ADVANCES IN
HIGH FREQUENCY STRATEGIES

ADVANCES IN HIGH FREQUENCY STRATEGIES

A dissertation presented by

Marcos M. López de Prado, Ph.D.

to

The Department of Financial Economics

in partial fulfillment of the requirements
for the degree of
Doctor of Science
in the subject of
Insurance and Financial Risks

Complutense University
Madrid, Spain

December, 2011

Advances in High Frequency Strategies

Complutense University
Madrid, 2011

Library of Congress Control Number: 2011904210
United States Copyright Office: 1-571135451
ISBN: 5-800081-485841

For related research, please visit http://ssrn.com/author=434076
Additional copies can be purchased at http://tinyurl.com/hfpin

A mis hijos.

"Far better an approximate answer to the right question,
which is often vague,
than an exact answer to the wrong question,
which can always be made precise."

John W. Tukey (1915-2000)
The future of data analysis
Annals of Mathematical Statistics, 33 (1), (1962), p.13.

CONTENTS

PREFACE

Motivation

My interest in High Frequency Finance began in early 2006, when I started to trade strategies with shorter holding periods. I quickly realized that many standard Microestructural, Financial and Econometric theories could not be exported into a framework where time and information have a different relationship. I knew the Low Frequency framework well, from my perspective of both a practitioner and an academic. The literature devoted to High Frequency Finance at that time was relatively small, fragmented, and in many cases developed by stretching Low Frequency models, forced to perform in an environment for which they had not been conceived.

Thus began the work on my second doctoral thesis, which covers topics with important practical implications for optimal execution, liquidity provision, market risk, the risk of market failure, … and of course alpha generation by High Frequency strategies. Most of the profits harvested by High Frequency trading nowadays can be attributed to speed. But as our trading speed reaches the limits of Physical feasibility, being fast is no longer enough. Some critics of High Frequency trading argue that a speed limit should be imposed on market participants. We have news for them: That has already been taken care of.

We are witnessing the dawn of a new paradigm in the science of investing. Those who ignore these principles are broadcasting their trading intentions, and become easy prey for predatory algorithms every day. In effect, they are paying a virtual tax for each transaction –enriching their competitors. Not to mention their underestimation of the risks induced by this new market microstructure. Some firms (and academic approaches) will evolve and adapt accordingly, but many will fade and vanish over time.

Acknowledgements

Several models presented here were developed during my term as Visiting Scholar at *Cornell University*. My collaboration with Prof. Dr. David Easley, Chairman of the Economics Department, and Prof. Dr. Maureen O'Hara, former President of the *American Finance Association* (AFA), resulted in the development of the *VPIN Flow Toxicity Metric*, and three patent applications.

INTRODUCTION

High Frequency Strategies

Recent legislative changes in the United States ("Regulation National Market System" of 2005, or "RegNMS") and Europe ("Markets in Financial Instruments Directive" or "MiFID", in force since November 2007), preceded by substantial technological advances in computation and communication, have revolutionized the financial markets.

Europe's MiFID fosters greater competition among *brokers*, with the objective of improving liquidity, cohesion and depth in financial markets. Similarly, U.S. RegNMS encourages competitiveness among *exchanges* by allowing market fragmentation. Cohesion is recovered through a mechanism for the consolidation of individual orders processed via multiple venues (NBBO, or "National Best Bid and Offer"). The result has been an "arms race" for developing the technology and quantitative methods that squeeze the last cent of profitability when serving the demands of market participants.

High frequency strategies are of a very diverse nature. We will follow the general description proposed by Aldridge (2010), defining *high frequency strategies* as those characterized by a brief investment horizon, which may range from a split of a second to several hours. A main advantage comes from placing numerous independent bets every day on the same instrument or portfolio, because as the "Fundamental Law of Active Management" postulates, a tiny predictive power on a sufficiently large number of independent bets yields a high Information Ratio (Grinold (1989)). The goal is to exploit the inefficiencies derived from the market's microstructure, such as rigidities, agents' idiosyncrasy, asymmetric information, etc. As a consequence of this higher frequency, the identification of opportunities, risk control, execution and other investment management activities must be automated. Not all *algorithmic trading* occurs in high frequency, but all high frequency requires algorithmic trading. This in turn has made it possible to interact directly with the exchange's auction mechanism (or "double auction order book").

High frequency traders often operate with proprietary capital, meaning that investors are also investment managers. Their actions are not derived from client orders but for their own benefit. Their servers reside in the proximity of the exchange's matching engine ("co-location"), with the purpose of minimizing the time that passes between the shipping of an order and the arrival of confirmation of reception by the exchange ("latency"). All of the above demands a considerable investment in terms of infrastructure and of course the development of trade secrets in the form of algorithms and quantitative models.

According to Herdershott, Jones and Menkveld (2011), high frequency strategies overall benefit the investment community by lowering trading costs, improving the information in the order book, eliminating arbitrage opportunities across markets, narrowing the bid-ask spread, adding liquidity, etc. Cartea and Penalva (2010) conclude that high frequency strategies increase market impact, with mixed results depending on the type of participant.

Fields of study

The introduction to each chapter presents a detailed analysis of the state of the art regarding the subject discussed. Nevertheless, there exists a set of general and shared themes referred to in multiple occasions across this study that we find convenient to introduce now for the sake of clarity.

Researching high frequency trading models encompasses a wide range of fields, which we could group in the areas of Financial Economics (measurement and management of risks), Statistics (estimation and forecasting of high frequency time series) and Economic Analysis (microstructure of financial markets, price formation and price discovery processes).

a) Statistics

Modern markets require real time pricing of products and risk management. Iati (2009) has estimated that over 70% of the volume of U.S. shares is transacted by high frequency participants. For the year 2010, TABB has estimated that number to be 60% in the U.S., and around 40% in Europa.

The generalization of electronic markets and automation of financial transactions have accelerated the decision-making process to the point of rendering many old standing economic models obsolete. Higher frequency not only means that a large number of actions are taken every day, but also that these actions occur in stochastic time (i.e., with uncertainty regarding the time gap between these decisions). If ten years ago it was difficult to find research based upon tick-by-tick datasets, nowadays it is extraordinary to encounter recently published papers that use daily series.

Goodhart and O'Hara (1997) describe many peculiarities characteristic of high frequency time series, finding that some of them are incompatible with the assumptions generally made by traditional statistical and econometric models (e.g., Hamilton (1994)). As we shall see in Chapter I, high frequency returns time series are serially conditioned (Bollerslev and Domowitz (1993)), are subject to stochastic volatility (Andersen and Bollerslev (1996)), are irregularly spaced, and follow non-Gaussian distributions with elevated kurtosis and asymmetry.

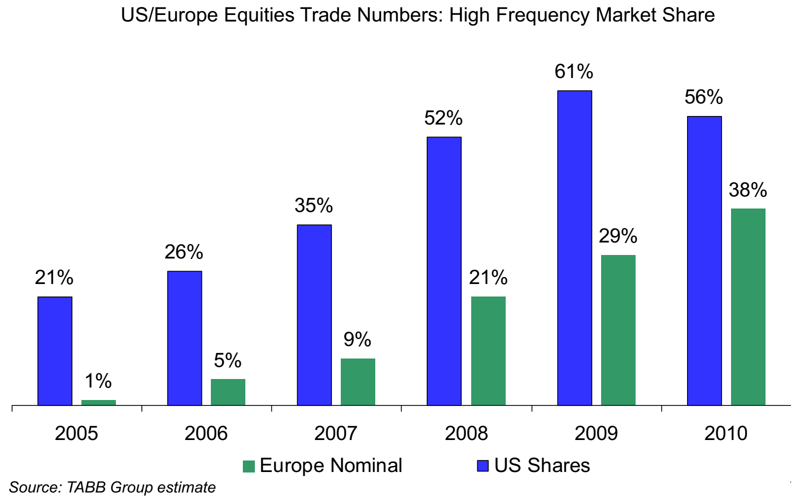

Figure 1 – Volume contributed by High Frequency Traders

Some of these problems can be partially dealt with through complex specifications, such as the ACD models proposed by Engle and Russell (1996), Engle and Russell (2005), or variants discussed by Bauwens and Giot (1998), Dufour and Engle (2000) among others. Refenes et al. (1996) and Bolland et al. (1998) apply Neural Networks. The complexity inherent to many of these models makes them prone to numerical instability. Furthermore, most of them lack any theoretical foundation, becoming purely empirical exercises whose conclusions can hardly be rationalized.

A different approach, consistent with microstructure theory, has attempted to model the impact that information arrival has on prices and bid-ask spreads. Some examples are Almeida et al. (1997), Goodhart et al. (1991), Low and Muthuswamy (1996). Most of this research has focused on FX, and its presence has decayed over the years. A possible explanation is that they relied on daily time series, and the aforementioned problems associated with the modeling of tick data prevented their evolution into the high frequency domain.

Ané and Geman (2000) re-discovered an idea of Clark (1973) that allows for a partial recovery of the properties assumed by most traditional statistical models on high frequency datasets. Instead of sampling by regular time intervals (chronological clock), the authors adopt a sampling subordinated to their measurement of stochastic volatility: The higher the volatility, the more samples are drawn per unit of time. The resulting high frequency time series are closer to normal. Unfortunately, the procedure requires the estimation of instantaneous volatility, which is inaccurate and not directly observable in real time.

Chapter I will address the same problem as Ané and Geman (2000). We present a simple sampling procedure that delivers high frequency time series nearly normal, plus it allows for a reduction of serial correlation and heteroskedasticity.

b) Economic Analysis

O'Hara (1995) describes the purpose of market microstructure theory and explains its motivation:

> *"The study of the process and outcomes of exchanging assets under a specific set of rules. While much of economics abstracts from the mechanics of trading, microstructure theory focuses on how specific trading mechanisms affect the price formation process."*

It is a relatively new area of research that combines elements of economic analysis (agents, expectations, utility maximization) and financial economics (valuation, risks, asset management).

As for this study, we begin with the basic model of sequential trading devised by Glosten and Milgrom (1985). Easley, Kiefer, O'Hara and Paperman (1996) much improved that model by recognizing the existence of participants with asymmetric information under event uncertainty. The outcome was the celebrated PIN ("Probability of Informed Trading") model, which allows market makers to monitor information asymmetry and thus avoid adverse selection. Although a few PIN estimation procedures exist for low frequency data (Easley, Engle, O'Hara and Wu (2008), Easley, Kiefer, O'Hara and Paperman (1996)), there was no possibility of measuring PIN in the high frequency domain because of the intractability of those datasets.

Chapter I is dedicated to solving the problem of estimating VPIN for high frequency data.

c) Financial Economics

The statistical and microstructure models developed in this book address a number of issues treated in the Financial Economics literature. In particular, our contributions aim to further the understanding of the following research topics:

- Models for valuation and risk measurement in the context of high frequency.
- Forecasting of toxicity-induced volatility, as derived from the high frequency series of transactions.
- Development of new financial instruments which may provide a hedge to the risks inherent in high frequency trading.
- A benchmark to assess brokers' performance on behalf of their clients.
- Dealing with *spreads* and computation of optimal hedges based on time series, in the high as well as the low frequency domain.

We believe that this study is pioneering in the aforementioned subjects. Chapter II provides evidence that volatility forecasting can be improved through microstructural models. Those models analyze the behavior of market makers operating under asymmetric information, and explain how their response to order flow toxicity is a source of volatility. Most volatility forecasting models do not incorporate in their specification a theory that explains the origin of volatility, thus they tend to treat volatility as an exogenous, generally univariate, filtered process. We are left with tools that attempt to forecast "something" without an understanding of why and how that "thing" comes to be. But how is an econometric model unsupported by a theory any better than a high-tech horoscope? Under these circumstances, we should consider the possibility that a large part of the results reported in the volatility forecasting literature are in fact spurious, numerically driven, and unrelated to any existing structure. Conversely, Chapter II presents a bivariate, dynamic equilibrium model that studies the interaction between toxicity (signal) and realized volatility (response). Out-of-sample results are superior to those derived from univariate specifications, especially in the context of forecasting beyond the immediate horizon. Empowered with this new analytical tool, Chapter III throws light upon the events of May 6th 2010 ('flash crash'). Chapter IV defines a futures contract that provides a hedge to market makers against the risk of order flow toxicity or adverse selection. Chapter V offers a new benchmark to assess the costs of trading under conditions of asymmetric information.

Chapter VI reviews and improves some of the most used methods for hedging portfolios, plus it introduces two new methods with superior characteristics. These methods are applicable in a High Frequency space, but also in Low Frequency. Chapter VII reviews the scientific literature on Sharpe ratio, offers its projection on the probabilistic space (*Probabilistic Sharpe ratio*), and develops a new methodology which determines the minimum track record length required to evidence investment skill at a preset confidence level. High Frequency returns are non-Gaussian, so a new *Sharpe Ratio Efficient Frontier* is needed to address this feature. Chapter VIII introduces the EF3M algorithm, with important applications in the modeling of financial data. This new estimation procedure is particularly useful in Financial applications, and satisfies the high frequency computational requirements established in Chapter VII.

Objectives

Cahan et al. (2010) argue that a large portion of the financial literature has devoted itself to measuring low frequency risks, to some extent ignoring the specific risks associated with high frequency trading. In particular, those authors call for research that may uncover the relationships between risks in both domains.

A first objective of the present study is to put forth a procedure for transforming high frequency data in order to comply with the assumptions of traditional econometric models. Once we are able to deal with high frequency series with relative simplicity, we address the second objective, namely to provide a measurement of the risk of toxicity in the high frequency order flow (our VPIN model). In particular, we present abundant empirical evidence of the bidirectional relationship between order flow toxicity and future volatility.

A third objective consists in answering the challenge formulated by Cahan et al. (2010). Consequently, we show how high frequency risks *spill over* into the low frequency domain and vice versa. We believe that this study is the first to propose such unified framework, in an attempt to explain the transmission of risk between both domains. A paradigmatic case of how high frequency risks unleash low frequency risks is evidenced by the *'flash crash'* of May 6th 2010, an episode studied in detail in Chapter III.

The fourth objective is to characterize market makers as sellers of an option to be adversely selected, at a premium determined by the range at which they are willing to provide liquidity. We will show that exchanges currently lack a mechanism or tool to protect market makers against the risk of adverse selection. As a solution, we propose a futures contract with VPIN as underlying, which could allow liquidity providers to dynamically manage toxicity risk, avoiding future repetitions of the *'flash crash'*.

The fifth objective is to propose a new benchmark for measuring brokers' efficiency when executing the clients' orders. We argue that VWAP does not incorporate all available information in connection to the level of flow toxicity, thus being an incomplete benchmark.

The sixth objective is to study the optimality of some of the most popular portfolio hedging methods. Solving the problem of portfolio hedging has critical applications in high frequency, not only for risk management but also in trading strategies. This study presents two new methodologies that overcome some of the caveats found in previous methods: DFO and MMSC.

The seventh objective is to analyze the characteristics of returns distributions which are responsible for "inflating" the Sharpe ratio. Such characteristics happen to be intrinsic to high frequency series, from which we can expect a certain upwards bias on the Sharpe ratios derived from high frequency strategies. The conclusions reached allow us to develop an alternative performance measure, named *Probabilistic Sharpe ratio*, which corrects the referred "inflation", and translates Sharpe ratio readings into probabilities of skillful investing. One application of this model is to answer the key question of *"how long should a track record be in order to have statistical confidence that its Sharpe ratio is above a given threshold?"* The empirical evidence we present indicates that, despite the high Sharpe ratios publicized for several hedge fund styles, in many cases they may not be high enough to indicate statistically significant investment skill beyond a moderate annual Sharpe ratio of 0.5 for the analyzed period, confidence level and track record length. This in turn leads to the concept of Sharpe Ratio Efficient Frontier. As we would like to model distributions matching the empirically observed first four moments, we develop the EF3M methodology for the exact fit of a mixture of two Gaussian distributions.

The analysis and management of investments in high frequency is an emerging field that will gradually attract the attention of more researchers. The present study confronts some of the most urgent questions on this subject, such as the measurement and control of high frequency risks, its contagion to the low frequency domain and the computation of hedges. Being a new field of research, the list of pending questions not addressed by this work, and for which (presently) no answer exists, is endless. In particular, we will not participate in the polemic regarding the social benefit or cost derived from high frequency trading. This is an extremely complex debate which would require a monographic book.

Methodology

We will employ an assortment of techniques drawn from mathematical and statistical analysis to answer questions derived from this confluence of areas (Finance, Statistics, Economic Analysis). In particular:

- Probability and econometric methods for time series
    - Chapter II estimates VAR and Granger-causality models in order to analyze VPIN's predictive power on volatility.
    - Chapter II estimates conditional distribution probabilities, correlation surfaces and threshold correlations.
    - Chapter VI makes use of cointegration and error correction models to estimate the optimal hedging vectors on spreads.
    - Chapter VII employs a mixture of Normal distributions to illustrate the inflationary effect that skewed and fat-tailed returns distributions have on Sharpe ratio.

    - Chapter VII develops a projection of Sharpe ratio in the probabilistic domain (PSR).
    - Chapter VIII develops the EF3M algorithm for the exact fit of a Mixture of two Gaussian distributions.
- Monte Carlo methods
    - Chapter I evaluates how accurately VPIN estimates PIN.
    - Chapter V simulates the impact that a variety of serial price correlation scenarios have on VPIN.
    - Chapter V simulates the performance of the VPINC execution algorithm in comparison to VWAP.
    - Chapter VIII applies a Monte Carlo to evaluate the Probability of Departure of an investment strategy.
- Linear algebra
    - Chapter II computes an analytical spectral decomposition of the VAR coefficients matrix.
    - Chapter VI develops a generalized PCA hedging procedure, not bounded by number of instruments or asset class.
- Differential calculus
    - Chapters I, II, V and VI apply optimization methods to identify the global maxima or minima on a variety of problems.
    - Chapter VI solves the analytical derivatives of the objective functions for the MMSC and DFO methods.
    - Chapter VI presents a customized algorithm for the optimization of the MMSC objective function.
    - Chapter VII computes the gradient of the "minimum track record length" due to sampling frequency.
- Differential equations and equations in differences
    - Chapter II develops a dynamic equilibrium model for the determination of the state of the VPIN-Volatility system, in discrete and continuous time.
- Historical simulations
    - Chapters I, II, V and VI estimate the historical performance of multiple variables based upon high frequency time series.
    - Chapter VII determines the minimum backtest size required in order to evidence skill subject to a predefined confidence level.

Organization and format

This study is organized in eight interrelated papers. Four have been peer-reviewed and accepted for publication at scientific journals (the first three co-authored with Profs. Easley and O'Hara):

- Chapter I is the basis for a paper accepted for publication in the *Review of Financial Studies* (forthcoming, 2012).
- Chapter III is the basis for a paper accepted for publication in the *Journal of Portfolio Management* (Winter 2011). http://www.iijournals.com/doi/abs/10.3905/jpm.2011.37.2.118
- Chapter IV is the basis for a paper accepted for publication in the *Journal of Trading* (Spring 2011). http://www.iijournals.com/doi/abs/10.3905/jot.2011.6.2.008
- Chapter VI is the basis for a paper accepted for publication in the *Journal of Investment Strategies* (Risk Journals, forthcoming, 2012).

According to *"The Social Science Research Network"* (SSRN), these papers have been downloaded by up to 35,000 social scientists members of this network (http://ssrn.com/author=434076). Some of the papers derived from this study have been included in the top 10 ranking for the most read papers in the history of SSRN, in the areas of Finance, Economic models and Econometrics:

- http://papers.ssrn.com/sol3/topten/topTenResults.cfm?groupingId=1152425&netorjrnl=jrnl
- http://papers.ssrn.com/sol3/topten/topTenResults.cfm?groupingId=1153655&netorjrnl=jrnl
- http://papers.ssrn.com/sol3/topten/topTenResults.cfm?groupingId=1153629&netorjrnl=jrnl

Compliant with Complutense University's regulations, we include an extensive summary of the findings and conclusions, written in Spanish.[1]

Seminars

A number of the models included in this work have been presented in international seminars:

1. Opening Keynote Speech at *"Best Execution USA 2010"* conference:
    a. Title: *"Flash Crash: Dissecting what happened and preventing it from happening again"*.
    b. Location: New York City
    c. Date: October 6th 2010.
    d. Host: Risk Magazine.
    e. Link: http://ev153.eventive.incisivecms.co.uk/static/day11
2. Seminar offered to CFTC Commissioners and Researchers:
    a. Title: *"The Microstructure of the Flash Crash"*.

[1] Art. 4.3 of UCM's Directive of 12/02/2010, developing the Royal Decree 1393/2007 of October 29th (B.O.E. 30/10/2007).

Carlin, Sousa Lobo and Viswanathan (2007) developed a model of how predatory trading can lead to episodic liquidity crises and contagion. We have found that, over the last three years, hundreds of extreme price actions can be associated with failures in the liquidity provision process. It is not our goal to demonstrate that predatory algorithms are to blame. We are content with making the case that: i) liquidity crises are becoming more recurrent and ii) that this is happening despite the extraordinary degree of sophistication achieved by high-frequency market makers. We believe that a plausible explanation to this apparent contradiction is that predatory algorithms are taking advantage of the commitment of high-frequency market makers, just as macro traders would not let pass an opportunity like U.K.'s ERM episode.

Brunnermeier and Pedersen (2005) theorized that predatory trading could amplify contagion and price impact in related markets. This amplification would not be driven by a correlation in economic fundamentals or by information spillovers, but rather by the composition of the holdings of large traders who must significantly reduce their positions.

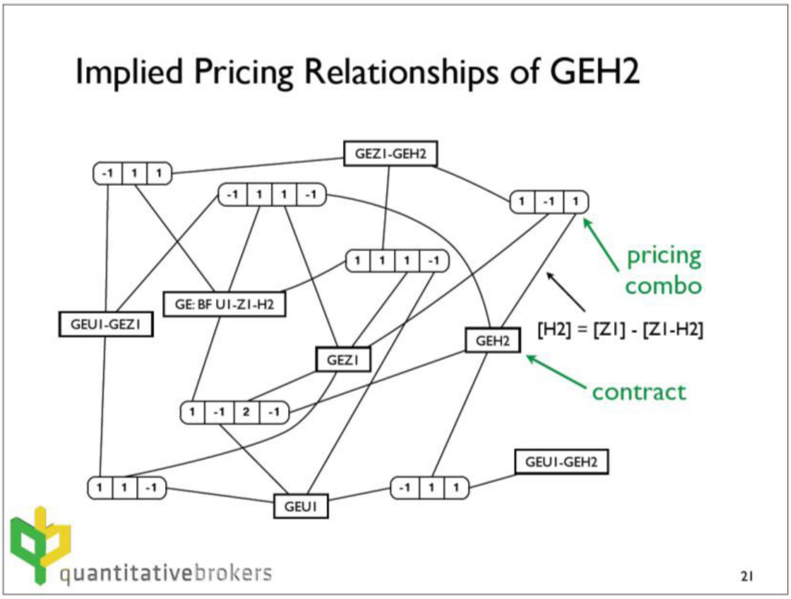

Figure 32 – Discovering hidden liquidity in the GEH2 contract

The dynamics of the order books are interrelated across multiple products. Figure 32 illustrates how, in order to decide at what level to place a client's order on Eurodollar short futures, Quantitative Brokers' algorithms analyze 6 different relationships in real time, searching for *hidden liquidity* (liquidity that, although is not displayed in that particular book, is implied by the liquidity present in the related books). Consequently, in order to operate on

2.4.1. Specification of the spill over mechanism

A system of equations in differences suits our problem, since we can only estimate VPIN in discrete (volume) time, i.e. updated every time a new volume bucket is completed.

The system is composed of two equations. The first one forecasts the absolute return over the next volume bucket as a function of the last absolute return and reading of log VPIN. The second equation forecasts the log VPIN as a function of the same variables as in the first equation. Evidently, the second equation is redundant for a unit forecasting horizon *k*, but as *k>1* the second equation allows us to incorporate in the first one the feedback mechanism.

Consider the system

$$p_1(t+1) = \beta_{1,1} p_1(t) + \beta_{1,2} p_2(t)$$
$$p_2(t+1) = \beta_{2,1} p_1(t) + \beta_{2,2} p_2(t)$$

with initial conditions $p_1(0)$, $p_2(0)$, where $p_1(t) = \left| \frac{P_t}{P_{t-1}} - 1 \right|$ and $p_2(t) = Ln(VPIN_t)$.[39]

Its matrix representation is $p(t+1) = \beta\ p(t)$, where $p(t+1) = \begin{bmatrix} p_1(t+1) \\ p_2(t+1) \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{bmatrix}$.

The key role played by the second equation can be illustrated with an example. Suppose that market makers are being impacted by excessive flow toxicity. As they widen their trading ranges, volatility increases. Those market makers who were ‘slow’ in adjusting will suffer losses and vanish, which will force the ‘fast’ market makers to re-adjust. Some of those ‘fast’ market makers may miss this re-adjustment, being driven away in a second wave of losses, and so on. At some point, the chain of events may be broken (e.g., by new market makers making their appearance, attracted by wide spreads), or lead to an ‘explosive’ state. The feedback dynamics do not affect volatility over the next bucket, but it adds information over the long run, explaining the mechanism by which high frequency risks *spill over* the low frequency domain. A univariate or single-equation specification fails to provide an explanation for such spillover.

[39] Centering these variables introduces an intercept, if so desired.

2.4.2. Eigenvalues of the characteristic matrix

We know that $\beta W = W\Lambda$, which leads to the eigenvalue equation $|\beta - I\Lambda| = 0$, where W is the matrix of eigenvectors and Λ is the matrix of eigenvalues.

$\begin{vmatrix} \beta_{1,1} - \lambda & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} - \lambda \end{vmatrix} = 0 \Rightarrow (\beta_{1,1} - \lambda)(\beta_{2,2} - \lambda) - \beta_{1,2}\beta_{2,1} = 0$, a second degree equation with roots in $\lambda_1 = \dfrac{Tr(\beta) + \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$; $\lambda_2 = \dfrac{Tr(\beta) - \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$, where $Tr(\beta)$ is the trace of β and $|\beta|$ its determinant. Thus, $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$.

2.4.3. Eigenvectors of the characteristic matrix

Λ has been found to make the matrix $\beta - I\Lambda$ singular. Let's compute β's eigenvectors by finding $\beta - I\Lambda$'s kernel.

For λ_1, we establish a system $\begin{bmatrix} \beta_{1,1} - \lambda_1 & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} - \lambda_1 \end{bmatrix} \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Elemental row operations will yield $\begin{bmatrix} 1 & \dfrac{\beta_{1,2}}{\beta_{1,1} - \lambda_1} \\ 0 & 1 \end{bmatrix}$, and so we reduce the system to:

$$w_{1,1} + w_{2,1}\frac{\beta_{1,2}}{\beta_{1,1} - \lambda_1} = 0$$

$$w_{2,1} = 1$$

Applying the especial solutions on the kernel allow us to conclude that $W = \begin{bmatrix} \dfrac{-\beta_{1,2}}{\beta_{1,1} - \lambda_1} & \dfrac{-\beta_{1,2}}{\beta_{1,1} - \lambda_2} \\ 1 & 1 \end{bmatrix}$.

Let's write $p(0) = Wc$, where c is, like before, the column vector that solves W for the initial conditions $p(0)$. Assuming that β has all independent eigenvectors, we know that β is diagonalizable and $\beta = W\Lambda W^{-1}$. Then, $p(1) = \beta p(0) = W\Lambda W^{-1}Wc = W\Lambda c$. Multiplying k times by β will yield $p(k) = \beta^k p(0) = W\Lambda^k c$, or what is the same, $p(k) = \begin{bmatrix} p_1(k) \\ p_2(k) \end{bmatrix} = c_1\lambda_1^k \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} + c_2\lambda_2^k \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix}$.

At the initial conditions, $p(0) = c_1 \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} + c_2 \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix}$, which is a system we already solved: $c_1 = p_2(0) - c_2$, $c_2 = \dfrac{p_1(0) - p_2(0)w_{1,1}}{w_{1,2} - w_{1,1}}$.

2.4.4. The solution

$$\begin{bmatrix} \left| \dfrac{P_{t+k}}{P_{t+k-1}} - 1 \right| \\ Ln\left(VPIN_{t+k}\right) \end{bmatrix} = c_1\lambda_1^k \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} + c_2\lambda_2^k \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix}$$, where:

- $\begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} = \begin{bmatrix} \dfrac{-\beta_{1,2}}{\beta_{1,1} - \lambda_1} & \dfrac{-\beta_{1,2}}{\beta_{1,1} - \lambda_2} \\ 1 & 1 \end{bmatrix}$.
- $\lambda_1 = \dfrac{Tr(\beta) + \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$; $\lambda_2 = \dfrac{Tr(\beta) - \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$, where $Tr(\beta)$ is the trace of β and $|\beta|$ its determinant.
- $c_1 = Ln\left(VPIN_0\right) - c_2$; $c_2 = \dfrac{\left| \dfrac{P_1}{P_0} - 1 \right| - Ln\left(VPIN_0\right) w_{1,1}}{w_{1,2} - w_{1,1}}$.

Our solution is analytical and the forecast can be estimated in a single calculation for any horizon (no sequential estimation is needed). This is an important advantage for the purpose of integrating our equations in optimization exercises.

2.4.5. Stability conditions

Now that we know how to estimate our dynamic system, we would like to understand what causes a crash from a mathematical standpoint. Later on, we will offer an interpretation from a market microstructure perspective.

The previous analysis is powerful in the sense of establishing the conditions for the system to be *stable*, *steady* or *explosive* in discrete time:

- *Stable state*: $\left|\lambda_i\right| < 1; i = 1,2$. Both eigenvalues must be smaller than one in absolute value. If imaginary eigenvalues exist, their real part must be smaller than one in absolute value.
- *Steady state*: $\exists i, j \| \lambda_i \right| = 1, \left|\lambda_j\right| < 1$. The absolute value of one eigenvalue is equal to one, and the other is not greater than one in absolute value (or their real part, being imaginary).
- *Explosive state*: $\exists i \| \lambda_i \right| > 1$. The absolute value of any eigenvalue is greater than one (or their real part, being imaginary).

2.5. The continuous time model

2.5.1. Model specification

Consider two variables with levels $p_1(t), p_2(t)$, mutually related by a system of differential equations:

$$\begin{aligned}\frac{dp_1(t)}{dt} &= \beta_{1,1} p_1(t) + \beta_{1,2} p_2(t) \\ \frac{dp_2(t)}{dt} &= \beta_{2,1} p_1(t) + \beta_{2,2} p_2(t)\end{aligned}$$

, with initial conditions $p_1(0)$, $p_2(0)$ where

$p_1(t) = \left|\frac{P_t}{P_{t-1}} - 1\right|$ and $p_2(t) = Ln(VPIN_t)$.

Its matrix representation is $\frac{dp(t)}{dt} = \beta\, p(t)$, where $\frac{dp(t)}{dt} = \begin{bmatrix} \frac{dp_1(t)}{dt} \\ \frac{dp_2(t)}{dt} \end{bmatrix}$,

$\beta = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{bmatrix}$, $p(t) = \begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix}$.

Let's assume that β is diagonalizable[40], i.e. it has linearly independent eigenvectors. Under this assumption, we will solve this system and, furthermore, study its dynamics, stability conditions and possible equilibrium.

[40] A non-necessary but sufficient condition is having all different eigenvalues.

We have already derived $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, with $\lambda_1 = \frac{Tr(\beta) + \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$ and $\lambda_2 = \frac{Tr(\beta) - \sqrt{Tr(\beta)^2 - 4|\beta|}}{2}$, as well as $W = \begin{bmatrix} \frac{-\beta_{1,2}}{\beta_{1,1}-\lambda_1} & \frac{-\beta_{1,2}}{\beta_{1,1}-\lambda_2} \\ 1 & 1 \end{bmatrix}$.

2.5.2. The solution

The general solution of this system has the form[41] $p(t) = c_1 e^{\lambda_1 t} \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} + c_2 e^{\lambda_2 t} \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix}$, all of which has been previously derived except for the $\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$. This we can compute thanks to the initial conditions of the system ($p(0)$). At *t=0*, we know that

$$\begin{bmatrix} \frac{-\beta_{1,2}}{\beta_{1,1}-\lambda_1} & \frac{-\beta_{1,2}}{\beta_{1,1}-\lambda_2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} p_1(0) \\ p_2(0) \end{bmatrix}.$$

Finally, $c_1 = p_2(0) - c_2$, $c_2 = \frac{p_1(0) - p_2(0) w_{1,1}}{w_{1,2} - w_{1,1}}$.

Note how similar the solution to the system in differences is to the solution of the system of differential equations. However, this similarity is only in structure, because β (and therefore also Λ, W and c) will have different values.

---

[41] For a simple proof, see that $\frac{dp(t)}{dt} = \beta\ p(t)$ can be rewritten in terms of its pure solutions. For example, $\frac{dp(t)}{dt} = \lambda_1 e^{\lambda_1 t} \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix}$ and $\beta\ p(t) = \beta e^{\lambda_1 t} \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix}$, which are one and the same since we know that $\lambda_1 \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} = \beta \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix}$ from linear algebra.

2.5.3. Stability conditions of the differential specification

Looking at the general form of the solution, $p(t) = c_1 e^{\lambda_1 t} \begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} + c_2 e^{\lambda_2 t} \begin{bmatrix} w_{1,2} \\ w_{2,2} \end{bmatrix}$, we ought to distinguish among three alternative outcomes in continuous time:

- *Stable state*: Both eigenvalues are negative. If imaginary eigenvalues exist, their real part must be negative.
- *Steady state*: If at least one eigenvalue is null and the other is negative (or their real part, if imaginary).
- *Explosive state*: If any eigenvalue is positive (or their real part, being imaginary).

2.6. Empirical results

Chapter I showed that a VAR model on log VPIN and Absolute Returns had more predictive power *over the next observation in-sample* than an AR model on Absolute Returns. We would like to test how much more predictive our dynamic system is in a *multi-horizon out-of-sample forecast*.

After computing VPINs on our standard *(V,n)=(50,250)* combination, we have fitted our dynamic system every 50 buckets (equivalent to a day on average) starting January 1st 2008 on samples of 250 buckets (encompassing 1 week worth on data on average). After every fit, we have computed out-of-sample forecasts 1, 2, …, 50 buckets forward (equivalent to 1 day ahead on average). We have compared each k-horizon forecast with the realized absolute return, which gives us the out-of-sample forecasting error.

Let's denote τ the bucket at which a fit has occurred. As discussed, we can compute the *k=1,…,50* forecasting errors that follow our fit at bucket τ as:

$$e_\tau(k) = \left|\frac{P_{\tau+k}}{P_{\tau+k-1}} - 1\right| - E_\tau\left[\left|\frac{P_{\tau+k}}{P_{\tau+k-1}} - 1\right|\right]$$

Table 3 and 4 compare the standard deviation of the out-of-sample forecasting errors of the autoregressive univariate specification with those of the dynamic system.

Two important aspects can be extracted from these results:

1. VPIN improves the single horizon forecast of volatility (*k=1*). The univariate forecast (20% StDev) is more unreliable than the bivariate forecast that includes VPIN (17%). We knew this from Section 2.3.
2. As forecasts are produced beyond the immediate horizon (*k>1*), our dynamic system's confidence does not significantly decay, while the confidence of the univariate forecast persistently deteriorates.

$$\frac{\partial^2 \sigma_{\Delta S}^2}{\partial {\omega_i}^2} = 2\sigma_{\Delta \mathrm{P}_i}^2 \tag{39}$$

$$\frac{\partial^2 \sigma_{L(S)}}{\partial {\omega_i}^2} = -\frac{1}{8}\left[\sigma_{L(S)}^2\right]^{-2}\left(\frac{\partial \sigma_{L(S)}^2}{\partial \omega_i}\right)^2 \frac{\partial^2 \sigma_{L(S)}^2}{\partial {\omega_i}^2} \tag{40}$$

$$\frac{\partial^2 \sigma_{L(S)}^2}{\partial {\omega_i}^2} = 2\sigma_{L(\mathrm{P}_i)}^2 \tag{41}$$

$$\begin{aligned}\frac{\partial^2 \rho_{\Delta S,L(S)}}{\partial {\omega_i}^2} = \left(\sigma_{\Delta S}\sigma_{L(S)}\right)^{-4}\Bigg[\Bigg(&\frac{\partial\left[\sigma_{\Delta S}\sigma_{L(S)}\right]}{\partial \omega_i}\frac{\partial \sigma_{\Delta S,L(S)}}{\partial \omega_i} \\ &- \frac{\partial^2\left[\sigma_{\Delta S}\sigma_{L(S)}\right]}{\partial {\omega_i}^2}\rho_{\Delta S,L(S)} \\ &- \frac{\partial \rho_{\Delta S,L(S)}}{\partial \omega_i}\frac{\partial\left[\sigma_{\Delta S}\sigma_{L(S)}\right]}{\partial \omega_i}\Bigg)\sigma_{\Delta S}^2\sigma_{L(S)}^2 \\ &- \frac{\partial\left[\sigma_{\Delta S}^2\sigma_{L(S)}^2\right]}{\partial \omega_i}\Bigg(\frac{\partial \sigma_{\Delta S,L(S)}}{\partial \omega_i}\sigma_{\Delta S}\sigma_{L(S)} \\ &- \frac{\partial\left[\sigma_{\Delta S}\sigma_{L(S)}\right]}{\partial \omega_i}\rho_{\Delta S,L(S)}\Bigg)\Bigg]\end{aligned} \tag{42}$$

$$\begin{aligned}&\frac{\partial^2 \widehat{DF}}{\partial {\omega_i}^2} \\ &= \frac{\partial^2 \rho_{\Delta S,L(S)}}{\partial {\omega_i}^2}\left(\left(\frac{1-\rho_{\Delta S,L(S)}^2}{T-2}\right)^{-\frac{1}{2}}\right. \\ &\left.+ \frac{1}{(T-2)}\left(\frac{1-\rho_{\Delta S,L(S)}^2}{T-2}\right)^{-\frac{3}{2}}\rho_{\Delta S,L(S)}^2\right) \\ &+ \frac{\partial \rho_{\Delta S,L(S)}}{\partial \omega_i}\left(\left(\frac{1-\rho_{\Delta S,L(S)}^2}{T-2}\right)^{-\frac{3}{2}}\rho_{\Delta S,L(S)}\frac{\partial \rho_{\Delta S,L(S)}}{\partial \omega_i}\right. \\ &+ \left(\frac{3}{(T-2)}\left(\frac{1-\rho_{\Delta S,L(S)}^2}{T-2}\right)^{-\frac{5}{2}} 2\rho_{\Delta S,L(S)}\frac{\partial \rho_{\Delta S,L(S)}}{\partial \omega_i}\right)\rho_{\Delta S,L(S)}^2 \\ &\left.+ \frac{2}{T-2}\rho_{\Delta S,L(S)}\frac{\partial \rho_{\Delta S,L(S)}}{\partial \omega_i}\left(\frac{1-\rho_{\Delta S,L(S)}^2}{T-2}\right)^{-\frac{3}{2}}\right)\end{aligned} \tag{43}$$

BIBLIOGRAPHY

- Admati, A. and P. Pfleiderer, (1988): "A Theory of Intra-day Patterns: Volume and Price Variability," Review of Financial Studies, 1(1), 3-40.
- Aldridge, I. (2010): "High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems", Wiley, New York.
- Alexander, C. (2001): "Option pricing with Normal Mixture returns". ISMA Centre Research paper.
- Alexander, C. (2004): "Normal mixture diffusion with uncertain volatility: Modeling short- and long-term smile effects". Journal of Banking & Finance 28 (12).
- Alexander, C.O. and I. Giblin (1997): "Multivariate embedding methods: Forecasting high-frequency data in the first INFFC", Journal of Computational Intelligence in Finance 5(6), 17-24.
- Almeida A. and C. Goodhart (1998): "The effect of macroeconomic new on high frequency Exchange rate behavior", Journal of Financial and Quantitative Analysis, 33(3).
- Almgren, R. (2003): "Optimal Execution with Nonlinear impact functions and trading-enhanced risk", Applied Mathematical Finance (10), 1-18.
- Almgren, R. and N. Chriss (2001): "Optimal Execution of portfolio transactions", Journal of Risk (3), 5-139.
- Andersen, T. and T. Bollerslev (1996): "DM-dollar volatility: Intraday activity patters, macroeconomic announcements and longer run dependencies", working paper Nr. 217, Kellogg Graduate School of Management, Northwestern University.
- Andersen, T., T. Bollerslev, F. Diebold and P. Labys (2003): "Modeling and Forecasting Realized Volatility", Econometrica, 71, 529-626.
- Ané, T. and H. Geman (2000): "Order flow, transaction clock and normality of asset returns", The Journal of Finance, 55: 2259–2284.
- Antelo Suárez, M. (1998): "Microeconomía II", Ed. Tórculo.
- Arnuk, L. and J. Saluzzi (2008): "Toxic Equity Trading Order Flow and Wall Street", Themis Trading LLC White Paper, December 17. http://www.themistrading.com/article_files/0000/0348/Toxic_Equity_Trading_on_Wall_Street_12-17-08.pdf

- Bauwens, L. and P. Giot (1998): "The logarithmic ACD model: An application to the bid-ask quotes process of two NYSE stocks", CORE working paper Nr. 9789, Centre for Operations Research and Economics, Catholic University of Louvain.
- Berkowitz, S., D. Logue and E. Noser (1988): "The total cost of transactions on the NYSE", Journal of Finance, 41, 97-112.
- Bethel, W., D. Leinweber, O. Ruebel and K. Wu (2011): "Federal Market Information Technology in the Post Flash Crash Era: Roles for Supercomputing". Lawrence Berkeley National Laboratory, http://ssrn.com/abstract=1939522
- Bishop, C. (2006): "Pattern recognition and machine learning". New York: Springer.
- Blanco, J.A. and H. Mueller (1988): "Put-Optionen als Instrumente der Portfolioinsurance: Investitionsstrategien fuer institutionelle Anleger?", Schweiz. Zeitschrift fuer Volkswirtschaft und Statistik, Heft 3.
- Bolland, P., J. Connor and A. Refenes (1998): "Application of Neural Networks to forecast high frequency data: Foreign Exchange", en C. Dunis and B. Zhou (ed.) "Non-Linear Modeling of High Frequency Financial Time Series", Wiley.
- Bollerslev, T. and I. Domowitz (1993): "Trading patterns and prices in the interbank foregin exchange market", Journal of Finance 48, 1421-1443.
- Bollerslev, T., R. Engle and D. Nelson (1994): "ARCH Models", in R. Engle and D. McFaddden (eds.), Handbook of Econometrics, Volume IV, 2959-3038. Amsterdam, North-Holland.
- Box, G.E. and G.C. Tiao (1977): "A canonical analysis of multiple time series", Biometrika 64(2), 355.
- Box, G., W. Hunter, J. Hunter and W. Hunter (1978): "Statistics for experimenters", Wiley.
- Brigo, D., F. Mercurio and G. Sartorelli (2002) "Lognormal-Mixture Dynamics under different means". UBM. Working paper.
- Brogaard, J. (2010): "High Frequency Trading and Its impact on Market Quality", Working Paper, Northwestern University.
- Brooks, C. and H. Kat (2002): "The Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors", Journal of Alternative Investments, 5(2), Fall, 26-44.
- Brunnermeier, M. and L.H. Pedersen (2005): "Predatory Trading", Journal of Finance, 40(4), August, 1825-1863
- Brunnermeier, M. and L. Pedersen (2009): "Market Liquidity and Funding Liquidity", Review of Financial Studies, 22(6), 2201-2238.
- Cahan, R., Y. Luo, J. Jussa and M. Álvarez (2010): "Signal Processing: Frequency Arbitrage", Deutsche Bank Quantitative Strategy, November 10.
- Cancelo de la Torre, J. R. (1987): "Álgebra lineal", Ed. Tébar.

- Carlin, B., M. Sousa Lobo and S. Viswanathan (2005): “Episodic Liquidity Crises: Cooperative and Predatory Trading”, Journal of Finance, 42(5), October, 2235-2274.
- Cartea, A. and J. Penalva (2010): “Where is the value in High Frequency Trading?”, SSRN, http://ssrn.com/abstract=1712765
- CFTC-SEC (2010): "Findings Regarding the Market Events of May 6, 2010", September 30.
- CFTC-SEC (2010): "Preliminary findings regarding the Market Events of May 6", 2010, May 18.
- Chaboud, A., Hjalmarsson, E., Vega, C. and Chiquoine, B. (2009): “Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market”, FRB International Finance Discussion Paper No. 980.
- Chlistalla, M. (2011): “High Frequency Trading: Better than its reputation?”, Deutsche Bank Research, February 7th.
- Christie, S. (2005): “Is the Sharpe Ratio Useful in Asset Allocation?”, MAFC Research Papers No.31, Applied Finance Centre, Macquarie University.
- Clark, Peter K. (1973): “A subordinated stochastic process model with finite variance for speculative prices”, Econometrica, 41, pp. 135-155.
- Clarke R., de Silva H., Thorley S. (2011): “Minimum-Variance Portfolio Composition”, The Journal of Portfolio Management, Winter.
- CME Group (2010): “Statement on the Joint CFTC/SEC Report Regarding the Events of May 6”, October 1.
- Cohen, C. (1967): “Estimation in Mixtures of Two Normal Distributions”. Technometrics, 9(1), pp. 15-28.
- Copeland, T. and D. Galai (1983): “Information effects on the bid-ask spread,” Journal of Finance, 38(5).
- Craigmile, P. and D. Titterington (1997): “Parameter estimation for finite mixtures of uniform distributions”. Communications in Statistics - Theory and Methods, 26 (8), pp. 1981-1995.
- Dacorogna, M., R. Gencay, U. Mueller, R. Olsen, O. Pictet (2001): “An introduction to High-Frequency Finance”. Academic Press, 1st edition.
- Day, N. (1969): “Estimating the components of a mixture of two normal distributions”, Biometrika 56, pp. 463-474.
- D’Aspremont, A. (2008): “Identifying small mean reverting portfolios”, Quantitative Finance (forthcoming).
- DeGennaro, R.P. and R.E. Shrieves (1995): “Public information releases, private information arrival and volatility in the FX market”, in HFDF-1: Fist International Conference on High Frequency Data in Finance, Volume 1. Olsen and Associates, Zurich.

- Dempster, A., N. Laird and D. Rubin (1977): “Maximum Likelihood from Incomplete Data via the EM Algorithm”. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), pp. 1–38.
- Deusaker, P. and T. Johnson (2011): “Market Liquidity and Flow-driven Risk”, Review of Financial Studies, 24(3), 721-753.
- Dickey, D.A. and W.A. Fuller (1979): “Distribution of the Estimators for Autoregressive Time Series with a Unit Root”, Journal of the American Statistical Association, 74, pp. 427–431.
- Donefer, B.S. (2010): “Algos Gone Wild: Risk in the World of Automated Trading Strategies”, The Journal of Trading, 5, 31-34.
- Dufour, A. and R. Engle (2000): “Time and the Price Impact of a trade”, Journal of Finance 55, 2467-2498.
- Easley, D. and M. O’Hara (1987): “Price, Trade Size, and Information in Securities Markets”, Journal of Financial Economics, 19.
- Easley, D. and M. O’Hara (1992a): “Adverse Selection and Large Trade Volume: The Implications for Market Efficiency”, Journal of Financial and Quantitative Analysis, 27(2), June, 185-208.
- Easley, D. and M. O’Hara (1992b): “Time and the process of security price adjustment”, Journal of Finance, 47, 576-605.
- Easley, D., Kiefer, N., and M. O’Hara (1997): “One Day in the Life of a Very Common Stock”, Review of Financial Studies, Fall.
- Easley, D., Kiefer, N., O’Hara, M. and J. Paperman (1996): “Liquidity, Information, and Infrequently Traded Stocks”, Journal of Finance, September.
- Easley, D., N. Kiefer and M. O’Hara (1997a): “The Information Content of the Trading Process”, Journal of Empirical Finance, No. 4.
- Easley, D., R. F. Engle, M. O’Hara and L. Wu (2008): “Time-Varying Arrival Rates of Informed and Uninformed Traders”, Journal of Financial Econometrics.
- Eisler, Zoltan, J.-P. Bouchaud and J. Kockelkoren (2011): “The Impact of order book events: Market orders, limit orders and cancellations”, working paper, August.
- Elliott, G., Rothenberg, T. J. & J. H. Stock (1996): “Efficient Tests for an Autoregressive Unit Root”, Econometrica, Vol. 64, No. 4., pp. 813–836.
- Engle, R. and C. Granger (1987): “Cointegration and error correction: Representation, estimate and testing”, Econometrica, 55, 251-276.
- Engle, R. and J. Lange (2001): “Predicting VNET: A model of the dynamics of market depth”, Journal of Financial Markets, 4, 113-142.

- Engle, R. and J. Russell (1996): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," Econometrica (1998) 66: 1127-1162.
- Engle, R. and J. Russell (2005): "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive conditional Multinomial-Autoregressive Conditional Duration Model," Journal of Business and Economic Statistics, 166-180, V23, No. 2
- Engle, R., (1996): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," Econometrica (1998) 66: 1127-1162.
- Fabozi, F., S. Focardi and C. Jonas (2011): "High-Frequency Trading: Methodologies and market impact", Review of Futures Markets, 19, 7-38
- Farlow, Stanley (1993): "Partial Differential Equations for Scientists and Engineers", Ed. Dover Publications, 1st Edition.
- Favre, L. and J. Galeano (2002): "Mean-Modified Value-at-Risk optimization with hedge funds". Journal of Alternative Investments, 5 (2), pp. 21–25.
- Foucault, T., O. Kadan and E. Kandel (2009): "Liquidity Cycles and Make/Take Fees in Electronic Markets". http://ssrn.com/abstract=1342799
- Foucault, T., O. Kadan and E. Kandel (2005): "Limit Order Book as a Market for Liquidity", Review of Financial Studies, 18(4), 1171-1217.
- Foucault T. (1999): "Order flow composition and trading costs in a dynamic limit order book", Journal of Financial Markets 2, 99-134.
- Ghysels, E., A. Harvey and E. Renault (1996): "Stochastic Volatility", in G.S. Maddala (ed.), Handbook of Statistics, Volume 14, Statistical Methods in Finance, 119-191. Amsterdam, North-Holland.
- Glosten, L. R. and P. Milgrom (1985): "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders", Journal of Financial Economics, 14, 71-100.
- Goodhart, C. and L. Figliouli (1991): "Every minute counts in financial markets", Journal of International Money and Finance 10, 23-52.
- Goodhart, C. and M. O'Hara (1997): "High Frequency Data in Financial Markets: Issues and Applications", Journal of Empirical Finance 4, 73-114.
- Gosh, Asim (1993): "Hedging with stock index futures: Estimation and forecasting with error correction model", The Journal of Futures Markets, October, p.743

- Grinold, R., R. Kahn (1999): "Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk", McGraw-Hill, 2nd Edition.
- Grinold, R. (1989): "The Fundamental Law of Active Management", Journal of Portfolio Management, 15(3), Spring, 30-37.
- Hamilton, J. (1994): "Time Series Analysis", Princeton.
- Hasbrouck, J. and G. Saar (2010): "Low Latency Trading", Working Paper, Cornell University.
- Hasbrouck, J. (2007): "Empirical Market Microstructure", Oxford.
- Hendershott, T., C. Jones amd A. Menkveld (2011): "Does Algorithmic Trading Improve Liquidity?", Journal of Finance , Journal of Finance, 66(1), 1-33.
- Hendershott, T. and R. Riordan, (2009): "Algorithmic Trading and Information", NET Institute Working Paper No. 09-08.
- Huang, J. and J. Wang (2009): "Liquidity and Market Crashes, Review of Financial Studies", 22 (7): 2607-2643.
- Hwang, S., S. Satchell (1999): "Modeling emerging markets risk premia using higher moments". International Journal of Finance and Economics, 4, pp. 271-296.
- Iati, R. (2009): "High Frequency Trading Technology", TABB Group.
- J.P. Morgan (1997): "RiskMetrics Technical Documents", 4th edition, New York.
- Jeria, D. and G. Sofianos (2008): "Passive orders and natural adverse selection", Street Smart, 33, September 4.
- Johansen, S. (1991): "Cointegration and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models", Econometrica, 59(6), November, pp.1551-1580.
- Jones, C.M., G. Kaul and M.L. Lipton (1994): "Transactions, Volume and Volatility", Review of Financial Studies 7(4), 631-651.
- Jurcenzko, E. and B. Maillet (2002): "The Four-Moment Capital Asset Pricing Model: Some Basic Results". EDHEC-Risk Institute, working paper.
- Kirilenko, A., A. Kyle, M. Samadi and T. Tuzun (2010): "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market", SSRN, Working paper.
- Kissell, R. and M. Glantz (2003): "Optimal trading strategies", American Management Association.
- Kokot, S. (2004): "The Econometrics of Sequential Trade Models", Lecture Notes in Economics and Mathematical Systems, Springer.
- Kwiatkowski, D., P. C. B. Phillips, P. Schmidt and Y. Shin (1992): "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root", Journal of Econometrics 54, 159–178.
- Kyle, Albert S. (1985): "Continuous Auctions and Insider Trading", Econometrica 53, 1315-1335.

- Lee, C.M.C. and M.J. Ready (1991): "Inferring trade direction from intraday data", The Journal of Finance, 46, 733-746.
- Littermann, R. and J. Sheinkman (1991): "Commons Factors Affecting Bond Returns", Journal of Fixed Income, June, pp 54-61.
- Lo, A. (2002): "The Statistics of Sharpe Ratios", Financial Analysts Journal, 36-52.
- López de Prado, M., D. Easley and M. O'Hara (2011a): "The Exchange of Flow Toxicity", Journal of Trading (forthcoming, Spring 2012).
- López de Prado, M. (2011b): "Exact Fit for a Mixture of 2 Gaussians: The EF3M algorithm", Cornell University, Johnson School Research Paper Series (39).
- López de Prado, M. (2011c): "Advances in Cointegration and Subset Correlation Hedging Methods", Cornell University, Johnson School Research Paper Series (40).
- López de Prado, M., D. Easley and M. O'Hara (2010a): "Flow Toxicity and Volatility in a High Frequency World", Review of Financial Studies (forthcoming).
- López de Prado, M., D. Easley and M. O'Hara (2010b): "The Microstructure of the Flash Crash: Flow toxicity, liquidity crashes and the Probability of Informed Trading", Journal of Portfolio Management (forthcoming, Winter 2011).
- López de Prado, M. (2008): "The Sharpe ratio Efficient Frontier". RCC at Harvard University, http://ssrn.com/abstract=1821643
- López de Prado, M. and C. Rodrigo Illera (2004): "Invertir en hedge funds: Análisis de su estructura, estrategias y eficiencia", Ed. Díaz de Santos, Madrid.
- López de Prado, M. and A. Peijan (2004): "Measuring Loss Potential of Hedge Fund Strategies", Journal of Alternative Investments, Vol.7 (1), pp.7-31.
- López de Prado, M. (2003): "Die Unterschätzte Risiken von Hedge Funds", Stocks, n° 29/30, pp.18-30, July.
- Lord, R. and A. Pelsser (2007): "Level-slope-curvature: Fact or artifact?", Applied Mathematical Finance, Vol. 14, No. 2, May, pp. 105-130.
- Low, A. and J. Muthuswamy (1996): "Information flows in high frequency Exchange Rates", en C. Dunis (ed.), Forecasting Financial Markets, Wiley.
- Luo, Y., R. Cahan, J. Jussa, Z. Chen and M. Álvarez (2011): "Minimum Variance: Exposing the 'magic'", Global Markets Quantitative Research, Deutsche Bank.
- Maddala, G.S. and I.M. Kim (2004): "Unit roots, cointegration and structural change", Cambridge.
- Madhavan, A. (2002): "VWAP Strategies", ITG, Working paper (Spring).

- Makov, U., A. Smith and D. Titterington (1985): "The Statistical Analysis of Finite Mixture Models". Wiley.
- Markowitz, H. (1952): "Portfolio selection", Journal of Finance, 7:77-91.
- Mascareñas, J. and L. Díez (1994): "Ingeniería Financiera. La Gestión en los mercados financieros internacionales", McGraw-Hill, Madrid.
- McWilliam, N. and K. Loh (2008): "Incorporating Multidimensional Tail-Dependencies in the Valuation of Credit Derivatives". Misys Risk Financial Engineering, working paper.
- Mertens, E. (2002): "Variance of the IID estimator in Lo (2002)", Working paper, University of Basel.
- Moulton, P. and A. Seydoux (1998): "Using Principal Components Analysis to Structure Butterfly Trades", Global Relative Value Research, Deutsche Bank.
- Newey, W. K. and K. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", Econometrica, Econometric Society, vol. 55(3), May, pp. 703-08.
- Newey, W. K. and K. West (1994): "Automatic lag selection in covariance matrix estimation", Review of Economic Studies 61, 631-653.
- O'Hara, M. (1995): "Market Microstructure Theory", Blackwell.
- O'Hara, M. (2010), "What is a quote?", Journal of Trading, Spring, p.10-15.
- Opdyke, J. (2007): "Comparing Sharpe ratios: so where are the p-values?", Journal of Asset Management 8 (5), 308–336
- Pearson, K. (1894): "Contributions to the mathematical theory of evolution". Philosophical Transactions of the Royal Society, 185, pp. 71-110.
- Phillips, P.C.B and P. Perron (1988), "Testing for a Unit Root in Time Series Regression", Biometrika, 75, 335–346
- Rebonato R. (2001). "Managing Model Risk' in Handbook of Risk Management", Ed. FT-Prentice Hall.
- Rebonato, R. (2010): "Coherent Stress Testing: A Bayesian Approach to the Analysis of Financial Stress", Ed. Wiley.
- Rebonato, R. (2004): "Volatility and Correlation: The Perfect Hedger and the Fox", Ed. Wiley Finance.
- Rebonato, R. (2011): "Plight of the Fortune Tellers: Why We Need to Manage Financial Risk Differently", Ed. Wiley Finance.
- Refenes, A., Y. Abu-Mostafa, J. Moody and A. Weigend (1996): "Neural Networks in Financial Engineering", Proceedings of the Third International Conference on Neural Networks in Capital Markets. World Scientific.

- Riobóo Almanzor, J.M. and C.P. del Oro Sáez (2000): "Representaciones gráficas de datos estadísticos", Ed. AC, Madrid.
- Roy, Arthur D. (1952): "Safety First and the Holding of Assets", Econometrica (July): 431–450.
- Said, S.E. and D. A. Dickey (1984): "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order". Biometrika 71, 599–607
- Scherer, B. (2010): "A New Look At Minimum Variance Investing", Working paper, SSRN: http://ssrn.com/abstract=1681306.
- Sharpe, W. F. (1966): "Mutual Fund Performance", Journal of Business, 39 (S1): 119–138.
- Shilov, Georgi E. (1977): "Linear Algebra", Ed. Dover Publications, 1st Edition.
- Steeley, J. M. (1990): "Modelling the dynamics of the term structure of interest rates", Economic and Social Review, 21, pp. 337–361.
- Stevens, G. (1998): "On the Inverse of the Covariance Matrix in Portfolio Analysis", Journal of Finance, Vol. 53, No. 5.
- Strang, Gilbert (1988): "Linear algrabra and its applications", Ed. Harcourth, 3rd Edition.
- Tapia, M., R. Pascual and A. Escribano (2004): "Adverse selection costs, trading activity and liquidity in the NYSE: An empirical analysis", Journal of Banking and Finance, Vol. 28, pp. 107-128.
- Tapia, M., M.A. Martínez and G. Rubio (2005): "Understanding liquidity: A closer look at the limit order book", Revista de Economía Aplicada, Vol. 38, pp. 95-109.
- Tapia, M., J. Gil-Bazo and D. Moreno (2007): "Price dynamics, informational efficiency and wealth distribution in continuous double auction markets", Computational intelligence, Vol. 23, pp. 176-196.
- Tapia, M., M. Espinosa and M. Trombetta (2008): "Disclosure and liquidity in a driven by orders market: Empirical evidence from panel data", Investigaciones Económicas, Vol. 23, pp. 339-370.
- Tashman, A. and R. Frey (2008): "Modeling risk in arbitrage strategies using finite mixtures". Quantitative Finance, 9(5), pp. 495-503.
- Tauchen, G. E. and M. Pitts (1983): "The Price Variability-Volume Relationship on Speculative Markets", Econometrica, 51, 485-505.
- Vidyamurthy, V. (2004): "Pairs trading: Quantitative methods and analysis", Wiley Finance.
- Wang, J. (2001): "Generating daily changes in market variables using a multivariate mixture of normal distributions". Proceedings of the 33rd winter conference on simulation, IEEE Computer Society, pp. 283–289.

- Xu, L. and M. Jordan (1996): "On Convergence Properties of the EM Algorithm for Gaussian Mixtures", Neural Computation (8), pp. 129-151.
- Zellner (1962): "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", Journal of the American Statistical Association, Vol. 57, No. 298 (June), pp. 348-368.

ABOUT THE AUTHOR

Marcos López de Prado is Head of Global Quantitative Research at *Tudor Investment Corp.*, where he has also led High Frequency Futures Trading and several strategic initiatives. He has been a Partner at *PEAK6 Investments*, where he was responsible for Statistical Arbitrage at the Futures division. Prior to that, he was Head of Quantitative Equity Research at *UBS Wealth Management*, and a Portfolio Manager at *Citadel Investment Group*. In addition to his 15 years of investment management experience, he has received several academic appointments, including Postdoctoral Research Fellow of *RCC at Harvard University*, Visiting Scholar at *Cornell University*, and Research Affiliate of CIFT at *Lawrence Berkeley National Laboratory* (U.S. Department of Energy's Office of Science). He holds a Ph.D. in Financial Economics (2003), a Sc.D. in Mathematical Finance (2011) from *Complutense University*, received the National Graduation Award in Economics by the Government of Spain (National Valedictorian, 1998), and was admitted into *American Mensa* with a perfect score.

Dr. López de Prado is a member of the editorial board of the Journal of Investment Strategies (Risk Journals). His research has resulted in three international patent applications, several papers listed among the most read in Finance (SSRN), publications in the Review of Financial Studies, Journal of Risk, Journal of Portfolio Management, etc. His current Erdös number is 3, with a valence of 2.

This dissertation is the author's second doctoral thesis. The first one dealt with Hedge Funds' risk management and portfolio construction, and was published in 2003.[112]

Tudor Investment Corporation has been ranked among the top-3 most profitable hedge funds in history.[113] Tudor's funds manage in excess of 12 billion dollars, of which about 1/3 is allocated to Systems Trading. Tudor's High Frequency group is comprised by leading researchers and developers coming from the fields of Physics, Mathematics, Engineering, Computer Science and Economics.

---

[112] Official doctoral record: https://educacion.gob.es/teseo/mostrarRef.do?ref=295773
[113] *"Hedge fund report card"*, Absolute Return Magazine, September 2009.

ABOUT COMPLUTENSE

Founded on May 20th, 1293 by Royal Charter of King Sancho IV of Castile, **Complutense University** (*"Universitas Complutensis"*) is one of the oldest universities in continuous operation. In the course of over 7 centuries of history, Complutense has made some of the most enduring contributions to Western civilization.

While at Complutense, Antonio de Nebrija published the first grammar of a modern language (1492), and the first dictionaries Latin-Spanish (1492) and Spanish-Latin (1495). He was also the first erudite to make claims of intellectual property.

By the year 1509 it already had five major Colleges: Medicine, Philology, Arts and Philosophy, Theology and Canon Law. One of its alumni, Cardinal Cisneros, attracted many of the world's foremost linguists and biblical scholars. The Complutensian Polyglot Bible (1514), one of the greatest academic works of the Renaissance, was the result of 15 years of interdisciplinary research. For instance, the Greek typefaces devised for this 6 volume *tractatus* constituted the template for the Greek fonts used nowadays (*Otter Greek*, *GFS Complutensian Greek*). This publication predates Erasmus' *Textus Receptus* (1516), which later became the basis for Oxford University's King James version (1611).

Other former pupils include renowned philosophers (Ortega y Gasset, Marías, de Soto), writers (Lope de Vega, Quevedo, Lorca), scientists (Ramón y Cajal, Ochoa, Cabrera, Terradas, Marañón), historians (Mariana, Menéndez y Pelayo, Menéndez Pidal), military leaders (Don John of Austria, Farnese) and international officials (Solana, Rato, Borrell, Oreja). Between 1857 and 1954, only Complutense had the authority to grant Doctor degrees in the Kingdom of Spain (Moyano act). Complutense was the first European University from which Albert Einstein accepted a Doctor of Science degree *"Honoris Causa"* (February 28th, 1923). In April 1933, Dr. Einstein also accepted a Faculty position at one of its Research Institutes.

In recent years, the roster of alumni comprises winners of the Prince of Asturias Prize (18), Miguel de Cervantes Prize (7), Nobel Prize (7), Solvay Conference, European Union Commissioners, Presidents of the EU Parliament, European Council Secretary General, ECB Executive Board members, NATO Secretary General, UNESCO Director General, IMF Managing Director, Heads of State, Prime Ministers, etc.

The **Real Colegio Complutense at Harvard University** (RCC) was founded as a joint cooperative institution to foster intellectual and scientific interaction between Harvard University and Complutense. It follows the tradition of the Royal Spanish College, founded in 1364 (and in operation ever since) to host Spanish Visiting Scholars at the University of Bologna. The RCC accord is the only one of its sort ever to have been approved by Harvard. The institution is directed jointly by the President of Harvard and the Rector of Complutense, with an academic council formed by 5 Harvard professors and 5 Complutense professors.

For additional information, please visit:
www.ucm.es
www.realcolegiocomplutense.harvard.edu