further empirical evidence. *International Journal of Forecasting*, **14**, pp 339–358.

[2001] Filipovic D., Separable term structures and the maximal degree problems. Manuscript, ETH Zurich, Switzerland.

[1906] Fisher I., The nature of capital and income. Macmillan, London.

[1997] Fisher A., Calvet L., Mandelbrot B.B., Multifractality of Deutschemark/US dollar exchange rates. Cowles Foundation Discussion Papers 1166, Cowles Foundation for Research in Economics, Yale University.

[1992] Flandrin P., Wavelet analysis and synthesis of fractional Brownian motion. IEEE *Trans. Inform. Theory*, **38**, (2), pp 910–917.

[2004] Focardi S.M., Kolm P.N., Fabozzi F.J., New kids on the block. *The Journal of Portfolio Management*, **2004**, pp 42–54.

[1966] Fogel L.J., Artificial intelligence through simulated evolution. John Wiley, New York.

[2008] Fok W.W.T., Tam V.W.L., Computational neural network for global stock indexes prediction. *Proceedings of the World Congress on Engineering*, **2**, pp 1171–1175.

[1990] Follmer H., Schweizer M., Hedging of contingent claims under incomplete information. In M.H.A. Davis and R.J. Elliott, editors, *Applied Stochastic Analysis*, pp 389–414, Gordon and Breach, London.

[1986] Fox R., Taqqu M.S., Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics*, **14**, (2), pp 517–532.

[1995] Frachot A., Factor models of domestic and foreign interest rates with stochastic volatilities. *Mathematical Finance*, **5**, pp 167–185.

[2001] Frachot A., Théorie et pratique des instruments financiers. Ecole Polytechnique, Janvier.

[1927] Frechet M., Sur la loi de probabilite de l'ecart maximum. *Annales de la Societe Polonaise de Mathematique*, **6**, pp 93–116.

[1987] French K.R., Schwert G.W., Stambaugh R.F., Expected stock returns and volatility. *Journal of financical Economics*, **19**, pp 3–29.

[1948] Friedman M., Savage L.J., The utility analysis of choices involving risks. *Journal of Political Economy*, **56**, pp 279–304.

[1989] Friedman B.M., Laibson D.I., Economic implications of extraordinary movements in stock prices. Brookings Papers on Economic Activity 2.

[1985] Frisch U., Parisi G., Fully developed turbulence and intermittency. in M. Ghil, R. Benzi, G. Parisi, ed., *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, Amsterdam, pp 84–88.

[2000] Frittelli M., The minimal entropy martingale measure and the valuation problem in incomplete markets. *Mathematical Finance*, **10**, (1), pp 39–52.

[2010a] Fruehwirth-Schnatter S., Fruehwirth R., Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz, Eds., *Statistical Modelling and Regression Structure*, Festschrift in Honour of Ludwig Fahrmeir, Heidelberg, Physica-Verlag, pp 111–132.

[2010b] Fruehwirth-Schnatter S., Wagner H., Stochastic model specification search for Gaussian and partially non-Gaussian state space models. *Journal of Econometrics*, **154**, pp 85–100.

[2010] Fuertes A., Miffre J., Rallis G., Tactical allocation in commodity futures markets: Combining momentum and term structure signals. *Journal of Banking and Finance*, **34**, (10), pp 2530–2548.

[1976] Fuller W.A., Introduction to statistical time series. John Wiley & Sons, New York.

[1993] Funahashi K., Nakamura Y., Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, **6**, pp 801–806.

[1997] Galluccio S., Caldarelli G., Marsili M., Zhang Y.C., Scaling in currency exchange. *Physica A*, **245**, pp 423–436.

[2011] Gao J.B., Hu J., Tung W-W., Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. PLoS One, 6.

[1985] Gardner E.S. Jr, Exponential smoothing: The state of the art. *Journal of Forecasting*, **4**, pp 1–28.

[1985] Gardner E.S. Jr, McKenzie E., Forecasting trends in time series. *Management Science*, **31**, pp 1237–1246.

[1988] Gardner E.S. Jr, McKenzie E., Model identification in exponential smoothing. *Journal of the Operational Research Society*, **39**, pp 863–867.

[1989] Gardner E.S. Jr, McKenzie E., Seasonal exponential smoothing with damped trends. *Management Science*, **35**, pp 372–376.

[1999] Gardner E.S. Jr, Note: Rule-based forecasting vs. damped-trend exponential smoothing. *Management Science*, **45**, pp 1169–1176.

[2006] Gardner E.S. Jr, Exponential smoothing: The state of the art - Part II. *International Journal of Forecasting*, **22**, pp 637–677.

[2009] Gardner E.S. Jr, Damped trend exponential smoothing: A modelling viewpoint. Working Paper, C.T. Bauer College of Business, University of Houston.

[1980] Garman M.B., Klass M.J., On the estimation of security price volatilities from historical data. *Journal of Business*, **53**, (1), pp 67–78.

[1979] Gasser T., Muller H.G., Kernel estimation of regression functions. in Gasser and Rosenblatt (eds), Smoothing techniques for curve estimation, Springer Verlag, Heidelberg.

[2008] Gastineau G., The short side of 130/30 investing for the conservative portfolio manager. *Journal of Portfolio Management*, **34**, (2), pp 39–52.

[2006] Gatev E., Goetzmann W.N., Rouwenhorst K.G., Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, **19**, (3), pp 797–827.

[2006] Gatheral J., The volatility surface: A practitioner's guide, John Wiley & Sons, Hoboken, New Jersey.

[1809] Gauss C.F., Theoria motvs corporvm coelestivm in sectionibvs conicis Solem ambientivm.

[1991] Geltner D., Smoothing in appraisal-based returns. *Journal of Real Estate Finance and Economics*, **4**, 327–345.

[1995] Geman H., El Karoui N., Rochet J., Changes of numeraire, change of probability measure and option pricing. *Journal of Applied Probability*, **32**, pp 443–458.

[2001a] Genacy R., Selcuk F., Whitcher B., Scaling properties of foreign exchange volatility. *Physica A: Statistical Mechanics and its Applications*, **289**, pp 249–266.

[2001b] Genacy R., Selcuk F., Whitcher B., Differentiating intraday seasonalities through wavelet multi-scaling. *Physica A*, **289**, pp 543–556.

[2003] Genacy R., Selcuk F., Whitcher B., Systematic risk and timescales. *Quantitative Finance*, **3**, (2), pp 108–116.

[2005] Genacy R., Selcuk F., Whitcher B., Multiscale systematic risk. *Journal of International Money and Finance*, **24**, pp 55–70.

[1993] George E.I., McCulloch R.E., Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, pp 881–889.

[1997] George E.I., McCulloch R.E., Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, pp 339–373.

[2000] Gers F.A., Schmidhuber J., Cummins F., Continual prediction with LSTM. *Neural Computation*,**12**, (10), pp 2451–2471.

[2002] Gers F.A., Schraudolph N., Schmidhuber J., Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, **3**, pp 115–143.

[1983] Geweke J., Porter-Hudak S., The estimation and application of long memory time series models. *Journal of Time Series Analysis*, **4**, pp 221–238.

[2006] Giggs M.S., Maier H.R., Dandy G.C., Nixon J.B., Minimum number of generations required for convergence of genetic algorithms. in *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*, Vancouver, Canada, pp 2580–2587.

[1999] Gijbels I., Pope A., Wand M.P., Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society*, Series B, **61**, pp 39–50.

[1960] Girsanov I., On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theror. Probability Appl.*, **5**, pp 285–301.

[2004] Glasserman P., Monte Carlo methods in financial engineering, Springer-Verlag, New York, Berlin, Heidelberg.

[1987] Glassman D., Exchange rate risk and transactions costs: Evidence from bid-ask spreads. *Journal of International Money and Finance*, **6**, pp 479–490.

[1943] Gnedenko B.V., Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, **44**, pp 423–453.

[2007] Gneiting T., Raftery A., Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, (477), pp 359–378.

[2000] Goll T., Kallsen J., Optimal portfolios for logarithmic utility. *Stochastic Processes and their Application*, **89**, (1), pp 31–48.

[1999] Gonghui Z., Starck J-L., Campbell J., Murtagh F., The wavelet transform for filtering financial data streams. *J. Comput. Intell. Finance*, pp 18–35.

[1952] Good I.J., Rational decisions. *Journal of the Royal Statistical Society*, Series B, **14**, (1), pp 107–114.

[2006] Gorton G., Rouwenhorst K.G., Facts and fantasies about commodities futures. *Financial Analysts Journal*, **62**, (2), pp 47–68.

[2014] Goudarzi A., Banda P., Lakin M.R., Teuscher C., Stefanovic D., A comparative study of reservoir computing for temporal signal processing. Working Paper, University of New Mexico and Portland State University.

[1980] Granger C.W.J., Joyeux R., An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis*, **1**, pp 15–29.

[1999] Granger C.W.J., Terasvirta T., A simple nonlinear time series model with missleading linear properties. *Economics Letters*, **62**, pp 161–165.

[1983] Grassberger P., Procaccia I., Characterization of strange attractors. *Physical Review Letters*, **48**.

[2005] Graves A., Schmidhber J., Framewise phoneme classification with bidirectional LSTM and other neural netowrk architectures. *Neural Networks*, **18**, (5-6), pp 602–610.

[2005] Graves A., Supervised sequence labelling with recurrent neural netowrks. PhD thesis, Technische Universitat Munchen.

[2011] Graves A., Mohamed A-R., Hinton G., Speech recognition with deep recurrent neural networks. *International Conference on Acoustics, Speech and Signal Processing*, pp 1033–1040.

[2004] Grech D., Mazur Z., Can one make any crash prediction in finance using the local Hurst exponent idea? *Physica A: Statistical Mechanics and its Applications*, **336**, (1), pp 133–145.

[2005] Grech D., Mazur Z., Statistical properties of old and new techniques in detrended analysis of time series. *Acta Physica Polonica B*, **36**, (8), pp 2403–2413.

[1995] Green P.J., Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, (4), pp 711–732

[1977] Greene M.T., Fielitz B.D., Long-term dependence in common stock returns. *Journal of Financial Econom.*, **4** pp 339–349

[2000] Greene W.H., Econometric analysis. 4th edition, Prentice-Hall: Upper Saddle River, New Jersey.

[2003] Gregoriou G.N., Gueyie J.P., Risk-adjusted performance of funds of hedge funds using a modified Sharpe ratio. *Journal of Alternative Investments*, **6** (Winter), pp 77–83.

[2003] Griffin j.M., Ji X., Martin S., Momentum investing and business cycle risk: Evidence from pole to pole. *Journal of Finance*, **58**, pp 2515–2547.

[1992] Grimmett G.R., Stirzaker D.R., Probability and random processes. Oxford Science Publications, Second Edition.

[1989] Grinol R., The fundamental law of active management. *Journal of Portfolio Management*, **15**, (3), pp 30–37.

[1994] Grinol R.C., Alpha is volatility times IC times score. *Journal of Portfolio Management*, **20**, (4), pp 9–16.

[2000] Grinol R.C., Kahn R.N., Active portfolio management: A quantitative approach for producing superior returns and controlling risk. 2nd Edition, McGraw-Hill.

[2000b] Grinol R.C., Kahn R.N., The efficiency and gains of long-short investing. *Financial Analysts Journal*, **56**, pp 40–53.

[1998] Groetsch C.W., Lanczo's generalised derivative. *American Mathematical Monthly*, **105**, (4), pp 320–326.

[1984] Grossmann A., Morlet J., Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis*, **15**, pp 723–736.

[2010] Gu G-F., Zhou W-X., Detrending moving average algorithm for multifractals. *Physical Review E*, **82**, 82.011136, pp –.

[2012a] Gurrieri S., Monte-Carlo pricing under a hybrid local volatility model. *GTC*.

[2012b] Gurrieri S., Monte-Carlo calibration of hybrid local volatility models. Working Paper, SSRN.

[2012c] Gurrieri S., An analysis of Sobol sequence and the Brownian bridge, Risk Management Department, Mizuho International, London.

[2006] Guthrie C., Equity market-neutral strategy. AIMA Canada Strategy Paper Series, June.

[2009] Gyorfi L., Kevei P., St. Petersburg portfolio games. in R. Gavalda, G. Lugosi, T. Zeugmann, S. Zilles (eds.) *Proceedings of Algorithmic Learning Theory*, (Lecture Notes in Artificial Intelligence 5809), pp 83–96.

[2011] Gyorfi L., Ottucsak G., Urban A., Empirical log-optimal portfolio selections: A survey. *World Scientific Review Volume*, **9**, pp 79–115.

[1910] Haar A., Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, **69**, (3), pp 331–371.

[1971] Hakansson N., Capital growth and the mean variance approach to portfolio selection. *The Journal of Financial and Quantitative Analysis*, **6**, (1), pp 517–557.

[1982] Hall P., Cross-validation in density estimation. *Biometrika*, **69**, pp 383–390.

[1693] Halley E., An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslau; with an attempt to ascertain the price of annuities upon lives. *Phil. Trans.*, **17**, pp 596–610.

[1986] Halsey T.C., Jensen M.H., Kadanoff L.P., Procaccia I., Shraiman B.I., Fractal measures and their singularities: The characterisation of stange sets. *Physical Review A*, **33**, (2), pp 1141–1151.

[1960] Halton J.H., On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, **2**, pp 84–90.

[1994] Hamilton J., Time series analysis. Princeton University Press, Princeton, New Jersey.

[2000] Hammer B., On the approximatio capability of recurrent neural networks. *Neurocomputing*, **31**, (1-4), pp 107–123.

[2001] Hand D., Mannila H., Smyth P., Principles of data mining. MIT Press.

[1982] Hansen L.P., Singleton K.J., Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, **50**, pp 1269–1286.

[1983] Hansen L.P., Singleton K.J., Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy*, **91**, pp 249–265.

[1991] Hansen L.P., Jagannathan R., Implications of security market data for models of dynamic economies. *Journal of Political Economy*, **99**, pp 225–262.

[2003] Hanson R., Combinatorial information market design. *Information Systems Frontiers*, **5**, (1), pp 105–119.

[2007] Hanson R., Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, **1**, (1), pp 3–15.

[1986] Hardle W., Marron J.S., Random approximations to an error criterion of nonparametric statistics. *Journal of Multivariate Analysis*, **20**, pp 91–113.

[1990] Hardle W., Applied nonparametric regression. Cambridge University Press, Cambridge.

[1991] Hardle W., Smoothing techniques with implementation in S. Springer-Verlag, NY.

[1997] Harrald P.G., Kamstra M., Evolving artificial neural networks to combine financial forecasts. *IEEE Transactions on Evolutionary Computation*, **1**, (1), pp 40–52.

[1979] Harrison J.M., Kreps D., Martingales and arbitrage in multi-period securities markets. *Journal of Economic Theory*, **20**, pp 381–408.

[1981] Harrison J.M., Pliska S.R., Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications*, **11**, pp 215–260.

[1983] Harrison J.M., Pliska S.R., A stochastic calculus model of continuous trading: Complete markets. *Stochastic Processes and Their Applications*, **15**, pp 313–316.

[1984] Harvey A.C., A unified view of statistical forecasting procedures. *Journal of Forecasting*, **3**, pp 245–275.

[1989] Harvey A.C., Forecasting, structural time series models and the Kalman filter. Cambridge University Press, UK.

[1994] Harvey A.C., Ruiz E., Shephard N., Multivariate stochastic variance models. *Review of Economic Studies*, **61**, pp 247–264.

[2002] Harvey C.R., Travers K.E., Costa M.J., Forecasting emerging market returns using neural networks. *Emerging Markets Quaterly*, , pp 1–12.

[2008] Hazan E., Kale S., Extracting certainty from uncertainty: Regret bounded by variation in costs. *COLT*, pp –.

[1994] Hauser M.A., Kunst R.M., Reschenhofer E., Modelling exchange rates: Long-run dependence versus conditional heterscedasticity. *Appl. Financial Econom.*, **4**, pp 233–239.

[1992] Heath D., Jarrow V., Morton A., Bond pricing and the term structure of interest rates : A new methodology. *Econometrica*, **60**, pp 77–105.

[2002] Henderson V., Hobson D., Substitute hedging. *Risk*, **15**, (5), pp 71–75.

[2004] Henderson V., Hobson D., Utility indifference pricing: An overview. Volume on Indifference Pricing, (ed. R. Carmona), Princeton University Press.

[1983] Hentschel H.G.E., Procaccia I., The infinite number of generalised dimensions of fractals and stange attractors. *Physica 8D*, pp 435–444.

[2000] Hill J.R., Pruitt G., Hill L., The ultimate trading guide. John Wiley & Sons, Wiley Trading Advantage.

[2004] Ho D-S., Lee C-K., Wang C-C., Chuang M., Scaling characteristics in the Taiwan stock market. *Physica A*, **332**, pp 448–460.

[1997] Hochreiter S., Schmidhuber J., Long short-term memmory. *Neural Computation*, **9**, (8), pp 1735–1780.

[2001] Hochreiter S., Bengio Y., Frasconi P., Schmidhuber J., Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. in A field guide to dynamical recurrent neural networks. IEEE Press.

[1989] Hodges S., Neuberger A., Optimal replication of contingent claims under transactions costs. *Review of Futures Markets*, **8**, pp 222–239.

[1997] Hodges S.D., A generalisation of the Sharpe ratio and its applications to valuation bounds and risk measures. Financial Options Research Centre, Working Paper, University of Warwick.

[1962] Holland J.H., Outline for a logical theory of adaptive systems. *Journal of the Association for Computing Machinery*, **9**, pp 297-314.

[1975] Holland J.H., Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor.

[1989] Holschneider M., Kronland-Martinet R., Morlet J., Tchamitchian P., A real-time algorithm for signal analysis with the help of the wavelet transform. in *Wavelets, time-frequency methods and phase space*, *Proceedings of the International Conference*, Marseille, J.M. Combes, A. Grossman, Ph. Tchamitchian, eds., Springer, Berlin, pp 286–297.

[1957] Holt C.C., Forecasting seasonal and trends by exponentially weighted moving averages. ONR Memorandum, **52**, Pittsburgh, PA: Carnegie Institute of Technology.

[2004a] Holt C.C., Forecasting seasonal and trends by exponentially weighted moving averages. *International Journal of Forecasting*, **20**, pp 5–10.

[2004b] Holt C.C., Author's retrospective on Forecasting seasonal and trends by exponentially weighted moving averages. *International Journal of Forecasting*, **20**, pp 11–13.

[1999] Hong H., Stein J., A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance*, **54**, pp 2143–2184.

[1989] Honik K., Stinchcombe M., White H., Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, (5), pp 359–366.

[1982] Hopfield J.J., Neural networks and physical systems with emergent collective computational abilities. *Proceedings of National Academy of Sciences USA*, **79**, pp 2554–2558.

[2011] Horvath M., Urban A., Growth optimal portfolio selection with short selling and leverage. *World Scientific Review Volume*, **9**.

[1980] Hosking J.R.M., The multivariate portmanteau statistic. *Journal of the American Statistical Association*, **75**, pp 602–608.

[1981] Hosking J.R.M., Fractional differencing. *Biometrika*, **68**, pp 165–176.

[1984] Hosking J.R.M., Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research*, **20**, (12), pp 1898–1908.

[1989] Hseih D.A., Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business*, **62**, –.

[2001] Hu K., Ivanov P.C., Chen Z., Carpena P., Stanley H.E., Effect of trends on detrended fluctualtion analysis. *Phys. Rev. E*, **64**, 011 114, pp 1–19.

[1988] Huang C., Litzenberger R.H., Foundations for financial economics. North-Holland.

[1995] Huang B.N., Yang C.W., The fractal structure in multinational stock returns. *Appl. Econom. Lett.*, **2**, pp 67–71.

[1998] Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceeding of the Royal Society A*, **454**, pp 903–995.

[2005] Hubner G., The generalised Treynor ratio. *Review of Finance*, **9**, No 3, pp 415–435.

[2007] Hubner G., How do performance measures perform? EDHEC Business School, March.

[1951] Hurst H.E., Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Engineers*, **116**, pp 770–799.

[1965] Hurst H.E., Black R.P., Simaika Y.M., Long-term storage: An experimental study. London, Constable.

[2010] Hurst B., Ooi Y.H., Pedersen L.H., Understanding managed futures. AQR Working Paper

[1989] Hurvich C.M., Tsai C.L., Regression and time series model selection in small samples. *Biometrika*, **76**, pp 297–307.

[1991] Hurvich C.M., Tsai C.L., Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, **78**, 499–509.

[1657] Huygens C., De ratiociniis in ludo aleae (On reckoning at games of chance). London, UK: T. Woodward.

[2001] Hyndman R.J., It's time to move from what to why. *International Journal of Forecasting*, **17**, pp 567–570.

[2002] Hyndman R.J., Koehler A.B., Snyder R.D., Grose S., A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, **18**, pp 439–454.

[2003] Hyndman R.J., Billah B., Unmasking the theta method. *International Journal of Forecasting*, **19**, pp 287–290.

[2006] Hyndman R.J., Koehler A.B., Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22**, pp 679–688.

[2008] Hyndman R.J., Khandakar Y., Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, **27**, Issue 3, pp 679–688.

[2008b] Hyndman R.J., Koehler A., Ord J.K., Snyder R.D., Forecasting with exponential smoothing: The state space approach. Springer-Verlag, Berlin.

[2012] Ihlen E.A.F., Introduction to multifractal detrended fluctuation analysis in Matlab. *Frontiers in Physiology*, **3** (141), pp 1–18.

[2013] Ihlen E.A.F., Multifractal analyses of response time series: A comparative study. *Behav. Res.*, **45**, pp 928–945.

[2014] Ihlen E.A.F., Vereijken B., Detection of co-regulation of local structure and magnitude of stride time variability using a new local detrended fluctuation analysis. *Gait & Posture*, **39**, pp 466–471.

[2011] Ilmanen A., Expected returns: An investor's guide to harvesting market rewards. John Wiley & Sons, West Sussex, Unated Kingdom.

[2012] Ilmanen A., Kizer J., The death of diversification has been greatly exaggerated. *Journal of Portfolio Management*, **38**, pp 15–27.

[2006] In F., Kim S., The hedge ratio and the empirical relationship between the stock and futures markets: A new approach using wavelet analysis. *Journal of Business*, **79**, (2), pp 799–820.

[2007] In F., Kim S., A note on the relationship between Fama-French risk factors and innovations of ICAPM state variables. *Finance Research Letters*, **4**, pp 165–171.

[2008] In F., Kim S., Marisetty V., Faff R., Analyzing the performance of managed funds using the wavelet multiscaling method. *Review of Quantitative Finance and Accounting*, **31**, (1), pp 55–70.

[2004] Ishii K., van der Zant T., Becanovic V., Ploger P., Identification of motion with echo state network. in *Proceedings of the OCEANS 2004 MTS/IEEE Conference*, **3**, pp 1205–1210.

[1996] Jackwerth J.C., Rubinstein M., Recovering probability distributions from option prices. *Journal of Finance*, **51**, pp 1611–1631.

[1995] Jacobs B.I., Levy K.N., More on long-short strategies. Letter to the editor, *Financial Analysts Journal*, **51**, (2), pp 88–90.

[1998] Jacobs B.I., Levy K.N., Starer D., On the optimality of long-short strategies. *Financial Analysts Journal*, **54**, pp 40–51.

[1999] Jacobs B.I., Levy K.N., Starer D., Long-short portfolio management: An integrated approach. *Journal of Portfolio Management*, **22**, pp 23–32.

[2005] Jacobs B.I., Levy K.N., Markowitz H.M., Portfolio optimization with factors, scenarios, and realistic short positions. *Operations Research*, pp 586–599.

[2006] Jacobs B., Levy K., Enhanced active equity strategies. *Journal of Portfolio Management*, **32**, pp 45–55.

[2007a] Jacobs B., Levy K., 20 myths about enhanced active 120-20 portfolios. *Financial Analysts Journal*, **63**, pp 19–26.

[2007b] Jacobs B., Levy K., Enhanced active equity portfolios are trim equitized long-short portfolios. *Journal of Portfolio Management*, **33**, pp 19–27.

[2004] Jacod J., Protter P., Probability essentials. Springer Second Edition, Universitext.

[1994] Jacquier E., Polson N.G., Rossi P., Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, **12**, pp 371–417.

[2001] Jaeger H., The echo state approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, German National Research Center for Information Technology.

[2007] Jaeger H., Echo state network. *Scholarpedia*, **2**, (9), pp 2330.

[2006] Jaffard B., Lashermes B., Abry P., Wavelet leaders in multifractal analysis. in *Wavelet Analysis and Applications*, ed. T. Qian, M.I. Vai, Y. Xu, pp 219–264.

[2001] Jain B.J., Pohlheim H., Wegener J., On termination criteria of evolutionary algorithms. GECCO 2001 - Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kauffmann, San Francisco.

[2005] Janosi I., Muller R., Empirical mode decomposition and correlation properties of long daily ozone records. *Physical Review E*, **71**, 056126.

[1991] Jansen D., de Vries C., On the frequency of large stock market returns: putting booms and busts into perspective. *Review of Economics and Statistics*, **23**, pp 18–24.

[1980] Jarque C.M., Bera A.K., Efficient test for Normality, homoskedasticity, and serial independence of regression residuals. *Economics Letters*, **6**, pp 255–259.

[1987] Jarque C.M., Bera A.K., A test for Normality of observations and regression residuals. *International Statistical Review*, **55**, pp 163–172.

[1990] Jegadeesh N., Evidence of predictable behavior of security returns. *Journal of Finance*, **45**, pp 881–898.

[1993] Jegadeesh N., Titman S., Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, **48**, (1), pp 65–91.

[2001] Jegadeesh N., Titman S., Profitability of momentum strategies: An evaluation of alternative explanations. *Journal of Finance*, **56**, (2), pp 699–720.

[1955] Jenkinson A.F., The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Quaterly Journal of the Royal Meteorological Society*, **81**, pp 158–171.

[1968] Jensen M.J., The performance of mutual funds in the period 1945-1964. *Journal of Finance*, **23**, No 2, pp 389–416.

[1999] Jensen M.J., Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter. *Journal of Forecasting*, **18**, pp 17–32.

[2008] Jiang F., Berry H., Schoenauer M., Supervised and evolutionary learning of echo state networks. in *Proceedings of 10th International Conference on Parallel Problem Solving from Nature*, **5199** of LNCS, pp 215–224, Springer.

[2012] Jizba P., Korbel J., Methods and techniques for multifractal spectrum estimation in financial time series. FNSPE, Czech Technical University, Prague.

[2008] Joe S., Kuo F.Y., Constructing Sobol sequences with better two-dimensional projections, *SIAM Journal of Scientific Computing*, **30**, pp 2635–2654.

[1988] Johansen S., Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, pp 231–254.

[2007] Johnson S., Kahn R., Petrich D., Optimal gearing. *Journal of Portfolio Management*, **33**, pp 10–20.

[2004] Jones M.C., Families of distributions arising from distributions of order statistics. *Test*, **13**, pp 143.

[2009] Jones M.C., Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, **6**, (1), pp 70–81.

[2001] Jorion P., Value at Risk. 2nd edition, McGraw-Hill.

[2014] Joshi M., Kooderive: Multi-Core Graphics Cards, the Libor market model, least-squares Monte Carlo and the pricing of cancellable swaps. Working Paper, SSRN.

[2007] Jurek J.W., Yang H., Dynamic portfolio in arbitrage. Working Paper, SSRN.

[1979] Kahneman D., Tversky A., Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, (2), pp 263–291.

[2000] Kahneman D., Tversky A., Choices, values and frames. Cambridge University Ptress.

[1958] Kaiser H.F., The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, pp187–200.

[1960] Kalman R.E., A new approach to linear filtering and prediction problems. Transactions of the ASME *Journal of Basic Engineering*, **82**, (D), pp 35–45.

[2008] Kanamura T., Rachev S.T., Fabozzi F.J., The application of pairs trading to energy futures markets. Working Paper.

[1996] Kandel S., Stambaugh R., On the predictability of stock returns: An asset allocation perspective. *Journal of Finance*, **51**, pp 385–424.

[2002] Kantelhardt J., Zschiegner S., Koscielny-Bunde E., Bunde A., Havlin S., Stanley H.E., Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A*, **316**, (1-4), pp 87–114.

[2004] Kaplan P., Knowles J., Kappa: A generalised downside-risk performance measure. *Journal of Performance Measurement*, **8**, (D), pp 42–54.

[2004] Karaboga D., Okdem S., A simple and global optimization algorithm for engineering problems: Differential evolution algorithm. *Turkish Journal of Electrical Engineering & Computer Sciences*, **12**, (1), pp 53–60.

[1991] Karatzas I., Lehoczky J.P., Shreve S.E., Xu GL., Martingale and duality methods for utility maximisation in an incomplete market. *SIAM Journal of Control and Optimisation*, **29**, pp 702–730.

[1997] Karatzas I., Shreve S., Brownian motion and stochastic calculus. Springer.

[1990] Kariya T., Tsukuda Y., Maru J., Testing the random walk hypothesis for Japanese stock price in S. Taylor's model. Working Paper, University of Chicago.

[1986] Keim D., Stambaugh R., Predicting returns in the stock and bond markets. *Journal of Financial Economics*, **17**, 357–390.

[1956] Kelly J.L., A new interpretation of information rate. *Bell System Technical Journal*, **35**, pp 917–926.

[1960] Keltner C., How to make money in commodities.

[1938] Kendall M., A new measure of rank correlation. *Biometrika*, **30**, pp 81–89.

[1948] Kendall M., Rank correlation methods. Charles Griffin & Company Limited

[1976] Kendall M.G., Stuart A., The advanced theory of statistics. Vol. 3, Hafner, New York.

[1976b] Kendall M.G., Time series. second edition, Charles Griffin and Company, London.

[1994] Kennedy D., The term structure of interest rates as a Gaussian random field. *Mathematical Finance*, **4**, pp 247–258.

[2004] Kim S-J., Choi J-S., Multifractal measures for the yen-dollar exchange rate. *Journal of the Korean Physical Society*, **44**, (3), pp 643–646.

[2005] Kim S., In F., The relationship between stock returns and inflation: New evidence from wavelet analysis. *Journal of Empirical Finance*, **12**, pp 435–444.

[2006] Kim S., In F., A note on the relationship between industry returns and inflation through a multiscaling approach. *Finance Research Letters*, **3**, pp 73–78.

[2009] Kim S-J., Koh K., Boyd S., Gorinevsky D., l_1 trend filtering. *SIAM Review*, **51**, (2), pp 339–360.

[1997] Kivinen J., Warmuth M., Exponential gradient versus gradient descent for linear predictors. *Journal of Information and Computation*, **132**, (1), pp 1–63.

[1981] Knuth D.E., The art of computer programming. Volume 2: Seminumerical Algorithms. Addison-Wesley, second edition, Reading, Massachusetts.

[1997] Knuth D.E., The art of computer programming. Volume 2: Seminumerical Algorithms. Addison-Wesley, third edition, Reading, Massachusetts.

[1992] Koedijk K.G., Stork P.A., de Vries C., Differences between foreign exchange rate regimes: The view from the tails. *Journal of International Money and Finance*, **11**, pp 462–473.

[1988] Koehler A.B., Murphree E.S., A comparison of results from state space forecasting with forecasts from the Makridakis competition. *International Journal of Forecasting*, **4**, pp 45–55.

[1978] Koenker R.W., Bassett G.W., Regression quantiles. *Econometrica*, **46**, pp 33–50.

[1933] Kolmogorov A.N., Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlin, Germany: Springer. [Transl. Foundations of the theory of probability by N. Morrison, 2nd edn. New York, NY: Chelsea, 1956].

[1941] Kolmogorov A.N., Dissipation of energy in a locally isotropic turbulence. *Dokl. Akad. Nauk.*, **32**, pp 141.

[1984] Kon S.J., Models of stock returns: A comparison. *Journal of Finance*, **39**, pp 147–165.

[1925] Kondratiev N., The Major Economic Cycles. English version Nikolai Kondratieff (1984). Long Wave Cycle. Guy Daniels. E P Dutton

[1992] Koza J.R., Genetic programming: On the programming of computers by means of natural selection. MIT Press, USA.

[1976] Kraus A., Litzenberger R., Skewness preference and the valuation of risky assets. *Journal of Finance*, **31**, pp 1085–1100.

[1990] Kreps D., A course in microeconomic theory. Princeton University Press, Princeton.

[2009] Kristoufek L., Fractality of stock markets: A comparative study. Diploma Thesis, Charles University, Faculty of Social Sciences, Prague.

[2013] Kristoufek L., Vosvrda M., Measuring capital market efficiency: Global and local correlations structure. *Physica A*, **392**, (1), pp 184–193.

[1980] Kumaraswamy P., A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46**, (1-2), pp 79–88.

[2012] Kusakci A.O., Can M., Constrained optimization with evolutionary algorithms: A comprehensive review. *Southeast Europe Journal of Soft Computing*, **1**, (2), pp 16–24.

[2000] Laloux L., Cizeau P., Potters M., Bouchaud J.P., Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, **3**, (3), pp 391–397.

[2004] Lampinen J., Storn R., Differential evolution. in *New Optimization Techniques in Engineering*, G.C. Onwubolu and B. Babu, Eds., Springer-Verlag, Berlin, pp 123–166.

[2006] Landa Becerra R., Coello Coello C.A., Cultured differential evolution for constrained optimization. *Computer Methods in Applied Mechanisand Engineering*, **195**, Jully, pp 4303–4322.

[1984] Lane G., Lane's stochastics. *Technical Analysis of Stocks and Commodities*, pp 87–90.

[1990] Lang K.J., Waibel A.H., Hinton G.E., A time-delay neural network architecture for isolated word recognition. *Neural Networks*, **3**, pp 33–43.

[1987] Lapedes A., Farber R., Nonlinear signal processing using neural networks: Prediction and modeling. Technical Report, LA-UR87-2662, Los Alamos, New Mexico.

[1991] Larrain M., Testing chaos and nonlinearities in T-bill rates. *Financial Analysts Journal*, **47**, (5), pp 51–62.

[1959] Latane H.A., Criteria for choice among risky ventures. *Journal of Political Economy*, **38**, pp 145–155.

[1995] Lau W.C., Erramilli A., Wang J.L., Willinger W., Self-similar traffic generation: The random midpoint displacement algorithm and its properties. *Proceedings IEEE International Conference on Communications ICC*, **1**, pp 466–472.

[1990] LeBaron B., Some relations between volatility and serial correlations in stock market returns. Working Paper, February.

[1992] LeBaron B., Some relations between volatility and serial correlations in stock market returns. *Journal of Business*, **65**, pp 199–219.

[2014] Lebovits J., Levy Vehel J., White noise-based stochastic calculus with respect to multifractional Brownian motion. *Stochastics*, **86**, (1), pp 87–124.

[1989] LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D., Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**, (4), pp 541–551.

[1991] L'Ecuyer P., Tables of maximality-equidistributed combined LFSR generators. Working Paper, DIRO, Universite de Montreal.

[1996] L'Ecuyer P., Combined multiple recursive random number generators. *Operations Research*, **44**, (5), pp 816–822.

[1996b] L'Ecuyer P., Maximally equidistributed combined Tausworthe generators. *Mathematics of Computation*, **65**, (213), pp 203–213.

[1998] L'Ecuyer P., Good parameters and implementations for combined multiple recursive random number generators. Working Paper, DIRO, Universite de Montreal, Canada, May 4.

[1999b] L'Ecuyer P., Tables of maximally-equidistributed combined LFSR generators. *Mathematics of Computation*, **68**, (225), pp 261–269.

[2000] L'Ecuyer P., Lemieux C., Variance reduction via lattice rules. *Management Science*, **46**, pp 1214–1235.

[2000b] L'Ecuyer P., Touzin R., Fast combined multiple recursive generators with multipliers of the form $a = \pm 2^q \pm 2^r$. in J.A. Joines, R.R. Barton, K. Kang, P.A. Fishwick, eds, *Proceedings of the 2000 Winter Simulation Conference*, pp 683–689.

[2006] L'Ecuyer P., Uniform random number generation. in S.G. Henderson and B.L. Nelson, edts, *Simulation, Handbook in Operations Research and Management Science*, **3**, pp 55–81, Elsevier, Amsterdam, The Netherlands.

[2007] L'Ecuyer P., Pseudorandom number generators. Working Paper, DIRO, Universite de Montreal, August 28.

[2007b] L'Ecuyer P., Simard R., TestU01: A C library for empirical testing of random number generators. *ACM Transaction on Mathematical Software*, **33**, (4), Article 22.

[1998] Lee B.S., Permanent, temporary and nonfundamental components of stock prices. *Journal of Finance and Quantitative Analysis*, **33**, pp 1–32.

[1990] Lehmann B., Fads, martingales and market efficiency. *Quaterly Journal of Economics*, **105**, pp 1–28.

[1976] LeRoy S.F., Efficient capital markets: Comment. *The Journal of Finance*, **31**, (1), pp 139–141.

[1981] LeRoy S.F., Porter R.D., The present value relation: Tests based on implied variance bounds. *Econometrica*, **49**, pp 557–574.

[1967] Levy R., Relative strength as a criterion for investment selection. *Journal of Finance*, **22**, pp 595–610.

[2002] Lewellen J., Momentum and autocorrelation in stock returns. *The Review of Financial Studies*, **15**, pp 533–563.

[1999] Li J., Tsang E.P.K., Improving technical analysis predictions: An application of genetic programming. In proceedings of Florida Artificial Intelligence Research Symposium.

[1981] Li W.K., McLeod A.I., Distribution of the residual autocorrelations in multivariate ARMA time series model. *Journal of the Royal Statistical Society*, Series B, **43**, pp 231–239.

[2013] Liao L., Wang C., Liu X., Discrete wavelet transform decomposition level determination exploiting sparseness measurement. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, **7**, (9), pp 691–694.

[1986] Lidl W.K., Niederreiter H., Introduction to finite fields and their applications. Cambridge University Press, Cambridge.

[1967] Lilliefors H.W., On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**, (318), pp 399–402.

[2009] Lin BH., Chen YJ., Negative market volatility risk premium: Evidence from the LIFFE equity index options. *Asia-Pacific Journal of Financial Studies*, **38**, (5), pp 773–800.

[1965] Lintner J., The valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets. *Review of Economics and Statistics*, **47**, No. 1, pp 13–37.

[1992] Littlestone N., Warmuth M.K., The weighted majority algorithm. University of California, Santa Cruz.

[1994] Littlestone N., Warmuth M., The weighted majority algorithm. *Info. and Computation*, **108**, (2), pp 212–261.

[2002] Liu J., Lampinen J., On setting the control parameter of the differential evolution method. in *Proc. 8th Int. Conf. Soft Computing*, pp 11–18.

[2005] Liu J., Lampinen J., A fuzzy adaptive differential evolution algorithm. *Soft Computing*, **3**, (6), pp 448–462.

[2007] Liu R., di Matteo T., Lux T., True and apparent scaling: The proximities of the Markov-switching multifractal model to long-range dependence. *Physica A*, **383**, pp 35–42.

[2008] Liu R., Multivariate multifractal models: Estimation of parameters and application to risk management. University of Kiel, PhD thesis.

[2009] Liu B., Fernandez F.V., De Jonghe D., Gielena G., Less expensive and high quality stopping criteria for MC-based analog IC yield optimization. Working Paper, ESAT-MICAS, Katholieke Universiteit Leuven and IMSE, CSIC and University of Sevilla.

[1978] Ljung G.M., Box G.E.P., On a measure of lack of fit in time series models. *Biometrika*, **65**, pp 297–303.

[1990] Lo A.W., MacKinlay A.C., When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, **3**, No 2, pp 175–205.

[1990b] Lo A.W., MacKinlay A.C., Data snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, **3**, No 2, pp 431–468.

[1991] Lo A.W., Long-term memory in stock market prices. *Econometrica*, **59**, (5), pp 1279–1313.

[2008] Lo A.W., Hedge funds, systemic risk, and the financial crisis of 2007-2008. Written Testimony of A.W. Lo, Prepared for the US House of Representative.

[2008] Lo A.W., Patel P.N., 130/30: The new long-only. *The Journal of Portfolio Management*, pp 12–38.

[1998] Lobato I., Savin N., Real and spurious long-memory properties of stock market data. *Journal of Business and Economics Statistics*, **16**, pp 261–283.

[1965] Loftsgaarden D.O., Quesenberry G.P., A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, **36**, pp 1049–1051.

[1990] Long J.B., The numeraire portfolio. *Journal of Financial Economics*, **26**, pp 29–69.

[1995a] Longerstaey J., Zangari P., Five questions about RiskMetrics. Morgan Guaranty Trust Company, Market Risk Research, JPMorgan.

[1995b] Longerstaey J., More L., Introduction to RiskMetrics. 4th edition, Morgan Guaranty Trust Company, New York.

[1996] Longerstaey J., Spencer M., RiskMetrics: Technical document. Fourth Edition, Morgan Guaranty Trust Company, New York.

[2008] Los C., Measuring the degree of financial market efficiency. *Finance India*, **22**, (4), pp 1281–1308.

[2006] Lu Dang Khoa N., Sakakibara K., Nishikawa I., Stock price forecasting using back propagation neural networks with time and profit based adjusted weight factors. *SICE-ICASE International Joint Conference*, Oct., Bexco, Busa, Korea, pp 5484–5488.

[1978] Lucas R., Asset prices in an exchange economy. *Econometrica*, **46**, pp 1429–1445.

[1990] Luede E., Optmization of circuits with a large number of parameters. *Archiv f. Elektr. u. Uebertr.*, Band (44), Heft (2), pp 131-138.

[2006] Lukosevicius M., Popovici D., Jaeger H., Siewert U., Time warping invariant echo state networks Technical Report No. 2, Jacobs University Bremen.

[2007] Lukosevicius M., Echo state networks with trained feedbacks. Technical Report No. 4, Jacobs University Bremen.

[2009] Lukosevicius M., Jaeger H., Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**, (3), pp 127–149.

[2012] Lukosevicius M., Jaeger H., Schrauwen B., Reservoir computing trends. *Kunstliche Intelligenz*, Springer, **26**, (4), pp 365–371.

[1996] Lux T., The stable Paretian hypothesis and the frequency of large returns: An examination of major German stocks. *Applied Economics Letters*, **6**, pp 463–475.

[2003] Lux T., The multi-fractal model of asset returns: Its estimation via GMM and its use for volatility forecasting. Economics Working Paper No. 2003-13, Christian-Albrechts-Universitat Kiel.

[2004] Lux T., Detecting multi-fractal properties in asset returns: The failure of the scaling estimator. *International Journal of Modern Physics*, **15**, pp 481–491.

[2013] Lux T., Morales-Arias L., Relative forecasting performance of volatility models: Monte Carlo evidence. *Quantitative Finance*, **13**, pp –.

[2012] Lye C-T., Hooy C-W., Multifractality and efficiency: Evidence from Malaysian sectoral indices. *Int. Journal of Economics and Management*, **6**, (2), pp 278–294.

[1995] Lyons T., Uncertainty volatility and risk-free synthesis of derivatives. *Appl. Math. Fin.*, **2**, pp 117–133.

[2002] Maass W., Natschlager T., Markram H., Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, **14**, (11), pp 2531–2560.

[1977] Mackey M.C., Glass L., Oscillation and chaos in physiological control systems. *Science*, **197**, pp 287–289.

[2004] Magdon-Ismail M., Atiya A., Maximum drawdown. *Risk Magazine*, October.

[2007] Maginn J.L., Managing investment portfolios: A dynamic process. 3rd ed, ed. C.I.I. Series, Wiley.

[1993] Maheswaran S., Sims C., Empirical implications of arbitrage-free asset markets. in P.C.B. Phillips, ed., *Models, Methods and Applications of Econometrics*, Cambridge, Basil Blackwell, pp 301–316.

[1987] Majani B.E., Decomposition methods for medium-term planning and budgeting. in S. Makridakis, S. Wheelwright, edt., *The handbook of forecasting: A manager's guide*, Whiley, New York, pp 219–237.

[2011] Maknickiene N., Vytautas Rutkauskas A., Maknickas A., Investigation of financial market prediction by recurrent neural network. *Innovative Infotechmologies for Science, Bussiness and Education*, **2**, (11), pp 3–8.

[1982] Makridakis S., Andersen A., Carbon R., Fildes R., Hibon M., Lewandowski R., Newton J., Parzen R., Winkler R., The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, **1**, pp 111–153.

[1989] Makridakis S., Wheelwright S.C., Forecasting methods for management. John Wiley & Sons.

[1991] Makridakis S., Hibon M., Exponential smoothing: The effect of initial values and loss functions on post-sample forecasting accuracy. *International Journal of Forecasting*, **7**, pp 317–330.

[2000] Makridakis S., Hibon M., The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, **16**, pp 451–476.

[1973] Malkiel B., A random walk down Wall Street. W.W. Norton & Company.

[1989] Mallat S.G., A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, (7), pp 674–693.

[1990] Mallat S.G., Hwang W.L., Singularity detection and processing with wavelets. Technical Report No. 549, Computer Science Department, New York University.

[1992] Mallat S.G., Hwang W.L., Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory*, **38**, pp 617–643.

[1992b] Mallat S.G., Zhong S., Complete signal representation with multiscale edges. *IEEE Trans.*, PAMI **14**, pp 710–732.

[1960] Mandelbrot B.B., The Pareto-Levy law and the distribution of income. *International Economic Revue*, **1**.

[1963a] Mandelbrot B.B., New methods in statistical economics. *The Journal of Political Economy*, **71**.

[1963] Mandelbrot B.B., The variation of certain speculative prices. *Journal of Business*, **36**, (4), pp 394–419.

[1964] Mandelbrot B.B., The variation of certain speculative prices. in P. Cootner, ed., The random character of stock prices. Cambridge: MIT Press.

[1967] Mandelbrot B.B., Taylor H.M., On the distribution of stock price differences. *Operations Research*, **15**, pp 1057–1062.

[1967] Mandelbrot B.B., Forecasts of future prices, unbiased markets and martingale models. *Journal of Business*, **39**, pp 242–255.

[1968] Mandelbrot B.B., van Ness J., Fractional Brownian motions, fractional noises and applications. *SIAM Review*, **10**, (4), pp 422–437.

[1969a] Mandelbrot B.B., Wallis J., Computer experiments with fractional Gaussian noises: Part 1, averages and variances. Water Resources Research 5.

[1969b] Mandelbrot B.B., Wallis J., Computer experiments with fractional Gaussian noises: Part 2, rescaled ranges and spectra. Water Resources Research 5.

[1974] Mandelbrot B.B., Intermittent turbulence in self similar cascades: Divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics*, **62**, (2), pp 331–358.

[1975] Mandelbrot B.B., Les objets fractals: forme, hasard et dimension. Paris, Flammarion.

[1975b] Mandelbrot B.B., Stochastic models for the earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands. *Pr. of the National Academy of Sciences USA*, **72**, pp 3825–3828.

[1982] Mandelbrot B.B., The fractal geometry of nature. W.H. Freeman and Company, New York.

[1989] Mandelbrot B.B., Multifractal measures, especially for the geophysicist. *Pure and Applied Geophysics*, **131**, pp 5–42.

[1997] Mandelbrot B.B., Fisher A., Calvet L., A multifractal model of asset returns. Cowles Foundation Discussion Paper No. 1164.

[2004] Mandelbrot B.B., The (mis)behavior of markets, a fractal view of risk, ruin and reward. Basic Books.

[2005] Manimaran P., Panigrahi P.K., Parikh J.C., Wavelet analysis and scaling properties of time series. *Phys. Rev. E*, **72**, 046120, pp 1–5.

[2009] Manimaran P., Panigrahi P.K., Parikh J.C., Multiresolution analysis of fluctuations in non-stationary time series through descrete wavelets. *Physica A*, **388**, pp 2306–2314.

[1945] Mann H.B., Nonparametric tests against trend. *Econometrica*, **13**, (3), pp 245–259.

[1995] Mantegna R.N., Stanley H.E., Scaling behaviour in thedynamics of an economic index. *Nature*, **376**, pp 46–49.

[2000] Mantegna R.N., Stanley H.E., An introduction to econophysics: Correlation and complexity in finance. Cambridge University Press, Cambridge.

[1952] Markowitz H.M., Portfolio selection. *Journal of Finance*, **7**, (1), pp 77–91.

[1952b] Markowitz H.M., The utility of wealth. *Journal of Political Economy*, **60**, pp 151–158.

[1959] Markowitz H.M., Portfolio selection, efficient diversification of investments. John Wiley and Sons, New York.

[1976] Markowitz H.M., Investment for the long run: New evidence for an old rule. *Journal of Finance*, **31**, (5), pp 1273–1286.

[1991] Marsaglia G., Zaman A., A new class of random number generators. *The Annals of Applied Probability*, **1**, pp 462–480.

[2003] Marsaglia G., Xorshift RNGs. *Journal of Statistical Software*, **8**, (14), pp 1–6.

[2004] Marsaglia G., Evaluating the normal distribution. *Journal of Statistical Software*, **11**, (4).

[1938] Marschak J., Money and the theory of assets. *Econometrica*, **6**, pp 311–325.

[2005] Martielli J.D., Quantifying the benefits of relaxing the long-only constraint. *SEI Investments Developments Inc.*, pp 1–18.

[1989] Martin P., McCann B., The investor's guide to fidelity funds: Winning strategies for mutual fund investors.

[2003] Matia K., Ashkenazy Y., Stanley H.E., Multifractal properties of price fluctuations of stocks and commodities. *Europhysics Letters*, **61**, (3), pp 422–428.

[2004] Matos J.A.O., Gama S.M.A., Ruskin H., Duarte J., An econophysics approach to the Portuguese stock index-psi-20. *Physica A*, **342**, (3-4), pp 665–676.

[2008] Matos J.A.O., Gama S.M.A., Ruskin H.J., Al Sharkasi A., Crane M., Time and scale Hurst exponent analysis for financial markets. *Physica A*, **387**, pp 3910–3915.

[1994] Matsumoto M., Kurita Y., Twisted GFSR generators II. *ACM Transactions on Modeling and Computer Simulation*, **4**, (3), pp 254–266.

[1998] Matsumoto M., Nishimura T., Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8**, (1), pp 3–30.

[1996] McCoy E.J., Walden A.T., Wavelet analysis and synthesis of stationary long-memory processes. *Journal of Computational and Graphical Statistics*, **5**, (1), pp 26–56.

[1943] McCulloch W., Pitts W., A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, pp 115–133.

[1984] McDonald J.B., Some generalized functions for the size distribution of income. *Econometrica*, **52**, pp 647–663.

[2008] McKay B., Evolutionary Algorithms. Encyclopedia of Ecology, Seoul National University, Seoul, Republic of Korea, pp 1464–1472.

[1959] McKenzie L., On the existence of general equilibrium for a competitive market. *Econometrica*, **27**, pp 54–71.

[1997] Mehrabi A.R., Rossamdana H., Sahimi M., Characterization of long-range correlations in complex distributions and profiles. *Phys. Rev.*, **56**, (1), pp 712–722.

[1990] Melino A., Turnbull S.M., Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, **45**, pp 239–265.

[1934] Menger K., Das unsicherheitsmoment in der wertlehre. *Journal of Economics*, **5**, pp 459–485.

[1969] Merton R.C., Lifetime portfolio selection under uncertainty: The continuous-time case. *Rev. Econom. Statist.*, **51**, pp 247–257.

[1971] Merton R.C., Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, **3**, pp 373–413.

[1973] Merton R.C., An intertemporal capital asset pricing model. *Econometrica*, **41** (5), pp 867–887.

[1981] Merton R.C., On market timing and investment oerformance: An equilibrium theory of value for market forecasts. *Journal of Business*, **54** (3), pp 363–406.

[2005] Meucci A., Risk and asset allocation. Springer Finance.

[2004a] Mezura-Montes E., Coello Coello C.A., Tun-Morales E.I., Simple feassibility rules and differential evolution for constrained optimization. *Third Mexican International Conference on Artificial Intelligence, MICAI, Lecture Notes in Artificial Intelligence*, pp 707–716.

[2004] Mezura-Montes E., Coello Coello C.A., A study of mechanisms to handle constraints in evolutionary algorithms. Workshop at the Genetic and Evolutionary Computation Conference, Seattle, Washington, ISGEC.

[2006] Mezura-Montes E., Velazquez-Reyes J., Coello Coello C.A., Modified differential evolution for constrained optimization. *IEEE Congress on Evolutionary Computation*, IEEE Press, pp 332–339.

[2006b] Mezura-Montes E., Coello Coello C.A., Velazquez-Reyes J., Munoz-Davila L., Multiple offspring in differential evolution for engineering design. *Engineering Optimization*, **00**, pp 1–33.

[1995] Michalewicz Z., Gnetic algorithms, numerical optimization and constraints. L. Eshelman, ed., *Proceeding of the Sixth International Conference on Genetic Algorithms*, San Mateo, pp 151–158.

[1993] Michaud R., Are long-short equity strategies superior? *Financial Analysts Journal*, **49**, pp 44–50.

[2007] Miffre J., Rallis G., Momentum strategies in commodity futures markets. *Journal of Banking and Finance*, **31** (6), pp 1863–1886.

[2003] Millen S., Beard R., Estimation of the Hurst Exponent for the Burdekin river using the Hurst-Mandelbrot rescaled range statistics. First Queensland Statistics Conference, Toowoomba, Australia.

[1991] Miller M.H., Financial innovations and market volatility. Cambridge, MA, Blackwell Publishing.

[2007] Miskiewicz J., Ausloos M., Delayed information flow effect in economy systems. An ACP model study. *Physica A: Statistical Mechanics and its Applications*, **382** (1), pp 179–186.

[1927] Mitchell W.C., Business cycles: The problem and its setting. New York, National Bureau of Economic Research.

[1997] Mitchell T.M., Machine learning. McGraw-Hill.

[1997] Modigliani F., Modigliani F., Risk-adjusted performance: How to measure it and why. *Journal of Portfolio Management*, **23** (2), pp 45–54.

[2003] Moler C., Van Loan C., Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, **45**, No. 1, pp 3-000.

[2001] Moody J., Saffell M., Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, **12** (4).

[1982] Mori T.F., Asymptotic properties of empirical strategy in favourable stochastic games. in Proc. Colloquia Mathematica Societatis Janos Bolyai 36 *Limit Theorems in Probability and Statistics*, pp 777–790.

[1984] Mori T.F., I-divergence geometry of distributions and stochastic games. in Proc. of the 3rd Pannonian Symp. on *Math. Stat*, (Reidel, Dordrecht), pp 231–238.

[1986] Mori T.F., Is the empirical strategy optimal? *Statistics and Decision*, **4**, pp 45–60.

[1984] Morozov V.A., Methods for solving incorrectly posed problems. Springer-Verlag, New York.

[2012] Moskowitz T., Ooi Y.H., Pedersen L.H., Time series momentum. *Journal of Financial Economics*, **104**, (2), pp 228–250.

[2006] Moyano L.G., de Souza J., Duarte Queiros S.M., Multi-fractal structure of traded volume in financial markets. *Physica A*, **371**, pp 118–121.

[1990] Muller U.A., Dacorogna M.M., Olsen R.B., Pictet O.V., Schwarz M., Morgenegg C., Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking and Finance*, **14**, pp 1189–1208.

[1997] Muller U.A., Dacorogna M.M., Dave R.D., Olsen R.B., Pictet O.V., von Weizsacker J.E., Volatilities of different time resolutions: Analyzing the dynamics of market components. *Journal of Empirical Finance*, **4**, pp 213–239.

[2009] Murguia J.S., Perez-Terrazas J.E., Rosu H.C., Multifractal properties of elementary cellular automata in a discret wavelet approach of MF-DFA. *EPL Journal*, **87**, pp 2803–2808.

[1999] Murphy J.J., Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Prentice Hall Press.

[2003] Murtagh F., Stark J.L., Renaud O., On neuro-wavelet modeling. Working Paper, School of Computer Science, Queen's University Belfast.

[1960] Muth J.F., Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, **55**, pp 299–306.

[1991] Muzy J.F., Bacry E., Arneodo A., Wavelets and multifractal formalism for singular signals: Application to turbulence data. *Phys. Rev. Lett.*, **67**, (25), pp 3515–3518.

[1993] Muzy J.F., Bacry E., Arneodo A., Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev. E*, **47**, (2), pp 875–884.

[1964] Nadaraya E.A., On estimating regression. *Theory of Probability and its Application*, **10**, pp 186–190.

[2006] Nagarajan R., Reliable scaling exponent estimation of long-range correlated noise in the presence of random spikes. *Physica A*, **366**, (1), pp 1–17.

[1995] Nason G.P., Silverman B.W., The stationary wavelet transform and some statistical applications. *Lecture Notes in Statistics*, **103**, pp 281–299.

[1997] Neely C., Weller P., Dittmar R., Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of Financial Quantitative Analysis*, **32**, pp 405–426.

[1990] Nelson D.B., ARCH models as diffusion approximations. *Journal of Econometrics*, **45**, pp 7–38.

[1991] Nelson D.B., Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, **59**, pp 347–370.

[1987] Newey W.K., West K.D., A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, (3), pp 703–708.

[1995] Niederreiter H., The multiple-recursive matrix method for pseudorandom number generation. *Finite Fields and their Applications*, **1**, pp 3–30.

[2013] Niere H.M., A multifractality measure of stock market efficiency in Asean region. *European Journal of Business and Management*, **5**, (22), pp 13–19.

[2011] Noman N., Bollegala D., Iba H., An adaptive differential evolution algorithm. *IEEE*, pp 2229–2236.

[2005] Norouzzadeh P., Jafari G.R., Application of multifractal measures to Tehran price index. *Physica A*, **356**, pp 609–627.

[2006] Norouzzadeh P., Rahmani B., A multifractal detrended fluctuation description of Iranian rial-US dollar exchange rate. *Physica A*, **367**, pp 328–336.

[2014a] NVIDIA, Introducing NVIDIA Tesla GPUs for computational finance. Available: $http : //www.nvidia.com/content/tesla/pdf/Finance_brochure_2014_fin2.pdf$.

[2014b] NVIDIA, CUDA Toolkit Documentation v6.5.

[2012] Oh G., A multifractal analysis of Asian foreign exchange markets. *European Physical Journal B*, pp 85–214.

[1998] Oksendal V, Stochastic differential equations. Springer Fifth Edition.

[2003] Olson D., Mossman C., Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, **19**, pp 453–465.

[2002] Onoz B., Bayazit M., The power of statistical tests for trend detection. Istanbul Technical University, Faculty of Civil Engineering.

[2004] Onwubolu G.C., Differential evolution for the flow shop scheduling problem. in *New Optimization Techniques in Engineering*, G.C. Onwubolu and B. Babu, Eds., Springer-Verlag, Berlin, pp 585–611.

[2009] Oppenheim A.V., Schafer R.W., Discrete-time signal processing. third edition, Prentice-Hall.

[1997] Ord J.K., Koehler A.B., Snyder R.D., Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, **92**, pp 1621–1629.

[1964] Osborne M.F.M, Brownian motion in the stock market. in P. Cootner, ed., The random character of stock market prices. Cambridge: MIT Press.

[1997] Oussaidene M., Chopard B., Pictet O.V., Tomassini M., Parallel genetic programming and its application to trading model induction. *Parallel Computing*, **23**.

[1997] Owen A.B., Scrambled net variance for integrals of smooth functions, *Annals of Statistics*, **25**, pp 1541–1562.

[2005] Oyama A., Shimoyama K., Fujii K., New constraint-handling method for multi-objective multi-constraint evolutionary optimization and its application to space plane design. *Evolutionary and Deterministic Methods for Design*.

[1494] Pacioli F.L.B., Summa de arithmetica, geometrica, proportioni et proportionalita. Venice, Italy: Paganino de Paganini.

[2001] Panas E., Estimating fractal dimension using stable distributions and exploring long memory through ARFIMA models in Athens stock exchange. *Applied Financial Economics*, **11**.

[2006] Panneton F., L'Ecuyer P., Matsumoto M., Improved long-period generators based on linear recurrences modulo 2. *ACM Transactions on Mathematical Software*, **32**, (1), pp 1–16.

[2000] Papagelis A., Kalles D., GA tree: Genetically evolved decision trees. ICTAI.

[1896] Pareto V., Cours d'economie politique. Droz, Geneva.

[1985] Parisi G., Frisch U., On the singularity structure of fully developed turbulence. in M. Ghil, R. Benzi, G. Parisi, eds., *Turbulence and Predictability in Geophysical Fluid Dynamics*, Proceedings of the International School of Physics, Amsterdam, pp 84–87.

[1985] Parker D.B., Learning logic report TR-47. MIT Press

[1980] Parkinson M., The extreme value method for estimating the variance of the rate of return. *Journal of Business*, **53**, (1), pp 61–65.

[2000] Pasquini M., Serva M., Clustering of volatility as a multiscale phenomenon. *European Physical Journal B*, **16**, (1), pp 195–201.

[1997] Paxson V., Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. *ACM SIGCOMM Computer Communication Review*, **27**, (5), pp 5–18.

[2010] Pedersen M.E.H., Tuning and simplifying heuristical optimization. PhD thesis, Computational Engineering and Design Group, School of Engineering Sciences, University of Southampton.

[1992] Peitgen H.O., Jurgens H., Saupe D., Chaos and fractals. Springer, New York.

[1969] Pegels C., Exponential forecasting: Some new variations. *Management Science*, **15**, pp 311–315.

[1995] Peltier R.F., Levy Vehel J., Multifractional Brownian motion: Definition and preliminary results. *INRIA*, Technical Report No. 2645.

[1994] Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A.L., Mosaic organization of DNA nucleotides. *Physical Review E*, **49**, (2), pp 1685–1689.

[2000] Percival D.B., Walden A.T., Wavelet methods for time series analysis. Cambridge University Press, Cambridge.

[2009] Pesaran M., Schleicher C., Zaffaroni P., Model averaging in risk management with an application to futures markets. *Journal of Empirical Finance*, **16** (2), pp 280–305.

[1996] Pesquet J-C., Krim H., Carfantan H., Time invariant orthonormal wavelet representations. *IEEE Trans. Signal Processing*, **44** (8), pp 1964–1970.

[1991-96] Peters E.E., Chaos and order in the capital markets. second edition, John Wiley & Sons.

[1994] Peters E.E., Fractal market analysis: Applying chaos theory to investment and economics. John Wiley & Sons.

[2011a] Peters O., The time resolution of the St Petersburg paradox. *Philosophical Transactions of the Royal Society*, **369**, pp 4913–4931.

[2011b] Peters O., Menger 1934 revisited. *Journal of Economics Literature*, **104** (2), pp 228–250.

[2011c] Peters O., Optimal leverage from non-ergodicity. *Quantitative Finance*, **11** (11), pp 1593–1602.

[2006] Pezier J., White A., The relative merits of investable hedge fund indices and of funds of hedge funds in optimal passive portfolios. ICMA Centre Discussion Papers in Finance, The University of Reading.

[1992] Pictet O., Real-time trading models for foreign exchange rates. *Neural Network World*, **6**, pp 713–744.

[2013] Piketty T., Le capital au XXIe siecle. Les livres du nouveau monde, Edition du Seuil.

[2005] Pirrong C., Momentum in futures markets. Working Paper.

[2002] Plerou V., Gopikrishnan P., Rosenow B., Amaral L.N., Guhr T., Stanley H.E., Random matrix approach to cross correlations in financial data. *Phys. Re.*, E65, 066126.

[2002] Pochart B., Bouchaud J.P., The skewed multifractal random walk with applications to option smiles. *Quantitative Finance*, **24**, pp 303–314.

[2007] Pole A., Statistical arbitrage: Algorithmic trading insights and techniques. Wiley Finance.

[1993] Pomerleau D.A., Knowledge-based training of artificial neural networks for autonomous robot driving. in J. Connell and S. Mahadevan, (eds.), *Robot Learning*, pp 19–43, Kluwe Academic Publishers.

[1964] Pratt J., Risk aversion in the small and in the large. *Econometrica*, **32**, pp 122–136.

[1992] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., Numerical recipes. 2nd ed. Cambridge, Cambridge University Press.

[2005] Price K., Storn R., Lampinen J., Differential evolution: A practical approach to global optimization. Springer, Heidelberg.

[1972] Priestley M.B., Chao M.T., Nonparametric function fitting. *Journal of the Royal Statistical Society*, **34**, Series B, pp 385–392.

[2004] Protter P., Stochastic integration and differential equations. 2nd edt., Springer Verlag.

[1994] Pulley L.B., Mean-variance approximation to expected logarithmic utility. *IEEE Transactions on Information Theory*, **40**, pp 409–418.

[2001] Qi M., Predicting US recessions with leading indicators vianeural network models. *International Journal of Forecasting*, **17**, pp 383–401.

[1993] Rabemananjara R., Zakoian J., Threshold ARCH models and asymmetries in volatility. *Journal of Applied Econometrics*, **8**, pp 31–49.

[1976] Radner R., Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica*, **40**, pp 289–303.

[2009] Rakhlin A., Lecture notes on online learning. Draft.

[2003] Ramamoorthy S., A strategy for stock trading based on multiple models and trading rules. Class Projects from CS395T, Agent Based E-Commerce at UT Austin, USA.

[2004] Ramamoorthy S., Subramanian H.K., Stone P., Kuipers B.J., Safe strategies for autonomous financial trading agents: A qualitative multiple-model approach. Department of Computer Sciences, University of Texas at Austin.

[2011] Ramirez-Chavez L., Coello Coello C., Rodriguez-Tello E., A GPU-based implementation of differential evolution for solving the gene regulatory network model inference problem. *The Fourth International Workshop On Parallel Architectures and Bioinspired Algorithms*.

[1995] Ramsey J.B., Usikov D., Zaslavsky G.M., An analysis of US stock price behaviour using wavelets. *Fractals*, **3**, (2), pp 377–389.

[1996] Ramsey J.B., Zhang Z., The application of wave from dictionaries to stock market index data. *Predictability of Complex Dynamical Systems*, ed Y.A. Kravtsov and J.B. Kadtke, Springer, pp 189–205.

[1997] Ramsey J.B., Zhang Z., The analysis of foreign exchange data using waveform dictionaries. *Journal of Empirical Finance*, **4**, pp 341–372.

[1998] Ramsey J.B., Lampart C., Decomposition of economic relationships by timescale using wavelets. *Macroeconomic Dynamics*, **2**, pp 49–71.

[1999] Ramsey J.B., The contribution of wavelets to the analysis of economic and financial data. *Phil. Trans. R. Soc.*, **357**, pp 2593–2606.

[2011] Raubenheimer H., Constraints on investment weights: What mandate authors in concentrated equity markets such as South Africa nee to know. *Investment Analysts Journal*, **74**, pp 39–51.

[2000] Ray B.K., Tsay R.S., Long-range dependence in daily stock volatilities. *Journal of Business & Economic Statistics*, **18**, pp 254–262.

[1995] Refenes A.N., Neural networks in capital markets. Wiley.

[1994] Refenes A.N., Zapranis A.D., Francis G., Stock performance modeling using neural network: A comparative study with regression models. *Neural Network*, **5**, pp 961–970.

[1997] Refenes A.N., Bentz Y., Bunn D.W., Burgess A.N., Zapranis A.D., Financial time series modeling with discounted least squares backpropagation. *Neurocomputing*, **14**, (2), pp 123–138.

[2009] Reid M.D., Williamson R.C., Surogate regret bounds for proper losses. in *ICML*.

[1997] Reiss R., Thomas M., Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. Birkhauser, Basel.

[2002] Renaud O., Starck J-L., Murtagh F., Wavelet-based forecasting of short and long memory time series. Cahiers du departement d'econometrie, No 2002.04, Faculte des sciences economiques et sociales, Universite de Geneve.

[2003] Renaud O., Starck J-L., Murtagh F., Prediction based on multiscale decomposition. *International Journal of Wavelets, Multiresolution and Information Processing*.

[2005] Renaud O., Starck J-L., Murtagh F., Wavelet-based combined signal filtering and prediction. *IEEE Transaction on Systems, Man, and Cybernetics*, **35**, (6), pp 1241–1251.

[1995] Rice J.A., Mathematical statistics and data analysis. Thomson Information, Second Edition.

[2000] Richards G.R., The fractal structure of exchange rates: measurement and forecasting. *Journal of International Financial Markets, Institutions and Money*, **10**, pp 163–180.

[1999] Riedi R.H., Crouse M.S., Ribeiro V.J., Baraniuk R.G., A multifractal wavelet model with application to network traffic. IEEE *Trans. Inform. Theory*, **45**, (3), pp 992–1018.

[1964] Roberts H.V., Stock market patterns and financial analysis: Methodological suggestions. in P. Cootner, ed., The random character of stock market prices. Cambridge: MIT Press.

[1995] Robinson P.M., Gaussian semiparametric estimation of long-range dependence. *The Annals of Statistics*, **23**, pp 1630–1661.

[1987] Robinson A.J., Fallside F., The utility driven dynamic error propagation network. Cambridge University Engineering Department

[1995] Rockafellar R.T., Convex analysis. Princeton University Press.

[1994] Rogers L.C.G., Satchell S.E., Yoon Y., Estimating the volatility of stock prices: A comparison of methods that use high and low prices. *Applied Financial Economics*, **4**, (3), pp 241–247.

[1997] Rogers L.C.G., Arbitrage with fractional Brownian motion. *Mathematical Finance*, **7**, (1), pp 95–105.

[1973] Roll R., Evidence on the growth-optimum model. *Journal of Finance*, **28**, pp 551–566.

[1980] Roll R., Ross S.A., An empirical investigation of the arbitrage pricing theory. *Journal of Finance*, **35**, (5), pp 1073–1103.

[2010] Roper M., Arbitrage free implied volatility surfaces. Working Papaer, School of Mathematics and Statistics, The University of Sydney, Australia.

[2001] Rosenberg J.V., Engle R.F., Empirical pricing kernels. Working Paper, Federal Reserve Bank of New York, Stern School of Business, New York University.

[2002] Rosenberg J.V., Engle R.F., Empirical pricing kernels. *Journal of Financial Economics*, **64**, pp 341–372.

[1956] Rosenblatt M., Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp 642–669.

[1976] Ross S.A., The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, **13**, pp 341–360.

[1978] Ross S., A simple approach to the valuation of risky streams. *Journal of Business*, **51**, pp 453–475.

[2005] Ross S., Neoclassical finance. Princeton University Press, Princeton, NJ.

[2013] Ross S., The recovery theorem. Working Paper, forthcoming *Journal of Finance*.

[2000] Rouge R., El Karoui N., Pricing via utility maximization and entropy. *Mathematical Finance*, **10**, pp 259–276.

[1952] Roy A.D., Safety first and the holding of assets. *Econometrica*, **20**, pp 431–449.

[2002] Rubinstein M., Markowitz's portfolio selection: A fifty-year retrospective. *The Journal of Finance*, **57**, (3), pp 1041–1045.

[1986] Rumelhart D., McClelland J., Parallel distributed processing. MIT Press, Cambridge, Mass.

[1986b] Rumelhart D.E., Hinton G.E., Williams R.J., Learning Internal Representations by error propagation. MIT Press, Cambridge, Mass, pp 318–362.

[1994] Rumelhart D., Widrow B., Lehr M., The basic ideas in neural networks *Communications of the ACM*, **37**, (3), pp 87–92.

[1993] Saito N., Beylkin G., Multiresolution representations using the autocorrelation functions of compactly supported wavelets. *IEEE Trans. Signal Processing*, **41**, (12), pp 3584–3590.

[2006] Saito M., Matsumoto M., SIMD-oriented fast Mersenne Twister: A 128-bit pseudorandom number generator. in A. Keller, S. Heinrich, H. Niederreiter *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pp 607–622.

[1960] Samuelson P.A., The St. Petersburg paradox as a divergent double limit. *International Economic Review*, **1**, pp 31–37.

[1965] Samuelson P.A., Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, **6**, pp 41–50.

[2009] Sang Y-F., Wang D., Wu J-C., Zhu Q.P., Wang L., Entropy-based wavelet de-noising method for time series analysis. *Journal of Entropy and Information Studies*, **11**, pp 1123–1147.

[2010] Sang Y-F., Wang D., Wu J-C., Entropy-based method of choosing the decomposition level in wavelet threshold de-noising. *Journal of Entropy and Information Studies*, **12**, pp 1499–1513.

[2005] Santana-Quintero L.V., Coello Coello C.A., An algorithm based on differential evolution for multi-objective problems. *International Journal of Computational Intelligence Research*, **1**, ISSN 0973-1873, pp 151–169.

[1988] Saupe D., Algorithms for random fractals. Springer-Verlag, New York.

[1954] Savage L.J., The foundations of statistics. Wiley, New York, second revised edition: Dover, New York, 1972.

[2007] Scherer B., Portfolio construction and risk budgeting. Third Edition, Risk Books, London

[1991] Schertzer D., Lovejoy S., Lavallee D., Schmitt F., Universal hard multifractal turbulence: Theory and observations. in R. Sagdeev, U. Frisch, F. Hussain, S. Moisdeev, N. Erokin, eds., *Nonlinear Dynamics of Structures*, World Scientific, Singapore, pp 213–235.

[1991b] Schertzer D., Lovejoy S., Scaling nonlinear variability in geodynamics: Multiple singularities, observables and universality classes. in D. Schertzer, S. Lovejoy, eds., *Nonlinear Variability and Geophysics: Scaling and Fractals*, Kluwer, Dordrecht, pp 41–82.

[2002] Schleicher C., An introduction to wavelets for economists. Bank of Canada, Working Paper 2002-3.

[1992] Schmidhuber J., A fixed size storage $O(n^3)$ time complexity learning algorithm for fully recurrent continually running network. *Neural Computation*, **4**, (2), pp 243–248.

[2007] Schmidhuber J., Wierstra D., Gagliolo M., Gomez F.J., Training recurrent networks by Evolino. *Neural Computation*, **19**, (3), pp 757–779.

[1999] Schmitt F., Schertzer D., Lovejoy S., Multifractal analysis of foreign exchange data. *Applied Stochastic Models and Data Analysis*, **15**, pp 29–53.

[2006] Schobel R., Veith J., An overreaction implementation of the coherent market hypothesis and option pricing. Tubinger Diskussionsbeitrag, No. 306.

[1946] Schoenberg I.J., Contribution to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math.*, **4**, pp 45–99, 112–141.

[2006] Scholz H., Wilkens M., The Sharpe ratio's market climate bias: Theoretical and empirical evidence from US equity mutual funds. Working Paper, Catholic University of Eichstaett-Ingolstadt.

[1969] Schonfeld P., Methoden der okonometrie. Verlag Franz Vahlen GmbH, Berlin und Frankfurt.

[2011] Schumann A.Y., Kantelhardt J.W., Multifractal moving average analysis and test of multifractal model with tuned correlations. *Physica A: Statistical Mechanics and its Applications*, **390**, pp 2637–2654.

[1927] Schumpeter J., The Explanation of the Business Cycle. *Economica*.

[1996] Schwager W.F., Technical analysis. Wiley.

[1995] Schwefel H.P., Evolution and optimum seeking. John Wiley & Sons.

[1987] Scott D.W., Terrell G.R., Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, (400), pp 1131–1146.

[2012] Segara R., Das A., Turner J., Performance of active extension strategies: Evidence from the Australian equities market. *Australasian Accounting, Business and Finance Journal*, **6**, (3), pp 3–24.

[2013] Segnon M., Lux T., Multifractal models in finance: Their origin, properties, and applications. Working Paper No. 1860, Kiel Institute for the World Economy.

[2001] Segura J.V., Vercher E., A spreadsheet modeling approach to the Holt-Winters optimal forecasting. *European Journal of Operational Research*, **131**, pp 375–388.

[2002] Shadwick W.F., Keating C., A universal performance measure. *Journal of Performance Measurement*, Spring, pp 59–84.

[1948] Shannon C.E., A mathematical theory of communication. *Bell System Technical Journal*, **27**, (3), pp 379–423.

[1949] Shannon C.E., Communication in the presence of noise. *Proc. I.R.E.*, **37**, pp 10–21.

[1964] Sharpe W.F., Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, **19**, (3), pp 425–442.

[1966] Sharpe W.F., Mutual fund performance. *The Journal of Business*, **39**, (1), pp 119–138.

[1970] Sharpe W.F., Portfolio theory and capital markets. New York: McGraw-Hill.

[1994] Sharpe W.F., The Sharpe ratio. *Journal of Portfolio Management*, **21**, pp 49–58.

[1992] Shensa M.J., Discrete wavelet transforms: Wedding the a trous and Mallat algorithms. *IEEE Transactions on Signal Processing*, **40**, pp 2464–2482.

[2003] Sherstov A., Automated stock trading in PLAT. Class Projects from CS395T, Agent Based E-Commerce at UT Austin, USA.

[1981] Shiller R.J., Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review*, **71**, pp 421–436.

[1989] Shiller R.J., Market volatility. Cambridge, MIT Press.

[2000] Shiller R.J., Irrational exuberance. Princeton, First Edition, Princeton University Press.

[2002] Shimizu Y., Thurner S., Ehrenberger K., Multifractal spectra as a measure of complexity in human posture. *Fractals*, **10**, pp 103–.

[1991] Shin Y., Ghosh J., The pi-sigma network: An efficient higher-order neural network for pattern classification and function approximation. *International Joint Conference on Neural Networks*.

[2000] Shin T., Han I., Optimal signal multi-resolution by genetic algorithms to support financial neural networks for exchange-rate forecasting. *Expert Syst. Appl.*, **18**, pp 257–269.

[1997] Shleifer A., Vishny R.W., The limits of arbitrage. *Journal of Finance*, **52**, (1), pp 35–55.

[1991] Siegelmann H.T., Sontag E.D., Turing computability with neural nets. *Applied Manthematics Letters*, **4**, (6), pp 77–80.

[1998] Simonsen I., Hansen A., Nes O.M., Determination of the Hurst exponent by use of wavelet transforms. *Phys. Rev.*, **58**, (3), pp 2779–2787.

[1995] Smith J.E., Nau R.F., Valuing risky projects: Option pricing theory and analysis. *Management Science*, **41**, (5), pp 795–816.

[2012] Smith J., The strategic case for momentum. Strategic View, Shroders.

[1980] Snedecor G.W., Cochran W.G., Statistical methods. Iowa State University Press, 7th edition, Ames, Iowa.

[1967] Sobol' I.M., On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, **7**, (4), pp 86–112.

[2000] Soltani S., Boichu D., Simard P., Canu S., The long-term memory prediction by multiscale decomposition. *Signal Processing*, **80**, (10), pp 2195–2205.

[1970] Sorenson H.W., Least-squares estimation: From Gauss to Kalman. *IEEE Spectrum*, **7**, pp 63–68.

[1998] Sorensen E.H., Mezrich J.J., Miller K.L., A new technique for tactical asset allocation. Chapter 12 in F.J. Fabozzi, erd., Active Equity Portfolio Management, Hoboken, John Wiley & Sons.

[2007] Sorensen E.H., Shi J., Hua R., Qian E., Aspects of constrained long-short equity portfolios. *The Journal of Portfolio Management*, **33**, (2), 12–20.

[1991] Sortino F.A., Van Der Meer R., Downside risk. *The Journal of Portfolio Management*, **17**, (4), 27–31.

[1999] Sortino F.A., Van Der Meer R., Plantinga A., The Dutch triangle: A framework to measure upside potential relative to downside risk. *The Journal of Portfolio Management*, **26**, pp 50–58.

[1981] Stein C.M., Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, (6), pp 1135–1151.

[1989] Sterge A.J., On the distribution of financial futures price changes. *Financial Analysts Journal*, **45**, (3), pp 75–78.

[1977] Stone C.J., Consistent nonparametric regression (with discussion). *Annals of Statistics*, **5**, pp 595–645.

[1995] Storn R., Price K., Differential evolution: A simple and efficient adaptive scheme for global optimization over continuous spaces. International Computer Science Institute, Berkeley, **TR-95-012**.

[1996] Storn R., System design by constraint adaptation and differential evolution. International Computer Science Institute, Berkeley, **TR-96-039**.

[1997] Storn R., Price K., Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, **11**, (4), pp 341–359.

[2000] Storn R., Digital filter design program. FIWIZ.

[2008] Storn R., Differential evolution research: Trends and open questions. in U.K. Chakraborty, editor, *Advances in Differential Evolution*, **1**, pp 1–31, Springer-Verlag, Berlin Heidelberg.

[2014] Stosic D., Stosic D., Stosic T., Stanley H.E., Multifractal analysis of managed and independent float exchange rates. Department of Physics, Boston University.

[1971] Strang G., Fix G., A Fourier analysis of the finite element variational method. in *Constructive Aspect of Functional Analysis*, Rome, Edizioni Cremonese, pp 796–830.

[1980] Strang G., Linear algebra and its applications. 2nd ed, Harcourt Brace Jovanovich, Chicago.

[1996] Strang G., Nguyen T., Wavelets and filter banks. Wellesley, Wellesley-Cambridge.

[1998] Struzik Z.R., Removing divergence in the negative moments of the multi-fractal partition function with the wavelet transformation. Working Paper INS-R9803, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.

[1999] Struzik Z.R., Local effective Holder exponent estimation on the wavelet transform maxima tree. in *Fractals: Theory and Applications in Engineering*, eds., M. Dekking, J. Levy Vehel, E. Lutton, C. Tricot, Springer Verlag, pp 93–112.

[2000] Struzik Z.R., Determining local singularity strengths and their spectra with the wavelet transform. *Fractals*, **8**, (2), pp 163–179.

[2002] Struzik Z.R., Siebes A., Wavelet transform based multifractal formalism in outlier detection and localisation for financial time series. *Physica A*, **309**, pp 388–402.

[2003] Struzik Z.R., Econophysics vs cardiophysics: The dual face of multifractality. Working Paper, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.

[2004] Subramanian H.K., Evolutionary algorithms in optimization of technical rules for automated stock trading. M.S. Thesis, The University of Texas at Austin.

[1998] Sutton R.S., Barto A.G., Reinforcement learning: An introduction. (Adaptive Computation and Machine Learning). The MIT Press, USA.

[1985] Sweet A.L., Computing the variance of the forecast error for the Holt-Winters seasonal models. *The Journal of Forecasting*, **4** pp 235–243.

[2007] Tabb L., Johnson J., Alternative investments 2007: The quest for alpha. Technical Report, Tabb Group.

[1981] Takens F., Detecting strange attractors in turbulence. *Springer Lecture Notes in Mathematics*, **898**, pp 366–381.

[2005] Tankov P., Calibration de modeles et couverture de produits derives. Working Paper, Universite Paris VII.

[1995] Taqqu M.S., Teverovsky V., Willinger W., Estimators for long-range dependence: An empirical study. *Fractals*, **3**, (4), pp 785–798.

[1999] Taqqu M.S., Montanari A., Teverovsky V., Estimating long-range dependence in the presence of periodicity: An empirical study. *Mathematical and Computer Modelling*, **29**, (10), pp 217–228.

[1965] Tausworthe R.C., Random numbers generated by linear recurrence modulo two. *Math. of Computation*, **19**, pp 201–209.

[1986] Taylor S.J., Modelling financial time series. New York, John Wiley & Sons.

[1994] Taylor S.J., Modeling stochastic volatility. *Mathematical Finance*, **4**, pp 183–204.

[2003] Taylor J.W., Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, **19**, pp 715–725.

[2004a] Taylor J.W., Volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting*, **20**, pp 273–286.

[2004b] Taylor J.W., Smooth transition exponential smoothing. *Journal of Forecasting*, **23**, pp 385–404.

[2004c] Taylor J.W., Forecasting with exponentially weighted quantile regression. Working Paper, Said Business School, University of Oxford, Park End St., Oxford.

[2007] Taylor G.W., Hinton G.E., Roweis S., Modeling human motion using binary latent variables. in *Advances in Neural Information Processing Systems*, **19**, pp 1345–1352, MIT Press, Cambridge.

[1867] Tchebichef P., Des valeurs moyennes. *Journal de mathematiques pures et appliquees*, **2**, (12), pp 177–184.

[1999] Teverovsky V., Taqqu M.S., Willinger W., A critical look at Lo's modified R/S statistic. *Journal of Statistical Planning and Inference*, **80**, pp 211–227.

[2004] Thadewald T., Buning H., Jarque-Bera test and its competitors for testing normality: A power comparison. School of Business & Economics Discussion Paper: Economics, No. 2004/9.

[1983] Tiao G.c., Tsay R.S., Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Annals of Statistics*, **11**, pp 856–871.

[1998] Tikhonov A.N., Leonov A.S., Yagola A.G., Nonlinear ill-posed problems. Chapman & Hall, London.

[1958] Tobin J., Liquidity preference as behavior towards risk. *The Review of Economic Studies*, **25**, pp 65–86.

[1990] Tong H., Non-linear time series: A dynamical system approach. Oxford University Press, Oxford.

[1965] Treynor J.L., How to rate management of investment funds. *Harvard Business Review*, **43**, (1), pp 63–75.

[1973] Treynor J.L., Black F., How to use security analysis to improve portfolio selection. *Journal of Business*, pp 66–85.

[1967] Trigg D.W., Leach D.H., Exponential smoothing with an adaptive response rate. *Operational Research Quarterly*, **18**, pp 53–59.

[1984] Tsay R.S., Tiao G.C., Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association*, **79**, 84–96.

[2002] Tsay R.S., Analysis of financial time series. John Wiley & Sons, Hoboken, New Jersey.

[1993] Tsitsiklis J., Bertsimas D., Simulated annealing. *Statistical Science*, **8**, (1), pp 10–15.

[2013] Turc J., Ungari S., Risk-premia strategies: a way to distance yourself from the crowd. Global Quantitative Research, Societe Generale.

[2003] Turiel A., Perez-Vicente C.J., Multifractal geometry in stock market time series. *Physica A*, **322**, pp 629–649.

[2006] Turiel A., Perez-Vicente C.J., Grazzini J., Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study. *Journal of Computational Physics*, **216**, pp 362–390.

[2008] Turiel A., Yahia H., Perez-Vicente C.J., Microcanonical multifractal formalism: a geometrical approach to multifractal systems. Part I: Singularity analysis. *Journal of Physics A: Mathematical and General*, **41**, 015501, pp –.

[1990] Turner A.L., Weigel E.J., An analysis of stock market volatility. Russel Research Commentaries, Frank Russell Company, Tacoma, WA.

[1990] Tversky A., The psychology of risk. in *Quantifying the market risk premium phenomena for investment decision making*, Charlottesville, Institute of Chartered Financial Analysts.

[1987] Ullah A., Nonparametric estimation of econometric funktionals. Unpublished Manuscript.

[2013] Ungari S., Turc J., Momentum strategies for rate: Bridging the gap between statistics and option theory. Global Quantitative Research, Societe Generale.

[1992] Unser M., Aldroubi A., Eden M., On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Trans. Information Theory*, **38**, (2), pp 864–872.

[1993] Unser M., Aldroubi A., Eden M., A family of polynomial spline wavelet transforms. *Signal Processing*, **30**, (2), pp 141–162.

[1999] Unser M., Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, pp 22–38.

[1996] Urzua C., On the correct use of omnibus tests for normality. *Economics Letters*, **53**, pp 247–251.

[1990] Vaga T., The coherent market hypothesis. *Financial Analysts Journal*, **46**, (6), pp 36–49.

[1992] Vaidyanathan P.P., Multirate systems and filter banks. Prentice Hall, New Jersey.

[2006] Vajda I., Analysis of semi-log-optimal investment strategies. in M. Huskova and M. Janzura (eds.), *Prague Stochastics* (MATFYZ-PRESS, Prague).

[1997] Vandewalle N., Ausloos M., Coherent and random sequences in financial fluctuations. *Physica A*, **246**, (3), pp 454–459.

[1998] Vandewalle N., Ausloos M., Crossing of two mobile averages: A method for measuring the robustness exponent. *Phys. Rev.*, **58**, pp 177–188.

[1998b] Vandewalle N., Ausloos M., Multi-affine analysis of typical currency exchange rates. *European Physical Journal B.*, **4**, (2), pp 257–261.

[1998c] Vandewalle N., Ausloos M., Boveroux PH., Detrended fluctuation analysis of the foreign exchange market. Working Paper.

[1977] Vasicek O., An equilibrium characterization of the term structure. *Journal of Financial Economics*, **5**, pp 177–188.

[1999] Vidakovic B., Statistical modeling by wavelets. Wiley, New York.

[2004] Vidyamurthy G., Pairs trading, quantitative methods and analysis. John Wiley & Sons, Canada.

[1996] Vokurka R.J., Flores B.E., Pearce S.L., Automatic feature identification and graphical support in rule-based forecasting: A comparison. *International Journal of Forecasting*, **12**, pp 495–512.

[1944] Von Neumann J., Morgenstern O., Theory of games and economic behavior. Princeton: Princeton University Press, second edition: Princeton UP, 1947.

[2011] Vovk V., Losing money with a high Sharpe ratio. Working Paper.

[1999] Wagman L., Stock portfolio evaluation: An application of genetic programming based technical analysis. Working Paper, Stanford University, California.

[1874-7] Walras L., Elements d'economie politique pure. Lausanne: Corbaz. Translated as: Elements of pure economics. Chicago: Irwin (1954).

[2006] Wang L., Liang Y., Shi X., Li M., Han X., An improved OIF Elman neural network and its applications to stock market. *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science **4251**, pp 21–28.

[2009] Wang Y., Liu L., Gu R., Analysis of efficiency for Shenzhen stock market based on multifractal detrended fluctuation analysis. *International Review of Financial Analysis*, **18**, pp 271–276.

[2011] Wang Y., Wei Y., Wu C., Analysis of the efficiency and multifractality of gold markets based on multifractal detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, **390**, pp 817–827.

[2011b] Wang Y., Wu C., Pan Z., Multifractal detrending moving average analysis on the US Dollar exchange rates. *Physica A*, **390**, pp 3512–3523.

[1964] Watson G.S., Smooth regression analysis. *Sankhya*, Series A, (26), pp 359–372.

[1951] Weibull W., A statistical distribution function of wide applicability. *J. Appl. Mech.-Trans.*, ASME, **18**, (3), pp 293–297.

[2008] Wendt H., Contributions of wavelet leaders and bootstrap to multifractal analysis: Images, estimation performance, dependence structure and vanishing moments. Confidence intervals and hypothesis tests. Docteur de l'Universite de Lyon, Ecole Normale Superieure de Lyon, Traitement du Signal - Physique.

[1974] Werbos P., Beyond regression: New tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University.

[1990] Werbos P., Backpropagation through time: What it does and how to do it. in *Proceedings of the IEEE*, **78**, (10), pp 1550–1560.

[2002] Weron R., Estimating long-range dependence: Finite sample properties and confidence intervals. *Physica A*, **312**, pp 285–299.

[2000] White H., A reality check for data snooping. *Econometrica*, **68**, pp 1097–1127.

[1997a] Whitelaw R.F., Stock market risk and return: An equilibrium approach. Working Paper, NYU, Stern School of Business.

[1997] Whitelaw R.F., Time-varying Sharpe ratios and market timing. Working Paper, NYU, Stern School of Business.

[1978] Wilder W., New concepts in technical trading systems. Trend Research, Greensboro, NC.

[1936] Williams J.B., Speculation and the carryover. *The Quaterly Journal of Economics*, **50**, (3), pp 436–455.

[1938] Williams J.B., The theory of investment value. North Holland Publishing, Amsterdam, reprinted: Fraser Publishing, 1997.

[1989] Williams R.J., Zipser D., A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, **1**, pp 270–280.

[1990] Williams R.J., Peng J., An efficient gradient-based algorithm for online training of recurrent network trajectories. *Neural Computation*, **2**, (4), pp 490–501.

[1992] Williams R.J., Zipser D., Gradient-based learning algorithms for recurrent networks and their computational complexity. in Back-propagation: Theory, architectures and applications. NJ:Erlbaum, edt. Y.Chauvin and D.E.Rumelhart, chapter 13, pp 433–486.

[1999] Willinger W., Taqqu M.S., Teverovsky V., Stock market prices and long-range dependence. *Finance and Stochastic*, **3**, pp 1–13.

[2005] Willmott C.J., Matsuura K., Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, **30**, pp 79–82.

[2012] Wilson C., Concave functions of a single variable. Mathematics for Economics, New York University.

[1960] Winters P.R., Forecasting sales by exponentially weighted moving averages. *Management Science*, **6**, pp 324–342.

[2011] Witkowski B., Building a solver for optimisation problems. BSc Thesis, University of Science and Technology in krakow.

[1997] Wolpert D.H., Macready W.G., No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1**, (1), pp 67–82.

[2003] Wong H., Ip W.C., Xie Z., Lui X., Modelling and forecasting by wavelets, and the application to exchange rates. *Journal of Applied Statistics*, **30**, (5), pp 537–553.

[1994] Wood A.A., Chan G., Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of Computational and Graphical Statistics*, **3**, (4), pp 409–432.

[2013] Woodie A., GPUs show big potential to speed pricing routines at banks. *HPC Wire*.

[1991] Wright A.H., Genetic algorithms for real parameter optimization. in Foundation of Genetic Algorithms, ed. G. Rawlins, *First Workshop on the Foundation of Gen. Alg. and Classified Systems*, Los Altos, CA, pp 205-218.

[2009] WST, Bloomberg uses GPUs to speed up bond pricing. WallStreet & Technology.

[2009] Wu Z., Huang N.E., Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, **1**, pp 1–41.

[2007] Xu Y., Yang L., Haykin S., Decoupled echo state networks with lateral inhibition. *Neural Networks*, **20**, (3), pp 365–376.

[2009] Xu S.J., Jin X.J., Predicting drastic drop in Chinese stock market with local Hurst exponent. *International Conference on Management Science and Engineering*, pp 1309–1315.

[2003] Yan W., Profitable, return enhancing portfolio adjustments: An application of genetic programming with constrained syntactic structure. MSc Computer Science Project 2002/2003, University College London, UK.

[2005] Yan W., Clack C.D., Evolving robust GP solutions for hedge fund stock selection in emerging markets. Working Paper, University College London, UK.

[2000] Yang D., Zhang Q., Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business*, **73**, (3), pp 477–491.

[1996] Yao J.T., Poh H.L., Equity forecasting: A case study on the KLSE index. *Neural Networks in Financial Engineering, Proceedings of 3rd International Conference on Neural Networks in the Capital Markets*, Oct 1995, London, A-P.N. Refenes, Y. Abu-Mostafa, J. Moody and A. Weigend, (eds.), World Scientific, pp 341–353.

[1998] Yao J.T., Tan C.L., A study on training criteria for financial time series forecasting. Working Paper.

[2000] Yao J.T., Tan C.L., Time dependent directional profit model for financial time series forecasting. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, July 2000, Como, Italy, **5**, pp 291–296.

[1991] Young T.W., Calmar ratio: A smoother tool. *Futures Magazine*, **20** (1), October.

[2009] Yuan Y., Zhuang X-T., Jin X., Measuring multifractality of stock price fluctuation using multifractal detrended fluctuation analysis. *Physica A*, **388**, (11), pp 2189–2197.

[2002] Yue S., Pilon P., Caradias G., Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrol.*, **259**, pp 254–271.

[1994] Zadeh L.A., Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, **37**, (3), pp 77–84.

[2002] Zaharie D., Critical values for the control parameters of differential evolution algorithms. in R. Matousek, P. Osmera, eds., *Proceedings of MENDEL 2002, 8th International Conference on Soft Computing*, Brno University of Technology, Faculty of Mechanical Engineering, pp 62–67, Institute of Automation and Computer Science.

[2001] Zhang B-L., Coggins R., Jabri M.A., Dersch D., Flower B., Multiresolution forecasting for futures trading using wavelet decompositions. *IEEE Trans. on Neural Networks*, **12**, (4), pp 765–775.

[2005] Zhang D., Jiang Q., Li X., A heuristic forecasting model for stock decision making. *Mathware and Soft Computing*, **12**, pp 33–39.

[2012] Zhang H., Rangaiah G.P., An efficient constraint handling method with integrated differential evolution for numerical and engineering optimization. *Computers and Chemical Engineering*, **37**, pp 74–88.

[1999] Zheng G., Stark J.L., Campbell J., Murtagh F., The wavelet transform for filtering financial data streams. *Journal of Computational Intelligence in Finance*, **7**, pp 18–35.

[2005] Zhu K., A statistical arbitrage strategy. Master's Thesis in Numerical Analysis, Royal Institute of Technology, Stockholm.

[2005] Zielinski K., Peters D., Laur R., Stopping criteria for single-objective optimization. In Proceedings of the Third International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore.

[2006] Zielinski K., Laur R., Constrained single-objective optimization using differential evolution. in *2006 IEEE Congress on Evolutionary Computation*, Vancouver, Canada, pp 223–230.

[2007] Zielinski K., Laur R., Stopping criteria for a constrained single-objective particle swarm optimization algorithm. *Informatica*, **31**, pp 51–59.

[2008] Zielinski K., Laur R., Stopping criteria for differential evolution in constrained single-objective optimization. in Chakrabotry, Ed., Advances in Differential Evolution, **143**, Springer-Verlag, Berlin, pp 111–138.

[2005] Ziemba W.T., The symmetric downside-disk Sharpe ratio. *The Journal of Portfolio Management*, **32**, (1), pp 108–122.

[2006] Zimmermann H.G., Grothmann R., Schaefer A.M., Tietz C., Identification and forecasting of large dynamical systems by dynamical consistent neural networks. in New Directions in Statistical Signal Processing: From systems to brain. MIT Press, edt S.Haykin, J.Principe, T.Sejnowski, J.McWhirter, pp 203–242.

[2003] Zinkevich M., Online convex programming and generalized infinitesimal gradient ascent. *ICML*, pp –.

[2006a] Zumbach G., Back testing risk methodologies from 1 day to 1 year. Technical Report, RiskMetrics Group.

[2006b] Zumbach G., The riskmetrics 2006 methodology. Technical Report, RiskMetrics Group.

[2007] Zunino L., Tabak B.M., Perez D.G., Garavaglia M., Rosso O.A., Inefficiency in Latin-American market indices. *The European Physical Journal B*, **60**, (1), pp 111–121.

[2008] Zunino L., Tabak B.M., Figlioa A., Perez D.G., Garavaglia M., Rosso O.A., A multifractal approach for stock market inefficiency. *Physica A*, **387**, pp 6558–6566.