Faculty of Science and Engineering
Department of Mathematics and Statistics

Measuring and Modelling the Volatility of Financial Time Series

Phan Anh Chuong Luong

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

February 2016

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: Phan Anh Chuong Luong

Date: February 2016

Measuring and modelling the volatility of financial time series

by

Phan Anh Chuong Luong

Submitted to the Department of Mathematics & Statistics
in February 2016, in fulfilment of the
requirements for the Degree of Doctor of Philosophy

Abstract

The primary purpose of this thesis is to investigate the measures and models of volatility for the financial time series. Whilst the field of study itself is not new, measuring and modelling of the financial volatility are still crucial and challenging tasks, as volatility is the heart of many financial applications. The thesis applies the historical volatility measure and the realised volatility measure to study the impact of sampling frequency on the volatility estimation and address the dependency of volatility on sampling frequency. It is suggested to employ stochastic delay equations to model the dependency of volatility on timescales. The thesis also addresses the impact of the underlying price movement on the volatility measure. By using the option price data, the "purified" implied volatility process is proposed to overcome this issue. In addition, regression techniques are applied on the future volatility to investigate the information contained in this new implied volatility process. It is shown that this new process contains information about future volatility. Furthermore, the heterogeneous autoregressive model and random forest algorithm are shown to improve the accuracy in forecasting of future volatility.

Publications and presentations during candidature

The following papers were published or accepted for publication during the PhD candidature:

Journal Articles

- Luong, C., & Dokuchaev, N. (2016). Modelling dependency of volatility on sampling frequency via delay equations. Annals of Financial Economics, TBA. In press. (Discussed in Chapter 3)
- Luong, C., & Dokuchaev, N. (2014). Analysis of market volatility via a dynamically purified option price process. Annals of Financial Economics, 9(03), 1450006. (Discussed in Chapter 4)

Conference presentations

- Luong, C., & Dokuchaev, N. (2014). Modelling dependency of volatility on time scales with delay equations. In 6th Australasian Actuarial Education and Research Symposium, December 8-10, 2014, Curtin University, Western Australia, Australia.
- Luong, C., & Dokuchaev, N. (2014). Dependency of volatility on sampling frequency via delay equations. In IMS-FPS 2014, July 3-5, 2014, University of Technology, Sydney, Australia.
- Luong, C., & Dokuchaev, N. (2013). On the implied volatility from a “purified” option price process. In Vietnam International Applied Mathematics Conference (VIAMC), Dec 19, 2013, Ho Chi Minh City: Vietnam Mathematical Society.

Acknowledgements

I would like to express my gratitude to my supervisor, Associate Professor Nikolai Dokuchaev, whose understanding, expertise and patience, added considerably to the success of this thesis. Without his guidance, encouragement and persistent helps, in addition to his financial support through the 2013-2016 Australian Research Council (ARC) grant of Australia DP120100928, this thesis would not have been possible.

I would like to thanks Professor Yong Hong Wu, the Chairperson of my Thesis Committee. He is always there to assist and give me valuable advice. I am also grateful to my Associate Supervisor Dr Honglei Xu for providing indispensable advice, information and encouragement.

I would like to acknowledge the financial supports that I have been provided with by Curtin University, specifically by the Department of Mathematics and Statistics - ARC Discovery Grant Scholarship.

I gratefully acknowledge Curtin University for the Support for Conference Attendance, pertaining to my attendance to the Vietnam International Applied Mathematics Conference (VIAMC 2013), Vietnam, December 19th - 20th, 2013; and the Fourth Institute for Mathematical Statistics - Finance, Probability and Statistics Workshop (IMS-FPS 2014), Sydney, July 3rd - 5th, 2014. I also gratefully acknowledge the Department of Mathematics & Statistics, Curtin University for the Support for Conference Attendance pertaining to my attendance to the 6th Australasian Actuarial Education and Research Symposium, Perth, Dec. 8th - 9th 2014.

I would like to acknowledge supports from the following organisations: the Securities Industry Research Centre of Asia-Pacific (SIRCA) for their provision of access to the Thomson Reuters Tick History (TRTH) financial data; The Nectar Research Cloud for their provision of computing infrastructure and services that allow me to develop computationally intensive models and collaboration environments.

I would like to thanks my friends and colleagues at Curtin University for our philosophical debates, exchanges of knowledge, and skills

Finally, I would like to thank my family and relatives for the support they provided me through my entire life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A good understanding of volatility plays a critical role in financial investment. It is the major measure of risk in modern finance and is at the heart of many financial applications such as pricing of options, hedging strategies, constructing optimal portfolios, and derivatives. With the increasing usage of financial volatility, both academia and financial practitioners have pioneered and ventured into developing methods of estimating and measuring the volatility of financial time series based on various models. However, some existing measures and models are limited by their underlying assumptions or have been proven not being able to withstand their assumptions in practice. In addition, most of the existing techniques and models have been developed under the low-frequency settings. With the availability of high-frequency data, it is of interest to study further the financial volatility under different sampling frequencies and to model their behaviours. It appears that further contributions can be made to the measures and models of univariate financial markets' volatility.

Measuring and modelling the financial volatility are crucial and challenging tasks. While the financial returns are observable from the price processes, the financial volatilities are not observed directly. It requires the use of some estimation techniques such as historical measures [6, 47] which are based on the historical returns movements, implied measures [37, 121] which are based on the imputed value of volatility from given financial models, or stochastic measures [117] as estimated from some underlying dynamic stochastic processes. Hence, it is essential to have a good understanding about these estimators and acknowledge their gaps in forecast performance of future volatility, especially when the economic outcome is often determined by the forecast performance of the volatility forecasting models. It is also a challenging task. As discussed, there are a number of estimators that can be used for measuring the financial volatility. One must select a volatility estimator that he/she think it can best describe the evolution of the underlying asset. In addition, the selected estimator needs to be able to model and forecast the future volatility accurately and effectively.

As for measuring the volatility of financial market, Engle[52] and Bollerslev [29] first proposed the ARCH model and the GARCH model. These models have been extended in numbers of directions based on the empirical evidences that the volatility process is non-linear, asymmetry and has long memory. Such extensions can be referred to EGARCH [97], GJR-GARCH [63], AGARCH [51], and TGARCH [123]. With the availability of high-frequency data, Andersen et al. [9] provide a framework for integration of high-frequency intraday data into historical volatility estimation, so-called realised volatility. This measure is built on the theory of continuous-time and arbitrage-free processes with the theory of quadratic variation. In comparison with the GARCH-type measures, realised volatility is more preferred as it is a model-free measure. Hence, it provides convenience for calculation, similar to the classical historical volatility [47].

However, volatility measured at different frequencies (e.g. hourly, intraday, daily or monthly) has different information content [91]. It is observed in [126] that the volatility increases when the sampling frequency increases for some financial time series. It is also seen that the volatility may decrease when more frequent samples are used for other financial time series. For the data generated from the classical Black-Scholes' [28] market model or the Heston's [70] stochastic volatility model, this effect does not take place, i.e. the volatility is independent of the sampling frequency. Therefore, the observed dependency of the volatility on sampling frequency requires modifications of the classical model. To address this, it is suggested to develop a continuous time model for stock prices such that the volatility calculated via different sampling frequency may take certain preselected values. It appears that Ito equations with delay feature this property. The presence of additional parameters that describe the delay enables us to capture the difference of volatility estimated at alternative sampling rates, unlike existing processes. In this thesis, we consider the simplest linear delay equation with only one delay term. This new delay term is found to be capable of matching the volatility from the simulated price process with the volatility from the historical data, on three different sampling frequencies.

Another measure of volatility that is studied in this thesis is the implied volatility. While the historical volatility and the realised volatility describe the volatility observed from the returns processes, the implied (or imputed) volatility is the theoretical measure which describe the volatility of an underlying financial instrument. This implied volatility is implicitly derived by inverting the option pricing models while other parameters are provided and kept unchanged. For example, given option pricing models, by using the observations of the price processes and the related values given by the market, one can derive the value of the volatility "implied" by the market. The implied volatilities are usually defined as the inverse of the Black-Scholes [28] pricing formula applied to the observed market prices. The option prices and the implied volatilities are fluctuating along with the underlying assets prices; they have a stochastic "random walk" type pattern of movement, similar to the stock prices. Since the implied volatility depends on the strike price and the expiration time, often one would find it difficult to decide which implied volatility to use among all possible versions of the implied volatility when describing the market expectations of the degrees of the future stock price deviations. Alternatively, one can assume that the entire implied volatility surface has a known initial values and follows a stochastic process. Such interpretations are as the implied volatility indexes by Whaley [121]ls VXO or Carr and Wu [37]'s VIX. For this purpose, the volatility indexes were created. However, the implied volatility has some limitations. Firstly, it is restricted by the assumptions made in the underlying model (i.e. Black-Scholes' model). This often has an artificially induced upward bias [38] on the estimation. Secondly, there is impact of the price movement on the option prices hence generating another type of bias on the implied volatility. While the prior issue is not easy to be addressed, it is suggested from this thesis that a "dynamically purified" option price process can be constructed to reduce the bias caused by the latter issue. Our study suggest that the "purified" implied volatility calculated from this process could be more informative than the traditional implied volatility such as VXO and VIX indexes. In theory, the proposed option price process can eliminate the impact of the stock price movements. However, this would only be possible if the option prices were available for continuous sets of strike prices and expiration times. In practice, we have to use finite sets of available prices. To restore this process from incomplete sets of available option prices, it is suggested to use the first order Taylor series interpolation and quadratic interpolation. From this approach, the proposed option price process requires less option data while maintaining the accuracy of the approximation. Furthermore, a number of regression models using the implied volatility index VIX and the implied volatility from the purified option prices as predictors of the future volatility were studied. It is found that for some selected dataset and constructed models, the new implied volatility has superior information about future volatility than that of the traditional implied volatility indexes.

Choosing a model to generate volatility forecasts also play an important role in financial applications as the forecast performance of the forecasting models can influence the economic outcomes. There have been many forecasting models that have been developed to predict the realised volatility, ranging from very simple to complex models; such as ARCH model and GARCH model, HAR (heterogeneous autogressive model for realised volatility) model (HAR) [43] and HEAVY (High frEquency bAsed VolatilitY) model [111]. This thesis focuses on the implementation of the HAR model. This model is developed in accordance with the heterogeneous market hypothesis proposed by Muller[96] and the long memory character of realised volatility by Andersen [9]. Empirical studies have shown that the HAR model has high forecasting performance on future volatility, especially for the out-of-sample data given different time horizons ([43], [82]). By incorporating this model with the proposed implied volatility process, it is suggested that the forecasting of volatility can further be improved via the use of machine learning techniques such as random forests algorithm.

This thesis is structured as follows:

Chapter 2 reviews definitions of volatility and their measures. The chapter also provides some empirical results with stylize facts of volatility.

Chapter 3 addresses the dependency of volatility on the sampling frequency. The chapter shows that delay equations allow to model the price processes with volatility that increases when the sampling rates increase, as well as the inverse phenomena where the volatility decreases with the increase in sampling frequencies. An empirical study is demonstrated whether analytical and simulation results apply to the proposed settings.

Chapter 4 introduces a dynamic estimation of implied volatility for financial time series. This implied volatility is inferred from an artificial "dynamically purified" price process which allows to eliminate the impact of the stock price movements. The chapter also investigates the information of this proposed implied volatility in forecasting of the future volatility, in comparison with the traditional implied volatility processes such as the volatility index VIX.

Chapter 5 studies the forecasting of realised volatility for financial time series using heterogeneous autoregressive model (HAR) and random forest, a machine learning algorithm. The chapter extends the existing HAR model by including the proposed "purified" implied volatility proposed in Chapter 4, and shows that it is possible to improve the forecast of both direction and magnitude of the realised volatility. The demonstration on the forecasting power of this new class of model is shown via experiment with historical high frequency financial data with different forecasting horizons.

Chapter 6 summarises the main contributions of this work and identifies potential directions for future research.

Chapter 2

Some preliminaries: volatility and its estimation

2.1 Rationale behind financial volatilities

In general, volatility is a statistical measure of the dispersion of a given time series. In finance, the time series is associated with the continuous price process. Volatility is an important measure in many financial applications such as asset pricing, derivatives pricing, portfolio optimizations and risk management. Estimating the volatility is a challenging task as volatility is not an observable process like the return process which can be obtained directly from the price process. The estimates of volatility can be categorised into three main classes: historical volatility, implied volatility and stochastic volatility. These classes of estimators can be either non-parametric (based on a scale estimator) or parametric (based on regression estimation of a flexible specification of the return process).

In this chapter, we provide backgrounds on the volatility estimators of the financial series that were used throughout this thesis. We will review some literature on the underlying theories, assumptions and robustness of these measures. Although there have been extensive reviews conducted such as [8], this chapter highlights some recent techniques that tackle the issues in measuring and forecasting financial volatility such as jump in price levels and market micro-structure noise in the financial data. The backgrounds included in this chapter result in the development of the price model discussed in Chapter 3, the construction of the new implied volatility process in Chapter 4 and the forecasting techniques proposed in Chapter 5.

2.2 Definition of volatility for continuous time price model

Let $P(t)$ denote the price process of an asset at time t, and suppose that the dynamic of the process $p(t) = \ln P(t)$ is described by the stochastic differential equation

$$dp(t) = \mu(t)dt + \sigma(t)dW(t). \tag{2.2.1}$$

This is the so-called Ito equation [76]. Here, $W(t)$ is a Wiener process (a standard Brownian motion [49]). In particular, $W(t)$ has Gaussian increments; $W(t+s) - W(s)$ is normally distributed with mean 0 and variance s, i.e. $W(t+s) - W(s) \sim N(0, s)$. The process $\mu(t)$ is called the drift and $\sigma(t)$ is called the diffusion coefficient. In financial applications, $\sigma(t)$ is known as the volatility of the price $P(t)$. Both $\mu(t)$ and $\sigma(t)$ can be random and must be independent of $W(t+s) - W(s)$.

In Chapter 3, we consider $\mu(t) = \lambda(p(t) - p(t - \varrho))$ for $\lambda > 0$ and $\varrho > 0$ for the delay equations. In Chapter 4, we consider $\mu(t) = a(t) - \sigma(t)^2/2$ for the underlying asset where $a(t)$ is the appreciation rate. We also consider a more general stochastic differential equation for the price process with jumps in Chapter 5 (see also Section 2.3.2 below).

2.3 Historical volatility

2.3.1 Classical estimator of volatility

In practice, it is not feasible to obtain $\sigma(t)$ at a given time t. However, it is possible to estimate the values of $\sigma(t)$ within a period $[t - \triangle t, t]$ for $\triangle t > 0$. Often, volatility is estimated by using the historical returns collected during a given period. We consider below estimation of $v(t)$ such that

$$v(t) = \frac{1}{\triangle t}\int_{t-\triangle t}^{t} \sigma^2(\tau)d\tau. \tag{2.3.1}$$

Let us consider constructing an estimate $v(t)$ from observed prices $P(t_k)$ with $t_k \in [t-\triangle t, t]$, $k = n_0, n_0+1, ..., n$ where the time t_k are observed at equal intervals (*equispaced*), $\delta = t_k - t_{k-1}$, $t_{n_0} = t - \triangle t$ and $t_n = t$ (or $\triangle t = (n - n_0)\delta$). The *sample variance* in discrete time series is estimated by

$$\hat{v}(t_n) = \frac{1}{\triangle t} \sum_{k=n_0+1}^{n} (E_n - R_k)^2, \tag{2.3.2}$$

where

$$\begin{aligned} E_n &= \frac{1}{n - n_0} \sum_{k=n_0+1}^{n} R_k, \\ R_k &= \ln P(t_k) - \ln P(t_{k-1}). \end{aligned}$$

The square-root value of equation (2.3.2) is often known as the *classical historical volatility*. It is important to note that the assumption of collecting data at equal interval of unit of time is crucial, as in practice, trades occur at discrete points in time.

2.3.2 Realised volatility

With the availability of high-frequency data, Andersen et al. [9] introduce an alternative measure of volatility using all available high-frequency intraday data. This measure is so-called *realised volatility*. Similar to the *classical historical volatility*, the *realised volatility* is a nonparametric expost estimate of return realisations over a fixed time interval. In particular, this measure was built on the theory of continuous-time and arbitrage-free processes with the theory of quadratic variation. Let us extend the assumptions used for *classical historical volatility* estimator to the followings.

Let $P(t)$ represent the asset price which is recorded at equally-spaced discrete points from the financial market within a given time interval $[t - \triangle t, t]$, where $0 \leqslant t - \triangle t \leqslant t \leqslant T$, $p(t) = \ln P(t)$ and $r(t, \triangle t) = p(t) - p(t - \triangle t)$. We assume that $P(t)$ is governed by the Ito equation (2.2.1) with $\mu(t)$ and $\sigma(t)$ represent the instantaneous conditional mean and volatility of the return respectively. As such,

$$r(t, \triangle t) = p(t) - p(t - \triangle t) = \int_{t-\triangle t}^{t} \mu(\tau) d\tau + \int_{t-\triangle t}^{t} \sigma(\tau) W(t), \tag{2.3.3}$$

and the quadratic variation is

$$QV(t, \triangle t) = \int_{t-\triangle t}^{t} \sigma^2(\tau) d\tau. \tag{2.3.4}$$

Following this result, the developments of $\mu(t)$ do not relate directly to the sample path variation of return. Andersen and Benzoni [4] explain that the term $\mu(t)(dt)$ is of lower order (second order) than the diffusion component $\sigma(t)dW(t)$. As a result, the cumulated returns from many high-frequency values over a short time interval can be neglected. Therefore, the variance in equation (2.3.1) coincides with the quadratic variation $QV(t, \triangle t)$. Now, let us assume that the time interval $[t - \triangle t, t]$ is observed evenly at $\frac{j}{n}$ steps in discrete time, i.e. $t - \triangle t + \frac{j}{n}, j = 1, ..., n \cdot \triangle t$. Under semi-martingale theory [68], the realised variance measure converges in probability to the quadratic variation as the sampling frequency n increases, i.e.

$$RV(t, \triangle t; n) \longrightarrow QV(t, \triangle t) \quad as \quad n \longrightarrow \infty. \tag{2.3.5}$$

Protter [103] suggests that the link between realised volatility measures is based on high-frequency returns and the underlying price process's quadratic variation, which was applied in the context of measuring empirical volatility by Andersen and Bollerslev [13].

Andersen et al. [9] suggests to estimate the *realised volatility* (RV) of $P(t)$ by

$$RV(t, \triangle t; n) = \sum_{j=1}^{n \cdot \triangle t} r(t - \triangle t + \frac{j}{n}, \frac{1}{n})^2. \tag{2.3.6}$$

For one-day period volatility estimate, the above formula can be rewritten as

$$RV_{t-\triangle t,t} = \sqrt{\sum_{j=0}^{M-1} r^2_{t-j\delta}}, \tag{2.3.7}$$

where $[t - \triangle t, t]$ is the duration of one day, $\delta = \frac{1}{M}$, $r_{t-j\delta} = p(t - j\delta) - p(t - (j + 1)\delta)$, and returns are assumed to be sampled evenly at δ time step with M observations in that time interval. Since we assume that the data is collected at equal-space, δ can be collected at second-, minute- and hour-frequencies. We will revisit this volatility estimator in Chapter 3, Section 3.2 and Chapter 5, Section 5.2.

While this estimation is simple and easy to compute for high-frequency data, the main challenges are at dealing with (1) jumps in the price level and (2) microstructure noise in high-frequency data.

Jumps in the price level

By the properties of Ito equation [76], the return process in equation (2.2.1) is path-wise continuous. This does not allow to model the markets where asset prices exhibit sudden discrete movements when unexpected events are introduced to the market. Such events cause jumps in the price. Let us discuss how the realised volatility can be modified to allow for the processes with jumps. Let the continuous time return process equation (2.2.1) with the presence of jumps in returns be

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t), \tag{2.3.8}$$

where $\kappa(t)$ denotes the magnitude of the jump in the returns process if there is a jump occur at time t, q is a Poisson process uncorrelated with W which follows the jump intensity $\lambda(t)$ such that $\mathbf{P}(dq(t) = 1) = \lambda(t)dt$ for positive and finite $\lambda(t)$. Here $\mathbf{P}$ is a probability measure. In addition to the volatility $\sigma(t)$, we are also interested in estimating $\kappa(t)$.

The quadratic variation process over the time interval $[t - \triangle t, t]$ is now

$$QV(t, \triangle t) = \int_{t-\triangle t}^{t} \sigma^2(\tau)d\tau + \sum_{t-\triangle t \leqslant \tau \leqslant t} J^2(\tau), \tag{2.3.9}$$

where $J(t) \equiv \kappa(t)dq(t)$ is the size of the jump and is non-zero if there is a jump at time t. The question arises is how to measure $\sum_{t-\triangle t \leqslant \tau \leqslant t} J^2(\tau)$.

Barndorff-Nielsen and Shephard [18] address this issue by introducing the estimation of *realised bi-power variation* (BV) where

$$BV(t, \triangle t; n) = \frac{\pi}{2} \sum_{i=2}^{n \cdot \triangle t} |r(t - \triangle t + \frac{i \triangle t}{n}, \frac{1}{n})||r(t - \triangle t) + \frac{(i-1)k}{n}, \frac{1}{n})|, \tag{2.3.10}$$

and show that the bi-power variation is robust to the presence of jumps as

$$\sum_{t-\triangle t \leqslant \tau \leqslant t} J^2(\tau) \equiv max\{RV(t, \delta; n) - BV(t, \triangle t; n), 0\}. \tag{2.3.11}$$

In fact, under semi-martingale theory [103], it can be showed that

$$RV(t, \triangle t; n) - BV(t, \triangle t; n) \longrightarrow \sum_{t-\triangle t \leqslant \tau \leqslant t} J^2(\tau) \quad \text{as } n \to \infty. \tag{2.3.12}$$

Microstructure noise in high-frequency data

As discussed, the approximation in equation (2.3.5) and equation (2.3.12) is converged as the sampling frequency n increases. This means that our realised volatility estimators equation (2.3.6) requires continuous price records. However, in practice, market prices are not continuous as trades occur randomly in discrete time. This results in:

- Sample distortions: since the time at which prices arrive is random, the discretisation of the price at a finite arrival time between $[t - \triangle, t]$ for each $\frac{j}{n}$ step will create distortions as one will need to decide which statistical price to record (i.e. maximum, minimum or average) and at which frequency is accepted as a proxy for the continuous series. In addition, the sample distortion bias increases with the jumps in price.

- Spurious autocorrelations [100]: this happens when we look at ultra-high frequency return series where the discreteness of price, rounding values, bid-ask spread, time-delay and data recording errors can significantly introduces biases that are correlated to each other and consequently inflate the realised volatility measures.

In order to tackle these problems, one can use alternative quadratic variance estimators that are less biased due to microstructure noise or sample price data at an optimal frequency. For example, Huang and Tauchen [72] and Andersen et al. [7] show that by using staggered returns, the bias generated by spurious correlations in returns due to the effect of noise such as the bid-ask spread is reduced. As such, the bi-power variation becomes

$$BV(t, \triangle t; n) = \frac{\pi}{2} \frac{n\triangle t}{n\triangle t - 1 - i} \sum_{i=2+j}^{n\cdot\triangle t} |r_{t_i}||r_{t_{i-1-j}}|, \tag{2.3.13}$$

where j is the offset chosen based on the order of the autocorrelation in the return process.

2.3.3 Other methods of estimating volatility

Other estimations of volatility include the estimations with extreme values of returns such as Brandt and Diebold [31] , Parkinson [99], Alizadah et al. [2], Garman and Klass [60] and Gallant et al. [58]. For tackling the microstructure noise issue, Zhang, Mykland and Aït-Sahalia [126] and Bandi and Russell [16] suggest to use optimal sampling schemes to remove the biases (TSAVGRV), whereas Barndorff-Nielsen et al. [21] suggest to use a kernel-based technique (KernelRV). Another estimation is introduced by Zhang [124] which suggests the use of adjusted realised volatility based on different time scales (TSRV). For the jumps correction, we have Andersen et al. [12] with the jump-robust volatility estimator using the nearest neighbour truncation (minRV and medRV), Gobbi and Mancini [99] with co-jumps coefficient between the diffusion parts given discrete observations (ThresholdRV), Boudt and Zhang [30] with the jump robust estimator using two time scale covariance (RTSRV).

The following table summarises the estimators of these measures with respect to their robustness in term of jumps correction and market microstructure noise.

Robustness of volatility estimators with respect to jumps and microstructure noise.

Estimator	Jump robust	Microstructure noise robust
RV (Andersen et al. [9])		
BVRV (Barndorff-Nielsen & Shephard [18])	✓	
minRV (Andersen et al. [12])	✓	
medRV (Andersen et al. [12])	✓	
ThresholdRV (Gobbi and Mancini [99])	✓	
TSAVGRV (Zhang et al. [126])		✓
TSRV (Zhang [124])	✓	✓
RTSRV (Boundt and Zhang [30])	✓	✓
KernelRV (Barndorff-Nielsen et al. [21])		✓

2.4 Implied volatility

While the historical volatility and the realised volatility describe the volatility observed from the returns processes, the implied (or imputed) volatility is the theoretical measure which describe the future volatility of an underlying financial instrument. This *implied volatility* is implicitly derived by inverting the option pricing models while other parameters are provided and kept unchanged. In other words, for a given option pricing models, by using the observations of the price processes and the related values given by the market, one can derive the value of the volatility “implied” by the market. The most common option pricing model used for computing the implied volatility is the Black and Scholes’ option price model [28]. We will discuss this model in detail in Chapter 4, Section 2.

From equation (4.2.4), it is seen that the call option and put option use the set of $\{x, t, \sigma, r, K\}$ as the input for their evaluations. The set $\{x, t, r, K\}$ are observable and can be obtained directly from the market data. However, σ is not observed directly and is often estimated by the market makers. For example, one may use the historical volatility as an input for the Black-Scholes option price, while others may use extreme values estimators or realised volatility. While there is no closed-form for inverting the parameter σ, the derivative of the Black-Scholes' model has a closed-form and non-negative. Therefore, it is possible to use Newton-Raphson algorithm to compute σ. Other techniques can also be used, such as secant method [3] and bisection method [42], to derive the implied volatility from Black-Scholes' option pricing model.

By plotting the implied volatility against the strike prices (moneyness) and time-to-maturity, we obtain the so-called *volatility surface*. The volatility surface allows us to study the dynamics of implied volatility and thus provide us with insights about market movements. One can also assume that the entire implied volatility surface has a known initial values and follows a stochastic process. Such interpretations are as the implied volatility indexes by Whaley [121] or Carr and Wu [37]. We will revisit the implied volatility in Chapter 4.

Chapter 3

Modelling dependency of volatility on sampling frequency

3.1 Introduction

Due to the growth of computing power and data storage capacity, the high-frequency market data is now available for analysis. This creates new computational challenges to both academics and practitioners.

As discussed in the Chapter 2, volatility calculated from different frequencies (or on the different time scales) contains different information. It was often observed that the volatility increases when the sampling frequency increases; see, e.g., [126]. It also appears that the volatility may decrease when more frequent samples are used; the summary of such observations is shown at Table A.4. For the data generated from the classical Black-Scholes market model, this effect does not take place, i.e. the volatility is independent of the sampling frequency. Therefore, the observed dependency of the volatility on sampling frequency requires modifications of the classical model.

It is commonly recognized that the dependence of volatility on different time scales is caused by the presence market micro-structure noise [6, 10, 17]. Market micro-structure noise refers to imperfections in the trading process of financial assets which causes the observed prices to differ from the underlying 'true' price. Previous studies suggest that one needs to use an optimal sampling frequency to ensure the most accurate volatility estimation for the underlying assets (see, e.g., [11]; [17]). Alternatively, by introducing bias-correction in measuring the volatility, such as the two time-scales estimator in [126], we can remove the dependency of volatility on the sampling frequency. For this method, the highest frequency sample is used and the market micro-structure noise is subtracted from a sub-sampled estimator that is calculated by using a 'sparse' sampling frequency.

In this chapter, we readdress the problem of modelling the price process to replicate the effect of volatility depending on the sampling frequency. We suggests to develop a continuous time model for stock prices such that the volatility calculated via different sampling may take certain preselected values. It appears that Ito equations with delay feature this property. The presence of additional parameters that describe the delay enables us to capture the difference of volatility estimated at alternative sampling rates, unlike other price processes such as Geometric Brownian Motion, mean-reversion model, Heston model, and others. So far, we considered the simplest linear delay equation with only one delay term. This new term provides us the capability to match the volatility from the simulated price process with the volatility from the historical data, on three different sampling frequencies.

This chapter is structured as following. Section 3.2 discusses the construction of our model. We then study the roles of each parameter in the model and summarise the findings via Monte-Carlo simulation. We also demonstrate how effectively the proposed model perform in Section 3.3. We later describe the steps for finding the model's parameters and reproduce the time-scale dependent volatility of some historical financial data. Some discussions of future developments are included afterwards.

3.2 The model for dependency of volatility on sampling frequency

We suggest to model the continuous time stock price process $S(t)$ via the following stochastic delay differential equation

$$\begin{aligned} &dR(t) = -\lambda(R(t) - R(t-\varrho))dt + \sigma dw(t), \quad t > 0, \\ &S(t) = e^{R(t)}. \end{aligned} \tag{3.2.1}$$

Here $R(t)$ is the return, $\lambda \in \mathbf{R}$, $\varrho > 0$, and $\sigma > 0$ are some constants, and $w(t)$ is a standard Wiener process [119].

$$dS(t) = S(t)[-\lambda(\log S(t) - \log S(t-\varrho)) + \sigma^2/2]dt + \sigma S(t)dw(t). \tag{3.2.2}$$

The choice of this particular model was based on the rationale that the presence of the mean reversion reduces the variance for the mean-reverting process [119]. This feature was used in financial modelling; see, e.g., [120, 109]. Under the mean-reverting settings, the return at time t tends to reverse to the long-term average of returns, and the variance of the process is lower than for a martingale with the same volatility. However, we found that the mean-reverting model is not particularly useful for the purpose of this paper, since it was difficult to justify a selection of a particular long-term return value. To overcome this, we considered model (3.2.1) with a delay term. One may say that, for the case where $\lambda > 0$, the process is pushed back to its past values at selected and fixed delay rate. Respectively, for the case where $\lambda < 0$, the process is pushed away from its past values; it appears that the case in which $\lambda < 0$ is also significant.

Existence, regularity, and non-arbitrage properties

Stochastic delay differential equations were widely studied, including quite general models with nonlinear dependence on the delayed term were allowed; see, e.g., [14, 77, 92, 93, 114], and the bibliography there. The first market model with a stochastic delay equation for the prices was introduced and investigated in [114], where no-arbitrage properties were established. Unfortunately, the results from [14, 77, 92, 93, 114] cannot be used for our relatively simple model (3.2.1), because the coefficients in equation (3.2.2) with log functions do not satisfy the conditions on regularity imposed therein. By this reason, non-arbitrage properties established in [14, 114] cannot be applied directly to our model (3.2.1). However, it appears that our model (3.2.1) still features the existence, regularity, and non-arbitrage properties. This can be shown as the following.

Assume that $R(t)|_{[-\varrho,0]}$ is a Gaussian process independent on $w(t)|_{t>0}$ and such that its second moment is bounded. Then equation (3.2.1) has a unique strong solution [88] on the time interval $(0, +\infty)$. This solution can be obtained consecutively on the intervals $[0, \varrho]$, $[\varrho, 2\varrho]$,...,$[(k-1)\varrho, k\varrho]$, ... This procedure produces a Gaussian process such that there exists a sequence $\{C_k\}_{k=1}^{\infty}$ such that $\mathbf{E}R(t)^2 \le C_k$ if $t \in [(k-1)\varrho, \varrho]$.

We will assume below that $R(t) = 0$ for $t < 0$. In this case, the appreciation rate for the stock price $S(t)$ is $a(t) = -\lambda(R(t) - R(t-\varrho)) + \sigma^2/2$, i.e., is a Gaussian process. It follows immediately that the Novikov's condition holds for any sufficiently small interval $[\theta, \theta+\varepsilon]$, i.e., $\mathbf{E}\exp\left(\frac{1}{2}\int_{\theta}^{\theta+\varepsilon}|a(s)|^2\sigma^{-2}ds\right) < +\infty$ if $\varepsilon > 0$ is sufficiently small. Therefore, by the Girsanov's Theorem, the process $S(t)$ can be transformed by a probability measure change into a Black-Scholes price process, with the volatility σ, "locally", i.e., on any sufficiently small time interval. This means the "true" volatility of S is σ, the same as for the Black-Scholes model; in the theory, this volatility that can be restored without error from the continuous time observations of the entire path of $S(t)|_{t\in[\theta,\theta+\varepsilon]}$ or $R(t)|_{t\in[\theta,\theta+\varepsilon]}$. In particular, this implies that the standard estimates for volatility converges to σ as sampling frequency converges to infinity (i.e., the sampling interval converges to zero). Therefore, in the limit case of infinite sampling frequency, or continuous time measurements, our price process is indistinguishable from the price process for the classical Black-Scholes market model. However, it appears that, for any given finite sampling frequency, the volatility estimates behave differently for our delay equations and for the Black-Scholes price process. We show below that, for any given finite sampling frequency, the presence of the delay term makes the volatility systematically underestimated if $\lambda > 0$ (respectively, overestimated if $\lambda < 0$); in both cases, the dependence of this systematic bias on the sampling frequency appears to be monotonic.

Further, it can be noted that it is not possible to use the Girsanov's Theorem [62] for an arbitrarily selected time interval $[0, T]$, because it is unclear if the Novikov's condition [98] holds if T is not small enough. A similar but simpler case where $S(t) = e^{G(t)}$, where G was a Gaussian Ornstein-Uhlenbek process, was studied in [46], where it was proved that the Novikov condition holds for an arbitrarily large interval, for this case. The method [46] relied on the Markov properties of the process and cannot be extended on our case of the equation with delay. Therefore, the existence of an equivalent martingale measure for an arbitrarily selected time interval $[0, T]$ is still an open question. However, it appears that the market with the suggested stock price $S(t)$ is arbitrage free with respect to the standard class of the self-financing strategies. More precisely, it appears that there is no a strategy such that $\mathbf{P}(X(T) \geq 0) = 1$, $\mathbf{P}(X(T) > 0) > 0$, $X(0) = 0$, where $X(t)$ is the corresponding wealth generated by a self-financing strategy. It can be shown as the following. Suppose that such a process exists. Let $N > 0$ be such that, for $\varepsilon = T/N$, $\mathbf{E}\exp\left(\frac{1}{2}\int_{k\varepsilon}^{(k+1)\varepsilon} |a(s)|^2\sigma^{-2}ds\right) < +\infty$ for all $k = 0, 1, .., N-1$. The market is equivalent to the Black-Scholes market on any time interval $[k\varepsilon, (k+1)\varepsilon]$. From the absence of an arbitrage for the market defined for $t \in [T-\varepsilon, T]$ considered on the conditional probability space given $\mathcal{F}_{T-\varepsilon}$, it follows that $\mathbf{P}(X(T-\varepsilon) > 0\,|\,\mathcal{F}_{T-\varepsilon}) = 1$. Taking backward steps, we obtain that $\mathbf{P}(X(T-k\varepsilon) > 0\,|\,\mathcal{F}_{T-k\varepsilon}) = 1$ for all $k = 2, ..., N$. Hence $X(0) > 0$, and the process $X(t)$ required to demonstrate the presence of arbitrage does not exists. This makes our price model applicable for derivatives pricing models.

Time discretisation and restrictions on the growth

To study the properties of the proposed model, let us discuss the results via Monte-Carlo simulation.

For the Monte-Carlo simulation, we have to replace stochastic delay differential equation (3.2.1) by the following stochastic delay difference equation with a given delay τ, such that

$$\begin{aligned} &R(t_k) = R(t_{k-1}) - \lambda(R(t_{k-1}) - R(t_{k-\tau}))\delta + \sigma\sqrt{\delta}\xi_k, \\ &S(t_k) = e^{R(t_k)}. \end{aligned} \tag{3.2.3}$$

Here, $k = 1, 2, ...$, $\delta = t_k - t_{k-1}$, and ξ_k are independent and identically distributed random variables from the standard normal distribution; $\tau > 0$ is an integer.

It will be sufficient to study the sample paths of solutions of (3.2.3) created by Monte-Carlo simulation as a substitution of (3.2.1), where $\varrho = \tau\delta$.

It can be noted that equation (3.2.3) represents a linear autoregression AR(τ) with the characteristic polynomial

$$z^\tau = z^{\tau-1} - \lambda\delta(z^{\tau-1} - 1).$$

This polynomial has a root $z = 1$. Therefore, the time series $\{R(t_k)\}$ does not converge to a stationary process as $k \to +\infty$. Let$\{z_1, ..., z_\tau\}$ be the roots of this polynomial, and let as select $z_1 = 1$. We will be using model (3.2.3) for the pairs (τ, λ) such that all other roots $\{z_2, ..., z_\tau\}$ are inside of the open disc $\mathbb{D} \overset{\Delta}{=} \{z \in \mathbf{C} : |z| < 1\}$, i.e.,

$$\{z_k\}_{k=2}^{\tau} \subset \mathbb{D}. \tag{3.2.4}$$

In this case, the series $R(t_k)$ features a moderate growth rate similar to the one for the returns in the Black-Scholes model. It can be noted that if $\lambda > 0$, then equation (3.2.4) holds for all $\tau \geq 2$. If $\lambda < 0$, then it may happen that (3.2.4) does not hold for some τ. However, it appears that (3.2.4) holds for small enough $|\lambda|$ and small enough τ. In particular, we found that, for $\tau \leq 11$, (3.2.4) holds for all $\kappa = \lambda\delta \geq -0.111$. For $\tau \leq 15$, (3.2.4) holds for all $\kappa = \lambda\delta \geq -0.075$. For $\tau \leq 20$, (3.2.4) holds for all $\kappa = \lambda\delta \geq -0.055$. For $\tau \leq 150$, (3.2.4) holds for all $\kappa = \lambda\delta \geq -0.006$. It appears that this range for the parameters allows to replicate the volatilities depending on the sampling frequencies similar to the ones observed for the historical data; in other words, it is sufficient for our purposes.

We discuss the choice of τ, σ and λ in the next section.

Volatility estimator

We will be using the classical estimator for volatility based on samples collected within $[t - \Delta t, t]$ interval, where $\Delta t > 0$ and is given.

Recall from Chapter 2, the volatility process is characterised by the integral

$$v(t) = \frac{1}{\Delta t} \int_{t-\Delta t}^{t} \sigma(s)^2 \mathrm{d}s.$$

From this section onward, we assume that the time points θ_k are equally spaced with sampling interval $\hat{\delta} = \theta_k - \theta_{k-1}$ and that $\theta_{m_0} = t - \Delta t$, $t_m = t$, and $\Delta t = (m - m_0)\hat{\delta}$.

Note that the choice of $\widehat{\delta}$ could be different from δ in (3.2.3). In our experiments, we consider only the cases where $\hat{\delta} = N\delta$ for some integer $N \geq 1$ such that $\{\theta_k\} \subset \{t_k\}$; if $N = 1$, then $\delta = \widehat{\delta}$ and $t_k = \theta_k$.

For a given choice of $\widehat{\delta}$, the estimate of $v(t)$ is calculated by:

$$\hat{\sigma}(t) = \sqrt{v(t)} = \left[\frac{1}{\Delta t} \sum_{k=m_0+1}^{m} (\overline{R}_m - R(\theta_k))^2 \right]^{1/2}, \tag{3.2.5}$$

where

$$\Delta t = (m - m_0)\hat{\delta}, \quad R(\theta_k) = \log S(\theta_k) - \log S(\theta_{k-1}), \quad \overline{R}_m = \frac{1}{m - m_0} \sum_{k=m_0+1}^{m} R(\theta_k).$$

The properties of this estimator are discussed in Chapter 2 and more in [113]. For the choices of the frequency data, it is well discussed in [1].

Note that the particular choice of this estimator is not crucial for our purposes; other estimators can be used instead. For instance, the estimator proposed by Andersen [10] gives very similar estimates on different frequencies. To illustrate this, we compared the annualized daily volatility estimated by the classical historical volatility and and the realised volatility for the SP500 index for 2008-2013. We estimated the mean absolute difference (MAD) between the two estimators as:

$$MAD = \frac{\Sigma |RV^{A}_{30sec} - RV^{C}_{30sec}|}{n},$$

where n is the number of observations in our sample. Here we have n = 1511 days. We found that $MAD_{30sec} = 0.000164$, $MAD_{5min} = 0.001647$ and $MAD_{15min} = 0.004826$.

Figure B.12 plot the time series for those volatility estimators for the selected dataset.

Monte-Carlo simulation of the process (3.2.3) and its volatility

We first simulate the process generated by (3.2.3) at some high frequency (small δ). We then compute the volatility at lower frequencies from the same simulated path. For instance, assuming that there are 6.5 trading-hours per day and 252 trading-days per year, we simulate a one-year sample path at 15-second frequency, i.e., $\delta = \frac{1}{252\times6.5\times60\times4} \approx 2.5437 \times 10^{-6}$. To measure the volatility at lower frequencies, we sub-sample the simulated path at 5-minute and 1-hour frequencies by using the tick-aggregation technique [122]. This is done by taking the last price realized before each new grid point.

In our Monte-Carlo simulation, we use the following selection criteria for the model's parameters:

1. Selection of δ: we used δ for 15-second data through out this experiment.
2. Selection of (τ, λ): we used $\tau = 5, 20,$ and 120. This corresponds to ϱ = 75sec, 5min, and 30min, respectively. In addition, we selected a variety of $\lambda \in [-2000, 20000]$ such that equation (3.2.4) holds.
3. Selection of σ: we used $\sigma = 0.3$.

We generated 100,000 instances for each combination of σ, τ and λ. To analyze the results statistically, we summarize the average and the standard deviation of the estimated volatility from these instances at each selected sampling frequency.

The results of the simulation experiments

For the prices simulated from the above settings, the measured volatility under different time-scale are summarized in Tables A.1 and A.2. Let us discuss the presented results.

Due the randomness of the data, short samples produces random estimates of the volatilities featuring significant variance. To decrease this variance for the demonstration purposes, we selected longer time series using the observations within 1-year window from the simulated data. In our experiments, we used samples with 393,120 observations, with 19,656 observations, and 1,638 observations for 15-sec, 5-min, and 1-hour data, respectively. Tables A.1 and A.2 show the impact of the standard deviation on the estimated volatility. Shorter time series would produce the same estimates for the volatilities but with higher variance (i.e. Table A.3).

For the case where $\lambda < 0$, we observe that the estimated volatility increases as the sampling frequency decreases. Given a fixed τ, the larger λ reduces the difference in the estimated volatility per sampling frequency. On the other hand, given a fixed λ, bigger τ results larger gaps in the estimated volatility at each sampling frequency. For example, with $\tau = 120$, we have that

$$\left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\kappa=-0.005} = 1.696 > \left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\kappa=-0.0025} = 1.255,$$

whereas with $\kappa = -0.005$,

$$\left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\tau=20} = 1.047 < \left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\tau=120} = 1.696.$$

For $\lambda > 0$, we have opposite results, i.e., the estimated volatility decreases as the sampling frequency decreases. Given a fixed τ (or a fixed λ), the larger λ (or τ) results in larger gaps in the estimated volatility as the sampling frequency decreases. For instance, $\tau = 120$,

$$\left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\kappa=0.0005} = 0.962 > \left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\kappa=0.005} = 0.729,$$

while with $\kappa = 0.005$,

$$\left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\tau=20} = 0.959 > \left.\frac{\sigma_{hour}}{\sigma_{5min}}\right|_{\tau=120} = 0.729.$$

These experiments demonstrate that, with the proposed model, we can replicate the volatility and time-scale dependence characteristic of financial time-series. The parameters within the model can be used to control the changes of the estimated volatility at different sampling frequency.

It is noted that we observed the estimated volatility $\hat{\sigma}_{15sec}$ is fairly close to the "true" volatility $\sigma = 0.3$ for all choices of (λ, τ).

As was mentioned above, the results of experiments are robust with respect to the choice of the volatility estimator. It appears that the results for numerical simulation are also robust with respect to variations of all other parameters. The average values for estimated volatilities are not changing significantly when we increased the number of Monte-Carlo trials, and they are changing very smoothly and systematically for different choices of $(\kappa, \tau, \lambda, \sigma, \delta)$ given that equation (3.2.4) on (τ, λ, δ) is satisfied. Condition equation (3.2.4) on (τ, λ, δ) is essential to ensure that the simulated process has a moderate growth.

3.3 Matching the time scale dependency of volatilities with real data

In this section, we discuss how to to obtain the parameters λ and τ in the proposed model for matching the volatility behavior in a given set of historical data.

3.3.1 Analysis of real data

In practice, the highest frequency financial time series available for analysis is the tick-data. Tick-data is recorded in discrete time that is not necessary equally spaced. Some data services however provide financial data which are sampled on given equispaced frequency (i.e. every 15-second or 5-minute). These samples are sub-samples of the tick-data and are aggregated using different weighting schemes.

For our empirical study, we estimate the volatility of financial time-series from different markets. As was discussed in Section 3.2, our statistical volatility estimator requires the data to be collected at equispace. Hence, we use the tick-data as the baseline data for each underlying assets. We then perform the data cleaning process to obtain samples at equal intervals. We used the datasets obtained from SIRCA - the Securities Industry Research Centre of Asia-Pacific [112] for the period 2008–2010.

We extended the scope of our inference analysis by selecting the top most traded indices and US stocks listed on Reuters Finance, including:

Category 1: Stock indexes: DAX (Deutsche Boerse AG German Stock Index, trading between 09:00 am and 17:45 pm CET), FTSE 100 (a share index of the 100 companies listed on the London Stock Exchange, trading between 08:00 am - 04:30 pm GMT); IBEX 35 (an index of the Spanish Continuous Market, opening between 09:00am - 05:30pm); SMI (Switzerland's blue-chip stock market index, from 09:00 to 5:30pm CET); S&P 500 (a stock market index based on the market capitalizations of the 500 largest companies having common stock listed on the NYSE, 09:30am till 04:00pm); and S&P 200 (a stock market index based on the market capitalizations of 200 large companies having common stock listed on the Australian Stock Exchange, operating from 10:00am to 04:00pm); and TSX 60 (stock market index of 60 large companies listed on the Toronto Stock Exchange).

Category 2: Individual company stock symbols include: AAPL, IBM, JPM, GE, GOOG, MSFT and XOM. These stocks are traded between 9:30am and 04:00pm on NYSE.

Methodology

We take the following steps to analyse our datasets.

Step 1. Obtain the tick-data for each stock/index from SIRCA,

Step 2. Perform data cleaning. By using the tick-data, we force these asynchronously and irregularly recorded series to a synchronized and equispaced time grid using the previous tick aggregation to obtain samples at different frequency

Step 3. Estimate the volatility using the formula discussed in Section 4.5.1 for the entire year to obtain the annualized volatility.

Step 4. Repeat Step 2-3 for each sampling frequency and each stock/index.

Table A.4 and A.5 show how the volatility varies when it is measured at different sampling frequency for the selected stocks and indexes. In this table, the volatility was calculated by applying the estimation from equation (3.2.5) for the observations collected during a whole year.

3.3.2 Matching the model's parameters with real prices

In this section, we demonstrate that it is possible to calibrate the proposed model with the historical data such that the volatilities match for three different sampling frequencies.

In our experiments, we considered data available at three sampling frequencies: 15-second, 5-minute and 1-hour.

The volatility of historical 15-sec data was accepted as σ in equation (3.2.3); this is the data sampled at the highest available frequency.

In the proposed model has parameters λ, τ, and σ. Our purpose is to select (σ, λ, τ) to ensure matching of the volatilities for the simulated process and for a set of historical data for 15-sec, 5-minute and 1-hour sampling. For this, we used the following simple and straightforward heuristic algorithm.

Step 1. Estimate the volatility at the selected sampling frequencies σ_{15sec}, σ_{5min} and σ_{1hour} using the historical data.

Step 2. Select a finite set of $\left(\lambda_i, \tau_j, \sigma_k\right)$ for the search space.

Step 3. For each triplet $\left(\lambda_i, \tau_j, \sigma_k\right)$, generate a path using equation (3.2.3) and estimate the volatilities $\hat{\sigma}_{15sec}(\lambda_i, \tau_j)$, $\hat{\sigma}_{5min}(\lambda_i, \tau_j)$ and $\hat{\sigma}_{1hour}(\lambda_i, \tau_j)$, calculated for the corresponding sampling frequencies.

Step 4. Among these outputs, find the values of $\left(\lambda_i, \tau_j, \sigma_k\right)$ that generate $\hat{\sigma}_{15sec}$, $\hat{\sigma}_{5min}$ and $\hat{\sigma}_{1hour}$ matching with the volatilities estimated from the historical data.

Step 5. If there are no matching values, extend the set $\left\{\left(\lambda_i, \tau_j, \sigma_k\right)\right\}$ and repeat steps 2-4.

It appears that selection $\sigma = \sigma_{15sec}$ allows to find satisfactory (λ, τ) for all our experiments. We suggest to this initial selection to reduce the search.

It can be noted that we do not trying to find the best matching (λ, τ, σ) via minimization of the fitting errors (residuals) for the paths of historical series as is usually done by the Least Square estimators. We match only the volatilities of the historical series and simulated series on the given set of sampling frequencies. Therefore, our simulation does not replicate other characteristics of the price evolution such as the rate of growth.

3.3.3 Numerical examples

We present some examples for both cases $\lambda > 0$ and $\lambda < 0$. The below table is extracted from Table A.4 where we only selected observations for S&P 500 Index and Google Stock in 2008. We use $\lambda < 0$ to reproduce the dependence on the sampling frequency for S&P 500 Index and we use $\lambda > 0$ to replicate that characteristic for Google stock prices.

Underlying assets	σ_{15sec}	σ_{5min}	σ_{1hour}
S&P 500 Index	0.2881	0.3654	0.3747
Google Stock	0.8114	0.6276	0.5937

In this samples, we considered the volatility calculated for an entire year time window. Shorter time periods prices their own volatility values that could also be implemented.

For $\lambda < 0$, the volatility of S&P 500 measured at different sampling frequency was $\sigma_{15sec} = 0.2881$, $\sigma_{5min} = 0.3654$ and $\sigma_{hour} = 0.3747$. Using the steps discussed above, we obtained the following parameters:

$$\begin{cases} \sigma = 0.2881 \\ \tau = 9 \\ \delta = \frac{1}{252\times 6.5\times 60\times 4} \\ \kappa = \lambda\delta = -0.0325, \lambda = -12776.4 \end{cases}$$

We substituted these parameters into equation 3.2.3, simulated 100,000 instances, and obtained the following average volatilities at each sampling frequency:

	S&P 500	$mean_{sim}$	sd_{sim}
σ_{15sec}	0.2881	0.2882	0.0003
σ_{5min}	0.3654	0.3615	0.0018
σ_{hour}	0.3747	0.3751	0.0065

Similarly, for $\lambda > 0$, the volatility of Google stock on NYSE in 2008 measured on alternative time-scales was $\sigma_{15sec} = 0.8114$, $\sigma_{5min} = 0.6276$ and $\sigma_{hour} = 0.5937$. We obtained parameters

$$\begin{cases} \sigma = 0.8114 \\ \tau = 10 \\ \delta = \frac{1}{252\times 6.5\times 60\times 4} \\ \kappa = \lambda\delta = 0.0445, \lambda = 17493.84 \end{cases}$$

which provides

	GOOG	$mean_{sim}$	sd_{sim}
σ_{15sec}	0.8144	0.8147	0.0010
σ_{5min}	0.6276	0.6302	0.0017
σ_{hour}	0.5937	0.5892	0.0101

To illustrate the search process of the matching values, we show in Tables A.6 and A.8

the errors corresponding to (λ_i, τ_j) used in calibrating S&P500 Index and Google stock respectively and calculated as

$$\epsilon_{(\lambda_i,\tau_j)} = \sqrt{(\sigma_{15sec} - \hat{\sigma}_{15sec})^2 + (\sigma_{5min} - \hat{\sigma}_{5min})^2 + (\sigma_{1hour} - \hat{\sigma}_{1hour})^2}. \tag{3.3.1}$$

For both tables, we selected $\sigma = \sigma_{15sec}$.

These tables show that selection of $\sigma = \sigma_{15sec}$ allows to find the according values of λ and τ to match the simulated volatility with sufficient accuracy on the given set of given sampling frequencies.

General properties of the simulated processes

As was mentioned above, we have to select the parameters (λ, τ) only among the pairs such that equation (3.2.4) holds, i.e., the characteristic polynomial for autoregression equation (3.2.3) does not have roots outside the unit circle. It appears that, under this restriction, the behavior of the simulated process with delay does not demonstrate any unusual and undesirable features such as excessive growth, and is quite similar to the behavior of the underlying processes, as well as to the behavior of standard Ito processes and autoregressions used to model the financial time series. Sample paths of the simulated price process and the underlying process are shown in Figure B.1. The distributions the returns of both processes are displayed in Figure B.2.

Time varying volatility and shorter time windows

To analyze the time varying volatility for historical prices, one could match the volatility for shorter time windows. In this case, we have to select (λ, τ, σ) separately for the corresponding time windows. For example, assume that we wish to match the quarterly data for Google stock prices during one year, and we calculate $\sigma^{(k)}_{15se}$, $\sigma^{(k)}_{5min}$, and $\sigma^{(k)}_{1hour}$ for each quarter, where $k \in \{1, 2, 3, 4\}$ represents a quarter. In this case, we have to select for each quarter $(\lambda^{(k)}, \tau^{(k)}, \sigma^{(k)})$ to match the corresponding volatilities. To replicate quarterly time depending $(\sigma^{(k)}_{15se}, \sigma^{(k)}_{5min}, \sigma^{(k)}_{1hour})$, it suffices to accept equation (3.2.3) with time dependent set of parameters $(\lambda, \tau, \sigma) = (\lambda^{(k)}, \tau^{(k)}, \sigma^{(k)})$ that depends on the particular quarter. The same approach can be used for other time windows, for instance, for weekly or daily volatilities. Again, we have to select the parameters (λ, τ, σ) for each time window separately, and accept equation (3.2.3) with piecewise constant parameters.

3.4 Discussion

The approach suggested in this chapter allows many modifications. We outline below some possible straightforward modifications as well as more challenging problems and possible applications that we leave for the future research.

1. Our purpose was to demonstrate a method of constructing a process $S(t)$ with preselected volatility measured at different sampling frequencies, for instance, to match the volatilities on these frequencies for a set of historical prices. In our experiments, the volatility depended monotonically on each parameter (λ, τ). This makes heuristic search relatively easy. More precise match can be achieved via smaller variations of the values of (λ, τ) around the values that we have demonstrated so far. In addition, these monotonic dependence can be simplified by using other search algorithms such as simulated annealing and genetic algorithm.

2. The models with $\lambda < 0$ requires some additional constraints on $|\lambda|$ and τ, to ensure that the characteristic polynomials for the autoregressions do not have roots outside of the unit circle, to avoid exponential growth of the solutions. These constraints are absent for the models with $\lambda > 0$.

3. It appears that the equations with one delay term can replicate volatilities for three sampling frequencies. To cover a setting with more different sampling frequencies, more delay terms may be necessary. We think that the inclusion of delay terms in the form of $\sum_{d=k-M}^{k} u_d R(t_d)$ in equation (3.2.3) can allow to model any volatility $\widehat{\sigma}(\hat{\delta})$ depending on the size of the sampling interval $\hat{\delta} = \delta, 2\delta, 3\delta, ...$, with an appropriate choice of M and $u \in \mathbf{R}^M$.

4. For the continuous time setting, we have a conjecture that the inclusion of a delay term $\int_{t-K}^{t} u(s)R(s)ds$ in equation (3.2.1) can allow to model any volatility $\widehat{\sigma}(\delta)$ continuously depending on the size of the sampling interval δ, with an appropriate choice of the function $u(s)$ and $K > 0$.

5. We did not consider pricing of options on the underlying time series and relations between the historical volatility and the implied volatility. Respectively, we did not analyse existence or uniqueness of a martingale measure that is used for pricing. Furthermore, we did not attempt neither to match the growth of the historical prices nor to ensure that the process $S(t) = e^{R(t)}$ is a martingale. We leave it for the future research. It can be noted that it is possible to match a given growth rate via

selection of $S(t) = e^{\beta t}e^{R(t)}$, with β as a new parameter to be selected. As we were trying to construct a model with the minimal number of the parameters, we excluded modelling of growth via selection of β and leave it for future research.

Chapter 4

Implied volatility via a dynamic purified option price process

4.1 Introduction

In the Chapter 2, Section 2.4, we discussed about the implied volatility. In this chapter, we introduce the construction of an implied volatility process and discuss the use of this new process in predicting the future volatility.

Often, the implied volatilities are defined as the inverse of the Black-Scholes [28] pricing formula applied to the observed market prices; given fixed and known asset price, strike price, future interest rate, and time-to-maturity, the implied volatility is uniquely defined by the option price. In fact, the option prices and the implied volatilities are fluctuating along with the underlying assets prices; they have a stochastic “random walk" type pattern of movement, similar to the stock prices. In addition, the implied volatility depends on the strike price and the expiration time. Therefore, one would find it difficult to decide which implied volatility to use among all possible versions of the implied volatility when describes the market expectations on the degrees of the future stock price deviations. To address this issue, volatility surface (i.e. Figure B.6) is used for analysing the implied volatility against strike price and time-to-maturity. Alternatively, one can consider the implied volatility surface to follow a stochastic process, i.e the implied volatility indexes. The volatility index VXO on the Chicago Board of Options Exchange (CBOE) [121] and the AVX on Australian Securities Exchange (ASX) [57] used the Black-Scholes-Merton’s framework as the underlying model to construct the implied volatility for S&P 100 and S&P 200 respectively. The implied volatility of these indexes used options at different strike prices and maturity dates to approximate the at-the-money implied volatility. However, this approach has some limitations. It is restricted by the assumptions made in the Black-Scholes' model and has an artificially induced upward bias [38]. With those limitations, Carr and Wu [37] introduced an alternative method for constructing the implied volatility using a model-free approach, which was then used by CBOE for constructing the volatility index VIX for S&P 500. The VIX index is constructed from the price of a portfolio including a number of out-the-money option prices that varies every day, depending on the number of options with non-zero bid price. While this implied volatility has a better known economic interpretation [37], the construction of the VIX process is complex and requires large samples of option data.

Motivated by these volatility processes, we suggest a modification in the approaches used for VXO, AVX, and VIX indexes with an aim to reduce the measurement errors and improve the computational robustness. We suggest to consider a "dynamically purified" option price process such that impact of stock price movements is reduced. This helps to separate the impact of the stock price movements from the changes in the market forecast of the future volatility. In effect, the implied volatility calculated form this process could be more informative than traditionally calculated implied volatility, similarly to the popular volatility indexes such as VXO and VIX indexes.

In theory, the dynamically 'purified' option price process eliminates the impact of the stock price movements. However, this would be possible if the option prices were available for continuous sets of strike prices and expiration times. In practice, we have to use only finite sets of available prices. In order to restore this process from incomplete sets of available option prices, we suggest to use a similar approach to the approach implemented in the calculation of the volatility index. Here, the implied volatilities from the missing option prices was replaced by linear combinations of implied volatilities using some observable options. However, instead of applying the linear interpolation on the implied volatility, we interpolate the missing options prices. In this paper, we discuss the use of both the first order Taylor series interpolation and quadratic interpolation. For this approach, the 'dynamically purified" option price process can be constructed using 18 observed option prices.

We study the statistical properties of the proposed process by using the S&P/ASX 200 Index Options data for the period from 1^{st} January 2010 to 31^{st} December 2012. For demonstration purposes, we consider only a special case of this process which represents the at-the-money implied volatility. It was found before that the classical volatility index VXO and VIX feature strongly negative correlations with the index return increments

[57, 121]. Our finding shows that the "dynamically purified" option price process has the same feature. In addition, the implied volatility of the purified option price process has a very strong positive correlation with the implied volatility index. This is an interesting result given that VIX process is calculated using very different data and methods.

As a possible application of the proposed process, we consider a simple model for forecasting the future volatility. We establish a number of regression models using the implied volatility index VIX and the implied volatility from the purified option prices as predictors of the future volatility. We find that for our selected dataset and constructed models, the forecasting ability of the new implied volatility is superior to that of the implied volatility index. Since calculation of the proposed process requires less option prices than calculation of the existing implied volatility index VIX, this process can be used as an alternative for VIX in some cases when there is not sufficient data to calculate VIX.

4.2 The model of "purified" option price process

Let us consider the diffusion model of a securities market consisting of a risk free bond or bank account with the price $B(t)$, $t \geq 0$, and a risky stock with the price $S(t)$, $t \geq 0$. The prices of the stocks evolve as

$$dS(t) = S(t)\left(a(t)dt + \sigma(t)dw(t)\right), \qquad t > 0, \tag{4.2.1}$$

where $w(t)$ is a Wiener process, $a(t)$ is an appreciation rate, $\sigma(t)$ is a random volatility coefficient. The initial price $S(0) > 0$ is a given deterministic constant. The price of the bond evolves as

$$B(t) = \exp\left(\int_0^t r(s)ds\right)B(0), \tag{4.2.2}$$

where $r(t) \geq 0$ is a random process and $B(0)$ is given.

We assume that $w(\cdot)$ is a standard Wiener process on a given standard probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is a set of elementary events, $\mathcal{F}$ is a complete σ-algebra of events, and $\mathbf{P}$ is a probability measure.

Let $\mathcal{F}_t$ be a filtration generated by the currently observable data. We assume that the process $(S(t), \sigma(t))$ is $\mathcal{F}_t$-adapted and that $\mathcal{F}_t$ is independent of $\{w(t_2) - w(t_1)\}_{t_2 \geq t_2 \geq t}$. In particular, this means that the process $(S(t), \sigma(t))$ is currently observable and $\sigma(t)$ does not depend on $\{w(t_2) - w(t_1)\}_{t_2 \geq t_2 \geq t}$. We assume that $\mathcal{F}_0$ is the P-augmentation of the set $\{\emptyset, \Omega\}$, and that $a(t)$ does not depend on $\{w(t_2) - w(t_1)\}_{t_2 \geq t_2 \geq t}$. For simplicity, we assume that $a(t)$

is a bounded process.

Option pricing terminology

Below is a list of terms that are commonly used in option pricing

- Long position, a portfolio is ‘long asset X’ if it has net positive holdings of contracts in asset X.
- Short position, a portfolio is ‘short asset X’ if it has net negative holdings of contracts in asset X (i.e. has short sales of contracts).
- Hedge, or ‘hedging portfolio’ is a portfolio that minimises the losses it might obtain, i.e. reducing the risk of adverse from price movements.
- In-the-money (ITM), a derivative contract that would have positive payout if settlement based on today’s market prices (e.g. a call option with very low strike).
- Out-of-the-money (OTM), a derivative contract that would be worthless if settlement based on today’s market prices (e.g. a call option with very high strike).
- At-the-money (ATM), a derivative contract exactly at it’s breaking point between ITM and OTM.
- Underlying, the stock, bond, ETF, exchange rate, etc. on which a derivative contract is written.
- Strike price, the price upon which a call or put option is settled.
- Maturity, the latest time at which a derivative contract can be settled.
- Exercise, the event that the long party decides to use a derivative’s embedded option (e.g. using a call option to buy a share of stock at lower than market value).

The Black-Scholes price

Let K be non-negative, i.e. $K > 0$. We shall consider two types of options: vanilla call and vanilla put, with payoff function $f(S(T)) = F(S(T), K)$, where $F(S(T), K) = (S(T) - K)^+$ or $F(S(T), K) = (K - S(T))^+$, respectively, with K be the strike price.

Let $T > 0$ be fixed. Let $H_{BS,c}(t, x, \sigma, r, K)$ and $H_{BS,p}(t, x, \sigma, r, K)$ denote Black-Scholes prices for the vanilla put and call options with the payoff functions $F(S(T), K)$ described above under the assumption that $S(t) = x$, $(\sigma(s), r(s)) = (\sigma, r)$ $(\forall s > t)$, where $\sigma \in (0, +\infty)$ is non-random. The Black-Scholes formula for a call option can be rewritten as

$$H_{BS,c}(t, x, \sigma, r, K) = x\Phi(d_+(t, x, \sigma, r, K)) - Ke^{-r(T-t)}\Phi(d_-(t, x, \sigma, r, K)), \qquad (4.2.3)$$
$$H_{BS,p}(t, x, \sigma, r, K) = H_{BS,c}(t, x, \sigma, r, K) - x + Ke^{-r(T-t)},$$

where

$$\Phi(x) \overset{\Delta}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}} ds,$$

and

$$\begin{aligned} d_+(x, t, \sigma, r, K) &\overset{\Delta}{=} \frac{\log(x/K) + (T-t)r}{\sigma\sqrt{(T-t)}} + \frac{\sigma\sqrt{(T-t)}}{2}, \\ d_-(x, t, \sigma, r, K) &\overset{\Delta}{=} d_+(x, t, \sigma, r, K) - \sigma\sqrt{(T-t)}. \end{aligned} \qquad (4.2.4)$$

Set

$$\tilde{S}(t) \overset{\Delta}{=} S(t)\exp\Big(-\int_0^t r(s)ds\Big).$$

The risk neutral pricing

We assume that there exist a risk-neutral measure Q such that the process $\tilde{S}(t)$ is a martingale under Q, i.e., $\mathbf{E}_Q\{\tilde{S}(T)\,|\mathcal{F}_t\} = \tilde{S}(t)$, where $\mathbf{E}_Q$ is the corresponding expectation.

The local risk minimization method, the mean variance hedging, and some other methods based on the risk-neutral valuation lead to the following pricing rule: given (a, σ, r), the option price is

$$P_{RN}(t, \sigma(\cdot), r(\cdot)) \overset{\Delta}{=} \mathbf{E}_Q\{e^{-\int_t^T r(s)ds} F(S(T))\,|\,\mathcal{F}_t\}, \qquad (4.2.5)$$

where Q is some risk neutral measure, and where $\mathbf{E}_Q$ is the corresponding expectation. Usually, Q is uniquely defined by (a, σ, r), and by the pricing method used.

For numerical simulation purposes, we assume that we have chosen one of these methods (for instance, local risk minimization method or mean variance hedging). Therefore, the risk neutral measure Q is uniquely defined by (a, σ, r) given the method of pricing.

For brevity, we shall denote by H_{BS} the corresponding Black-Scholes prices for different options, i.e., $H_{BS} = H_{BS,c}$ or $H_{BS} = H_{BS,p}$, for vanilla call, vanilla put respectively. Let

$$v(t) \overset{\Delta}{=} \frac{1}{T-t}\int_t^T \sigma(s)^2 ds, \qquad \rho(t) \overset{\Delta}{=} \frac{1}{T-t}\int_t^T r(s)ds.$$

Let us consider dynamically adjusted parameters $T = T(t) = t + \tau$ and $K = K(t) = \kappa S(t)$, where $\kappa \in (0, +\infty)$ and $\tau > 0$ are some parameters, t is the current time. In this case, $F(S(T)) = F(S(T), K) = S(t)F(Y(t + \tau), \kappa)$, where

$$Y(T) = S(t + \tau)/S(t).$$

By rule (4.2.5), the option price given (a, σ, r), is

$$\begin{aligned} P_{RN}(t, \sigma(\cdot), r(\cdot)) &\overset{\Delta}{=} \mathbf{E}_Q\{e^{-\int_t^T r(s)ds} F(S(T), K) \,|\, \mathcal{F}_t\} \\ &= S(t)\mathbf{E}_Q\{e^{-\int_t^{t+\tau} r(s)ds} F(Y(t+\tau), \kappa) \,|\, \mathcal{F}_t\}, \end{aligned}$$

where Q is some risk neutral measure, and $\mathbf{E}_Q$ is the corresponding expectation.

Let

$$G(t) \overset{\Delta}{=} \frac{P_{RN}(t, \sigma(\cdot), r(\cdot))}{S(t)}.$$

By the definitions,

$$G(t) = \mathbf{E}_Q\{e^{-\int_t^{t+\tau} r(s)ds} F(Y(t+\tau), \kappa) \,|\, \mathcal{F}_t\}. \tag{4.2.6}$$

Suppose that $v(t)$ and $\rho(t)$ are $\mathcal{F}_t$-measurable. In this case,

$$\begin{aligned} H_{BS,c}(t, 1, \sqrt{\mathbf{v}(t)}, \rho(t), \kappa_c) &= G_c(t), \\ H_{BS,p}(t, 1, \sqrt{\mathbf{v}(t)}, \rho(t), \kappa_p) &= G_p(t) \end{aligned} \tag{4.2.7}$$

for call and put options respectively, where κ_c and κ_p are defined similarly to κ. Therefore, for a general case, we can accept that the implied volatility $\sigma_{imp}(t)$ and the implied average forward risk-free rate $\rho_{imp}(t)$ at time t can be inferred from the system

$$\begin{cases} H_{BS,c}(t, 1, \sigma_{imp}(t), \rho_{imp}(t), \kappa_c) = G_C(t), \\ H_{BS,p}(t, 1, \sigma_{imp}(t), \rho_{imp}(t), \kappa_p) = G_P(t). \end{cases} \tag{4.2.8}$$

The following lemma from [45] is a generalization for random $r(\cdot)$ of the lemma from Hull and White [75]:

Lemma 4.2.1 *Let $t \in [0, T)$ be fixed. Let $v(t)$ and $\rho(t)$ be $\mathcal{F}_t$-measurable. Then*

$$\mathbf{E}_Q\{e^{-\int_t^T r(s)ds} F(S(T))|\mathcal{F}_t\} = H_{BS}(t, S(t), \sqrt{v(t)}, \rho(t), K).$$

Clearly, $\frac{1}{T-t}\int_t^T \sigma(s)^2 ds$ and $\frac{1}{T-t}\int_t^T r(s)ds$ are not $\mathcal{F}_t$-measurable in the general case of stochastic (r, σ), and the assumptions of Lemma 4.2.1 are not satisfied.

Corollary 4.2.1 *Assume that $H_{BS} = H_{BS,c}$ and $H_{BS} = H_{BS,p}$. Consider a market model with pricing rule (4.2.5). Let (σ, r) does not depend on w under Q. Then $P_{RN}(t) = \mathbf{E}_Q\{H_{BS}(t, S(t), \sqrt{v(t)}, \rho(t), K) \,|\, \mathcal{F}_t\}$, where (v, ρ) are defined in Lemma 4.2.1.*

By rule (4.2.5), the option price given (a, σ, r), is

$$\begin{aligned} P_{RN}(t, \sigma(\cdot), r(\cdot)) &\overset{\Delta}{=} \mathbf{E}_Q\{e^{-\int_t^T r(s)ds} F(S(T), K) \,|\, \mathcal{F}_t\} \\ &= S(t)\mathbf{E}_Q\{e^{-\int_t^{t+\tau} r(s)ds} F(Y(t+\tau), \kappa) \,|\, \mathcal{F}_t\}, \end{aligned}$$

where Q is some risk neutral measure, and where $\mathbf{E}_Q$ is the corresponding expectation.

Let

$$G(t) \overset{\Delta}{=} \frac{P_{RN}(t, \sigma(\cdot), r(\cdot))}{S(t)}. \tag{4.2.9}$$

It follows that

$$G(t) = \mathbf{E}_Q\{e^{-\int_t^{t+\tau} r(s)ds} F(Y(t+\tau), \kappa) \,|\, \mathcal{F}_t\}. \tag{4.2.10}$$

Assume that $v(t)$ and $\rho(t)$ are $\mathcal{F}_t$-measurable in this case,

$$\begin{aligned} H_{BS,c}(t, 1, \sqrt{\mathbf{v}(t)}, \rho(t), \kappa_i) &= G_C(t), \\ H_{BS,p}(t, 1, \sqrt{\mathbf{v}(t)}, \rho(t), \kappa_i) &= G_P(t), \end{aligned} \tag{4.2.11}$$

for call and put options respectively.

The observations of option prices with dynamic adjusted strike price $K = \kappa S(t)$ with a fixed κ and a fixed period t can be useful for econometrics purposes even without calculation of the implied parameters. In particular, some features of the evolution law for implied parameters $(\sigma(t), r(t))$ can be restored directly from the observations of the processes $G(t)$. For instance, if $\rho(t)$ is a non-random process then the implied volatility $\sqrt{v(t)}$ can be calculated from equation (4.2.11) for call and put options. In addition, as the impact of the stock price movements is damped, one may expect that $G(t)$ is a relatively smooth process. Thus, the study of the process $G(t)$ will be of interest.

Up to the end of this chapter, we will assume that $\rho(t)$ is non-random and known. This is an usual assumption since the risk-free rate is relatively stable.

4.3 Approximation of incomplete option data

In practice, option prices are available only for finite sets of possible option prices different strikes and time-to-maturity. Therefore, it is not possible to collect the prices $P_i(t)$ of the options at the exact strike prices $K_i = \kappa_i S(t)$ with fixed κ_i and t. In order to study the process $G(t)$ described above, we have to use the prices $\tilde{P}_i(t)$ of the corresponding options with the closest available strike prices $\tilde{K}_i(\tilde{t})$.

From this section onward, let $\tilde{P}_C$ and $\tilde{P}_P$ be the values of call and put options observed on the market.

4.3.1 Parametric approximation for absent option prices

Delta of the strike

The price change of the option price P with respect to K, when other factors remaining constant, is called the delta of the strike

$$\triangle K = \frac{\delta P}{\delta K}.$$

From equation (4.2.3) and equation (4.2.4), the delta of the strike for a call and a put are:

$$\frac{\delta P_C}{\delta K} = e^{-\tau\triangle t}\Phi(d_+),$$

and

$$\frac{\delta P_P}{\delta K} = e^{-\tau\triangle t}\Phi(-d_-),$$

receptively.

Theta of the option (time-decay)

Theta of an option is defined as the rate of change of its price P with respect to time t, while all other factors remaining constant:

$$\Theta = \frac{\delta P}{\delta t}.$$

From equation (4.2.3) and equation (4.2.4), it can be shown that

$$\theta_C = \frac{\delta P_C}{\delta t} = -\frac{S\sigma}{2\sqrt{\triangle t}}\Phi'(d_+) - \tau K e^{-\tau \triangle t}\Phi(d_-) < 0,$$

and

$$\theta_P = \frac{\delta P_P}{\delta t} = -\frac{S\sigma}{2\sqrt{\triangle t}}\Phi'(d_+) + \tau K e^{-\tau \triangle t}\Phi(-d_-) < 0.$$

Let's assume that one wishes to approximate the missing call option price P^*_C at strike price K^* with time-to-maturity $\triangle t^*$. The nearest available strike price is $\tilde{K}$ with $\triangle \tilde{t}$ has a value of $\tilde{P}_C$. The first order approximation can be used such that:

$$P^*_C \approx \tilde{P}_C - m_1(\tilde{K} - K),$$

where $m_1 = \dfrac{\delta P_C}{\delta K}$, the delta of the strike for a call option. Thus, P^*_C can be approximated as followed:

$$P^*_C(\tilde{t}) \approx \tilde{P}_C - S e^{-\tau \triangle \tilde{t}}\Phi(d_2)(S - \tilde{K}). \tag{4.3.1}$$

We then need to adjust this approximated value $P^*_C(\tilde{t})$ in order to match $\triangle \tilde{t}$ with $\triangle t^*$. The equation (4.3.1) can be extended further by using the first-order-approximation on $\triangle t$ where $m_2 = \theta = -\dfrac{\delta V}{\delta \triangle t}$, the time-decay of the option.

However, this approach can only help calculating $G(t)$ for the case of small value $|K - \tilde{K}|$. In addition, one would have to obtain the implied parameters of the underlying option pricing model before estimating the missing option data with this method.

4.3.2 Quadratic approximation of absent option prices

For the construction of the implied volatility suggested in [121], the linear approximation of the implied volatility was used for determining the implied volatility of at-the-money options. We instead suggest using quadratic approximation (spline interpolation of degree two) for estimating the missing option price data before computing the implied volatility.

As fitting the option price surface often leads to numerical difficulties [79], only some available option prices near the targeted κ_i will be used. Now let:

1. K^* be the strike price of the missing option;
2. K_{j-1} be the strike price that is the second closest to and below; K^*
3. K_j be the strike price that is the closest to and below K^*;
4. K_{j+1} be the strike price that is the closest to and just above K^*;
5. K_{j+2} be the strike price that is the second closest to and above K^*;
6. T_1, T_2 and T_3 be the first-nearby, second-nearby and third-nearby expiration dates.

The selection criteria are as following:

▲ For $|K^* - K_j| < |K^* - K_{j+1}|$:

		1st Strike	2nd Strike	3rd Strike
Call Options	1st nearby	$\tilde{P}^{K_{j-1}}_{c,1}$	$\tilde{P}^{K_j}_{c,1}$	$\tilde{P}^{K_{j+1}}_{c,1}$
	2nd nearby	$\tilde{P}^{K_{j-1}}_{c,2}$	$\tilde{P}^{K_j}_{c,2}$	$\tilde{P}^{K_{j+1}}_{c,2}$
	3rd nearby	$\tilde{P}^{K_{j-1}}_{c,3}$	$\tilde{P}^{K_j}_{c,3}$	$\tilde{P}^{K_{j+1}}_{c,3}$
Put Options	1st nearby	$\tilde{P}^{K_j}_{p,1}$	$\tilde{P}^{K_{j+1}}_{p,1}$	$\tilde{P}^{K_{j+2}}_{p,1}$
	2nd nearby	$\tilde{P}^{K_j}_{p,2}$	$\tilde{P}^{K_{j+1}}_{p,2}$	$\tilde{P}^{K_{j+2}}_{p,2}$
	3rd nearby	$\tilde{P}^{K_j}_{p,3}$	$\tilde{P}^{K_{j+1}}_{p,3}$	$\tilde{P}^{K_{j+2}}_{p,3}$

▲ *For* $|K^* - K_j| > |K^* - K_{j+1}|$

		1st Strike	2nd Strike	3rd Strike
Call Options	1st nearby	$\tilde{P}^{K_j}_{c,1}$	$\tilde{P}^{K_{j+1}}_{c,1}$	$\tilde{P}^{K_{j+2}}_{c,1}$
	2nd nearby	$\tilde{P}^{K_j}_{c,2}$	$\tilde{P}^{K_{j+1}}_{c,2}$	$\tilde{P}^{K_{j+2}}_{c,2}$
	3rd nearby	$\tilde{P}^{K_j}_{c,3}$	$\tilde{P}^{K_{j+1}}_{c,3}$	$\tilde{P}^{K_{j+2}}_{c,3}$
Put Options	1st nearby	$\tilde{P}^{K_{j-1}}_{p,1}$	$\tilde{P}^{K_j}_{p,1}$	$\tilde{P}^{K_{j+1}}_{p,1}$
	2nd nearby	$\tilde{P}^{K_{j-1}}_{p,2}$	$\tilde{P}^{K_j}_{p,2}$	$\tilde{P}^{K_{j+1}}_{p,2}$
	3rd nearby	$\tilde{P}^{K_{j-1}}_{p,3}$	$\tilde{P}^{K_j}_{p,3}$	$\tilde{P}^{K_{j+1}}_{p,3}$

▲ *For* $|K^* - K_j| = |K^* - K_{j+1}|$

		1st Strike	2nd Strike	3rd Strike	4th Strike
Call/Put Options	1st nearby	$\tilde{P}^{K_{j-1}}_{c,1}$	$\tilde{P}^{K_{j}}_{c,1}$	$\tilde{P}^{K_{j+1}}_{c,1}$	$\tilde{P}^{K_{j+2}}_{c,1}$
	2nd nearby	$\tilde{P}^{K_{j-1}}_{c,2}$	$\tilde{P}^{K_{j}}_{c,2}$	$\tilde{P}^{K_{j+1}}_{c,2}$	$\tilde{P}^{K_{j+2}}_{c,2}$
	3rd nearby	$\tilde{P}^{K_{j-1}}_{c,3}$	$\tilde{P}^{K_{j}}_{c,3}$	$\tilde{P}^{K_{j+1}}_{c,3}$	$\tilde{P}^{K_{j+2}}_{c,3}$

Since the approximation uses limited data points, we suggest to apply the centring and scaling transformation of the data. This will in turn improve the numerical properties of the quadratic approximation. For example, in order to approximate the option price with strike price K^* by using n nearby available options, we find the coefficients of the quadratic equation in:

$$\hat{K}^* = \frac{K^* - \mu_K}{\sigma_K},$$

where

$$\mu_K = \tfrac{1}{n}\sum_{j=1}^{j=n} K_j, \qquad \sigma_K = \sqrt{\tfrac{1}{n-1}\sum_{j=1}^{j=n}(K_j - \mu_K)^2}.$$

The Algorithm

- Approximate $\tilde{P}^{K^*}_{c,i}$ and $\tilde{P}^{K^*}_{p,i}$, where $i = T_1, T_2, T_3$: for each expiration date, apply the quadratic approximation on option prices at different strike prices to find the approximated option value at K^*,
- Approximate $\tilde{P}^{K^*}_{c,t^*}$ and $\tilde{P}^{K^*}_{p,t^*}$: for each type of option, apply the quadratic approximation on option prices at different time-to-maturity to find the approximated option value with fixed $\triangle t^*$ trading-day time horizon, by using the approximated option prices at $\tilde{K}_{T_1}$, $\tilde{K}_{T_2}$ and $\tilde{K}_{T_3}$,
- The G process for call-put options are constructed by the following equations:

$$G_C(t) = \frac{\tilde{P}^{K^*}_{c,t^*}}{S_{(t)}}, \qquad G_P(t) = \frac{\tilde{P}^{K^*}_{p,t^*}}{S_{(t)}}.$$

For the rest of this paper, the G process is defined as the average of the dynamically purified call option and put option processes, with

$$G(t) = \frac{G_C(t) + G_P(t)}{2}. \tag{4.3.2}$$

It is noted that the approximations from our approach are found to fall within the optimal bounds of option prices when using the convex optimization approach suggested by Bertsimas & Popescu [24].

Analysis of impact of missing prices via Monte-Carlo simulation

Let us study the effectiveness of the quadratic approximation via Monte-Carlo simulation. Particularly, in this experiment, we simulated the time series for $S(t)$ that evolved from the Geometric Brownian Motion and the option price process P_t followed the Black-Scholes' model with the following settings:

- The instantaneous stock price movements was characterized by

  $$dS = \mu S\, dt + \sigma S\, dz$$

  where $S_0 = 3000$, $\mu = 0$ and $\sigma = 0.3$. Since $\mu = 0$, $P_C = P_P$ in this framework.

- Strike prices: The range of the strike price is simulated with a gap of 10 points for each increment.

- Expiration dates: Three different expiration dates for each option type were used with $\triangle t_1 = 0.08$, $\triangle t_2 = 0.18$, $\triangle t_3 = 0.26$ and $\triangle t^* = 0.12$.

The purpose of this experiment is to compare the actual option prices P_t of at-the-money options (i.e. $K_t^* = S_t$) estimated from the Black-Scholes' model with the approximated prices $\hat{P}_t$ that followed the quadratic approximation. We also computed the root-mean-square error as

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(P_t - \hat{P}_t)^2}{n}}.$$

Figure B.3 shows the approximated missing option price using the quadratic approximation method against the actual option price computed from Black-Scholes' formula. Figure B.4 also illustrates the estimated price path and bounds of the missing options using Bertsimas and Popescu's convex optimization approach [24].

4.4 Analysis of the dynamic purified option price process

4.4.1 The data

The S&P/ASX 200 index options are traded on the ASX with the underlying asset being the S&P/ASX 200 index. These option contracts was first listed on 31st March, 2001 and are European in exercise style, with quarterly expiry cycles: March, June, September and December. The exercise prices are set at intervals of 25 index points with new exercise prices automatically created as the underlying index oscillates. The S&P/ASX 200 index options are cash settled and the settlement amount is based on the opening prices of the stocks in the underlying index on the morning of the last trading date. The below table summarises the features of these index options.

S&P/ASX 200 Index Options Features

Underlying asset	ASX approved indexes (currently the ASX 200 Index)
Exercise style	European
Settlement	Cash settled based on the opening prices of the stocks in the underlying index on the morning of the last trading date.
Expiry day	The third Thursday of the month, unless otherwise specified by ASX.
Last trading day	Trading will cease at 12 noon on expiry Thursday. This means trading will continue after the settlement price has been determined.
Premium	Expressed in points
Strike price	Expressed in points
Index multiplier	A specified number of dollars per point e.g. AUD 10
Contract value	The exercise price of the option multiplied by the index multiplier

The daily data for S&P/ASX 200 Index options and S&P/ASX 200 Index were obtained from SIRCA - the Securities Industry Research Centre of Asia-Pacific [112]. In this experiment, we used the last price for each trading day for both option prices and index level. We reported the statistical summary for the daily return and volatility of the index from January 2010 to December 2012 with 757 observations in Table below. When computing the implied volatility from the G-process, we used the overnight interest rate and bank bill swap (BBSW) rates to interpolate the risk-free rate, with a fixed 22-trading-day time horizon. These rates are obtained from the Reserve Bank of Australia [105].

4.4.2 Some statistical properties of the purified option price process

We constructed the G-process for the selected period with a fixed time-horizon of 22 trading days ($\triangle t^* = \frac{22}{252}$) and $K^* = S$ ($\kappa = 1$) for the at-the-money options. The available option data were selected as discussed in Section 4.3.2. Table A.11 provides a summary of statistics for the purified option price process and their logarithmic series. Figure B.7 plots the S&P/ASX 200 level against this dynamically purified option price process G. It is noted that the purified option price process has a relatively small standard deviation.

To examine the properties of this new process against the index level, let's define:

- Increments for the dynamic log index level:

$$r_t = \triangle \ln S_t = \ln \frac{S_t}{S_{t-1}},$$

- Increments for the purified option price process:

$$q_t = \triangle G_t = G_t - G_{t-1}.$$

Figure B.8 plots the cross-correlations between S&P 200 index returns at different leads and lags against daily changes in the G process, with the two dash-dotted lines denoting the 95% confidence band. We observed a strongly negative instantaneous correlation between q_t and r_t. In effect, it was found that corr(q_t, r_t) = -0.8410 for the whole period, while the correlation estimates at other leads and lags are smaller. Also, the statistical standard deviation of r_t and q_t were 27.76% and 4.35% respectively. A breakdown by years for these results is provided in table A.12 .

4.5 Forecasting the market volatility with the purify implied volatility

As discussed, option prices reflect the expectations of the future movements of the underlying assets. Therefore, the volatility implied from the options prices may contain useful information about the future stock market volatility. In this section, we look at the forecasting power of the implied volatility derived from the purified option prices against the traditional volatility index VIX, and their relationship with the future volatility.

We introduce the following processes:

- FV_t: the rolling ex-post (future) volatility measured at 22-trading-day windows, estimated by using the stock index prices at $s = t, t+1, ..., t+22$.
- VIX_t: the non-parametric 22-trading-day volatility index S&P/ASX 200, using mid prices for S&P/ASX 200 put/call options.
- IV_t^G: the 22-trading-day till expiration implied volatility computed from the purified option prices, constructed by using the at-the-money options, average of $IV_t^{G_C}$ and $IV_t^{G_P}$.

For the future volatility, we estimate:

$$FV_t = \left(\frac{252}{\Delta t} \sum_{k=t}^{t+\Delta t} (\overline{R} - R(t_k))^2 \right)^{1/2}, \tag{4.5.1}$$

where

$$R(t_k) = \log S(t_k) - \log S(t_{k-1}), \qquad \overline{R} = \frac{1}{\Delta t} \sum_{k=t}^{t+\Delta t} R(t_k), \qquad \Delta t = 22.$$

Next, the volatility index VIX is derived from the near term and next term options on the S&P/ASX 200 using the out-of-money option. The overnight RBA rate, 1-month, 2-month and 3-month BBSW rates are used to interpolate the risk free rates at each maturity. The general formula to calculate this implied volatility is:

$$\sigma^2 = \frac{2}{T} \sum_i \frac{\triangle K_i}{K_i^2} e^{RT} O(K_i) - \frac{1}{T} \left(\frac{F}{K_0} - 1 \right)^2, \tag{4.5.2}$$

where: σ: implied volatility, T: time to expiration, F: forward index level, K_i: strike price of the i^{th} out-of-the-money option, $\triangle K_i$: interval between strike prices, K_0 = F, R: risk-free interst rate, $O(K_i)$: strike mid-price of each option with strike K_i. More details about the construction of the S&P/ASX 200 VIX process can be found at [107]. Here we simply obtain the raw VIX data from SIRCA.

The Black-Scholes' model 4.2.4 is then used to derive the implied volatility from the at-the-money purified call/put option prices with fixed 22-trading-day time horizon. For the risk-free rate, we interpolate the RBA and BBSW rates similar to that was used for the volatility index as discussed above.

In Figure B.9, we present a time series plot of the three volatilities over the next 22 trading days. It is observed that both IV_t^G and VIX_t could track FV_t's movements, therefore can be used as predictors of the future volatility. We estimate the correlation between those volatility measures. We observe that $corr(FV_t, VIX_t) = 0.8511$ and $corr(FV_t, IV_t^G) = 0.8992$ respectively.

To examine the information content in the new implied volatility process IV_t^G and compare its ability in forecasting the future volatility against the implied volatility index, we consider the following multiple regressions:

Model (1)

$$FV_t = \eta + \alpha FV_{t-\Delta t} + \beta_0 VIX_t + \varepsilon_t;$$

Model (2)

$$FV_t = \eta + \alpha FV_{t-\Delta t} + \beta_0 VIX_t + \beta_1 IV_t^G + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$ are the residual errors of each model. It is noted that Model 1 is based on the conventional multiple regression model of the volatility with the inclusion of the implied volatility index [40] and $FV_{t-\Delta t}$ is the non-overlapped estimation of the future volatility. We extend Model 1 by adding the implied volatility derived from the purified option prices. If these predictors contain some information about the future volatility, the coefficients α and β_i should be statistically significant.

We compute the residuals (the difference of the observed and the actual values) by:

$$\varepsilon_t = \widehat{FV_t} - FV_t,$$

where $\widehat{FV_t}$ is the predicted future volatility and FV_t is the observed future volatility. To compare the accuracy among the models, we compute the root-mean-square-error:

$$RMSE = \sqrt{MSE} = \left(\frac{1}{n} SSE\right)^{1/2} = \left(\frac{1}{n} \sum_{t=1}^{n} (\widehat{FV_t} - FV_t)^2\right)^{1/2}. \tag{4.5.3}$$

for each model. Here, n is the number of observations in the dataset, i.e. 757 observations for our selected sample. The model with smaller RMSE would suggest that the predicted values on average are closer to the observed values, hence is a better model.

We also include the values from Akaike information criterion (AIC) test and Bayesian information criterion (BIC) test to measure the relative quality of the two models. Both criteria are capable of dealing with the trade-off between the goodness of fit and the complexity of the model as more variables are introduced . These criteria are based on a high log-likelihood value, but the penalty term of BIC ($k \ln n$) is potentially much more stringent than that of AIC ($2k$). These information criteria are estimated by:

$$AIC = n \ln MSE + 2k, \qquad BIC = n \ln MSE + k \ln n,$$

where n is the number of observations, k is the number of estimated parameters and MSE is given by equation (4.5.3). The Durbin–Watson (DW) statistic is also reported as a diagnostic check for the independence of errors in regression by detecting the presence of autocorrelation in the residuals, given by:

$$DW = \frac{\sum\limits_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum\limits_{t=1}^{n} \varepsilon_t^2},$$

where $DW < 2$ suggests positive autocorrelation, $DW = 2$ for no autocorrelation and $DW > 2$ for negative autocorrelation.

In our experiment, the data is split into two periods. The first subset includes data points from 01/01/2010 to 31/05/2011. This dataset is used as ‘in-sample’ data for determining the models’ parameters. The rest of the data from 01/06/2011 to 31/12/2012 is then used as ‘out-of-sample’ data for checking the efficiency of the models (1) and (2).

Usually, the coefficients of the regression models can be found by using the Ordinary Least Square (OLS) estimation method. However, previous study showed that the residuals computed from OLS can be highly autocorrelated for such models with Durbin-Watson test values are less than 1 [40]. This will raise the possibility of a spurious regression phenomenon [102] in our prediction. Therefore, the OLS estimation for the coefficients is inconsistent. The Feasible Generalized Least Squares (FGLS) estimation can be used an alternative consistent estimates of those models in the presence of autocorrelated errors. From our experiments, we observed that with OLS estimation, the coefficients of the predictive variables are statistically significant (see Table A.13). This suggests that the selected predictors are useful for predicting the future volatility. More-over, from the model (2), we observe that the implied volatility from the purified option prices plays a more important role in forecasting the future volatility in comparison with the past volatility and the volatility index. However, when examine the Durbin-Watson statistical results, it is observed that the residuals from both models are autocorrelated (0.3619 and 0.3914 for model (1) and (2) respectively). As previously discussed, the estimated coefficients from OLS will be inconsistent. To overcome this, we apply the Cochrane-Orcutt feasible generalised least square (CO-FGLS) estimation method [41] for these models (1) and (2) by modelling the first-order autoregressive on the error terms. Table A.13 also reports the CO-FGLS estimates of these specifications. Hence, models (1) and (2) are adjusted as following:

Model (1′)

$$\begin{aligned} FV_t &= 0.2393 - 03950\, FV_{t-\Delta t} - 0.0468\, VIX_t + \varepsilon_t; \\ \varepsilon_t &= 0.9951\, \varepsilon_{t-1} + e_t, \end{aligned}$$

Model (2′)

$$\begin{aligned} FV_t &= 0.2376 - 0.3890\, FV_{t-\Delta t} - 0.0468\, VIX_t + 0.0569\, IV_t^G + \varepsilon_t; \\ \varepsilon_t &= 0.9949\, \varepsilon_{t-1} + e_t, \end{aligned}$$

with e_t being the input noise. As a result, the Durbin-Watson's tests after using Cochrane-Orcutt's transformation are close to two, indicating that the CO-FGLS procedure has eliminated the autocorrelation of residuals. This confirms the relevance of model (1) and (2).

In terms of accuracy of each model, the RMSE for the in-sample data from Model (1) is 0.0284 and 0.0263 from Model (2). With the out-sample data, RMSE are 0.0171 and 0.0168 for Models (1) and (2) respectively. The RMSE estimates are further improved via Cochrane-Orcutt estimation, with RMSE of model (2′) be 0.0072 vs RMSE of model (1′) be 0.0087. This suggests that the inclusion of the implied volatility from the purified option prices improves the accuracy of our forecast. This is also in agreement with the AIC and BIC tests from Table A.13.

In conclusion, for the selected dataset, we found that the implied volatility from the proposed process contents useful information about the movement of future volatility and can be used to improve the accuracy in forecasting future volatility.

4.6 Discussion

In this chapter, we propose the use of a process G which represents the “dynamically purified" option price process where the impact of the stock price movement is reduced. The process is constructed by using observation of the market option prices. In our experiments, we constructed the process G for the stock index S&P 200 using the at-the-money options. We observed that there is a stable and strongly negative contemporaneous correlation between the increments of stock price return and the increments of G. In additions, we observed a strong correlation between the implied volatility computed from the at-the-money purified option price process and the non-parametric out-the-money implied volatility index VIX. This is an interesting feature since the VIX is calculated using very different data and methods. Similar to VIX , the implied volatility from the purified option prices can be used directly in volatility forecast. We found that the use of the implied volatility from the purified option prices can help improve the accuracy in predicting the future volatility in some experiments with a set of linear regression models. Besides, the new process G can be constructed using observations of just 18 option prices. This is significantly fewer prices than what VIX requires. Therefore, this process can be used to replace VIX in some cases when there is no sufficient data to calculate VIX or one interests in different ranges of strike prices.

Chapter 5

A mixed model for forecasting volatility with high-frequency financial data

5.1 Introduction

Chapter 4 discussed the construction of the purified implied volatility and a simple future volatility forecasting model. We will further incorporate the volatility forecasting models based on the observed realised volatility, the implied volatility from the purified option price process, and non-traditional regression techniques for predicting the future volatility.

As discussed, there has been an enormous body of research on forecasting volatility. Engle[52] and Bollerslev [29] first proposed the ARCH model and the GARCH model for forecasting volatility. These models have been extended in numbers of directions based on the empirical evidences that the volatility process is non-linear, asymmetry and has long memory. Such extensions can be referred to EGARCH [97], GJR-GARCH [63], AGARCH [51], and TGARCH [123].

With the appearance of high-frequency data, Andersen[9] introduced the realized volatility (RV) as discussed in Chapter 2, Section 2.3.2. It was found that the distribution of the standardized exchange rate series and stock returns were almost Gaussian when using the realised volatility, and the logarithm of the realised volatility was also nearly Gaussian. In comparison with the GARCH-type measures, realised volatility is more preferred as it is a model-free measure. Hence, it provides convenience for calculation. In addition, the realised volatility takes the high-frequency data into consideration and exhibits the long memory property. There have been many forecasting models that have been developed to predict the realised volatility. Among those models, the heterogeneous autogressive model for realised volatility (HAR) by Corsi [43] is one to name. The HAR-RV model was developed in accordance with the heterogeneous market hypothesis proposed by Muller[96] and the long memory character of realised volatility by Andersen [9]. Empirical studies have shown that the HAR model has high forecasting performance on future volatility, especially for the out-of-sample data given different time horizons [43, 82].

Another volatility proxy that is often used for forecasting the volatility is the implied volatility indexes as discussed in Chapter 4. We have seen that the implied volatility also contains information about the future volatility.

In this chapter, we will show that the prediction of future volatility can be further improved by using that process via the HAR model and random forest algorithm. We implement the use of classification and regression trees models - machine learning techniques - with the aim to improve the accuracy the forecast of realised volatility. This proposed model is constructed to predict both the direction and the magnitude of realised volatility.

5.2 Volatility measures, HAR model and Random forests algorithm.

5.2.1 Volatility measures

Our focus is on predicting the future realised volatility using high-frequency data via a mixed model with heterogeneous autoregressive for realised and the inclusion of the 'purified' implied volatility. The realised volatility measure for high-frequency data is as discussed in Chapter 2, Section 2.3.2. For the construction of the purified implied volatility, please refer Chapter 4, Section 4.5.

5.2.2 Heterogeneous autoregressive model for realised volatility

Corsi([43], [44]) proposed the heterogeneous autoregressive model for realised volatility as an extension of the Heterogenous ARCH (HARCH) class of models analysed by Muller et al. [96], which recognizes the presence of heterogeneity in the traders. The idea stems from "Fractal Market Hypothesis" [101], "Interacting Agent View" [90] and "Mixture of distribution" hypothesis [5] in the realised volatility process.

It is noted that the definition of the realised volatility involves two time parameters: (1) the intraday return interval $\triangle$, (2) the aggregation period one day. For the heterogeneous autoregressive model of realised volatility [43], it is considered that the latent realised volatility viewed over time horizons longer than one day. The n days historical realised volatility at time t (i.e. $RV_{t-n,t}$) is estimated as an average of daily realised volatility between $(t-n)$ and t. The daily HAR is expressed by

$$RV_{t,t+1} = \beta_0 + \beta_D RV_{t-1,t} + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \varepsilon_{t,t+1}, \tag{5.2.1}$$

where $W = 5$ days, $M = 22$ days and $RV_{t-5,t}, RV_{t-22,t}$ present the average realised volatility of the last 5 days and 22 days respectively. The HAR model can be extended by including the jump component proposed by Barndorff-Nielsen and Shephard [22]. Hence, the general form of the model is

$$RV_{t,t+k} = \beta_0 + \beta_D RV_{t-1,t} + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \beta_J J_{t-k,t} + \varepsilon_{t,t+k} \tag{5.2.2}$$

Most recently, the heterogeneous structure was extended with the inclusion of the leverage effect observed by [26] - the asymmetry in the relationship between returns and volatility [44]. For a given period of time, the leverage level at time t is measured as the average aggregated negative and positive returns during that period where

$$r^+_{t-k,t} = \frac{1}{M}\sum_{j=0}^{M-1} r_{t-j\triangle,t} I_{\{r_{t-k,t},\dots,r_{t,t} \geqslant 0\}}; \quad r^-_{t-k,t} = \frac{1}{M}\sum_{j=0}^{M-1} r_{t-j\triangle,t} I_{\{r_{t-k,t},\dots,r_{t,t} \leqslant 0\}},$$

with M is the number of observation between t-k, t and $\triangle$ is the time step. Therefore, one would include the leverage effect as a predictor for the realised volatility in the next k days as following

$$\begin{aligned} RV_{t,t+k} = \beta_0 &+ \beta_D RV_{t-1,t} + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} \\ &+ \beta_J J_{t-k,t} + \alpha_P r^+_{t-k,t} + \alpha_N r^-_{t-k,t} + \ \varepsilon_{t,t+k}. \end{aligned} \tag{5.2.3}$$

Often, the coefficients $\beta_0, \beta_D, \beta_W, \beta_M, \beta_J, \alpha_P, \alpha_N$ are obtained by using the Ordinary-Least-Square estimation for linear regression models.

From this section onwards, let r^* represent the leverage effect and let HAR-JL denote the heterogeneous autoregressive model with the jump component and leverage effect.

5.2.3 Random forests algorithm

Breiman [32] introduced the random forests algorithm as an ensemble approach that can also be thought of as a form of nearest neighbour predictor. The random forest starts with a standard machine learning technique called “decision trees". We provide a brief summary of this algorithm in this section.

Decision trees

Decision trees algorithm is an approach that uses a set of binary rules to calculate a target class or value. Different from predictors like linear or polynomial regression where a single predictive formula is supposed to hold over the entire data space, decision trees aim to sub-divide the data into multiple partitions using recursive method, then fit simple models for each cell of the partition. Each decision tree has three levels:

- Root node: entry points to a collection of data,
- Inner nodes: a set of binary questions where each child node is available per possible answer,
- Leaf nodes: response to the decision to take if reached.

For example, in order to predict a response or class Y from inputs $X_1, X_2, ..., X_n$, a binary tree is constructed based on the information from each input. At the internal nodes in the tree, a test to one of the inputs is run for a given criterion with logical outcomes: **TRUE** or **FALSE**. Depending on the outcome, a decision is drawn to the next sub-branches corresponding to **TRUE** or **FALSE** responses. Eventually, a final prediction outcome is obtained at the leaf node. This prediction aggregates or averages all the training data points which reach that leaf. Figure B.10 illustrates the binary tree concept.

Algorithm 1 describes how a decision trees can be constructed using CART [33]. This algorithm is computationally simple and quick to fit the data. In addition, as it requires no parametric, no formal distributional assumptions are required. However, one of the main disadvantages of tree-based models is that they exhibit instability and high variance, i.e. a small change in the data can results in very different series of split, or over-fitting. To overcome such a major issue, we use an alternative ensemble approach known as random forests algorithm.

Algorithm 1 CART algorithm for building decision trees.

```
1: Let $N$ be the root node with all available data.
2: Find the feature $F$ and threshold value $T$ that split the samples assigned to $N$ into
   subsets $I_{TRUE}$ and $I_{FALSE}$, to maximise the label purity within these subsets.
3: Assign the pair (F, T) to $N$.
4: If $I(s)$ are too small to be split, attach a ‘child’ leaf nodes $L_{TRUE}$ and $L_{FALSE}$ to $N$ and
   assign the leaves with the most present label in $I_{TRUE}$ and $I_{FALSE}$ respectively.
   If subset $I(s)$ are large enough to be split, attach child nodes $N_{TRUE}$ and $N_{FALSE}$ to $N$,
   then assign $I(s)$ to them respectively.
5: Repeat step 2 - 4 for the new node $N = N_{TRUE}$ and $N = N_{FALSE}$ until the new subsets
   can no longer be split.
```

Random forests

A random forest can be considered as a collection or ensemble of simple decision trees that are selected randomly. It belongs to the class of so-called bootstrap aggregation or bagging technique which aims to reduce the variance of an estimated prediction function. Particularly, a number of decision trees are constructed and random forests will either “vote" for the best decision (classification problems) or “average" the predicted values (regression problems). Here, each tree in the collection is formed by firstly selecting at random, at each node, a small group of input coordinates (also called features or variables hereafter) to split on and, secondly, by calculating the best split based on these features in the training set. The tree is grown using CART algorithm to maximum size, without pruning. Using random forests can lead to significant improvement in prediction accuracy (i.e. better ability to predict new data cases) in comparisons with a single decision tree as discussed in the previous section. Algorithm 2 [32] details how the random forests can be constructed.

Algorithm 2 Random Forests

```
1: Draw a number of bootstrap samples from the original data ($n_{tree}$) to be grown.
2: Sample N cases at random with replacement to create a subset of the data. The subset
   is then split into in-bag and out-of-bag samples at a selected ratio (i.e. 7:3).
3: At each node, for a preselected number $m^{(1)}$, $m$ predictor variables ($m_{try}$) are chosen at
   random from all the predictor variables.
4: The predictor variable that provides the best split, according to some objective function,
   is used to build a binary split on that node.
5: At the next node, choose another m variables at random from all predictor variables.
6: Repeat 3 - 5 until all nodes are grown.
```

Note: $^{(1)}$ - For $m = 1$, the algorithm uses random splitter selection. m can also be set to the total number of predictor variables which is known as Breiman's bagger parameter. If m is much less than the number of predictor variables, Breiman [32] suggests three possible values for m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$ and $2\sqrt{m}$. In this paper, we set m equal to the maximum number of variables of interest used in the proposed model.

Applications of random forests algorithm can be found in machine learning, pattern recognitions, bio-infomatics and big data modelling. Recently, a number of financial literatures have applied random forests algorithm in forecasting stock price as well as developing investment strategies found in [118] and [104]. Here, we introduce an application of the random forests algorithm in forecasting the realised volatility.

5.3 A mixed model for forecasting realised volatility with HAR and random forests

5.3.1 Forecasting the direction of realised volatility

We define two states of the world outcome on the volatility direction as “UP" and “DOWN". Let D_δ be the direction of the realised volatility observed at the time δ, such that

$$D_\delta = \begin{cases} UP & \text{if } \dfrac{RV_\delta}{RV_{\delta-1}} > 1, \\ DOWN & \text{if } \dfrac{RV_\delta}{RV_{\delta-1}} < 1. \end{cases} \tag{5.3.1}$$

In order to forecast the direction of realised volatility, we use a set of predictors (or technical indicators) which are derived from the historical price movement of the underlying asset and its realised volatility. Below is the list of technical indicators we use for forecasting the realised volatility's direction.

1. Average True Range (ATR): ATR is an indicator that measures volatility using the high-low range of the daily prices. ATR is based on n-periods and can be calculated on an intraday, daily, weekly and monthly basis. It is noted that ATR is often used as a proxy of volatility. To estimate ATR_t, we are required to compute the “true range" (TR) such that

$$TR_\delta = max\{H_\delta - L_\delta, |H_\delta - C_{\delta-1}|, |L - C_{\delta-1}|\}, \tag{5.3.2}$$

where H_δ, L_δ, $C_{\delta-1}$ are the current highest return, the current lowest return and the previous last return of a selected period respectively, with absolute values to ensure TR_δ is always positive. Hence, the average true range within n-days is:

$$ATR_{\delta-n,\delta} = \frac{(n-1)ATR_{\delta-n-1,\delta} + TR_\delta}{n}. \tag{5.3.3}$$

2. Close Relative To Daily Range(CRTDR): The location of the last return within the day's range is a powerful predictor of next-returns. Here, CRTDR is estimated by

$$CRTDR_\delta = \frac{C_\delta - L_\delta}{H_\delta - L_\delta}. \tag{5.3.4}$$

   where, H_δ, L_δ and C_δ are the high-low-close returns at time δ for a selected time period using high frequency returns.

3. Exponential Moving Average of realised volatility (EMARV): Exponential moving averages reduce the lag effect in time-series by applying more weight to recent prices. The weighting applied to the most recent price depends on the number of periods n in the moving average and the weighting multiplier κ. The formula for EMARV of n-periods is as following:

$$EMARV_{\delta-n,\delta} = RV_\delta - \kappa \times EMARV_{\delta-n-1,\delta} + EMARV_{\delta-n-1,\delta}. \tag{5.3.5}$$

4. Moving average convergence/divergence oscillator (MACD) measure of realised volatility: MACD is one of the simplest and most effective momentum indicators. It turns two moving averages into a momentum oscillator by subtracting the longer moving average (m-days) from the shorter moving average (n-days). The MACD fluctuates above and below the zero line as the moving averages converge, cross and diverge. We estimate the MACD for realised volatility as:

$$MACDRV_{\delta,m,n} = EMARV_{\delta,m} - EMARV_{\delta,n}. \tag{5.3.6}$$

5. Relative Strength Index for realised volatility (RSIRV): this is also a momentum oscillator that measures the speed and change of volatility movements. We define RSIRV as

$$RSIRV_{\delta-n,\delta} = 1 - \frac{1}{1 + \frac{\overline{RV}^{+}_{\delta-n,\delta}}{\overline{RV}^{-}_{\delta-n,\delta}}}, \tag{5.3.7}$$

where $\overline{RV}^{+}_{\delta-n,\delta}$ is the average increase in volatility and $\overline{RV}^{-}_{\delta-n,\delta}$ is the average decrease in volatility within the n-days.

The steps we take to forecast the volatility direction are listed in Algorithm 3.

Algorithm 3 Forecasting the direction of realised volatility

```
1: Obtain the direction of the realised volatility.
2: Compute the above technical indicators for each observation.
3: Split the data into a training set and a testing set.
4: Apply the random forests algorithm to the training set to develop the pattern solution
   of the realised volatility using the above indicators.
5: Use the solution from Step 4 to predict the direction of the testing set.
```

Figure B.11 demonstrates a possible decision tree that was built for forecasting the direction of realised volatility D_δ using the above steps. In this example, node #4 can be reached when RSI-RV(5)$\geqslant$ 0.5 and TR(10) < 0.0084, with 19% of the in-sample data falls into this category and 91% of these observations are classified as“DOWN". Likewise, node #27 is reached when RSI-RV(5)$\leqslant$ 0.5, $r^+ \geqslant 0.014$ and $0.0049 \leqslant TR(10) < 0.0072$. In random forests, we can construct similar trees but with different structures to classify the direction of the realised volatility based on the information from other predictors.

Let $\widehat{D}_{t,t+k}$ denote the predicted direction of realised volatility at time $t + k$, using the information up to time t from the Algorithm 3.

5.3.2 Forecasting the level of realised volatility

To forecast the level of realised volatility, we consider the heterogeneous autoregression model as discussed in Section 5.2.2. We further include the purified implied volatility and the predicted direction of the future volatility as new predictive variables. Particularly, the Model 5.2.3 is extended to

$$\begin{aligned} RV_{t,t+k} = \beta_0 + \beta_D RV_{t-1,t} + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \beta_J J_{t-k,t} + \alpha r^*_{t-k,t} \\ + \gamma PV_{t-k,t+22} + \kappa \widehat{D}_{t,t+k} + \varepsilon_{t,t+k}. \end{aligned} \tag{5.3.8}$$

We also consider the logarithmic form of this model as the logarithmic of the realised volatility is often believed to be a smoother process. Thus, we model $\log RV$ as

$$\begin{aligned}\log RV_{t,t+k} = \beta_0 + \beta_D \log RV_{t-1,t} + \beta_W \log RV_{t-5,t} + \beta_M \log RV_{t-22,t} + \beta_J \log(1 + J_{t-k,t}) \\ + \alpha \log |r^*_{t-1,t}| + \gamma log(PV_{t-k,t+22}) + \kappa \widehat{D}_{t,t+k} + \varepsilon_{t,t+k}, \qquad (5.3.9)\end{aligned}$$

where $k = \{1, 5, 22\}$ for 1-day, 5-day and 22-day time horizons. We use $\log(1 + J_{t-k,t})$ instead of $\log(J_{t-k,t})$ to allow for the cases where $J_{t-k,t} = 0$ and the leverage effect is measured by $\log |r^*_{t-1,t}|$ to allow for the average aggregated negative returns.

The parameters in models 5.3.8 and 5.3.9 (HAR-JL-PV-D) are fitted using the random forests regression algorithm. It is important to note that for the in-sample data, we replace $\widehat{D}_{t,t+k}$ with the actual direction $D_{t,t+k}$ to measure the impact of the direction variable in forecasting the realised volatility.

5.4 Numerical experiments and results

5.4.1 Data summary

We demonstrate the proposed model b analysing the S&P ASX 200 Index high frequency returns data and its realised volatility. Our dataset is collected from SIRCA for the period 1^{st} January, 2008 to 31^{st} December, 2014. The Australian Stock Exchange is open between 10:00 am to 4:00 pm. We collect the tick-by-tick S&P 200 levels hence the prices are not recorded at equispaced time points. We use the previous tick aggregation method to force the observed prices into an equispaced grid, i.e. taking the last price realized before each grid point and obtain the 15-second frequency data. The daily realised volatility (with 1762 observations) is then estimated using these 15-second prices.

Table A.14 provides a summary of the 15-second realised volatility measured using different time-windows. It is observed that both non-logarithmic and logarithmic series are skewed and non-normal. This suggests that the Ordinary Least Square estimation approach will not be applicable for our dataset. As a result, we compare the maximum likelihood estimation (MLE) with the random forests algorithm instead. In terms of correlation coefficients between the series, we observed that the computed realised volatility exhibits the long memory effect. Further, the purified implied volatility is strongly correlated with realised volatility measures, which indicates that PV can be a useful predictor of realised volatility.

5.4.2 Measuring errors

Since the paper focuses on forecasting both the realised volatility's direction and its magnitude, we use the following measures to compare each model.

Classification problem

In forecasting the direction of the realised volatility, the classification problem consists of only two stages. We measure the accuracy of the forecast as follows.

Let's define

- True positive (TP): the number of days that are observed with "Down" signals were correctly predicted.
- False positive (FP): the number of days that are observed with "Down" signals were predicted with "Up" signals.
- False negative (FN): the number of days that are observed with "Up" signals were predicted with "Down" signals.
- True negative (TN): the number of days that are observed with "Up" signals were correctly predicted.
- Accuracy: the proportion of the total number of correct prediction

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{5.4.1}$$

Regression problem

We split our data into two subsets: the training (in-sample) data and the test (out-of-sample) data. Since random forests algorithm is used, we measure the accuracy of the model proposed method for training data and test data separately.

Measuring error for training data

For the random forests algorithm, an estimate of the error rate can be obtained based on the training as following:

1. For each bootstrap, predict the out-of-bag values using the tree grown within the bootstrap sample.

2. Aggregate the OOB predictions and calculate the mean square error rate by

$$MSE_{OOB} = \frac{1}{m}\sum_{t=1}^{n}\left\{RV_t - \overline{RV_t}^{OOB}\right\}^2 \tag{5.4.2}$$

   where m is the number of observations in the OOB data (i.e. $m < N$) and $\overline{RV_t}^{OOB}$ is the average of the OOB predictions for the t^{th} observation.

3. Estimate the percentage variance explained as a measure of goodness of fit by

$$1 - \frac{MSE_{OOB}}{\sigma^2_{RV}} \tag{5.4.3}$$

   where σ^2_{RV} is the variance of the OOB sample.

Measuring error for test data

Let RV_t denote the t^{th} observation, $\widehat{RV_t}$ denote its forecast, and k be the number of data points observed in the selected period. The error measures include

- Mean absolute error

$$MAE = \frac{1}{k}\sum_{t=1}^{k}|RV_t - \widehat{RV_t}|. \tag{5.4.4}$$

- Mean absolute percentage error

$$MAPE = \frac{1}{k}\sum_{t=1}^{k}\frac{|RV_t - \widehat{RV_t}|}{RV_t}. \tag{5.4.5}$$

- Root mean square error

$$RMSE = \sqrt{\frac{1}{k}\sum_{t=1}^{k}(RV_t - \widehat{RV_t})^2}. \tag{5.4.6}$$

- Root mean square percentage error

$$RMSPE = \sqrt{\frac{1}{k}\sum_{t=1}^{k}\left(\frac{RV_t - \widehat{RV_t}}{RV_t}\right)^2}. \tag{5.4.7}$$

5.4.3 Results of experiments

Table A.15 compares the in-sample forecast results of the proposed model using random forest algorithm. For the selected time horizons, the inclusion of purified implied volatility shows improvement in the forecast accuracy against the original HAR-JL model (based on RMSE measure and % OOB variance explained), where the logarithmic RV series perform better than the non-logarithmic RV series. It is also observed that the direction indicator can further improve the forecast results. Such improvement is most significant for the 1-day forecast (with 75.58 % and 80.65% variance explained for RV and log RV in comparison with 57.81% and 61.66% from the HAR-JL model respectively). For the 5-day and 22-day in-sample forecasts, we observe slight improvements in RMSE with better goodness of fit.

In forecasting the direction of the out-sample realised volatility, we obtain the accuracy of the hit-rate at 80.05%, 72.85% and 65.22% for 1-day, 5-day and 22-day forecast respectively. This suggests our classification model can perform better for the short-term forecasts than long-term forecasts. This can be explained by the fact that long-term forecast require not only technical indicators but also fundamental indicators and long-term expectations from the market.

Table A.16 provides summaries of the forecast errors for the out-sample data. For consistent comparison, we take the exponential transformation of the predicted values of the log RV to obtain the predicted RV. In general, the out-of-sample performances of the proposed model are in favours with the in-sample performances. The MAPE and RMSPE for 1-day forecast of the RV from the HAR-JL-PV-D are reduced by 8% and 11% respectively, while the MAPE and RMSPE for 5-day and 22-day are reduced by 3% and 5%. When comparing the HAR-JL-PV model against the HAR-JL-D model, it can be seen that the the forecast errors are smaller for the HAR-JL-PV model for these time horizons. This is anticipated as we found that the forecast of long-term direction is less accurate for 5-day and 22-day forecasts. However, the HAR-JL-D model still performs better than the HAR-JL alone and the HAR-JL-PV-D model provides the best fits.

We present in Figure B.13 the actual S&P200's realised volatility measured under different time horizons from 1^{st} January, 2014 to 31^{st} December, 2014, with the predicted realised volatility using maximum likelihood estimation for the HAR-JL model (left panel) and using random forests estimation for the HAR-JL-PV-D model (right panel).

5.5 Discussion

This chapter introduces an application of the random forests algorithm in forecasting the realised volatility. For the classification problem, the algorithm shows that it is possible to forecast directions of the realised volatility . For the regression problem, with its non-linear structure, the technique was able to reduce the forecasting errors from volatility clustering systematically under different time horizons. The empirical results of S&P 200 shows that the existing HAR model framework was improved by including the purified implied volatility and applying this machine leaning technique. We suggest to further investigate the roles of the purified implied volatility and random forests algorithm in other high frequency models of volatility.

Concluding remarks

The motivation for this thesis stems from the critical role of volatility in financial investment. With the use of high frequency data, this leads to how different the new settings for estimating and modelling financial volatility, in comparison with the traditional methods. Extensive research had been carried out in many aspects of the financial time series and includes analysis of estimators, errors reduction and predictability enhancement. This thesis aims to examine some of the estimators of financial volatility and investigate its results for different sampling frequency.

The main contributions of this thesis is the comparison of the effectiveness of different volatility measures and models. Additional contributions include the development of a process for modelling the dependency of volatility on sampling frequency, the construction of a new implied volatility process via reducing the impact of price movement on the volatility estimation, and some improvement in accuracy for the forecasting of future volatility.

For modelling the dependence of volatility on sampling frequency, it is showed that the prescribed dependence of the volatility on the sampling frequency can be achieved using delay equations for the underlying prices. These equations allow to model the price processes with volatility that increases when the sampling rates increase, as well as the inverse effect where the volatility decreases with the increase in sampling frequencies. While higher dimensional multi-timescale dependence with delay equations have yet to be explored, the proposed framework is the foundation for this research area. However, we did not consider pricing of options on the underlying time series and relations between the historical volatility and the implied volatility. We leave this for future research.

For the proposed implied volatility measure, it is showed that the artificial "dynamically purified" price process allows to eliminate the impact of the stock price movements on the implied volatility. The complete elimination would be possible if the option prices were available for continuous sets of strike prices and expiration times. However, in practice, only finite sets of prices are available. To overcome the incompleteness of the available option prices, the first order Taylor series extrapolation and quadratic interpolation were examined. It is also showed that this new implied volatility process can also be used as a proxy for forecasting of the future volatility, in comparison with the traditional implied volatility process. While the construction of the "purified" implied volatility using daily financial data, we have yet to explore the effectiveness of this approach for the intra-day estimations and forecasts. Further investigation can be followed from this direction.

For the forecasting of the volatility, with the aim to improve the forecast accuracy, the mixed model of heterogeneous autoregressive model and random forest algorithm was studied for different forecasting horizons. It is showed that for the classification problem, the proposed algorithm was able to forecast directions of the realised volatility; and for the regression problem, with its non-linear structure, the technique was able to reduce the forecasting errors from volatility clustering systematically under different time horizons. With the extended heterogeneous autoregressive model via the inclusion of the "purified" implied volatility, it is found that the forecast of both direction and magnitude of the realised volatility were improved. It will definitely be of interest to know if the choice of other machine learning technique will be significant using this innovative approach in forecasting the future volatility.

Appendix A

Tables

Table A.1 Simulation of model 3.2.3 for $\lambda < 0$.

$\kappa = \lambda\delta\tau$	5			20			120		
		mean	sd		mean	sd		mean	sd
$\kappa = -0.0005$ $\lambda = -196.56$	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3000	0.0003
	σ_{5min}	0.3004	0.0015	σ_{5min}	0.3014	0.0015	σ_{5min}	0.3015	0.0010
	σ_{hour}	0.3009	0.0052	σ_{hour}	0.3027	0.0052	σ_{hour}	0.3140	0.0030
		mean	sd		mean	sd		mean	sd
$\kappa = -0.0025$ $\lambda = -982.8$	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3002	0.0003
	σ_{5min}	0.3026	0.0015	σ_{5min}	0.3073	0.0015	σ_{5min}	0.3101	0.0016
	σ_{hour}	0.3029	0.0053	σ_{hour}	0.3141	0.0054	σ_{hour}	0.3892	0.0069
		mean	sd		mean	sd		mean	sd
$\kappa = -0.005$ $\lambda = -1965.6$	σ_{15sec}	0.3002	0.0003	σ_{15sec}	0.3006	0.0003	σ_{15sec}	0.3000	0.0003
	σ_{5min}	0.3052	0.0016	σ_{5min}	0.3153	0.0019	σ_{5min}	0.3341	0.0025
	σ_{hour}	0.3059	0.0049	σ_{hour}	0.3301	0.0057	σ_{hour}	0.5667	0.0050

Table A.2 Simulation of model 3.2.3 $\lambda > 0$.

$\kappa = \lambda\delta\tau$	5			20			120		
		mean	sd		mean	sd		mean	sd
$\kappa = 0.0005$ $\lambda = 196.56$	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3000	0.0003	σ_{15sec}	0.3000	0.0003
	σ_{5min}	0.3000	0.0015	σ_{5min}	0.2985	0.0015	σ_{5min}	0.2987	0.0015
	σ_{hour}	0.3000	0.0050	σ_{hour}	0.2971	0.0050	σ_{hour}	0.2873	0.0052
		mean	sd		mean	sd		mean	sd
$\kappa = 0.005$ $\lambda = 1965.6$	σ_{15sec}	0.3003	0.0003	σ_{15sec}	0.3004	0.0003	σ_{15sec}	0.3003	0.0003
	σ_{5min}	0.2948	0.0015	σ_{5min}	0.2866	0.0015	σ_{5min}	0.2912	0.0015
	σ_{hour}	0.2941	0.0037	σ_{hour}	0.2749	0.0047	σ_{hour}	0.2123	0.0037
		mean	sd		mean	sd		mean	sd
$\kappa = 0.05$ $\lambda = 19656$	σ_{15sec}	0.3003	0.0003	σ_{15sec}	0.3038	0.0004	σ_{15sec}	0.3067	0.0004
	σ_{5min}	0.2559	0.0012	σ_{5min}	0.2148	0.0007	σ_{5min}	0.2848	0.0021
	σ_{hour}	0.2503	0.0020	σ_{hour}	0.1594	0.0019	σ_{hour}	0.1045	0.0020

Table A.3 The average and standard deviation of the annual volatility using different windows for $\kappa = 0.005$ ($\lambda = 1965.6$) and $\tau = 120$.

Annual volatility	Mean	Standard Deviation
1-day windows	σ_{15sec}= 0.3002 σ_{5min} = 0.2923 σ_{hour} = 0.2111	σ_{15sec}= 0.0054 σ_{5min} = 0.0254 σ_{hour} = 0.0792
5-day windows	σ_{15sec}= 0.3002 σ_{5min} = 0.2926 σ_{hour} = 0.2143	σ_{15sec}= 0.0022 σ_{5min} = 0.0098 σ_{hour} = 0.02482
22-day windows	σ_{15sec}= 0.3002 σ_{5min} = 0.2926 σ_{hour} = 0.2134	σ_{15sec}= 0.0010 σ_{5min} = 0.0045 σ_{hour} = 0.0113

Table A.4 Volatility of stock indexes under different sampling frequency.

Stock Index	2008	2009	2010
DAX	$\sigma_{15sec} =$ 0.3569	$\sigma_{15sec} =$ 0.2476	$\sigma_{15sec} =$ 0.1755
	$\sigma_{5min} =$ 0.3928	$\sigma_{5min} =$ 0.2682	$\sigma_{5min} =$ 0.1854
	$\sigma_{hourly} =$ 0.3959	$\sigma_{hourly} =$ 0.2777	$\sigma_{hourly} =$ 0.1888
FTSE 100	$\sigma_{15sec} =$ 0.2660	$\sigma_{15sec} =$ 0.1822	$\sigma_{15sec} =$ 0.1346
	$\sigma_{5min} =$ 0.3462	$\sigma_{5min} =$ 0.2313	$\sigma_{5min} =$ 0.1686
	$\sigma_{hourly} =$ 0.3606	$\sigma_{hourly} =$ 0.2344	$\sigma_{hourly} =$ 0.1708
IBEX 35	$\sigma_{15sec} =$ 0.3289	$\sigma_{15sec} =$ 0.2428	$\sigma_{15sec} =$ 0.2549
	$\sigma_{5min} =$ 0.3569	$\sigma_{5min} =$ 0.2446	$\sigma_{5min} =$ 0.2772
	$\sigma_{hourly} =$ 0.3620	$\sigma_{hourly} =$ 0.2571	$\sigma_{hourly} =$ 0.2847
SMI	$\sigma_{15sec} =$ 0.3265	$\sigma_{15sec} =$ 0.2106	$\sigma_{15sec} =$ 0.1603
	$\sigma_{5min} =$ 0.3421	$\sigma_{5min} =$ 0.2171	$\sigma_{5min} =$ 0.1469
	$\sigma_{hourly} =$ 0.3513	$\sigma_{hourly} =$ 0.2278	$\sigma_{hourly} =$ 0.1564
S&P 500	$\sigma_{15sec} =$ 0.2881	$\sigma_{15sec} =$ 0.1952	$\sigma_{15sec} =$ 0.1300
	$\sigma_{5min} =$ 0.3654	$\sigma_{5min} =$ 0.2507	$\sigma_{5min} =$ 0.1771
	$\sigma_{hourly} =$ 0.3747	$\sigma_{hourly} =$ 0.2601	$\sigma_{hourly} =$ 0.1808
S&P 200	$\sigma_{15sec} =$ 0.2124	$\sigma_{15sec} =$ 0.1464	$\sigma_{15sec} =$ 0.1055
	$\sigma_{5min} =$ 0.2796	$\sigma_{5min} =$ 0.1805	$\sigma_{5min} =$ 0.1288
	$\sigma_{hourly} =$ 0.3367	$\sigma_{hourly} =$ 0.2100	$\sigma_{hourly} =$ 0.1530
TSX 60	σ_{15sec} 0.3345	$\sigma_{15sec} =$ 0.2256	$\sigma_{15sec} =$ 0.1246
	σ_{5min} 0.3884	$\sigma_{5min} =$ 0.2602	$\sigma_{5min} =$ 0.1341
	σ_{hourly} 0.4207	$\sigma_{hourly} =$ 0.3884	$\sigma_{hourly} =$ 0.1352

Table A.5 Volatility of company stocks under different sampling frequency.

Stock Symbol	2008	2009	2010
AAPL	$\sigma_{15sec} = 0.7032$ $\sigma_{5min} = 0.6347$ $\sigma_{hourly} = 0.5832$	$\sigma_{15sec} = 0.3508$ $\sigma_{5min} = 0.3373$ $\sigma_{hourly} = 0.3267$	$\sigma_{15sec} = 0.3180$ $\sigma_{5min} = 0.2915$ $\sigma_{hourly} = 0.2738$
IBM	$\sigma_{15sec} = 0.5245$ $\sigma_{5min} = 0.4507$ $\sigma_{hourly} = 0.3932$	$\sigma_{15sec} = 0.3085$ $\sigma_{5min} = 0.2766$ $\sigma_{hourly} = 0.2622$	$\sigma_{15sec} = 0.2179$ $\sigma_{5min} = 0.2005$ $\sigma_{hourly} = 0.1812$
JPM	$\sigma_{15sec} = 0.9068$ $\sigma_{5min} = 0.8217$ $\sigma_{hourly} = 0.7487$	$\sigma_{15sec} = 0.7274$ $\sigma_{5min} = 0.6741$ $\sigma_{hourly} = 0.6586$	$\sigma_{15sec} = 0.3238$ $\sigma_{5min} = 0.3039$ $\sigma_{hourly} = 0.2873$
GE	$\sigma_{15sec} = 0.7163$ $\sigma_{5min} = 0.6220$ $\sigma_{hourly} = 0.5790$	$\sigma_{15sec} = 0.7137$ $\sigma_{5min} = 0.6040$ $\sigma_{hourly} = 0.5844$	$\sigma_{15sec} = 0.4021$ $\sigma_{5min} = 0.3124$ $\sigma_{hourly} = 0.2919$
GOOG	$\sigma_{15sec} = 0.8114$ $\sigma_{5min} = 0.6276$ $\sigma_{hourly} = 0.5937$	$\sigma_{15sec} = 0.3543$ $\sigma_{5min} = 0.3053$ $\sigma_{hourly} = 0.2944$	$\sigma_{15sec} = 0.3353$ $\sigma_{5min} = 0.2820$ $\sigma_{hourly} = 0.2587$
MSFT	$\sigma_{15sec} = 0.7032$ $\sigma_{5min} = 0.6347$ $\sigma_{hourly} = 0.5832$	$\sigma_{15sec} = 0.3908$ $\sigma_{5min} = 0.3373$ $\sigma_{hourly} = 0.3267$	$\sigma_{15sec} = 0.3180$ $\sigma_{5min} = 0.2915$ $\sigma_{hourly} = 0.2738$
XOM	$\sigma_{15sec} = 0.5071$ $\sigma_{5min} = 0.4908$ $\sigma_{hourly} = 0.4876$	$\sigma_{15sec} = 0.2911$ $\sigma_{5min} = 0.2708$ $\sigma_{hourly} = 0.2620$	$\sigma_{15sec} = 0.2156$ $\sigma_{5min} = 0.2083$ $\sigma_{hourly} = 0.1834$

Table A.6 The values of error in calibrating S&P500 historical data for some (λ_i, τ_j).

$\lambda\tau$	5	6	7	8	9	10	11	12	13
-16707.6	0.0396	0.0120	0.0147	0.0500	0.0716	0.0845	0.1260	0.1637	0.2335
-12776.4	0.0593	0.0557	0.0268	0.0108	**0.0098**	0.0306	0.0554	0.0674	0.1214
-8845.2	0.0736	0.0645	0.0507	0.0427	0.0378	0.0291	0.0246	0.0217	0.0251
-4914.0	0.0852	0.0878	0.0813	0.0739	0.0716	0.0691	0.0698	0.0612	0.0626
-982.8	0.1108	0.1080	0.1091	0.1063	0.1001	0.1051	0.1040	0.1049	0.1017

Table A.7 Measures of $\hat{\sigma}_{15sec} - \sigma_{15sec}$ in calibrating S&P500 historical data.

$\lambda\tau$	5	6	7	8	9	10	11	12	13
-16707.6	0.0012	0.0015	0.0019	0.0026	0.0028	0.0038	0.0042	0.0052	0.0065
-12776.4	0.0007	0.0010	0.0007	0.0012	0.0001	0.0017	0.0018	0.0020	0.0031
-8845.2	0.0003	0.0008	-0.0001	0.0003	0.0007	0.0009	0.0003	0.0013	0.0011
-4914.0	-0.0001	-0.0004	0.0002	0.0002	0.0005	0.0004	0.0004	-0.0004	0.0008
-982.8	-0.0003	0.0002	0.0006	0.0003	0.0001	-0.0001	-0.0001	-0.0002	0.0003

Table A.8 The values of error in calibrating GOOG historical data for some (λ_i, τ_j).

$\lambda\tau$	5	6	7	8	9	10	11	12	13
17886.96	0.0919	0.0661	0.0456	0.0307	0.0154	0.0052	0.0128	0.0137	0.0231
17690.40	0.0797	0.0681	0.0482	0.0317	0.0114	0.0145	0.0081	0.0119	0.0254
17493.84	0.0946	0.0651	0.0504	0.0332	0.0238	**0.0030**	0.0079	0.0173	0.0246
17297.28	0.0967	0.0633	0.0460	0.0369	0.0199	0.0113	0.0076	0.0102	0.0210
17100.72	0.0840	0.0678	0.0448	0.0349	0.0228	0.0141	0.0089	0.0105	0.0196

Table A.9 Measures of $\hat{\sigma}_{15sec} - \sigma_{15sec}$ in calibrating GOOG historical data.

$\lambda\tau$	5	6	7	8	9	10	11	12	13
17886.96	0.0015	0.0034	0.0047	0.0039	0.0044	0.0051	0.0067	0.0053	0.0076
17690.40	0.0015	0.0024	0.0046	0.0055	0.0049	0.0062	0.0065	0.0066	0.0055
17493.84	0.0031	-0.0001	0.0059	0.0054	0.0053	0.0002	0.0049	0.0057	0.0073
17297.28	0.0021	0.0018	0.0044	0.0033	0.0046	0.0053	0.0048	0.0056	0.0070
17100.72	0.0025	0.0036	0.0036	0.0051	0.0051	0.0066	0.0055	0.0051	0.0055

Table A.10 Statistical summary of returns for Australian stock index S&P 200 (AXJO) at different sampling frequency for the period 1^{st} January, 2008 to 31^{st} December, 2013.

Asset	Frequency	Mean	SD	Skewness	Kurtosis	Min	Max
AXJO	Daily	-0.0002446	0.0131496	-0.3767034	2.863	-0.0714057	0.0458126
AXJO	60-min	-0.0000406	0.0054814	-0.3073395	18.048	-0.0515180	0.0627551
AXJO	30-min	-0.0000204	0.0038930	-0.2759757	39.654	-0.0597412	0.0682942
AXJO	15-min	-0.0000102	0.0027673	-0.3159879	74.005	-0.0532332	0.0659015
AXJO	5-min	-0.0000034	0.0013278	-0.2473133	80.334	-0.0309539	0.0391501
AXJO	1-min	-0.0000007	0.0004838	-0.2353127	224.649	-0.0249049	0.0261786

Table A.11 Summary statistics for at-the-money call/put options price process and their average.

Series	Mean	Std. Dev.	Skew.	Kurt.	Min.	Max.
$ATM_{G_C(t)}$	0.0158	0.0057	1.2610	2.4451	0.0061	0.0435
$ATM_{G_P(t)}$	0.0160	0.0055	1.3886	2.4133	0.0073	0.0422
$ATM_{G_C(t)}$	0.0159	0.0052	1.5093	2.6646	0.0083	0.0382
$\ln(ATM_{G_C(t)})$	-4.2038	0.3406	0.1209	0.0639	-5.0995	-3.135
$\ln(ATM_{G_P(t)})$	-4.1860	0.3113	0.4517	0.0713	-4.9199	-3.1653
$\ln(ATM_{G(t)})$	-4.1863	0.2946	0.6461	0.2068	-4.8036	-3.2649

Table A.12 Summary statistics for daily log return of S&P 200 and the 22 trading-day G' process.

		$\triangle \ln S_t$		$\triangle G'_t$		$Cross - correlation$
Period	No. Obs	Mean	SD(annual)	Mean	SD(annual)	corr$(r_t, q_t), lag = 0$
2010	253	-0.0001	15.80%	-0.0005	2.12%	-0.8167
2011	251	-0.0006	19.56%	0.001	3.42%	-0.8575
2012	253	0.0005	11.85%	-0.0025	1.50%	-0.7984
All	757	-0.0001	27.76%	-0.0005	4.35%	-0.8410

Table A.13 Regression results for in-sample data.

	Dependent variable: FV_t			
	OLS Estimation		Cochrane–Orcutt Estimation	
Coefficients	Model (1)	Model (2)	Model (1′)	Model (2′)
η	−0.0285	−0.0236	0.2393	0.23763
$FV_{t-\triangle t}$	0.1117	0.0980	−0.3950	−0.3890
VIX_t	0.8126	0.2807	−0.0084	−0.0468
IV_t^G		0.7060		0.0569
ρ			0.9951	0.9949
Durbin-Watson	0.3619	0.3914	1.9810	1.9834
AIC	-2581.402	-2631.773	-3633.377	-3633.845
BIC	-2563.841	-2609.821	-3616.362	-3617.576
RMSE	0.0284	0.0263	0.0087	0.0072
Note:	all coefficients are significant at p =1%.			

Table A.14 Statistical summary of S&P/ASX 200's 15-second Realised Volatility at different time horizons from 1-January, 2008 to 31-December, 2014, and their correlation matrix.

Series	Mean	Std. Dev.	Skew.	Kurt.	Min.	Max.	$RV_{t-1,t}$	$RV_{t-5,t}$	$RV_{t-22,t}$	PV
$RV_{t-1,t}$	0.1335	0.0848	2.4957	8.9530	0.0328	0.7811	1	0.8441	0.7523	0.7757
$RV_{t-5,t}$	0.1335	0.0721	2.0481	5.4748	0.0484	0.5453	0.8441	1	0.9042	0.8919
$RV_{t-22,t}$	0.1331	0.0664	1.8304	3.9311	0.0593	0.4228	0.7523	0.9042	1	0.9180
PV	0.1614	0.0705	1.5181	2.8461	0.0698	0.5004	0.7757	0.8919	0.9180	1
Series	Mean	Std. Dev	Skew.	Kurt.	Min.	Max.	$\log RV_{t-1,t}$	$\log RV_{t-5,t}$	$\log RV_{t-22,t}$	$\log PV$
$\log RV_{t-1,t}$	-2.1588	0.5139	0.5678	0.2336	-3.4184	-0.2471	1	0.8548	0.7739	0.7936
$\log RV_{t-5,t}$	-2.1244	0.4499	0.6619	0.0960	-3.0274	-0.6064	0.8548	1	0.9124	0.8972
$\log RV_{t-22,t}$	-2.113	0.4213	0.7156	-0.0407	-2.8248	-0.8608	0.7739	0.9124	1	0.9017
$\log PV$	-1.9044	0.3893	0.5190	-0.3229	-2.6618	-0.6923	0.7936	0.8972	0.9017	1

Table A.15 Forecasting error of the realised volatility for the in-sample data from 1-Jan, 2008 to 31-Dec, 2013.

		1-day				5-day				22-day			
		HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D	HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D	HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D
RV	RMSE	0.0996	0.0957	0.0509	0.0502	0.0378	0.0336	0.0326	0.0295	0.0383	0.0323	0.0339	0.0287
	% OOB Var	57.81	59.61	74.68	75.58	79.28	80.28	81.13	81.81	74.47	76.44	78.09	79.44
log RV	RMSE	0.0031	0.0029	0.0018	0.0018	0.002	0.0011	0.0010	0.0010	0.0011	0.0010	0.0009	0.0001
	% OOB Var	61.66	63.12	80.39	80.65	80.55	82.70	83.25	84.83	77.48	81.97	80.05	83.12

Table A.16 Forecasting error of the realised volatility for the out-sample data from 1-Jan, 2014 to 31-Dec, 2014.

		1-day				5-day				22-day			
		HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D	HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D	HAR-JL	HAR-JL-PV	HAR-JL-D	HAR-JL-PV-D
RV	MAE	0.0212	0.0205	0.0176	0.0171	0.0147	0.0135	0.0137	0.0127	0.0184	0.0142	0.017	0.0137
	MAPE	0.2715	0.2516	0.2042	0.1974	0.1814	0.1573	0.1670	0.1500	0.2245	0.1630	0.2094	0.1576
	RMSE	0.0285	0.0277	0.0247	0.0235	0.0192	0.0182	0.0180	0.0168	0.0223	0.0182	0.0209	0.0176
	RMSPE	0.3610	0.3245	0.2709	0.2568	0.2352	0.2025	0.2181	0.1926	0.2745	0.2046	0.2602	0.1973
log RV	MAE	0.0206	0.0201	0.0170	0.0165	0.0143	0.0135	0.0130	0.0129	0.0170	0.0138	0.0156	0.0135
	MAPE	0.2525	0.2331	0.1947	0.1878	0.1740	0.1553	0.1574	0.1481	0.2058	0.1576	0.1881	0.1532
	RMSE	0.0279	0.0280	0.0239	0.0233	0.0185	0.0185	0.0175	0.0175	0.0206	0.0177	0.0191	0.0174
	RMSPE	0.3250	0.2929	0.2573	0.2454	0.2230	0.1980	0.2116	0.1913	0.2499	0.1958	0.2310	0.1912

Note: as random forests algorithm requires a random selection process, for consistent comparison across models, we preset the random seed to a specific value before applying the algorithm to each of the above models. Table A.15 reported the average accuracy based on 10 folds cross-validation for the in-sample data.

Appendix B

Figures

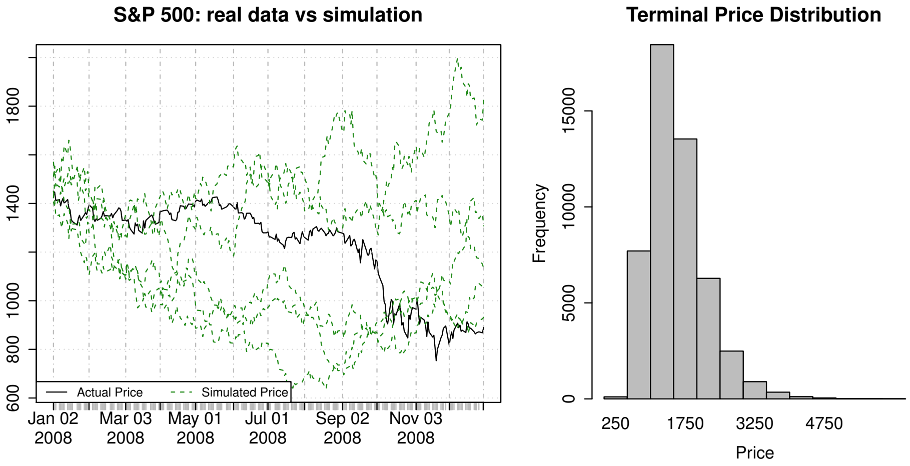

Figure B.1: The Actual Price vs the Simulated Price for SP500 from January to December, 2008; and the terminal price distribution of 100,000 instances.

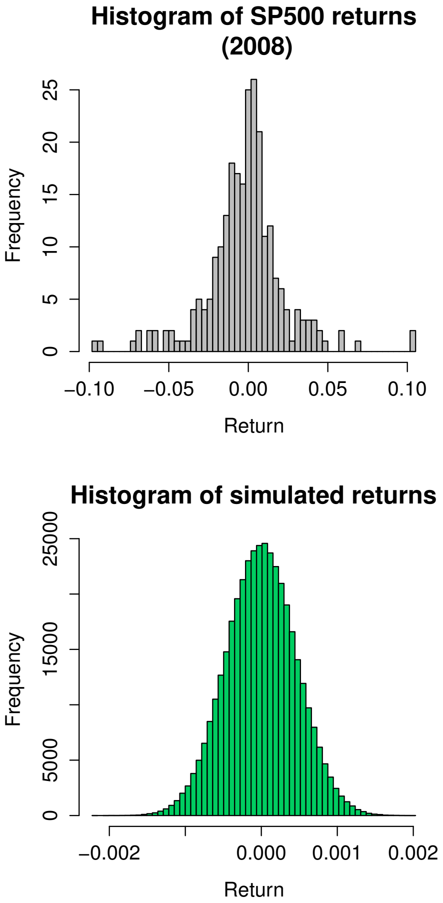

Figure B.2: Histograms and Q-Q Plots for the historical S&P500 (2008) vs the simulated process from the proposed model.

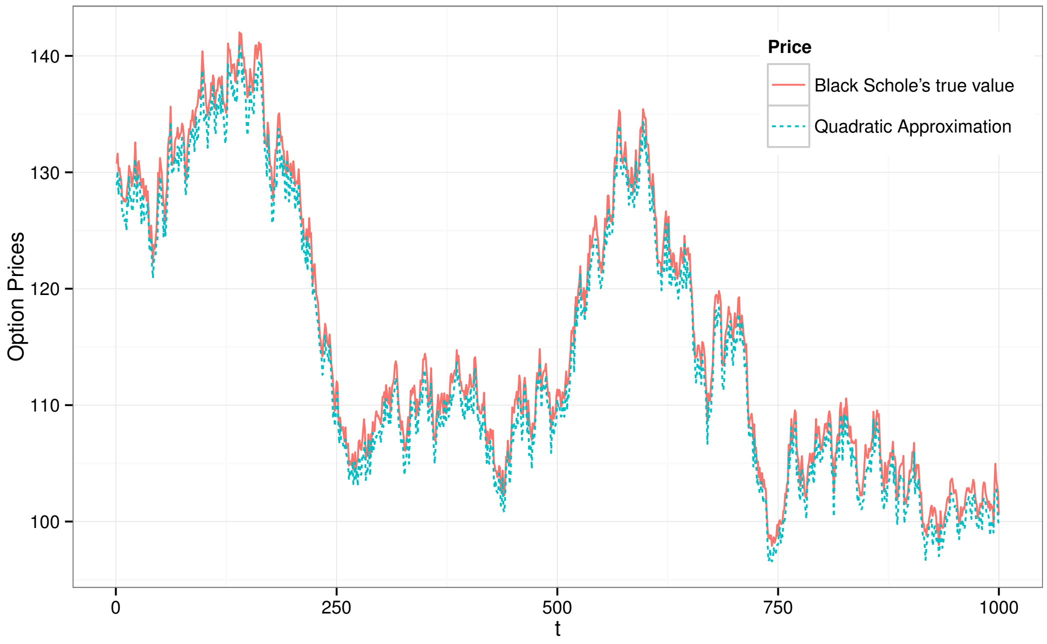

Figure B.3: True simulated option values for the at-the-money options vs the approximated option prices with the same strike prices and time-to-maturity. The approximated values was estimated by using the algorithm discussed in Section 4.3.2.

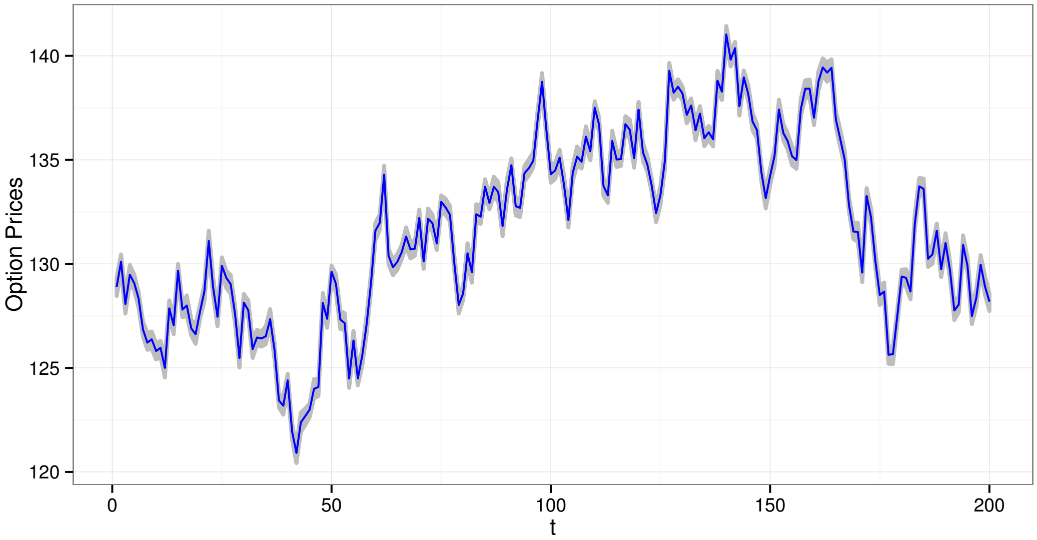

Figure B.4: The approximated call option prices were found by using the algorithm discussed in Section 4.3.2. The shaded area are the optimal bounds of option prices which were constructed by analysing the convexity of nearby option prices, according to Bertsimas and Popescu [24].

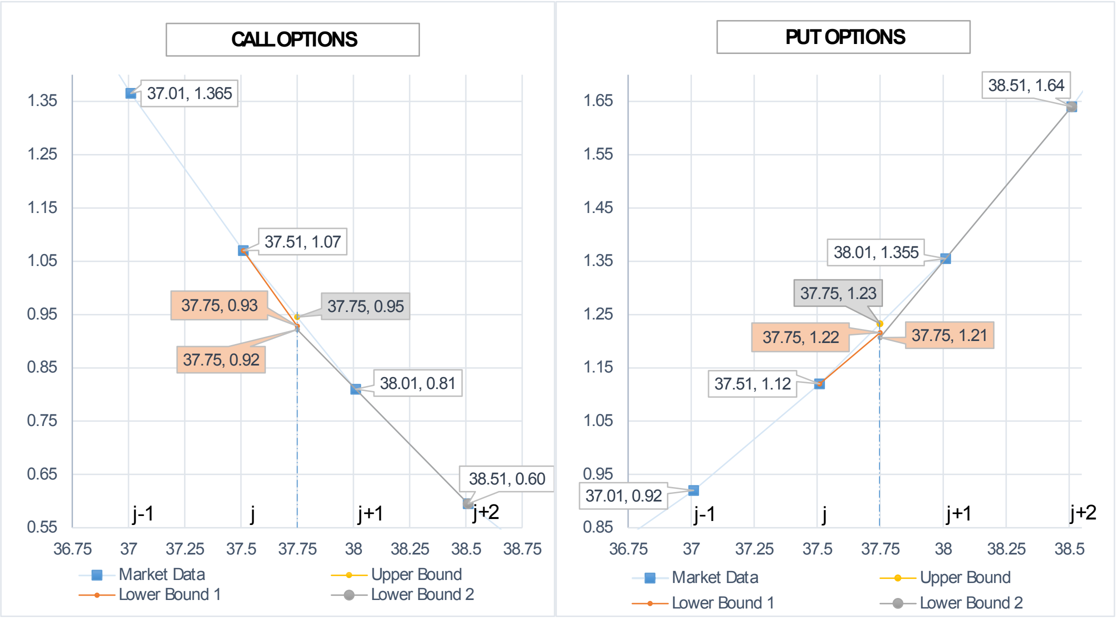

Figure B.5: Bounds on call/put options by using the convexity and monotonicity of the available option prices according to Dimitris and Popescu [24]. The underlying asset was BHP.AX as at 8th Feb, 2012 and the selected options are the BHP March12 options with strike K = [37.01, 37.51, 38.51, 38.51] as shown.

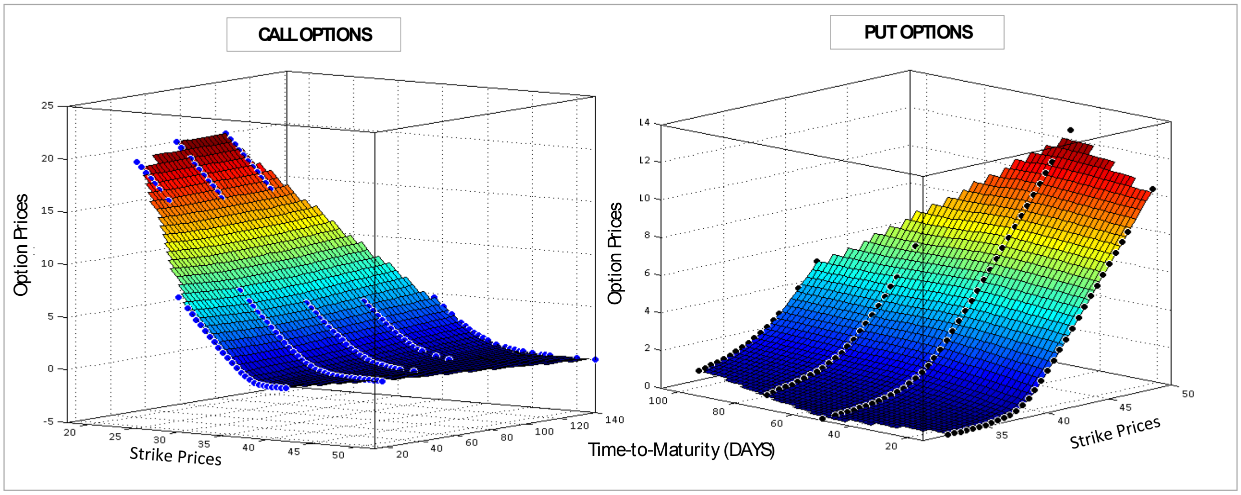

Figure B.6: Option price surface with different strike prices and time-to-maturity (days) of the call options and put options recorded for BHP.AX as at 8^{th} Feburary 2012 (last prices of the trading days). This figure showed that there is a nonlinear relationship between the option price and strike price, especially for in-the-money options. This data was obtained from SIRCA.

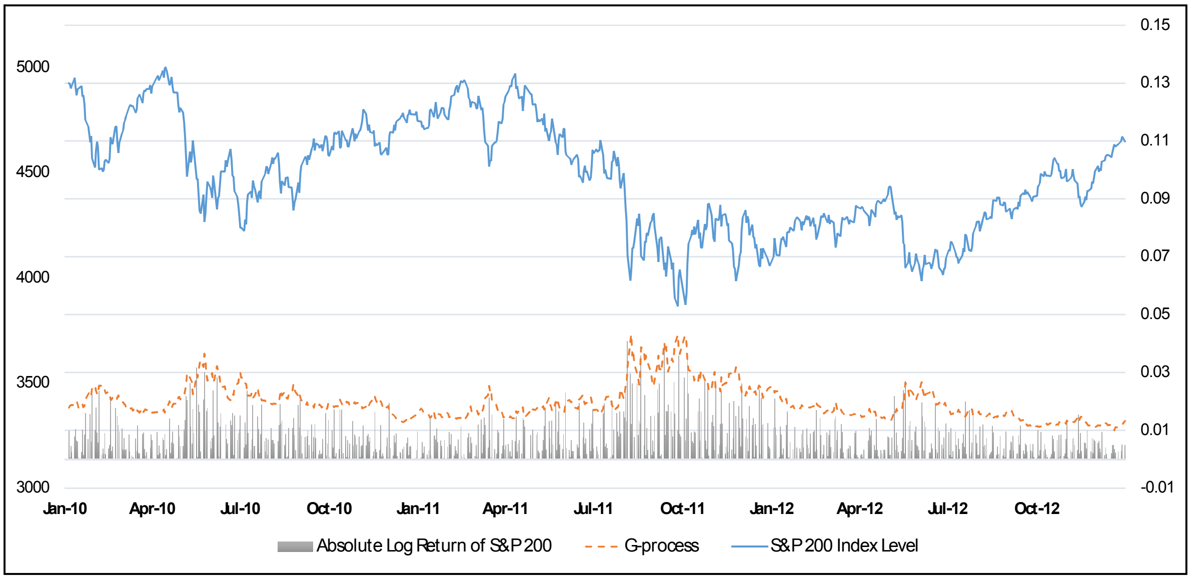

Figure B.7: The daily S&P/ASX 200 index level with its absolute log return and the at-the-money purified option price process G.

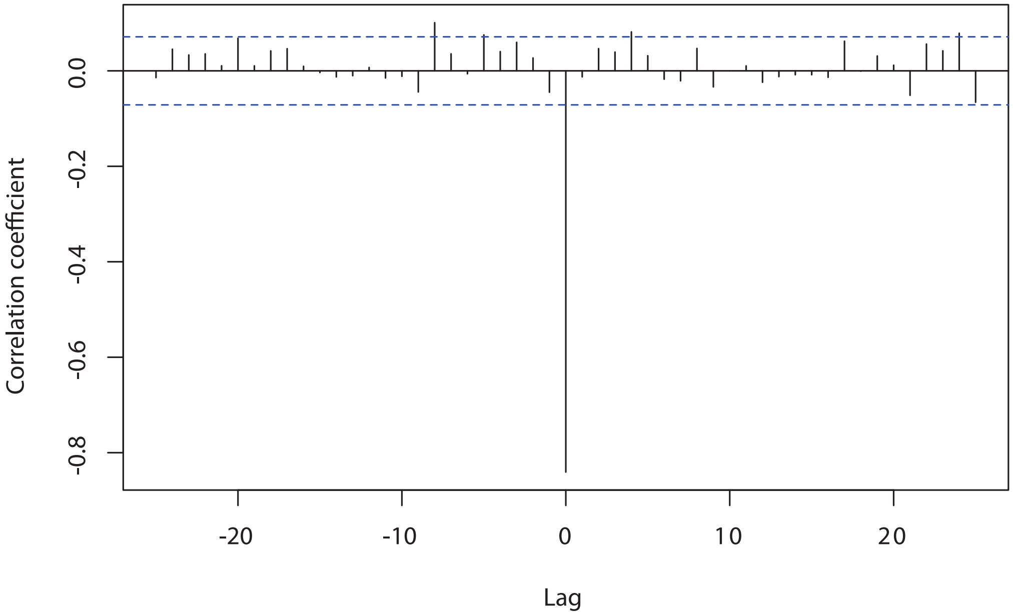

Figure B.8: Cross-correlations between daily returns of the S&P 200 index level and daily changes in G.

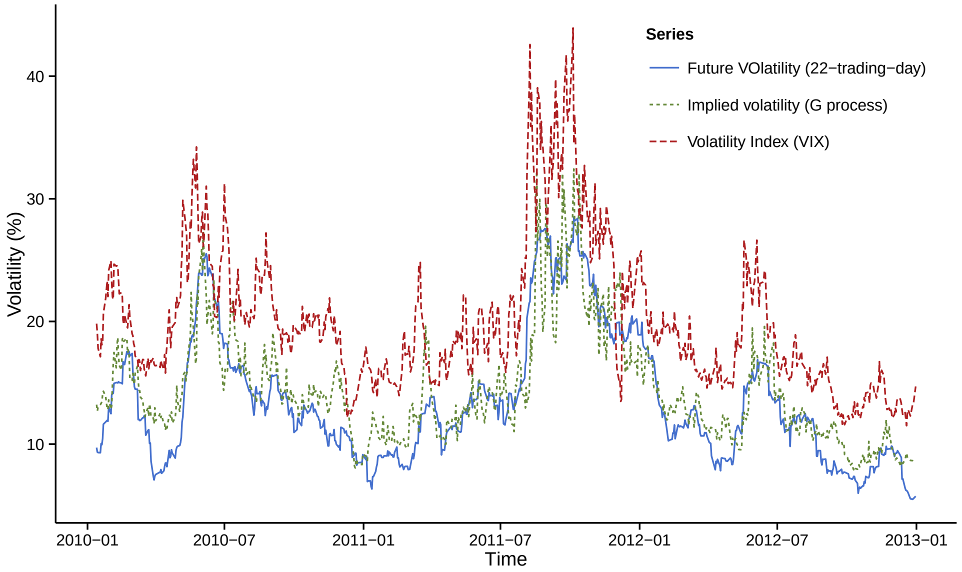

Figure B.9: Future volatility, volatility index and implied volatility of the purified at-the-money option price process for S&P/ASX 200 from 01/01/2010 to 31/12/2012.

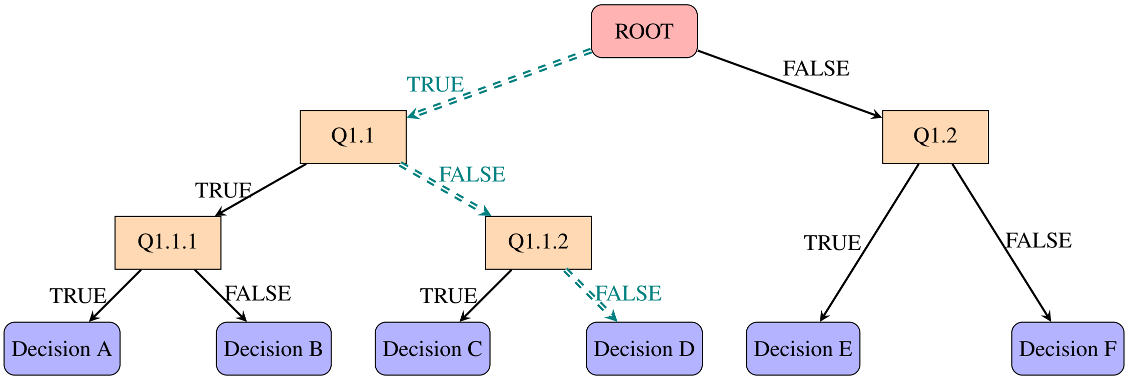

Figure B.10: A binary tree - starts from the root node, multiple criteria are selected based on the information from each input. A decision is drawn at a particular leaf, i.e. Decision D, if all criteria along its path “==" are satisfied .

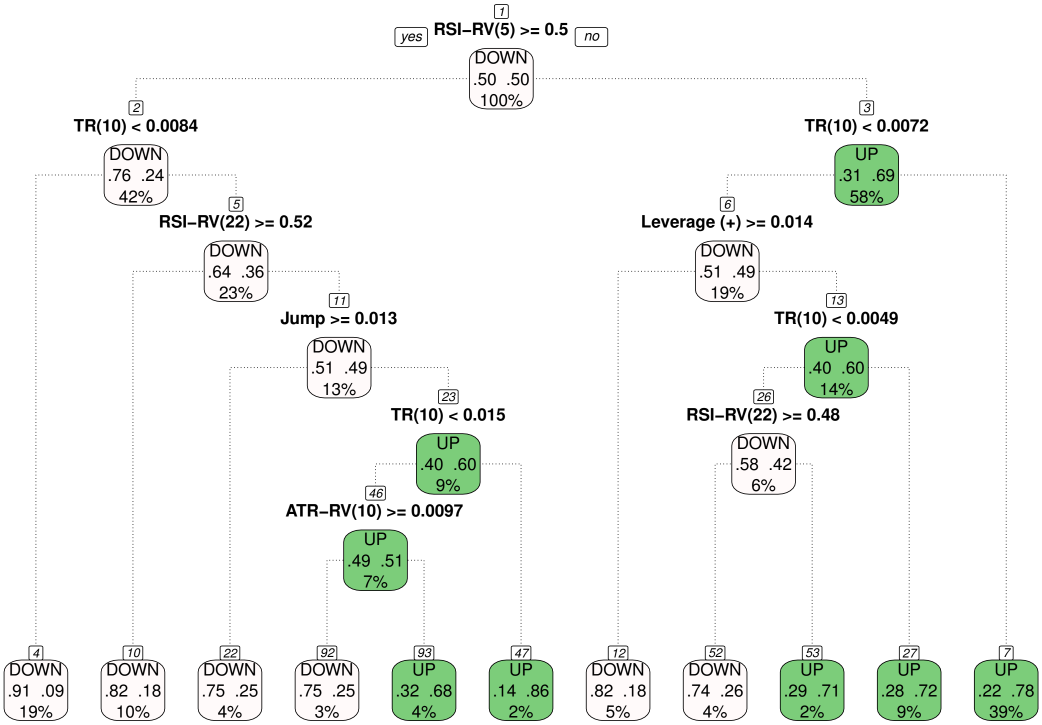

Figure B.11: A possible decision tree for classifying the daily realised volatility direction using the technical indicators from the previous day.

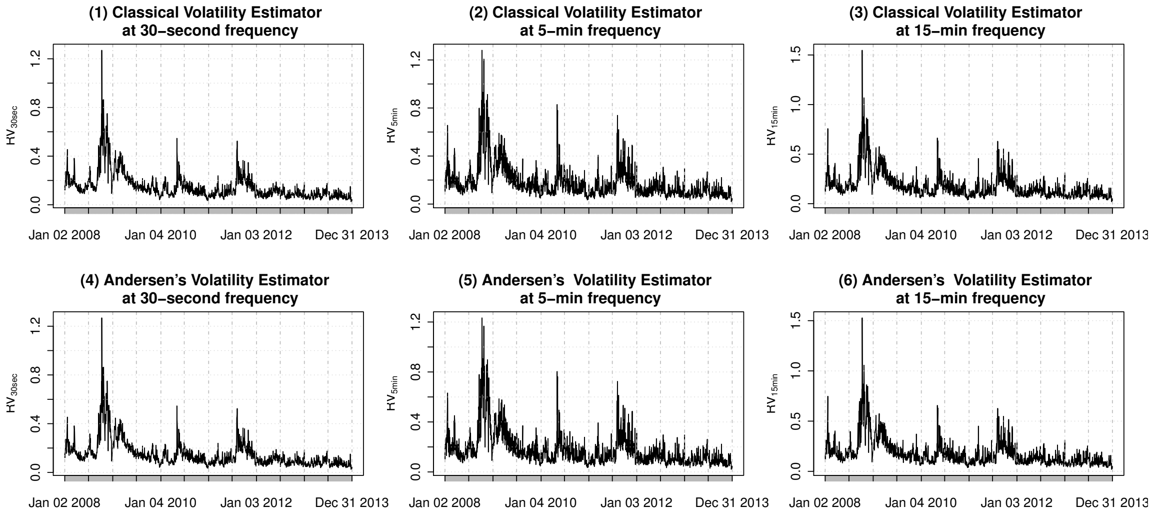

Figure B.12: Annualised daily volatility estimates with classical estimator vs Andersen's realised volatility estimates for SP500 from 2008-2013.

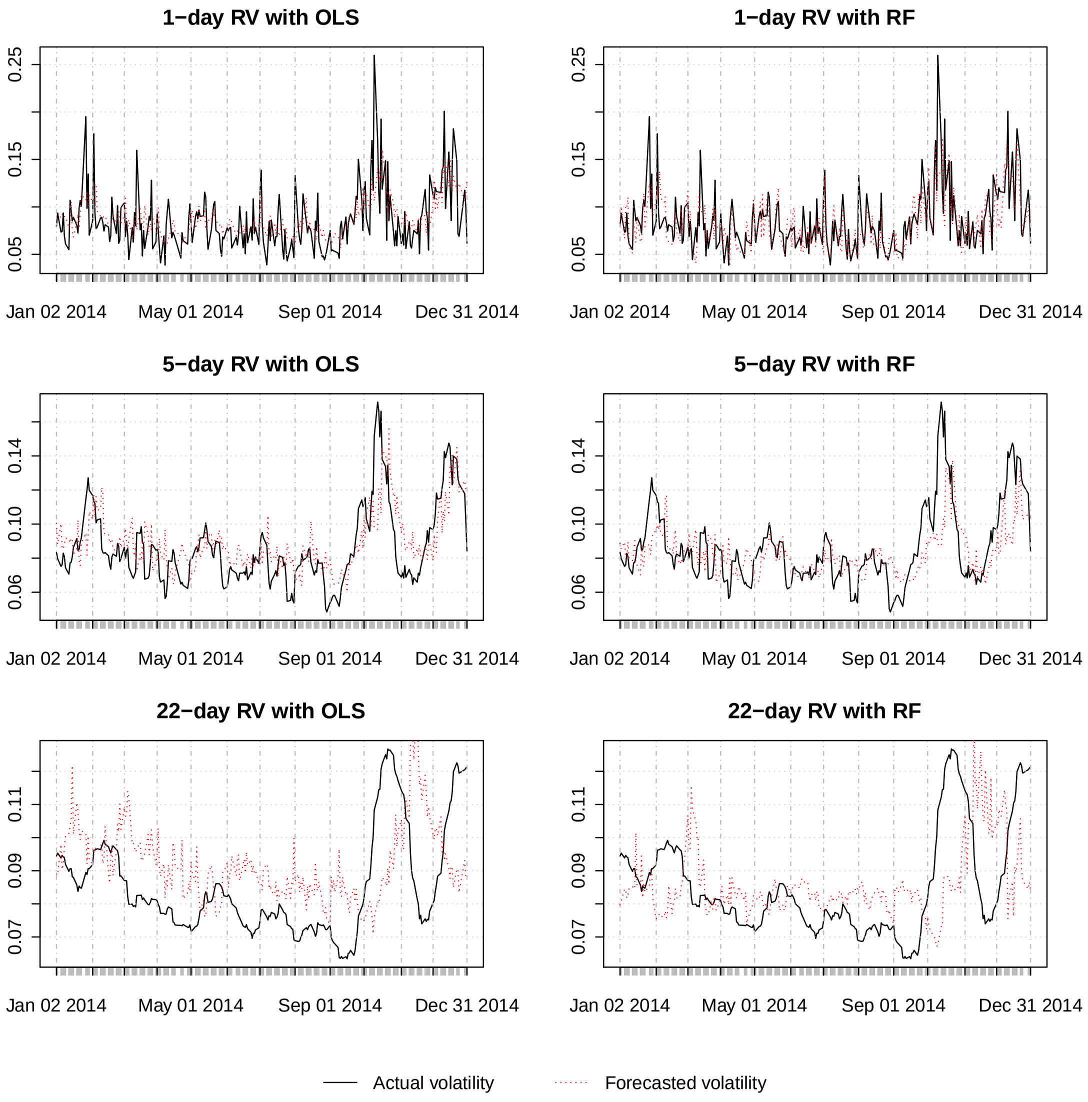

Figure B.13: Predicted vs Actual realised volatility using HAR-JL-PV-D model with maximum likelihood estimation and random forests estimation.

Bibliography

[1] Ait-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. Review of Financial studies, 18(2), 351-416.

[2] Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. The Journal of Finance, 57(3), 1047-1091.

[3] Allen, M. B., & Isaacson, E. L. (2011). Numerical analysis for applied science (Vol. 35). John Wiley & Sons.

[4] Andersen, T. G., & Benzoni, L. (2008, July). Realized volatility. FRB of Chicago Working Paper No. 2008-14. http://dx.doi.org/10.2139/ssrn.1092203

[5] Andersen, T. G., & Bollerslev, T. (1997). Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns. The Journal of Finance, 52(3), 975-1005.

[6] Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 885-905.

[7] Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. The review of Economics and Statistics, 89(4), 701-720.

[8] Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. Handbook of Economic Forecasting, 1, 777-878.

[9] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579-625.

[10] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579-625.

[11] Andersen, T. G., Dobrev, D., & Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics, 169(1), 75-93.

[12] Andersen, T. G., Dobrev, D., & Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics, 169, 1, 75-93.

[13] Andersen, T.G., & Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. International Economic Review, 39, 885-905.

[14] Arriojas, M., Hu, Y., Mohammed, S. E., & Pap, G. (2007). A delayed Black and Scholes formula. Stochastic Analysis and Applications, 25(2), 471-492.

[15] Aruchunan, E., Muthuvalu, M. S., & Sulaiman, J. (2015). Quarter-sweep iteration concept on conjugate gradient normal residual method via second order quadrature-finite difference schemes for solving fredholm integro-differential equations. Sains Malaysiana, 44(1), 139-146.

[16] Bandi F, & Russell, JR (2005) Realized covariation, realized beta, and microstructure noise. University of Chicago.

[17] Bandi, F. M., & Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. The Review of Economic Studies, 75(2), 339-369.

[18] Barndorff-Nielsen, O. E., & Shephard, N. (2004). Measuring the impact of jumps in multivariate price processes using bipower covariation. Discussion paper, Nuffield College, Oxford University.

[19] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: trades and quotes. The Econometrics Journal, 12(3), C1-C32.

[20] Barndorff-Nielsen, O. E., & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. Journal of Financial Econometrics, 2(1), 1-37.

[21] Barndorff-Nielsen, O.E., Hanse, P.R., Lunde, A, & Shephard, N. (2008). Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. Econometrica, 76, 6, 1481–1536.

[22] Barndorff-Nielsen, O.E.& N. Shephard.(2001). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society, Series B, 64.

[23] Bauer, R. (2012). Fast calibration in the Heston model.

[24] Bertsimas, D., & Popescu, I. (2002). On the relation between option and stock prices: a convex optimization approach. Operations Research, 50(2), 358-374.

[25] Black, F. (1976). Studies of stock price volatility changes, proceedings of the 1976 meetings of the business and economic statistics section. 177-191. In American Statistical Association.

[26] Black, F. (1976). The pricing of commodity contracts. Journal of Financial Economics 3, 167–179.

[27] Black, F., & Scholes, M. (1972). The valuation of option contracts and a test of market efficiency. Journal of Finance, 27, 399-418.

[28] Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. The Journal of Political Economy, 637-654.

[29] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

[30] Boudt, K., & Zhang, J. (2015). Jump robust two time scale covariance estimation and realized volatility budgets. Quantitative Finance, 15, 6, 1041-1054.

[31] Brandt, M.W., & Diebold, F.X. (2006). A no-arbitrage approach to range-based estimation of return covariances and correlations. Journal of Business, 79, 61-73.

[32] Breiman, L.(2001). Random Forests Machine Learning, 45(1), 5-32.

[33] Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. (1984). Classification and Regression Trees. Wadsworth and Brooks.

[34] Brenner, M., & Subrahmanyan, M. G. (1988). A simple formula to compute the implied standard deviation. Financial Analysts Journal, 44(5), 80-83.

[35] Campbell, J. Y., Lo, A. W. C., & MacKinlay, A. C. (1997). The econometrics of financial markets, 2, 149-180. Princeton, NJ: Princeton University Press.

[36] Canina, L., & Figlewski, S. (1993). The informational content of implied volatility. Review of Financial studies, 6(3), 659-681.

[37] Carr, P., & Wu, L. (2006): A tale of two indexes. Journal of Derivatives, 13, 13–29.

[38] Carr, P., & Lee, R. (2003). At-the-money implied as a robust approximation of the volatility swap rate. In Bloomberg LP Working paper.

[39] Bin, C. (2007). Calibration of the heston model with application in derivative pricing and hedging. Master's thesis, Department of Mathematics, Technical University of Delft, Delft, The Netherlands.

[40] Christensen, B. J., & Prabhala, N. R. (1998). The relation between implied and realized volatility. Journal of Financial Economics, 50(2), 125-150.

[41] Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American Statistical Association, 44(245), 32-61.

[42] Corliss, G. (1977). Which root does the bisection algorithm find?. Siam Review, 19(2), 325-327.

[43] Corsi, F. (2003). A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7, 2, 174-196.

[44] Corsi, F., & Reno, R. (2009). HAR volatility modelling with heterogeneous leverage and jumps. Available at SSRN 1316953.

[45] Dokuchaev, N. (2006). Two unconditionally implied parameters and volatility smiles and skews. Applied Financial Economics Letters, 2, 199-204.

[46] Dokuchaev, N. (2007). Mean-reverting market model: speculative opportunities and non-arbitrage. Applied Mathematical Finance, 14(4), 319-337.

[47] Dokuchaev, N. (2014). Volatility estimation from short time series of stock prices. Journal of Nonparametric Statistics, 26(2), 373-384.

[48] Duffie, D. (2010). Dynamic asset pricing theory. Princeton University Press.

[49] Durrett, R. (2010). Probability: theory and examples. Cambridge University Press.

[50] Elerian, O., Chib, S., & Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. Econometrica, 69(4), 959-993.

[51] Engel, R. F. (1990). Discussion: stock market volatility and the crash. Review of Financial Studies, 3, 103-106.

[52] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society, 987-1007.

[54] Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. Journal of Business & Economic Statistics, 19(2), 177-191.

[54] Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. Journal of Business & Economic Statistics, 19(2), 177-191.

[55] Evans, D. J. (1987). The alternating group explicit (AGE) matrix iterative method. Applied Mathematical Modelling, 11(4), 256-263.

[56] Evans, D. J., & Yousif, W. S. (1988). The modified alternating group explicit (MAGE) method. Applied Mathematical Modelling, 12(3), 262-267.

[57] Frijns, B., Tallau, C., & Tourani-Rad, A. (2010). The information content of implied volatility: evidence from Australia. Journal of Futures Markets, 30(2), 134-155.

[58] Gallant, A.R., Hsu, C., Tauchen, G.E. (1999). Using daily range data to calibrate volatility diffusions and extract the forward integrated variance. Review of Economics and Statistics, 81, 617-631.

[59] Galligani, I., & Ruggiero, V. (1990). The arithmetic mean method for solving essentially positive systems on a vector computer. International Journal of Computer Mathematics, 32(1-2), 113-121.

[60] German, M., & Klass, M. (1980). On the Estimation of security Price volatility from historical data. Journal of Business, 53, 67-69.

[61] Giot, P. (2005). Implied volatility indexes and daily Value at Risk models. The Journal of derivatives, 12(4), 54-64.

[62] Girsanov, I. V. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. Theory of Probability & Its Applications, 5(3), 285-301.

[63] Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance, 48(5), 1779-1801.

[64] Hadjidimos, A. (2000). Successive overrelaxation (SOR) and related methods. Journal of Computational and Applied Mathematics, 123(1), 177-199.

[65] Hadjidimos, A. (2000). Successive overrelaxation (SOR) and related methods. Journal of Computational and Applied Mathematics, 123(1), 177-199.

[66] Hansen, P., & Lunde, A. S. G. E. R. (2010). Forecasting volatility using high frequency data. Oxford Handbook of Economic Forecasting', Oxford University Press.

[67] Harri, A., & Brorsen, S.W. (2009). The overlapping data problem. Quantitative and Qualitative Analysis in Social Sciences, 3, 3, 78-115.

[68] He, S. W., & Yan, J. A. (1992). Semimartingale theory and stochastic calculus. Taylor & Francis.

[70] Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. Review of Financial Studies, 6(2), 327-343.

[70] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. The Review of Financial Studies, 6, 327-343.

[71] Hin, L. Y., & Dokuchaev, N. (2014). On the implied volatility layers under the future risk-free rate uncertainty. International Journal of Financial Markets and Derivatives, 3(4), 392-408.

[72] Huang, X., & Tauchen, G. (2005). The relative contribution of jumps to total price variance. Journal of Financial Econometrics, 3(4), 456-499.

[73] Hudson, R. S., & Gregoriou, A. (2015). Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. International Review of Financial Analysis, 38, 151-162.

[74] Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. The journal of finance, 42(2), 281-300.

[75] Hull, J., & White. A. (1987). The pricing of options on assets with stochastic volatilities. Journal of Finance, 42, 281-300.

[76] Itô, K. (1944). 109. Stochastic Integral. Proceedings of the Imperial Academy, 20(8), 519-524.

[77] Ivanov, A. F., Kazmerchuk, Y. I., & Swishchuk, A. V. (2003). Theory, stochastic stability and applications of stochastic delay differential equations: a survey of results. Differential Equations Dynam. Systems, 11(1-2), 55-115.

[78] Jacquier, E., Polson, N. G., & Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics, 20(1), 69-87.

[79] Jiang, G. J., & Tian, Y. S. (2005). The model-free implied volatility and its information content. Review of Financial Studies, 18(4), 1305-1342.

[80] Johnson, H., & Shanno, D. (1987). Option pricing when the variance is changing. Journal of Financial and Quantitative Analysis, 22(02), 143-151.

[81] Kastner, G. (2015). Dealing with stochastic volatility in time series using the R package stochvol. Journal of Statistical Software. http://cran.r-project.org/web/packages/stochvol/vignettes/article.pdf.

[82] Khan, M. A. I. (2011). Financial volatility forecasting by nonlinear support vector machine heterogeneous autoregressive model: evidence from Nikkei 225 stock index. International Journal of Economics and Finance, 3(4), p138.

[83] Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. The Review of Economic Studies, 65(3), 361-393.

[84] Klebaner, F.C. (2005). Introduction to Stochastic Calculus With Applications. Second Edition, London, Imperial College Press.

[85] Koksal, B. (2009). A comparison of conditional volatility estimators for the ISE National 100 Index Returns. Journal of Economic and Social Research, 11(2), 1-28.

[86] Kovachev, Y. (2014). Calibration of stochastic volatility models.

[87] Kress, R., Numerical Analysis, New York: Springer, 1998.

[88] Liptser, R., & Shiryaev, A. N. (2013). Statistics of random Processes: I. general Theory (Vol. 5). Springer Science & Business Media.

[89] Luong, C., & Dokuchaev, N.(2014). Analysis of market volatility via a dynamically purified option price process. Annals of Financial Economics Vol. 09, No. 03, 1450006 (2014)

[90] Lux, T., & Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market. Nature, 397(6719), 498-500.

[91] Masset, P. (2011). Volatility stylized facts. Available at SSRN 1804070.

[92] Mao, X., & Shah, A. (1997). Exponential stability of stochastic differential delay equations. Stochastics: An International Journal of Probability and Stochastic Processes, 60(1-2), 135-153.

[93] Mao, X., Koroleva, N., & Rodkina, A. (1998). Robust stability of uncertain stochastic differential delay equations. Systems & Control Letters, 35(5), 325-336.

[94] Meucci, A. (2010). Quant nugget 2: Linear vs. compounded returns–common pitfalls in portfolio management. GARP Risk Professional, 49-51.

[95] Mikhailov, S., & Nogel, U. (2004). Heston's stochastic volatility model: Implementation, calibration and some extensions. John Wiley and Sons.

[96] Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V., & von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components. Journal of Empirical Finance, 4(2), 213-239.

[97] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. econometrica: Journal of the Econometric Society, 347-370.

[98] Novikov, A. A. (1980). On conditions for uniform integrability of continuous nonnegative martingales. Theory of Probability & Its Applications, 24(4), 820-824.

[99] Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. Journal of Business, 61-65.

[100] Pearson, K. (1896). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the royal society of London, 60(359-367), 489-498.

[101] Peters, E. (1994). Fractal market analysis. A Wiley Finance Edition. John Wiley & Sons, New York.

[102] Phillips, P. C. (1986). Understanding spurious regressions in econometrics. Journal of Econometrics, 33(3), 311-340.

[103] Protter, P. (1990). Stochastic integration and differential equations: a New Approach. Springer.

[104] Qin, Q., Wang, Q. G., Li, J., & Ge, S. S. (2013). Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market. Journal of Intelligent Learning Systems and Applications, 5(1), 1.

[105] Reserve Bank of Australia. (2013):Interest rates and yields - Money market - daily - 1976 to 2013 - F1. Retrieved from http://www.rba.gov.au/statistics/tables/

[106] Roberts, G. O., & Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. Biometrika, 88(3), 603-621.

[107] S&P Dow Jones Indices. (2014): S&P/ASX 200 VIX Methodology. Retrieved from http://au.spindices.com/

[108] Sahimi, M. S., Ahmad, A., & Bakar, A. A. (1993). The iterative alternating decomposition explicit (IADE) method to solve the heat conduction equation. International Journal of Computer Mathematics, 47(3-4), 219-229.

[109] Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. The Journal of Finance, 52(3), 923-973.

[110] Shephard, N. (2008). Stochastic volatility. New Palgrave Dictionary of Economics (2nd ed.)

[111] Shephard, N., & Sheppard, K. (2010). Realising the future: forecasting with high frequency based volatility (HEAVY) models. Journal of Applied Econometrics, 25(2), 197-231.

[112] SIRCA. (2013-2015). Thomson Reuters Tick History. Retrieved from http://www.sirca.org.au/.

[113] Slepaczuk, R., & Zakrzewski, G. (2009). High-frequency and model-free volatility estimators. Available at SSRN 2508648.

[114] Stoica, G. (2005). A stochastic delay financial model. Proceedings of the American Mathematical Society, 133(6), 1837-1841.

[115] Sugiyama, S. (1969). On the stability problems on difference equations, Bull. Sci. Eng. Research Lab. Waseda Univ. 45, 140-144.

[116] Sewell, M. V. (2010). The application of intelligent systems to financial time series analysis (Doctoral dissertation, PhD thesis, PhD dissertation, Department of Computer Science, University College London, University of London).

[117] Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes - a study of daily sugar prices 1961-79. In O. D. Anderson (Ed.), Time Series Analysis: Theory and Practice, 1, 203-226. Amsterdam: North-Holland.

[118] Theofilatos, K., Likothanassis, S., & Karathanasopoulos, A. (2012). Modeling and Trading the EUR/USD Exchange Rate Using Machine Learning Techniques. Engineering, Technology & Applied Science Research, 2(5), 269.

[119] Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the Brownian motion. Physical review, 36(5), 823.

[120] Vasicek, O. (1977). An equilibrium characterization of the term structure. Journal of Financial Economics, 5(2), 177-188.

[121] Whaley, R. E. (2000). The investor fear gauge. The Journal of Portfolio Management, 26(3), 12-17.

[122] Yan, B., & Zivot, E. (2003). Analysis of high-frequency financial data with S-PLUS. UWEC-2005-03.

[123] Zakoian, J. M. (1994). Threshold heteroskedastic models. Journal of Economic Dynamics and control, 18(5), 931-955.

[124] Zhang, L. (2011). Estimating covariation: epps effect, microstructure noise. Journal of Econometrics, 160, 1, 33-47.

[125] Zhang, L., Mykland, P. A., & Ait-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association, 100, 472, 1394-1441.

[126] Zhang, L., Mykland, P. A., & Ait-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association, 100, 472, 1394-1441.