

Washington University in St. Louis
Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations

Arts & Sciences


Spring 5-15-2016

Spot Volatility Estimation of Ito Semimartingales Using Delta Sequences

Weixuan Gao

Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

 Part of the [Applied Statistics Commons](#), [Finance and Financial Management Commons](#), [Other Statistics and Probability Commons](#), and the [Statistical Models Commons](#)

Recommended Citation

Gao, Weixuan, "Spot Volatility Estimation of Ito Semimartingales Using Delta Sequences" (2016). *Arts & Sciences Electronic Theses and Dissertations*. 701.

https://openscholarship.wustl.edu/art_sci_etds/701

This Thesis is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics

Spot Volatility Estimation of Itô Semimartingales Using Delta Sequences

by

Weixuan Gao

A thesis presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the
degree of Master of Arts

May 2016
Saint Louis, Missouri

copyright by

Weixuan Gao

2016

Contents

- List of Tables iii
- List of Figures iv
- Acknowledgments v
- Abstract vii
- 1 Introduction 1**
 - 1.1 Overview of Model 1
 - 1.2 Microstructure Noise and Jumps 2
 - 1.3 Background Knowledge 3
 - 1.3.1 Brief Introduction to Stochastic Processes 3
 - 1.3.2 Stopping Time 7
 - 1.3.3 Itô Integral 7
 - 1.4 Euler Method 10
- 2 Estimation Method 11**
 - 2.1 The Model 11
 - 2.2 Assumptions 13
 - 2.3 Asymptotic Theory 19
 - 2.4 Robustness to Microstructure Noise Effects 22
 - 2.5 Robustness to Jumps 24
- 3 Simulation Experiments 26**
- 4 Application using real market data 29**
 - 4.1 Introduction to TAQ Database 29
 - 4.2 Data Preprocessing 31
 - 4.3 Estimation 33
 - 4.4 Analysis 34
- 5 Conclusion 36**
- References 38**

List of Tables

3.1	Simulated MSE for six delta sequences	28
4.1	Quote data from TAQ database	30
4.2	Trade data from TAQ database	30

List of Figures

3.1	The Frequency of Relative Error for Six Delta Sequences	28
4.1	AAPL stock price on July 1, 2014	31
4.2	Estimated spot volatility for AAPL averaged 128 trading dates using threshold estimator (2.7) with double exponential kernel	32
4.3	Estimated spot volatility for AAPL averaged 128 trading dates using threshold estimator (2.35) with double exponential kernel	34
4.4	Estimated spot volatility for AAPL in the presence of microstructure noise after exclude 16 days which is detected to have jumps by using $C - Tz$ test	35

Acknowledgments

I would first like to thank my thesis advisors Professor José E. Figueroa-López and Professor Renato Feres. The doors to their offices were always open whenever I ran into a trouble spot or had questions about my research and writing. Their valuable time and insights were indispensable to this thesis. I would also like to thank my academic advisor, Professor Todd Kuffner, who taught my first class after I came to America. I can still remember his impressive class. Besides, he helps me a lot both in my academic and personal lives.

Finally, I must thank my parents for their continuous support during my years of study and the process of researching and writing this thesis. This thesis would have not been possible without them.

Weixuan Gao

Washington University in Saint Louis
May 2016

Dedicated to my parents.

ABSTRACT OF THE

Spot Volatility Estimation of Itô Semimartingales Using Delta Sequences

by

Weixuan Gao

Master of Arts in Statistics

Washington University in St. Louis, 2016

Professor José E. Figueroa-López, Chair

Abstract: This thesis studies a unifying class of nonparametric spot volatility estimators proposed by Mancini et. al.(2013). This method is based on delta sequences and is conceived to include many of the existing estimators in the field as special cases. The thesis first surveys the asymptotic theory of the proposed estimators under an infill asymptotic scheme and fixed time horizon, when the state variable follows a Brownian semimartingale. Then, some extensions to include jumps and financial microstructure noise in the observed price process are also presented. The main goal of the thesis is to assess the suitability of the proposed methods with both high-frequency simulated data and real transaction data from the stock market. In conclusion, double exponential kernel shows the best properties when estimating. Besides, the theorem is robust with the presence of jumps and microstructure noise and the U-shape curves of intraday spot volatility are achieved.

Keyword: Spot volatility, High-frequency estimation, Delta sequences, Microstructure noise, Kernel estimator

Chapter 1

Introduction

1.1 Overview of Model

During the past several decades, continuous-time models for the dynamics of asset returns have been widely used in financial and economical literature. Volatility is one the main components in these models. As a main measure of risk in market, volatility plays a primary role in the asset pricing formulas of several derivatives. It is therefore not surprising that understanding volatilities and their dynamics has attracted much attention. With the development of technology and information, larger amount of financial data sets is gradually not hard to achieve any longer and, as a result, has increasingly been widely used in plenty of econometric research. The availability of high-frequency intraday data of price returns has provided new opportunities to empirically study the volatility process. Particularly, the realized volatility measure, which provides an estimate of the integrated volatility, has drawn much attention. Many other measures have been also introduced, such as range-based volatility and power volatility (see e.g. Alizadeh et al. [1]). However, in many applications, the recovery of the spot (or instantaneous) volatility is desirable, and interest has gradually been moving to the study of spot volatility during the last few years. For example, one

can achieve efficient estimation of the integrated volatility by using the integration of the preliminary spot volatility (see Jacod and Rosenbaum [17]).

Currently there are a number of methods to estimate spot volatility. Kernels estimation (Kristensen [21]) is one of the most popular methods, while Fourier estimator of spot volatility introduced by Malliavin and Mancino [23] is of kernel type. More recent literature has also explored the volatility estimation in the presence of jumps and microstructure noise (e.g., Johannes [20]), Hoffmann et al [14], Ogawa and Sanfelici [13], Ait-Sahalia [32], Mykland and Zhang [3] and many more).

This thesis studies a unifying class of estimators for the spot volatility of a univariate semimartingale proposed by Mancini et al [25], which, with appropriate adjustment, also allow jumps and microstructure noise in the price process.

1.2 Microstructure Noise and Jumps

In contrast to low frequency (daily, weekly, or longer) financial datasets, high-frequency datasets are identified by a large amount of intraday observations and contain so-called market microstructure noise. Since high-frequency estimation heavily depends on an accurate description of the stock price dynamics during a very short time, high-frequency data can capture a variety of friction inherent in the trading process.

The computation of the realized volatility at first, which is a sum of squared intraday return, is done under the premise that the prices are observed continuously and without measurement error (e.g. Merton [26]). Unfortunately, the realized volatility suffers from a well-known bias

problem, that becomes worse as the frequency of data increases. The source of this bias problem is attributed to the the so-called market microstructure noise. Thus, there is a trade-off between bias and variance, when choosing the sampling frequency. Microstructure noise has many sources. For example, bid-ask bounce effect, order arrival latency, asymmetry of information, and discreteness of price changes.

Besides microstructure noise, discontinuities or “jumps” are also believed to be a primary component of financial asset prices. We should note that large jumps do not basically come to markets regularly, but their arrivals tend to depend on market information. For example, the advent of unanticipated news may have a great influence on the valuation of certain financial assets. Since incorporating jumps related to market information can greatly influence the accuracy of our model, identifying jumps is essential.

1.3 Background Knowledge

1.3.1 Brief Introduction to Stochastic Processes

A stochastic process is a collection of variable evolving with time, representing a process of evaluation. More precisely, a stochastic process is a collection of random variables $\{X_t\}$ indexed by time t . For a discrete time process, t takes values on an increasing countable sequence. In contrast, for a continuous time process, t takes values in $[0, \infty)$. For each $t \in T$ fixed, we have a random variable $\omega \rightarrow X_t(\omega)$, for $\omega \in \Omega$. On the other hand, if we fix $\omega \in \Omega$, the function

$$t \rightarrow X_t(\omega); \quad t \in T$$

can be regarded as the path of X_t . Sometimes we can also consider the process as

$$(t, \omega) \rightarrow X(t, \omega)$$

which is a function of two variables mapping from $T \times \Omega$ into \mathbf{R}^n . $X(t, \omega)$ is jointly measurable in (t, ω) .

We now introduce the most important class of continuous-time processes.

Definition 1.1 $\{W_t\}_{t \geq 0}$ is called a *Brownian processes or Wiener processes*, if the following conditions are satisfied.

(1) Every increment $W(t) - W(s)$ over an interval of length $t-s$ is normally distributed with mean 0 and variance $t - s$, that is

$$W(t) - W(s) \sim N(0, t - s)$$

(2) For every pair of disjoint of time intervals $[t_1, t_2]$ and $[t_3, t_4]$, with $t_1 \leq t_2 \leq t_3 \leq t_4$, the increments $W(t_4) - W(t_3)$ and $W(t_2) - W(t_1)$ are independent.

(3) $W(0) = 0$

(4) $t \rightarrow W(t)$ is continuous for all $t \in [0, \infty)$.

Note that Property 2 tells us that, the value of $W(s)$ gives no further knowledge of $W(t) - W(s)$ with $t > s$, Formally, if $0 \leq t_0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq t$, then $P(X_t = y | X_{t_0} = x_0, X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n) = P(X_t = y | X_{t_n} = x_n)$ which is known as the Markov property of Brownian Motion.

Brownian motion is continuous but nowhere differentiable. Fix $x \in \mathbf{R}^n$, for $y \in \mathbf{R}^n$, the function

$$p(t, x, y) = (2\pi t)^{-1/2} \exp\left(-\frac{|x - y|^2}{2t}\right) \quad (1.1)$$

is the (transition) density of W_{t+s} given that $W_s = x$, where $s, t \geq 0$. We now introduce a fundamental class of discrete time processes.

Definition 1.2 A Poisson process N is an integer-valued process such that

(1) $N_0 = 0$;

(2) Independent increments: $X_{t_1} - X_{t_0}, \dots, X_{t_n} - X_{t_{n-1}}$ are independent for any $0 \leq t_0 \leq \dots \leq t_n$;

(3) $N_t - N_s$ has Poisson distribution with parameter $\lambda(t - s)$, for any $s < t$;

(4) Its paths are càdlàg, that is, right-continuous with left-limits.

The parameter λ is called the intensity of the process.

Definition 1.3

(1) Let $0 < \gamma \leq 1$. A function $f: [0, T] \rightarrow \mathbf{R}$ is called uniformly Hölder continuous with exponent $\gamma > 0$, if there exists a constant K , for all $s, t \in [0, T]$, such that

$$|f(t) - f(s)| \leq K|t - s|^\gamma, \quad (1.2)$$

(2) f is Hölder continuous with exponent $\gamma > 0$ if for all $s, t \in [0, T]$, f satisfies condition (1.2) above, for some constant K .

The following results gives conditions for a process to admit a continuous version:

Theorem 1.1 (Kolmogorov's continuity theorem) Suppose that the process $X = \{X_t\}_{t \geq 0}$ satisfies the following condition: For all $T \geq 0$ there exist positive constants α, β, D , such that

$$E\{|X_t - X_s|^\beta\} \leq D|t - s|^{1+\alpha}; 0 \leq s, t \leq T.$$

Then there exists a continuous version of X .

Take Brownian motion as an example. If $W(\cdot)$ is an n -dimensional Brownian motion. For all integers $m = 1, 2, \dots$, we have

$$E(|W(t) - W(s)|^{2m}) = C|t - s|^m,$$

thus the hypotheses of Kolmogorov's theorem hold for $\beta = 2m, \alpha = m - 1$. The process $W(\cdot)$ is thus Hölder continuous for exponents such that $0 < \gamma < \frac{1}{2}$.

1.3.2 Stopping Time

A stopping time τ and its corresponding σ -algebra \mathcal{F}_τ can be defined both for discrete time process and continuous process.

Definition 1.4 *Suppose a non-negative (probablyly infinite) random variable τ satisfies*

$$\{\tau \leq t\} \in \mathcal{F}_t,$$

for every $t \geq 0$, then τ is an $\{\mathcal{F}_t\}$ -stopping time. And we call \mathcal{F}_τ stopped σ -algebra if

$$\mathcal{F}_\tau = \{A \cap \mathcal{F}; A \cap \{\tau \leq t\} \in \mathcal{F}_t, t \geq 0\}.$$

There are some important properties for stopping time that we need to know.

Preposition 1.1 *1. \mathcal{F}_τ is a σ -algebra.*

2. For a second stopping time, $\min\{\tau, \sigma\}$ is \mathcal{F}_τ -measureable.

3. Let $\{\tau_n\}_{n \geq 1}$ is a sequence of stopping time. If \mathcal{F}_t is right continuous, then $\inf_n \tau_n$, $\liminf_{n \rightarrow \infty} \tau_n$, and $\limsup_{n \rightarrow \infty} \tau_n$ are \mathcal{F}_t -stopping times.

1.3.3 Itô Integral

First, consider a mathematical finance case. Let $X(t)$ be the price of an asset at time t . Denote the increment of price during a short period of time $[t, t + \Delta t)$ by $\Delta X = X(t + \Delta t) -$

$X(t)$. Now consider the returns of this asset that is defined as $\Delta X/X$. Then we can model it as

$$\frac{\Delta X}{X} = \text{deterministic contribution} + \text{stochastic contribution.} \quad (1.3)$$

The deterministic contribution is attributed to the interest rate of non-risky activity, and thus is defined to be proportional to time t with a constant rate μ :

$$\text{deterministic contribution} = \mu\Delta t. \quad (1.4)$$

As for the stochastic contribution, it can be assumed to be related to the variation of noise and the variation of market (so-called volatility). Denote the variation of noise by $\Delta W = W(t + \Delta t) - W(t)$ and enable it proportional to the market volatility σ , then we have

$$\text{stochastic contribution} = \sigma\Delta W. \quad (1.5)$$

Intuitively, it is natural to assume the noise follows a Gaussian distribution, that is, $\Delta W \sim N(0, \Delta t)$, which indicates that X is a Brownian motion. Combine the two formulas, and we have

$$\frac{\Delta X}{X} = \mu\Delta t + \sigma\Delta W \quad (1.6)$$

As $\Delta t \rightarrow 0$, it is reasonable we rewrite the formula above in a differential form as

$$dX(t) = \mu X(t)dt + \sigma X(t)dW(t) \quad (1.7)$$

Since we have mentioned that Brownian motion is nowhere differentiable, we switch to its integral form

$$X(t) = X(0) + \mu \int_0^t X(\mu) d\mu + \sigma \int_0^t X(u) dW_u. \quad (1.8)$$

The term $\int_0^t X(u) dW_u$ is known as stochastic integral with respect to the Brownian motion.

Definition 1.5 Let $B_t(\omega)$ be an n -dimensional Brownian motion. Define $\mathcal{F}_t = \mathcal{F}_t^{(n)}$ to be the σ -algebra generated by random variables $\{B_i(s)\}_{1 \leq i \leq n, 0 \leq s \leq t}$. That is, \mathcal{F}_t is the smallest σ -algebra containing all sets of the form

$$\{w; B_{t_1}(\omega) \in F_1, \dots, B_{t_k}(\omega) \in F_k\},$$

where $t_j \leq t$ and $F_j \in \mathbf{R}^n$ are Borel sets, $j = 1, 2, \dots$

\mathcal{F}_t can be understood as the history of B up to time t . Broadly function $h(\omega)$ is \mathcal{F}_t -measurable if it can be written as the pointwise a.s. limit of sums of functions of the form

$$g_1(B_{t_1})g_2(B_{t_2}) \cdots g_k(B_{t_k}),$$

where g_1, \dots, g_k are continuous bounded functions with $t_j \leq t$ for $j \leq k$, and $k = 1, 2, \dots$

That Function h is \mathcal{F}_t -measurable intuitively means that the value of $h(\omega)$ is decided by the values of $B_s(\omega)$ for $s \leq t$. Note that $\{\mathcal{F}_t\}$ is increasing for all t , that is, $\mathcal{F}_s \subset \mathcal{F}_t$ for any $s < t$.

Definition 1.6 Let $\{\mathcal{N}_t\}_{t \geq 0}$ be an increasing family of σ -algebras of subsets of Ω . The process $g(t, \omega) : [0, \infty) \times \Omega \rightarrow \mathbf{R}^n$ is called $\{\mathcal{N}_t\}$ -adapted if the variable

$$w \rightarrow g(t, \omega)$$

is \mathcal{N}_t -measurable for each $t \geq 0$.

Lemma 1.1 (*The Itô isometry*) Suppose $\phi(t, \omega)$ is \mathcal{F}_t -adapted and $E\left[\int_S^T \phi(t, \omega)^2 dt\right] < \infty$, then

$$E\left[\left(\int_S^T \phi(t, \omega) dB_t(\omega)\right)^2\right] = E\left[\int_S^T \phi(t, \omega)^2 dt\right], \quad (1.9)$$

for any $0 \leq s \leq t < \infty$.

1.4 Euler Method

Euler method is one of the most popular methods to simulate the solution of stochastic differential process. The Euler method approximates X by a continuous stochastic process Y fulfilling the iteration that

$$Y_{t_{i+1}} = Y_{t_i} + b(t_i, Y_{t_i})(t_{i+1} - t_i) + \sigma(t_i, Y_{t_i})(W_{t_{i+1}} - W_{t_i}), \quad (1.10)$$

where $i = 0, 1, \dots, N - 1$ and with the same initial value $X_{t_0} = Y_{t_0}$. And we usually take the time increment $\Delta t = t_{i+1} - t_i$ is set to be constant. Furthermore, the process between any two times t_i and t_{i+1} is considered to be linear interpolation, and is defined as

$$Y(t) = Y_{t_i} + \frac{t - t_i}{t_{i+1} - t_i}(Y_{t_{i+1}} - Y_{t_i}), t \in [t_i, t_{i+1}). \quad (1.11)$$

Chapter 2

Estimation Method

In this chapter we preview the method studied in this thesis, closely following the paper Mancini et. al. [24], where the method was first proposed.

2.1 The Model

Suppose the logarithmic price $X = \{X_t\}$ is defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, P)$ and X is the solution of the following differential equation

$$dX_t = \mu_t dt + \sigma_t dW_t, \tag{2.1}$$

where $W = \{W_t\}_{t \geq 0}$ is a standard Brownian motion defined on the filter probability space, $\{\mu_t\}_{t \geq 0}$ and $\{\sigma_t\}_{t \geq 0}$ are adapted stochastic processes. The process $\{\sigma_t\}$ is called the spot volatility process, and $\{\mu_t\}$ is the drift process.

The realized volatility is the most common popular method to gain the information of volatility, which concentrate on the quadratic variation of $\{X_t\}$. The quadratic variation at time $t > 0$ is defined as

$$[X]_t = \lim_{\Delta \rightarrow 0} \sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})^2 \quad (2.2)$$

where $0 = t_0 < t_1 < t_2 < \dots < t_n = t$ is any partition of $[0, T]$, with $\Delta := \max_{i=1, \dots, n} |t_i - t_{i-1}|$, (cf. Protter (2004 Thm II22)) . The alternative representation of quadratic variation is the time integral of the variance process:

$$[X]_t = \int_0^t \sigma_s^2 ds \quad (2.3)$$

However, normally a full recovery is not available. We can use a high-frequency discrete sample $\{X_t\}_{i=1, \dots, n}$ over the interval $[0, T]$ to estimate the quadratic variation. As shown from (2.2), a natural finite-sample estimate of (2.3) is

$$[\widehat{X}]_t = \sum_{i=1}^n I\{t_{i-1} < t\} (\Delta X_{t_{i-1}})^2, \quad (2.4)$$

where $\Delta X_{t_{i-1}} = X_{t_i} - X_{t_{i-1}}$, $i = 1, \dots, n$, which is called the realized quadratic variation of X over the sample time $t_1 \leq t_2 \leq \dots \leq t_n$. On the other hand, the estimator above can be viewed as the average of the quadratic variation weighted by the function $I\{t_{i-1} < t\}$.

Now we consider a function with the following property:

(a) $\int \delta(x) dx = 1,$

(b) $\delta(x) = 0,$ except $x = 0,$

which is a weighted function giving all its mass to one point. Then, we can easily derive one of its important properties,

$$\int_{-\infty}^{\infty} \delta(x - a)\phi(x)dx = \phi(a), \quad (2.5)$$

for a certain function $\phi(x)$. The $\delta(x)$ function is called Dirac delta function in mathematics. Based on the property above, if we weighted the delta function on the quadratic variation, we may achieve the information of volatility at any exact time. However, Dirac delta function does not actually exist. But it can be generated by sequence of function $\{\delta_n(x)\}$, with

$$\lim_{n \rightarrow \infty} \delta_n(x) = \delta(x), \quad (2.6)$$

for $x \in \mathbf{R}$, which is delta sequence. Then, we can define the main estimator used in this thesis:

$$\hat{\sigma}_{n,f}^2(\bar{t}) = \sum_{i=1}^n f_n(t_{i-1} - \bar{t})(\Delta X_i)^2, \quad (2.7)$$

where $\{f_n(\cdot)\}$ is a delta sequence. Its definition and properties will be introduced in the next section.

2.2 Assumptions

The first assumption is a type of local weak Hölder continuity conditions on $t \rightarrow \sigma_t^2$.

Assumption 2.1 *Supposed $\bar{t} \in [0, T]$ and $\varepsilon > 0$ are fixed, let $B_\varepsilon(\bar{t}) = [\bar{t} - \varepsilon, \bar{t} + \varepsilon]$, and assuming that there exists a constant $\Gamma > 0$, then we assume that a sequence of stopping times $\tau_m \uparrow \infty$, and constant $C_{\bar{t}}^{(m)}$ such that for all m , for $u \in B_\varepsilon(\bar{t})$. We have*

$$E_{\min\{u,s\}}[|\sigma_u - \sigma_s|^2] \leq C_{\bar{t}}^{(m)}|u - s|^\Gamma. \quad (2.8)$$

The next assumption gives conditions of the time of observation which partition $[0, T]$.

Assumption 2.2 *Suppose there are $n + 1$ observations for the process X at time $0 = t_0 < t_1 < \dots < t_n = T$, with T fixed. Set $\Delta_i = t_i - t_{i-1}$ and $\bar{\Delta}_n = \frac{T}{n}$. Then we assume $\max_{i=1, \dots, n} \Delta_i = O(\bar{\Delta}_n)$ and the quadratic variation of time up to $t \leq T$, $H(t) = \lim_{n \rightarrow \infty} \sum_{t_i \leq t} H_n(t)$, where*

$$H_n(t) = \frac{1}{\bar{\Delta}_n} \sum_{t_i \leq t} (\Delta_i)^2. \quad (2.9)$$

We require that H is Lebesgue-almost everywhere differentiable in $[0, T]$, with H' bounded such that for some $K \geq 0$, and

$$|H'(t_i) - \frac{\Delta_i}{\bar{\Delta}_n}| \leq K \Delta_i, \quad (2.10)$$

for any t_i in which H is differentiable, $i = 1, 2, \dots, n$

When the observations are equally spaced, we have $\Delta_i = \bar{\Delta}_n$, $H(t) = t$, $H'(t) = 1$, and (2.7) is satisfied. If the observations are getting more concentrated around t , $H'(t)$ tends to be smaller than 1. Conversely, if the observations are sparse around t , we have $H'(t) > 1$. The assumption $\max_{i=1, \dots, n} \Delta_i = O(\bar{\Delta}_n)$ guarantees the partition does not vary asymptotically too much with regarding to the equally spaced partition.

We now introduce what δ -sequence is, which generalizes kernel functions.

Definition 2.1 *$\{f_n\}_{n \in \mathbf{N}}$ is a sequence of functions with $f_n : D \rightarrow \mathbf{R}$, and $D \subseteq \mathbf{R}$ a given set and $0 \in \mathring{D}$, is said to be a delta sequence when for all the processes $\{\sigma_t\}$ satisfying Assumption 2.1,*

$$\int_0^T f_n(s - \bar{t}) \sigma_s^2 ds = (\sigma^2)_{\bar{t}}^* + R_n^{(\sigma^2)}(\bar{t}), \quad (2.11)$$

$$\frac{1}{f_n(0)} \int_0^T f_n^2(s - \bar{t}) \sigma_s^2 ds = c_f(\sigma^2)_{\bar{t}}^* + o_p(1), \quad (2.12)$$

$$\frac{1}{f_n^2(0)} \int_0^T f_n^4(s - \bar{t}) \sigma_s^2 ds = O_p(f_n(0)), \quad (2.13)$$

Where $R_n^{(\sigma^2)}(\bar{t}) = o_p(1)$ and

$$(\sigma^2)_t^* = (\psi_f^+ \sigma_t^2 + \psi_f^- \sigma_{t^-}^2) I_{\{t \in (0, T)\}} + \psi_f^- \sigma_T^2 I_{\{t = T\}} + \psi_f^+ \sigma_0^2 I_{\{t = 0\}}. \quad (2.14)$$

where $\int_{x < 0} f_n(x) dx \rightarrow \psi_f^-$ and $\psi_f^+ = 1 - \psi_f^-$. For symmetric delta sequences, we have $\psi_f^+ = \psi_f^- = \frac{1}{2}$.

$\overset{\circ}{D}$ above represents the interior of D . When estimating the value at boundaries ($\bar{t} = 0$ or $\bar{t} = T$), we would only weight for the delta sequence at the right and left of \bar{t} .

For two sequences $\{A_n\}$ and $\{B_n\}$, $A_n = O_p(B_n)$ means that there exists N such that for all $n > N$, there exists a constant $\eta > 0$ satisfying that $P(|A_n| > \eta|B_n|) < \varepsilon$, for any $\varepsilon > 0$. As for the notation o_p , if $\{B_n\}$ is a positive real sequence goes to zero as $n \rightarrow \infty$, we can say $A_n = o_p(B_n)$, if $\frac{A_n}{B_n} \rightarrow 0$ in probability.

In (1), note that $(\sigma^2)_t^* = \sigma_t^2$ if σ_t^2 is continuous at $\bar{t} \in (0, T)$. When estimating the value at boundaries ($\bar{t} = 0$ or $\bar{t} = T$), we would only weight for the delta sequence at the right and left of \bar{t} .

Condition 2.11 is very close to the normal definition of delta sequence, which has already been introduced to estimate probability density since 1979 by G. Walter and J. Blum [9]. Condition 2.12 and 2.13 are used to guarantee the central limit theorem.

Additional conditions are stated below.

Assumption 2.3 Suppose $F = \{f_n, n \in \mathbf{N}\}$ is a delta sequence converging to Dirac delta function, we assume that $f_n(0) \rightarrow +\infty$ and $\int_D f_n(x)dx \rightarrow 1$ as $n \rightarrow \infty$, and further,

(i) $\sup_{x \in D} |f_n(x)| \leq C f_n(0)$ for some constant C .

(ii) f_n satisfies Lipschitz condition in a neighborhood of 0 with a Lipschitz constant L_n such that $L_n \sqrt{\overline{\Delta}_n} / f_n(0) \rightarrow 0$; further either $f_n \geq 0$ or $\overline{\Delta}_n^{\Gamma/2} \sum_i |f_n(t_{i-1} - \bar{t}) \Delta_i| \rightarrow 0$.

(iii) There exists a constant $M_\varepsilon > 0$ that does not depend on n such that

$$\sup_{x \in B_\varepsilon^c(0)} |f_n(0)| \leq M_\varepsilon. \quad (2.15)$$

To make the Definition 2.1 of delta sequence more straightforward and sufficient to verify without mentioning the process of $\{\sigma_t\}$, the following proposition is introduced.

Proposition 2.1 Suppose f_n is a sequence of nonnegative functions from D to \mathbf{R} , with $D \subset \mathbf{R}$ and $0 \in \overset{\circ}{D}$. f_n satisfying conditions (i)-(iii), and further, as $n \rightarrow \infty$,

(iv)

$$\int_D f_n(x)dx \rightarrow 1, \quad (2.16)$$

(v) there exists a sequence $\varepsilon_n \rightarrow 0$ such that

$$\int_{-\varepsilon_n}^{\varepsilon_n} f_n(x)dx \rightarrow 1, \quad (2.17)$$

(vi) there exists a real constant c_f such that

$$\int_D \frac{f_n^2(x)}{f_n(0)} dx \rightarrow c_f. \quad (2.18)$$

Then $\{f_n\}$ is a delta function.

Example 1 (Kernels) Let $K : \mathbf{R} \rightarrow \mathbf{R}$ be continuously differentiable and $\{h_n\}$ be a positive sequence with $h_n \rightarrow 0$. We define

$$f_n(x) = \frac{1}{h_n} K\left(\frac{x}{h_n}\right). \quad (2.19)$$

The sequence $\{h_n\}$ is usually called bandwidth. Then, $f_n(x)$ is a delta sequence. Since we have $f_n(0) = \frac{1}{h_n} K(0)$, $f_n(0)$ can also be understood as the inverse of the bandwidth. The kernel estimators was used to estimate spot volatility by Kristensen [21] and can be used to generate a class of delta sequences.

In this case, we can reformulate Assumption 2.3:

Assumption 2.4

(1) $\int_{-\infty}^{+\infty} K(x) dx = 1$ and $\int_{-\infty}^{+\infty} K^2(x) dx = c_2$ (that is, $c_f = \frac{c_2}{K(0)}$).

(2) $\sup_{x \in \mathbf{R}} |K(x)| \leq CK(0)$.

(3) K is differentiable almost everywhere and K' is bounded. Also, h_n satisfies that

$$\sup_{x \in \mathbf{R}} |K'(xh_n^{-1})| \sqrt{\Delta_n/h_n^3} \rightarrow 0. \quad (2.20)$$

(4) $\sup_{x \in B_\varepsilon(\bar{t})} \left| \frac{1}{h_n} K\left(\frac{x}{h_n}\right) \right| \leq M_\varepsilon$, where M_ε is a constant that does not depend on n .

The following are classical examples,

(1) (*Gaussian kernel*) $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, thus $c_2 = \frac{1}{2\sqrt{\pi}}$ and $c_f = \frac{1}{\sqrt{2}}$.

(2) (*Epanechnikov kernel*) $K(x) = \frac{3}{4}(1-x^2)I_{\{|x| \leq 1\}}$, thus $c_2 = \frac{3}{5}$ and $c_f = \frac{4}{5}$.

(3) (*Indicator kernel*) $K(x) = \frac{1}{2}I_{\{|x| \leq 1\}}$, then $c_2 = \frac{1}{2}$ and $c_f = 1$.

(4) (*Double exponential kernel*) $K(x) = \frac{e^{-|x|}}{2}$, then $c_2 = 1$, $c_f = 2$.

Example 2 In Fourier analysis, trigonometric functions are often used to approximate the delta Dirac function.

(1) (*Dirichlet sequence*) Let $g_n(x) = \frac{1}{2\pi} D_{N_n}(x)$, $x \in [-\pi, \pi]$, where

$$D_N(x) := \sum_{|h| \leq N} e^{ihx} = \frac{\sin\left(\left(N + \frac{1}{2}\right)x\right)}{\sin \frac{x}{2}}, \quad (2.21)$$

and $\{N_n\}$ is a diverging sequence. Note that the Dirichlet sequence can be negative at some points.

(2) (*Fejér sequence*) The delta sequence is given by $f_n(x) = \frac{1}{2\pi} F_{N_n}(x)$ with domain $(-\pi, \pi)$, where $\{F_N\}$ is the Fejér sequence

$$F_N(x) := \sum_{|s| \leq N} \left(1 - \frac{|s|}{N+1}\right) e^{isx} = \frac{1}{N+1} \left(\frac{\sin\left(\frac{N+1}{2}x\right)}{\sin \frac{x}{2}}\right)^2, \quad (2.22)$$

and $\{N_n\}$ is a diverging sequence.

Now we can verify the properties of Fejér sequence and Dirichlet sequence for all N :

$$(a) \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} F_N(x) dx = 1, \quad \frac{1}{F_N(0)} \int_{-\pi}^{\pi} F_N^2(x) dx = \frac{4\pi}{3},$$

$$(b) \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} D_N(x) dx = 1, \quad D_N^2(x) = (2N + 1)F_{2N}(x).$$

Thus, the integration of f_n and g_n is 1 and we have $c_f = \frac{2}{3}$ and $c_g = 1$. Since when $\varepsilon \leq |x| \leq \pi$, we have $\frac{1}{\sin(x)} \leq \frac{1}{\sin(\frac{\varepsilon}{2})}$, it is easy to prove the conditions (iv) and (vi) in Proposition 2.1 for f_n . Furthermore, when $\varepsilon = \varepsilon_n \rightarrow 0$, we can obtain

$$\int_{\varepsilon_n \leq |x| \leq \pi} F_{N_n}(x) dx \leq \frac{1}{N_n + 1} \frac{1}{\sin^2(\frac{\varepsilon_n}{2})} 2(\pi - \varepsilon_n). \quad (2.23)$$

When $\varepsilon_n^2 N_n \rightarrow \infty$, (2.23) above converges to 0. Thus the Property 2.1 (v) is proved. Therefore $\{f_n\}$ is a delta sequence

2.3 Asymptotic Theory

We have already introduced the estimator of volatility $\hat{\sigma}_{n,f}^2(\bar{t}) = \sum_{i=1}^n f_n(t_{i-1} - \bar{t})(\Delta X_i)^2$, for $\bar{t} \in D$ in Section 2.1. The following result of Mancini et. al. [24] derives the asymptotic distribution of the estimator $\hat{\sigma}_{n,f}^2(\bar{t})$.

Theorem 2.1 *Suppose Assumptions 2.1, 2.2, and 2.4 hold, as $n \rightarrow \infty$, and $f_n(0) \rightarrow \infty$ in a way that $f_n(0)\bar{\Delta}_n \rightarrow 0$. Then any $\bar{t} \in [0, T]$, we have*

$$\hat{\sigma}_{n,f}^2(\bar{t}) \xrightarrow{p} (\sigma^2)_{\bar{t}}^*, \quad (2.24)$$

where $(\sigma^2)_{\bar{t}}^*$ is given in (2.14). Furthermore, if we have $R_n^{(\sigma^2)}(\bar{t}) = o_p(\sqrt{f_n(0)\overline{\Delta}_n})$, then we can obtain

$$\frac{1}{\sqrt{f_n(0)\overline{\Delta}_n}}(\hat{\sigma}_{n,f}^2 - (\sigma^2)_{\bar{t}}^*) \longrightarrow \mathbf{MN}(\mathbf{0}, 2c_f(H'\sigma^4)_{\bar{t}}^*), \quad (2.25)$$

where the convergence is stable in law.

Above, $\mathbf{MN}(0, V)$ represents a multivariate normal distribution with mean 0 and stochastic variance V . The stable convergence in law mentioned is one mode of convergence that was introduced by Rényi (1963). It is stronger than the convergence in law, and its important feature is that A_n is any sequence of variables converging in probability to a limit A on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbf{P})$, whereas if another variables B_n converge stably in law to B , then the pair (A_n, B_n) converges to the pair (A, B) stably in law.

Brief Outline of the proof of theorem convergence in law: Without loss generality, we can set $\mu_t \equiv 0$. We have

$$\begin{aligned} \frac{1}{\sqrt{f_n(0)\overline{\Delta}_n}}(\hat{\sigma}_{n,f}^2 - (\sigma^2)_{\bar{t}}^*)^2 &= \frac{1}{\sqrt{f_n(0)\overline{\Delta}_n}} \left(\sum_{i=1}^n f_n(t_{i-1} - \bar{t}) \Delta(X_i)^2 - (\sigma^2)_{\bar{t}}^* \right) \\ &= \frac{1}{\sqrt{f_n(0)\overline{\Delta}_n}} \left(\sum_{i=1}^n f_n(t_{i-1} - \bar{t}) (\Delta(X_i)^2 - \int_{t_{i-1}}^{t_i} \sigma^2(s) ds) \right. \\ &\quad \left. + O_{a.s.}(L_n \overline{\Delta}_n) - R_n^{(\sigma^2)}(\bar{t}) \right) \\ &= \sum_{i=1}^n U_i + O_{a.s.} \left(L_n \sqrt{\frac{\overline{\Delta}_n}{f_n(0)}} - \frac{R_n^{(\sigma^2)}(\bar{t})}{\sqrt{f_n(0)\overline{\Delta}_n}} \right), \end{aligned}$$

where, for $i = 1, \dots, n$.

$$U_i := f_n(t_{i-1} - \bar{t}) \sqrt{f_n(0)\overline{\Delta}_n} \left(\left(\int_{t_{i-1}}^{t_i} \sigma_s dW_s \right)^2 - \int_{t_{i-1}}^{t_i} \sigma_s^2 ds \right),$$

Since we have assumed the $L_n\sqrt{\overline{\Delta_n}/f_n(0)} \rightarrow 0$ and $R_n^{(\sigma^2)}(\bar{t}) = o_p(\sqrt{f_n(0)\overline{\Delta_n}})$ in assumptions, the last two terms above converge to zero in probability. Therefore it is sufficient to prove the asymptotic theorem stable in law for $\sum_{i=1}^n U_i$.

According to Theorem IX. 7.28 in Jacod J. and Shiryaev A. ([19]), it is enough to prove the following conditions to achieve our goal:

$$\begin{aligned} (i) \sum_{i=1}^n E_{i-1}[U_i] &\xrightarrow{p} 0, & (ii) \sum_{i=1}^n E_{i-1}[U_i^2] &\xrightarrow{p} V_{\bar{t}}, \\ (iii) \sum_{i=1}^n E_{i-1}[U_i^4] &\xrightarrow{p} 0, & (iv) \sum_{i=1}^n E_{i-1}[U_i \Delta Z_i] &\xrightarrow{p} 0, \end{aligned}$$

Where $E_{i-1}[\cdot]$ represents $E[\cdot|\mathcal{F}_{t_{i-1}}]$. To be clear, we just provide the proof the condition (i) and (ii) here,

By Itô isometry, condition (i) is easily proved,

$$\sum_{i=1}^n E_{i-1}[U_i] = \sum_{i=1}^n \frac{f_n(t_{i-1} - \bar{t})}{\sqrt{f_n(0)\overline{\Delta_n}}} E_{i-1}\left[\left(\int_{t_{i-1}}^{t_i} \sigma_s dW_s\right)^2 - \int_{t_{i-1}}^{t_i} \sigma_s^2 ds\right] = 0.$$

The proofs of conditions (ii), (iii) and (iv) are complicated and we do not show in this paper (see Mancini et. al. [25]).

Remark: When the samples size is small, we can use the the estimator

$$\hat{\sigma}_{n,f}^2(\bar{t}) = \frac{\sum_{i=1}^n f_n(t_{i-1} - \bar{t})(\Delta X_i)^2}{\sum_{i=1}^n f_n(t_{i-1} - \bar{t})\Delta_i}. \quad (2.26)$$

As $n \rightarrow \infty$, $\sum_{i=1}^n f_n(t_{i-1} - \bar{t})\Delta_i \rightarrow 1$, thus we can immediately achieve the same result as in Theorem 2.1. Furthermore, this estimator can also be used to remove some boundary effect.

For example, when using a symmetric kernel delta sequence to estimate σ_τ^2 at $\tau = T$ will lead to $E[\hat{\sigma}_T^2] = \frac{1}{2}\sigma_T^2 + o(1)$ as $h \rightarrow 0$, which is called boundary or edge effect.

2.4 Robustness to Microstructure Noise Effects

As we have introduced in the introduction part, bid-ask bounces, discreteness of price changes and rounding, trades occurring on different markets or networks, and some other may cause microstructure noise. Thus, it is important to deal the situation when microstructure noise is present. To make the result simple, we consider logarithmic asset prices observed at equispaced times t_0, t_1, \dots, t_n , and the microstructure noise is supposed to be additive and independent and identically distributed (i.i.d). More concretely, we have

Assumption 2.5 *Supposed that observations are equally spaced ($\Delta_i = \bar{\Delta}_n$). Let*

$$X_{t_i} = Y_{t_i} + \varepsilon_i, \quad (2.27)$$

where X_{t_i} is the observed logarithmic asset prices, Y_{t_i} is the unobservable efficient prices satisfying the Assumption 2.1, and ε_i represents the microstructure noises component. Assume that the noise process $\{\varepsilon_i\}_{i=0,1,\dots,n}$ is independent of Y and i.i.d., with $E[\varepsilon_i] = 0$ and $E[\varepsilon_i^8] < +\infty$. Set that $V_\varepsilon = E[\varepsilon_i^2]$ and $\Lambda_\varepsilon = E[\varepsilon_i^4]$.

Lemma 2.1 *Assume Assumption 2.4 and 2.5 hold. If $R_n^{(\sigma^2)}(\bar{t}) = o_p(\sqrt{\bar{\Delta}_n f_n(0)})$ and $f_n(0)\bar{\Delta}_n \rightarrow \infty$, we have*

$$\frac{1}{\sqrt{f_n(0)\bar{\Delta}_n}} \left(\frac{1}{2} \bar{\Delta}_n \hat{\sigma}_{n,f}^2(\bar{t}) - V_\varepsilon \right) \xrightarrow{d} \mathbf{N}\left(0, \frac{1}{2} c_f(\Lambda_\varepsilon + V_\varepsilon^2)\right). \quad (2.28)$$

Then, we can easily derive that our original estimator has a microstructure-induced bias, and

$$E[\hat{\sigma}_{n,f}^2(\bar{t}) - \sigma^2(\bar{t})] = \frac{2V_\varepsilon}{\bar{\Delta}_n} + o_p\left(\frac{1}{\bar{\Delta}_n}\right), \quad (2.29)$$

and a consistent estimate of the noise variance can be obtained as

$$\hat{V}_\varepsilon = \frac{1}{2} \bar{\Delta}_n \hat{\sigma}_{n,f}^2(\bar{t}). \quad (2.30)$$

Following the two-scale approach in Zhang et al. [32], in order to obtain a consistent spot variance estimator with asymptotically normal distribution, we set

$$\hat{\sigma}_{n,\bar{n}}^{2,TS}(\bar{t}) = \frac{1}{\bar{n}} \sum_{i=1}^{n-\bar{n}+1} f_n(t_{i-1} - \bar{t}) \left((X_{t_{i+\bar{n}-1}} - X_{t_{i-1}})^2 - (X_{t_i} - X_{t_{i-1}})^2 \right). \quad (2.31)$$

Then we have the following asymptotic theorem.

Theorem 2.2 *Suppose Assumption 2.4 and 3.1 hold, that, $\bar{n}f_n(0)\bar{\Delta}_n \rightarrow 0$ as $\bar{n} \rightarrow \infty$, and $R_n^{(\sigma^2)}(\bar{t}) = o_p(1)$. Then we have*

$$\hat{\sigma}_{n,\bar{n}}^{2,TS}(\bar{t}) \xrightarrow{p} (\sigma^2)_{\bar{t}}^*. \quad (2.32)$$

Furthermore, if we have $R_n^{(\sigma^2)}(\bar{t}) = o_p\left(\sqrt{f_n(0)\bar{\Delta}_n\bar{n}}\right)$ and $\bar{n} = c(\bar{\Delta}_n)^{-\frac{2}{3}}$ with $c \in \mathbf{R}$, then

$$\frac{1}{\sqrt{f_n(0)(\bar{\Delta}_n)^{1/3}}} \left(\hat{\sigma}_{n,\bar{n}}^{2,TS}(\bar{t}) - (\sigma^2)_{\bar{t}}^* \right) \rightarrow \mathbf{MN}\left(\mathbf{0}, 2c_f(V_\varepsilon^2 + c(\sigma^4)_{\bar{t}}^*)\right), \quad (2.33)$$

where the convergence is stable in law.

2.5 Robustness to Jumps

To learn about the stochastic features of irregular jump arrivals, it is crucial to develop a robust test to detect jumps. This is our main assumption

Assumption 2.6 *Assume process $\{X_t\}$ is adapted and defined on $[0, T]$. Let*

$$X_t = Y_t + J_t, \quad (2.34)$$

where $\{Y_t\}$ satisfies Assumption 2.1 and $dJ_t = c_J(t)dN_t$, where $\{N_t\}$ is a nonexplosive Poisson counting process with adapted intensity $\{\lambda_t\}$, the jumps occur at times $\tau_1, \dots, \tau_{N_t}$, and the sizes of jumps $c_J(\tau_j)$ are i.i.d. and satisfy that $P[c_J(\tau_j) = 0] = 0$, for $j = 1, \dots, N_t$.

Following the non-parametric threshold estimation proposed in C. Mancini [24], we have our threshold estimator

$$\hat{\sigma}_{n,f}^2(\bar{t}) = \sum_{i=1}^n f_n(t_{i-1} - \bar{t})(\Delta X_i)^2 I_{\{(\Delta X_i)^2 \leq \theta_n\}}, \quad (2.35)$$

where $\{\theta_n\}$ is a suitable sequence converging to 0 used to disentangle the discontinuous variation induced by the jumps. If convergence of $\{\theta_n\}$ to zero is much slower than the modulus of continuity of the Brownian paths, then the disentangling is possible as stated by the following theorem

Theorem 2.3 *Let assumptions 2.2, 2.4, and 2.6 hold. If as $n \rightarrow \infty$, $f_n(0) \rightarrow \infty$, $\theta_n \rightarrow 0$, $f_n(0)\bar{\Delta}_n \rightarrow 0$, and $\theta_n/(\bar{\Delta}_n \log(\frac{1}{\bar{\Delta}_n})) \rightarrow \infty$, we have*

$$\hat{\sigma}_{n,f}^2(\bar{t}) \xrightarrow{p} (\sigma^2)_{\bar{t}}^*. \quad (2.36)$$

If we further have $R_n^{(\sigma^2)}(\bar{t}) = o_p(\sqrt{f_n(0)\bar{\Delta}_n})$, then we can gain

$$\frac{1}{\sqrt{f_n(0)\bar{\Delta}_n}}(\hat{\sigma}_{n,f}^2(\bar{t}) - (\sigma^2)_{\bar{t}}^*) \longrightarrow \mathbf{MN}(\mathbf{0}, 2c_f(H'\sigma^4)_{\bar{t}}^*), \quad (2.37)$$

and the convergence above is stable in law.

Chapter 3

Simulation Experiments

The goal of this section is to compare different delta sequences. The bias of the estimator in theorem 2.1 is decided both by the regularity of σ and the choice of kernel. For a simple example, if f_n is an indicator function, the bias is $O(f_n(0)^{-\Gamma/2})$, and since the variance is $O(f_n(0)\overline{\Delta}_n)$, the optimal form of $f_n(0)$ is proportional to $(\overline{\Delta}_n)^{-\frac{1}{1+\Gamma}}$, and the estimator convergence at the rate of $n^{\frac{1}{4}}$.

We consider the following Heston model (Heston [12])

$$dX_t = \mu_t dt + \sigma_t dW_{1,t}, \quad (3.1)$$

$$d\sigma_t^2 = \beta(\alpha - \sigma_t^2)dt + \kappa_\sigma \sigma_t dW_{2,t}, \quad (3.2)$$

where $W_{1,t}$ and $W_{2,t}$ are standard Brownian motions, and $\text{corr}(dW_{1,t}, dW_{2,t}) = \rho dt$. Equation 3.1 describes the dynamics stock price at time t . Equation 3.2 is the process of the variance following a square root process: α is the long run mean variance, β stands for the speed of mean reversion, and κ_σ represents the parameter that determines the volatility of variance process.

Choose $T = 1/12$ one month as time period. We set $\rho = -0.315$, $\mu = 0$, $\beta = 1.05$, $\alpha = 0.0945$, $\kappa_\sigma = 0.095$, $X_0 = \log(100)$, and $\sigma_0^2 = 0.25$ (these parameters are based on experiments in A. Stoep [29]). 1000 paths are generated with 5 minutes as step size, which means $n = 252 * 12 * 6.5/12 = 1638$. The choice of $f_n(0)$ is crucial for the main theorem, however, this thesis does not explore this part, and set $f_n(0) = 500$ for all the delta sequences used.

Based on Euler method, the paths can be generated as follows:

$$s_{i+1,j} = s_{i,j} + \mu s_{i,j} \bar{\Delta}_n + s_{i,j} \sigma_{i,j} \sqrt{\bar{\Delta}_n} Z_s, s_{0,j} = S(t_0), \quad (3.3)$$

$$\sigma_{i+1,j}^2 = \sigma_{i,j}^2 + \beta(\alpha - \sigma_{i,j}^2) \bar{\Delta}_n + \kappa_\sigma \sigma_t \sqrt{\bar{\Delta}_n} Z_\sigma, \sigma_{0,j}^2 = \sigma^2(t_0), \quad (3.4)$$

for $i = 0, \dots, n$, and $j = 1, 2, \dots, 1000$, where $Z_s = Z_1$, $Z_\sigma = \rho Z_1 + (1 - \rho^2)^{\frac{1}{2}} Z_2$, with Z_1 and Z_2 two independent variables satisfying normal distribution.

Six delta sequences are involved: double exponential, Dirichlet, Fejér, indicator, Epanechnikov, and Gaussian. For each delta sequence, 1000 paths are generated. Figure 3.1 shows that at the time $\bar{t} = 1/24$ which is the mid point of time T , the distribution of relative root errors of estimator in Theorem 2.1. Table 3.1 compares the MSE of different delta sequences, which is defined as

$$MSE = \frac{1}{N} \sum_{j=1}^n \left(\sum_{i=1}^n \left(\hat{\sigma}_j^2(t_i) - \sigma^2(t_i) \right)^2 \right) \quad (3.5)$$

As Figure 3.1 shows, double exponential kernel has the best property when estimating, with the least variance. The other five sequences, however, do not appear to have great differences with each other. Further, the MSE of the six delta sequences during one month

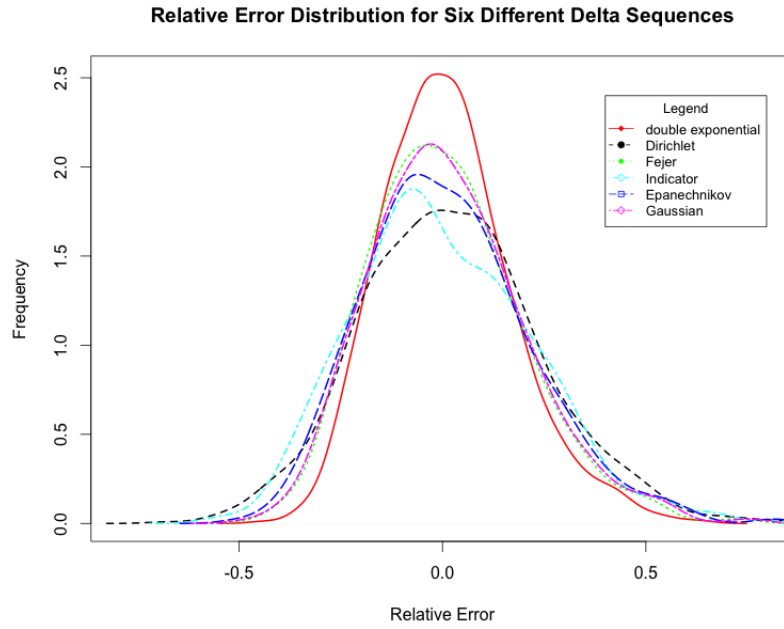


Figure 3.1: The Frequency of Relative Error for Six Delta Sequences

with 5 minutes returns vary from 2.5 to 6.0, and double exponential shows the optimal MSE value.

Table 3.1: Simulated MSE for six delta sequences

	Double Exponential	Dirichlet	Fejér	Indicator	Epanechnikov	Gaussian
MSE	2.566	4.994	3.444	5.082	4.048	3.560

Chapter 4

Application using real market data

4.1 Introduction to TAQ Database

The Trade and Quote (TAQ) database from WRDS (Wharton Research Data Services) contains consolidated intraday transactions data for all securities listed on the American Stock Exchange (AMEX), New York Stock Exchange (NYSE), Nasdaq National Market System (NMS), SmallCap issues, as well as stocks trades on Arca.

There are two categories of data within the TAQ, which are *quotations* and *transactions*. Since early 1990s, the NYSE has been firstly distributing its ultra high-frequency data sets. In 1993, the trades, order, and quotes (TORQ) database, containing a three months sample of data, was released (Hasbrouck, 1992).

Quote data includes information regarding the best trading conditions available on the exchange. Table 4.1 displays some sample records from the quote database with an explanation of the various fields. However, the quote table does not contain any information of the quality of the reported data. Trade data involves information regarding the orders executed on the exchange. Table 4.2 displays few sample records from the trade database. Some information

Table 4.1: Quote data from TAQ database

SYMBOL	DATE	TIME	BID	OFR	BIDSIZ	OFRSIZ	MODE	EX	MMID
AAPL	20140102	8:00:00	1	0	1	0	12	Z	
AAPL	20140102	8:00:00	1	0	2	0	12	Z	
AAPL	20140102	8:00:00	1	0	1	0	12	C	
AAPL	20140102	8:00:00	1	0	2	0	12	C	
AAPL	20140102	8:00:00	556.2	556.9	1	1	12	K	

in the database stands for the quality of the recorded ticks, according to which, wrong or inaccurate ticks are able to be removed: e.g. the CORR field indicate the correction of a tick, and that the field of COND is equal “Z” or “G” indicates a trade reported at a later time.

Table 4.2: Trade data from TAQ database

SYMBOL	DATE	TIME	PRICE	SIZE	G127	CORR	COND	EX
AAPL	20140102	4:00:00	561.02	9	0	0	@	P
AAPL	20140102	4:00:00	560.00	5	0	0	@	P
AAPL	20140102	4:00:00	558.19	86	0	0	@	P
AAPL	20140102	4:00:00	558.19	64	0	0	@	P
AAPL	20140102	4:00:00	557.00	5	0	0	@	P

Considering the large amount of data recorded in the TAQ database, we need to identify the information we are interested in and abandon those wrong and unrelated data. As reported by Faltenberry (2002), errors can be present both in automatic and semiautomatic trading systems. As the velocity of transactions increases, the frequency of errors will increase in reporting system. Therefore, the primary goal in data cleaning is to eliminate the erroneous data, and it is also equally important to deal with outliers and those incompatible data with normal trading market activity.

In this chapter, we apply the proposed theorem to analyze the high-frequency transaction of AAPL in 2014.

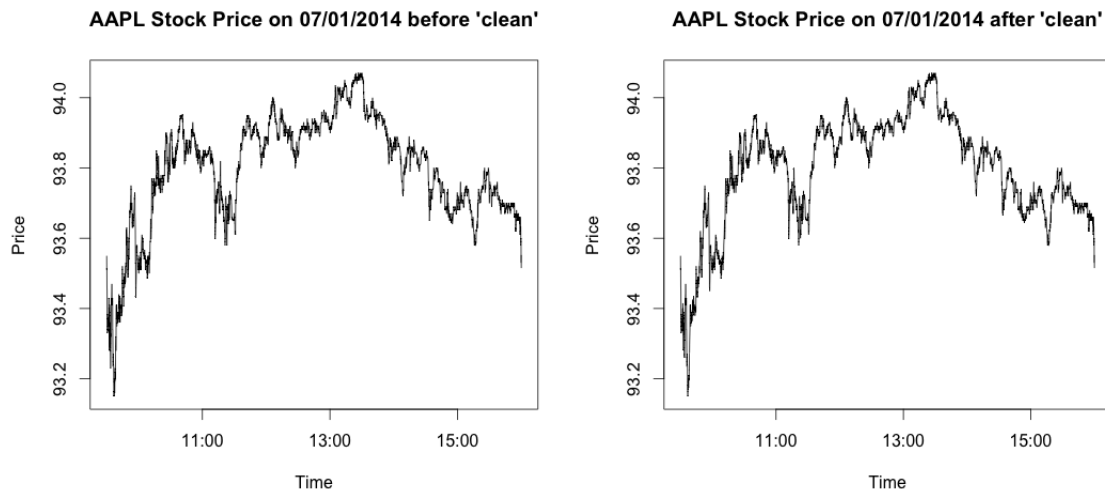


Figure 4.1: AAPL stock price on July 1, 2014

4.2 Data Preprocessing

In order to obtain a clean sample we need to identify and discard the records that are not of interest using available information. The cleaning process is as follows.

Trades were kept if they were regular way trades, that is, trades that had no stated conditions ($\text{COND}='*' \text{ or } \text{COND}=' '$). And CORR indicating that the trade was “regular” or original data which was latter corrected ($\text{CORR}=0, 1, 2$). Secondly, to reduce the effect of outliers, further filter was needed: the 99th percentile of these daily absolute differences $|\Delta X|$ was obtained. Then, if the difference of a price from the prior price was more than twice the 99th percentile of that days absolute differences and this difference’s sign was reversed on the following trade, this trade was eliminated. Figure 4.1 is the comparison of stock price before and after cleaning. Left is the raw data of AAPL market price on July 1, 2014. Right is market price after cleaning.

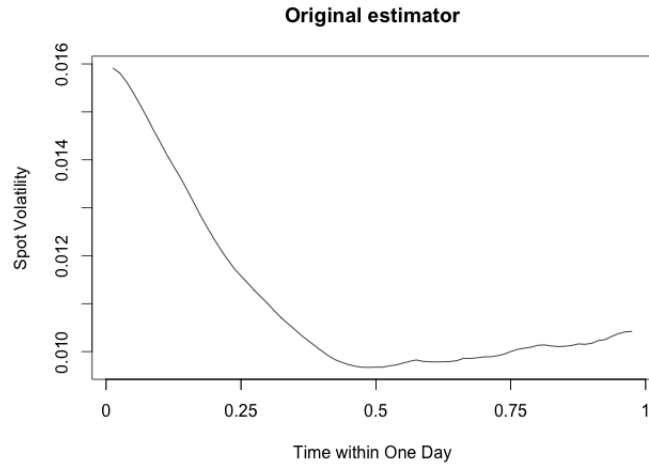


Figure 4.2: Estimated spot volatility for AAPL averaged 128 trading dates using threshold estimator (2.7) with double exponential kernel

Next we focus our attention on the second half year of 2014, from July 1, 2014 to December 31, 2014. Transactions are recorded during 128 trading days between 9:30am to 4:00pm and are interpolated to a 5-minutes grid. 79 prices are recorded every trading day. However, since prices typically are not recorded at equispaced but our model used relies on equispaced returns, the previous tick aggregation method is used, which forces prices to be an equispaced time grid by taking the last price realized before each grid point. Furthermore, in the case that there are a lot of prices within the same second, the last price in the second is recorded. During calculation, we set one day as time unit.

4.3 Estimation

Firstly, the proposed original estimator (2.7) and threshold estimator (2.35) are used to estimate spot volatility as applications. Double exponential kernel is chosen. Set 45 minutes as bandwidth and one day as unit.

The threshold sequence used is chosen according to Jacod and Todorov [18], $\theta_n = 4\bar{\Delta}_n^\omega \sqrt{BPV}$, where $\omega = 0.49$, and BPV denotes the bi-power variation within one day, which is defined as

$$BPV = \sum_{j=2}^n |\Delta X_{j-1}| |\Delta X_j|. \quad (4.1)$$

The process of estimation is executed day by day. Then, average all volatility estimations of 128 days by the same time of each day. Figure 4.3 is volatility $\{\sigma_t\}$ estimated.

Furthermore, estimator (2.30) in the presence of noise is used and set $\bar{n} = 18$. However, without detecting the jump, the figure of spot volatility performs badly and not accurate. Then $C - Tz$ test is used to identify the day that has a jump and then delete the day's data (see Corsia et. al. [6]). By setting the significant of jump detection at 99%, 16 days are detected to have jumps. After excluding these days data, the estimation of spot volatility is executed day by day and then average. The figure of spot volatility $\{\sigma_t\}$ is shown in figure 4.4.

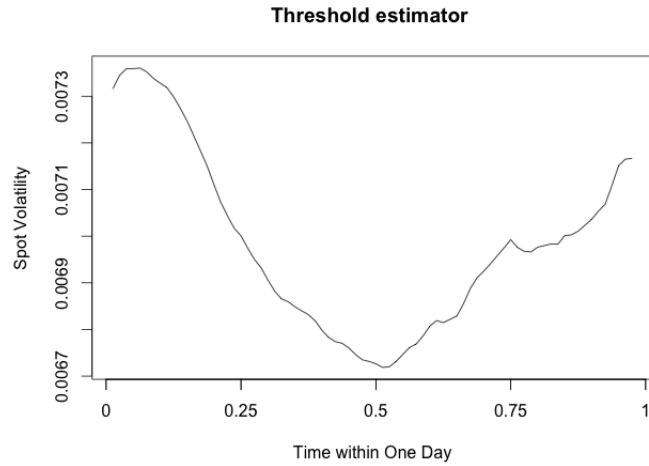


Figure 4.3: Estimated spot volatility for AAPL averaged 128 trading dates using threshold estimator (2.35) with double exponential kernel

4.4 Analysis

Comparing Figure 4.2 and Figure 4.3, we can see both of the two curves significantly decrease during the first several hours after the market opening time till to the noon. Then the figure of original estimator slightly increases. However, the figure of threshold estimator has a great increase in the afternoon. Since most of the big jumps of the dataset are detected to happen in the morning instead of randomly distributed, it will greatly affect the shape of volatility process estimated. But we can still observe the U-shape of volatility curves from the two figures.

The curves of Figure 4.4 and 4.3 are very similar. Both of these two figures sharply decrease from the open time of market to the time around noon reaching the lowest level. After noon, the spot volatility began to increase till to the market close time with a little fluctuation.

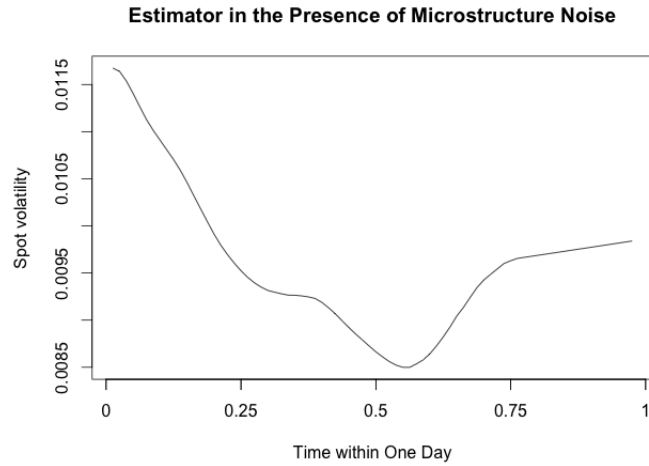


Figure 4.4: Estimated spot volatility for AAPL in the presence of microstructure noise after exclude 16 days which is detected to have jumps by using $C - Tz$ test

This kind of U-shape curves can well describe the dynamics of volatility within a day, which is described as volatility smiles.

On the other hand, Since we are using 5 minutes price return, microstructure noise does not have great influence on volatility estimation. Microstructure is softened in this time grid. Besides, as \bar{n} increase, the edge effect increases greatly. When estimating the points that are close to the end side, the estimator appeared to have great deviation.

Chapter 5

Conclusion

This thesis studies and summarizes the spot volatility estimation proposed by Mancini et. al. [24], which unifies a class of spot volatility estimator constructed by delta sequences. Two classes of delta sequences are mainly mentioned, kernel estimator as well as non-kernel estimator (e.g. Dirichlet sequence). Then a full asymptotic theory is proposed under some gentle assumptions and hypothesis. The extended situations with the presence of jump and microstructure noise are also studied.

To compare different delta sequences, data are simulated by Heston Model using Euler method with some restriction. The distributions of relative error of each estimator and the averaged MSE are performed. The distribution of different delta sequences appeared to be similar and close, among which, double exponential sequence tends to be a little better.

In the final part, the theory is applied to real market data. Because most of the big jumps in the dataset are detected to happen during the morning, the shape of volatility using original estimator without considering jumps and microstructure noise is greatly affected. The curve of intraday volatility greatly decreases in the morning and has a slight increase in the afternoon. The threshold estimator and the estimator in the presence of microstructure

noise are also used to test the suitability of the method. Both of two estimators' figures appear in traditional U-shape intraday volatility pattern. Because of using 5 minutes price return, microstructure noise does not show a great influence on volatility estimation.

References

- [1] S. Alizadeh, M. W. Brandt, and F. X. Diebold. Range-based estimation of stochastic volatility models. *Journal of Finance*, 57:1047–1092, 2002.
- [2] Y. At-Sahalia and J. Jacod. Test for jumps in a discretely observed process. *The Annals of Statistics*, 37(1):184–222, 2009.
- [3] Y. At-Sahalia, P. A. Mykland, and L. Zhang. Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, pages 160–175, 2011.
- [4] K. Boudt, J. Cornelissen, and S. Payseur. Highfrequency: Toolkit for the analysis of highfrequency financial data in r.
- [5] C.T. Brownlee and G.M. Gallo. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics and Data Analysis*, 2006.
- [6] F. Corsia, D. Pirinoc, and R. Rend. Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159(2):276–288, December 2010.
- [7] C. V. Eeden. Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Annals of the Institute of Statistical Mathematics*, November 1985.
- [8] J. E. Figueroa-Lopez, S. R. Lancette, K. Lee, and Y. Mi. Estimation of nig and vg models for high frequency financial data. *Handbook of High-Frequency Data in Finance*, 2012.
- [9] J. Blum G. Walter. Probability density estimation using delta sequences. *The Annals of Statistics*, 7(2):328–340, 1979.
- [10] P. R. Hansen and A. Lunde. Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, August 2005.
- [11] N. Hautsch. *Econometrics of Financial High-Frequency Data*. Springer, 2012.
- [12] S. L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Oxford Journal*, 6(2):327–343, 1993.

- [13] N. Hoang-Long and O. Shigeyoshi. A central limit theorem for the functional estimation of the spot volatility. *Monte Carlo Methods and Applications*, 15(4):353–380, 2009.
- [14] M. Hoffmann, A. Munk, and J. Schmidt-Hieber. Adaptive wavelet estimation of the diffusion coefficient under additive error measurements. *Statistics Theory*, July 2010.
- [15] S. M. Iacus. *Simulation and Inference for Stochastic Differential Equations with R Examples*. Springer Series in Statistics, 2012.
- [16] J. Jacod. On continuous conditional gaussian martingales and stable convergence in law. In *Sminaire de Probabilits XXXI*, volume 1655 of *Lecture Notes in Mathematics*, pages 232–246. Springer Berlin Heidelberg, April 1997.
- [17] J. Jacod and M. Rosenbaum. Quarticity and other functionals of volatility: Efficient estimation. *Annals of Statistics*, 41(3):1462–1484, 2013.
- [18] J. Jacod and V. Todorov. Do price and volatility jump together? *The Annals of Applied Probability*, 20(4):1425–1469, 2010.
- [19] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. 0072-7830. Springer-Verlag Berlin Heidelberg, 2 edition, 2010.
- [20] M. Johannes. The statistical and economic role of jumps in continuous-time interest rate models. *The Journal of Finance*, LIX(1), February 2004.
- [21] D. Kristensen. Nonparametric filtering of the realized spot volatility: a kernel-based approach. *Econometric Theory*, pages 60–73, 2010.
- [22] S. S. Lee and P. A. Mykland. Jumps in financial markets: A new nonparametric test and jump dynamics. *Oxford University Press on behalf of The Society for Financial Studies*, 2007.
- [23] P. Malliavin and Maria Elvira Mancino. A fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics*, 37(4):1982–2010, 2009.
- [24] C. Mancini. Non parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 2009.
- [25] C. Mancini, V. Mattiussi, and R. Ren. Spot volatility estimation using delta sequences. *Finance and Stochastic*, 2015.
- [26] R. C. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, pages 323–361, 1980.
- [27] N. Moodley. The heston model: A practical approach with matlab code. Bachelor of Science Honours, Programme in Advanced Mathematics of Finance, 2005.

- [28] P. A. Mykland and L. Zhang. Anova for diffusions and itô process. *The Annals of Statistics*, 34(4):1931–1963, 2006.
- [29] A.W.V.d. Stoep, L. A. Grzelak, and C. W. Oosterlee. The heston stochastic-local volatility model: Efficient monte carlo simulation. *International Journal of Theoretical and Applied Finance*, 17(7), 2014.
- [30] R. S. Tsay. *An introduction to analysis of financial data with R*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2013.
- [31] G. S. Watson and M. R. Leadbetter. *Hazard analysis II*. Springer, July 1964.
- [32] L. Zhang, P. A. Mykland, and Y. AT-SAHALIA. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Stat. Assoc.*, pages 1394–1411, 2005.
- [33] Y. Zu and H. P. Boswijk. Estimating spot volatility with high-frequency financial data. Technical report, City University London, Department of Economics, May 2014.
- [34] B. ksendal. *Stochastic Differential Equations: An Introduction with Application*. Springer, six edition, 2010.