Using High Frequency Stock Market Index Data to Calculate, Model & Forecast Realized Return Variance

Roel C.A. Oomen*

This Version: April 2001. (First Version January 2000)

Abstract

The objective of this paper is to calculate, model, and forecast realized return variance, using high frequency stock market index data. The approach taken differs from the existing literature in several aspects. First, it is shown that the decay of the serial dependence of high frequency returns with the sampling frequency, is consistent with an ARMA process under temporal aggregation. This finding has important implications for the modelling of high frequency returns and the optimal choice of sampling frequency when calculating realized variance. Second, motivated by the outcome of several test statistics for long memory in realized variance, it is found that the realized variance series can be modelled as an ARFIMA process. Significant exogenous regressors include lagged returns and contemporaneous trading volume. Finally, the ARFIMA's forecasting performance is assessed in a simulation study. Although it outperforms representative GARCH models, the simplicity and flexibility of the GARCH may outweigh the modest gain in forecasting performance of the more complex and data intensive ARFIMA model.

Keywords: High Frequency Data; Realized Return Variance; Market Microstructure; Temporal Aggregation; Fractional Integration; GARCH.

JEL Codes: C51, C52, C53, G12, G13

---

*European University Institute, Department of Economics Via dei Roccettini 9, 50016 San Domenico di Fiesole, Italy. Email: roel.oomen@iue.it. This paper has greatly benefitted from discussions with Frank Diebold, Marcelo Fernandes, George Jiang, Mark Lauer, Bent Sørensen and Juan Toro. I am particularly grateful for many detailed comments and suggestions of Søren Johansen. I thank *Logical Information Machines, Inc.* who kindly provided the data needed for the analysis. The research was partly supported by the Netherlands Organization for Scientific Research (NWO) grants SIR 12-4010 and R-46-447. Errors are mine.

1 Introduction

A crucial element in the theory and practice of derivative pricing and financial risk management is the estimation and modelling of asset return variance. Both the Stochastic Volatility and the Autoregressive Conditional Heteroskedasticity (ARCH) class of models have become widely established and successful approaches, in both the theoretical and the empirical literature, to the modelling of the variance process. The objective of this paper, however, is to explore the extent to which the use of the increasingly available intra-daily data on financial assets can be used to improve or facilitate the estimation and modelling of return variance. To this end Merton (1980) notes that the variance (over a fixed interval) of an iid random variable can be estimated arbitrarily accurate by the sum of squared realizations, provided that the data is available at a sufficiently high sampling frequency. Empirical studies, such as French, Schwert and Stambaugh (1987), Hsieh (1991), Taylor and Xu (1997), Andersen and Bollerslev (1998), make use of this insight and employ intra-daily return data to estimate daily return variance by simply summing up squared intra-daily returns. In the literature this variance measure is referred to as "realized variance" or more commonly "realized volatility" (variance being volatility squared). In a recent paper Andersen et al. (2001a) has shown in a continuous time setting that when the return process follows a special semi-martingale, the sum of squared returns will yield a consistent estimate for the integrated variance of the return process:

> "*The mechanics are simple - we compute daily realized volatility simply by summing up squared returns - but the theory is deep: by sampling intra-day returns sufficiently frequently, the realized volatility can be made arbitrarily close to the underlying integrated volatility, the integral of instantaneous volatility over the interval of interest, which is a natural volatility measure" - (Andersen et al. (2001a))*

Although the work by Merton (1980) and Andersen et al. (2001a) is taken as a starting point for the calculation and analysis of the realized variance measure, the present study is distinguished from the existing literature in several ways. First and foremost, the choice of sampling frequency and the impact that market micro structure induced autocorrelations have on realized variance are discussed in considerable detail; issues to which has been paid surprisingly little attention so far (a notable exception is Bai, Russell, and Tiao (2000)). The serial dependence of high frequency returns is analyzed and it is found that the autocorrelation structure (magnitude and rate of decay) of returns at different sampling frequencies is consistent with the results on temporal aggregation of an ARMA process. This finding has important implications for the choice of optimal sampling frequency when calculating the realized variance measure. 10 years of minute by minute *FTSE-100* index returns are employed to illustrate that when the sampling frequency is not carefully chosen, solely summing up squared returns can seriously under- or overestimate the average daily return variance. Second, a realized variance series for the *FTSE-100* index returns is constructed using 25 minute return data which is modelled subsequently. It is found that an ARFIMA specification, including some exogenous variables such as lagged returns and trading volume, models the series well. This extends the work of Andersen and Bollerslev (1998) which uses the realized variance estimates for evaluation of the forecasting performance of their GARCH model or that of Andersen et al. (2000a,b) which analyzes the properties of the data. The regression coefficients of the lagged return variable are used to test for the presence of Black's leverage effect. Third and finally, the forecasting performance of the ARFIMA model for realized variance is assessed in a simulation study. The results indicate that the ARFIMA model for realized variance outperforms representative GARCH-class models. It is noted, however, that the simplicity of the GARCH together with its flexibility to account for persistence in return variance, may outweigh the modest gain in forecasting performance of the more complex and data intensive ARFIMA model.

The remainder of this paper is structured as follows. Section 2 discusses the calculation of realized variance and proposes a model for daily as well as intra-daily returns. The realized variance series is constructed and a careful analysis shows that most of the stylized facts, as documented in the recent literature, can be confirmed for the *FTSE-100* index data. Section 3 models realized variance as an ARFIMA process. Section 4 compares the forecasting performance of the ARFIMA model for realized variance with conventional GARCH type models. Section 5 concludes.

2 Realized Variance

The term "realized variance" refers to the sum of squared intra-period returns, being an estimator for the average or integral of instantaneous variance over the interval of interest. In fact, in a continuous time framework, it has been shown by Andersen et al. (2001a) that when the return process is assumed to follow a special semi-martingale the realized variance measure can be made arbitrarily close to the integral of instantaneous variance, provided that the intra-period returns are sampled at a sufficiently high frequency. In the present context, however, the focus will be on a discrete time model which has the advantage that the impact of market micro structure effects, present in returns sampled at high frequency, on the realized variance measure can be analyzed in a straightforward fashion. It will be shown that when intra-period returns are serially correlated, the realized variance measure will yield a biased estimator of the average true variance over the interval of interest. Throughout the remainder of the paper the interval of interest is set to one trading day.

Let $S_{t,j}$ denote the j^{th} intra day$-t$ price of the security under consideration and $\Psi_{t,j}$ be the sigma field generated by $\{S_{a,b}\}_{a=-\infty,b=0}^{a=t,b=j}$. Under the assumption of N *equally time-spaced* intra-daily observations of S $(j=1,\ldots,N)$, the daily return is defined as:

$$R_t = \ln S_{t,N} - \ln S_{t-1,N},$$

$t=1,\ldots,T$. At sampling frequency f, we can construct $N_f = \frac{N}{f}$ intra-daily returns:

$$R_{f,t,i} = \ln S_{t,if} - \ln S_{t,(i-1)f},$$

for $i=1,\ldots,N_f$ and $S_{t,0} = S_{t-1,N}$. In the following, it is assumed that the asset's *(excess) return at the daily frequency* can be characterized as:

$$R_t = \sigma_t \varepsilon_t,$$

where $\varepsilon_t \sim$iid $\mathcal{N}(0,1)$ and σ_t^2 represents the day$-t$ return variance. Note that $E_{\Psi_{t,0}}\left[R_t^2\right] = \sigma_t^2$ and that $V_{\Psi_{t,0}}\left[R_t^2\right] = 2\sigma_t^4$. Now consider the situation in which intra-daily returns, at sampling frequency f, are uncorrelated and can be characterized as:

$$R_{f,t,i} = \sigma_{f,t,i}\varepsilon_{f,t,i},$$

where $\varepsilon_{f,t,i} \sim$iid $\mathcal{N}\left(0, N_f^{-1}\right)$ and $R_t = \sum_{i=1}^{N_f} R_{f,t,i}$ by definition. Since intra-daily returns are assumed to be uncorrelated, it directly follows that $E_{\Psi_{t,0}}\left[\sum_{i=1}^{N_f} R_{f,t,i}^2\right] = E_{\Psi_{t,0}}\left[R_t^2\right]$ and hence $\sigma_t^2 = N_f^{-1}\sum_{i=1}^{N_f}\sigma_{f,t,i}^2$. As a results, two unbiased estimators for the average day$-t$ return variance exist, namely the squared day$-t$ return and the sum of squared intra day$-t$ returns. It is noted, however, that while $V_{\Psi_{t,0}}\left[R_t^2\right] = 2\sigma_t^4$ the following holds:

$$E_{\Psi_{t,0}}\left[\left(\sum_{i=1}^{N_f}\sigma_{f,t,i}^2\varepsilon_{f,t,i}^2\right)^2\right] = 3N_f^{-2}\sum_{i=1}^{N_f}\sigma_{f,t,i}^4 + 2N_f^{-2}\sum_{i=1}^{N_f-1}\sum_{j=i+1}^{N_f}\sigma_{f,t,i}^2\sigma_{f,t,j}^2,$$

and therefore:

$$V_{\Psi_{t,0}}\left[\sum_{i=1}^{N_f} R_{f,t,i}^2\right] = \frac{2}{N_f}\sum_{i=1}^{N_f}\frac{\sigma_{f,t,i}^4}{N_f} < \frac{2}{N_f}\left[\sum_{i=1}^{N_f}\frac{\sigma_{f,t,i}^2}{\sqrt{N_f}}\right]^2 = V_{\Psi_{t,0}}\left[R_t^2\right].$$

In words, the average daily return variance can be estimated more accurately by summing up squared intra-daily returns rather than calculating the squared daily return. In addition, when returns are observed (and uncorrelated) at any arbitrary sampling frequency, it is possible to estimate the average daily variance free of measurement error as $\lim_{N_f\to\infty} V_{\Psi_{t,0}}\left[\sum_{i=1}^{N_f} R_{f,t,i}^2\right] = 0$. The only (weak) requirement on the dynamics of the intra-daily return variance for the above to hold is that $\sum_{i=1}^{N_f}\sigma_{f,t,i}^4 \propto N_f^{1+c}$ where $0 \leq c < 1$. Finally, note that although the daily realized variance measure employs intra-daily return data, there is no need to take the (well documented) pronounced intra-day variance pattern of the return process into account. This feature of the realized variance measure contrasts sharply with popular parametric variance models which generally require the explicit modelling on intra-daily regularities in return variance.

The focus in the remainder of this section will be on how the increasingly available high frequency financial data can be used for the purpose of variance estimation. In particular, minute by minute *FTSE-100* index level data[1] will be used to investigate whether the method of calculating the realized variance measure, being the sum of squared intra-daily returns, will yield satisfactory results. To this end, the decomposition of the daily return into the sum of N_f intra-daily returns can be used to derive the following expression:

$$R_t^2 = \left[\sum_{i=1}^{N_f} R_{f,t,i}\right]^2 = \sum_{i=1}^{N_f} R_{f,t,i}^2 + 2\sum_{i=1}^{N_f-1}\sum_{j=i+1}^{N_f} R_{f,t,i}R_{f,t,j}. \tag{1}$$

When the assumption of uncorrelated returns at sampling frequency f is satisfied, the second term on the right hand side of expression (1) is zero in expectation and the realized variance measure will therefore yield an unbiased estimate of the average day$-t$ return variance. However, as noted by, for example, French et al. (1987), when the returns are positively correlated, solely summing up squared returns will underestimate the average daily variance;

[1]The dataset contains minute by minute data on the *FTSE-100* index level, starting May 1, 1990 and ending January 11, 2000. Trading hours are 08:30-16:30, Monday to Friday until September 14, 1994 and 8:30-16:00 afterwards (minute data from 08:35 until 16:10 is available). The total number of observations is just over 1.1 million.

the cross multiplication of returns will be positive on average. The reverse will occur for negatively autocorrelated returns.

To illustrate this, a dataset containing minute by minute data on the level of the *FTSE-100* stock market index is employed to calculate the 10 year average (1990-2000) of the two terms on the right hand side of expression (1) for sampling frequencies between 1 and 45 minutes. The results are displayed in Figure 1.1. It is clear that the first term, the realized variance measure, increases with a decrease in sampling frequency while the second term, the summation of cross multiplied returns, decreases. The positivity of the second term indicates that the *FTSE-100* returns are positively correlated, introducing a *downward bias* into the realized variance measure (up to 35% when using minute by minute data!), while its decreasing pattern demonstrates that this dependence, and consequently the bias, diminishes when sampling is done less frequent. This term can therefore be interpreted as the autocovariance induced bias of the realized variance measure and will be referred to as the "*autocovariance bias factor*" in the remainder of the paper.

Although, in the context of efficient markets, the finding of correlated intra-daily returns may at first sight appear puzzling, it has a sensible explanation in the context of the market microstructure literature[2]. One of the most prominent hypotheses which can be used to explain the observed positive autocorrelation in stock index returns is non-synchronous trading. The basic idea is that when individual stocks contained in an index do not trade simultaneously, the contemporaneous positive autocorrelation among the components will induce serial correlation in the index returns. Intuitively, when the index components incorporate non-synchronously the shocks to a common factor driving their price, this will result in a sequence of correlated price changes at the aggregated or index price level. This phenomenon, which is consistent with the above empirical findings[3], obviously disappears when sampling frequency decreases.

It has been shown that average daily return variance can be estimated consistently by the realized variance measure, provided that the intra-daily returns are serially uncorrelated. When the intra-daily returns are correlated, realized variance will either overestimate (with negative correlation) or underestimate (positive correlation) the average daily return variance.

[2]See e.g. Campbell, Lo and MacKinlay (1998), Lequeux (1999), Madhavan (2000) or Wood (2000).

[3]The reverse would occur for a single asset. The serial correlations of returns, if present, would likely be negative thereby introducing an upward bias in the realized variance measure. The negative autocorrelation can be attributed to the bid-ask bounce; in a market where no new information arrives, the stock price is expected to bounce between the bid and the ask price whenever a trade occurs.

Correcting for the bias term, it is known after all, is not desirable as this is equivalent to using the squared daily return to estimate daily realized variance. Hence, when the intra-daily return data at the highest frequency available is serially correlated, one will need to aggregate the returns down to a frequency at which the correlation has disappeared. Plotting both the sum of squared intra-daily returns and the autocovariance bias factor versus the sampling frequency, as is done in Figure 1.1, proves a very helpful and easily implementable strategy to determine the frequency[4] at which the correlation has died off. The "optimal" sampling frequency is chosen as the highest available frequency for which the autocovariance bias term has disappeared. Based on these observations, the sampling frequency for estimating realized variance will be set to 25 minutes ($f = 25$).

2.1 Serial Correlation, Time Aggregation & Sampling Frequency

Prior to the estimation and analysis of realized variance, a closer look is taken at the autocovariance bias term in relation to the dynamic properties of the intra-daily returns at different sampling frequencies. Table I (and Figure 1.1) reports some descriptive statistics for the *FTSE-100* return data at different sampling frequencies. Both the order and the magnitude of the autocorrelations decrease with a decrease in sampling frequency. The Box-Ljung statistic comes down from around 800 for minute data, to around 20 for daily data (the 95% critical value of this test is 18.31). The Durbin Watson test statistic increases from about 1.6 for the minute data to 2.0 for the daily data. Finally, it is noted that the first 20 autocorrelations calculated for the minute by minute data appear significant while only the first order autocorrelation of daily data is significant. These findings suggest that a realistic statistical model for intra-daily returns should have a more flexible structure than the standard model for daily returns. In the remainder of this section it will be shown that modelling intra-daily returns as an ARMA process is a natural and, as it turns out, successful choice for it is well suited to account for the serial dependence of returns at various sampling frequencies. From a market micro structure point of view, the *AR* part will arguably be able to capture any autocorrelation induced by non-synchronous trading while the *MA* part will account for potential negative first order autocorrelation induced by the bid ask bounce. Moreover, within this framework it can be shown that the decreasing order and magnitude of autocorrelations

[4] Independent and concurrent work of Andersen, Bollerslev, Diebold and Labys (2000), Corsi, Zumbach, Müller and Dacorogna (2001) have proposed a closely related approach for determining the optimal sampling frequency.

with the sampling frequency is a consequence of temporal aggregation of the return process.

Suppose that returns at the highest sampling frequency, R_1 (the t subscript is momentarily dropped for convenience), can be described as an ARMA(p,q) process:

$$\alpha(L) R_{1,i} = \beta(L) \varepsilon_{1,i},$$

where $\alpha(L)$ and $\beta(L)$ are lag polynomials of lengths p and q respectively. Consider the case where all the reciprocals of the roots of $\alpha(L) = 0$, denoted by $\theta_1, ..., \theta_p$, lie inside the unit circle. The model through which the returns at an arbitrary sampling (or aggregation) frequency can be represented is derived using the results of Wei (1981) on temporal aggregation[5] (see appendix for a summary). In particular, when R_1 follows an ARMA(p,q) process as given above, the returns sampled at frequency f, denoted by R_f, can be represented by an ARMA(p,r) process:

$$\prod_{j=1}^{p}\left(1-\theta_j^f L^f\right) R_{f,i} = \prod_{j=1}^{p} \frac{1-\theta_j^f L^f}{1-\theta_j L}\frac{1-L^f}{1-L}\beta(L)\varepsilon_{f,i},$$

where r equals the integer part of $p + \frac{q-p}{f}$, $\varepsilon_{f,i} = \sum_{j=0}^{f-1} \varepsilon_{1,fi-j}$. Due to the invertibility of the AR polynomial, the above model can be written as an MA(∞) process with parameters $\{\psi_j\}_{j=0}^{\infty}$ and $\psi_0 = 1$. Let φ_h^f denote the h^{th} autocovariance of the temporally aggregated returns at frequency f, for which it turns out that:

$$\varphi_h^f = E\left[R_{f,i}R_{f,i-h}\right] \propto \sum_{j=0}^{\infty}\left[\left(\sum_{i=\max(0,j-f+1)}^{j}\psi_i\right)\left(\sum_{i=j+1+f(h-1)}^{j+fh}\psi_i\right)\right], \tag{2}$$

As the ψ_j coefficients decay exponentially fast with j, the serial correlation disappears under temporal aggregation. To see this, let $\psi_j = w\delta^j$ for $|\delta| < 1$ and w some positive constant. It can now be shown that:

$$\varphi_h^f \propto \sum_{j=0}^{\infty}\left[\sum_{i=0}^{j} w\delta^i \sum_{i=j+f(h-1)}^{j+fh} w\delta^i\right] \leq \frac{w^2}{(1-\delta)^3}\delta^{f(h-1)}$$

from which it can be seen that the serial correlation disappears when either the sampling frequency, f, or the displacement, h, increases. In fact, Wei (1981) has shown that the limit

[5] Temporal aggregation for ARMA models is discussed in Brewer (1973), Tiao (1972), Wei (1981), Weiss (1984) and the VARFIMA in Marcellino (1999).

model of an ARMA(p,q) process under temporal aggregation is an ARMA(0,0) or equivalently white noise.

Note that these theoretical properties of the ARMA process appear very much in accordance with the reported empirical properties of the return process at different sampling frequencies. More specifically, at high sampling frequencies the ARMA model can account for the observed serial dependence while at lower sampling frequencies these dependencies die off as a consequence of temporal aggregation of the return process. In addition, as the limit model of the ARMA(p,q) model is an ARMA(0,0) under temporal aggregation, the model specification for returns at the intra-daily frequency does not necessarily conflict with the model for daily returns.

The above expression for the autocovariance function of the ARMA process can be used to check the consistency of the model with the properties of the data by comparing the temporal aggregation implied decay of the autocovariance bias term with the empirically observed one. To this end various ARMA models are estimated using the minute by minute returns and it is found that an ARMA(6,0) model yields satisfactory results[6] with uncorrelated residuals and relatively stable coefficients over time. Using solely one set of ARMA(6,0) parameters for the minute data, autocovariances for the estimated return process at various sampling frequencies are "*implied*" using expression (2). It is noted that:

$$E_{\Psi_{t,0}}\left[\sum_{i=1}^{N_f-1}\sum_{j=i+1}^{N_f} R_{f,t,i}R_{f,t,j}\right] = \sum_{h=1}^{N_f-1}\left(N_f-h\right)\varphi_h^f, \tag{3}$$

Hence, the "aggregation implied" autocovariance estimates can be used to calculate the "aggregation implied" autocovariance bias term as in expression (3). Figure 1.2 in the appendix demonstrates that the empirical and theoretically implied curves are remarkably close. The implications of this finding are twofold. First, it shows that the ARMA model is a good description of the return data sampled at different frequencies; the decay of the (market - microstructure - induced) serial dependencies in high frequency returns is consistent with the decay of an ARMA process under temporal aggregation. Second, relying on the close correspondence between the empirical and theoretically implied autocovariance bias factor one can locate the optimal frequency, that is, the highest sampling frequency available for which the

[6] Although the residuals are highly leptokurtic and heteroskedastic, and hence the MLE is not efficient, the parameter estimates are consistent (see e.g. Amemiya (1985)). Moreover, the efficiency loss should be unimportant given the large amount of data.

autocovariance bias term has died off, using solely a *single* set of ARMA parameter estimates.

2.2 Stylized Facts of Realized Variance

High frequency data have already been analyzed extensively by a number of studies[7]. The results regarding the characteristics of high frequency financial data obtained so far (relevant for this study) can be summarized as follows. First, the unconditional distribution of daily returns is not skewed, but it does exhibit excess kurtosis. Daily returns are not autocorrelated (except for the first order in some cases). Second, the unconditional distributions of realized variance (calculated as the sum of squared intra-daily returns sampled at frequencies between 5 and 20 minutes depending on the dataset used) and variance are distinctly non-normal and extremely right skewed, whereas the natural logarithm of the standard deviation is close to Gaussian. Third, the log of the realized variance displays a high degree of (positive) autocorrelation which dies out very slowly. Fourth, realized variance does not seem to have a unit root, but there is clear evidence of fractional integration[8], roughly of order 0.40. Fifth[9], daily returns standardized by the realized variance measure are (nearly) Gaussian.

In order to complement and widen the focus of the research in this area European stock market index data (minute by minute data on the *FTSE-100* from May 1990 until January 2000) is utilized. As mentioned above, daily realized variance on the *FTSE-100* is estimated using the high frequency data sampled at $f = 25$. This results in a total of 2445 realized variance estimates which are reported in Figure 2.1. In the appendix, some descriptive statistics of realized variance, the log of realized variance, the daily return, and daily return standardized by realized variance are reported in Table II. The stylized facts are essentially confirmed for the dataset under study. More specifically, the distribution of daily returns has fat tails but is not very skewed and variance clustering is clearly present in the return series (not reported). The strong evidence found regarding the normality of the daily returns standardized by realized variance indicates that similar results of Andersen et al. (2001b) on exchange rate data can be extended to stock market index data. The unconditional distribution of the realized

[7] See e.g. Anderson, Bollerslev, Diebold and Labys (2000a,b), Froot & Perold (1995), Goodhart & O'Hara (1997), Hsieh (1991), Lequeux (1999), Stoll & Whaley (1990), Zhou (1996).

[8] See e.g. Baillie (1996), Bollerslev and Mikkelsen (1996, 1999), Breidt et. al. (1998), Comte and Renault (1998), Liu (2000), Lo (1991).

[9] In a multivariate setting it is found that the distribution of correlations between realized variance is close to normal with positive mean, and that the autocorrelations of realized correlation decays extremely slow.

variance is significantly skewed and exhibits severe kurtosis, while the unconditional distribution of log realized variance is much less skewed and displays significantly reduced kurtosis. Furthermore, the correlogram for the realized variance measure decays only very slowly but the Augmented Dickey Fuller test[10] strongly rejects the null hypothesis of a unit root (Table II, Figure 2.3). This last observation is usually indicative for fractional integration.

3 Modelling Realized Variance

Having calculated and analyzed the realized variance measure, the focus is on finding a statistical model that captures the main characteristics of this time series. Some observations from the previous section have to be taken into account when deciding on the modelling strategy. First of all, the benefits of the log transformation of realized variance indicate that log realized variance should be modelled instead of the original series. Second, the absence of a unit root and the highly persistent autocorrelation point into the direction of fractional integration (see appendix for a summary on the concept of fractional integration). This section further explores the characteristics of the data and finds that the log of realized variance series can be modelled well using a fractionally integrated ARMA model. By means of an application, the existence of the Black leverage effect is tested for within the ARFIMA framework. The present section concludes with a discussion of the impact that structural breaks have on the empirical findings.

3.1 Fractional Integration & Realized Variance

Driven by the remarkable resemblance between the correlogram displayed in Figure 2.3 in the appendix and the theoretically implied correlogram for fractionally integrated process, the focus is on fractionally integrated models. Prior to estimation, some informal tests are performed to strengthen this motivation.

The fractional difference operator is applied to the log of the realized variance series for various values of d (the sequence given by expression (5) in the appendix is truncated at $h = 1000$). It is found that autocorrelations are drastically reduced for values of d between

[10] Augmented Dickey Fuller test: $\triangle x_t = \alpha + \beta x_{t-1} + \sum_{i=1}^{n} \gamma_i \triangle x_{t-i} + \varepsilon_t$. Rejection of $H_0 : \beta = 0$, implies that x_t is $I(0)$. The specification of the lag length, which we set equal to 5, assumes that ε_t is white noise. The critical value of this test equals -2.865 at 5% confidence level and -3.439 at 1%.

0.25 and 0.45 and that they fall nicely in between the 95% confidence bounds ($\pm 2N^{-1/2}$ where N denotes the number of observations). For instance, see Figure 2.4 in the appendix where d is set equal to 0.40. Moreover, the long 'waves' in log realized variance disappear almost entirely after fractional differencing (Figure 2.1 versus 2.2). A supplementary test consists of plotting log autocorrelations versus the log of displacements. It is known that for a fractionally integrated process the autocorrelation function decays at a hyperbolic rate as opposed to the autocorrelation function of an $I(0)$ process which decays at an exponential rate. Therefore the log of the autocorrelation function will yield a linear relationship in terms of log displacement, i.e. $\log \varphi_h \propto (2d-1) \log h$. Figure 2.5 in the appendix shows that plotting $\log \varphi_h$ against $\log h$ yields a linear relationship up to approximately $h = 100$. An OLS regression is performed to determine the slope. Using the complete sample ($h = 250$) implies that $d \approx 0.37$. Ignoring the last 150 autocorrelations implies that $d \approx 0.43$. Although the graph does not strongly indicate a linear relationship, the results are not taken as a rejection of fractional integration. The final check on the presence and degree of fractional integration is done in the frequency domain. Two standard tests are employed. The first one has been developed by Geweke and Porter-Hudak (1983, GPH hereafter), while the second one by Robinson (1995). A short summary of both the estimators can be found in the appendix. To implement both methods a bandwidth parameter m, controlling the range of periodic frequencies used, has to be set. Although it is required that m grow at a slower rate than the sample size T, this does not guide us as to what the value of m should be. In the present study, d is estimated for a range of m between[11] 25 and 275. The results of this estimation are summarized in Figure 2.6 where the GPH estimates together with the Robinson estimates are plotted as a function of m. For small m, the two alternative estimates both fall into the non-stationary region while for large m (above 150) they are both below 0.5. Although it is clear from this that the value for d will be close to 0.5, it is difficult to judge on the stationarity of the process as the choice of m is relatively arbitrary. In summary, the reported test results provide good reasons to believe that the (log) realized variance series is fractionally integrated. The results are ambiguous as to what the value for d will be, although it seems clear that it will most likely be close to 0.50.

[11] The sample size is 2445 and hence the range of m is between $T^{0.40}$and $T^{0.70}$. This is in line with e.g. Bollerslev, Cai and Song (2000) who sets $m = T^{0.50}$ or Dittmann and Granger (2000) who set $m = T^{0.8}$.

3.2 Empirical Results

Motivated by the preliminary tests discussed above, the focus of the modelling approach will center around the ARFIMA specification. Consider the following model:

$$\alpha(L)(1-L)^d\left[\ln\widetilde{\sigma}^2_{25,t}-\pi' X_t\right]=\beta(L)\varepsilon_t,$$

where $\widetilde{\sigma}^2_{25,t}$ denotes the day$-t$ realized variance measure calculated using 25 minute intra-daily returns, X_t is a vector of exogenous variables, $\alpha(L)$ is a lag polynomial of order p, $\beta(L)$ a lag polynomial of order q and ε_t is a residual term. Under the assumptions that the roots of $\alpha(L)$ and $\beta(L)$ are outside the unit circle, roots of $\alpha(L)$ are simple, residuals are iid Normal and $d<\frac{1}{2}$, the ARFIMA model parameters are estimated in Ox[12] using the maximum likelihood estimator of Sowell (1992). The model could alternatively be estimated with the popular and easily implementable two-step procedure in which the fractional parameter is estimated in the first step (by e.g. the GPH or Robinson estimator), while the remaining ARMA coefficients are estimated in the second step on the fractionally differenced data by ordinary least squares. As it has been found that the ARMA coefficients are generally estimated inaccurate or biased this way (see e.g. Smith et al. (1997)), the Sowell procedure is preferred as it allows for the simultaneous estimation of the model parameters.

In order to address the concern that the long memory may be induced by infrequent structural breaks[13] (see e.g. Granger (1999), Diebold and Inoue (1999) and Granger and Hyung (1999)) the above model is estimated on various subsamples. Table 3 in the appendix, reports the estimation results for the ARFIMA model (where $p=q=1$) for two different samples, namely Sample I which runs from May 1, 1990 until June 15, 1997 (1800 observations) and Sample II which is the full sample (2444 observations). As the fractional parameter remains highly significant for the different subsamples considered, it is argued that the realized variance series clearly exhibit a long memory feature that is not caused by structural breaks. For both samples, the fractional parameter d is indeed between 0.40 and 0.50, confirming the preliminary analysis above. In fact, based on the $t-$statistic one cannot reject that $d>0.5$

[12] See Doornik and Ooms (1998) for documentation on the Arfima package.

[13] A simple and representative model that can cause long memory is the stochastic break model which takes the following form: $y_t=u_t+\varepsilon_t$, where $u_t=u_{t-1}+q_{t-1}\eta_t, \varepsilon_t\sim iid\ \mathcal{N}(0,\sigma^2_y)$, $\eta_t\sim iid\ \mathcal{N}(0,\sigma^2_u)$ and q_t equals 0 with probability p and 1 with probability $1-p$. Diebold and Inoue (1999) note that in order to achieve a slowly declining autocorrelation function, whatever the model may be, the key idea is to let p decrease with the sample size so that regardless of the sample size, realizations of the process tend to have just a few breaks.

at a 95% confidence level: the realized variance series may be non-stationary. Regarding the inclusion of exogenous variables[14], lagged returns and contemporaneous trading volume are considered. This choice of variables is motivated by Black (1976) and Lamoureux and Lastrapes (1990a) respectively.

Black's leverage states that negative returns have a larger impact on future variance than do positive returns. One possible rationale for this is that when equity value decreases the debt-to-equity ratio increases, thereby raising the riskiness of the firm as manifested by an increase in future variance. Modifying the above model as follows:

$$\pi' X_t = \zeta_1^+ R_{t-1}^+ + \zeta_1^- R_{t-1}^- + ... + \zeta_m^+ R_{t-m}^+ + \zeta_m^- R_{t-m}^-,$$

where R^+ (R^-) is a vector containing the positive (negative) daily returns and zero otherwise, allows an assessment of the explanatory power of lagged returns while the relative magnitude of the ζ coefficients will indicate whether a leverage effect is present. From the estimation results reported in the appendix (Table III) it can be observed that the log likelihood value drastically increases when lagged returns are added and that the AIC information criteria drops from about 0.8 to 0.5. Moreover, coefficients on negative returns are consistently above coefficients on positive returns. Although this observation is indicative for the presence of the leverage effect, its significance is tested for by reformulating the model as follows:

$$\zeta_1 R_{t-1} + \overline{\zeta}_1 \left| R_{t-1} \right| + ... + \zeta_h R_{t-h} + \overline{\zeta}_h \left| R_{t-h} \right| .$$

Note that $\zeta_i = \frac{1}{2}(\zeta_i^+ - \zeta_i^-)$ and $\overline{\zeta}_i = \frac{1}{2}(\zeta_i^+ + \zeta_i^-)$. Where ζ_h is significantly negative[15], it can be concluded that the leverage effect is significant at horizon h. Estimation results[16] indicate that the leverage effect, as proposed by Black (1976), is present at the first three horizons and is insignificant afterwards. Note that although the statistical significance of this finding does not necessarily imply an economic significance, it does provide support for the GJR-GARCH and EGARCH specifications which explicitly account for the asymmetric effect that returns have on future variance.

[14]Glosten, Jagannathan and Runkle (1993) find that the short term interest rate has a significant positive effect on stock market volatility. The 1 month UK Interbank rate is added, but for the dataset under study it does not appear to be a significant regressor. The data frequency (daily) in the present analysis may well be too high for the interest rate to have a significant influence.

[15]Note that $r = r^+ - r^-$ and $|r| = r^+ + r^-$. Therefore, $\zeta_i^+ = \zeta_i + \overline{\zeta}_i$ and $\zeta_i^- = \overline{\zeta}_i - \zeta_i$. For leverage to be present it is required that $\zeta_i^- - \zeta_i^+ = -2 \cdot \zeta_i > 0$ or $\zeta_i < 0$.

[16]The estimates for ζ_1, ζ_2 and ζ_3 are -4.60 (4.21), -5.58 (5.13) and -3.81 (3.49) respectively. t-values are in parenthesis.

The inclusion of (log) contemporaneous trading volume has been suggested by Lamoureux and Lastrapes (1990a). They discuss a model in which heteroskedasticity results from time dependence in the process which governs the information flow to the market. Taking trading volume as a proxy for the information arrival rate they argue (and show empirically) that (i) trading volume is positively related to return variance and (ii) the persistence of return variance decreases (or disappears) when trading volume is accounted for. Based on the estimation results reported in the appendix, where λ_0 denotes the regression coefficient of log contemporaneous trading volume, it is found that trading volume further improves the fit of the model and, consistent with the theory, has a significant positive effect on variance. The fractional parameter d does, however, not decrease upon inclusion of trading volume indicating that the persistence of the variance process remains unchanged upon conditioning on trading volume.

4 Forecasting Realized Variance

Ultimately the goal of analyzing and modelling variance is to use the resulting model for return variance forecasting. Obviously, this is of great interest to the fields of risk management and derivative pricing. This section assesses the forecasting ability of the ARFIMA model for daily log (realized) variance and takes a standard GARCH(1,1) model, implemented with daily returns, as a benchmark. As the dataset in this study does not allow us to study the forecasting performance of the models in depth, a simulation study is undertaken.

4.1 Simulation Design

The approach taken is as follows. In every simulation run a time series of 2750 daily log (realized) return variances is generated according to an ARFIMA(0,d,0) process:

$$(1-L)^d \left[\ln \sigma_t^2 - \mu\right] = \varepsilon_t,$$

where $\varepsilon_t \sim$i.i.d.$\mathcal{N}\left(0, \triangle_\varepsilon^2\right)$. The rationale for excluding the autoregressive and moving average terms, which govern the short run dynamics of the process, is that including or excluding them is unlikely to change the results qualitatively. Motivated by the empirical results of the last section the mean $\mu = 10$, the fractional parameter $d = 0.45$ and the residual variance $\triangle_\varepsilon^2 = 0.15$. To obtain forecasts of $\ln \sigma^2$, note that $E_{T-1}\left[(1-L)^d \left(\ln \sigma_T^2 - \mu\right)\right] = 0$ or equivalently $E_{T-1}\left[\ln \sigma_T^2 - \mu\right] = -\sum_{h=1}^{\infty} \frac{\Gamma(h-d)}{\Gamma(h+1)\Gamma(-d)} \left(\ln \sigma_{T-h}^2 - \mu\right)$. This autoregressive representation of

$\ln \sigma_T^2$ can be used to recursively compute the $s-$step ahead forecasts (where the infinite AR polynomial is truncated at $h = 2500$ lags for practical implementation). The $s-$step ahead forecast of log realized variance will have a forecast error $\bar{\epsilon}_{t,s} = E_t\left[\ln \sigma_{t+s}^2\right] - \ln \sigma_{t+s}^2$, that is normally distributed with mean 0 and variance $\triangle_s^2$. As a consequence the forecast error for realized variance[17], $\epsilon_{t,s} = E_t\left[\exp(\ln \sigma_{t+s}^2)\right] - \sigma_{t+s}^2$, will have a log normal distribution with a mean equal to $\exp\left(\frac{1}{2}\triangle_s^2\right)$. Given that the focus is on the realized variance forecasts (and not the log of it), this "bias" is corrected for by using an estimate of $\triangle_s^2$.

In every simulation run the fractional parameter of the ARFIMA model is estimated using the GPH estimator. A density of the estimates is given in Figure 3.4. To assess the forecasting performance of the ARFIMA model for realized variance the GARCH(1,1) model, motivated by its widespread usage in variance modelling and forecasting, is taken as a benchmark. Daily return data, which are needed for the implementation of the GARCH model, are constructed by multiplying the simulated realized variance series with an iid standard normal random variable. To avoid re-estimation of the GARCH parameters at every simulation run α_1 and β_1 are fixed while the intercept, ω, is set such that the unconditional mean of the GARCH process, $\frac{\omega}{1-\alpha_1-\beta_1}$, equals the estimated unconditional variance of the return process. Regarding the choice of α_1 and β_1, two cases are considered. In the first model specification, which is referred to as "GARCH1", $\alpha_1 = 0.90$ and $\beta_1 = 0.05$. In the "GARCH2" specification the persistence of the variance process is increased by increasing α_1 to 0.94 while leaving β_1 unchanged. The $s-$step ahead forecast of daily realized variance generated by the GARCH model, $\widehat{\sigma}_{T+1|T}^2$, is given by $\widehat{\sigma}_{T+s|T}^2 = \frac{\omega\left(1-(\alpha+\beta)^{s-1}\right)}{1-\alpha-\beta} + (\alpha+\beta)^{s-1}\,\widehat{\sigma}_{T+1|T}^2$. Note that the long run forecast tends to the unconditional mean of the variance process exponentially fast. Finally, the information set for forecasting consists of the first 2500 observations while the forecasting horizon, s, takes on values between 1 and 250 (one day up to one year). The simulation is repeated $k = 5000$ times.

4.2 Performance Measurement and Simulation Results

For notational convenience let $\widehat{\Sigma}_{s,i|T} = \sum_{j=1}^{s} \widehat{\sigma}_{T+j,i|T}^2$; the forecasted cumulative realized variance over a specified forecast horizon s starting at period T. The subscript $i = 1, \ldots, k$ denotes the simulation run. Analogously, define $\Sigma_{s,i|T} = \sum_{j=1}^{s} \sigma_{T+j,i}^2$ as the cumulative true

[17] Note that the simulated realized variance series will be log normal and have an unconditional mean equal to $\exp\left[\mu + \frac{1}{2}c\triangle_\varepsilon^2\right]$ where $c = \sum_{h=0}^{\infty}\left[\frac{\Gamma(h-d)}{\Gamma(h+1)\Gamma(-d)}\right]^2$. For $d = 0.45$ we have $c \approx 2.5$ (we cut of the infinite summation at $h = 2500$) which corresponds to an annual return volatility of about 15%.

variance over a specified forecast horizon s starting at period T. The forecasting performance of each model is measured by computing the sum of squared error criteria: $SSE(s) = \sqrt{\sum_{i=1}^{k}\left[\widehat{\Sigma}^2_{s,i|T} - \Sigma^2_{s,i|T}\right]^2}$. To measure potential consistent over or under estimation the ratio of forecasted and actual average realized variance are computed for the different forecasting horizons: $RFA(s) = \sum_{i=1}^{k} \widehat{\Sigma}^2_{s,i|T} / \sum_{i=1}^{k} \Sigma^2_{s,i|T}$.

The simulation results are reported in the appendix. Figure 3.1 demonstrates that the forecasts for all models and all horizons are virtually unbiased. For forecasts up to a year the maximum under or over estimation of the accumulated realized variance is about 2.5% of the actual value. The sum of squared error criteria, which is used to assess the models' forecasting performance, is plotted in Figure 3.2. Not surprisingly, it is found that the ARFIMA model does the best job for all horizons considered. However, the difference in SSE between the ARFIMA and GARCH model for forecasting horizons up to say 3 months is small. For longer forecasting horizons the ARFIMA clearly outperforms the GARCH. This can be attributed to the ARFIMA model's ability to account for the long memory property of the log realized variance series. The standard deviations of forecasting errors are plotted in Figure 3.3; the ARFIMA model has the lowest forecast error variance for all horizons. Finally, the standard deviation, skewness and kurtosis of the forecast errors are calculated (not reported). The results indicate that both skewness and kurtosis are almost identical for the different models. Moreover, they decrease with the forecast horizon; skewness from about -1.5 for daily horizon to -0.5 for yearly horizon. Kurtosis from 8.5 for daily horizon to about 4.5 for yearly horizon. Some complimentary simulations are performed to investigate the model and parameter estimation risk but it is found that for both the ARFIMA and the GARCH model this risk appears to be negligible. The naive and simple estimation procedure of the GARCH yields sensible forecasts while the ARFIMA model's forecasting performance is insensitive to "noisy" input values of the fractional parameter.

The results reported above suggest that the ARFIMA model, as expected, outperforms the representative GARCH model according to all the criteria used. Several points should, however, be noted. The forecasts of the ARFIMA model are obtained from a truncated infinite autoregression, generally including a thousand lags or more. In contrast, forecasts of the GARCH model can be constructed using solely one lagged return and conditional variance estimate. Moreover, the GARCH model can be implemented with daily returns while the ARFIMA model requires intra-daily return data for the calculation of the realized variance measure. It is therefore clear that the ARFIMA model will be more difficult to implement in practise than the GARCH model. In addition, for a given forecasting horizon, the persistence of the misspecified GARCH can be controlled for by fixing the parameters to the appropriate values so as to minimize the difference in forecasting performance with respect to the ARFIMA model; while for short horizon forecasting the flexibility of the GARCH is more important than its persistence, the opposite is true for long horizon forecasting where a high persistence is needed to counter the long memory property of the variance process (i.e. "GARCH1 versus "GARCH2" in Figure 3.2 and 3.3).

5 Summary

Minute by minute data on the *FTSE-100* stock market index have been employed to calculate and analyze daily realized variance. It has been shown that in order to implement the "sum of squared returns" approach to calculate realized variance, a careful check on serial dependence in high frequency returns is required so as to avoid serious biases in the resulting measure for average daily return variance; i.e. for the *FTSE-100* data this (downward) bias amounts up to 35% for the minute data. In addition, it has been shown that the decay of serial dependence with the sampling frequency is consistent with an ARMA model under temporal aggregation. This finding can be used to set the "optimal sampling frequency", that is, the highest possible sampling frequency at which the autocovariance bias factor is negligible. Motivated by several test statistics for the presence of long memory, the realized variance time series is modelled as an ARFIMA model. Lagged returns and trading volume are found to be significant regressors. The Black leverage effect is tested for and found to be present at horizons of one to three days. This finding is supportive for asymmetric GARCH models such as the EGARCH and GJR GARCH model. Contemporaneous trading volume is helpful in explaining the variation in realized variance, although the persistence of the process remains unchanged upon inclusion of this variable. In a simulation study the forecasting performance of the ARFIMA model for daily (realized) variance is assessed and it is found that it outperforms conventional GARCH type models. It should be noted, however, that although the ARFIMA model works best, its implementation requires much more data than the GARCH. The relatively small loss at short horizons together with the flexibility of the GARCH to account for persistence in the variance process, make it a reasonable alternative to the complicated and data intensive ARFIMA model for realized variance.

6 References

Amemiya, T. (1985), "*Advanced Econometrics*", Harvard University Press, MA

Andersen, T.G. and T. Bollerslev (1998), "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts", *International Economic Review* 39, No. 4, 885-905

Andersen, T.G., T. Bollerslev, F.X. Diebold and H. Ebens (2000), "The Distribution of Stock Return Volatility," forthcoming *Journal of Financial Economics*

Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys (2000), "Great Realizations," *Risk*, 13, 105-108.

Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys (2001a), "Exchange rate returns standardized by realized volatility are (nearly) Gaussian", forthcoming in *Multinational Finance Journal*

Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys (2001b), "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association* 96, 42-55

Bai, X., J.R. Russell, and G. Tiao (2000), "Beyond Merton's Utopia I: Effects of Dependence and Non-Normality on Variance Estimates Using High-Frequency Data", mimeo Chicago Graduate School of Business

Baillie, R.T. (1996), "Long Memory Processes and Fractional Integration in Econometrics", *Journal of Econometrics* 73, 5-59

Baillie, R.T., T. Bollerslev and H.O. Mikkelsen (1996), "Modeling and pricing long memory in stock market volatility", *Journal of Econometrics* 73, 151-84

Black, F. (1976), Studies of stock price volatility changes, Proceeding from the American Statistical Association, Business and Economic Statistics Section, 177-181

Bollerslev, T., J. Cai and F.M. Song (2000), "Intra-day Periodicity, Long Memory Volatility and Marcoeconomic announcement effects in the US Treasury Bond Market", *Journal of Empirical Finance* 7, 37-55

Bollerslev, T. and H.O. Mikkelsen (1996), "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics* 74, 3-30

Breidt, F.J., N. Crato and P. de Lima (1998), "The Detection and Estimation of Long Memory in Stochastic Volatility", *Journal of Econometrics* 83, 325-348

Comte, F. and E. Renault (1998), "Long Memory in Continuous Time Stochastic Volatility Models", *Mathematical Finance* 8, 291-323

Corse, F., G. Zumback, U.A. Müller and M. Dacorogna (2001), "Consistent High-Precision

Volatility from High-Frequency Data", Olsen Group Working Paper

Diebold, F.X. and A. Inoue (1999), "Long Memory and Regime Switching", forthcomming Journal of Econometrics 2001

Dittmann, I. and C.W.J. Granger (2000), "Properties of Nonlinear Transformations of Fractionally Integrated Processes", UCSD discussion paper 2000-07.

Doornik, J.A. and M. Ooms (1998), "A Package for Estimation, Forecasting and Simulating Arfima Models: Arfima package 1.0 for Ox", mimeo Nuffield College, Oxford.

French, K.R., Schwert, G.W. and Stambaugh, R.F. (1987), "Expected Stock Returns and Volatility", *Journal of Financial Economics* 19, 3-29

Froot, K.A. and A.F. Perold (1995), "New Trading Practices and Short-Run Market Efficiency", *Journal of Futures Markets* 15(7), 731-765

Geweke J. and S. Porter-Hudak (1983), "The Estimation and Application of Long Memory Time Series Models", *Journal of Time Series Analysis* 4(4), 221-238

Glosten, Jagannathan and Runkle (1993), "On the relation between the expected value and the volatility of the nominal excess return on stocks", *Journal of Finance* 48, 1779-1801

Goodhart, C.A.E. and M. O'Hara (1997), "High Frequency Data in Financial Markets: Issues and Applications", *Journal of Empirical Finance* 4, 73-114

Granger, C.W.J. (1980), "Long memory relationships and the aggregation of dynamic models", *Journal of Econometrics* 14, 227-238

Granger, C.W.J. (1981), "Some properties of time series data and their use in econometric model specification", *Journal of Econometrics* 16, 121-130

Granger, C.W.J. (1999), "Aspects of Research Strategies for Time Series Analysis", Outline of presentation to the conference on New Developments in Time Series Economics at the Cowless Foundation, New Haven.

Granger, C.W.J. and N. Hyung (1999), "Occasional Structural Breaks and Long Memory", UCSD discussion paper 99-14.

Granger, C.W.J. and R. Joyeux (1980), "An introduction to long-memory time series and fractional differencing", *Journal of Time Series Analysis*, 1

Hsieh, D.A. (1991), "Chaos and Nonlinear Dynamics: Application to Financial Markets", *Journal of Finance* 46(5), 1839-1877

Lamoureux, C.G. and W.D. Lastrapes (1990), "Heteroskedasticity in Stock Return Data: Volume versus GARCH effects", Journal of Finance 45(1), 221-229

Lamoureux, C.G. and W.D. Lastrapes (1990b), "Persistence in Variance, Structural Change and the GARCH Model", *Journal of Business & Economic Statistics* 6(2), 225-234

Lequeux, P. (1999), "Financial Markets Tick by Tick: Insights in Financial Market Micro Structure", John Wiley & Sons Ltd, England

Liu, M. (2000), "Modeling Long Memory in Stock Market Volatility", *Journal of Econometrics* 99, 139-171

Lo, A.W. (1991), "Long Memory in Stock Market Prices", *Econometrica* 59(5), 1279-1313

Madhavan, A. (2000), "Market Microstructure: A Survey", Journal of Financial Markets 3, 205-258

Merton, R.C. (1980), "On Estimating the Expected Return on the Market; An Exploratory Investigation", *Journal of Financial Economics* 8, 323-361

Nelson, D.B. (1991), "Conditional Heteroskedasticity in asset returns: A new approach", *Econometrica* 59, 347-70

Robinson, P.M. (1995a), "Log-Periodogram Regression of Time Series with Long Range Dependence", *Annals of Statistics* 23(3), 1048-1072

Robinson, P.M. (1995b), "Gaussian Semiparametric Estimation of Long Range Dependence", *Annals of Statistics* 23, 1630-61

Smith, J., N. Taylor and S. Yadav (1997), "Comparing the Bias and Misspecification in ARFIMA models", *Journal of Time Series Analysis*, 18(5), 507-527

Sowell, F. (1992), "Maximum likelihood estimation of stationary univariate fractionally integrated time series models", *Journal of Econometrics* 53, 165-88

Taylor, S.J. and X. Xu (1997), "The Incremental Volatility Information in One Million Foreign Exchange Quotations", *Journal of Empirical Finance* 4, 317-340

Tiao, G.C. (1972), "Asymptotic Behaviour of Temporal Aggregates of Time Series", *Biometrika* 59, 525-531

Wei, W.W.S. (1981), "Effects of Systematic Sampling on ARIMA Models", *Communications in Statistical-Theoretical Mathematics* A 10, 2389-2398

Wood, R.A. (2000), "Market Microstructure Research Databases: History and Projections", *Journal of Business & Economic Statistics* 18(2), 140-145

Zhou, B. (1996), "High Frequency Data and Volatility in Foreign-Exchange Rates", *Journal of Business & Economic Statistics* 14(1), 45-52

A Appendix: Empirical Results

A.1 Descriptive Statistics and Estimation Results

Frequency	Skew	Kurtosis	BL (10)	DW	φ_1	φ_2	φ_3	φ_5	φ_{10}	φ_{15}
1 Min	30.6	1432	814	1.70	0.152	0.131	0.115	0.072	0.032	0.023
10 Min	10.1	410	308	1.70	0.151	0.071	0.049	-	-	-
25 Min	5.26	202	144	1.76	0.118	-	-	-	-	-
1 Hour	4.34	142	53.1	1.94	0.031	-	0.030	-	-0.029	-
1 Day	0.04	5.24	12.6	2.00	0.069	-	-	-	-	-

Table 1: Descriptive statistics of FTSE-100 returns at different sampling frequencies. BL(10) denotes the Box-Ljung Statistic with 10 autocorrelations (18.31 critical value at 95% confidence bound). DW denotes the Durbin Watson test. ρ_j denotes the j^{th} autocorrelation. The statistics are calculated using 10,000 observations for sampling frequencies down to one hour and 2,400 observations for the daily frequency. The 95% confidence bounds are therefore $\pm$ 0.02 and $\pm$0.042 respectively. Entries which contain the "-" symbol are not significant based on these bounds.

	Mean	Std.Dev.	Skewness	Kurtosis	ADF(5)
Realized Variance	8.54E-5	2.57E-4	21.2	596	-16.2
Log of Realized Variance	-9.98	0.962	0.558	4.11	-8.83
Daily Return	4.60E-4	9.28E-3	0.0628	5.29	-21.8
Standardized Daily Return	0.091	1.09	0.0362	2.23	-22.3

Table 2: Descriptive statistics for realized return variance and daily returns. The column "ADF(5)" reports the augmented dickey fuller test including a constant and 5 lags

Model	Sample	d	α_1	β_1	ζ_1^+	ζ_1^-	ζ_2^+	ζ_2^-	λ_0	LogL	AIC/T
ARFIMA	I	0.446 (9.20)	0.422 (5.92)	-0.636 (8.41)	-	-	-	-	-	-723	0.81
+ Returns	I	0.415 (7.52)	0.302 (4.63)	-0.597 (7.33)	33.6 (20.1)	39.4 (21.5)	14.8 (8.86)	18.0 (9.77)	-	-453	0.51
+ Volume	I	0.481 (19.8)	0.317 (6.39)	-0.677 (15.3)	27.7 (16.8)	36.4 (20.6)	11.7 (7.31)	16.9 (9.64)	0.379 (13.6)	-366	0.42
ARFIMA	II	0.487 (29.2)	0.401 (7.24)	-0.642 (13.1)	-	-	-	-	-	-968	0.80
+ Returns	II	0.484 (24.2)	0.327 (7.12)	-0.662 (16.0)	25.5 (20.5)	32.3 (24.2)	10.6 (8.52)	14.9 (11.2)	-	-649	0.54
+ Volume	II	0.489 (32.1)	0.339 (8.06)	-0.684 (19.4)	21.5 (17.7)	29.4 (22.7)	8.62 (7.22)	13.9 (10.9)	0.358 (15.3)	-537	0.45

Table 3: ARFIMA estimation results. Sample I runs from May 2, 1990 until June 15, 1997. Sample II runs from May 2, 1990 until January 11, 2000. d, α_1, and β_1 denote the ARFIMA(1,d,1) model parameters. ζ_i^+, ζ_i^-, and λ_0, are the regression coefficients of corresponding to the i^{th} lag of positive and negative returns and contemporeneous trading volume respectively. The last two columns contain the value of the log likelihood function and the Akaike information criteria. t-Statistics are reported in parenthesis.

A.2 Micro Structure & Sampling Frequency

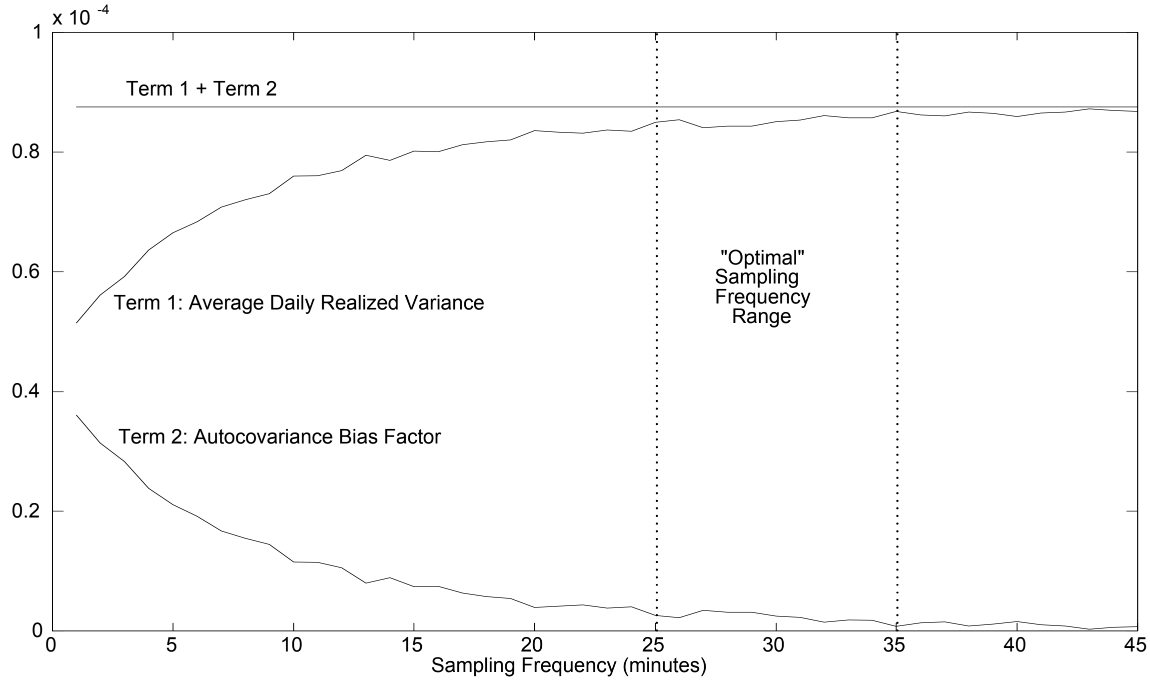

Figure 1.1: 1990-2000 average daily realized FTSE-100 variance and autocovariance bias factor

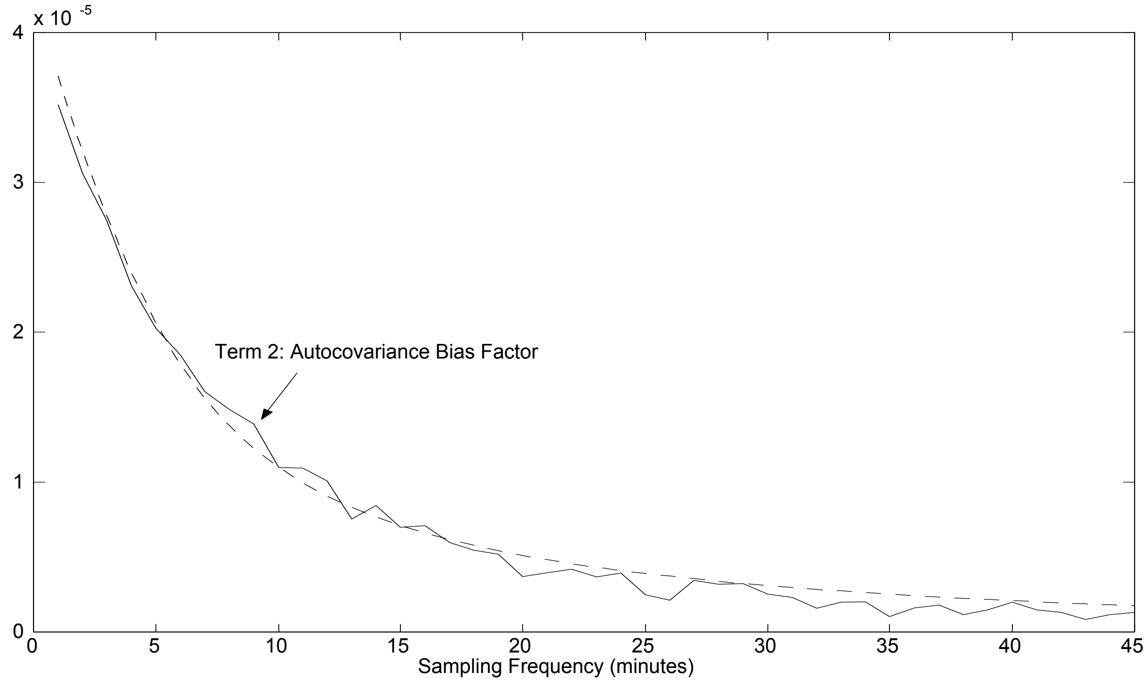

Figure 1.2: Empirical (solid line) and aggregation implied (dotted line) "Autocovariance Bias Factor"

A.3 The *FTSE-100* Realized Variance

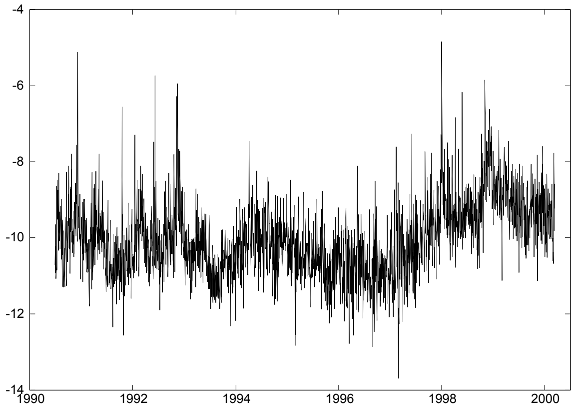

Figure 2.1: Log of realized variance

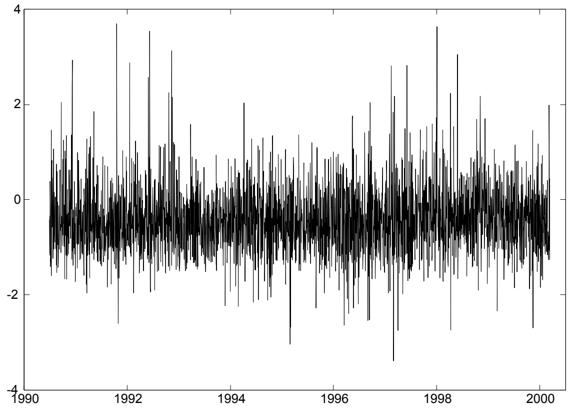

Figure 2.2: Fractionally differenced (using d=0.40) log realized variance series.

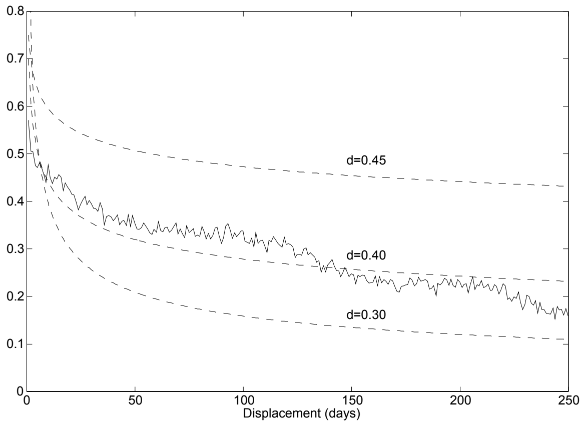

Figure 2.3: Correlogram log realized variance plus fractional implied correlations.

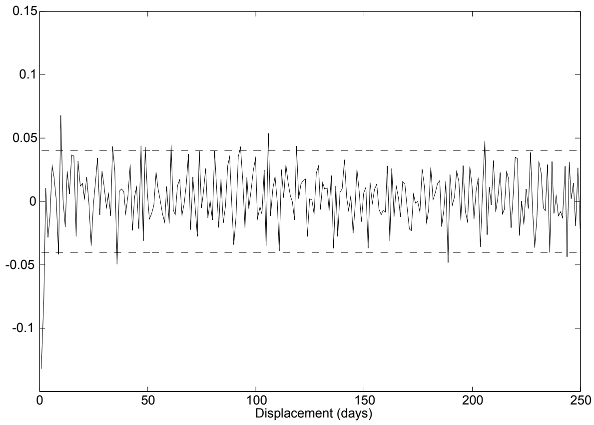

Figure 2.4: Correlogram of fractionally differenced log realized variance series

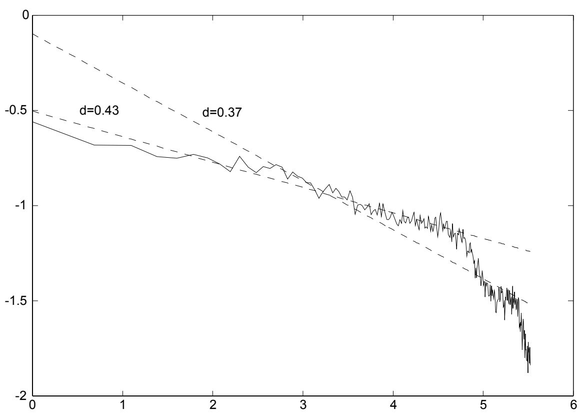

Figure 2.5: $\log \varphi_h$ versus $\log h$

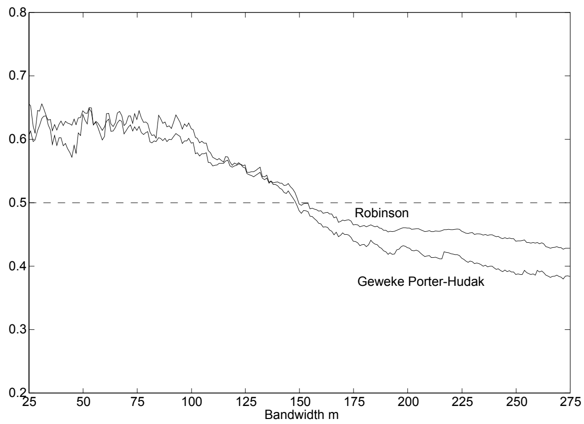

Figure 2.6: GPH & Robinson estimate for d as a function of the bandwidth m.

A.4 Forecasting Performance

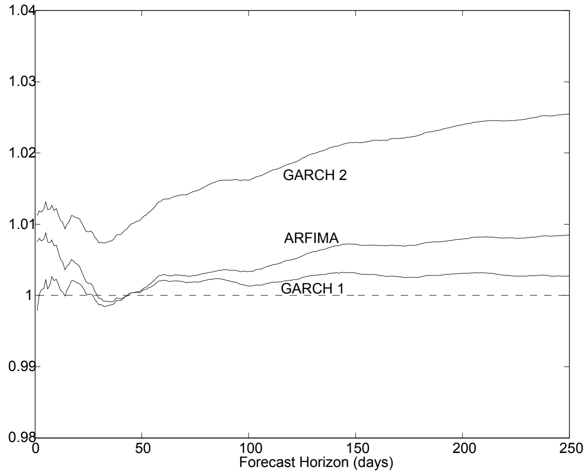

Figure 3.1: Forecasted divided by actual average realized variance

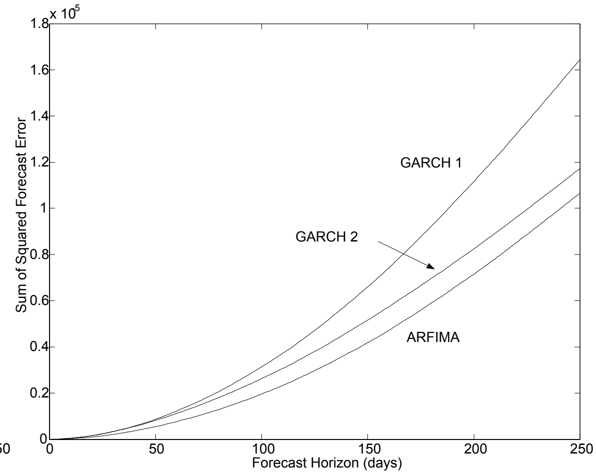

Figure 3.2: Sum of squared errors of forecasted cummulative realized variance

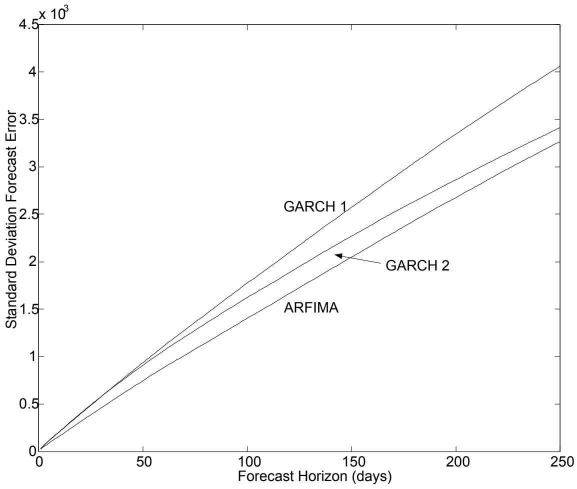

Figure 3.3: Forecast error standard deviation

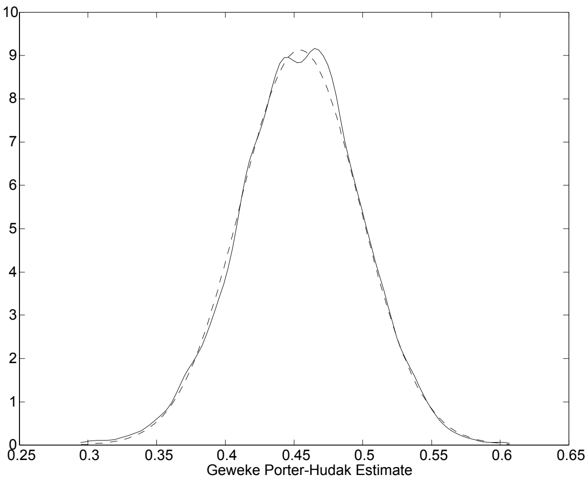

Figure 3.4: Density Geweke Porter Hudak estimates

B Appendix: Theory

B.1 Temporal Aggregation & Systematic Sampling

Consider an ARMA(p,q) process

$$\alpha(L) X_t = \beta(L) \varepsilon_t,$$

where $\alpha(L)$ and $\beta(L)$ are lag polynomials of lengths p and q respectively. Consider the case where the reciprocals of the roots of $\alpha(L) = 0$, $\theta_1, ..., \theta_p$, all lie inside the unit circle.

When X_t is a stock variable, Wei (1981) showed that by applying the operator $S(L)$

$$S(L) = \prod_{j=1}^{p} \frac{1 - \theta_j^f L^f}{1 - \theta_j L}$$

to ARMA(p,q) process, the ARMA(p,r) model through which the filtered subseries may be represented is obtained. In this case, f denotes the systematic sampling frequency. When X_t is a flow variable, Wei (1981) showed that by applying the operator $T(L)$

$$T(L) = S(L) \frac{1 - L^f}{1 - L}$$

to ARMA(p,q) process, the ARMA(p,r) model through which the filtered subseries may be represented is obtained. In this case, f denotes the temporal aggregation frequency. The moving average lag length r is given by the integer part of $p + \frac{q-p}{f}$ and for large f will be equal to p or $p - 1$, depending on whether $q \geq p$ or $q < p$.

Note that although the order of the autoregressive part of the ARMA model remains the same under systematic sampling or temporal aggregation, the magnitude of the autoregressive coefficients decrease exponentially with f. Hence, in the limit the autoregressive part will disappear. Moreover, for temporal aggregation, the term $\frac{1-L^f}{1-L}$ will dominate the MA part and hence the limit model is an ARMA(0,0) or equivalently white noise.

B.2 Fractional Integration

A time series, X_t is said to be fractionally integrated of order d, if after applying the difference operator $(1 - L)^d$ it follows a stationary ARMA(p,q) process where p and q are finite nonnegative integers. The concept was developed by Granger (1980, 1981) and Granger and Joyeux (1980). A typical feature of a fractionally integrated or long memory process is that the effect of a shock to the process is highly persistent but decays over time. This, as opposed to $I(1)$ processes where a shock has infinite persistence or at the other extreme $I(0)$-processes in which the effect of a shock to the system decays exponentially fast. Moreover,note that the fractional difference operator takes care of the long run dynamics while the ARMA structure can account for the short run dynamics. The ARFIMA(p,d,q) model can be written as

$$\alpha(L)(1-L)^d X_t = \beta(L)\varepsilon_t, \tag{4}$$

where $\alpha(L)$ is a lag polynomial of order p and $\beta(L)$ of order q. Note that the ARMA(p,q) and Integrated ARMA(p,q) models are special cases of (4) for $d = 0$ and $d = 1$ respectively. Using a binomial[18] or Taylor-like expansion (around $L = 0$), the fractional difference operator can be expressed as follows:

$$\begin{aligned}(1-L)^d &= 1 - dL - \frac{1}{2}d(1-d)L^2 - \frac{1}{6}d(1-d)(2-d)L^3 - \ldots \\ &= \sum_{h=0}^{\infty} \frac{\Gamma(h-d)}{\Gamma(h+1)\Gamma(-d)} L^h \\ &= 1 + \sum_{h=1}^{\infty} \frac{\prod_{n=1}^{h}(n-1-d)}{h!} L^h \end{aligned} \tag{5}$$

Using Stirling's formula[19] it can be shown that:

$$\frac{\prod_{n=1}^{h}(n-1-d)}{h!} = \frac{\Gamma(h-d)}{\Gamma(-d)\Gamma(h+1)} \propto h^{-d-1}$$

for h large. Moreover, for $d < \frac{1}{2}$ and $d \neq 0$, it can be shown that:

$$\varphi_h = corr(X_t, X_{t-h}) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(h+d)}{\Gamma(h+1-d)} \underset{h \text{ large}}{\propto} h^{2d-1} \tag{6}$$

and hence the decay of the correlogram is hyperbolic, as opposed to an exponential decay for an $I(0)$ process. For $d = 0$, $\varphi_h = 0$ for $h > 0$. The process is stationary and long memory for $0 < d < 0.5$. The process is stationary and intermediate memory when $-0.5 < d < 0$. For $d \geq 0.5$, the process is non-stationary.

---

[18]The Binomial Theorem states that $(1-L)^d = \sum_{h=0}^{\infty} \binom{d}{h} (-1)^h L^h$ where $\binom{d}{h}(-1)^h = \Gamma(h-d)/[\Gamma(-d)\cdot\Gamma(h+1)]$ and $\Gamma(\cdot)$ is the Gamma function for which it holds that $\Gamma(x) = (x-1)\cdot\Gamma(x-1)$

[19]Stirling's formula states that $\Gamma(h) \approx \sqrt{2\pi}\cdot h^{h-1/2}e^{-h}$ for h large and hence $\frac{\Gamma(h+a)}{\Gamma(h+b)} \approx h^{a-b}$.

B.2.1 Log-Periodogram Regressions

Several two step procedures have been proposed to estimate the fractionally integrated ARMA model. The main idea is to estimate the fractional parameter, d, in the first step while in the second step this estimate can be used to fractionally difference the observed series, transforming it into an ARMA process of which the parameters can be obtain straightforwardly by ordinary least squares. Geweke Porter-Hudak (1983, GPH hereafter) propose to estimate d by a log-periodogram regression which is described below. Consider $(1-L)^d X_t = \varepsilon_t$ where ε_t is a stationary linear process with spectral density function $f_\varepsilon(\lambda)$ which is finite, bounded away from zero and continuous on the interval $[-\pi, \pi]$. The spectral density function of $\{X_t\}$ is $f(\lambda) = \frac{\sigma^2}{2\pi}\left[4\sin^2\lambda\right]^{-d} f_\varepsilon(\lambda)$ or equivalently

$$\ln f(\lambda) = \ln \frac{\sigma^2 f_\varepsilon(0)}{2\pi} - d \ln\left[4\sin^2(\lambda/2)\right] + \ln \frac{f_\varepsilon(\lambda)}{f_\varepsilon(0)}. \tag{7}$$

Define the harmonic frequency $\lambda_j = \frac{2\pi j}{T}$ where T is the sample size and let $I(\lambda_j)$ be the periodogram at λ_j which is given by

$$I(\lambda_j) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} X_t e^{i\lambda_j t}\right|^2.$$

Rearranging terms and evaluating expression (7) at λ_j close to zero (the term $\ln \frac{f_\varepsilon(\lambda)}{f_\varepsilon(0)}$ is therefore negligible) the following expression is obtained:

$$\ln I(\lambda_j) = \ln \frac{\sigma^2 f_\varepsilon(0)}{2\pi} - d \ln\left[4\sin^2(\lambda_j/2)\right] + \ln \frac{I(\lambda_j)}{f(\lambda_j)}. \tag{8}$$

Therefore, the coefficient d can now be estimated as the slope coefficient in a least squares regression of $\ln I(\lambda_j)$ on a constant and $\ln\left[4\sin^2(\lambda_j/2)\right]$ for $j = 1 + l, \ldots, m \ll T$. GPH set $l = 0$ and require that the bandwidth parameter m increases at a slower rate than the sample size. In many practical applications m is set to equal to the square root of the sample size T. Robinson (1995a) provides the asymptotic behavior of the estimator. In addition, Robinson (1995b) proposed an alternative estimator which is derived under weaker conditions and proved to be asymptotically more efficient than the GPH estimator. This estimator, for the fractional parameter, is given by the value of d that minimizes the following objective function:

$$Q(c, d) = \frac{1}{m}\sum_{j=1}^{m}\left[\ln\left(c\lambda_j^{-2d}\right) + \frac{\lambda_j^{2d}}{c} I(\lambda_j)\right],$$

where $c > 0$ and d is restricted to lie between $-\frac{1}{2}$ and $\frac{1}{2}$.

B.2.2 Exact Maximum Likelihood

Motivated by the finding that the ARMA parameters are generally not estimated accurately using a two step method, Sowell (1992) proposes an exact maximum likelihood procedure that estimates all model parameters simultaneously. Let Σ denote the covariance matrix of the joint distribution of $\{X_t\}_{t=1}^{t=T}$ where X_t is assumed to follow an ARFIMA(p,d,q) process as given by expression (4). The model parameters can be estimated by maximizing the following log-likelihood function over the parameter space

$$\log L\left(d,\alpha,\beta,\sigma_{\varepsilon}^{2}\right)=-\frac{T}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left|\Sigma\right|-\frac{1}{2}X'\Sigma^{-1}X,$$

where $\Sigma_{ij}=\varphi_{|i-j|}$ for $i,j=1,2,\ldots,T$, is expressed in terms of the model parameter. Sowell (1992) shows that the covariance matrix, needed for the estimation, is given by

$$\varphi_{s}=\sigma_{\varepsilon}^{2}\sum_{l=-q}^{q}\sum_{j=1}^{p}\vartheta\left(l\right)\varkappa_{j}C\left(d,p+l-s,\varphi_{j}\right)$$

where

$$\vartheta\left(l\right)=\sum_{s=\max[0,l]}^{\min[q,q-l]}\beta_{s}\beta_{s-l}$$

and

$$\varkappa_{j}^{-1}=\theta_{j}\prod_{i=1}^{p}\left(1-\theta_{i}\theta_{j}\right)\prod_{m\neq j}\left(\theta_{j}-\theta_{m}\right)$$

where θ_j denote the roots of $\alpha(L)$ and are assumed to lie outside the unit disk. Finally, the expression for C is given by

$$C\left(d,h,\theta\right)=\frac{\Gamma\left(1-2d\right)\Gamma\left(d+h\right)\left[\theta^{2p}F\left(d+h,1;1-d+h;\theta\right)+F\left(d-h,1;1-d-h;\theta\right)-1\right]}{\Gamma\left(1-d+h\right)\Gamma\left(1-d\right)\Gamma\left(d\right)}$$

where $F\left(a,b;c;e\right)$ is the hypergeometric function.