FOURIER METHOD FOR THE MEASUREMENT OF UNIVARIATE AND MULTIVARIATE VOLATILITY IN THE PRESENCE OF HIGH FREQUENCY DATA

Chanel Malherbe

June 4, 2007

Abstract

The increasing availability of intraday data as well as the increase in computational power has created an interest in the calculation of volatility and correlation using high frequency data instead of the more commonly used daily/weekly/monthly data that is available. However, on a high frequency scale, financial time series are not evenly spaced or synchronous, and therefore standard methods of calculating volatility and correlation cannot be directly applied. This dissertation evaluates a method proposed by Malliavin and Mancino [22] that uses a method based on Fourier series analysis to calculate the univariate and multivariate volatility. This method is compared with the traditional methods of estimating volatility and cross-correlation from tick-by-tick (high frequency) data in the context of an emerging market.

In the case of evenly spaced data, we found that the Fourier method compares very well with classical methods and provide smoother estimates. In the case of high frequency data, we confirmed and extended the results of Iori [34] and found that the Fourier method gives better results than the realised volatility estimator in terms of generating smooth estimates with a lower bias and root mean squared error, which are also less sensitive to the choice of returns time scale. It is conceptually superior to methods that use interpolation and is also model independent. In addition, the Fourier method guarantees a positive definite matrix, which is not the case with other classical methods.

The dataset analysed in this paper is the two-and-a-half year tick-by-tick trades executed on the JSE Stock Exchange from May 2002 till October 2004. The dataset was provided by Deutsche Bank South Africa.

Acknowledgement

I am most grateful to my supervisor, Dr. Diane Wilcox, for her guidance, patience and helpful suggestions through the course of this research. I also thank Dr. Tim Gebbie for interesting discussions and his insightful comments and suggestions. I am also deeply grateful to my family for their support and encouragement throughout my period of study.

We also acknowledge and thank Pieter Snyman, Deutsche Securities and Roland Rousseau, Deutsche Bank South Africa for providing the 3 years of matched trade-by-trade data as flat files and for providing their proprietary data for research purposes at the University of Cape Town. The price matching from the high frequency tick-by-tick datasets to the high frequency trade-by-trade datasets was carrier out by Deutsche Securities. The data preprocessing was carried out by Dr. Tim Gebbie and Dr. Diane Wilcox at the University of Cape Town.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The volatility of a financial instrument is a measure of the uncertainty about the returns provided by the instrument and is therefore one of the most important parameters in asset pricing, optimal asset allocation and risk management. Unfortunately, volatility cannot be directly observed and must therefore be estimated from historical data. While Black, Scholes and Merton made the assumption that volatility is constant when they established their theory of option pricing [12], recent financial econometrics literature has shown that volatility is in fact time-varying with a high degree of persistence [39]. Since the seminal paper by Engle [10], the development of new tools for volatility measurement, modelling and forecasting has therefore been, and still is, a topic of active research.

Volatility can be computed through parametric or nonparametric methods. We use the framework proposed by Andersen and Bollerslev [39] to categorise the various volatility modelling procedures.

In the case of parametric methods, the expected volatility is modelled using a functional form of variables observed in the market. Parametric methods include *discrete-time volatility models* such as ARCH models, where the volatility depends on past returns and other directly observable variables only, *stochastic volatility models*, where past returns as well as latent state variables are used, and *continuous-time volatility models*, where the instantaneous volatility is explicitly modelled with possible additional volatility dynamics. In addition to these three separate model classes, so-called *implied volatility* approaches also figure prominently in the literature. The implied volatilities are based on a parametric volatility model for the returns as defined above, along with an asset pricing model, such as the Black-Scholes model, and an augmented information set consisting of option prices and/or term structure variables. Implied volatility methods depend on a liquid derivatives market, which is not available for most derivatives traded on the

JSE.

Nonparametric methods address the computation of the historical volatility without assuming a functional form of the volatility. As volatility changes over time, its computation through nonparametric methods focuses on small time windows (daily, weekly, monthly) and high frequency data is employed. In the case where volatility is measured using high frequency data most, if not all, of the emphasis in the literature has so far been on nonparametric methods.

In this paper, the focus of the study will be on measuring volatility using high frequency data and nonparametric methods. This is a topic that has gained much interest from both the professional and academic communities in recent years. Jungbacker [3] defines high frequency data as data obtained by collecting all prices in a certain period. In finance literature, this usually means that observations are recorded trade-by-trade. The trade-by-trade data is regarded as the ultimate high frequency collection of prices. In a time scale of seconds, multiple trades may even occur within the same time interval, although this is unlikely. It is, however, more likely that many prices will be 'missing', since trades do not take place every second in most financial markets, especially not in emerging markets such as South Africa.

But why do we want to measure volatility using high frequency data? Zumbach, Corsi and Trapletti [15] give a simple explanation via the following question: Given one realisation of a random walk with constant drift μ and volatility σ, what are the minimal sufficient statistics to estimate the variables μ and σ? For the drift, only the start and end points are necessary. Thus, there is no need to use high frequency data. However, for the volatility, all the increments help in getting a better estimate, and the omission of any of the original data implies a loss of information. Hence, we should use data at the highest frequency available, i.e. trade-by-trade data. This point was also proved in this paper, where it was shown that more accurate volatility estimates are obtained from high frequency data than daily data. It should, however, be noted that high frequency data should be handled with care, because the observed asset returns could contain error or noise.

A common practise is to estimate volatility from the sum of the frequently-sampled squared returns, known as the *realised volatility*. Most of these methods assume that price series are evenly spaced. In [39], Andersen and Bollerslev have, however, shown that the squared daily returns provide a poor approximation of realised volatility. More accurate results are obtained with the sum of squared intraday returns. Unfortunately, returns, i.e. high frequency data, is not necessarily evenly spaced. To overcome this problem, most classical methods throw away some data by making use of some sort of imputation or interpolation scheme to create evenly spaced time series from the unevenly spaced high frequency data. This could lead to biased estimates [8]. In addition, the calculation of the volatility involves a sort of numerical derivative. This is not always desirable, since it is well doc-

umented that differentiation of empirical functions often generate a large instability [22].

In this dissertation, we specifically investigate a nonparametric method proposed by Malliavin and Mancino [22] based on Fourier analysis to measure volatility exploiting high frequency observations. The Fourier method enables the measurement of instantaneous volatility and the so-called integrated volatility. A key point is that volatility is reconstructed as a function of time. This feature is essential when stochastic derivation of volatility along the time evolution is performed as in contingent claim pricing-hedging [41], [27]. An additional benefit of this method is that it can be generalised to the multivariate case, i.e. the measurement of cross-correlations or covariances.

The Fourier method has been evaluated in the literature, but so far, all simulations have been based on models chosen to represent liquid stock markets, such as stocks from the S&P100. We evaluate this method in the South African context, where time series are less liquid. The results obtained are compared to that obtained from the realised volatility estimator, which will be discussed in the next chapter. The comparison is done through Monte Carlo experiments. The results on empirical data are then reported.

1.2 The Literature

This section briefly refers the reader to some recent literature related to volatility estimation, especially realised volatility, and then gives a comprehensive overview of papers written on the Fourier method.

Volatility has been one of the most active areas of research in empirical finance and time series econometrics during the past decade. Andersen and Bollerslev [39] provide a unified framework for categorising the various volatility concepts, measurement procedures and modelling procedures. They define the different types of volatility and thoroughly survey the parametric and nonparametric approaches to volatility modelling. Bollerslev [4] provides a selective summary of the most important developments in the field of econometrics over the past two decades, focusing primarily on ARCH and GMM models. Aboura [1] reviews local and implied volatility models, while Barndorff-Nielsen and Shephard [30] provide a comprehensive overview of the econometrics of nonparametric estimation of the components of the variation of asset prices. Litterman and Winkelmann [36] give an overview of the problems encountered in the multivariate case.

There exists a large body of literature on the estimation of the volatility through the quadratic variation formula called realised volatility, starting with papers by Andersen and Bollerslev [38] and Barndorff-Nielsen and Shephard [29]. For a comprehensive overview of the topic the reader can refer to [31], which recalls the fundamental result that realised volatility is

a consistent estimator of quadratic variation and derives the rate of convergence of realised volatility to quadratic variation. In [28], the moments and asymptotic distribution of the realised volatility error is derived when intraday data is used. In [35], Hansen and Lunde investigate the behaviour of realised volatility in the presence of market micro-structure noise, while Bandi and Russel [13] show that the realised volatility estimator diverges to infinity almost surely when noise plays a role, as in a realistic price formation mechanism.

The volatility estimator to be evaluated in this paper, known as the *Fourier estimator*, was first introduced by Malliavin and Mancino [22]. While most of the classical estimators are built on evenly sampled observations, they proposed an estimator based on Fourier series analysis that overcomes many of the problems encountered when volatility is measured using high frequency data, without using interpolation to convert the unevenly spaced time series to evenly spaced series. This method is based on the integration of the time series, rather than on its differentiation, and is nonparametric. This makes the method well suited to financial market applications, in particular for the analysis of high frequency time series and for the computation of covariances.

In [23], the results obtained in [22] are generalised by obtaining an identity which connects the Fourier transform of the volatility with the Fourier transform of the variation of the price process. Moreover, they provide the almost sure convergence of the proposed estimator to the volatility function and provide a limit distribution theory for the error. This method has the additional advantage of always providing a positive estimate of the instantaneous volatility. This is not the case with the original Fourier method. Both cases, however, provide identical positive estimates for the integrated volatility.

The Fourier method was first evaluated by Reno [37], where it was used in an attempt to better understand the dynamics underlying the Epps effect - the phenomenon of a drop in correlations among stocks when the sampling horizon is decreased. (See [11] for more information on this phenomenon.) This was also the first time the Fourier estimator was used to estimate correlations. It was found that the method is well suited to the time structure of high frequency data and computed correlations in a precise way when tested on Monte Carlo bivariate experiments.

In [6], simulations of the continuous-time GARCH model are used by Barucci and Reno to show that the method performs well in computing integrated volatility. It is also shown that the forecasting performance of the GARCH model is improved with respect to what is established when classical methods are employed.

In the paper by Hog and Lunde [9], an estimator of integrated volatility using wavelet methods is derived. They demonstrate that the performance of this estimator is indeed as good as the Fourier based method, but compu-

tationally the wavelet based method is preferable. The main disadvantage of this method is that it cannot be extended to the multivariate case.

In [26], Nielsen and Frederiksen consider the wavelet method in comparison with the Fourier estimator and realised volatility. The estimation methods are compared in a simulation study using a number of different data generating processes, which reveals a general robustness toward persistence or jumps in the latent stochastic volatility process. They found that, when looking at the root mean squared errors (RMSE), the Fourier method is superior to the other two estimators, while having only a slightly higher bias.

In [33] and [34], Precup and Iori compare the Fourier method with realised volatility using linear interpolation and the weighted realised volatility method proposed by Dacorogna et al [25]. In [33], the univariate case is examined, while [34] looks at the multivariate case. The three methods are tested on simulated data and actual trades time series, where it is found that the Fourier method is better than the other two methods in generating smooth estimates, which are not oversensitive to the choice of the returns time scale.

In [19],[20] and [21], Kanatani proposes a new framework building on the theory of quadratic variation. In this context, the Fourier method, realised volatility and a method named raw data realised volatility are evaluated and an explicit link between these methods are derived.

1.3 Outline of the Dissertation

The outline of the rest of the dissertation is as follows: Chapter 2 covers the theoretical background. We start by looking at important concepts and definitions related to the study of volatility and then take a detailed look at the Fourier method and its derivation as proposed by Malliavin and Mancino [22] in Chapter 3. In Chapter 4, the results of the estimator in the univariate setting are considered by evaluating the performance of the different estimators using Monte Carlo experiments. The behaviour of the estimators when empirical data is used is also considered. Chapter 5 examines the multivariate case, while Chapter 6 concludes. Note that the appendix contains a glossary of terms.

Chapter 2

Theoretical Background

In this chapter, we review some of the fundamental definitions and theorems surrounding the concepts of volatility and correlation that will later be used or quoted in this research.

2.1 Volatility Basics

Assume, as is done in most modern finance theory, that all measurable economic data p are driven by semi-martingales. We remind the reader of the definition as expressed in [31].

Definition 1 *Suppose $p(t)$ is a stochastic process and that for ease of exposition we assume that $p(0) = 0$. Then $p(t)$ is said to be a semi$-$martingale if it is decomposable as*

$$p(t) = \alpha(t) + m(t), \quad \alpha(0) = m(0) = 0,$$

where $\alpha(t)$, a drift term, is a process with bounded variation paths i.e. the real function α on $[a, b]$ is such that

$$\sup_k \sum |\alpha(x_i) - \alpha(x_{i-1})| < \infty,$$

where the supremum is taken over all subdivisions k of $[a, b]$, and $m(t)$ is a local martingale.

Now, specifically assume that p is such a univariate semi-martingale which has its Itô stochastic differential equation (SDE) given by

$$dp = \sigma dx + \beta dt, \tag{2.1}$$

where σ and β are random processes satisfying some hypothesis. In this case, they are uniformly bounded functions depending upon time. No parametrisation of the process is assumed. We call σ^2 the *instantaneous* or *spot*

volatility, while β is a drift function representing expected return. The x is an independent Brownian motion, where Brownian motion is defined as follows:

Definition 2 *A one-dimensional stochastic process $(B_t \to t \geq 0)$ is called a standard Brownian motion or a Wiener process if and only if:*

- ***Continuous sample paths:*** *The map $t \to B_t(\omega)$ is continuous for almost all ω.*
- ***Independent increments:*** *Given $0 \leq t_0 < t_1 < \ldots t_n$, the random variables*
$$B_{t_k} - B_{t_{k-1}}, \quad k = 1, \ldots, n \tag{2.2}$$
are mutually independent.
- ***Normally distributed increments:*** *If $0 \leq s < t$, then*
$$B_t - B_s \approx N(0, t-s). \tag{2.3}$$

Malliavin and Mancino [22] refer to Equation 2.1 as the Bachelier paradigm. Under these assumptions and using Itô calculus, the instantaneous volatility of the process p, at time t, where $0 \leq t \leq T$, is obtained by

$$\sigma^2 \equiv \lim_{h \to 0} \frac{1}{h} [E[(p(t+h) - p(t))^2 | F_t], \tag{2.4}$$

where $E[.|F_t]$ denotes the conditional expectation operator with respect to the σ-field F_t generated by the full observation of the process until time t [23].

Since volatility measurement takes place over discrete time intervals, volatility is mainly computed relying upon the quadratic variation formula. We recall the definition.

Definition 3 *Let $p(t)$ denote any semi-martingale. The quadratic variation process $< p >_t$, $t \in [0, T]$ associated with $p(t)$ is formally defined by*

$$< p >_t \equiv \lim \sum_{i=1}^{N} (p(s_i) - p(s_{i-1}))^2 \tag{2.5}$$

where $0 = s_0 < s_1 < \ldots < s_N = T$ and the limit is taken for $\max_i |s_i - s_{i-1}| \to 0$ as $N \to \infty$.

The following theorem summarises the relationship between instantaneous volatility (σ^2), integrated volatility (σ^2_{INT}) and quadratic variation ($< p >_t$) under the Bachelier paradigm (Equation 2.1) as shown by Malliavin and Mancino [22].

Theorem 1 *Assume that the instantaneous volatility function σ and the drift term β are uniformly bounded, then almost surely we have the identity*

$$< p >_t = \int_0^t \sigma^2(t)dt, \tag{2.6}$$

where

$$\sigma^2_{INT} = \int_0^{2\pi} \sigma^2(t)dt \tag{2.7}$$

is known as the integrated volatility.

Therefore, for each path of the asset p, $< p >_t$ converges to the integrated volatility. Thus, a single path of the asset - which is all that is available - is sufficient to calculate volatility using the quadratic variation.

In this paper, the object of interest is the *integrated (cumulative) volatility*, denoted by σ^2_{INT}. This is the relevant volatility measure for option pricing and the parameter of interest for econometricians.

From the above, we know that, if we have $M+1$ evenly spaced observations of a process $q(jT/M)_{j=0}^M$, the integrated volatility can be measured as follows:

$$\sigma^2_{INT}(M) = \sum_{j=1}^{M} (q(\frac{jT}{M}) - q(\frac{(j-1)T}{M}))^2. \tag{2.8}$$

This is known as the *realised volatility* of the process q(t).

The definition and analysis of realised volatility in financial return series has attracted considerable interest in the literature. See, for example, [39] and the references therein for a review.

Note that some authors use the term realised variance for the quantity (2.8) while the term realised volatility is then used for the square root of this equation. We shall use the name realised volatility when referring to (2.8).

By definition, realised volatility is a consistent estimator of integrated volatility. This result was strengthened when Shephard [29] showed that the convergence of realised volatility to integrated volatility happens at a rate of $\sqrt{M}$. However, although the calculation of the realised volatility is well defined and rather simple, a number of unresolved issues exist with respect to application of the rule and interpretation of the results in the high frequency data domain. The main problem is that the data input is a time series with homogeneous, i.e. equal, spacing between ticks. When we work with daily/weekly/monthly data, this requirement is easily satisfied. However, high frequency data is often unevenly spaced with the trade-by-trade data having different frequencies which may or may not overlap.

Most methods to solve this problem make use of some form of imputation or interpolation. In Dacorogna et al [25], some interpolation techniques,

including linear interpolation and previous-tick interpolation, are discussed. We adopt their notation.

When constructing $M+1$ evenly spaced data points $q(jT/M)_{j=0}^{M}$ from $p(t_k)_{k=0}^{N}$, where N is the number of raw observations and M is the number of evenly spaced data points, the data manipulation is as follows:

$$q(\frac{jT}{M}) = \begin{cases} (1-\rho_j)p(t_M^-) + \rho_j p(t_j^+) & \text{linear interpolation} \\ p(t_j^-) & \text{previous-tick interpolation} \end{cases}$$

where

$$\begin{aligned} \rho_j &= \frac{(jT/M) - t_j^-}{(t_j^+ - t_j^-)}, \\ t_j^- &= max(t_i : t_i \leq jT/M), \\ t_j^+ &= min(t_i : t_i \geq jT/M). \end{aligned} \tag{2.9}$$

Thus, when using previous-tick interpolation, the time-line is split into evenly spaced intervals and the last observation is used inside each interval. Note that the number of observations, N, should preferably be much higher than the number of evenly spaced data points, M, to avoid using the same observation more than once. In interpolation methods, the point corresponding to the evenly spaced interval between the observed market prices in a given interval is taken. While these methods give homogeneous and equally spaced time series, there is the possibility that they will introduce a false bias into the univariate and multivariate volatility calculations. This was shown by Barucci and Reno [7]. Kanatani [20] calculates the theoretical bias and notes that the bias is more pronounced when the time window is more finely divided, the interpolated time point is far from the observed time points, the intervals are coarsely-sampled or the volatility is large. In the case of previous-tick interpolation, it was shown that the realised volatility is unbiased. We will therefore use this method in our implementation of realised volatility when high frequency data is used.

If high frequency data is available, the granularity of the realised volatility estimator is determined by the choice of M. For example, if we have data for every minute of a 24 hour day, i.e. 1440 observations, choosing $M = 288$ will correspond to 5 minute returns. While additional information could be obtained when the value of M is high, it could also increase the sensitivity towards micro-structure effects in the market, for example measurement errors. In [42], Mykland, Ait-Sahalia and Zhang look at the optimal sampling frequency in the presence of market micro-structure noise, and conclude that the answer is to model the noise term and then sample as frequently as possible. In our implementation of realised volatility we will

restrict ourselves to modelling realised volatility in its standard form, using different values for M.

Another problem with using realised volatility is that the calculation of the instantaneous volatility involves a numerical derivative, while it is well documented that differentiation of empirical functions often generates a large instability [22]. In addition to the above, it was shown that, since not all of the available data is employed, the power of statistical tests are reduced.

An extension to realised volatility was examined by Kanatani [19]. Like the realised volatility estimator, this method is also based on the quadratic variation formula. While the formal definition for realised volatility is stated in terms of equally spaced observations, the extension makes use of all the raw data and does not assume that the data points are evenly spaced. This is given by

$$\sigma_R^2 = \sum_{i=1}^{N} (\Delta p(t_i))^2, \tag{2.10}$$

where $\Delta p(t_i) = p(t_i) - p(t_{i-1})$. To distinguish the standard realised volatility estimator given by (2.8) from (2.10), we employ the same terminology as [19] and refer to this estimator as the *raw data realised volatility* or RDRV. This estimator is based on the results of Andersen and Bollerslev [40], where it was shown that the classical realised volatility estimator is a consistent estimator of integrated volatility and that the consistency result does not require the observations to be evenly spaced, only that the maximum distance between observations goes to zero in the limit. This requirement can, however, not always be effectively realised in an illiquid market such as the South African market. In addition, the granularity of the raw data realised volatility estimator cannot be adjusted.

In this paper, we will focus on the method proposed by Malliavin and Mancino [22] and extended in [23], where volatility is calculated in the context of Fourier series. This method is discussed in the next chapter.

2.2 Correlation Basics

The concept of realised volatility also has a multidimensional counterpart, which can be defined as follows: Given two processes p_1 and p_2, the instantaneous correlation of the processes at time t, where $0 \leq t \leq T$, is obtained by

$$\Sigma^2 \equiv \lim_{h \to 0} \frac{1}{h} [E[(p_1(t+h) - p_1(t))(p_2(t+h) - p_2(t))]|F_t], \tag{2.11}$$

where $E[.|F_t]$ denotes the conditional expectation operator with respect to the σ-field F_t generated by the full observation of the two processes until time t.

In addition, the concept of quadratic variation can also be extended to the multivariate case. This is called the quadratic covariation and is derived from the quadratic variation by the polarization identity

$$< p_1, p_2 >= \frac{1}{4}(< p_1 + p_2, p_1 + p_2 > - < p_1 - p_2, p_1 - p_2 >).$$

Theorem 1 can be extended to show integrated covariation can be calculated using the quadratic covariation in the same way that integrated volatility can be calculated using the concept of quadratic variation [22].

Chapter 3

The Fourier Estimator

3.1 The Fourier Estimator

In this section, we present the method proposed by Malliavin and Mancino [22], henceforth called the Fourier method.
This method needs only the assumption that the quadratic variation of the economic time series of interest is bounded. This will be true if we suppose that all measurable economic data p are driven by semi-martingales (See Definition 2.1), which have their Itô stochastic differential given by the Bachelier Paradigm (See Equation 2.1).

Using Itô calculus, the instantaneous volatility matrix of the process p is obtained from:

$$\Sigma^2 \equiv \lim_{h\to 0}[E\frac{[(p_1(t+h)-p_1(t))(p_2(t+h)-p_2(t))}{h}]|F_t], \tag{3.1}$$

where $E[.|F_t]$ denotes the conditional expectation operator with respect to the σ-field F_t generated by the full observation of the two processes until time t [23]. The relationship between $\Sigma(t)$ and the diffusion process

$$dp^j = \sum_i \sigma_i^j dx^i + \beta^j dt \tag{3.2}$$

is given by

$$\Sigma^{j,k}(t) = \sum_i \sigma_i^j(t)\sigma_i^k(t). \tag{3.3}$$

For the rest of this chapter, we will focus on the univariate case, and will therefore omit the indices.

The main idea behind the Fourier method is to construct the volatility matrix Σ by applying the classical Fourier-Féjer inversion formula to its Fourier coefficients, which are in turn obtained from the Fourier coefficients of dp.

The Fourier coefficients of dp are defined as

$$a_0(dp) = \frac{1}{2\pi}\int_0^{2\pi} dp(t), \tag{3.4}$$

$$a_k(dp) = \frac{1}{\pi}\int_0^{2\pi} \cos(kt)dp(t), \tag{3.5}$$

$$b_k(dp) = \frac{1}{\pi}\int_0^{2\pi} \sin(kt)dp(t). \tag{3.6}$$

Similarly, the formulae for the Fourier coefficients of the volatility are given by

$$a_0(\Sigma) = \frac{1}{2\pi}\int_0^{2\pi} \Sigma(t)dt, \tag{3.7}$$

$$a_k(\Sigma) = \frac{1}{\pi}\int_0^{2\pi} \cos(kt)\Sigma(t)dt, \tag{3.8}$$

$$b_k(\Sigma) = \frac{1}{\pi}\int_0^{2\pi} \sin(kt)\Sigma(t)dt. \tag{3.9}$$

The instantaneous volatility matrix Σ can then be reconstructed from its Fourier coefficients by the Fourier-Féjer inversion formula [22] given by

$$\Sigma(t) = \lim_{N\to\infty}\sum_{k=0}^{N}(1-\frac{k}{N})(a_k(\Sigma)\cos(kt) + b_k(\Sigma)\sin(kt)), \forall t \in (0, 2\pi). \tag{3.10}$$

The advantage of using this inversion formula instead of any of the many other inversion formulae is that, if Σ is a positive function, the approximation given by Equation 3.10 will again be positive [22]. We refer to 3.10 as the Fourier estimator.

For numerical applications, the formulae for the Fourier coefficients for the return series dp and the volatility Σ need to be discrete.

First, we need to rescale our irregularly spaced observations $[t_1, \ldots t_n]$ into the interval $[0, 2\pi]$ using the formula

$$\tau_j = \frac{2\pi(t_j - t_1)}{(t_n - t_1)}, \quad j = 1, \ldots n. \tag{3.11}$$

The integrals for the Fourier coefficients for dp can then be computed by integration by parts to obtain

$$a_k(dp) = \frac{p(2\pi) - p(0)}{\pi} + \frac{k}{\pi}\int_0^{2\pi} \sin(kt)p(t)dt, \tag{3.12}$$

$$b_k(dp) = -\frac{k}{\pi}\int_0^{2\pi} \cos(kt)p(t)dt. \tag{3.13}$$

This expression is numerically stable, since it involves the integration, rather than the differentiation of p, which is used in most classical methods [22]. Since we are working with discrete financial time series and observations are therefore finite, we need an assumption on how data are connected before we can compute the integrals in Equations 3.12 and 3.13. In [6], the Fourier method was implemented using linearly interpolated prices in the interval $[t_i, t_{i+1}]$, instead of assuming the price to be constant, but it was found that this resulted in a downward biased estimator. This is due to the fact that linear interpolation induces spurious autocorrelation. We will therefore make the assumption that $p(t) = p(t_i)$ in the interval $[t_i, t_{i+1}]$.

With the choice of piecewise constant prices, the integral becomes

$$a_k(dp) \approx \frac{p(2\pi) - p(0)}{\pi} + \frac{1}{\pi} \sum_{i=1}^{N-1} [\cos(kt_i) - \cos(kt_{i+1})] p(t_i), \quad (3.14)$$

$$b_k(dp) \approx \frac{1}{\pi} \sum_{i=1}^{N-1} [\sin(kt_i) - \sin(kt_{i+1})] p(t_i). \quad (3.15)$$

The next step, which is the main result of the paper by [22], is the calculation of the Fourier coefficients of the volatility. We present the theorem and the complete proof, which is an expansion of the proof given by [22].

Theorem 1 *Assume that all measurable economic data follow the Bachelier paradigm (Definition 2.1). Then, for any fixed, strictly positive integer n_0, the Fourier coefficients of the volatility are given by the following formulae:*

$$a_0(\Sigma) = \lim_{S \to \infty} \frac{\pi}{S + 1 - n_0} \sum_{s=n_0}^{S} (a_s^2(dp) + b_s^2(dp)), \quad (3.16)$$

$$a_q(\Sigma) = \lim_{S \to \infty} \frac{2\pi}{S + 1 - n_0} \sum_{s=n_0}^{S} (a_s(dp) a_{s+q}(dp)), \forall q > 0, \quad (3.17)$$

$$b_q(\Sigma) = \lim_{S \to \infty} \frac{2\pi}{S + 1 - n_0} \sum_{s=n_0}^{S} (a_s(dp) b_{s+q}(dp)), \forall q \geq 0. \quad (3.18)$$

Proof 1 Let p_t be a logarithmic asset price that is generated by the diffusion:

$$dp(t) = \beta(t)dt + \sigma(t)dx, \quad 0 \leqq t \geqq T.$$

If we ignore the drift term, this equation simplifies to

$$dp(t) = \sigma(t)dx.$$

This is acceptable, because it implies an efficient market [19], and also because it can be proved that the contribution of the drift term to the formulae 3.16, 3.17 and 3.18 is zero [22].

We introduce the Gaussian variables

$$G_k(t) := a_k(dp(t)), \quad G'_k(t) := b_k(dp(t)).$$

From a corollary to the Martingale Representation theorem, we know that $G_k(t)$ is a martingale and from the properties of the stochastic integral, we know that $G_k(t)$ is a Brownian motion [32]. Therefore, $G_k(t)$ is a Gaussian random variable, and $E(G_k(t)) = 0$.

Then

$$\begin{aligned} a_k(dp) &= \frac{1}{\pi}\int_0^{2\pi} \cos(kt)dp(t) \\ &= \frac{1}{\pi}\int_0^{2\pi} \cos(kt)\sigma(t)dx. \end{aligned}$$

Substituting this into the expression for the covariance, we get

$$\begin{aligned} E(G_kG_l) &= E(a_k(dp)a_l(dp)) \\ &= E[\frac{1}{\pi^2}\int_0^{2\pi} \cos(kt)\sigma(t)dx \int_0^{2\pi} \cos(ls)\sigma(s)dx]. \end{aligned}$$

We substitute the stochastic integrals with their simple function approximations

$$\begin{aligned} \int_0^{2\pi} \cos(kt)\sigma(t)dx(t) &= \int \cos(kt)\sigma(t)I_{[0,t]}(t)dx(t) \\ &= \sum_i \cos(ki)\sigma(i)(x(t_{i+1}) - x(t_i)). \end{aligned}$$

Thus, the value of the Itô stochastic integral is the Riemann sum of the path of $\cos(kt)\sigma(t)$, evaluated at the left end points of the intervals $[t_{i-1}, t_i]$, with respect to Brownian motion [24]. The expression for the covariance now takes the form

$$\begin{aligned} E(G_kG_l) &= E[\frac{1}{\pi^2}(\sum(\ldots)dx)(\sum(\ldots)dx)] \\ &= E[\frac{1}{\pi^2}(\sum(\ldots)dx.dx)]. \end{aligned}$$

Using the Itô multiplication rules (See [2]):

$$\begin{aligned} dt.dt &= (dt)^2 = 0, \\ dt.dx_r &= 0, \\ dx_r.dx_q &= 0, \quad \text{for} \quad r \neq q \\ dx_r.dx_q &= dt, \quad \text{for} \quad r = q, \end{aligned}$$

the equation simplifies to

$$\begin{aligned} E(G_k G_l) &= E[\frac{1}{\pi^2}(\sum(\sigma^2 \cos(kt)\cos(lt))dt)] \\ &= \frac{1}{\pi^2}\int_0^{2\pi} \Sigma(t)\cos(kt)\cos(lt)dt. \end{aligned}$$

Using the trigonometric identity

$$\cos(kt)\cos(lt) = \frac{1}{2}(\cos(k-l)t + \cos(k+l)t)$$

and the definition of the Fourier coefficients of the volatility, we obtain the following expression for the covariance:

$$\begin{aligned} E(G_k G_l) &= \frac{1}{\pi^2}\int_0^{2\pi} \Sigma(t)\cos(kt)\cos(lt)dt \\ &= \frac{1}{2\pi^2}\int_0^{2\pi} \Sigma(t)[(\cos(k-l)t + \cos(k+l)t)]dt \\ &= \frac{1}{2\pi^2}\int_0^{2\pi} \Sigma(t)\cos(k-l)tdt + \frac{1}{2\pi^2}\int_0^{2\pi} \Sigma(t)\cos(k+l)tdt \\ &= \frac{1}{2\pi}(a_{|k-l|}(\Sigma) + a_{k+l}(\Sigma)). \end{aligned} \tag{3.19}$$

For $q > 0$, consider the random variables

$$U_S^q := \frac{1}{S}\sum_{k=1}^{S} G_k G_{k+q}.$$

Using (3.19) we get

$$\begin{aligned} E(U_S^q) &= E(\frac{1}{S}\sum_{k=1}^{S} G_k G_{k+q}) \\ &= \frac{1}{S}\sum_{k=1}^{S} E(G_k G_{k+q}) \\ &= \frac{1}{S}\sum_{k=1}^{S} \frac{1}{2\pi}(a_{|k-k-q|}(\Sigma) + a_{k+k+q}(\Sigma)) \\ &= \frac{1}{S}\sum_{k=1}^{S} \frac{1}{2\pi}(a_q(\Sigma) + a_{2k+q}(\Sigma)) \\ &= \frac{1}{2\pi}a_q(\Sigma) + \frac{1}{S}\sum_{k=1}^{S} \frac{1}{2\pi}a_{2k+q}(\Sigma) \\ &\approx a_q(\Sigma) + R_S, \end{aligned}$$

where $|R_S| = \frac{1}{S}|\sum_{k=1}^{S} a_{2k+q}(\Sigma)|$.

Now,

$$\frac{1}{S}\sum_{k=1}^{S} 1.a_k(\Sigma) \leqq \frac{1}{S}(\sum_{k=1}^{S} 1)^{\frac{1}{2}}(\sum_{k=1}^{S} a_k^2(\Sigma))^{\frac{1}{2}} \tag{3.20}$$

$$= \frac{1}{\sqrt{S}}(\sum_{k=1}^{S} a_k^2(\Sigma))^{\frac{1}{2}} \tag{3.21}$$

$$= \leqq \frac{1}{\sqrt{S}}||\Sigma||_{L^2} \tag{3.22}$$

and hence the LHS converges to 0 as $S \to 0$. Therefore, $R_S \to 0$.
Next, we want to compute

$$E((U_S^q)^2) = \frac{1}{S^2}\sum_{0\leq k,k'\leq S} E(G_k^2 G_{k'+q}^2).$$

Consider an $\mathbf{R^2}$-valued normal variable (G_1, G_2) and denote

$$\lambda_i := E(G_i^2), \quad \mu := E(G_1 G_2),$$

and define $Z := G_2 - \frac{\mu}{\lambda_1}G_1$.

By substituting the definition of Z into the equation, we can see that $E(G_1 Z) = 0$, which implies that the covariance of $G_1 Z$ is zero, which in turn implies that G_1 and Z are independent [24]. Using this fact, as well as the fact that $E(G_1) = E(G_2) = 0$, we get

$$\begin{aligned}
E(G_1^2 G_2^2) &= E(G_1^2(Z^2 + 2\frac{\mu}{\lambda_1}G_1 Z + \frac{\mu^2}{\lambda^2}G_1^2)) \\
&= E(G_1^2 Z^2 + \frac{2\mu}{\lambda_1}G_1^3 Z + \frac{\mu^2}{\lambda_1^2}G_1^4) \\
&= E(G_1^2 Z^2) + \frac{\mu^2}{\lambda_1^2}E(G_1^4) \\
&= E(G_1^2)E(Z^2) + \frac{\mu^2}{\lambda_1^2}E(G_1^4) \\
&= E(G_1^2)E(G_2^2 - \frac{2\mu}{\lambda_1}G_1 G_2 + \frac{\mu^2}{\lambda_1^2}G_1^2) + \frac{\mu^2}{\lambda_1^2}E(G_1^4) \\
&= E(G_1^2)E(G_2^2) - 2\frac{\mu}{\lambda_1}E(G_1^2)E(G_1 G_2) \\
&\quad + \frac{\mu^2}{\lambda_1^2}E(G_1^2)E(G_1^2) + \frac{\mu^2}{\lambda_1^2}E(G_1^4).
\end{aligned} \tag{3.23}$$

Now, using the characteristic function for Gaussian variables

$$\begin{aligned} E(\exp(i\xi G_k(t)) &= \exp(-\frac{1}{2}\xi^2\sigma^2 + i\xi\eta) \\ &= \exp(-\frac{1}{2}\xi^2 E(G_k(t)^2)), \end{aligned}$$

where $G_k(t) \approx N(\eta, \sigma^2)$, $\sigma^2 = E(G_k(t)^2)$ and $\eta = 0$.

Using the Taylor expansion

$$\exp(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots + \frac{x^n}{n!},$$

the left-hand side of Equation 3.24 becomes

$$E(1 + i\xi G_k - \frac{1}{2}\xi^2 G_k^2 - \frac{i\xi^2 G_k^3}{3!} + \frac{1}{4!}G_k^4 + \ldots) \tag{3.24}$$

and the right-hand side becomes

$$\begin{aligned} &(1 - \frac{1}{2}\xi^2 E(G_k^2) + \frac{(-\frac{1}{2}\xi^2 E(G_k^2))^2}{2} + \ldots \\ &= 1 - \frac{1}{2}\xi^2 E(G_k^2) + \frac{1}{8}\xi^4 E(G_k^2)^2 + \ldots \quad . \end{aligned} \tag{3.25}$$

Equating Equations 3.24 and 3.25, we see the following:

- The first terms of both sides are equal to one.
- The second term of the left hand side is zero.
- The third term of the left hand side is equal to the second term of the right hand side.
- The fourth term of the left hand side is zero.
- The fifth term of the left hand side is equal to the third term of the right hand side.

This leaves us with

$$\begin{aligned} E(\frac{1}{24}\xi^4 G_k^4) &= \frac{1}{8}\xi^2 E(G_k^2)^2 \\ \therefore E(G_k^4) &= 3E(G_k^2)^2. \end{aligned}$$

If we now revisit Equation 3.23, we see that

$$E(G_1^2 G_2^2) = E(G_1^2)E(G_2^2) - 2\frac{\mu}{\lambda_1}E(G_1^2)E(G_1 G_2)$$

$$
\begin{aligned}
&+\frac{\mu^2}{\lambda_1^2}E(G_1^2)E(G_1^2)+\frac{\mu^2}{\lambda_1^2}E(G_1^4)\\
&= E(G_1^2)E(G_2^2)-2\frac{\mu}{\lambda_1}E(G_1^2)E(G_1G_2)\\
&\quad +4\frac{\mu^2}{\lambda_1^2}E(G_1^2)E(G_1^2)\\
&= E(G_1^2)E(G_2^2)-2\frac{\mu}{\lambda_1}\lambda_1\mu+4\frac{\mu^2}{\lambda_1^2}\lambda_1\lambda_1\\
&= E(G_1^2)E(G_2^2)+2\mu^2.
\end{aligned}
$$

Putting all of this together, we get

$$
\begin{aligned}
E((U_S^q-E(U_S^q))^2) &= E((U_S^q)^2-2E(U_S^q)E(U_S^q)+(E(U_S^q))^2)\\
&= E((U_S^q)^2)-(E(U_S^q))^2\\
&= (\frac{1}{S^2}\sum_{0\le k,k'\le S}E(G_k^2G_{k'+q}^2))-a_q(\Sigma)^2\\
&= \frac{1}{S^2}\sum_{0\le k,k'\le S}(E(G_k^2)E(G_{k'+q}^2)+2(E(G_kG_{k'+q}))^2-(a_q(\Sigma))^2)\\
&= \frac{1}{S^2}\sum_{0\le k,k'\le S}(E(G_k^2)E(G_{k'+q}^2)+\frac{1}{2\pi^2}(a_{k-k'+q}(\Sigma)+a_{k+k'+q}(\Sigma))^2-(a_q(\Sigma))\\
&= \frac{1}{S^2}\sum_{0\le k,k'\le S}(E(U_S^q))^2+\frac{1}{2\pi^2}(a_{k-k'+q}(\Sigma)+a_{k+k'+q}(\Sigma))^2-(a_q(\Sigma))^2\\
&= \frac{1}{S^2}\sum_{0\le k,k'\le S}(a_q(\Sigma))^2+\frac{1}{2\pi^2}(a_{k-k'+q}(\Sigma)+a_{k+k'+q}(\Sigma))^2-(a_q(\Sigma))^2\\
&= \frac{1}{2\pi^2N^2}\sum_{0\le k,k'\le S}(a_{|k-k'+q|}(\Sigma)+a_{k+k'+q}(\Sigma))^2\\
&\le \frac{1}{S}||\Sigma||_{L^2}^2.
\end{aligned}
\tag{3.26}
$$

Since $\lim_{S\to\infty}\frac{1}{S}||\Sigma||_{L^2}^2=0$ it follows that $\lim_{S\to\infty}(U_S^q-E(U_S^q))=0$ a.s., and since $\lim_{S\to\infty}E(U_S^q)=a_q(\Sigma)$ (as we have previously proved) (3.17) follows.

The proofs for *(3.18)* and *(3.16)* can be derived in a similar manner.

By polarisation of the one-dimensional result, the Fourier estimator can be extended to the multivariate case. The following formulae shows how the Fourier coefficients of the volatility matrix Σ_{ij} can be calculated:

$$
a_0(\Sigma_{ij}) = \lim_{S\to\infty}\frac{\pi}{S+1-n_0}\sum_{s=n_0}^{S}(a_s(dp_i)a_s(dp_j)+b_s(dp_i)a_s(dp_j)),
$$

$$a_k(\Sigma_{ij}) = \lim_{S\to\infty} \frac{\pi}{S+1-n_0} \sum_{s=n_0}^{S} \frac{1}{2}(a_s(dp_i)a_{s+k}(dp_j) + a_s(dp_j)a_{s+k}(dp_i)), \tag{3.27}$$

$$b_k(\Sigma_{ij}) = \lim_{S\to\infty} \frac{\pi}{S+1-n_0} \sum_{s=n_0}^{S} \frac{1}{2}(a_s(dp_i)b_{s+k}(dp_j) + a_s(dp_j)a_{s+k}(dp_i)). \tag{3.28}$$

The coarseness of the estimator is governed by the value of S. Choosing a higher value for S corresponds to choosing a finer grid when using realised volatility. By including only the lowest S frequencies, high frequency noise can be ignored. In theory, this therefore renders this estimator invariant to market micro-structure effects. Note that, while not all the Fourier coefficients are used in the calculation of the volatility, all the Fourier coefficients are used in the calculation of the return series, i.e. no data is thrown away.

According to Reno and Barucci [6], the smallest wavelength that can be evaluated when computing the Fourier coefficients of the return series dp to avoid aliasing effects is twice the smallest distance between two consecutive prices. In the case of equally spaced data, it will correspond to $k = \frac{N}{2}$. This is known as the Nyquist frequency [1]. The first n_0 terms can also be omitted, since this makes the estimator too sensitive to the drift term [22]. Iori [34] notes that the highest wave harmonic that can be analysed is $S = \frac{N}{\tau}$, where τ denotes the lower bound of the time gap between two consecutive trades. In the case of a highly liquid market, τ will be equal to 1, which indicates that the lower bound between trades is equal to one second. This is equivalent to using the Nyquist frequency.

It is important to note that the volatility given by the Féjer inversion formula is the instantaneous volatility. Most nonparametric methods proposed in the literature focus on integrated volatility. As shown herein before, it is possible to reconstruct the integrated volatility from the instantaneous volatility using the formula

$$\sigma^2_{INT} = \int_0^{2\pi} \sigma^2(t)dt. \tag{3.29}$$

As a matter of fact, by the identity

$$2\pi a_0(\sigma^2) = \int_0^{2\pi} \sigma^2(t)dt \tag{3.30}$$

[1]The Nyquist frequency is half the sampling frequency of a discrete signal processing system. It is sometimes called the critical frequency.

and using the definition of a_0, we have

$$\sigma^2_{INT} = a_0(\sigma^2) = \lim_S \frac{\pi}{S+1-n_0} \sum_{s=n_0}^{S} a_s^2(dp) + b_s^2(dp). \tag{3.31}$$

Similarly, in the multivariate case, this formula can be used to reconstruct the instantaneous volatility matrix Σ_{ij} by the Fourier-Féjer inversion formula.

To calculate the integrated volatility matrix Σ^{ij}_{INT} we use the fact that

$$\begin{aligned} \Sigma^{ij}_{INT} &= 2\pi a_0(\Sigma_{ij}) \\ &= \frac{\pi^2}{S+1-n_0} \sum_{s=1}^{S} (a_s(dp_i)a_s(dp_j) + b_s(dp_i)b_s(dp_j)). \end{aligned} \tag{3.32}$$

3.2 Extension of the Fourier method

Although satisfactory results were obtained from the Fourier method proposed in [22], it has the disadvantage that positivity of the instantaneous volatility function is not necessarily maintained. In an extension to this method discussed in [23], this point is addressed by providing an expansion of the volatility function as a series of positive trigonometric polynomials.

In this method, the formula for the Fourier coefficients of the return series are prolongated to all integers k by parity for a_k and by imparity for b_k. The new formulae are given by

$$a_0^\star = b_0^\star = 0, \tag{3.33}$$

$$a_k^\star = \begin{cases} a_k(dp) & \text{for } k > 0 \\ a_{-k}(dp) & \text{for } k < 0 \end{cases} \tag{3.34}$$

$$a_b^\star = \begin{cases} b_k(dp) & \text{for } k > 0 \\ -b_{-k}(dp) & \text{for } k < 0 \end{cases} \tag{3.35}$$

The following theorem, proved by Malliavin and Mancino [23], shows how the Fourier coefficients for the volatility can be constructed from the Fourier coefficients for the return series.

Theorem 2 *Consider the process p satisfying the Bachelier paradigm and $p(0) = p(2\pi)$. Define, for $0 \leq q \leq 2N$,*

$$\alpha_q(N) = \frac{1}{2N+1} \sum_{-N}^{N-q} (a^\star_{q+s}a^\star_s + b^\star_{q+s}b^\star_s),$$

$$\beta_q(N) = \frac{1}{2N+1} \sum_{-N}^{N-q} (-a^\star_{q+s}b^\star_s + b^\star_{q+s}a^\star_s).$$

Then the trigonometric polynomial having coefficients $\alpha_q(N), \beta_q(N), 0 \leq q \leq 2N$, *is positive. Denote by* $a_q(\sigma^2), b_q(\sigma^2)$ *the Fourier coefficients of* $\sigma^2(t)$. *Then, for all fixed* $q \leq 0$, *the following convergence in probability holds:*

$$\lim_{N \to +\infty} \alpha_q(N) = \frac{1}{\pi} a_q(\sigma^2), \quad \lim_{N \to +\infty} \beta_q(N) = \frac{1}{\pi} b_q(\sigma^2).$$

From these Fourier coefficients, the function $\sigma^2(t)$ can once again be reconstructed using the Fêjer inversion formula.

Note that this extension to the Fourier method influences only the instantaneous volatility, but not the integrated volatility. Both methods provide equal, positive estimates for the integrated volatility.

3.3 Scaling

One important concept that has not been mentioned so far is the fact that users may want their historical volatility in scaled form, i.e. there could be a need to calculate the expected volatility over a certain time interval from the returns of another time interval, for example, calculating annualised volatility from hourly returns. This issue is discussed in [25]. They state that, through a Gaussian scaling law, $v^2 \propto dt$, where v is defined as follows:

$$v = [\frac{1}{n} \sigma_{INT}]^{\frac{1}{2}}, \tag{3.36}$$

where $\sigma^\star$ is the integrated volatility, n is the number of observations and dt is the time step of the returns used in the calculation of the volatility. The following definition of scaled volatility is then obtained:

$$v_{scaled} = \sqrt{\frac{\Delta t_{scale}}{\Delta t}} v, \tag{3.37}$$

where Δt_{scale} is the time interval of the expected volatility. If we want to calculate annualised volatility, it can be calculated as follows:

$$v_{annualised} = \sqrt{\frac{1\text{year}}{\Delta t}} v. \tag{3.38}$$

If the required volatility should be in percent, the result obtained from the above formulae should be multiplied by hundred percent. As noted in the previous section, the coarseness of the the Fourier estimator is controlled by S, the user-defined number of Fourier coefficients included in the estimation [26]. S is related to M for realised volatility by $S = M/2$. Therefore, if we are measuring intraday volatility using the Fourier estimator, dt can be determined from S and the volatility in percentage calculated using Equation 3.37.

Chapter 4

Estimation of univariate integrated volatility

In this chapter, the estimation of univariate volatility when high frequency data is available is considered. In particular, the behaviour of the univariate Fourier estimator is examined in comparison with the classical method of realised volatility, which was discussed in Chapter 2. Firstly, the methods are evaluated using simulated evenly spaced and high frequency data series with known parameters. The behaviour of the estimator when applied to empirical data is then examined. The multivariate case is considered in the next chapter.

4.1 Simulation Setup

In this section, we describe the simulation setup used in the comparison of the different volatility estimation methods. The objective of the simulation exercise is to determine whether the Fourier estimator performs well in comparison with classical methods.

For all the simulations, data is generated using a straightforward diffusion process. To simplify the analysis, we assume that there is no drift. This is assumption is acceptable, because it implies an efficient market [19], and also because it can be proved that the contribution of the drift term to the formulae 3.16, 3.17 and 3.18 is zero [22].
This price model can be simulated by the differential equation

$$dp(t) = \sigma(t)p(t)dW(t) \tag{4.1}$$

in the following way:

$$\begin{aligned} p_1 &= 100, \\ p_{i+1} &= p_i.e^{r_i}, \\ r_i &\approx N(0, \sigma_i). \end{aligned} \tag{4.2}$$

We then evaluate $u(t) = \log(p(t))$.

Since we are interested in the behaviour when volatility is constant, we simulate the price model given by Equation 4.1 with a constant standard deviation of σ. Note that, in the same way that the relationship between monthly and annual volatility is given by $\sigma_{monthly} = \sqrt{12}\sigma_{annual}$, taking N data points with a volatility σ is equivalent to taking $\sigma_i = \sigma\sqrt{\frac{2\pi}{N}}$ in Equation 4.2.

To simulate the high frequency, unevenly spaced data from the evenly spaced simulated time series generated by Equation 4.2, a sample series using exponentially distributed interval sizes is extracted. The choice is motivated by the fact that the empirical distribution of $t_i - t_{i-1}$ can be approximated with an exponential shape [6]. This is discussed in more detail in Section 4.4.1. This process is then repeated for the equivalent of N days, which results in a different number of irregularly spaced observations each day. Different values for the exponential mean in the extraction are used to simulate stocks with different liquidity levels.

The relevance of the results on the volatility of the Monte Carlo simulations of the price process presented here obviously depends on how good the model used to simulate the price process duplicates the properties of empirical data. Yet, a direct comparison is impossible precisely because the volatility of the empirical data is an unknown quantity. Although a direct check of the model is not possible, we also include tests on empirical data to evaluate the behaviour of the estimator.

4.2 Evaluation Methodology

In our first simulation, discussed in Section 4.3.1, we evaluate some of the assumptions made during the derivation of the Fourier method. In particular, we examine the accuracy of the estimation of the Fourier coefficients of the return series obtained from the discretised formula proposed by Malliavin and Mancino [22] and shown in Equations 3.14 and 3.15. The results are compared with those obtained from using trapezoidal integration.

The results of the Fourier estimator of the integrated volatility using both evenly spaced (Section 4.3.2) and high frequency data (Section 4.3.3) are then discussed. The effect of the different parameters included in the estimation is also examined. This includes the effect of the size of the data series, denoted by N, as well as the liquidity of the series determined by the value of β, the exponential mean used during the generation of the unevenly spaced data. In addition, we look at the performance of the estimator over different time scales, i.e. different number of Fourier coefficients included in the estimation, denoted by S.

To evaluate the performance of the Fourier estimator within the scenarios described above, a method proposed by Hog and Lunde [9] is used, where

the standardised errors of the estimators are evaluated. The formula for this statistic is given by

$$\pi_n = \frac{\sigma_n^2 - \text{true volatility}}{\text{true volatility}}, \tag{4.3}$$

where σ_n^2 denotes the estimated volatility.
When simulating an evenly spaced time series with constant volatility, calculating the true simulated volatility is straightforward. However, since true volatility is unobservable, evaluating the performance of a volatility model is not always as clear-cut. It is, however, possible to construct benchmark criteria to assess the performance of different estimators. The equation

$$\text{true volatility} \approx \sum_{t=2}^{N} [p_d(t) - p_d(t-1)]^2 \tag{4.4}$$

is often used to estimate the 'true' integrated volatility on day d, where p_d is the original evenly spaced data series generated before the exponentially distributed observations are extracted and N is the number of observations. Naturally, if evenly spaced data is used, this 'true' volatility is equal to the realised volatility estimator.

The mean and root mean squared errors (RMSE) of π_d is then used as evaluation criteria, which can be interpreted as relative bias and RMSE of the estimators respectively. They are defined as

$$Bias = \frac{1}{N} \sum_{n=1}^{N} \pi_n \tag{4.5}$$

and

$$RMSE = [\frac{1}{N} \sum_{n=1}^{N} \pi_n^2]^{\frac{1}{2}}, \tag{4.6}$$

where N is the number of days included in the estimation.

4.3 Results on Simulated Data

In this section, we present the results of the study on univariate simulated data. Section 4.3.1 covers the results on the accuracy of the discretised formulae used in the calculation of the Fourier coefficients of the returns series, Section 4.3.2 considers the results on evenly spaced data while Section 4.3.3 covers the results on high frequency data.

4.3.1 Evaluation of the discretised formula in the Fourier method

From Chapter 3, we know that the Fourier coefficients of the return series dp are given by the formulae

$$\begin{aligned} a_0(dp) &= \frac{1}{2\pi}\int_0^{2\pi} dp(t), \\ a_k(dp) &= \frac{1}{\pi}\int_0^{2\pi} \cos(kt)dp(t), \\ b_k(dp) &= \frac{1}{\pi}\int_0^{2\pi} \sin(kt)dp(t). \end{aligned}$$

As proposed by Malliavin and Mancino [22] and discussed in the previous chapter, the integrals for the Fourier coefficients for dp can be computed by integration by parts and, when making the assumption that prices are piecewise constant, we obtain

$$\begin{aligned} a_k(dp) &\approx \frac{p(2\pi)-p(0)}{\pi} + \frac{1}{\pi}\sum_{i=1}^{N-1}[\cos(kt_i)-\cos(kt_{i+1})]p(t_i), \\ b_k(dp) &\approx \frac{1}{\pi}\sum_{i=1}^{N-1}[\sin(kt_i)-\sin(kt_{i+1})]p(t_i). \end{aligned}$$

The performance of this approximation is evaluated in this section using time series with constant volatility. For comparison, the integrals were also calculated using trapezoidal integration.

The same testing methodology as proposed by Barucci, Mancino and Reno [8] is used. In this paper, the distribution of the Fourier coefficients are examined in the evaluation of the suitability of the approximation, making use of the fact that the theoretical distribution of the Fourier coefficients are given by

$$a_k(dp), b_k(dp) \approx N(0, \frac{\sigma}{\sqrt{\pi}}), \quad k \geq 1. \tag{4.7}$$

The results of the Monte Carlo simulation study, consisting of a thousand iterations with N = 100 and $\sigma = 0.01$ are shown in Figure 4.1. From this figure we can see that the Fourier coefficients for the return series, denoted by a_k, are normally distributed with a mean consistent with zero and standard deviation consistent with $\frac{\sigma}{\sqrt{\pi}} = 0.0056$. This indicates that the mathematical approximations made do not destroy normality or the expected standard deviation.

In addition, from Barucci, Mancino and Reno [8], we know that the theoretical distribution of the Fourier coefficients of the volatility are given by

$$a_k(\Sigma), b_k(\Sigma) \approx (0, 8\frac{\sigma^4}{N}), \quad k \geq 1. \tag{4.8}$$

Figure 4.2, where the results of the Monte Carlo experiment are displayed, confirms the suitability of this approximation by showing that the coefficients are normally distributed with mean consistent with zero and standard deviation consistent with $8\frac{\sigma^4}{N}$. We do, however, note that this approximation cannot be used in cases where k is a multiple of $N-1$, since the frequency of the cosine wave used in the approximation coincides with the evenly spaced points in time. This will cause the result to always be equal to zero, which is not true for the theoretical case.

In addition to the above approximation, tests were also performed using the trapezoidal method of integration. This was done using the Matlab function 'trapz' to calculate the integrals. This method provides satisfactory results for values of k where $\frac{N}{10} \lesssim k \lesssim N - \frac{N}{10}$, but is not as stable for values of k outside of these boundaries. Therefore, for the rest of this study, as is done in most of the literature on this topic, the discretised formula for evaluating the Fourier coefficients, given by Equation 4.7, is used.

In [8], the distribution of the Fourier estimators for the return series and the volatility is investigated. Here they show that the theoretical volatility of $a_0(\Sigma)$, which gives the estimator for the integrated volatility, is given by

$$Var[a_0(\Sigma)] = 2\frac{\sigma^2}{N}. \tag{4.9}$$

Figure 4.3 shows the volatility of the distribution is obtained from the Monte Carlo samples for different choices of N, compared to the expected volatility as a function of N. It shows that the function given by Equation 4.9 gives a satisfactory explanation of the observed function. Barucci, Mancino and Reno [8] remark that the precision of the estimator is given by $\approx \frac{\sqrt{2}}{\sqrt{N}}$.

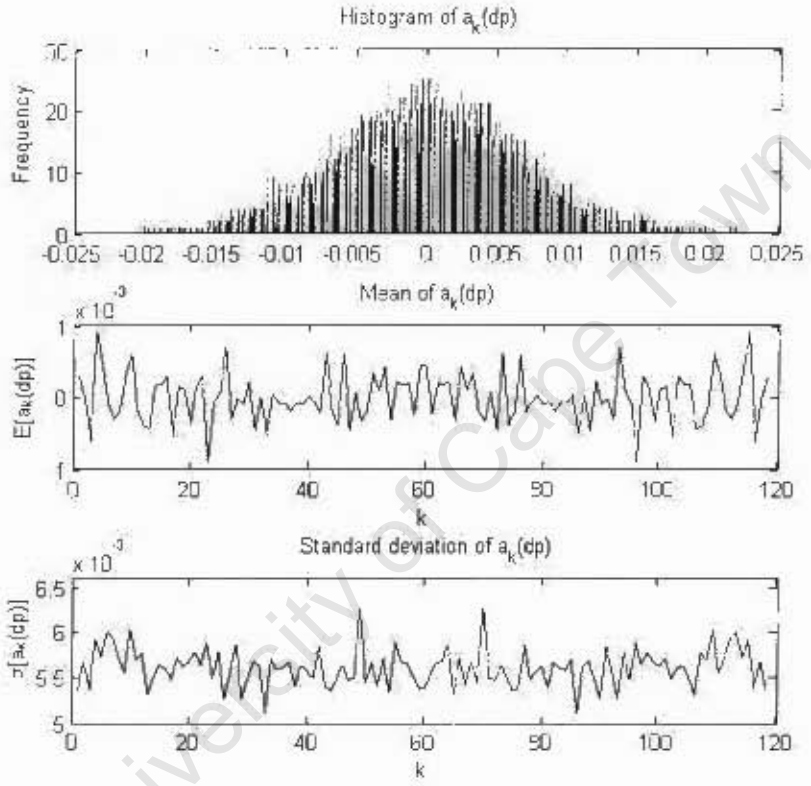

Figure 4.1: Results of the estimation of $a_k(dp)$ when using the discretised formula shown in Equation 4.7 for the case where $N = 100, \sigma = 0.01$ and $1 \leq k \leq 100$. This shows that the mathematical approximations do not destort the normality and standard deviation. a) *Histogram showing distribution of* a_k: This figure confirms that the coefficients are normally distributed. Note that this is a standard histogram and the colour variation does not carry meaning. b) *Mean of* a_k: The mean is consistent with zero. c) *Standard deviation of* a_k: The standard deviation is consistent with $\frac{\sigma}{\sqrt{\pi}} = 0.0056$.

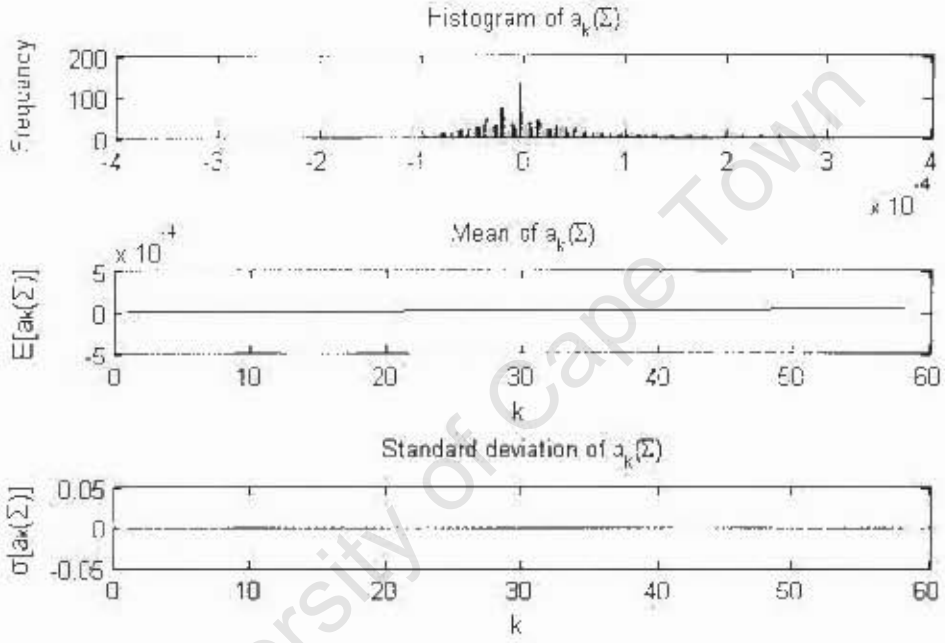

Figure 4.2: Results of the estimation of $a_k(\Sigma)$, calculated by Equation 3.16, when previous-tick interpolation was used for the calculation of the Fourier coefficients of the return series with $N = 100, \sigma = 0.01$ and $1 \leq k \leq 100$. This shows that the mathematical approximations do not destort the normality and standard deviation. a) *Histogram showing distribution of* $a_k(\Sigma)$: This figure shows that the coefficients are normally distributed. Note that this is a standard histogram and the colour variation does not carry meaning. b) *Mean of* $a_k(\Sigma)$: The mean is consistent with zero. c) *Standard deviation of* $a_k(\Sigma)$: The standard deviation is consistent with $8\frac{\sigma^4}{N} \approx 0$.

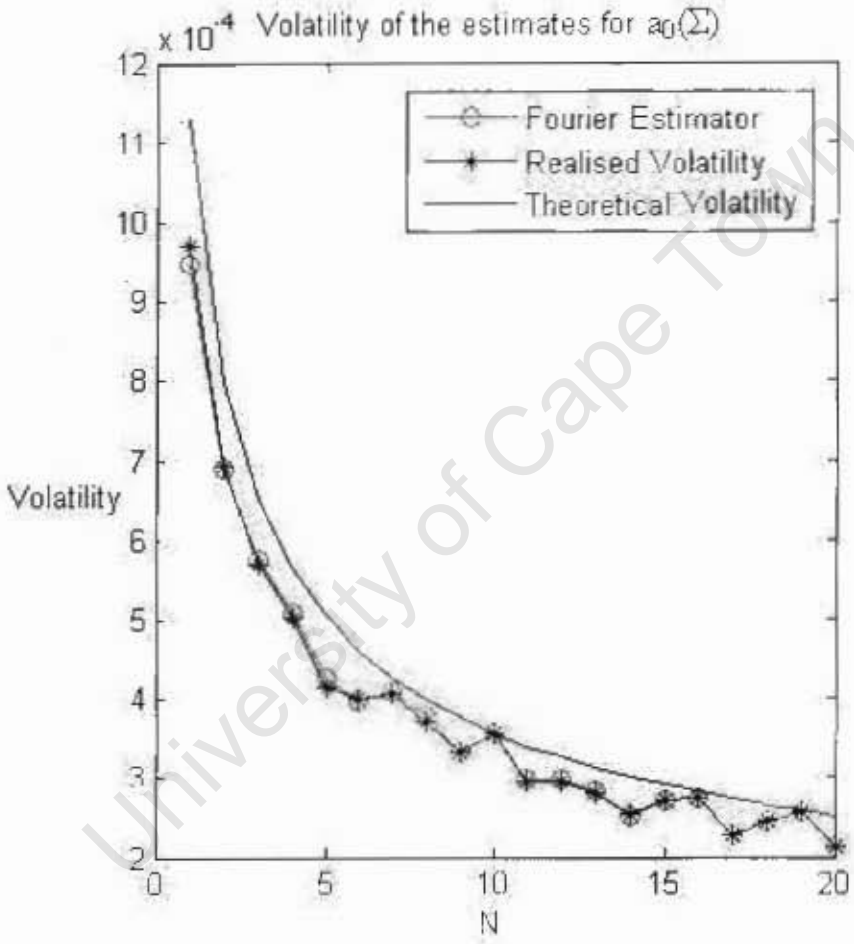

Figure 4.3: Volatility of the estimated values for $a_0(\Sigma)$ in comparison with the theoretical volatility, given by $2\frac{\sigma^2}{N}$, when a Monte Carlo study with evenly spaced data and different values for N, the size of the data series, is used.

4.3.2 Analysis of simulated evenly spaced data

In this section, the results of the volatility estimators are examined when evenly spaced simulated data is used. In the next section, high frequency simulated data is considered.

In our first simulation, we look at the effect of the size of the dataseries, denoted by N, on the results of the different estimators. This is done through a Monte Carlo study where volatility is constant. The Nyquist frequency is used to determine the number of Fourier coefficients included in the estimation. The results of this simulation are shown in Table 4.1 and Figure 4.4.

Table 4.1 shows the relative bias (Equation 4.5) and the RMSE (Equation 4.6) of the estimators when looking at 1000 replications using evenly spaced data of variable size N, ranging between 100 and 1000 data points. Here we can see that the accuracy of both estimators increase with the size of N and that the Fourier estimator gives slightly better results than the realised volatility estimator when looking at both bias and RMSE. Both estimators have a slight negative bias.

Table 4.1: Sample bias and RMSE of the Fourier estimator and realised volatility for different values of N when using evenly spaced data with constant volatility.

	Bias		RMSE	
N	Fourier	RV	Fourier	RV
100	-0.0073	-0.0112	0.0722	0.0724
200	-0.0066	-0.0093	0.0520	0.0517
300	-0.0017	-0.0032	0.0414	0.0411
400	-0.0031	-0.0042	0.0328	0.0330
500	-0.0025	-0.0034	0.0329	0.0329
600	-0.0005	-0.0013	0.0293	0.0294
700	-0.0013	-0.0018	0.0254	0.0254
800	-0.0013	-0.0018	0.0254	0.0254
900	-0.0005	0.0000	0.0245	0.0246
1000	-0.0005	-0.0010	0.0232	0.0232

In this simulation, we used a constant volatility of $\sigma = 0.01$. Tests using different values for the constant volatility showed that higher constant volatility does not imply less accuracy (higher RMSE). Note that this would not necessarily have been the case if linear interpolation was used (See Section 2.1 or [20]).

Figure 4.4 shows the distribution of the standardised errors, calculated by Equation 4.4, for the case where $N = 1000$, for the Fourier estimator

and realised volatility. The kurtosis, skewness and standard deviation of the distribution are also shown in this figure. From this we see that the Fourier estimator performs well in comparison with realised volatility. The difference in performance is, however, relatively small.

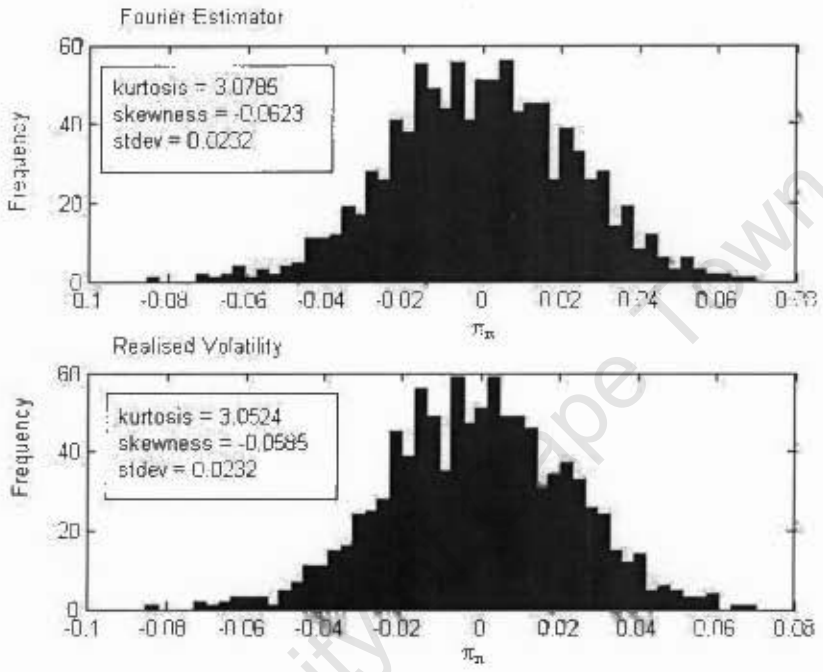

Figure 4.4: Distribution of the standardised errors, π_n, between the true volatility and the Fourier estimator and realised volatility respectively when volatility is constant and the data is evenly spaced with N = 1000. The kurtosis, skewness and standard deviation of the relative errors are also shown.

4.3.3 Analysis of simulated high frequency data

In Table 4.2, the results of the Monte Carlo study using simulated high frequency data with constant volatility is presented. As described in Section 4.1, the unevenly spaced data series is created by extracting data from an evenly spaced data series with exponentially distributed intervals with mean denoted by γ. Therefore, the smaller γ, the more liquid the stock. The table shows the bias and RMSE obtained from the Fourier estimator and realised volatility for different values of γ.

Table 4.2: Sample bias and RMSE of the Fourier estimator and realised volatility for different values of γ when using high frequency data with constant volatility.

	Bias		RMSE	
γ	Fourier	RV	Fourier	RV
14.0000	-0.0006	-0.0002	0.0255	0.0370
34.0000	-0.0048	-0.0037	0.0405	0.0563
54.0000	-0.0015	-0.0038	0.0521	0.0695
74.0000	-0.0021	-0.0052	0.0600	0.0807
94.0000	-0.0038	-0.0062	0.0673	0.0914
114.0000	-0.0083	-0.0060	0.0740	0.0992
134.0000	-0.0053	-0.0096	0.0790	0.1059
154.0000	-0.0050	-0.0082	0.0875	0.1206
174.0000	-0.0082	-0.0119	0.0925	0.1234
194.0000	-0.0091	-0.0162	0.0982	0.1303
214.0000	-0.0131	-0.0105	0.0996	0.1349
234.0000	-0.0123	-0.0177	0.1059	0.1402

Different values for the mean γ of the exponential distribution are used to simulate stocks with different liquidity levels. From Table 4.2, we can see that both estimators are more accurate for smaller values of γ. This is an important observation, since we know that the South African market is less liquid than, for example, the S&P100, which implies that results in emerging markets, such as South Africa, could be less accurate. This result is not unexpected, since a smaller value of γ will result in a larger number of data points, which was already shown to increase the accuracy of both estimators when using evenly spaced data. In addition, we can see that the Fourier estimator outperforms the realised volatility estimator when looking at both the bias and RMSE for different values of γ.

Figure 4.4 shows the distribution of the standardised errors, calculated by Equation 4.4, for the case where $N = 1000$ and γ is 423 for the Fourier estimator and realised volatility. This value for γ was chosen, since it is the mean time between ticks for the dataset under question in this paper. The kurtosis, skewness and standard deviation of the distribution are also shown

in this figure. From this we see that the Fourier estimator outperforms the realised volatility in the case of high frequency data.

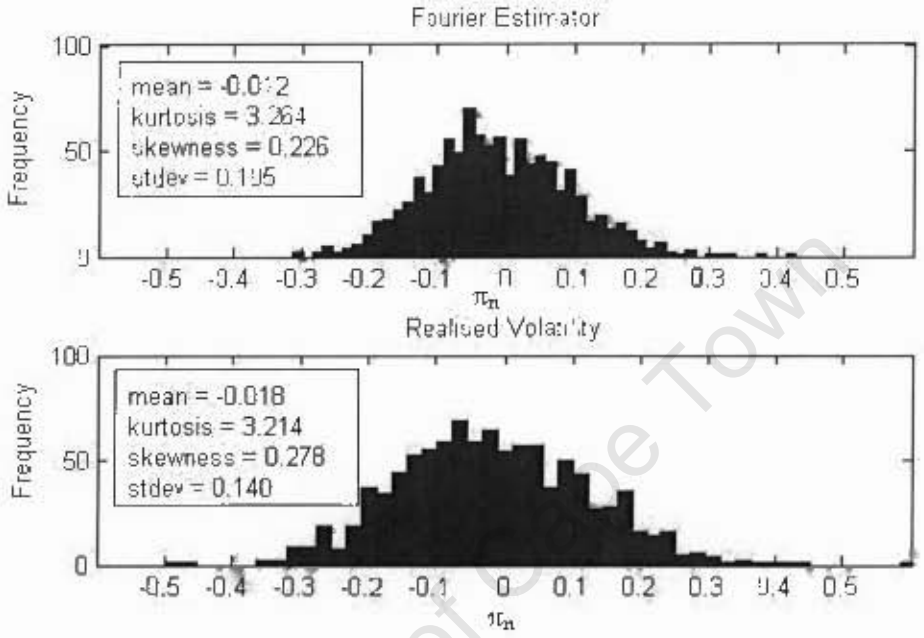

Figure 4.5: Distribution of the standardised errors, π_n, between the true volatility and the Fourier estimator and realised volatility when volatility is constant and the data is exponentially distributed with $\gamma = 234$.

In our next simulation, we look at the behaviour of the volatility measures on different time scales. In the Fourier method, the time scale is varied using the variable τ, where $S = \frac{N}{2\tau}$ for the Fourier method and $M = \frac{N}{\tau}$ for realised volatility estimator. Iori [34] investigated this subject in the multivariate setting using generated time series with mean exponential rates between 3 and 20. These rates are based on the average means of S&P100 stocks in September 2002. Since the South African market is less liquid, we repeat this exercise for the univariate case using exponential means corresponding to the values observed on the JSE.

Figures 4.6 shows the results obtained from unevenly spaced time series with constant volatility over different time scales, ranging from 0 to 120 minutes. Here 'Fourier-100' denotes the results obtained from the Fourier estimator on a time series with $\gamma = 100$. The same applies for the other legend entries. Here we can see that the mean time between ticks has a negligible influence on the performance of the estimator for larger time scales. This figure also shows that the Fourier method takes on a time step function on scales greater than an hour. This is explained by Iori [34] as follows: The value of k, used in the reconstruction of the return series (Equation 3.14), is

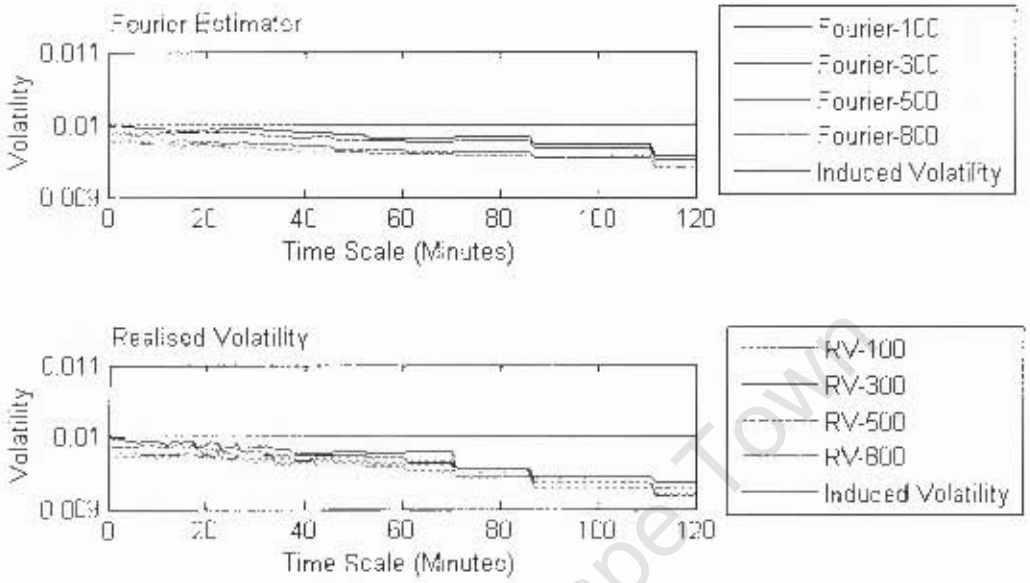

Figure 4.6: Results obtained from the Fourier estimator and realised volatility when using a simulated process with constant volatility of $\sigma = 0.01$ over different time scales, ranging from 0 minutes to two hours.

inversely proportional to the difference in time scale for consecutive k values. Therefore, as k decreases, the time scale difference increases, which leads to the step function. In addition, the simultaneous decay of all the volatility spectra from the induced level on time scales greater than 60 minutes can be attributed to the fact that volatility was initially set at the 1 second time step of the original series.

In [21], Kanatani shows that the Fourier estimator is equal to the raw data realised volatility estimator (RDRV), discussed in Section 2.1, when the number of Fourier coefficients is sufficiently large, i.e. the time scale is sufficiently small. This implies that, for very small time scales, the Fourier estimator and realised volatility should converge. This can also be observed from the figure above.

4.4 Empirical Analysis

In this section, the performance of the different volatility estimators is evaluated when empirical data is used. Firstly, the data that will form part of the analysis is analysed in Sections 4.4.1 and 4.4.2. The results of the empirical analysis are discussed in Section 4.4.3.

4.4.1 The data

The dataset analysed in this paper is the set of two-and-a-half year tick-by-tick trades executed on the JSE Stock Exchange from May 2002 till October 2004. Each quote comes with a time stamp rounded to the nearest second as well as the price of the trade. The volume of the trade is not included. This data has been recorded and provided by Deutsche Bank South Africa [5].

We focus our study on the 9 shares listed in Table 4.3. These are stocks from a number of different sectors on the JSE chosen for their high liquidity levels.

Table 4.3: JSE-traded shares used in this analysis. The share name, share code and sector are shown.

Share Name	Share Code	Sector
Anglo American	AGL	Resources
BHP Billiton	BIL	Resources
FirstRand	FSR	Financials
Gold Fields	GFI	Resources
Harmony Gold Mining	HAR	Resources
MTN Group	MTN	Non-Cyclical Services
Richemont Securities	RCH	Cyclical Consumer Goods
Standard Bank Group	SBK	Financials
Sasol	SOL	Resources

To determine the distribution of time between ticks, we examine the logarithm of the time between ticks. The probability density functions of the log of the time between ticks from our selection of stocks are shown in Figure 4.7. The fact that the distribution of the logarithm of the time appears normal shows that the distribution of the times are exponential. This therefore validates our method of simulating high frequency data explained in Section 4.1. An exponential mean of 558.61 seconds is observed, in contrast with a mean of 14 seconds observed on the S&P100 index, indicating a market that is comparatively illiquid.

As a matter of interest, we also examined the distribution of the trades during the course of the day. Figure 4.4.1 shows the number of ticks per series per time window , while Table 4.4 shows the average number of seconds between ticks as well as the average number of executed trades on these stocks during the time period of the available data, which is an indication of the liquidity of the stocks.. As expected, for most shares, the number of trades increases sharply in the period between 9:00 and 11:00, while there is a sharp decline at 13:00. It is interesting to note that, on average, the

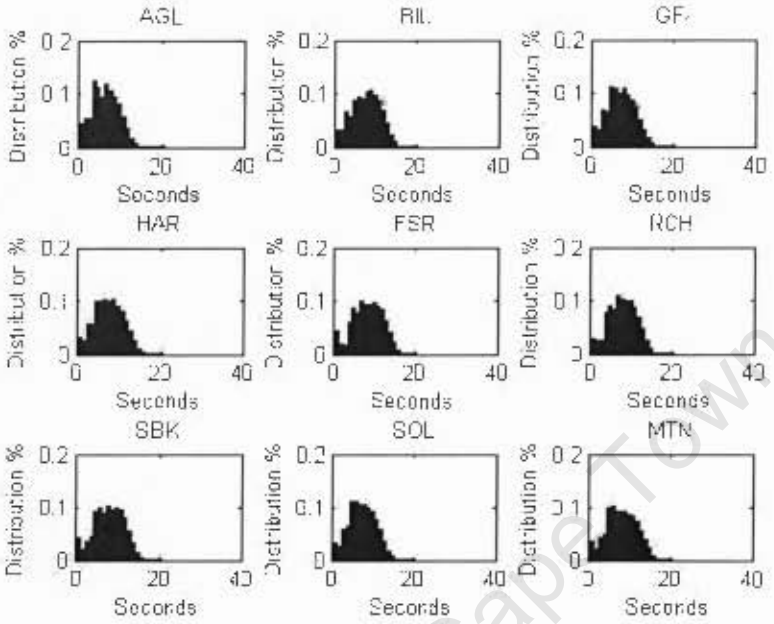

Figure 4.7: Probability density function of the logarithm of the time between ticks, measured in number of seconds.

highest volume of trades occurs in the last two hours of the business day. This is possibly due to day-traders closing out their positions for the day.

4.4.2 Data Filtering

When research is performed on high frequency historical data, bad quotes could negatively influence the results of the study (see [25] for a complete discussion on the topic of data cleaning). It is therefore necessary to ensure that the dataset that is used is clean. Data cleaning could be a very technical, demanding topic and task. For this study, we have therefore kept to a number of basic filtering rules to clean the data. The following rules, based on the filtering methodology used by Hog and Lunde in [9], were used:

- Remove prices equal to zero.
- Remove prices stamped outside 9:30 am and 4:00 pm.
- For a running window of 100 prices, remove all prices that are more than two standard deviations different from the median absolute deviation about the mean (MADAM) in their window.

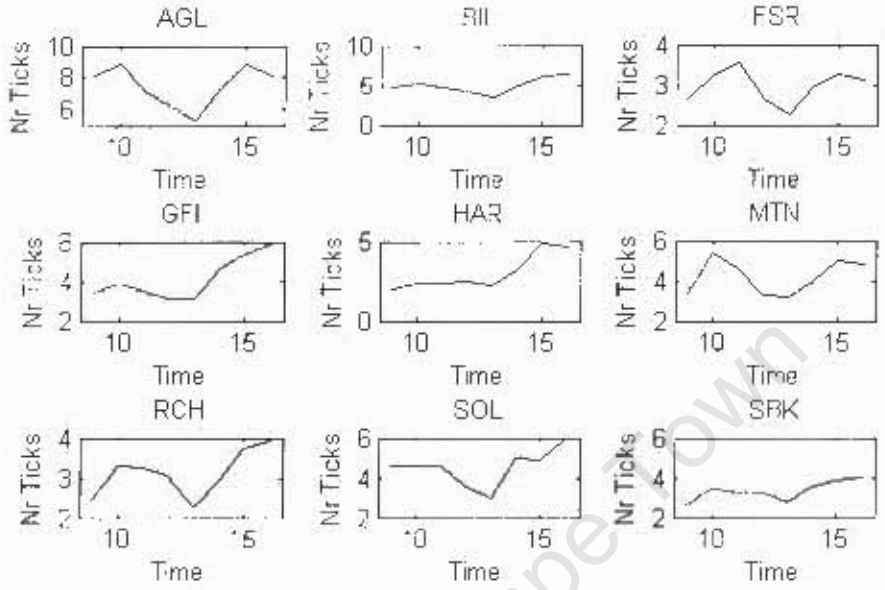

Figure 4.8: Number of ticks per share per time window, ranging between 9:30 am and 16:00 pm.

- If a price went up (or down) by more than two standard deviations and then back down (or up) again within 10 trades, all of the quotes in this interval are removed.

The following table shows the results of the filtering exercise.

4.4.3 Results of empirical analysis

In this section, the results of the empirical analysis are discussed. The performance of the estimators when using both evenly spaced and high frequency data is considered.

Results on evenly spaced data

In this section, we compare the results of the Fourier estimator and realised volatility when evenly spaced, weekly data is used. In this analysis, we measure the integrated volatility of nine stocks over a rolling window of data with window size equal to 30 days. The results are shown in Figure 4.9. From this figure we can see that the Fourier estimator compares well with realised volatility and gives smoother results.

Table 4.4: Average number of seconds between ticks as well as the average number of ticks that occurred on any given day during the time period in question is shown

Series	Average number of seconds between ticks	Average number of ticks per day
AGL	203	426
BIL	385	224
FSR	548	280
GFI	308	208
HAR	414	157
MTN	558	214
RCH	403	184
SBK	469	273
SOL	315	154

Table 4.5: Number of ticks per time series before and after filtering.

Series	Ticks before filtering	Ticks after filtering
AGL	383477	303083
BIL	201692	158274
FSR	141894	100526
GFI	252517	191865
HAR	187972	139777
MTN	139359	97938
RCH	192855	150080
SBK	165958	117944
SOL	246514	186990

Results on high frequency data

For the first step of our analysis of the results on high frequency data, we compute the average of the intraday integrated volatility of our time series using both the Fourier estimator and realised volatility for a rolling window with a size of one hour, rolled every fifteen minutes over 255 trading days. The results are shown in Figure 4.10.

From these figures, we can see that the Fourier estimator gives results comparible to that of realised volatility. Since the stocks from the JSE used in this report is not as liquid as, for example, the S&P100, we note that care should be taken in calculating volatility using such small time frames, since we have already shown that both estimators give more accurate results when larger time series are used. We will therefore focus on complete sets of intraday data for the rest of our analysis.

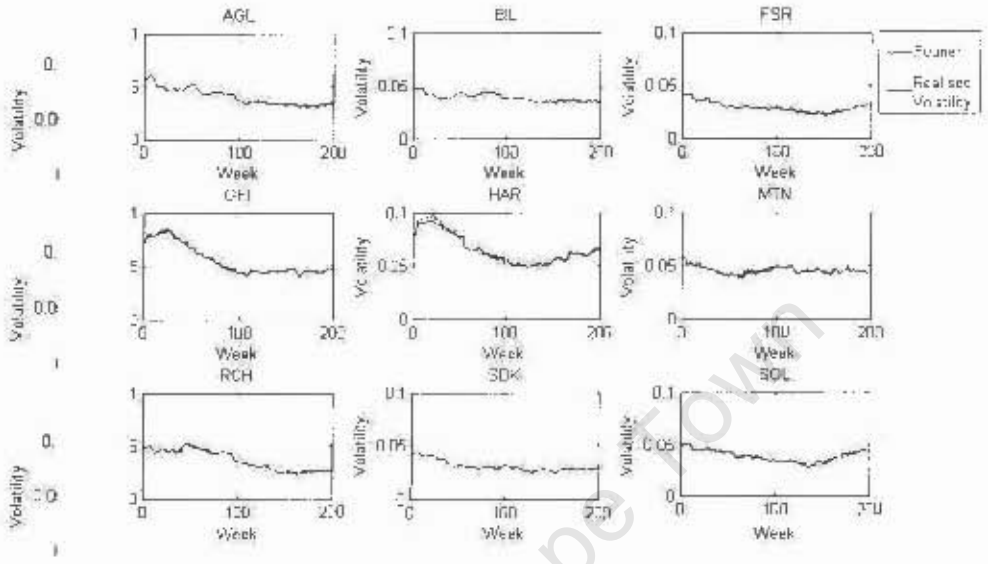

Figure 4.9: Integrated volatility over 255 rolling windows with a window size of 30 days when evenly spaced data is used. The blue line represents realised volatility while the red line represents the Fourier estimator.

In our next simulation, intraday volatility is calculated over different time scales, i.e. the number of Fourier estimators included in the estimation is varied to determine the return scale. The analysis is performed on all nine of the stocks with time scales ranging between 1 minute and 2 hours. The average over 30 days is taken. The results of the simulation are shown in Figure 4.11.

From this figure we can see that the Fourier estimator performs well in comparison with realised volatility, with the Fourier method providing slightly smoother estimates. Both estimators shows an inverse relationship between volatility and the size of the time scale, i.e. the volatility decreases as the time scale increases.

To get an objective view on how the two methods compare, we calculate the average absolute error between the methods using the formula

$$\text{Average absolute error} - \sum_{d=1}^{\text{Nr of days}} \frac{|\sigma^{S}_{Fourier}(d) - \sigma^{M}_{RV}(d)|}{\sigma_{RV}(d)} \qquad (4.10)$$

where $\sigma^{S}_{Fourier}$ represents the integrated volatility calculated using the Fourier estimator with S Fourier coefficients and σ^{M}_{RV} represents the integrated volatility calculated using realised volatility with M intervals.

Since ticks stamped between 9:30 am and 16:00 pm have been removed,

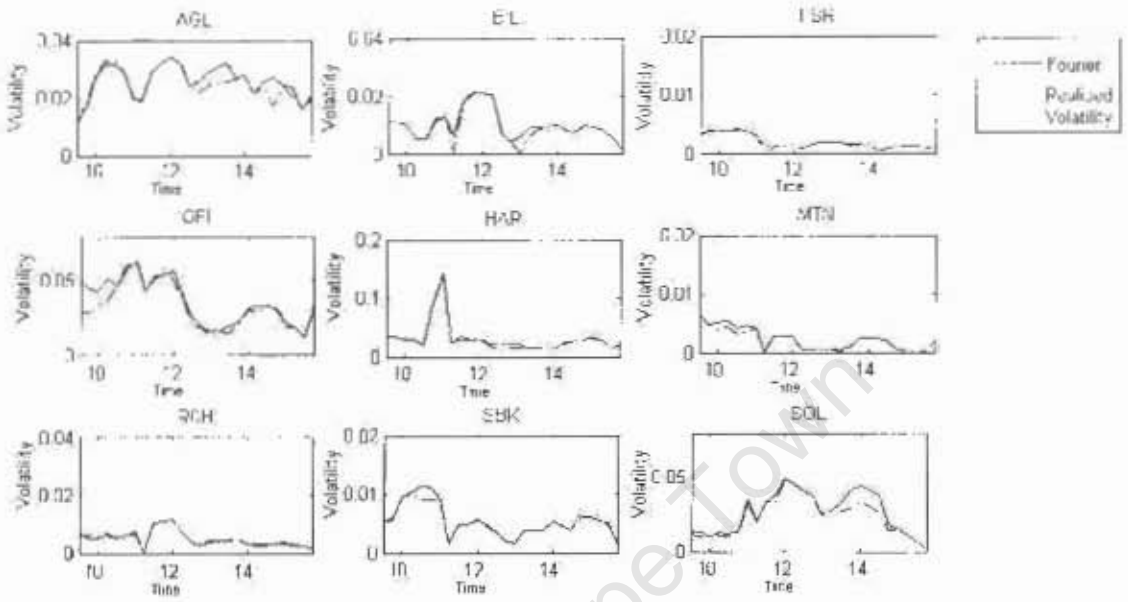

Figure 4.10: Average integrated volatility of a rolling window with a size of one hour, rolled every fifteen minutes, over 255 days using the Fourier estimator (red solid line) and realised volatility (blue dotted line).

6.5 hours of trading per day are left, which equals 23400 seconds. When calculating volatility using the Fourier estimator, $S = \{6.5, 13, 26, 39, 78, 195\}$ was used, which corresponds to using hourly, 30-minutes, 15-minutes, 5-minutes and 2-minute returns respectively.

The results of this study are shown in Figure 4.12, where the average absolute error against the number of Fourier coefficients included in the estimation is shown for each stock. This figure clearly shows that the difference between the two methods decreases when the sampling frequency increases. This was also proved by Kanatani [19].

Table 4.6 summarises the mean absolute error between the different methods for all the stocks.

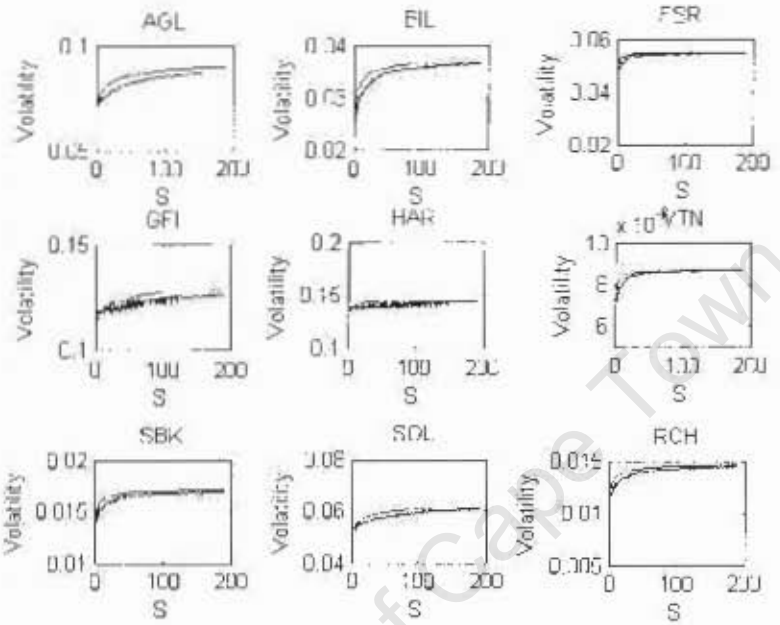

Figure 4.11: Results obtained from the two estimators when using real intraday data over different time scales. The average volatility over 255 days are shown.

Table 4.6: Mean absolute error between methods

Stock Code	Mean absolute error
AGL	0.0360
BIL	0.0283
FSR	0.0080
GFI	0.0219
HAR	0.0141
MTN	0.0086
RCH	0.0236
SBK	0.0159
SOL	0.0224

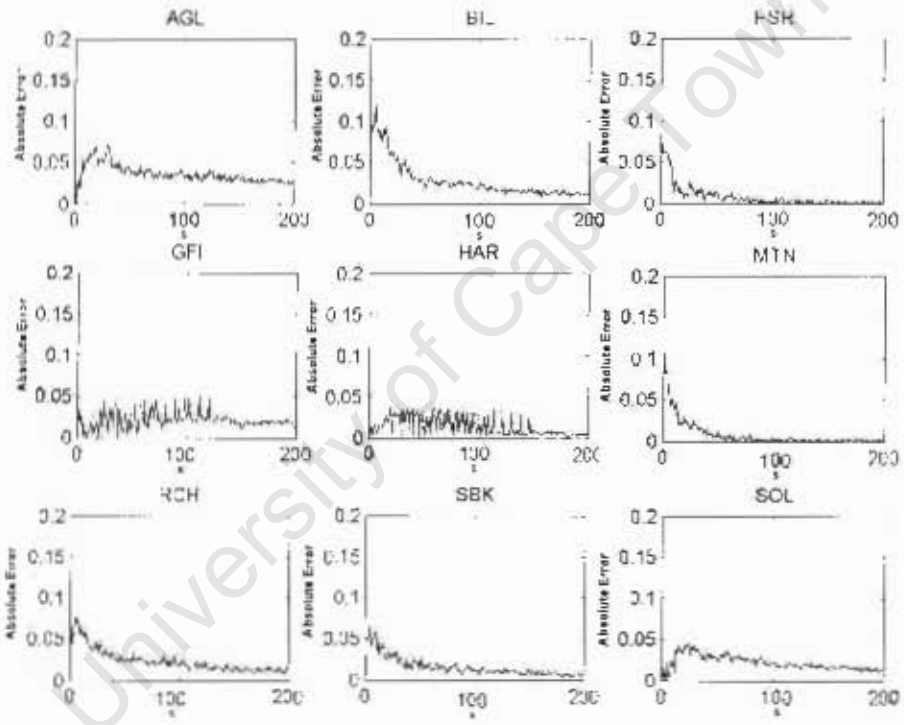

Figure 4.12: The absolute error between realised volatility and the Fourier estimator for different values of S.

4.5 Summary

In this chapter, we evaluated the univariate Fourier estimator in comparison with realised volatility. From the results of our simulations, we have found that:

1. The discretised formulae used to calculate the Fourier coefficients of the return series and the volatility, which are used in the Fourier method, is consistent with the theory.

2. The Fourier estimator performs well in comparison with realised volatility when evenly spaced, synchronous, simulated data is used.

3. The accuracy of the Fourier estimator increases with the size, i.e. the number of observations, denoted by N, of the data series .

4. The accuracy of the Fourier estimator increases with the liquidity, γ, of the data series.

5. The Fourier estimator outperforms the realised volatility estimator when high frequency simulated data is used.

6. The Fourier estimator is less sensitive to the choice of the return scale. The return scale is determined by the number of Fourier coefficients, S, included in the estimation in the case of the Fourier estimator and the number of evenly spaced data points, M, included in the calculation of Realised Volatility.

7. The Fourier estimator provides smoother results than realised volatility when empirical data is used.

8. The absolute error between the Fourier estimator and realised volatility decreases as the return time scale decreases.

Chapter 5

Estimation of multivariate integrated volatility

In this chapter, the estimation of correlation between financial time series is considered in the case where high frequency data is available. The estimation of correlation between the returns of different financial time series plays an important role in fields such as risk management, portfolio allocation, trading and hedging. We will examine the behaviour of the multivariate Fourier estimator, which was derived in Chapter 3, in comparison with that of the linear correlation coefficient.

Correlation is often estimated using the *linear correlation coefficient*, also known as the *Pearson coefficient*, which is the multivariate version of realised volatility and is seen as the basic measurement of the dependence between variables. It is defined as follows:

$$\rho(p_k, p_l) \equiv \frac{\sum_{i=1}^{n}(p_k(i) - \overline{p_k})(p_l(i) - \overline{p_l})}{\sqrt{\sum_{i=1}^{n}(p_k(i) - \overline{p_k})^2 (p_l(i) - \overline{p_l})^2}} \tag{5.1}$$

with the sample means defined by

$$\overline{p_k} \equiv \sum_{i=1}^{n} \frac{p_k(i)}{n} \quad \overline{p_l} \equiv \sum_{i=1}^{n} \frac{p_l(i)}{n}, \tag{5.2}$$

where p_k and p_l are two equally spaced, i.e. homogeneous, return series. By definition, $\rho(p_k, p_l)$ can vary from -1 (completely anti-correlated pair of stocks) to 1 (completely correlated pair of stocks). When $\rho(p_k, p_l) = 0$, the two stocks are uncorrelated.

In the previous chapter, the problems encountered when volatility is estimated from high frequency data were discussed. These problems also apply in the multivariate case. In addition, we also face the problem of asynchronous time series when working with multivariate high frequency data, i.e. the price of different stocks do not necessarily change at the same time.

To avoid these issues when calculating the Pearson coefficient, the unequal and irregularly spaced high frequency raw time series are converted to evenly spaced, synchronous data series using interpolation. As discussed in the previous section, various interpolation methods are available, but the ones most widely used in the literature are linear interpolation and previous-tick interpolation. Since Barucci and Reno [7] have shown that linear interpolation causes a downward bias, which becomes more profound as the frequency of the data increases, while previous-tick interpolation was shown to be unbiased if the diffusion is observed at $t_0 = 0$ and $t_N = T$, we will use the latter. The Pearson method still entails two main drawbacks due to the fact that it requires data to be both evenly spaced and synchronous: in some intervals of time no observations are available and some observations are thrown away. This is especially problematic when large covariance matrices are calculated.

Another method for measuring correlation is the multivariate version of the raw data realised volatility estimator (See Equation 2.10), which was examined by Kanatani [21]. In this case, the unevenly sampled observations $p_i(t_k^i)_{k=0}^{N_i}$ and $p_j(t_k^j)_{k=0}^{N_j}$ are used and the correlation is defined as follows:

$$\Sigma(i,j) = \frac{\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta p_i(t_k^i)\Delta p_j(t_l^j)I(A)}{\sqrt{(\sum_{k=1}^{N_i}\Delta p_i(t_k^i)^2)(\sum_{k=1}^{N_j}\Delta p_j(t_k^j)^2)}} \tag{5.3}$$

where $A = (t_k^i, t_{k-1}^i) \cap (t_l^j, t_{l-1}^j) \neq \emptyset$ and $I(.)$ denotes the indicator function. We refer to this as *raw data realised correlation* or RDRC. In this case, we assume that $\Delta t \to 0$, i.e. the time between ticks is very small. This implies that this method cannot be used to calculate the correlation over different time scales and can only be used for comparison with other methods when looking at the smallest time scale, which is 1 second in the case of high frequency data.

In the previous chapter, the univariate Fourier estimator was examined. By polarisation of the one-dimensional result (Equations 3.16, 3.17 and 3.18), the Fourier estimator can be extended to the multivariate case. Since this method is based on the integration instead of the differentiation of the return series, it should provide the necessary robustness to address the issues related to high frequency data. The following formulae, which reiterate Equations 3.27, 3.28 and 3.29, show how the Fourier coefficients of the volatility matrix Σ_{ij} can be calculated:

$$a_0(\Sigma_{ij}) = \lim_{S\to\infty}\frac{\pi}{S+1-n_0}\sum_{s=n_0}^{S}(a_s(dp_i)a_s(dp_j) + b_s(dp_i)a_s(dp_j)),$$

$$a_k(\Sigma_{ij}) = \lim_{S\to\infty}\frac{\pi}{S+1-n_0}\sum_{s=n_0}^{S}\frac{1}{2}(a_s(dp_i)a_{s+k}(dp_j) +$$

$$
\begin{aligned}
& a_s(dp_j)a_{s+k}(dp_i)), \\
b_k(\Sigma_{ij}) &= \lim_{S\to\infty} \frac{\pi}{S+1-n_0} \sum_{s=n_0}^{S} \frac{1}{2}(a_s(dp_i)b_{s+k}(dp_j) + \\
& a_s(dp_j)a_{s+k}(dp_i)).
\end{aligned}
$$

The instantaneous volatility matrix Σ_{ij} can then be reconstructed by the Fourier-Fêjer inversion formula shown in Equation 3.10.

To calculate the integrated volatility matrix $\Sigma^{\star}_{ij}$ we can once again use the fact that

$$
\begin{aligned}
\Sigma^{\star}_{ij} &= 2\pi a_0(\Sigma_{ij}) \\
&= \frac{\pi^2}{S} \sum_{q=1}^{S} (a_q(dp_i)a_q(dp_j) + b_q(dp_i)b_q(dp_j)). \qquad (5.4)
\end{aligned}
$$

The Fourier multivariate estimator was first implemented by Reno [37], where it was used to investigate the Epps effect. Kanatani [21] examined the Fourier estimator in comparison with optimally weighted realised volatility, while Iori and Precup [33], [34] compared the Fourier multivariate estimator, realised volatility and the co-volatility weighting proposed by Dacorogna et al [25] using simulated data as well as empirical data from the S&P100.

In this chapter, the Fourier multivariate estimator is evaluated in the South African context, where empirical data from stocks traded on the JSE data is used in conjunction with simulated series with the same characteristics as stocks traded on the JSE.

5.1 Evaluation Methodology

Without loss of generality, we set the number of assets as two and test the Fourier method through a Monte Carlo study on a simulated bivariate process using the methodology proposed by [25].

This is done by first simulating two uncorrelated, normally distributed, i.i.d. random time series $A(t)$ and $B(t)$ of size N = 86000 (number of seconds in a day), each with zero mean and standard deviation of 0.01. It can be confirmed that A and B are still normally distributed with correlation of κ. A third series $C(t)$, which is a linear combination of $A(t)$ and $B(t)$, is then created as follows:

$$
C(t) \equiv \kappa A(t) + \sqrt{(1-\kappa^2)}B(t), \quad \text{where } 0 \leqq t \leqq N \qquad (5.5)
$$

where N is the number of observations in each series and the correlation coefficient κ is selected such that $0 \leq \kappa \leq 1$. This corresponds to the

following constant volatility GARCH(1,1) model:

$$\begin{aligned} dp_1(t) &= \sigma dW_1(t), \\ dp_2(dt) &= \sigma dW_2(t), \\ corr(dW_1, dW_2) &= \kappa, \end{aligned}$$

where $0 \leq \kappa \leq 1$. In our simulation, we chose $\sigma = 0.01$ and $dt = 1/86000$. This process is repeated a 1000 times to simulate a 1000 days of trading. [8] states that a reasonable approximation of the error on the estimate of κ is

$$\Delta\kappa = (1 - \kappa^2)\sqrt{\frac{1}{N}}.$$

The time interval between trades in JSE equities approximately follows an exponential distribution with rate parameter γ in the range 43 seconds (very liquid stock) to 3311 seconds (least liquid stock), while the mean time between ticks is equal to 423 seconds. [34] found that stocks from the S&P100 are much more liquid, with the time interval between trades ranging between 1 and 67 seconds.

To imitate actual trading patterns, we sampled the simulated process using the exponential distribution, as described in Section 4.1, and varied γ with distributions observed on the JSE for both synchronous and asynchronous series.

5.2 Results on Simulated Data

In our first simulation, we repeat the Monte Carlo simulation described above with different values for the correlation κ, where $\kappa = \{0.1, 0.2, \ldots, 0.8, 0.9, 1\}$. The Fourier estimator is then used to calculate the correlation between these simulated synchronous and asynchronous time series. The results are compared with that obtained from the Pearson coefficient. In this simulation, the Nyquist frequency is used in calculating the Fourier correlation. In the case of the Pearson coefficient, the estimator was tested using both linear and previous tick interpolation to convert the data into equally spaced series.

The results of this simulation are shown in Table 5.1, where the constant multiplier κ from Equation 5.5, which determines the correlation between the two series, and the mean values obtained by the two estimators are shown. The Fourier estimator consistently approximates the correlation between distributions A and C successfully. The results of Table 5.1 are depicted in Figures 5.1 (synchronous data) and 5.2 (asynchronous data). In the simulation on synchronous, unevenly spaced data, the Fourier estimator outperforms the Pearson coefficient, which has a slightly negative bias. As

Table 5.1: Results of a Monte Carlo simulation of the Fourier and Pearson correlation estimation methods. The table shows the induced correlation κ in the first column and the estimated correlations for the two methods for synchronous and asynchronous in the other columns.

	Synchronous Data		Asynchronous Data	
Multiplier κ	Fourier	Pearson	Fourier	Pearson
0.0	0.000	0.000	0.000	0.000
0.1	0.100	0.085	0.067	0.080
0.2	0.199	0.147	0.212	0.089
0.3	0.299	0.257	0.286	0.229
0.4	0.400	0.370	0.364	0.400
0.5	0.500	0.434	0.463	0.444
0.6	0.600	0.551	0.574	0.533
0.7	0.699	0.662	0.662	0.680
0.8	0.799	0.754	0.768	0.773
0.9	0.900	0.874	0.868	0.852
1.0	1.000	1.000	0.963	0.987

expected, the relative errors obtained when asynchronous data is used is larger than that obtained from synchronous data.

Table 5.2 reports the bias and root mean squared error (RMSE) of each estimator from the 1000 replications using synchronous and asynchronous data. The bias and RMSE were calculated as follows:

$$\begin{aligned} \text{mean} &= \frac{1}{N}\sum_{t=1}^{N}(k - \Sigma_{ij}^{\star}(t)), \\ \text{RMSE} &= \frac{1}{N}\sum_{t=1}^{N}(k - \Sigma_{ij}^{\star}(t))^2, \end{aligned}$$

where $\Sigma_{ij}^{\star}(t)$ denotes the estimator under question at time t.

Figure 5.3 shows the distribution of the standardised errors for both methods when using synchronous and asynchronous data with an induced correlation of 0.3. The kurtosis, skewness and standard deviation are also shown. A correlation of 0.3 was chosen as an approximation for the mean correlation value of stocks on short time scales. Here we see that the standardised errors obtained from the Fourier estimator follows a normal distribution with mean close to zero. This figure shows that the standard deviation of the relative error obtained from the Fourier method is significantly lower than that obtained from the Pearson method when using both synchronous and asynchronous data. The Fourier method appears to be unbiased, while the Pearson estimator is negatively biased. Both estimators, as expected, provide better results when synchronous data is used.

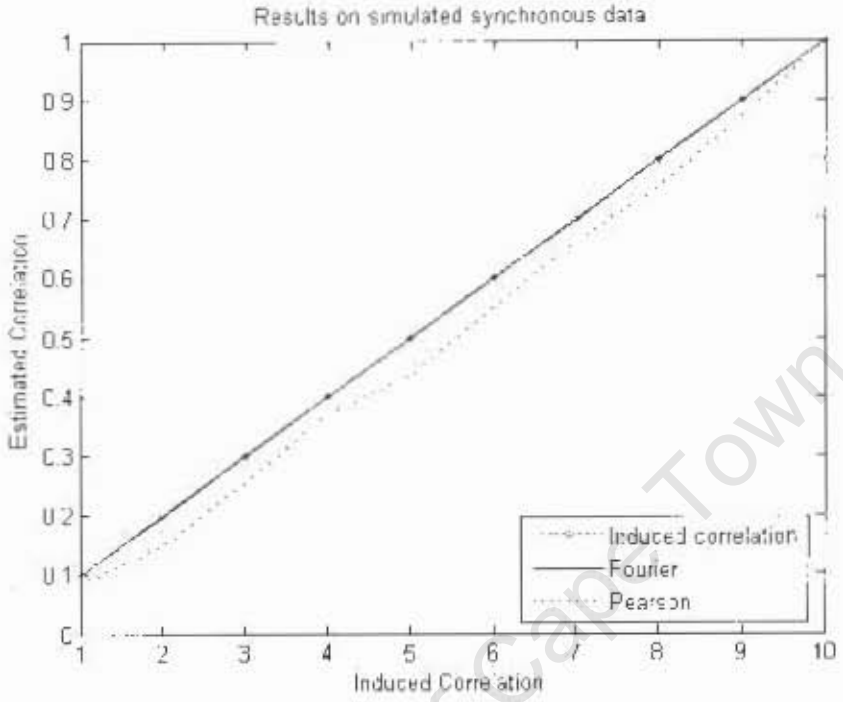

Figure 5.1: Results of the Fourier and Pearson estimators on simulated synchronous data. The figure shows the induced correlation, ranging between correlations of 0.1 and 1, in comparison with the estimated values from the two methods. A Monte Carlo study with 1000 simulations was used.

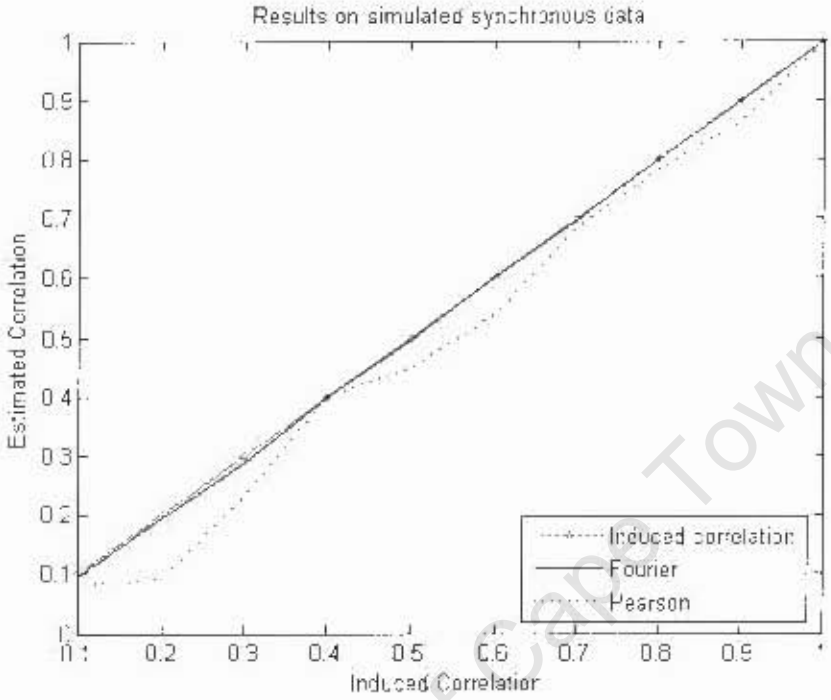

Figure 5.2: Results of the Fourier and Pearson estimators on simulated, unevenly spaced, asynchronous data. The figure shows the induced correlation, ranging between correlations of 0.1 and 1, in comparison with the estimated values from the two methods. A Monte Carlo study with 1000 simulations was used.

Table 5.2: Sample bias and RMSE of the Fourier and Pearson estimators for different values of κs when using synchronous and asynchronous unevenly spaced simulated data.

	Synchronous Data				Asynchronous Data			
	Bias		RMSE		Bias		RMSE	
k	Fourier	Pearson	Fourier	Pearson	Fourier	Pearson	Fourier	Pearson
0.1	-0.049	-0.049	0.005	0.246	-0.331	-0.086	0.047	0.242
0.2	0.026	-0.026	0.004	0.315	0.061	-0.027	0.034	0.312
0.3	-0.045	-0.045	0.003	0.267	-0.049	0.051	0.042	0.266
0.4	0.001	0.001	0.004	0.139	-0.090	0.020	0.035	0.135
0.5	-0.009	-0.009	0.003	0.211	-0.075	-0.026	0.022	0.210
0.6	0.008	0.008	0.002	0.134	-0.044	0.009	0.022	0.133
0.7	-0.003	-0.003	0.001	0.104	-0.055	-0.011	0.017	0.105
0.8	-0.001	-0.001	0.001	0.061	-0.041	-0.005	0.012	0.062
0.9	-0.001	-0.001	0.000	0.024	-0.036	-0.003	0.004	0.026

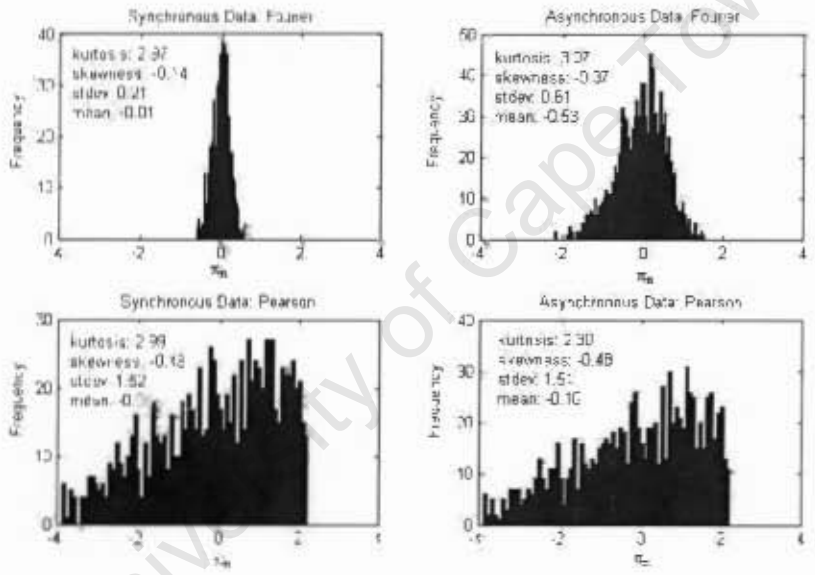

Figure 5.3: Distribution of standardised error, calculated using Equation 4.4, for both the Fourier and Pearson estimators when simulated synchronous and asynchronous data with a correlation of 0.3 is used.

In our next simulation, we look at the behaviour of the correlation measures on different time scales. Iori [34] investigated this issue using generated time series with mean exponential rates between 3 and 20. These rates are based on the average means of S&P100 stocks traded in September 2002. Since the South African market is less liquid, we repeat this exercise using exponential means corresponding to the values observed on the JSE. Here 'Fourier-100' once again denotes the results obtained from the Fourier estimator on two synchronous time series with $\gamma = 100$, while 'Fourier-100-300' denotes asynchronous series with $\gamma 1 = 100$ and $\gamma 2 = 300$. The same applies for the Pearson estimator.

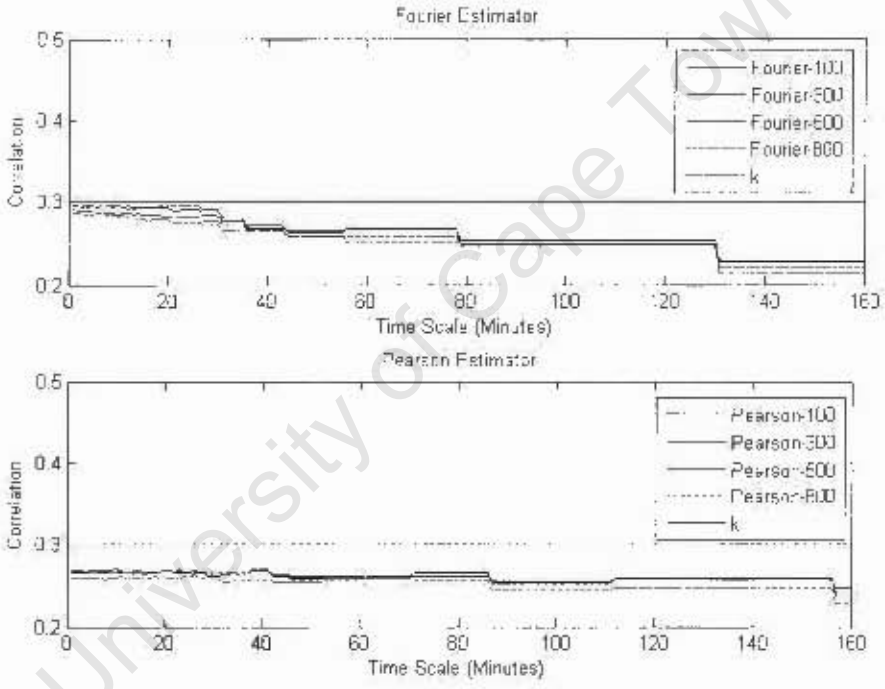

Figure 5.4: Analysis of simulated bivariate synchronous processes with an induced correlation of 0.3. The results of the Fourier and Pearson estimators are shown as a function of the time scale.

Figure 5.4 shows the results obtained from unevenly spaced synchronous time series. Here we can see that the mean time between ticks does not have a significant influence on the performance of the estimator. This figure also shows that the Fourier method takes on a time step function on scales greater than an hour. As mentioned in the previous chapter, this is due to the fact that the value of k, used in the reconstruction of the return series, is inversely proportional to the difference in time scale for consecutive k values. In addition, the simultaneous decay of all the correlation spectra from the induced level can be attributed to the fact that correlation was

initially set at the 1 second time step of the original series. We also observe that the Fourier estimator outperforms the Pearson estimator on very small time scales. The Pearson estimator is, however, less sensitive to the choice of time scale.

In [21], Kanatani shows that the Fourier estimator is equal to the raw data realised correlation estimator (RDRC), which is also equal to the Pearson coefficient, when the number of Fourier coefficients is sufficiently large, i.e. the time scale is sufficiently small. This can also be observed from the figure above.

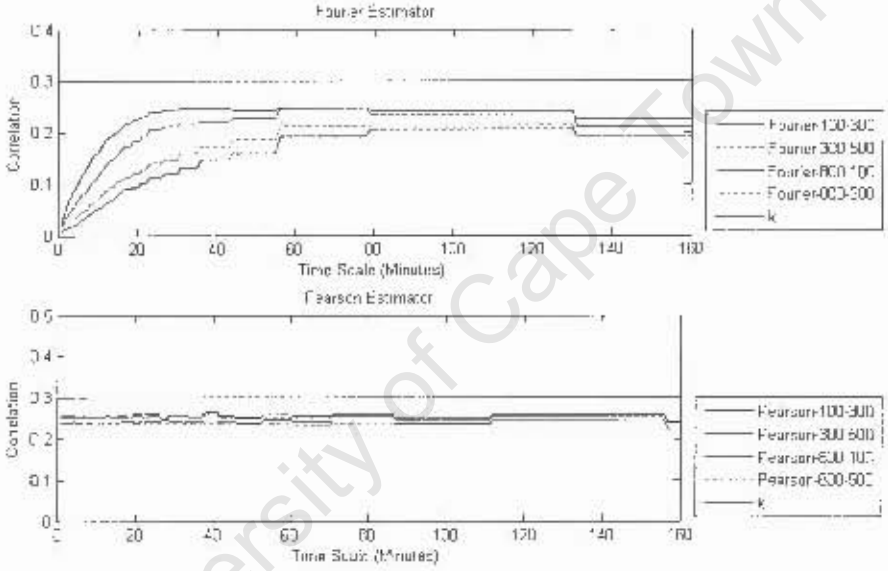

Figure 5.5: Analysis of simulated bivariate asynchronous processes with an induced correlation of 0.3. The results of the Fourier and Pearson estimators are shown as a function of the time scale.

When looking at the results on asynchronous data as shown in Figure 5.5, we observe that the Fourier estimator only reaches its optimal level at a time scale of about 60 minutes, while the Pearson estimator converges significantly faster. In [34], it was found that, for asynchronous series, the correlation spectra on shorter time scales are dependent on the exponential rates. This can also be observed in the figure above. The higher the mean rate of a series, the faster it deviates from the induced correlation level on a short time scale. With the dataset used in this paper, simulated to reflect the behaviour of stocks included in the S&P100 index, the series start to stabilise after 10 minutes, indicating that the time it takes for a series to stabilise is linked to the liquidity of the series. The initial deviation or decay is explained by the fact that, for exponentially distributed intertrade times, the proportion and magnitude of large positive deviations from the mean

increase with the mean itself. In other words, the standard deviation of intertrade times are higher in series with a large mean between intertrade times.

In both cases, the results obtained from the estimators under question have a negative bias, which can then be explained by the asynchronous nature of the data.

The effect of asynchronous data was also studied in [8]. In this paper they found that, when using only the synchronous data points, i.e. those data points which occur at the same time, the right correlation is obtained, but with a larger volatility. However, when using asynchronous data, a negatively biased estimate is obtained. They conclude that only data which come in the same time are therefore meaningful and that asynchronous data points cannot fully reveal correlations. This is a serious limitation, since the data set is reduced significantly when only considering synchronous data points. However, the fact that the high frequency data are not equally spaced does not affect the power of the Fourier algorithm, while it does influence the Pearson estimator.

5.3 Empirical Analysis

In this section, the behaviour of the Fourier and Pearson estimators are examined using empirical data from the JSE. The same stocks used in the previous chapter and examined in Section 4.4.1 are used. Section 5.3.1 looks at results obtained from evenly spaced data, while Section 5.3.2 considers the high frequency case. We replicate and confirm the results of Malliavin and Mancino [22] by evaluating the results obtained from evenly spaced data and replicate results from Iori [34] to evaluate the high frequency case. We add to this literature by evaluating the behaviour of the eigenvalues of the estimated correlation matrices to get a better understanding of the temporal stability of the matrices calculated using the Fourier method.

5.3.1 Results on evenly spaced data

To show that the method performs well in computing correlation for evenly spaced, synchronous financial time series, the method has been applied to compute the correlation between the chosen stocks using weekly data from January 2001 till January 2006. Figure 5.6, which is a subset of the results, display the correlation between Anglo American (AGL) and Billiton (BIL), Gold Fields (GFI), Standard Bank (SBK) and Sasol (SOL) respectively, calculated using both the Fourier method and the Pearson method. A rolling window with a window size of 52 weeks, which is rolled forward on a weekly basis, is used.

The estimate according to the Fourier methodology is in line with the one obtained with the Pearson method and confirms the results obtained

from simulated data discussed in the previous section. It also shows that the Fourier method gives smoother results than those obtained from the Pearson method. It was also confirmed that the Fourier estimator always provides positive definite matrices, while this is not the case with the Pearson method.

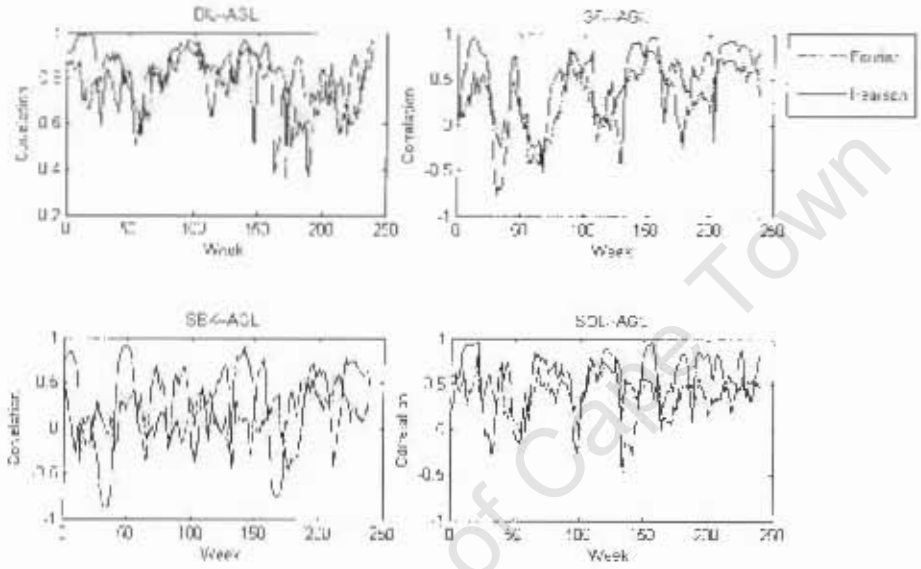

Figure 5.6: Correlation between AGL and four other shares, namely BIL, GFI, SBK and SOL. The study was performed over a rolling window of evenly spaced, weekly data from the JSE between January 2001 (Week 1) and January 2006 (Week 250).

5.3.2 Results on high frequency data

In this section, the behaviour of the correlation estimators are examined in the presence of high frequency data. We start by looking at the behaviour of the estimators when the correlation between stocks from the same sector is considered and then look at the correlation between stocks from different sectors. The temporal evolution of the correlation matrices is then examined by looking at the stability of the eigenvalues and the distribution of the difference between the correlation estimates between consecutive time scales.

Figure 5.7 shows the time scale evolution of the Fourier and Pearson correlations between 3 gold stocks from the Resources sector, namely Anglo Gold (ANG), Gold Fields (GFI) and Harmony Gold Mining (HAR). These stocks are all relatively liquid in the South African market with intertrade times of 126.26 (ANG), 51.98 (GFI), and 88.72 (HAR) for the time period under question. The average of the intraday volatility over the different time scales for every day of the month is calculated. This figure shows the results for January 2004.

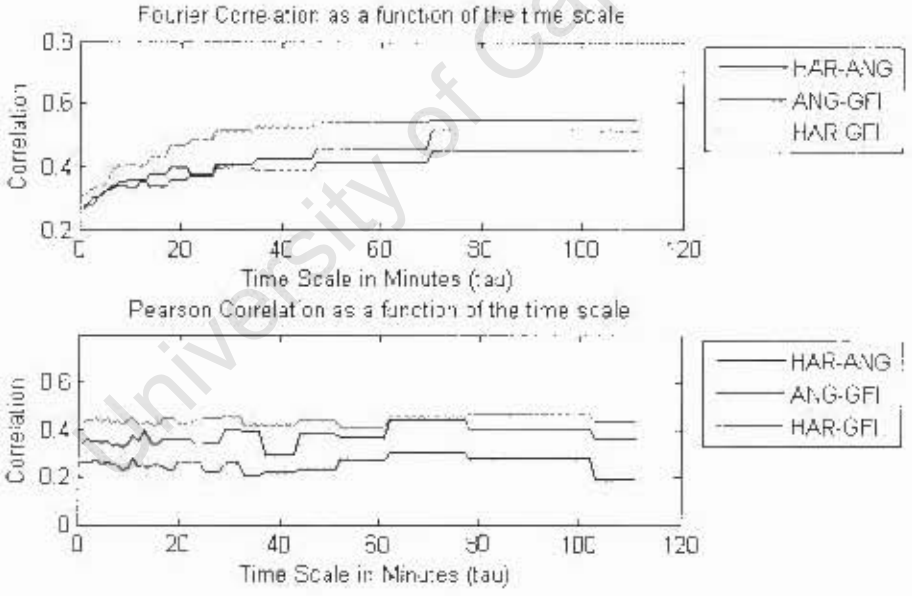

Figure 5.7: Average correlation as a function of the time scale as computed with the Fourier and Pearson methods between three gold stocks from the Resources sector, namely Anglo Gold (ANG), Gold Fields (GFI) and Harmony Gold Mining (HAR). High frequency data from January 2004 was used. The time scale ranges from 1 minute to two hours.

As expected, both estimators show significant correlation between the three pairs of stocks. The rise in the variation of the estimates with an increased time scale could be attributed to a loss of statistical power as

the number of observations included in the estimators are decreased. In all cases, the Fourier correlation method provides a smoother spectrum than the Pearson method. In addition to this test, tests were also performed on high frequency data with a length of one month (Note that in the previous test, the length of data was one day, with the tests repeated for every day of a month). This resulted in higher correlation estimates between the chosen stocks than that obtained from using daily data. This is as expected, since it was already shown by Epps [11] that the correlation among stocks drop when the sampling horizon is decreased. This is referred to as the Epps effect in the literature.

Figure 5.8 shows the time scale evolution of the correlation between 3 stocks from the Financial sector, namely First Rand (FSR), Standard Bank (SBK) and Nedbank (NED).

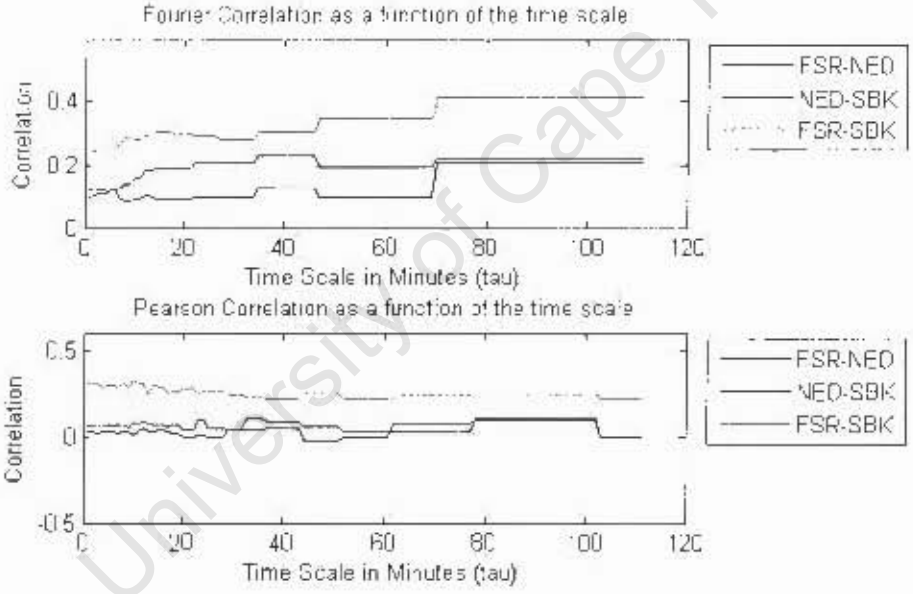

Figure 5.8: Average correlation as a function of the time scale as computed with the Fourier and Pearson methods between three stocks from the Financial sector, namely First Rand (FSR), Standard Bank(SBK) and Nedbank (NED). High frequency data from January 2004 was used. The time scale ranges from 1 minute to two hours.

In this case, the Fourier estimator shows significant correlation between the pairs of stocks while the Pearson estimator shows a strong correlation between FSR and SBK, while low, and sometimes even negative, correlation between NED and SBK and FSR and NED.

Figure 5.10 shows the time scale evolution of the correlation between 3 stocks from different sectors. The following stocks were chosen: Anglo Gold from the Resources sector (ANG), Standard Bank (SBK) from the

Financial sector and MTN (MTN) from the Non-Cyclical Services sector. We note that the Epps effect is slightly more pronounced in this figure where we estimate the correlation between stocks from different sectors than the previous figures where we looked at the correlation of stocks from the same sector. In [14], Mantegna, Bonanno and Lillo also noted that the correlation decreases faster for intra-sector stocks than for inter-sector stocks.

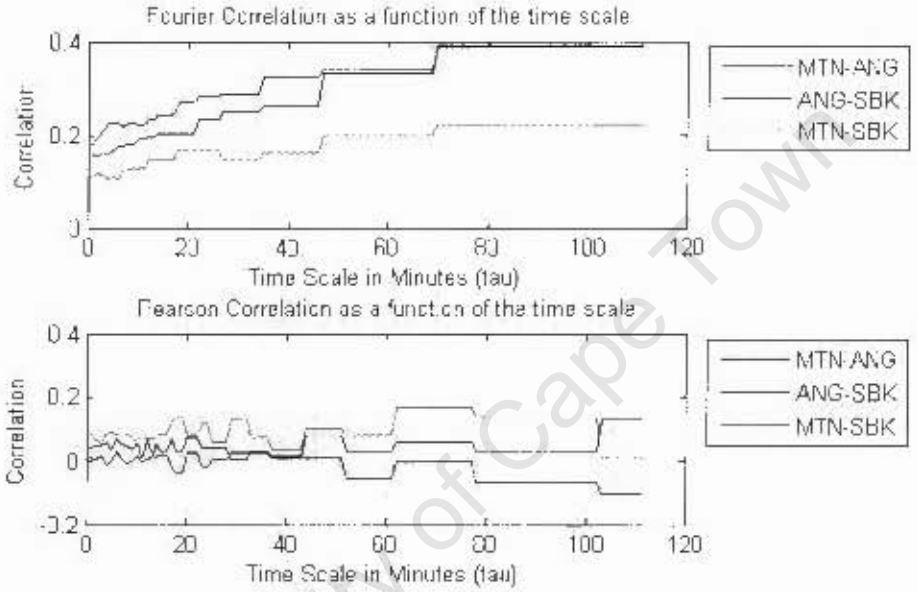

Figure 5.9: Average correlation as a function of the time scale as computed with the Fourier and Pearson methods between three stocks from different sectors, namely Anglo Gold (ANG) from the Resources sector, Standard Bank(SBK) from the Financial Sector and MTN (MTN) from the Non-Cyclical Services sector. High frequency data from January 2004 was used. The time scale ranges from 1 minute to two hours.

In our next study, we examine the temporal evolution of the correlation matrices by evaluating the behaviour of the eigenvalues of the correlation matrices across different time scales. Figure 5.10 shows the eigenvalues of the correlation matrices calculated by the Fourier and Pearson methods as a function of the time scale, while Figure 5.11 shows the standard deviation of the eigenvalues. From these figures, we can see that the larger eigenvalues obtained from the Fourier method are more stable (has a smaller standard deviation), than those obtained from the Pearson method, indicating that the temporal stability of correlation matrices calculated using the Fourier method exceeds that of matrices calculated using the Pearson method.

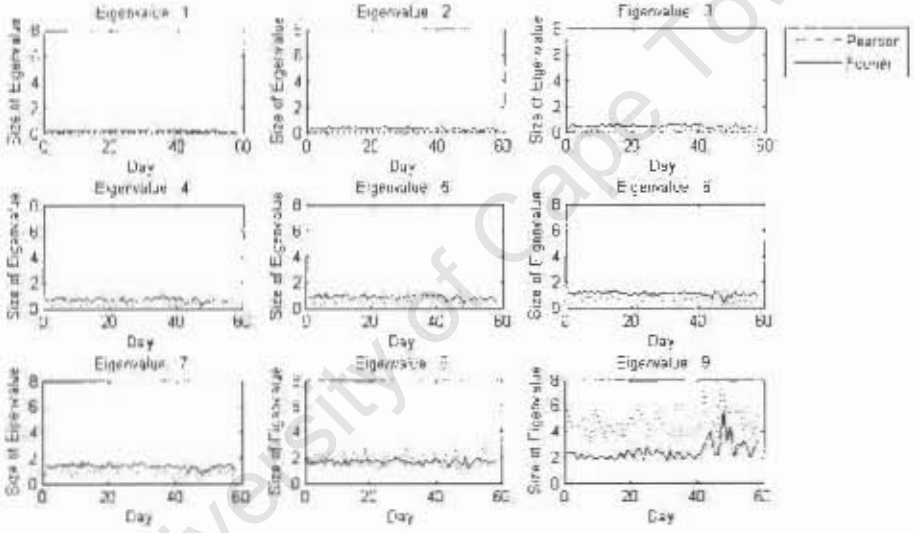

Figure 5.10: Time evolution of the eigenvalues of the estimated correlation matrices ordered by size. Eigenvalue 1 is the smallest eigenvalue and eigenvalue 9 the largest.

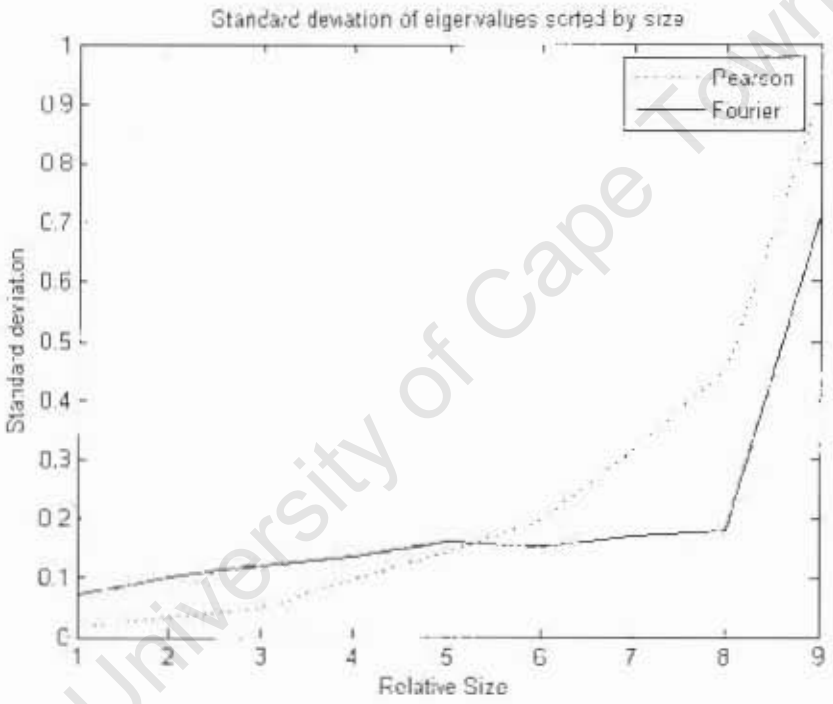

Figure 5.11: Standard deviation of the the eigenvalues of correlation matrices ordered by size. Eigenvalue 1 is the smallest eigenvalues and eigenvalue 9 the largest.

Another way of looking at the overall temporal evolution of the correlation matrices is to compute their consecutive differences defined as $c_{ij}(t) - c_{ij}(t-1)$, where c_{ij} is the correlation between stocks i and j at time scale t. This study was also performed on S&P100 data in [34]. The results are shown in Figure 5.12. From this figure we can see that the Fourier method has a standard deviation significantly smaller than that of the Pearson method. The tails of the distribution are also shorter. The largest variation between consecutive time scales when the Fourier method is used was 0.1854 and -0.1732. The Pearson method results in values 0.6614 and -0.4634, which are significantly larger.

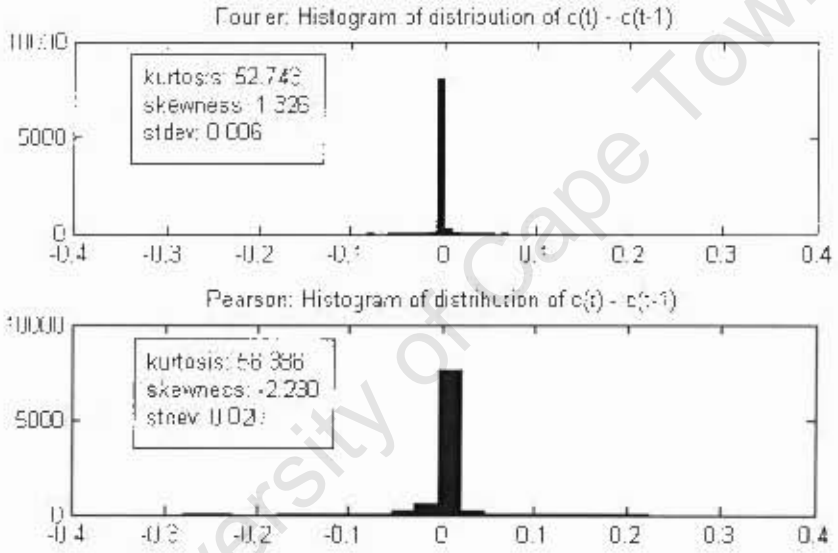

Figure 5.12: Distribution of the difference between the correlation estimates between consecutive time scales.

5.4 Summary

In this chapter, we evaluated the multivariate Fourier estimator in comparison with the Pearson estimator. From the results of our simulations, we have found that:

1. The Fourier method performs well in comparison with the Pearson coefficient when evenly spaced, synchronous, simulated data is used.

2. The Fourier method outperforms the Pearson method in terms of bias and RMSE when unevenly spaced, asynchronous, simulated data is used.

3. When using synchronous simulated data, the Fourier method outperforms the Pearson method on very small time scales in terms of accuracy. The Pearson method is, however, less sensitive to the time scale.

4. When using asynchronous simulated data, the Pearson coefficient converges faster to its optimal level of optimization when the time scale is increased.

5. The Fourier estimator provides smoother results than the Pearson estimator when evenly spaced empirical data is used.

6. The correlation matrices calculated using the Fourier estimator provide more stable eigenvalues than those calculated using the Pearson estimator.

7. The standard deviation of entries of the consecutive difference between correlation matrices is smaller for the Fourier estimator and has shorter tails.

Chapter 6

Conclusion and suggestions for future work

6.1 Concluding remarks

In this dissertation, we evaluated the method for estimating integrated univariate and multivariate volatility proposed by Malliavin and Mancino[22], referred to as the Fourier method. This method was compared to realised volatility, in the case of measuring univariate volatility, and the Pearson coefficient, in the case of measuring multivariate volatility.

To achieve our goal, we evaluated the method using both simulated time series and empirical financial data. In both cases, evenly spaced and high frequency data were considered.

In the case of evenly spaced data, we found that the Fourier method compares very well with classical methods and provide smoother estimates. In the case of high frequency data, we confirmed the results of Iori [34] and found that the Fourier method gives better results than the realised volatility estimator in terms of generating smooth estimates with a lower bias and root mean squared error, which are also less sensitive to the choice of returns time scale. The Fourier method is also model independent and guarantees a positive definite correlation matrix, which is not the case with other classical methods. We therefore conclude that the Fourier estimator is well suited to the time structure of high frequency data.

The implications of a better measure for the volatility are far reaching. In broad terms, a good tick-by-tick volatility estimator enlarges our information set about a given time series. This will lead to better forecasts, both because the information set in the past is better, and because the integrated volatility to be forecasted is known more accurately. In turn, this will lead to better risk management, portfolio optimization or option pricing.

6.2 Future work

For the purpose of this research, we made the assumption that volatility is constant over time when simulating stock prices. In future work, this assumption should be relaxed and other data generating mechanisms, such as the GARCH model, Ornstein-Uhlenbeck model, and other models with memory and jump processes should be implemented.

In this dissertation, we evaluated the correlation matrices in terms of their temporal stability. It would be an interesting exercise to use these correlation matrices in portfolio theory problems, such as the optimal asset allocation problem, to determine the success in comparison with classical methods.

Recent work of Mykland, Ait-Sahalia and Zhang [42] used the Fourier method in addition to certain bias correction methods. This has not been evaluated in the South African contect.

Appendix A

Program listing

The programs used to calculate the Fourier univariate and multivariate volatility are listed here. All other programs used in the simulation exercises and calculation of classicial methods for calculating volatility and correlation are available on request. All software was written in Matlab [17].

A.1 Volatility estimation using the Fourier method

```
function var = ftvar(data,N,varargin)
% FTCOR calculates correlation of high frequency data
%
% VAR = FTVAR(data,N,nrfc,n0) returns a matrix containing the
% integrated volatility calculated by the method proposed by
% Malliavin and Mancino in their paper 'Fourier series method
% for measurement of multivariate volatilities'.
%
% data      -  struct containing dates (data.dates) and
%              prices (data.PRC) for each series.
% N         -  length of time scale
% nrfc      -  number of Fourier coefficients to be
%              included in estimation
% n0        -  number of Fourier coefficients that should be
%              omitted
% pricetype -  char specifying whether prices are log prices
%              or not
%
% See also ftvar.m, ftcovar.m, pearson.m
%
% $Author Chanel Malherbe

%% Assign the defaults
```

```
%Number of Fourier coefficients to be included in
%the estimation defaults to the Nyquist frequency
nrfc = N/2;
%Number of Fourier coefficients to exclude from the
%beginning
n0 = 1;
%Specify whether data series is price or log of price
pricetype = 'logprice';

%% Manage polimorphism
switch nargin
    case 3
        nrfc = varargin{1};
    case 4
        nrfc = varargin{1};
        n0 = varargin{2};
    case 5
        nrfc = varargin{1};
        n0 = varargin{2};
        pricetype = varargin{3};
    otherwise
        if (nargin > 4);
            error(['BadInputArguments',...
   'Incorrect Input Arguments']);
        end
end

%% Check the validity of the input
%Check that the number of fourier coefficients is larger
%than zero
if (nrfc <= 0);
    nrfc = 1;
    warning('Number of Fourier coefficients is zero');
end
%Check that n0 is positive
if (n0 <= 0); n0 = 0;warning('N0 is zero'); end
%Check that data is a strucutre
if ~isstruct(data); error('Incorrect data type');end
%Check whether price is log price, if not, take logarithm
if strcmp(pricetype,'price')
    for i = 1:length(data)
        data(i).PRC = log(data(i).PRC);
    end
end
```

```
%% Calculate the volatility
kc = [1:round(nrfc)];
for i = 1:length(data);
    Ni = length(data(i).dates);
    tsi = rescale(data(i).dates);
    reti = diff(data(i).PRC);
    tki = tsi(2:Ni)*kc;
    fca(:,1) =  (1/pi)*(reti'*cos(tki))';
    fcb(:,1) = -(1/pi)*(reti'*sin(tki))';
    var(i) = (pi^2/(nrfc + 1 - n0))*(sum(fca(n0:round(nrfc),1) ...
        .*fca(n0:round(nrfc),1)) + sum(fcb(n0:round(nrfc),1) ...
        .*fcb(n0:round(nrfc),1)));
end
```

A.2 Correlation estimation using the Fourier method

```
function C = ftcor(data,N,varargin)
% FTCOR calculates correlation of high frequency data
%
% COR  = FTCOR(data,N,nrfc,n0) returns a S-by-S matrix
% containing the pairwise linear correlation calculated
% by the method proposed by Malliavin and Mancino in
% their paper 'Fourier series method for measurement
% of multivariate volatilities'.
%
% data      - struct containing dates (data.dates) and
%             prices (data.PRC) for each series.
% N         - length of time scale
% nrfc      - number of fourier coefficients to be included
%             in estimation
% n0        - number of fourier coefficients that should
%             be emitted
% pricetype - char specifying whether prices are log prices
%             or not
%
% See also ftvar.m, ftcovar.m, pearson.m
%
% $Author Chanel Malherbe

%% Assign the defaults
%Number of Fourier coefficients to be included in the
%estimation defaults to the Nyquist frequency
```

```
nrfc = N/2;
%Number of Fourier coefficients to exclude from
%the beginning
n0 = 1;
%Specify whether data series is price or log of price
pricetype = 'logprice';

%% Manage polimorphism
switch nargin
    case 3
        nrfc = varargin{1};
    case 4
        nrfc = varargin{1};
        n0 = varargin{2};
    case 5
        nrfc = varargin{1};
        n0 = varargin{2};
        pricetype = varargin{3};
    otherwise
        if (nargin > 4);
            error(['BadInputArguments', ...
               'Incorrect Input Arguments']);
        end
end

%% Check the validity of the input
%Check that the number of fourier coefficients is
%larger than zero
if (nrfc <= 0);
    nrfc = 1;
    warning('Number of Fourier coefficients is zero');
end
%Check that n0 is positive
if (n0 <= 0); n0 = 0;warning('N0 is zero'); end
%Check that data is a strucutre
if ~isstruct(data); error('Incorrect data type');end
%Check whether price is log price, if not,
%take logarithm
if strcmp(pricetype,'price')
    for i = 1:length(data)
        data(i).PRC = log(data(i).PRC);
    end
end
```

```
%Calculate the correlation matrix
kc = [1:round(nrfc)];
for i = 1:length(data);
    Ni = length(data(i).dates);
    tsi = rescale(data(i).dates);
    tki = tsi(2:Ni)*kc;
    fca(:,1) =  (1/pi)*(data(i).PRC'*cos(tki))';
    fcb(:,1) = -(1/pi)*(data(i).PRC'*sin(tki))';
    var1 = (pi^2/(nrfc + 1 - n0))*(sum(fca(n0:round(nrfc),1)...
        .*fca(n0:round(nrfc),1)) + sum(fcb(n0:round(nrfc),1)...
        .*fcb(n0:round(nrfc),1)));
    for j = 1:i
        %Determine the size of the input series p1 and p2
        Nj = length(data(j).dates);
        %Calculate the difference between consecutive prices
        tsj = rescale(data(j).dates);
        tkj = tsj(2:Nj)*kc;
        fca(:,2) =  (1/pi)*(data(j).PRC'*cos(tkj))';
        fcb(:,2) = -(1/pi)*(data(j).PRC'*sin(tkj))';
        %Calculate the integrated volatility and covolatility over
%the entire time window
        covar = (pi^2/(nrfc + 1 - n0))*(sum(fca(n0:round(nrfc),1)...
            .*fca(n0:round(nrfc),2)) + sum(fcb(n0:round(nrfc),1)...
            .*fcb(n0:round(nrfc),2)));
        var2 =  (pi^2/(nrfc + 1 - n0))*(sum(fca(n0:round(nrfc),2)...
            .*fca(n0:round(nrfc),2)) + sum(fcb(n0:round(nrfc),2)...
            .*fcb(n0:round(nrfc),2)));
        %Calculate the correlation
        if ((var1>0)&&(var2>0))
            C(i,j) = covar/(sqrt(var1*var2));
        else
            C(i,j) = 0;
        end
        if ~(i==j); C(j,i) = C(i,j);end
    end
end
```

Appendix B

Glossary of Terms

Autocorrelation: A mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except that the same time series is used twice - once in its original form and once lagged one or more time periods.

Bid Price: The price that a buyer is willing to pay for an asset.

Black-Scholes model: The first complete mathematical model for pricing European options on stocks, developed by Fischer Black, Myron Scholes and Robert Merton. It examines the market price, strike price, volatility, time to expiration and interest rates. It is limited to only certain kinds of options.

Calibration: Method for implying a model's parameters from the prices of actively traded options.

Correlation: A statistical measure of the simultaneous change in value of two random numeric variables. Correlation is computed into what is known as the correlation coefficient, which ranges between -1 and +1.

Correlation coefficient: A statistic in which the covariance is scaled to a value between minus one (perfect negative correlation) and plus one (perfect positive correlation).

Covariance: A measure of the degree to which returns on two assets move in tandem. A positive covariance means that asset returns move together; a negative covariance means they vary inversely.

Cross-correlation: A statistical measure timing the movements and prox-

[0]This glossary was compiled using definitions from [16] and [18].

imity of alignment between two different information sets of a series of information.

Derivative: An instrument whose price depends on, or is derived from, the price of one or more underlying assets. This includes, for example, futures contracts, forward contracts, options and swaps.

Diffusion process: Model where value of asset changes continuously without jumps.

Econometrics: The application of statistical theories to economic ones for the purpose of forecasting future trends.

GARCH model: A model for forecasting volatility where the variance rate follows a mean-reverting process.

Generalised Wiener process: A stochastic process where the change in a variable in each short time period of length δt has a normal distribution with mean and variance, both proportional to δt.

Geometric Brownian Motion: A stochastic process often assumed for asset prices where the logarithm of the underlying variable follows a generalised Wiener process.

High frequency data: Data obtained by collecting all prices in a certain period.

Historic volatility: A volatility estimated from historical data.

Implied volatility: Volatility implied from an option price using the Black-Scholes or a similar model.

Ito's Lemma: A result that enables the stochastic process for a function of a variable to be calculated from the stochastic process for the variable itself.

Ito process: Statistical assumptions about the behavior of security prices.

Kurtosis: A statistical measure used to describe the distribution of observed data around the mean.

Lognormal distribution: A variable has a lognormal distribution when the logarithm of the variable has a normal distribution.

Martingale: A zero-drift stochastic process.

Monte Carlo simulation: A mathematical modeling procedure for randomly sampling changes in variables. For a model that has several parameters with statistical properties, pick a set of random values for the parameters and run a simulation. Then pick another set of values, and run it again. Run it many times (often 10,000 times) and build up a statistical distribution of outcomes of the simulation. This distribution of outcomes is then used to answer whatever question you are asking.

Nonparametric volatility methods: Methods for addressing the computation of historical volatility without assuming a functional form of the volatility.

Non-stationary model: A model where the volatility parameters are a function of time.

Normal distribution: In statistics, a theoretical frequency distribution for a set of variable data, usually represented by a bell-shaped curve symmetrical about the mean.

Option: The right to buy or sell an asset.

Parametric volatility methods: Methods where the expected volatility is modelled using a functional form of the variables observed in the market.

Pearson coefficient: A type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale.

Realised volatility: Method for the measurement of volatility where the daily volatility is measured by a sum of short-term intraday squared returns, say for example, at a 10 minute time horizon.

Semi-martingale: A stochastic process which can be decomposed into a drift term and a local martingale.

Skewness: A statistical term used to describe a situation's asymmetry in relation to a normal distribution.

Standard deviation: A measure of the variation in a distribution, equal to the square root of the arithmetic mean of the squares of the deviations from the arithmetic mean; the square root of the variance.

Stochastic: Involving or containing a random variable or variables; involving chance or probability.

Stochastic process: An equation describing the probabilistic behaviour of a stochastic variable.

Stochastic variable: A variable whose future value is uncertain.

Variance: The dispersion of a variable. The square of the standard deviation.

Variance-covariance matrix: A matrix showing variances of, and covariances between, a number of different market variables.

Volatility: A statistical measure of the dispersion of returns for a given security or market index. Volatility can either be measured by using the standard deviation or variance between returns from that same security or market index. Commonly, the higher the volatility, the riskier the security.

Bibliography

[1] S. Aboura, *A Survey on Implied Theories: The Volatility Models*, (2003).

[2] W.A. Brock A.G. Malliaris, *Stochastic Methods in Economics and Finance*, Elsevier Science Publishers B.V., 1991.

[3] S.J. Koopman B. Jungbacker, *Model-based measurement of actual volatility in high-frequency data*, Discussion paper TI, 05-002/4 (2005).

[4] T. Bollerslev, *Financial econometrics: Past developments and future challenges*, Journal of Applied Econometrics **100** (2001), 41–51.

[5] "JSE Stock Exchange data", *"high frequency data"*, "May 2002 - October 2004".

[6] R. Reno E. Barucci, *On measuring volatility and the GARCH forecasting performance*, Journal of International Financial Markets, Institutions and Money **12** (2001), 182–200.

[7] ———, *On Measuring Volatility of Diffusion Processes with High Frequency Data*, Economics Letters **72** (2002), 371–378.

[8] R.Reno E. Barucci, M.E. Mancino, *Volatility Estimate via Fourier Analysis*, Finanza Computazionale, Atti della Scuola Estiva (2000), pp.273–291.

[9] A. Lunde E. Hog, *Wavelet estimation of integrated volatility*, Computing in Economics and Finance,Society for Computational Economics **274** (2003).

[10] R.F. Engle, *Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation*, Econometrica **45** (1982), 987–1007.

[11] T. Epps, *Comovements in Stock Prices in the Very Short Run*, Journal of the American Statistical Association (1979), 291–298.

[12] M. Scholes F. Black, *The valuation of option contract and a test of market efficiency*, Journal of Finance **27** (1972), 399–418.

[13] J.R. Russel F.M. Bandi, *Microstructure noise, realized volatility, and optimal sampling*, Econometric Society Latin American Meetings,220, Econometric Society (2004).

[14] R.N. Mantegna G. Bonanno, F. Lillo, *High-frequency cross-correlation in a set of stocks*, Quantitative Finance **1** (2001), 96–104.

[15] A. Trapletti G. Zumbach, F. Corsi, *Efficient Estimation of Volatility using High Frequency data*, Electronic paper (2002).

[16] J.C. Hull, *Options, futures, and other derivatives*, fifth edition ed., Prentice Hall, 2003.

[17] The Mathworks Inc, *Matlab student version*, 1996.

[18] Investopedia, *www.investopedia.com*, 2006.

[19] T. Kanatani, *High frequency data and realized volatility*, Ph.D. thesis, Graduate School of Economics, Kyoto University, Yoshide Honmachi, Skayo-Ku, Kyoto 6068501, Japan, December 2004.

[20] ———, *Integrated volatility measuring from unevenly sampled observations*, Economics Bulletin **3** (2004), no. 36, 1–8.

[21] ———, *Optimally weighted realized volatility*, CAEA Discussion Paper No.26 (2004).

[22] P. Malliavin M.E. Mancino, *Fourier series method for measurement of multivariate volatilities*, Finance and Stochastics **6** (2002), no. 1, 49–61.

[23] ———, *Harmonic analysis methods for nonparametric estimation of volatility*, Working Paper (2003), Presented at the First Italian Congress of Econometric and Empirical Economics, 2005.

[24] T. Mikosch, *Elementary Stochastic Calculus with Finance in View*, World Scientific Publishing Co. Pte. Ltd, 1998.

[25] R.Gencay M.M. Dacorogna, R.B. Olsen U. Muller, and O.V. Pictet, *An introduction to high-frequency finance*, Academic Press, 2000.

[26] P.H. Frederiksen M.O. Nielsen, *Finite sample accuracy of integrated volatility estimators*, Working Paper, Cornell University (2004).

[27] S. Yhan O. Ledoit, P. Santa-Clara, *Relative pricing of options with stochastic volatility*, Finance, Paper 9-98 (2002).

[28] N. Shephard O.E. Barndorff-Nielsen, *Econometric analysis of realised volatility and its use in estimating stochastic volatility models*, Journal of the Royal Statistical Society **Series B** (2001), no. 64.

[29] ———, *Econometric analysis of realized volatility and its use in estimating stochastic volatility models*, Journal of the Royal Statistical Society Series B **64** (2002), 253–280.

[30] ———, *Variation, jumps, market frictions and high frequency data in financial econometrics*, Prepared for the invited sumposium on Financial Econometrics, 9th World Congress of the Econometric Society, August 2005.

[31] N. Shepherd O.E. Barndorff-Nielsen, *Estimating quadratic variation using realised volatility*, Journal of Applied Econometrics **17** (2000), 457–477.

[32] B. Oksendal, *Stochastic differential equations*, Springer, New York, 1998.

[33] G. Iori O.V. Precup, *A comparison of high-frequency cross-correlation measures*, Physica A **344/1-2** (2004), 252–256.

[34] G.Iori O.V. Precup, *Cross-Correlation Measures in the High-Frequency Domain*, European Journal of Finance, Submitted (2006).

[35] A. Lunde P.R. Hansen, *An unbiased measure of realized variance*, (2004).

[36] K. Winkelmann R. Litterman, *Estimating covariance matrices*, Goldman Sachs Risk Management Series, January 1998.

[37] R. Reno, *A closer look at the Epps effect*, International Journal of Theoretical and Applied Finance **6** (2003), no. 1, 87–102.

[38] T. Bollerslev T. Andersen, *Answering the skeptics: Yes, standard volatility models do provide accurate forecasts*, International Economic Review **39** (1998), 885–905.

[39] ———, *Parametric and nonparametric volatility measurement*, NBER Working Paper No. T0279, 2002.

[40] T. Bollerslev T.G. Andersen and P. Labys F.X. Diebold, *The distribution of exchange rate volatility*, Journal of the American Statistical Association **96** (2001), 42–55.

[41] J. Xu, *Pricing and hedging options under stochastic volatility*, Master's thesis, The University of British Columbia, 2005.

[42] P.A. Mykland Y. Ait-Sahalia and L. Zhang, *How often to sample a continuous-time process in the presence of market micro-structure noise*, The Review of Financial Studies **18** (2005), no. 2.