4-16-2015 12:00 AM

Online Nonparametric Estimation of Stochastic Differential Equations

Xin Wang, *The University of Western Ontario*

Supervisor: Duncan Murdoch, *The University of Western Ontario*
Joint Supervisor: Matt Davison, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

ONLINE NONPARAMETRIC ESTIMATION OF STOCHASTIC DIFFERENTIAL EQUATIONS

(Thesis format: Monograph)

by

Xin Wang

Graduate Program in Statistics and Actuarial Science

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

The advent of the big data era presents new challenges and opportunities for those managing portfolios, both of assets and of risk exposures, for the financial industry. How to cope with the volume of data to quickly extract actionable information is becoming more crucial than ever before. This information can be used, for example, in pricing various financial products or in calculating risk exposures to meet (ever changing) regulatory requirements.

Stochastic differential equations are often used to model the risk factors in finance. Given the presumption of a functional form for the coefficients of these equations, the required parameters can be calibrated using a large body of statistical techniques which have been developed over the past decades. However, the price to pay for this convenience is the problems that occur if an incorrect functional form is used. To avoid this problem of misspecification, nonparametric regression has recently become important in finance. In order to adequately estimate local structures, large sample sizes are always required and so nonparametric regression is computationally intensive.

This thesis finds new ways to decrease the computational cost of non-parametric methods for estimating stochastic differential equations. Motivated by stochastic approximations, we propose online nonparametric methods to estimate the drift and diffusion terms of typical financial stochastic differential equations. Both stationary and non-stationary processes are considered and this thesis provides asymptotic properties of the estimators. For the stationary case, quadratic convergence, strong consistency, and asymptotic normality of the estimators are established; for the non-stationary case, weak consistency of the estimators is proved.

In addition to numerical examples, we also apply our methods to market risk management. We work from up to date examples based on the most recent Basel Committee documents for a wide range of risk factors from equity, foreign exchange, interest rates, and commodity prices. The advantages and disadvantages of applying our new statistical techniques to these risk management problems are also discussed.

Keywords: online nonparametric estimation, stochastic differential equations, financial modeling

Acknowledgements

When writing this part, I come to realize that my student life will become memories of the past. Recalling the last two years, from a novice in statistics and finance to writing this thesis, I know how I benefited from the guidance and support of many people.

I greatly appreciate my supervisors, Dr. Murdoch and Dr. Davison, for their valuable guidance, patience and enthusiasm during the years of my doctoral research. Dr. Murdoch's insight and knowledge in statistics broadened my horizons and deepened my understanding in statistics, while Dr. Davison led me into the realm of finance and gave me an intuition beneath the complicated concepts and techniques. They are my most valuable assets.

I would also like to thank my committee members, Dr. Reg Kulperger, Dr. Lars Stentoft, Dr. Hubert Pun and Dr. Adam Kolkiewicz, for their constructive feedback. Many thanks are given to my colleagues in TD Bank and office mates in the university for their kindness and support.

Also, I am very grateful to my parents for their endless love, encouragement, support and patience throughout my life.

Finally, my deep love goes to my wife Sulin Cheng, for her sacrifice and support, always by my side whenever in my busy or dark times, all that keep me going.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis proposes online nonparametric estimators for stochastic differential equations, especially for time-homogeneous diffusion models. For the stationary case, we establish quadratic convergence, strong consistency and asymptotic normality of our estimators. Numerical examples and a case study are used to validate effectiveness and efficiency of our methods. For the non-stationary case, a class of second-order stochastic differential equations is considered and online estimators are proposed and studied.

In this chapter, we first present background materials and some motivation for this thesis. Then the ideas and our contributions are discussed. Finally the organization of this thesis is outlined.

1.1 Background and Motivation

Stochastic differential equations (SDEs) are an essential tool to describe the randomness of a dynamic system. For example, physicists use this tool to model the time evolution of particles due to thermal fluctuations (Sobczyk, 2001) and ecologists study two interacting populations such as predator and prey by SDEs (Allen, 2007). In the financial system, many different SDEs have been developed to model a particular financial product or class of products, e.g. geometric Brownian motion (GBM) by Osborne (1959) for modeling stock prices or stock indices, and for modeling interest rates the Vasicek model (Vasicek, 1977), the CIR model (Cox et al., 1985) and the CKLS model (Chan et al., 1992). Financial institutions make use of SDEs to price their derivatives or measure the risks of their portfolios. For example, when pricing financial products, one needs to specify the form of SDEs driving the appropriate randomness and then estimate the parameters of interest in the equation to generate future scenarios; when measuring risks, one needs to calculate shocks based on these generated scenarios to obtain Value-at-Risk or Expected Shortfall. Therefore, this presumption of functional forms is always considered as a parametric method.

In the last two decades, nonparametric regression has attracted more academic and professional attention. The reason for this growing attention is that nonparametric regression is distribution-free, i.e. requiring little prior information on the data generation, so misspecification in the parametric method can be avoided. Now nonparametric regression has become a vital area in statistics (Härdle, 1990; Li and Racine, 2006) and it has many successful applications in finance (Campbell et al., 1988; Tsay, 2005). When nonparametric regression is applied in finance, an approach is to study the general SDE

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t \tag{1.1}$$

where $a(t, X_t)$ and $b^2(t, X_t)$ are the functions of our concern, called the drift and diffusion coefficients respectively[1]. They represent the expected return and volatility of the underlying variable, and are important factors to price assets, manage risks and choose portfolios. Through discretization of the equation, one can derive nonparametric estimators for the drift and diffusion coefficients by kernel methods (Fan and Gijbels, 1996) or smoothing splines (Wahba, 1990), and use these estimators to fulfill the financial purpose.

However, relaxing the assumption on the data generation in nonparametric regression does not come at no cost. In fact, compared to its parametric counterparts, nonparametric regression is computationally intensive. One reason is that each nonparametric estimator is the result of multiple local fits. Modern computers have drastically reduced the running time of the methods making them more practically available than ever before. But nonparametric regression uses the data themselves to tell the story, so larger sample sizes are always required to keep local structures for estimation. This implies that the computational cost is still quite large. In addition, nonparametric estimation methods have lower rate of convergence. This could be a serious impediment in some financial applications, where often time series exhibit nonstationary behavior that prevents us from using longer data sets. Thus we need new ways to lower the computational cost as well as make accurate estimates, which is our motivation of this thesis.

1.2 Ideas and Contributions

If we have the current data items of $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and must estimate the value of y at x where $x \neq x_i$ for $1 \leq i \leq n$, in this case we can use nonparametric regresion (such as kernel regression) to achieve this task. But at the moment a new observation (x_{n+1}, y_{n+1}) is available, then in order to obtain the real-time estimate, we have to use nonparametric regression again

[1]Sometimes $b(t, X_t)$ is called the diffusion (Fan and Zhang, 2003) but some works refer to $b^2(t, X_t)$ as the diffusion (Aït-Sahalia, 1996; Jiang and Knight, 1997). This thesis adopts the latter notation.

for all $n+1$ data items. It is clear that the complexity[2] of this procedure is at least $O(n)$. When a great deal of data are available sequentially and real-time estimation is required, it is not hard to imagine that nonparametric regression is not adequate to this job because the computational cost is quite large. This often happens in real cases. Financial institutions need to calculate Value-at-Risk and Expected Shortfall so as to meet the regulator's capital requirement, but in order to fulfill this task, they calibrate the parameters in the model by combining existing observations and the use of Monte Carlo simulation to generate scenarios. Often large financial institutions have very complicated portfolios. This means that they obtain millions of new observations each business day and so should use all their existing data plus the new data to repeat their calculations. Therefore how to obtain real-time results for a huge quantity of time series data is an important issue in practice.

However note that the previous estimate of the value of interest contains important historical information and can be used to derive the new one, so each time it is not necessary to use all current and all historical data. In this thesis, one of our main contributions is to propose an incremental way of computing nonparametric estimates for SDEs (called online methods[3]) and apply our methods in finance. In our methods, new data are used to update the previous estimate to yield the new one, hence the complexity of each update is $O(1)$ which is far better than $O(n)$ when n is large. Our methods can meet real-time demand in financial institutions.

This thesis studies the diffusion process as described in (1.1) and proposes online kernel estimators for the drift and diffusion. Here our main focus is on the diffusion process driven by Brownian motion instead of by other more general Lévy processes because such kind of processes are widely used in practice. We consider the time-homogeneous case, that is, the drift and diffusion do not depend on the time t directly

$$dX_t = a(X_t)dt + b(X_t)dW_t \tag{1.2}$$

In this thesis, we study both stationary and non-stationary processes. For example, GBM is a non-stationary time-homogeneous process; the Vasicek model, CIR model and CKLS model are stationary time-homogeneous processes. In addition, we theoretically prove quadratic convergence, strong consistency and asymptotic normality of online estimators for the stationary case, and weak consistency for the non-stationary case. By simulation we validate effectiveness and efficiency of our methods for both the stationary and non-stationary cases. We also test these new methods in real applications to calculate Value-at-Risk and Expected Shortfall for market risk management.

[2]In computer science, "big O" and "small o" are widely used notations to measure the computational complexity. Given two sequences $\{a_n\}$ and $\{b_n\}$, then $a_n = O(b_n)$ means $|a_n| \leq c|b_n|$ where c is some positive constant, whereas $a_n = o(b_n)$ means $a_n/b_n \to 0$ as $n \to \infty$. Thus $a_n = O(1)$ indicates a_n is bounded and $a_n = o(1)$ implies $a_n \to 0$. Additionally in probability theory, given a random variable sequence $\{X_n\}$, then $X_n = o_p(1)$ means X_n converges to zero in probability, i.e. $\lim_{n\to\infty} P(|X_n| > \varepsilon) = 0$ for any $\varepsilon > 0$.

[3]In order to differentiate, conventional nonparametric regression are called offline.

There are several things to note in this thesis. First, we propose online kernel-type estimators because in practice kernel methods are easy to implement. But similar ideas can be used to derive estimators of other types such as smoothing splines. Second, we are concerned with high-frequency data. With the development of modern technology, more data are measured every minute, so called one minute bars, and have become available than ever before. So it is of practical significance to propose online estimators for high-frequency data. Third, we build up estimators on a discrete-time sample of observations. Although the continuous sample path has been considered for many years (Rao, 1999), it is impossible to obtain with digital continuous-time observations in real applications. Thus our estimators are derived from discrete-time observations. Fourth, previous studies on nonparametric estimation of SDEs give offline estimators. To the best of our knowledge online nonparametric estimation has never been applied to SDEs, therefore our work bridges the gap between these areas and supplies feasible estimators for financial practice. Additionally, we give the rate of mean squared errors and asymptotic normality for further inference such as constructing confidence intervals.

1.3 Outline of This Thesis

This thesis is organized as follows. Chapter 2 begins by supplying basic concepts and measures in market risk management, then provides a brief introduction to kernel methods including kernel density estimators, kernel regression estimators and the bandwidth choice. We also discuss the applications of these methods in finance through empirical examples. Chapter 3 gives a literature review on recent studies of nonparametric estimation of SDEs. Some necessary preliminaries are included in the appendices. Next follows Chapter 4, in which online nonparametric estimators of the drift and diffusion in the stationary time-homogeneous diffusion process are developed. Quadratic convergence, strong consistency and asymptotic normality are also established for our proposed estimators. In Chapter 5, we give numerical examples to validate effectiveness and efficiency of our methods. In addition, an empirical case study of US 3-month treasury bill yields is also considered in this chapter, where Value-at-Risk and Expected Shortfall are calculated for financial practice. Chapter 6 studies online estimators in a non-stationary time-homogeneous process. After some transformation, the non-stationary process can be represented as a second-order SDE. Similarly weak consistency of estimators is also proved, and simulations and applications are also considered. The seventh and final chapter summarizes this thesis and points to ideas for further work.

Chapter 2

Kernel Methods in Risk Management

Since the 2007 financial crisis, the importance of the internal control has become clear not just to risk management but to the entire world. Yet risk management failures continue. For example, JP Morgan suffered large trading losses in 2012 for its ineffectiveness and failure of risk management in controlling trading activities, so-called the "London Whale" case[1]. A similar case occurred in 2013 at Everbright Securities, a Chinese Brokerage, because of the lack of risk management systems for monitoring trading errors[2]. On the other hand, there have been many studies of kernel methods in different areas such as finance (Fan and Yao, 2013), economics (Li and Racine, 2006) and meteorology (Xu, 2008), but little attention is paid to their applications for risk management. In this chapter, we first introduce the basic concepts and methods of both market risk management and kernel methods. Then we give an example to illustrate how to apply kernel methods to measure market risk.

2.1 Market Risk Management

As is shown in Figure 2.1, the prices of most financial products (e.g. stocks, bonds and their derivatives) fluctuate all the time. When financial institutions include these products into their portfolios, they have to use approaches to measure the risks of the exposure to these risk factors.

2.1.1 Risk Measures

Value-at-Risk (VaR) is a widely used measure of the market risk of losses on a portfolio. Let X denote the profit-and-loss (PnL) of a portfolio over a time period t, then the VaR of the portfolio is defined as follows:

$$VaR_\alpha = \sup\{-x : P_t(X > x) \leq \alpha\} \tag{2.1}$$

[1] See details in `http://en.wikipedia.org/wiki/2012_JPMorgan_Chase_trading_loss`.

[2] See details in `http://www.bloomberg.com/news/2013-08-16/everbright-securities-investigates-trading-system-error-1-.html`.

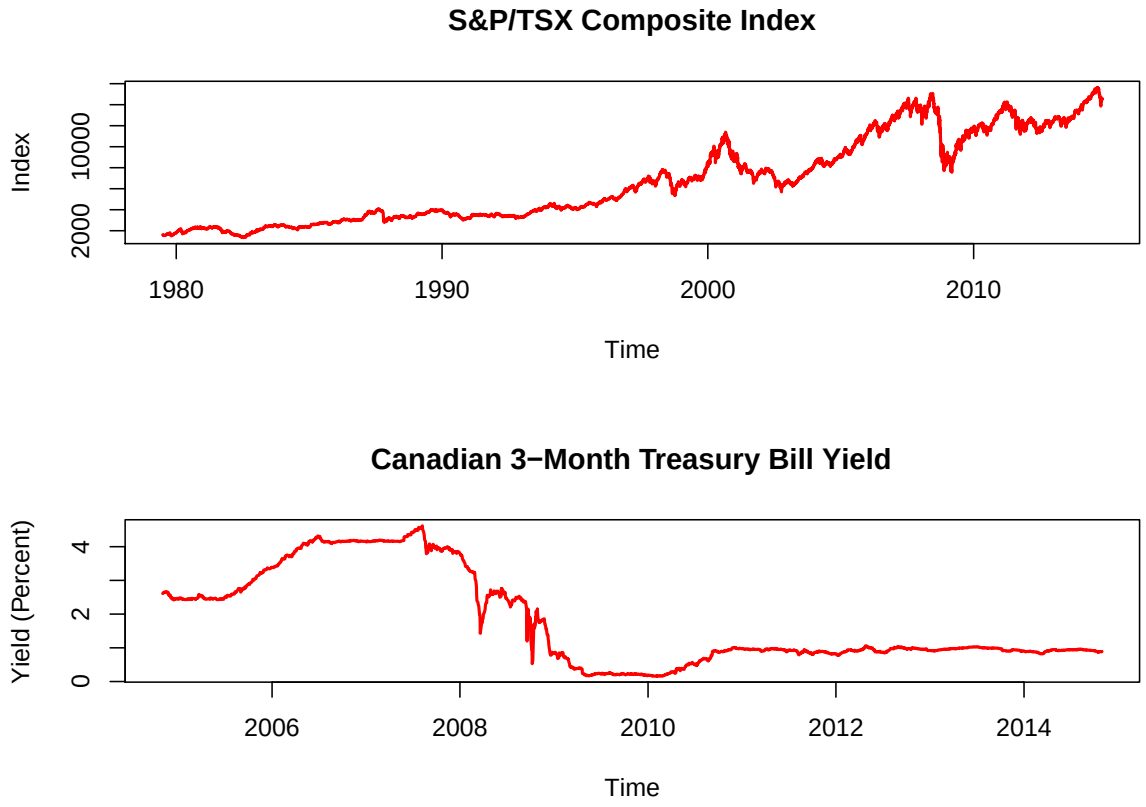

Figure 2.1: S&P/TSX Composite Index during the period from Jun 29, 1979 to Nov 2, 2014 (top) and Canadian 3-Month Treasury Bill Rate during the period from Nov 2, 2004 to Oct 31, 2014 (bottom).

where $\alpha \in (0, 1)$ is a significance level, e.g. taking 99% or 99.9%, see demonstration in Figure 2.2. From the definition, it can be seen that VaR estimation has the prediction of tail losses as its primary goal.

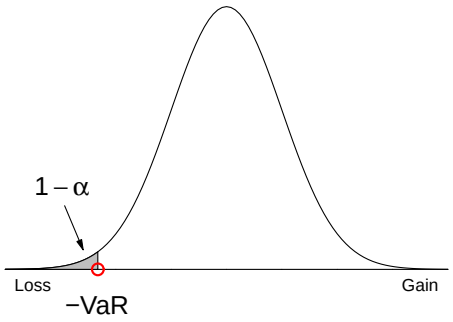

Figure 2.2: Demonstration of VaR, where the percentage of the shaded area is $1 - \alpha$.

As a risk measure, VaR describes how bad the PnL is likely to get, but is often criticized on the grounds that it is not sensitive to the shape of the PnL distribution's tail (Hull, 2012)

and not a coherent risk measure (Schied, 2006). Hence the Basel Committee suggest using the Expected Shortfall (ES) as an alternative. This ES metric is also called the conditional VaR and is given by

$$ES_\alpha = E[-X|X \leq -VaR_\alpha] \tag{2.2}$$

Different from VaR, the above definition of ES is about the expected loss given that the PnL is "bad". It can be found that ES is more sensitive to the shape of tail events because it provides more information on the tail.

Note that (2.1) and (2.2) are related to the distribution function of X, which unfortunately is usually unknown to us. In this case, it is often practical to use order statistics of X to calculate VaR and ES instead. Let $X_1, X_2, \ldots, X_n$ be n independently and identically distributed (i.i.d.) samples of X, and the k-th order statistic denoted by $X_{(k)}$. Then an empirical way to calculate VaR and ES (Hull, 2012) is listed as below

$$VaR_\alpha = -X_{(\lceil n(1-\alpha)\rceil)} \tag{2.3}$$

$$ES_\alpha = \frac{-1}{\lceil n(1-\alpha)\rceil} \sum_{i=1}^{\lceil n(1-\alpha)\rceil} X_{(i)} \tag{2.4}$$

For example, if there are $n = 100,000$ PnL scenarios, then 99.9% VaR is the 100th worst value and 97.5% ES is the average of the first 2,500 values in ordered PnL scenarios.

2.1.2 Approaches

The PnL distribution can be used to calculate VaR and ES according to (2.1) and (2.2). This part introduces two main approaches, historical simulation and the Monte Carlo method, about estimation of the PnL distribution based on a time series of observations. Let $\{X_t\}$ denote the time series of some risk factor with liquidity horizon τ and time horizon T. For example, the Basel Committee prescribe a liquidity horizon for the interest rate be 20 days and a time horizon be at least one year.

Historical Simulation

According to the work by Mehta et al. (2012), the majority of banks surveyed use historical simulation as their main approach. This is because of its greater simplicity with fewer simulations and more importantly, no additional presumption on the distribution of the asset returns. Thus it is considered as a nonparametric method. The procedure of historical simulation is listed as below (see demonstration in Figure 2.3)

Algorithm 1: Historical Simulation

input : A time series of observations $\{x_t\}$ where $t = 1, 2, \ldots, N$, liquidity horizon τ, time horizon T, and confidence levels α and β for VaR and ES respectively

output: A times series of $\{VaR_{\alpha,t}\}$ and $\{ES_{\beta,t}\}$

$i \leftarrow T + \tau$;

while $i \leq N$ **do**

  $j \leftarrow 1$;

  while $j \leq T - 1$ **do**

    Calculate the shock of the j-th scenario $\triangle x_j = x_{i-T+j} - x_{i-T+j-\tau}$;

    $j \leftarrow j + 1$;

  end

  Sort scenarios by increasing order and denote them by $\triangle x_{(1)}, \triangle x_{(2)}, \ldots, \triangle x_{(T-1)}$;

  Calculate $n_\alpha = \lceil (T-1)(1-\alpha) \rceil$ and $n_\beta = \lceil (T-1)(1-\beta) \rceil$;

  $VaR_{\alpha,i} \leftarrow -\triangle x_{(n_\alpha)}$ and $ES_{\beta,i} \leftarrow -\sum_{k=1}^{n_\beta} \triangle x_{(k)} / n_\beta$;

  $i \leftarrow i + 1$;

end

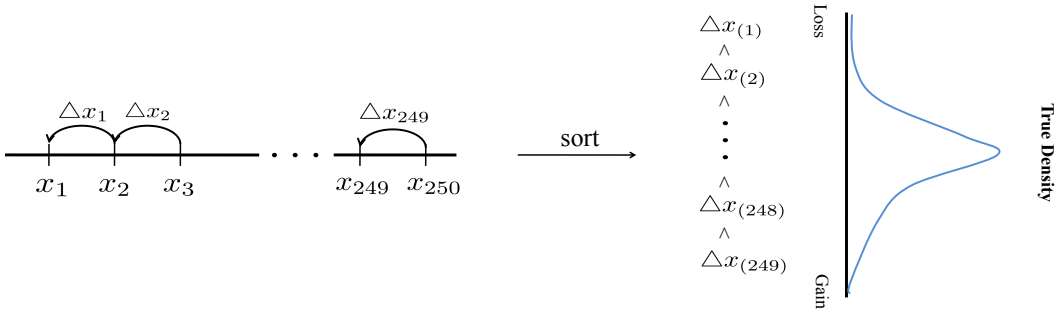

Figure 2.3: Demonstration of historical simulation, where $\tau = 1$, $T = 250$ and $N = 251$. So in this case, the 99% VaR is $\triangle x_{(3)}$ and 97.5% ES is the average of $\triangle x_{(1)}$ to $\triangle x_{(7)}$.

Algorithm 1 illustrates that the essential idea of historical simulation is recurrence of past events in the same way as before, i.e. by using previous scenarios to predict the PnL distribution. A breach event describes the case in which the actual PnL exceeds the estimated VaR (or ES) loss. The Basel Committee require that for every 250 trading days, the breaches are in green zone if the number is smaller than or equal to 4, in yellow zone if it is between 5 and 9, and in red zone if it is no smaller than 10. Financial institutions have the goal of developing justifiable statistical methods which, while accurately reflecting market reality, result in as few breaches as possible. This allows them to meet regulatory capital requirements with as little reserve capital as possible, hence increasing the return on equity on their balance sheets. We apply the procedure in Algorithm 1 to S&P/TSX Composite Index in Figure 2.1, and calculate the daily 99% VaR and the 97.5% ES to make comparison with true PnLs (see in Figure 2.4).

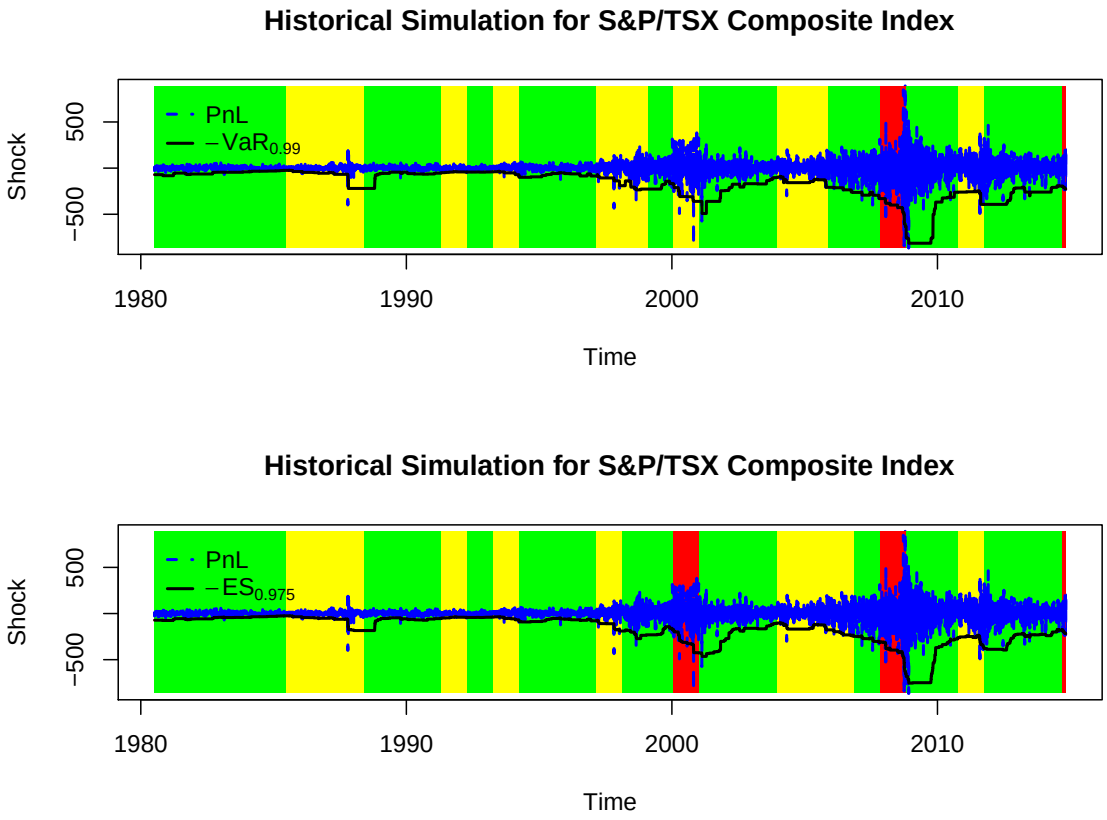

Figure 2.4: Historical simulation for S&P/TSX Composite Index where daily shocks are used for demonstration and 99% VaR and 97.5% ES are calculated to compare with PnL. The darkest bar represents the red zone, the lightest bar represents the yellow zone and the rest represents the green zone.

Despite its simplicity and distribution-free assumption, there are three restrictions of historical simulation. First, it uses equal weights for all PnLs. But more recent experience might be more important or, alternatively, experience observed at past times judged to be similar to the present in the business cycle might be deemed more important. Thus weighted historical simulation has been studied (Boudoukh et al., 1998). Second, historical simulation uses the past observations to yield its predictions. This means that the number of scenarios are limited to those actually experienced, a major restriction. Hence multiyear time horizon is often used (Mehta et al., 2012). Third, independent and identical distribution of the shocks is assumed in using historical simulation, but it could be violated for the shocks with overlapped time interval. So a filtered method has been devised for correlated data (Barone-Adesi et al., 1999).

Monte Carlo Method

Different from historical simulation, the Monte Carlo method needs to specify the underlying process for the risk factor and then uses observations to estimate the parameters in the model.

Therefore, it can provide a comprehensive picture of risks in the tail distribution. In addition, once the model is specified, as many scenarios as one likes can be obtained by the Monte Carlo method.

There is no general procedure for the Monte Carlo method because it is different from case to case. We use a simple example to illustrate its application to calculate VaR and ES, where GBM is used to model S&P/TSX Composite Index. The introduction to SDEs including GBM can be seen in Appendix A. For GBM $dX_t = \mu X_t dt + \sigma X_t dW_t$, the parameters are estimated by (A.13), that is,

$$\hat{\mu} = \frac{m + s^2/2}{\triangle} \quad \text{and} \quad \hat{\sigma} = \frac{s}{\sqrt{\triangle}}$$

where $m = n^{-1} \sum_{i=0}^{n-1} R_i$, $s^2 = (n-1)^{-1} \sum_{i=0}^{n-1} (R_i - m)^2$ and $R_i = \log(X_{i+1}) - \log(X_i)$ is the log-return over the time horizon as is considered. Then the procedure of the Monte Carlo method for GBM is listed as below

Algorithm 2: The Monte Carlo method for GBM

input : A time series of observations $\{x_t\}$ where $t = 1, 2, \ldots, N$, discretization step size $\triangle$, time horizon T, sample size n, and confidence levels α and β for VaR and ES respectively

output: A times series of $\{VaR_{\alpha,t}\}$ and $\{ES_{\beta,t}\}$

$i \leftarrow T + 1$;

while $i \leq N$ **do**

  $j \leftarrow 1$;

  while $j \leq T - 1$ **do**

    Calculate the log-returns $r_j = \log(x_{i-j}) - \log(x_{i-j-1})$;

    $j \leftarrow j + 1$;

  end

  Calculate $m_i = (T-1)^{-1} \sum_{k=1}^{T-1} r_k$ and $s_i^2 = (T-2)^{-1} \sum_{k=1}^{T-1} (r_k - m_i)^2$;

  Base on (A.13) to calculate $\hat{\mu}_i$ and $\hat{\sigma}_i$;

  Based on (A.9), generate samples $\hat{x}_{i,1}, \hat{x}_{i,2}, \ldots, \hat{x}_{i,n}$;

  Calculate the PnL $\triangle x_j = \hat{x}_{i,j} - x_{i-1}$;

  Calculate $n_\alpha = \lceil (T-1)(1-\alpha) \rceil$ and $n_\beta = \lceil (T-1)(1-\beta) \rceil$;

  $VaR_{\alpha,i} \leftarrow -\triangle x_{(n_\alpha)}$ and $ES_{\beta,i} \leftarrow -\sum_{k=1}^{n_\beta} \triangle x_{(k)} / n_\beta$;

  $i \leftarrow i + 1$;

end

It can be found that the core idea of the Monte Carlo method is to estimate parameters in the presumed model based on the past observations, and then generate samples and calculate PnLs for VaR and ES. Figure 2.5 demonstrates the application of the Monte Carlo method to calculation of VaR and ES for S&P/TSX Composite Index. At each time step, 100,000 PnL scenarios are simulated. Note that compared to the results by historical simulation in

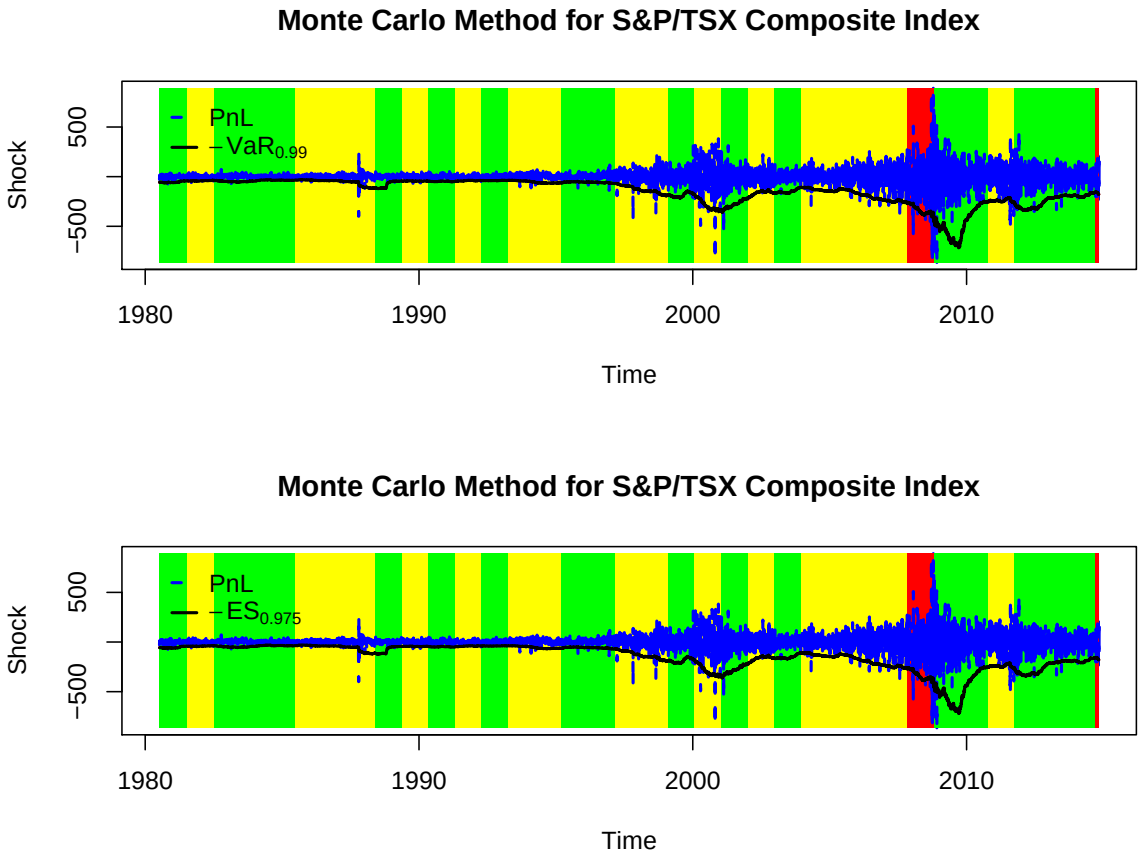

Figure 2.5: The Monte Carlo method for S&P/TSX Composite Index where 99% VaR and 97.5% ES are calculated to compare with PnL. The darkest bar represents the red zone, the lightest bar represents the yellow zone and the rest represents the green zone.

Figure 2.4, the Monte Carlo method gives more breaches of VaR and ES but a smaller amount of reserve capital (see in Figure 2.5). From the perspective of the regulators, they are more concerned with the number of breaches so as to avoid systemic risk[3].

Although the Monte Carlo method is regarded as the better theoretical approach, it suffers from two main problems. One of them is the computational complexity. For example, for each risk factor, 100, 000 simulated scenarios could be generated. As a result, the running time for a portfolio including thousands of factors is unacceptable. So in practice sampling for a longer period is often applied (Glasserman, 2003). Another problem is misspecification of the underlying process. The remedy could rely on one's prior experience and trial-and-error for new assumptions and models.

[3]Systemic risk refers to the risk that an event triggers a collapse of the financial system, whereas systematic risk refers to overall market risk.

2.2 Kernel Methods

As described above, financial risk managers are concerned with the distribution of risk factors, estimation of parameters or how these risk factors vary over time for the purpose of managing market risks.

To estimate the distribution, one can assume a specific form for the variables and validate the assumption by some criterion. For example, changes to the logarithm of the stock prices are usually assumed to be normally distributed, so graphical techniques (e.g. histogram and Q-Q plot) or statistical test (e.g. Jarque-Bera test and Shapiro-Wilk test) can be used to determine the normality of the distribution. To estimate the parameters, one can specify the model first, apply the maximum likelihood method and finally test the validity of the assumption. For example, the interest rate is supposed to follow the square-root CIR process but many researchers devote a great deal of attention to studying the plausible form of the drift and diffusion (Andersen and Lund, 1997; Conley et al., 1997; Chapman et al., 1999; Ang and Bekaert, 2002). To forecast future values, one can use the ARMA-GARCH model to fit the data and then make prediction based on the model, but the linearity or nonlinearity of the terms must be tested when the model is used.

Parametric methods as described above cannot avoid the problem of misspecification. Although changes to the logarithm of the stock prices are assumed to be normally distributed, these log returns are often observed to be skewed and to have fat tails. As a result, this misspecification may result in large estimation bias and the assumption of log-normality could be violated. In this case, one has to propose a new assumption and test it again. Also, except the CIR model and other descriptions of the dynamic, the actual forms to characterize the short-term interest rate are still on exploration. This trial-and-error process largely depends on one's prior experience. Therefore it is more or less inevitable to use the wrong form by parametric methods.

In such cases, the nonparametric technique could be an alternative to its parametric counterpart. One of its advantages is that instead of requiring prior information on specifying the parametric form, the technique lets the data speak of the appropriate functional form so that misspecification can be avoided. This is why nonparametric regression has received growing attention from academia and industry. But the nonparametric technique is not a panacea because they do result in higher computational costs. In addition, some criticize that the methods are "black-box" and lack intuitive interpretation. In fact, nonparametric and parametric methods are complementary to each other from a practical perspective.

Generally speaking, there are two branches of nonparametric regression: kernel regression and smoothing splines. The former is regarded as a local method in that the local polynomial is used to approximate the data, whereas the latter is considered as a global method in that a group of basis functions are constructed and smoothness is imposed globally. Compared to

smoothing splines, kernel regression is easy to implement and popular in real applications. So this thesis mainly focuses on kernel-type estimators.

2.2.1 Kernel Density Estimators

As mentioned above, one of the concerns with density estimation in finance is related to characterizing tail events such as calculation of VaR or ES. So in this part density estimation by kernel methods is introduced.

Suppose that the data $X_1, X_2, \ldots, X_n$ come from a common density function f. We know that the distribution function $F(x)$ is equal to $P(X \leq x)$, so the empirical distribution function is expressed as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where $I(\cdot)$ is the indicator function. If it is also assumed that X is continuous, then $F(x) = \int_{-\infty}^{x} f(u)du$, that is, $f(x) = dF(x)/dx$. Thus we can use an empirical distribution function to derive the estimator of the density, that is, for a small positive constant h_n

$$\hat{f}_n(x) = \frac{\hat{F}_n(x + h_n/2) - \hat{F}_n(x - h_n/2)}{h_n} = \frac{1}{nh_n} \sum_{i=1}^{n} I(|X_i - x| \leq h_n/2)$$

By letting $K(x) = I(|x| \leq 1/2)$, the above expression is rewritten as

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) \tag{2.5}$$

Here h_n is called the bandwidth (or smoothness parameter) and $K(\cdot)$ is called the kernel function. In addition to the uniform kernel $K(x) = I(|x| \leq 1/2)$, we can also take other forms for $K(\cdot)$, which results in different estimators. Informally, $K(\cdot)$ is the kernel function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$. $K(\cdot)$ is non-negative if $K(x) \geq 0$ and symmetric if $K(x) = K(-x)$. The j-th moment of $K(\cdot)$ is defined as $m_j = \int_{-\infty}^{\infty} x^j K(x)dx$ and the order of $K(\cdot)$ is defined as $\inf\{j : m_j \neq 0\}$. For example, if $m_1 = 0$ and $m_2 > 0$, then $K(\cdot)$ is a second-order kernel. It can be verified that the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is a symmetric, non-negative and second-order kernel function; this kernel is commonly used in practice. The Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)_+$ is also commonly used in real applications.

2.2.2 Kernel Regression Estimators

Regression analysis is one of the most widely used statistical tools to estimate the relationships among variables. For example, in the capital asset pricing model, we can use simple linear regression to estimate the parameter (namely β) in the model. In addition, regression can be

used to assist interpolation and extrapolation for missing data in some data sources.

Given the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the simple linear regression model is given by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where β_0, β_1 are parameters, and the errors $\varepsilon_1, \varepsilon_1, \ldots, \varepsilon_n$ satisfy $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. More generally, one would like to specify nonlinear relationships such as the exponential or logarithm function between X and Y. But as mentioned above, it is rare to know the true functional form in real applications, so lack of prior information could lead to inconsistent estimation by the presumed model. Nonparametric regression avoids this problem by freeing the assumption of the functional form about the data generation process. In the nonparametric regression model, we are concerned about estimation of $g(x)$ such that

$$Y_i = g(X_i) + \varepsilon_i \tag{2.6}$$

where $g(\cdot)$ satisfies some regularity conditions such as smoothness and moment conditions.

We can use Taylor expansion to approximate $g(X_i)$ in a neighborhood of x

$$g(X_i) = g(x) + (X_i - x)g'(x) + \frac{1}{2}(X_i - x)^2 g''(x) + o\left(|X_i - x|^2\right)$$

The functional form of $g(\cdot)$ is usually unknown, so by letting $\beta_{j,x}$ denote the estimate of the j-th order derivative of $g(x)$, kernel regression uses the weighted least squares technique to fit the data at the neighborhood of x by a polynomial of degree p, which minimizes

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_{j,x} (X_i - x)^j \right)^2 K_{h_n}(X_i - x) \tag{2.7}$$

where $K_{h_n}(x) = K(x/h_n)/h_n$. Thus the estimator $\hat{\beta}_{0,x}$ gives the estimated value of $g(x)$, and similarly we can approximate the j-th order derivative of $g(x)$ by $\hat{\beta}_{j,x}$.

Moreover let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^p \\ 1 & (X_2 - x) & \ldots & (X_2 - x)^p \\ \ldots & \ldots & \ldots & \ldots \\ 1 & (X_n - x) & \ldots & (X_n - x)^p \end{pmatrix} \qquad \boldsymbol{\beta}_x = \begin{pmatrix} \beta_{0,x} \\ \beta_{1,x} \\ \vdots \\ \beta_{p,x} \end{pmatrix} \tag{2.8}$$

and $\boldsymbol{W}_x = \text{Diag}\{K_{h_n}(X_1 - x), K_{h_n}(X_2 - x), \ldots, K_{h_n}(X_n - x)\}$, then the least-squares estimate to (2.7) at x is given by

$$\hat{\boldsymbol{\beta}}_x = \left(\boldsymbol{X}^T \boldsymbol{W}_x \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y} \tag{2.9}$$

thus the estimator of $g(x)$ is

$$\hat{g}_n(x) = e_1^T \hat{\boldsymbol{\beta}}_x = e_1^T \left(\boldsymbol{X}^T \boldsymbol{W}_x \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}$$

where $e_1^T = (1, 0, \ldots, 0)^T$. When $p = 0$ in (2.7), the estimator is reduced to be locally constant (also known as the Nadaraya-Watson estimator) proposed by Nadaraya (1964) and Watson (1964), which can be rewritten as

$$\hat{g}_n(x) = \frac{\sum_{i=1}^{n} Y_i K_{h_n}(X_i - x)}{\sum_{i=1}^{n} K_{h_n}(X_i - x)} \tag{2.10}$$

The local linear estimator, i.e. $p = 1$ in (2.7), is proposed by Fan (1993) with the closed form

$$\hat{g}_n(x) = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i + n^{-2}}$$

where $w_i = K\left(\frac{x - X_i}{h_n}\right)[s_{n,2} - (x - X_i)s_{n,1}]$ with

$$s_{n,j} = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)(x - X_i)^j$$

For general p, there is no such closed form for $\hat{g}_n(x)$. Fan and Gijbels (1996) gave a detailed study of this general case but advocated that the local linear estimator is enough in practice.

Moving beyond the local constant and local linear estimators, Hall et al. (1999) and Cai (2002) proposed a weighted Nadaraya-Watson estimator given by

$$\hat{g}_n(x) = \frac{\sum_{i=1}^{n} p_i(x) Y_i K_{h_n}(X_i - x)}{\sum_{i=1}^{n} p_i(x) K_{h_n}(X_i - x)}$$

where $p_i(x)$ are weights and satisfy: (1) $p_i(x) \geq 0$; (2) $\sum_{i=1}^{n} p_i(x) = 1$; (3) $\sum_{i=1}^{n} p_i(x) K_{h_n}(X_i - x) = 0$. Hall et al. (1999) and Cai (2002) proved that the weighted Nadaraya-Watson estimator and the local linear estimator have the same asymptotic distribution.

Although the Nadaraya-Watson estimator is easy to implement and popular in practice, it suffers from the boundary effect, i.e. its bias has lower order at the boundary than in the interior domain. From the definition of the Nadaraya-Watson estimator, it can be found that the estimator can make use of two-sided sample in the interior but only a one-sided sample can be used at the boundary. Thus lack of sample results in the big bias of the estimator at the boundary of the

domain, which is so called the boundary effect. To overcome this problem, many methods have

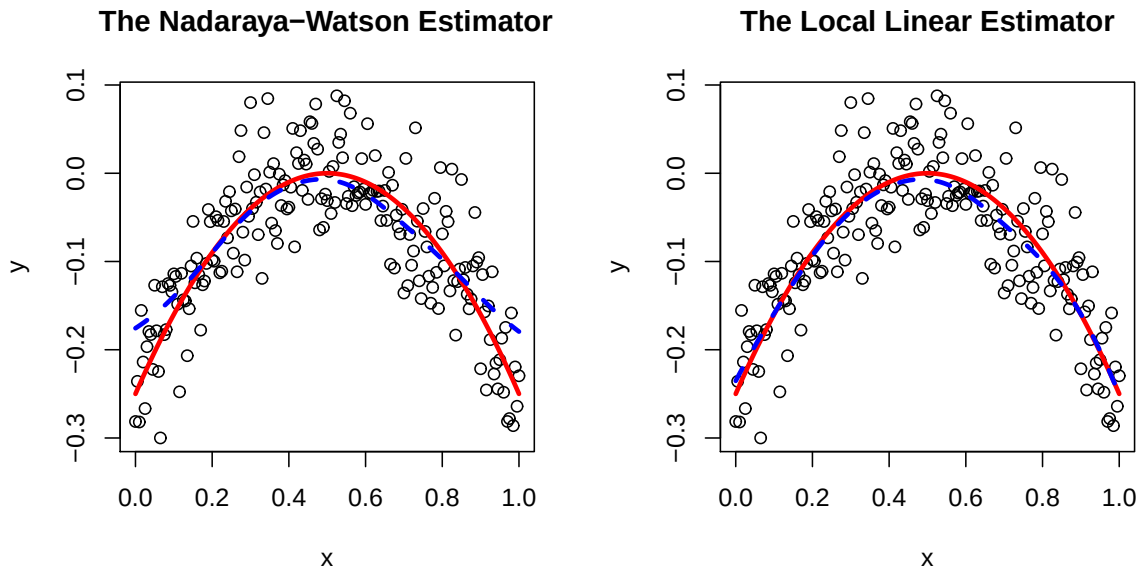

Figure 2.6: Boundary effect for the Nadaraya-Watson estimator and the local linear estimator, where the data are generated by $y = -(x-0.5)^2 + 0.05\varepsilon$ with $\varepsilon \sim N(0,1)$. The solid line is the true curve $y = -(x-0.5)^2$, and the dotted line gives the fitted values.

been proposed to remove the boundary bias such as the geometrical method (Hall and Wehrly, 1991) and the boundary correction approach (Gray and Schucany, 1972; Rice, 1984). In addition, the local linear estimator and the weighted Nadaraya-Watson estimator as mentioned above have automatically boundary adaption (see in Figure 2.6).

2.2.3 Choice of Bandwidth and Kernel Function

Relative to the kernel function, the choice of the bandwidth h_n is critical to the performance of the estimator (Wand and Jones, 1995). From the definition of the kernel function, it can be found that the bandwidth controls how many sample points are included in estimation. In fact, there is a tradeoff between bias and variance (see Figure 2.7). Larger h_n will include more sample points such that the estimator is not sensitive to the randomness, so the variance can be reduced. But in this case the estimator tends to be further away from those local points, as a result there is a larger bias. Similarly smaller h_n will result in smaller bias but larger variance. So it is very important to have a reliable choice of bandwidth which trades off between these two extremes.

In theoretical and practical settings, several approaches of choosing the h_n constant have been proposed. One of them is cross-validation (CV), a fully automatic data-driven technique proposed by Rudemo (1982), Stone (1984) and Bowman (1984). Note that different h_n's corre-

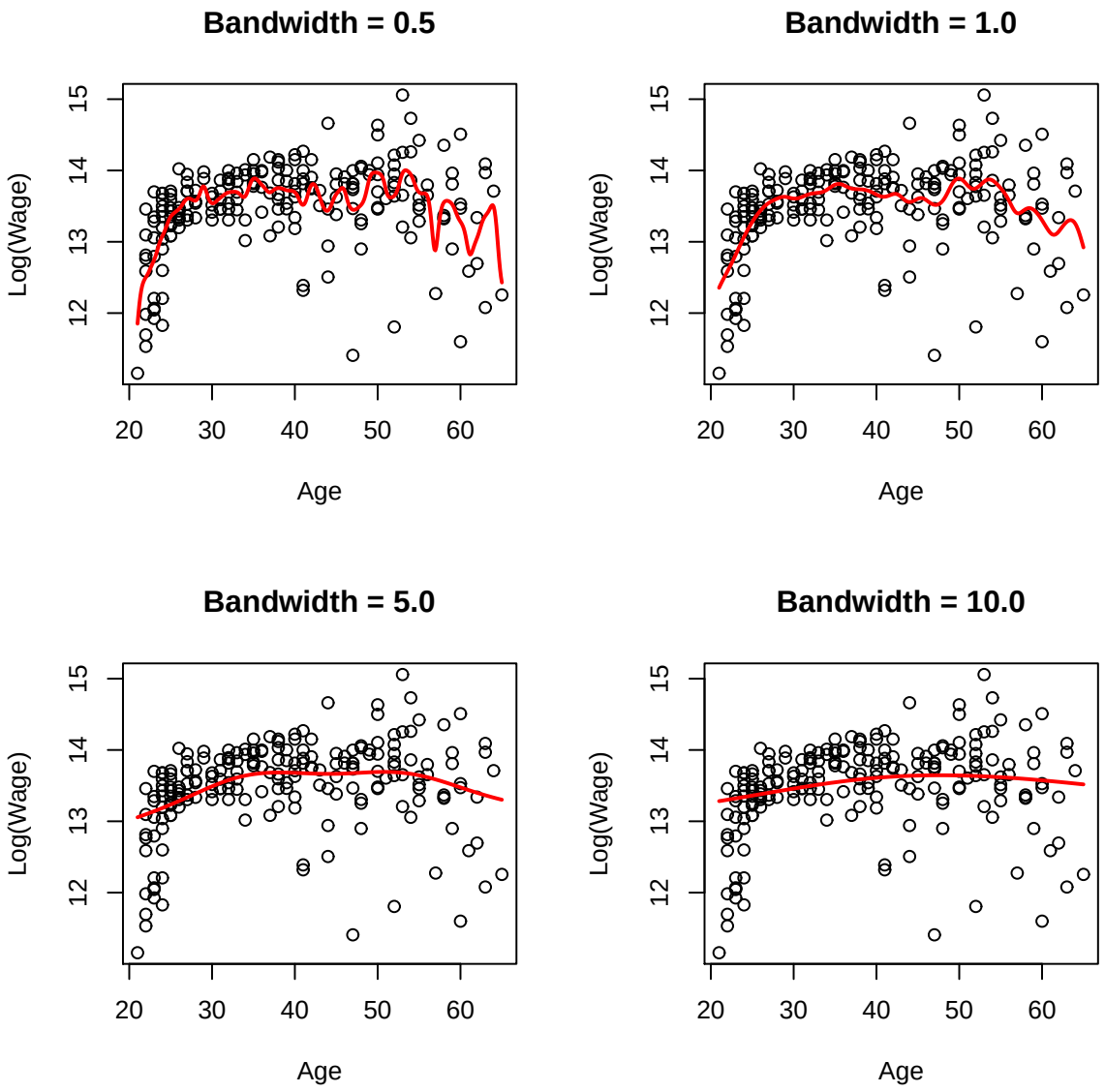

Figure 2.7: Four Nadaraya-Watson estimators for Canadian male wage data on 1971, where the bandwidth are $h_n = 0.5, h_n = 1.0, h_n = 5.0$ and $h_n = 10.0$.

spond to a variety of estimators, and the goal of CV is to validate how accurately the estimator will perform in real applications, so one can use CV to choose the optimal h_n based on some criterion such as mean squared errors (MSE). This technique begins by splitting the whole set of observations into the training dataset and the testing dataset. The motivation is that usually we can only access the dataset with a limited size which could be far smaller than expected. One simple way is to use the entire data for the training and testing purpose, but this will lead to overfitting. So in this case it is better to split the data into subsets. Then for each h_n, the estimator derived from the training dataset is evaluated on the testing dataset and the optimal bandwidth could be chosen based on evaluation. For example, given the nonparametric model

(2.6), each time the leave-one-out CV uses one observation as the testing dataset and the remaining as the training dataset, and repeats until each observation is used to test (see Figure 2.8). In other words, given the fixed bandwidth h_n, let $\hat{Y}_{-i}$ denote the fitted value on X_i obtained by the trained model without considering X_i, i.e. $\hat{Y}_{-i} = \hat{g}_{n,-i}(X_i)$ where $\hat{g}_{n,-i}(x)$ is constructed on the entire dataset excluding (X_i, Y_i). For the case of the Nadaraya-Watson estimator, $\hat{g}_{n,-i}(x)$ is given by

$$\hat{g}_{n,-i}(x) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j K_{h_n}(X_j - x)}{\sum_{\substack{j=1 \\ j\neq i}}^{n} K_{h_n}(X_j - x)}$$

After that, the optimal bandwidth is chosen as

$$h_n^{\text{opt}} = \arg\min_{h_n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{-i})^2$$

In addition, the leave-one-out CV can be generalized to the leave-p-out CV (Shao, 1993; Zhang, 1993), and partial data splitting schemes have been proposed in practice including k-fold CV introduced by (Geisser, 1975).

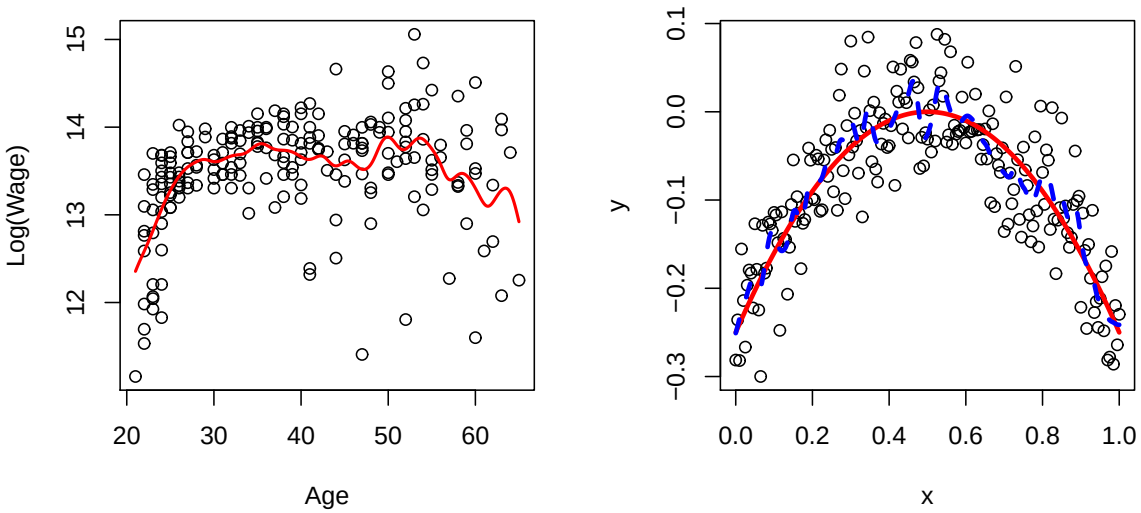

Figure 2.8: The leave-one-out CV is performed on the same Canadian male wage data seen in Figure 2.7 (left) and $y = -(x - 0.5)^2$ (right). The optimal bandwidth is $h_n^{\text{opt}} = 1$ for the left and $h_n^{\text{opt}} = 0.01$ for the right.

Other ways to select the constant bandwidth include the rule-of-thumb approach (Silverman, 1986; Li and Racine, 2006). For the kernel density estimator, suppose that the l-th order

kernel is used, then it can be calculated that

$$\begin{aligned} \text{Bias}[\hat{f}_n(x)] &= \frac{1}{l!} f^{(l)}(x) h_n^l m_l + o(h_n^l) \\ \text{Var}[\hat{f}_n(x)] &= \frac{f(x) \int_R K^2(u)du}{nh_n} + O\left(\frac{1}{n}\right) \end{aligned}$$

which leads to the asymptotic mean squared error (AMSE) given by

$$\text{AMSE}[\hat{f}_n(x)] = \left[\frac{1}{l!} f^{(l)}(x) h_n^l m_l\right]^2 + \frac{f(x) \int_R K^2(u)du}{nh_n}$$

with this definition the asymptotic mean integrated squared error (AMISE) can be calculated as

$$\text{AMISE}[\hat{f}_n(x)] = \int_R \text{AMSE}[\hat{f}_n(x)]dx = \left[\frac{h_n^l m_l}{l!}\right]^2 \int_R [f^{(l)}(u)]^2 du + \frac{\int_R K^2(u)du}{nh_n}$$

Note that the above AMISE is a function of h_n, which implies that the optimal h_n can be taken to minimize the AMISE. By taking the derivative of AMISE with respect to h_n and setting it to zero, we have

$$h_n^{\text{opt}} = \left\{\frac{(l!)^2 \int_R K^2(u)du}{2l m_l^2 \int_R [f^{(l)}(u)]^2 du}\right\}^{1/(2l+1)} \times n^{-1/(2l+1)}$$

It is noted that h_n^{opt} is related to $\int_R [f^{(l)}(u)]^2 du$ but $f(x)$ is an unknown function. So Silverman (1986) suggested to use a plausible candidate such as the normal density to replace $f(x)$. This results in the rule-of-thumb for the bandwidth $h_n^{\text{opt}} = C\hat{\sigma} n^{-1/(2l+1)}$ where C is some constant and $\hat{\sigma}$ is the sample standard deviation. If the standard normal kernel is used, the optimal bandwidth is $\hat{\sigma} n^{-1/5}$.

Note that given the number of observations n, the bandwidth in (2.7) is constant, neither incorporating the location of x nor that of X_i. As Fan and Gijbels (1996) pointed out, this constant bandwidth may not estimate curves with a complicated shape very well. Thus they introduced variable bandwidth $h/\alpha(X_i)$ such that (2.7) can be written as

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j\right)^2 K_{h_n/\alpha(X_i)}(X_i - x)$$

where $\alpha(\cdot)$ is positive and reflects the difference of each data point. Then AMISE is minimized to obtain the optimal bandwidth h_n and $\alpha(\cdot)$. They found that by using variable bandwidth, AMISE can be reduced more than by using constant bandwidth. Meanwhile the local linear regressor with variable bandwidth shares the advantage of having no boundary effect. Similar ideas have also been applied to kernel regression (Müller and Stadtmüller, 1987; Schucany, 1995).

Recently the choice of kernel functions has also been studied because it is found that classical methods with symmetric kernels have significant bias errors on the boundary (Mackenzie and Tieu, 2004). Michels (1992) used an asymmetric gamma kernel function to reduce the bias in estimation, and found that the use of asymmetric kernels can lead to better predictions in a time series model for environmental data. Chen (2002b) used asymmetric kernels in local linear regression and claimed that the flexible shape of asymmetric kernels supplies advantages of having finite variance and resistance to sparse design. Abadir and Lawford (2004) studied the class of optimal asymmetric kernels in the sense of the mean integrated squared error (MISE) and analyzed its main properties.

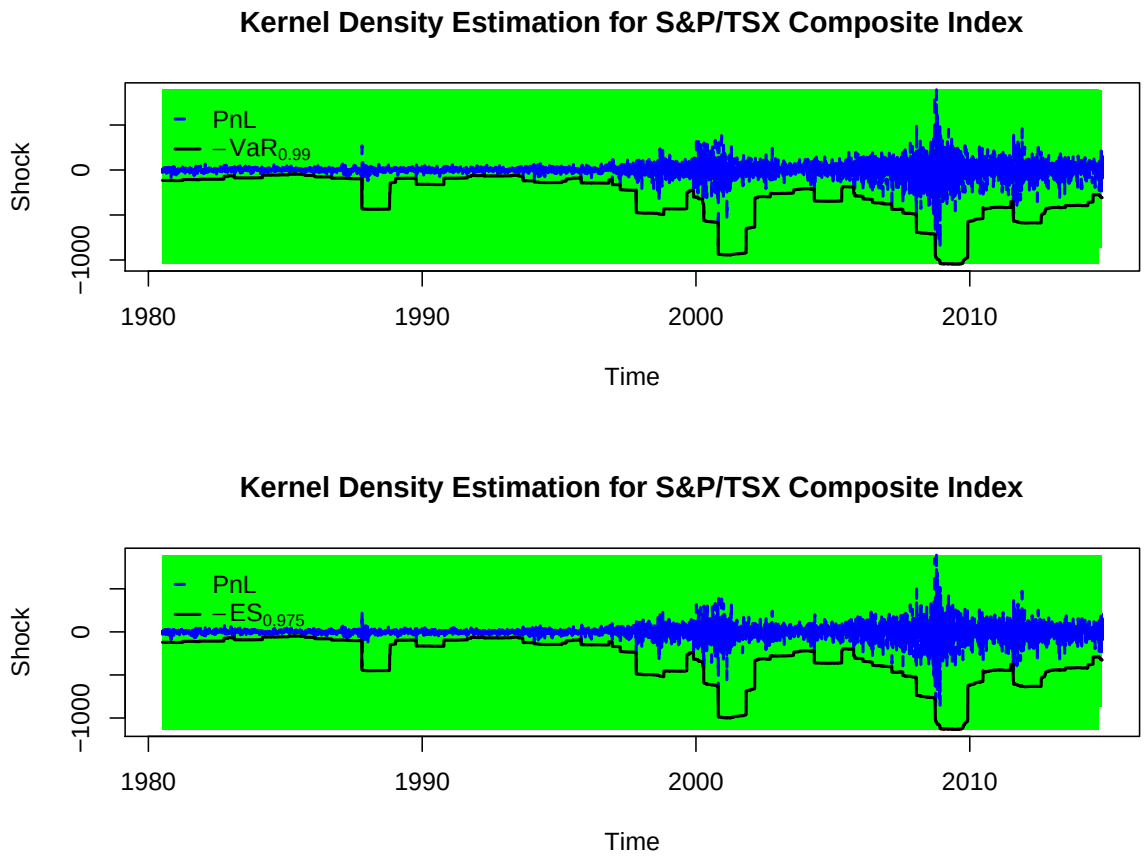

Figure 2.9: Historical simulation for S&P/TSX Composite Index where daily shocks are used for demonstration and 99% VaR and 97.5% ES are calculated to compare with PnL. All trading days are in green zone.

2.3 Example

In this section, a simple example is provided to demonstrate the application of kernel methods in market risk management. As we have mentioned, the essential step in calculating VaR and

ES is to predict the PnL distribution, so the kernel density estimator (2.5) can be used to fulfill the task. For illustration, historical simulation with the same parameters as in Figure 2.4 is used to generate the predicted shocks. Then instead of (2.3) and (2.4), we apply (2.5) to estimate the density of these shocks and make use of the original definition (2.1) and (2.2) to find VaR and ES. Figure 2.9 illustrates this application. It is noted in this case that for comparison with Figure 2.4, the results given by the kernel density estimator are more conservative as there are no breaches in the whole period. But we must admit that financial institutions might not be satisfied with these results because they imply higher capital requirements and so a lower return on equity.

In this chapter, we have discussed the industrial practice of calculating risk measures and the application of kernel methods to estimate the distribution of the profit and loss. The next chapter will provide a broader picture of the methodology by reviewing the recent developments of kernel regression in finance.

Chapter 3

Literature Review

In the last few decades, nonparametric regression has attracted growing attention from academia and industry because it requires little prior information on the process which generates the data. Nonparametric estimation for continuous-time models has been studied for many years (Tuan, 1981; Rao, 1985; Soulier, 1998; Spokoiny, 2000; Papaspiliopoulos et al., 2012), but the assumption of continuous time is unreasonable in real applications. Therefore the purpose of this chapter is to review recent developments of nonparametric regression based on discrete-time observations for estimation of the drift and diffusion in stochastic differential equations (SDEs). For those wishing a review, a brief introduction of SDEs, in particular diffusion processes, is provided in Appendix A.

3.1 Time-Homogeneous Diffusion Models

Florens-Zmirou (1993) left the drift restriction-free and proposed the following kernel estimator for the time-homogeneous diffusion by using the uniform kernel

$$\hat{b}_n^2(x) = \frac{\sum_{i=1}^{n-1} I(|X_i - x| < h_n) n (X_{i+1} - X_i)^2}{\sum_{i=1}^{n} I(|X_i - x| < h_n)}$$

For high-frequency data, i.e. in the limit as the discretization step size tends to zero, the author proved quadratic convergence and asymptotic normality of the estimator by expanding the transition density. By assuming the mean-reverted drift, Aït-Sahalia (1996) proposed a nonparametric estimator for the diffusion in the time-homogeneous case by using the Kolmogorov forward equation with time-stationary transition density, i.e. $\frac{\partial p(X_{t+h}|X_t)}{\partial t} = 0$, where $p(X_{t+h}|X_t)$ is the transition density of X_{t+h} given X_t in (A.5) in Appendix A. The Kolmogorov forward

equation can yield

$$b^2(X_t) = \frac{2}{p(X_t)} \int_0^{X_t} a(y)p(y)dy$$

where $p(y)$ is the stationary density function of the time series $\{X_t\}$ and can be approximated by the kernel density estimator as mentioned in Chapter 2. Note that the Kolmogorov forward equation with the time-stationary transition density can also provide a relationship

$$a(X_t) = \frac{1}{p(X_t)} \frac{d}{dX_t}[b^2(X_t)p(X_t)] \tag{3.1}$$

so Jiang and Knight (1997) first proposed a kernel estimator of $b^2(x)$ given by

$$\hat{b}^2(x) = \frac{\sum_{i=1}^{n-1} \frac{(X_{i+1}-X_i)^2}{\triangle} K_{h_n}(X_i - x)}{\sum_{i=1}^{n} K_{h_n}(X_i - x)}.$$

Based on (3.1) an estimator of $a(x)$ is proposed by

$$\hat{a}(x) = \frac{1}{p(x)} \frac{d}{dx}[b^2(x)p(x)] = \frac{1}{2}\left[\frac{d\hat{b}^2(x)}{dx} + \hat{b}^2(x)\frac{\hat{p}'(x)}{\hat{p}(x)}\right]$$

where $\hat{p}(x)$ is the estimator of the density function given by $\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(X_i - x)$. The central limit theorems for $\hat{a}(x)$ and $\hat{b}^2(x)$ are established and valid conditionally on the path passing through x. The methods are also applied to estimation of the short-term interest rate. Arapis and Gao (2006) specified Jiang and Knight's method by using a Gaussian kernel function to derive the closed form of the estimators of $a(x)$ and $b^2(x)$.

Stanton (1997) applied the infinitesimal generator (Øksendal, 2003) of the time-homogeneous diffusion model to expand the function $f(t, X_t)$ of t and X_t, where

$$\mathcal{L}f(t,x) = \lim_{\tau \downarrow t} \frac{E(f(\tau, X_\tau)|X_t = x) - f(t,x)}{\tau - t} = \frac{\partial f(t,x)}{\partial t} + a(x)\frac{\partial f(t,x)}{\partial x} + \frac{1}{2}b^2(x)\frac{\partial^2 f(t,x)}{\partial x^2}.$$

Then $E[f(t+\triangle, X_{t+\triangle})|X_t]$ can be expressed in the form of a Taylor expansion

$$E[f(t+\triangle, X_{t+\triangle})|X_t] = f(t,X_t) + \triangle\mathcal{L}f(t,X_t) + \frac{1}{2}\triangle^2\mathcal{L}^2 f(t,X_t) + \cdots + \frac{1}{n!}\triangle^n\mathcal{L}^n f(t,X_t) + O(\triangle^{n+1})$$

Then the first and second order approximations are

$$\mathcal{L}f(t,x) = \frac{1}{\triangle}E[f(t+\triangle, X_{t+\triangle}) - f(t,X_t)|X_t = x] + O(\triangle)$$

$$\mathcal{L}f(t,x) = \frac{1}{2\triangle}\{4E[f(t+\triangle, X_{t+\triangle}) - f(t,X_t)|X_t = x] - E[f(t+2\triangle, X_{t+2\triangle}) - f(t,X_t)|X_t = x]\} + O(\triangle^2)$$

By taking $f(t,x) = x$, implying $\mathcal{L}f(t,x) = a(x)$, first and second order approximations of $a(x)$ are

$$\begin{aligned}\tilde{a}(x) &= \frac{1}{\triangle}E[X_{t+\triangle} - X_t|X_t = x] + O(\triangle)\\ \tilde{a}(x) &= \frac{1}{2\triangle}\{4E[X_{t+\triangle} - X_t|X_t = x] - E[X_{t+2\triangle} - X_t|X_t = x]\} + O(\triangle^2).\end{aligned}$$

By taking $f(t,x) = (x - X_t)^2$, implying $\mathcal{L}f(t,x) = 2a(x)(x - X_t) + b^2(x)$ so $\mathcal{L}f(t,X_t) = b^2(X_t)$, first and second order approximations of $b^2(x)$ can be given by

$$\begin{aligned}\tilde{b}^2(x) &= \frac{1}{\triangle}E[(X_{t+\triangle} - X_t)^2|X_t = x] + O(\triangle)\\ \tilde{b}^2(x) &= \frac{1}{2\triangle}\left\{4E[(X_{t+\triangle} - X_t)^2|X_t = x] - E[(X_{t+2\triangle} - X_t)^2|X_t = x]\right\} + O(\triangle^2),\end{aligned}$$

where $\tilde{a}(x)$ and $\tilde{b}^2(x)$ can be estimated using data $X_i, i = 0,\ldots,n$. Stanton applied the above methods to estimate the drift and diffusion of the short-term rate and the market price of the interest rate risk by using the daily three- and six-month treasury bill data. The author claimed that higher order estimation should outperform lower order estimation. However, Fan and Zhang (2003) found that this claim may not always hold true. They extended Stanton's method and gave the general high order estimation of the drift and diffusion coefficients by adding weights

$$\mathcal{L}f(t,X_t) = \frac{1}{\triangle}\sum_{i=1}^{k} a_{k,i}E_t[f(t + i\triangle, X_{t+i\triangle}) - f(t,X_t)] + O(\triangle^k)$$

where $a_{k,i} = (-1)^{i+1}\binom{k}{i}/i$. Then local polynomial regression is used to derive the estimator of $a(x)$ and $b^2(x)$, and the asymptotic behaviors are obtained. They found that the asymptotic biases of the higher order estimators can be reduced but the asymptotic variances increase with the order of the estimator. However, the nonnegativity of the local linear estimator of the diffusion cannot be guaranteed, so some researchers proposed different methods to overcome this shortcoming of local polynomial estimation. One method is through logarithmic transformation to retain nonnegativity (Ziegelmann, 2002), i.e. the estimator of the diffusion is given by

$$(\hat{\beta}_0,\hat{\beta}_1) = \arg\max_{\beta_0,\beta_1}\sum_{i=1}^{n-1}\left[\frac{(X_{i+1} - X_i)^2}{\triangle} - e^{\beta_0+\beta_1(x-X_i)}\right]^2 K_{h_n}(x - X_i)$$

Similar ideas can be also found in (Yu and Jones, 2004). In Yu and Jones's method, one can use the Euler scheme to approximate the time-homogeneous case by (A.8), that is,

$$X_{i+1} = X_i + a(X_i)\triangle_i + b(X_i)\sqrt{\triangle_i}\varepsilon_i \tag{3.2}$$

Let $Y_i = X_{i+1} - X_i$, then the likelihood function for the process (3.2) is written as

$$\prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi b^2(X_i)\triangle_i}} \exp\left\{-\frac{[Y_i - a(X_i)\triangle_i]^2}{2b^2(X_i)\triangle_i}\right\}$$

and the log-likelihood function is proportional to

$$\sum_{i=1}^{n-1}\left[\log(b^2(X_i)\triangle_i) + \frac{(Y_i - a(X_i)\triangle_i)^2}{b^2(X_i)\triangle_i}\right].$$

By letting $a(X_i) = \alpha_0 + \alpha_1(x - X_i)$ and $\log b^2(X_i) = \beta_0 + \beta_1(x - X_i)$, one can propose kernel estimators of the drift and diffusion and use optimization procedures to determine the parameters $\alpha_0, \alpha_1, \beta_0$ and β_1. Note that the above logarithmic transformation can retain nonnegativity of the estimator of $b^2(x)$. But the transformation introduces an extra bias such that the resulting estimators may not have closed-form representation. Cai (2001) proposed a weighted Nadaraya-Watson estimator to capture both advantages from local constant and local polynomial methods by defining the weights

$$w_i(x) \geq 0 \quad \sum_{i=1}^{n} w_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n}(X_i - x)w_i(x)K_{h_n}(x - X_i) = 0$$

so one can use the weighted version of the Nadaraya-Watson method to estimate the diffusion by

$$\hat{b}^2(x) = \frac{\sum_{i=1}^{n-1} w_i(x)K_{h_n}(x - X_i)\frac{(X_{i+1}-X_i)^2}{\triangle}}{\sum_{i=1}^{n} w_i(x)K_{h_n}(x - X_i)}$$

then the constrained optimization technique is used to determine the weights. Xu (2010) extended Cai's method in more general settings. In addition, Arfi (2008) proved that the estimators by Stanton (1997) have strong consistency under some regular conditions.

For low-frequency data, with the fixed discretization step size, Nicolau (2003) quantified the bias of the Florens-Zmirou (1993) and Jiang and Knight (1997) estimators for the diffusion in the time-homogeneous case. Meanwhile, based on the quantified bias, Nicolau proposed a bias adjustment method to partially attenuate the distortion. In addition, weak consistency and asymptotic normality are obtained for the estimators. Gobet et al. (2004) proposed kernel estimators of the drift and diffusion under the assumption of ergodicity and proved quadratic convergence by the spectral analysis of the associated Markov semigroup.

However as pointed by Phillips (1973) and Hansen and Sargent (1983), it is harder to estimate the drift than the diffusion except by imposing stronger conditions on the drift. This is similar to the well-known "aliasing problem", that is, different continuous-time paths may

be indistinguishable in discrete time points. This is seen by the Cameron-Martin-Girsanov transformation to give an unnoticeable change in the drift. So the estimates of the drift do not have the same precision as those of the diffusion. In order to overcome this problem and obtain accurate estimates, Bandi and Phillips (2003) proposed estimators of the drift and diffusion as $\triangle \to 0$ and $n\triangle \to \infty$ given by

$$\hat{a}(x) = \frac{\sum_{i=1}^{n} K_{h_n}(x - X_i)\tilde{a}(X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)} \quad \text{and} \quad \hat{b}^2(x) = \frac{\sum_{i=1}^{n} K_{h_n}(x - X_i)\tilde{b}^2(X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)}$$

where

$$\tilde{a}(x) = \frac{\sum_{i=1}^{n-1} I(|x - X_i| < b_n)\frac{(X_{i+1} - X_i)}{\triangle}}{\sum_{i=1}^{n} I(|x - X_i| < b_n)} \quad \text{and} \quad \tilde{b}^2(x) = \frac{\sum_{i=1}^{n-1} I(|x - X_i| < b_n)\frac{(X_{i+1} - X_i)^2}{\triangle}}{\sum_{i=1}^{n} I(|x - X_i| < b_n)}$$

Without assuming the stationarity of the process, the authors derived the limit theory of the estimators in more general settings. In fact, Cai and Hong (2009) pointed out that Bandi and Phillips's estimators can be considered as a two-step kernel smoothing method, where the estimators on the first step are used to derive the ones on the second step. One can extend this idea to obtain a multi-step kernel method, but with the increase of steps, it is harder to figure out the relationship between steps and bandwidths. Arfi (1998) studied a particular diffusion model under the assumption of ergodicity. The author proposed a kernel estimator of the diffusion and proved strong consistency of the estimator as $\triangle \to 0$.

In addition, Chesney et al. (1993) considered the transformation $Y_t = \exp(X_t)$ and $f(Y_t)$, and then applied the Milstein scheme to obtain the approximation of Y_t and $f(Y_t)$. After some algebraic operations, the author approximated $b^2(X_t)$ by

$$\tilde{b}^2(X_t) = \frac{2}{\alpha\triangle}\left[\frac{Y_{t+\triangle}^{1+\alpha} - Y_t^{1+\alpha}}{(1+\alpha)Y_t^{1+\alpha}} - \frac{Y_{t+\triangle} - Y_t}{Y_t}\right]$$

then a kernel smoother can be used to estimate $\tilde{b}^2(x)$. Note that the choice of the power α must satisfy some technical conditions to guarantee the positivity of the estimator.

In addition to theoretical studies as mentioned above, there have been many empirical analyses on the form of the drift and diffusion for the short-term interest rate. Aït-Sahalia (1996) and Stanton (1997) claimed that the drift should have a nonlinear form. However Chapman and Pearson (2000) asked whether the drift is actually nonlinear or it is due to the bias of the estimator. Although the form of the drift is inconclusive in the paper, the authors found significant biases in the estimators of Aït-Sahalia (1996) and Stanton (1997) for fitting the linear drift. One reason for such biases is the original effect caused by the kernel estimator. In order

to overcome the problem of the boundary effect, great progress has been made in proposing the methods to improve the performance of the estimator of the drift in the origin bound. The local linear method (Fan and Zhang, 2003) and asymmetric kernels (Gospodinov and Hirukawa, 2012) can be used for this purpose.

3.2 Second-Order Diffusion Models

For a non-stationary process $\{X_t\}$, one can isolate temporal components to make the process stationary. However this detrending requires more assumptions and extra steps, so in practice differencing is a popular approach to remove a mean trend from the non-stationary process (Box et al., 1994). For example, while raw financial data usually exhibits non-stationarity and non-normality, it is practical to use differencing to model the stationary shocks because there are many theoretical methods applicable for stationary processes.

In order to accommodate this idea, Nicolau (2007) considered a second-order diffusion model for the process $\{Y_t\}$ given by

$$\begin{cases} dY_t = X_t dt \\ dX_t = a(X_t)dt + b(X_t)dW_t \end{cases} \tag{3.3}$$

where $\{X_t\}$ is a stationary process and $\{Y_t\}$ is a differentiable integrated process. Note that the above equations can be written as $d\,(dY_t/dt) = a(X_t)dt + b(X_t)dW_t$, so this is called a second-order equation. Nicolau (2007) obtained the proxy of X_t by

$$\tilde{X}_i = \frac{Y_i - Y_{i-1}}{\triangle}$$

because usually the process $\{X_t\}$ is unobservable. Then the author proposed kernel estimators of the drift and diffusion based on the proxy data, which are given by

$$\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - \tilde{X}_i) \qquad \hat{a}_n(x) = \frac{A_n(x)}{\hat{p}_n(x)} \qquad \hat{b}_n^2(x) = \frac{B_n(x)}{\hat{p}_n(x)}$$

where

$$\begin{aligned} A_n(x) &= \frac{1}{n\triangle_n}\sum_{i=1}^{n} K_{h_n}(x - \tilde{X}_i)\left(\tilde{X}_{i+1} - \tilde{X}_i\right) \\ B_n(x) &= \frac{3}{2n\triangle_n}\sum_{i=1}^{n} K_{h_n}(x - \tilde{X}_i)\left(\tilde{X}_{i+1} - \tilde{X}_i\right)^2 \end{aligned}$$

and proved weak consistency and asymptotic normality of the estimators. After that, Nicolau (2008) applied the proposed estimators to make an empirical analysis of the stock price and

exchange rate.

Wang and Lin (2011) extended Nicolau's idea to propose local linear estimators of the drift and diffusion for (3.3), which are given by

$$\begin{aligned}(\hat{\alpha}_0, \hat{\alpha}_1) &= \arg\min_{\alpha_0,\alpha_1} \sum_{i=1}^{n} \left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle_n} - \alpha_0 - \alpha_1(x - \tilde{X}_i) \right)^2 K_{h_n}(x - \tilde{X}_i) \\ (\hat{\beta}_0, \hat{\beta}_1) &= \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(\frac{\frac{3}{2}(\tilde{X}_{i+1} - \tilde{X}_i)^2}{\triangle_n} - \beta_0 - \beta_1(x - \tilde{X}_i) \right)^2 K_{h_n}(x - \tilde{X}_i)\end{aligned}$$

Then weak consistency is established in the paper and simulation results are used to validate the theory. However nonnegativity of the local linear estimator of the diffusion may not be guaranteed, thus similar to (Cai, 2001), Wang et al. (2012) proposed a weighted kernel estimator, and studied weak consistency and asymptotic normality of the estimators.

3.3 Time-Inhomogeneous Diffusion Models

The time-homogeneous model has its limitations and may not capture time varying characteristics of diffusion models. For example, economic and market conditions could vary from time to time (Cheng and Wang, 2007). So many efforts have been made to explicitly characterize the time inhomogeneity. Such kind of models include Ho and Lee (1986)'s model

$$dX_t = a(t)dt + b(t)dW_t,$$

Hull and White (1990)'s model

$$dX_t = [a_0(t) + a_1(t)X_t]dt + \sigma dW_t,$$

Black et al. (1990)'s model

$$dX_t = \left[a(t)X_t + \frac{b'(t)}{b(t)} X_t \ln X_t \right] dt + b(t)X_t dW_t,$$

and Black and Karasinski (1991)'s model

$$dX_t = [a_0(t)X_t + a_1(t)X_t \ln X_t]\, dt + b(t)X_t dW_t.$$

Note that all the above models can be written as the following general form

$$dX_t = [a_0(t) + a_1(t)f(t, X_t)]dt + b(t)[g(X_t)]^{\gamma(t)} dW_t. \tag{3.4}$$

For example, when $a_1(t) = 0$ and $\gamma(t) = 0$, (3.4) is the Ho and Lee model; when $f(t, X_t) = X_t$ and $\gamma(t) = 0$, (3.4) is the Hull and White model. Also (3.4) includes the popular time-homogeneous diffusion models. For example, when $a_0(t) = 0, a_1(t) = \mu, f(t, X_t) = X_t, b(t) = \sigma, g(X_t) = X_t$ and $\gamma(t) = 1$, (3.4) is the GBM; when $a_0(t) = ab, a_1(t) = -a, f(t, X_t) = X_t, b(t) = c$ and $\gamma(t) = 0$, (3.4) is the Vasicek model; when $a_0(t) = ab, a_1(t) = -a, f(t, X_t) = X_t, b(t) = c, g(X_t) = X_t$ and $\gamma(t) = 1/2$, (3.4) is the CIR model.

Fan et al. (2003) studied the refined model of (3.4) by assuming the linearity of $f(t, X_t)$ and $g(X_t)$ which is given by

$$dX_t = [a_0(t) + a_1(t)X_t]dt + b(t)X_t^{\gamma(t)}dW_t. \tag{3.5}$$

The discretized form of (3.5) is written as

$$X_{i+1} - X_i = [a_0(t_i) + a_1(t_i)X_i]\triangle_i + b(t_i)X_i^{\gamma(t_i)}\sqrt{\triangle_i}\varepsilon_i,$$

where the step $\triangle_i = t_{i+1} - t_i$, and note that

$$a(t, X_t) = \lim_{\triangle\to 0} E\left[\frac{X_{t+\triangle} - X_t}{\triangle}\middle| X_t\right] \quad \text{and} \quad b^2(t, X_t) = \lim_{\triangle\to 0} E\left[\frac{(X_{t+\triangle} - X_t)^2}{\triangle}\middle| X_t\right],$$

using a locally constant approximation of $a_0(t)$ and $a_1(t)$, the authors proposed the following estimator of the drift

$$(\hat{\alpha_0}, \hat{\alpha_1}) = \arg\min_{\alpha_0,\alpha_1} \sum_{i=1}^{n-1}\left[\frac{X_{i+1} - X_i}{\triangle_i} - \alpha_0 - \alpha_1 X_i\right]^2 K_{h_n}(t_i - t_0)$$

in which $\hat{a}_0(t) = \hat{\alpha_0}$ and $\hat{a}_1(t) = \hat{\alpha_1}$ depend on the time t. Let $\hat{a}(t, X_t) = \hat{a}_0(t) + \hat{a}_1(t)X_t$ and $R_i = \frac{1}{\sqrt{\triangle_i}}[X_{i+1} - X_i - \hat{a}(t_i, X_i)\triangle_i]$ and assume locally constant approximation of $b(t)$ and $\gamma(t)$, the authors proposed the estimator of the diffusion based on the quasi-likelihood function

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \sum_{i=1}^{n-1} K_{h_n}(t_i - t_0)\left[\ln(\beta^2 X_i^{2\gamma}) + \frac{R_i^2}{\beta X_i^{2\gamma}}\right]$$

then $\hat{b}(t) = \hat{\beta}$ and $\hat{\gamma}(t) = \hat{\gamma}$ depend on the time t. The authors applied the proposed estimators to study the short-term treasury bill data. However, the process $\{X_t\}$ may not be stationary due to the time-varying drift and diffusion, thus the authors did not obtain the asymptotic properties of the estimators, instead validating their results by simulations and real applications. Wang and Xiao (2013) generalized Fan et al. (2003)'s work for the more general time-inhomogeneous model (3.4) where the forms of $f(\cdot)$ and $g(\cdot)$ are given. The authors proposed local linear estimators and under regular conditions, weak consistency and asymptotic normality are established for the estimator of the diffusion. In addition, Yu et al. (2009) proposed the penalized

smoothing estimators for the drift and diffusion in the time-inhomogeneous case and proved the asymptotic properties of the estimators.

3.4 Remark

In addition to kernel estimation in diffusion processes, many efforts have been made to explore nonparametric estimation in jump-diffusion processes for modeling the behaviors with discontinuities or jumps caused by unpredicted events (Lobo, 1999; Bollerslev and Zhou, 2002; Liu et al., 2002; Johannes, 2004). For example, Bandi and Nguyen (2003) proposed the estimators for the time-homogeneous jump-diffusion process and provided the asymptotic properties for the estimators. Hanif (2013) studied local linear estimation for jump-diffusion models by using asymmetric kernels to overcome the boundary effect of kernel methods. Song and Lin (2013) applied the empirical likelihood method to make inference of the second-order jump-diffusion model based on Nadaraya-Watson estimators. Schmisser (2014) used the penalized least squares method to propose two adaptive estimators of the drift and characterized the bounds for the risks of both estimators. Song et al. (2013) proposed weighted Nadaraya-Watson estimators of a second-order jump-diffusion model and gave the asymptotic properties of these estimators. Cai and Hong (2009) provided detailed summary of the developments of nonparametric regression in these areas. However, all methods mentioned above are offline, namely taking all past observations in each calculation. When there are a great amount of data, the computing time is unbearable as discussed in Chapter 1. So in order to accelerate the procedure, it is necessary to put forward an incremental approach with linear complexity. This is what we will do in Chapter 4.

Chapter 4

Asymptotic Theory of Online Estimators for SDE

4.1 Introduction

The purpose of this chapter is twofold. First, we propose online kernel estimators of the drift and diffusion in the time-homogeneous stationary diffusion models for discrete-time sequential observations. Second, we study large sample properties of the estimators by establishing their quadratic convergence, strong consistency and asymptotic normality.

In this chapter, we start from time-homogeneous stationary diffusion models because many such models have been developed to describe stochastic behaviors of a wide range of risk factors including interest rates, for example the Vasicek (VAS) model (Vasicek, 1977) and the Cox, Ingersoll and Ross (CIR) model (Cox et al., 1985):

$$\begin{aligned} \text{(VAS):}\quad dX_t &= (\alpha_0 + \alpha_1 X_t)dt + \sigma dW_t \\ \text{(CIR):}\quad dX_t &= (\alpha_0 + \alpha_1 X_t)dt + \sigma X_t^{1/2} dW_t \end{aligned}$$

where X_t is the risk factor we are concerned with and W_t is a Brownian motion. Moreover the study of time-homogeneous stationary models can provide us with valuable intuition for the non-stationary case. The proposed online estimators of the drift and diffusion are new in that previous work on nonparametric estimation of diffusion models developed offline procedures, and our estimators can meet real-time demands of managing risks which are often required in practice. Here we concentrate on high-frequency data with sufficiently long time interval, i.e. the discretization step size $\triangle \to 0$ and the time period $T = n\triangle \to \infty$, because there are more one minute bars available with the development of modern technology. In this case, we establish the asymptotic properties of the proposed estimators. In this chapter, we mainly present the theoretical analysis of the online technique which will be useful for further inference, and

leave numerical examples and a case study to the next chapter.

This chapter is organized as follows. In Section 4.2, we derive online kernel estimators for the drift and diffusion. Section 4.3 presents the theoretical analysis of online estimators. In this section, some assumptions are made and useful lemmas from previous work are provided for our proof. Then we prove quadratic convergence, strong consistency and asymptotic normality respectively. A brief conclusion is contained in Section 4.4.

4.2 Method

Suppose the discrete-time sequential observations $\{X_i\}_{i\geq 0}$ at the time $t_i \geq 0$ and the discretization step size $\triangle_i = t_{i+1} - t_i$, where $\{X_i\}_{i\geq 0}$ are from the time-homogeneous stationary diffusion model

$$dX_t = a(X_t)dt + b(X_t)dW_t \tag{4.1}$$

where $\{W_t\}$ is the Wiener process and $a(\cdot), b^2(\cdot)$ are called the drift and diffusion respectively. The infinitesimal generator gives the precise form of the drift and diffusion by

$$a(X_t) = \lim_{\triangle\to 0} \frac{1}{\triangle} E[(X_{t+\triangle} - X_t)|X_t] \qquad \text{and} \qquad b^2(X_t) = \lim_{\triangle\to 0} \frac{1}{\triangle} E[(X_{t+\triangle} - X_t)^2|X_t]$$

Then Nadaraya-Watson kernel regression gives the estimator $\sum_{i=0}^{n} \frac{X_{i+1}-X_i}{\triangle_i} K_{h_n}(x-X_i) / \sum_{i=0}^{n} K_{h_n}(x-X_i)$ for the drift and $\sum_{i=0}^{n} \frac{(X_{i+1}-X_i)^2}{\triangle_i} K_{h_n}(x-X_i) / \sum_{i=0}^{n} K_{h_n}(x-X_i)$ for the diffusion. It can be found that as n increases, more observations are included for estimation and the past ones are used over and over again. This results in inefficiency of the offline estimators. However inspired by the idea of stochastic approximation (which is provided in Appendix B), we can estimate the drift and diffusion more efficiently by a semi-recursive procedure. In this thesis, the online estimator of $a(x)$ is proposed as follows:

$$\hat{g}_n(x) = \hat{g}_{n-1}(x) + \frac{1}{n}[Y_n K_{h_n}(x - X_n) - \hat{g}_{n-1}(x)] \tag{4.2}$$

$$\hat{f}_n(x) = \hat{f}_{n-1}(x) + \frac{1}{n}[K_{h_n}(x - X_n) - \hat{f}_{n-1}(x)] \tag{4.3}$$

$$\hat{a}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)} \tag{4.4}$$

where $Y_n = \frac{X_{n+1}-X_n}{\triangle_n}$ and $K_{h_n}(x) = K(x/h_n)/h_n$ with the kernel function $K(\cdot)$, and the online estimator of $b^2(x)$ is as follows:

$$\hat{d}_n(x) = \hat{d}_{n-1}(x) + \frac{1}{n}[Z_n K_{h_n}(x - X_n) - \hat{d}_{n-1}(x)] \tag{4.5}$$

$$\hat{b}^2_n(x) = \frac{\hat{d}_n(x)}{\hat{f}_n(x)} \tag{4.6}$$

where $Z_n = \frac{(X_{n+1}-X_n)^2}{\triangle_n}$. From a practical view, it is required to give an initial value to trigger the online estimator so as to make it work. We suggest to use the offline estimator with a small number of observations to decide $\hat{g}_0(x)$, $\hat{d}_0(x)$ and $\hat{f}_0(x)$.

It is worth mentioning the difference between the "online" and "offline" methods. The offline procedure stores all the samples and each time as new samples become available, the algorithm must recalculate the estimate at x using all samples. In contrast, the online procedure stores just the most recent sample and uses this to update the estimate at x. Therefore the proposed online estimators have the attractive properties of saving memory and running time. In addition, another technical difference is that the offline procedure applies one bandwidth to all current history samples and the online version uses a bandwidth for the most recent sample. But as Huang (2011) pointed out, online methods could have a larger bias but a smaller asymptotic variance even though the same convergence rate could be achieved as their offline counterparts. It is because the choice of the bandwidth usually depends on the number of samples. On average, online methods use a bigger bandwidth.

4.3 Theoretical Analysis

4.3.1 Assumptions and Preliminary Lemmas

In this part, we will prove the asymptotical properties of the estimators $\hat{a}_n(x)$ and $\hat{b}_n^2(x)$ in (4.4) and (4.6). We need the following assumptions:

A1 For the time-homogeneous case, $a(\cdot)$ and $b^2(\cdot)$ have at least first-order bounded derivatives and $\int_0^T |a(X_t)|dt < \infty$ and $\int_0^T |b^2(X_t)|dt < \infty$ almost surely which imply that $a(\cdot)$ and $b^2(\cdot)$ are bounded almost surely for $X_t, t \in [0, T]$. In addition, $a(\cdot)$ and $b(\cdot)$ satisfy the linear growth condition, i.e. for some constant $C > 0$ we have

$$|a(x)| \leq C(1 + x^2)^{1/2} \quad \text{and} \quad 0 \leq b(x) \leq C(1 + x^2)^{1/2}$$

and satisfy the uniform Lipschitz condition in x, i.e. for $x, y \in \mathbb{R}$ and $t \in [0, T]$, there is some positive constant D such that $|a(x)-a(y)| \leq D|x-y|$ and $|b(x)-b(y)| \leq D|x-y|$. Based on (Jiang and Knight, 1997), this assumption ensures that (4.1) has a unique solution

$$X_t = X_0 + \int_0^t a(X_t)dt + \int_0^t b(X_t)dW_t$$

A2 The stationary process $\{X_t\}$ has the bounded marginal distribution $f(x)$ with at least first-order bounded derivative on the support $\{x \in R : f(x) > 0\}$. In addition, for any t and s, X_t and X_s have the joint density $f(x, y)$ which satisfies $\sup_{x,y \in R} |f(x, y)| < \infty$.

A3 Assume that

$$\lim_{x\to+\infty}\sup\left(\frac{a(x)}{b(x)}-\frac{b'(x)}{2}\right)<0 \quad \text{and} \quad \lim_{x\to-\infty}\sup\left(\frac{a(x)}{b(x)}-\frac{b'(x)}{2}\right)>0$$

Based on (Nicolau, 2003), this assumption implies that the process is ρ-mixing with exponential decay, i.e., the mixing coefficient $\rho_k = O(e^{-rk})$ for some positive constant r. The definition of a ρ-mixing process can be seen in Appendix C.

A4 The kernel function satisfies

$$\sup_{x\in R}|K(x)|<\infty \text{ and } \int_R |K(x)|dx<\infty \tag{4.7}$$

$$\int_R K(x)dx = 1 \tag{4.8}$$

and for all $l \geq 0$

$$c_l = \int_R x^l K(x)dx < \infty \tag{4.9}$$

A5 The step of discretization $\triangle_n = \triangle = O(k^{-1})$ where $k/n \to 0$ as $k \to \infty$ and $n \to \infty$, implying that $\triangle \to 0$ and $n\triangle \to \infty$. Note that k represents the sampling frequency which is determined by the current development of technology. With the development of communication techniques, we can sample higher frequency of observations (i.e. $k \to \infty$). The intuition behind $k/n \to 0$ and $n\triangle \to \infty$ is that we can always obtain a sufficiently large sample of high-frequency data as long as we sample for sufficiently long time[1].

A6 The bandwidth $h_n \to 0, nh_n \to \infty$ and $nh_n\triangle \to \infty$ as $k \to \infty$ and $n \to \infty$. We also assume

$$\beta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{h_i}{h_n} < \infty \tag{4.10}$$

$$\alpha_l = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_n}{h_i}\right)^l < \infty \quad \text{for } l = 1/2, 1, 3/2, 2, 3 \tag{4.11}$$

Note that (4.10) and (4.11) can always be satisfied. For example, if $h_n = cn^{-\gamma}$ is taken, then

$$\beta = \frac{1}{n}\sum_{i=1}^{n}\frac{h_i}{h_n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^{-\gamma} \leq \int_0^1 x^{-\gamma}dx = \frac{1}{1-\gamma}$$

$$\alpha_l = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_n}{h_i}\right)^l \leq \frac{1}{1+l\gamma}$$

[1]Note that this assumption is not enough to sample for longer time at a low frequency, as then the data will be too far apart in state space to get a good Taylor series type estimate.

Therefore in this case, $0 < \gamma < 1$ can guarantee that (4.10) and (4.11) hold true. In addition, we assume that the item in the expression $O(\triangle^k)$ or $o(\triangle^k)$ is uniform in its domain for any $k \geq 0$.

The following lemmas owing to Bochner (Wheeden and Zygmund, 1977), Toeplitz (Loève, 1977), and Fan and Zhang (2003) are useful to our later proof.

Lemma 4.3.1 (Bochner)**.** *Suppose $K(x)$ satisfies (4.7) and $h_n \to 0$ as $n \to \infty$. Let $g(x) \in L^1$, then*

$$\lim_{n\to\infty} \int_R K_{h_n}(x-u)g(u)du = g(x) \int_R K(u)du$$

Lemma 4.3.2 (Toeplitz)**.** *Let $\{a_{n,k}\}_{n,k=1}^{\infty}$ be a matrix of real numbers satisfying: (1) $\lim_{n\to\infty} a_{n,k} = 0$ for each k; (2) $\sum_k |a_{n,k}| \leq C < \infty$ for each n and (3) $\lim_{n\to\infty} \sum_k a_{n,k} = A < \infty$. Let $\{x_n\}$ be a sequence of real numbers with a finite limit x as $n \to \infty$. Then*

$$\sum_k a_{n,k}x_k \to Ax \quad as\ n \to \infty$$

In addition, similar to Taylor expansion, for a function $m(\cdot) \in L^{2(p+1)}$, we have $E[m(X_{t+\triangle})|X_t] = \sum_{j=0}^{p} \mathcal{L}^j m(X_t)\frac{\triangle^j}{j!}$ where the operator $\mathcal{L} = a\frac{d}{dx} + \frac{1}{2}b^2\frac{d^2}{dx^2}$ is based on the infinitesimal generator. Then there is the following lemma

Lemma 4.3.3 (Fan and Zhang)**.** *For the time-homogeneous diffusion process (4.1),*

$$E[(X_{t+\triangle} - X_t)|X_t] = a(X_t)\triangle + O(\triangle^2) \tag{4.12}$$

$$E[(X_{t+\triangle} - X_t)^2|X_t] = b^2(X_t)\triangle + O(\triangle^2) \tag{4.13}$$

$$E[(X_{t+\triangle} - X_t)^3|X_t] = 3b^2(X_t)\left[a(X_t) + \frac{1}{2}(b^2)'(X_t)\right]\triangle^2 + O(\triangle^3) \tag{4.14}$$

$$E[(X_{t+\triangle} - X_t)^4|X_t] = 3b^4(X_t)\triangle^2 + O(\triangle^3) \tag{4.15}$$

4.3.2 Quadratic Convergence

In this section we establish quadratic convergence of the proposed estimators. Our proof partially follows Masry (1986)'s argument for the totally recursive estimator of the density function, but goes beyond it by considering the semi-recursive estimators for diffusion models.

Proposition 4.3.4. *Under **A1-A6**, we have*

$$E\hat{g}_n(x) = a(x)f(x) - (a(x)f'(x) + a'(x)f(x))c_1\beta h_n + O(h_n^2 + \triangle) \tag{4.16}$$

Proof. Note that (4.2) can be rewritten as

$$\begin{aligned}\hat{g}_n(x) &= \hat{g}_{n-1}(x) + \frac{1}{n}[Y_n K_{h_n}(x - X_n) - \hat{g}_{n-1}(x)] = \frac{n-1}{n}\hat{g}_{n-1}(x) + \frac{1}{n}Y_n K_{h_n}(x - X_n) \\ &= \frac{n-1}{n}\left[\frac{n-2}{n-1}\hat{g}_{n-2}(x) + \frac{1}{n-1}Y_{n-1}K_{h_{n-1}}(x - X_{n-1})\right] + \frac{1}{n}Y_n K_{h_n}(x - X_n) \\ &= \frac{n-2}{n}\hat{g}_{n-2}(x) + \frac{1}{n}[Y_{n-1}K_{h_{n-1}}(x - X_{n-1}) + Y_n K_{h_n}(x - X_n)] = \ldots\ldots = \frac{1}{n}\sum_{i=1}^{n} Y_i K_{h_i}(x - X_i)\end{aligned}$$

thus taking expectation on both sides of the above expression yields

$$\begin{aligned}E\hat{g}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i K_{h_i}(x - X_i)] = \frac{1}{n}\sum_{i=1}^{n} E\{E[Y_i K_{h_i}(x - X_i)|X_i]\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\{K_{h_i}(x - X_i)E[Y_i|X_i]\} = \frac{1}{n}\sum_{i=1}^{n}\int_R K_{h_i}(x - z)E[Y_i|X_i = z]f(z)dz\end{aligned}$$

Then, Lemma 4.3.3 implies $E[X_{i+1} - X_i|X_i = z] = a(z)\triangle + O(\triangle^2)$ and thus

$$E\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\int_R K_{h_i}(x - z)[a(z) + O(\triangle)]f(z)dz$$

Take $z = x - h_i u$ and apply Taylor expansion to yield

$$\begin{aligned}E\hat{g}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\int_R K(u)[a(x - h_i u) + O(\triangle)]f(x - h_i u)du \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_R K(u)[a(x) - h_i u a'(x) + O(h_i^2 + \triangle)][f(x) - h_i u f'(x) + O(h_i^2)]du \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_R K(u)\{a(x)f(x) - h_i u[a'(x)f(x) + a(x)f'(x)] + O(h_i^2 + \triangle)\}du \\ &= a(x)f(x) - [a(x)f'(x) + a'(x)f(x)]\frac{h_n}{n}\sum_{i=1}^{n}\frac{h_i}{h_n}\int_R uK(u)du + O(h_i^2 + \triangle) \\ &= a(x)f(x) - [a(x)f'(x) + a'(x)f(x)]c_1\beta h_n + O(h_n^2 + \triangle)\end{aligned}$$

□

Proposition 4.3.5. *Under **A1-A6**, the following statement holds true*

$$E\hat{f}_n(x) = f(x) - f'(x)c_1\beta h_n + O(h_n^2 + \triangle) \tag{4.17}$$

Proof. Similar to the proof of Proposition 4.3.4. □

Using Proposition 4.3.4 and 4.3.5, we can characterize the ratio $\frac{E\hat{g}_n(x)}{E\hat{f}_n(x)}$ by the following statement:

Proposition 4.3.6. *Under **A1-A6**,*

$$\frac{E\hat{g}_n(x)}{E\hat{f}_n(x)} - a(x) = -a'(x)c_1\beta h_n + O(h_n^2 + \triangle)$$

Proof. It can be seen that

$$\begin{aligned}
\frac{E\hat{g}_n(x)}{E\hat{f}_n(x)} - a(x) &= \frac{a(x)f(x) - (a(x)f'(x) + a'(x)f(x))c_1\beta h_n + O(h_n^2 + \triangle)}{f(x) - f'(x)c_1\beta h_n + O(h_n^2)} - a(x) \\
&= \frac{-a'(x)f(x)c_1\beta h_n + O(h_n^2 + \triangle)}{f(x) - f'(x)c_1\beta h_n + O(h_n^2)} \\
&= -a'(x)c_1\beta h_n + O(h_n^2 + \triangle)
\end{aligned}$$

□

In addition, we refer to the following lemma for the expectation and variance of ratio of two random variables (Elandt-Johnson and Johnson, 1980)

Lemma 4.3.7. *For the random variables X and Y,*

$$\begin{aligned}
E\left(\frac{X}{Y}\right) &= \frac{\mu_X}{\mu_Y} + O(Var(X) + Var(Y) + Cov(X, Y)) \\
Var\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\left\{Var(X) - 2\frac{\mu_X}{\mu_Y}Cov(X, Y) + \left(\frac{\mu_X}{\mu_Y}\right)^2 Var(Y)\right\}
\end{aligned}$$

where $\mu_X = EX$ and $\mu_Y = EY$.

In order to obtain the bias $E\hat{a}_n(x) - a(x)$, it suffices to calculate $Var[\hat{g}_n(x)]$, $Var[\hat{f}_n(x)]$ and $Cov[\hat{g}_n(x), \hat{f}_n(x)]$.

Proposition 4.3.8. *Under **A1-A6**,*

$$Var[\hat{g}_n(x)] = O\left(\frac{1}{nh_n\triangle}\right)$$

Proof. Here we have

$$\begin{aligned} Var[\hat{g}_n(x)] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} Y_i K_{h_i}(x-X_i)\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} Y_i K_{h_i}(x-X_i)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i K_{h_i}(x-X_i)] + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} Cov[Y_i K_{h_i}(x-X_i), Y_j K_{h_j}(x-X_j)] \\ &\triangleq I_n + R_n \end{aligned}$$

First, note that

$$\begin{aligned} I_n &= \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i K_{h_i}(x-X_i)] = \frac{1}{n^2}\sum_{i=1}^{n}\left\{E[Y_i K_{h_i}(x-X_i)]^2 - E^2[Y_i K_{h_i}(x-X_i)]\right\} \\ &\triangleq I_{n1} - I_{n2} \end{aligned}$$

where

$$I_{n1} = \frac{1}{n^2}\sum_{i=1}^{n} E[Y_i K_{h_i}(x-X_i)]^2 = \frac{1}{n^2}\sum_{i=1}^{n}\int_R K_{h_i}^2(x-z) E[Y_i^2|X_i=z] f(z)dz$$

Lemma 4.3.3 implies $E[Y_i^2|X_i=z] = \frac{b^2(z)}{\triangle} + O(1)$, thus

$$\begin{aligned} I_{n1} &= \frac{1}{n^2}\sum_{i=1}^{n}\int_R K_{h_i}^2(x-z)\left[\frac{b^2(z)}{\triangle} + O(1)\right] f(z)dz \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{h_i}\int_R \frac{1}{h_i} K^2\left(\frac{x-z}{h_i}\right)\left[\frac{b^2(z)}{\triangle} + O(1)\right] f(z)dz \end{aligned}$$

It can be checked that $K^2(x)$ satisfies the assumption for Lemma 4.3.1 as well, so as $i \to \infty$

$$\int_R \frac{1}{h_i} K^2\left(\frac{x-z}{h_i}\right)\left[b^2(z) + O(\triangle)\right] f(z)dz \to [b^2(x) + O(\triangle)] f(x)\int_R K^2(u)du$$

By Lemma 4.3.2 where $a_{n,i} = \frac{h_n}{nh_i}$ for $i \le n$ and $a_{n,i} = 0$ for $i > n$ and from (4.11)

$$\lim_{n\to\infty}\sum_{i=1}^{n} a_{n,i} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{h_n}{h_i} < \infty$$

thus

$$\begin{aligned} I_{n1} &= \frac{1}{nh_n\triangle}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{h_n}{h_i}\int_R \frac{1}{h_i} K^2\left(\frac{x-z}{h_i}\right)\left[b^2(z) + O(\triangle)\right] f(z)dz\right] \\ &= O\left(\frac{1}{nh_n\triangle}\right) \end{aligned}$$

In addition,

$$I_{n2} = \frac{1}{n^2}\sum_{i=1}^{n} E^2[Y_i K_{h_i}(x-X_i)] = \frac{1}{n^2}\sum_{i=1}^{n}\left[\int_R K_{h_i}(x-z)(a(z)+O(\triangle))f(z)dz\right]^2$$

By Lemma 4.3.1 and $\int_R K(x)dx = 1$, we have $\left\{\int_R K_{h_i}(x-z)[a(z)+O(\triangle)]f(z)dz\right\}^2 \to a^2(x)f^2(x)$, and by Lemma 4.3.2 so $I_{n2} = O(1/n)$. Therefore we can obtain $I_n = I_{n1} - I_{n2} = O\left(\frac{1}{nh_n\triangle}\right)$.

For the other part,

$$\begin{aligned} |R_n| &\le \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} |Cov[Y_i K_{h_i}(x-X_i), Y_j K_{h_j}(x-X_j)]| \\ &\le \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} E^{1/2}[Y_i K_{h_i}(x-X_i)]^2 E^{1/2}[Y_j K_{h_j}(x-X_j)]^2 |\rho(|j-i|)| \end{aligned}$$

where note that

$$\begin{aligned} E|Y_i K_{h_i}(x-X_i)|^2 &= \int_R K_{h_i}^2(x-z)\left[\frac{b^2(z)}{\triangle} + O(1)\right]f(z)dz \\ &= \frac{1}{h_i\triangle}\int_R \frac{1}{h_i}K^2\left(\frac{x-z}{h_i}\right)[b^2(z)+O(\triangle)]f(z)dz \triangleq \frac{1}{h_i\triangle}q_i(x) \end{aligned}$$

by Lemma 4.3.1 it follows that

$$q_i(x) = \int_R \frac{1}{h_i}K^2\left(\frac{x-z}{h_i}\right)[b^2(z)+O(\triangle)]f(z)dz \to b^2(x)f(x)\int_R K^2(u)du$$

thus by Cauchy-Schwartz inequality

$$\begin{aligned} |R_n| &\le \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} |\rho(|j-i|)|\left[\frac{1}{h_i\triangle}q_i(x)\right]^{1/2}\left[\frac{1}{h_j\triangle}q_j(x)\right]^{1/2} \\ &\le \frac{1}{nh_n\triangle}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{h_n}{h_i}|\rho(i)|q_i(x)\right]^{1/2}\left[\frac{1}{n}\sum_{j=1}^{n}\frac{h_n}{h_j}|\rho(j)|q_j(x)\right]^{1/2} \end{aligned}$$

note that $|\rho(n)|q_n(x) \to 0$ and by Lemma 4.3.2, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\frac{h_n}{h_i}|\rho(i)|q_i(x) &= o(1) \\ \frac{1}{n}\sum_{j=1}^{n}\frac{h_n}{h_j}|\rho(j)|q_j(x) &= o(1) \end{aligned}$$

implying that $|R_n| = o\left(\frac{1}{nh_n\triangle}\right)$. So

$$0 \leq Var[\hat{g}_n(x)] = I_n + R_n \leq I_n + |R_n| = O\left(\frac{1}{nh_n\triangle}\right)$$

Therefore

$$Var[\hat{g}_n(x)] = O\left(\frac{1}{nh_n\triangle}\right)$$

□

Similarly, we can prove that $Var[\hat{f}_n(x)] = O\left(\frac{1}{nh_n}\right)$ and $Cov[\hat{g}_n(x), \hat{f}_n(x)] = O\left(\frac{1}{nh_n\triangle^{1/2}}\right)$.

Theorem 4.3.9. *Under **A1**-**A6**,*

$$E[\hat{a}_n(x) - a(x)]^2 = O\left(h_n^2 + \triangle^2 + h_n\triangle + \frac{1}{nh_n\triangle}\right) \to 0$$

that is, $\hat{a}_n(x) \xrightarrow{L^2} a(x)$.

Proof. Note that

$$E[\hat{a}_n(x) - a(x)]^2 = [E\hat{a}_n(x) - a(x)]^2 + Var[\hat{a}_n(x)]$$

From Propositions 4.3.6 and 4.3.8 and Lemma 4.3.7, we have

$$\begin{aligned} E\hat{a}_n(x) - a(x) &= \frac{E\hat{g}_n(x)}{E\hat{f}_n(x)} - a(x) + O(Var[\hat{g}_n(x)] + Var[\hat{f}_n(x)] + Cov[\hat{g}_n(x), \hat{f}_n(x)]) \\ &= O\left(h_n + \triangle + \frac{1}{nh_n\triangle}\right) \end{aligned}$$

and

$$\begin{aligned} Var[\hat{a}_n(x)] &= Var\left[\frac{\hat{g}_n(x)}{\hat{f}_n(x)}\right] \\ &= \frac{1}{E^2\hat{f}_n(x)}\left\{Var[\hat{g}_n(x)] - 2\frac{E\hat{g}_n(x)}{E\hat{f}_n(x)}Cov[\hat{g}_n(x), \hat{f}_n(x)] + \left(\frac{E\hat{g}_n(x)}{E\hat{f}_n(x)}\right)^2 Var[\hat{f}_n(x)]\right\} \\ &= O\left(\frac{1}{nh_n\triangle}\right)\frac{1}{E^2\hat{f}_n(x)}\left(1 - \frac{E\hat{g}_n(x)}{E\hat{f}_n(x)}\right)^2 = O\left(\frac{1}{nh_n\triangle}\right) \end{aligned}$$

Therefore, we have

$$\begin{aligned} E[\hat{a}_n(x) - a(x)]^2 &= \left[O\left(h_n + \triangle + \frac{1}{nh_n\triangle}\right)\right]^2 + O\left(\frac{1}{nh_n\triangle}\right) \\ &= O\left(h_n^2 + \triangle^2 + h_n\triangle + \frac{1}{nh_n\triangle}\right) \to 0 \end{aligned}$$

that is, $\hat{a}_n(x) \xrightarrow{L^2} a(x)$. □

Using similar steps, we can obtain quadratic convergence of $\hat{b}_n^2(x)$. Here we give just a sketch of the proof. To do so, the following statements are useful:

Lemma 4.3.10. *Under **A1-A6**, then*

(1) $E[(X_{t+\triangle} - X_t)^8|X_t] = 105b^8(X_t)\triangle^4 + O(\triangle^5)$

(2) $E\hat{d}_n(x) = b^2(x)f(x) - [b^2(x)f'(x) + (b^2)'(x)f(x)]c_1\beta h_n + O(h_n^2 + \triangle)$

(3) $Var[\hat{d}_n(x)] = O\left(\frac{1}{nh_n}\right)$ *and* $Cov[\hat{d}_n(x), \hat{f}_n(x)] = O\left(\frac{1}{nh_n}\right)$

Proof. For (1), from the definition of the infinitesimal generator $\mathcal{L} = a\frac{d}{dx} + \frac{1}{2}b^2\frac{d^2}{dx^2}$, by letting $m(x) = (x - X_t)^8$, we find that

$$\mathcal{L}m(X_t) = 0 \quad \text{and} \quad \mathcal{L}^2m(X_t) = 0 \quad \text{and} \quad \mathcal{L}^3m(X_t) = 0$$

and

$$\mathcal{L}^4m(X_t) = \frac{8!}{2^4}b^8(X_t)\frac{\triangle^4}{4!} + O(\triangle^5)$$

that is, the approximation of order $\triangle^4$ is

$$E[(X_{t+\triangle} - X_t)^8|X_t] = 105b^8(X_t)\triangle^4 + O(\triangle^5)$$

Actually, for any integer $m \geq 1$, it can be calculated that

$$E[(X_{t+\triangle} - X_t)^{2m}|X_t] = \frac{(2m)!}{2^m}b^{2m}(X_t)\frac{\triangle^m}{m!} + O(\triangle^{m+1})$$

For (2), Lemma 4.3.1, Lemma 4.3.2 and Lemma 4.3.3 imply that

$$\begin{aligned} E\hat{d}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} E[Z_iK_{h_i}(x - X_i)] = \frac{1}{n}\sum_{i=1}^{n}\int_R K_{h_i}(x - x)[b^2(z) + O(\triangle)]f(z)dz \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_R K(u)[b^2(x) - h_iu(b^2)'(x) + O(h_i^2 + \triangle)][f(x) - h_iuf'(x) + O(h_i^2)]du \\ &= b^2(x)f(x) - [b^2(x)f'(x) + (b^2)'(x)f(x)]c_1\beta h_n + O(h_n^2 + \triangle). \end{aligned}$$

For (3), we have

$$\begin{aligned} Var[\hat{d}_n(x)] &= \frac{1}{n^2}\sum_{i=1}^{n} Var[Z_iK_{h_i}(x - X_i)] + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}} Cov[Z_iK_{h_i}(x - X_i), Z_jK_{h_j}(x - X_j)] \\ &\triangleq I_n + R_n \end{aligned}$$

then note that

$$\begin{aligned} I_n &= \frac{1}{n^2}\sum_{i=1}^{n} Var[Z_i K_{h_i}(x-X_i)] = \frac{1}{n^2}\sum_{i=1}^{n}\{E[Z_i K_{h_i}(x-X_i)]^2 - E^2[Z_i K_{h_i}(x-X_i)]\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left\{\int_R K_{h_i}^2(x-z)[3b^4(z)+O(\triangle)]f(z)dz - \left[\int_R K_{h_i}(x-z)[b^2(z)+O(\triangle)]f(z)dz\right]^2\right\} \\ &= O\left(\frac{1}{nh_n}\right). \end{aligned}$$

For R_n, note that

$$\begin{aligned} |R_n| &\leq \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} |Cov[Z_i K_{h_i}(x-X_i), Z_j K_{h_j}(x-X_j)]| \\ &\leq \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} E^{1/2}[Z_i K_{h_i}(x-X_i)]^2 E^{1/2}[Z_j K_{h_j}(x-X_j)]|\rho(|j-i|)| \end{aligned}$$

where

$$E[Z_i K_{h_i}(x-X_i)]^2 = \frac{1}{h_i}\int_R \frac{1}{h_i}K^2\left(\frac{x-z}{h_i}\right)[b^4(z)+O(\triangle)]f(z)dz = O(1/h_i)$$

so

$$|R_n| \leq \frac{1}{nh_n}\left[\frac{1}{n}\sum_{i=1}^{n} O(h_n/h_i)|\rho(i)|\right]^{1/2}\left[\frac{1}{n}\sum_{j=1}^{n} O(h_n/h_j)|\rho(j)|\right]^{1/2} = o\left(\frac{1}{nh_n}\right)$$

then we have

$$0 \leq Var[\hat{d}_n(x)] = I_n + R_n \leq I_n + |R_n| \leq O\left(\frac{1}{nh_n}\right).$$

therefore

$$Var[\hat{d}_n(x)] = O\left(\frac{1}{nh_n}\right)$$

Similarly, it can be obtained that $Cov[\hat{d}_n(x), \hat{f}_n(x)] = O\left(\frac{1}{nh_n}\right)$. □

Then quadratic convergence of $\hat{b}_n^2(x)$ is illustrated by the following theorem

Theorem 4.3.11. *Under **A1-A6**,*

$$E[\hat{b}_n^2(x) - b^2(x)]^2 = O\left(h_n^2 + \frac{1}{nh_n}\right) \to 0$$

that is, $\hat{b}_n^2(x) \xrightarrow{L^2} b^2(x)$.

Proof. First of all, we have

$$\begin{aligned}\frac{E\hat{d}_n(x)}{E\hat{f}_n(x)} - b^2(x) &= \frac{b^2(x)f(x) - [b^2(x)f'(x) + (b^2)'(x)f(x)]c_1\beta h_n + O(h_n^2 + \triangle)}{f(x) - f'(x)c_1\beta h_n + O(h_n^2)} - b^2(x) \\ &= \frac{-(b^2)'(x)f(x)c_1\beta h_n + O(h_n^2 + \triangle)}{f(x) - f'(x)c_1\beta h_n + O(h_n^2)} \\ &= -(b^2)'(x)c_1\beta h_n + O(h_n^2 + \triangle)\end{aligned}$$

then

$$\begin{aligned}E\hat{b}_n^2(x) - b^2(x) &= \frac{E\hat{d}_n(x)}{E\hat{f}_n(x)} - b^2(x) + O\left(Var[\hat{d}_n(x)] + Var[\hat{f}_n(x)] + Cov[\hat{d}_n(x) + \hat{f}_n(x)]\right) \\ &= O\left(h_n + \frac{1}{nh_n}\right)\end{aligned}$$

and

$$\begin{aligned}Var[\hat{b}_n^2(x)] &= Var\left[\frac{\hat{d}_n(x)}{\hat{f}_n(x)}\right] \\ &= \frac{1}{E^2\hat{f}_n(x)}\left\{Var[\hat{d}_n(x)] - 2\frac{E\hat{d}_n(x)}{E\hat{f}_n(x)}Cov[\hat{d}_n(x), \hat{f}_n(x)] + \left(\frac{E\hat{d}_n(x)}{E\hat{f}_n(x)}\right)^2 Var[\hat{f}_n(x)]\right\} \\ &= O\left(\frac{1}{nh_n}\right)\frac{1}{E^2\hat{f}_n(x)}\left(1 - \frac{E\hat{d}_n(x)}{E\hat{f}_n(x)}\right)^2 = O\left(\frac{1}{nh_n}\right)\end{aligned}$$

Therefore, we have

$$E[\hat{b}_n^2(x) - b^2(x)]^2 = \left[O\left(h_n + \frac{1}{nh_n}\right)\right]^2 + O\left(\frac{1}{nh_n}\right) = O\left(h_n^2 + \frac{1}{nh_n}\right) \to 0$$

From the above theorems, we can find that $\hat{b}_n^2(x)$ has a faster quadratic convergence rate than $\hat{a}_n(x)$, implying that the diffusion is more easily estimated than the drift. This is consistent with the result of the argument by Jiang and Knight (1999). □

4.3.3 Strong Consistency

In this part, we establish strong consistency, that is, $\hat{a}_n(x) \xrightarrow{a.s.} a(x)$ and $\hat{b}^2(x) \xrightarrow{a.s.} b^2(x)$. First, strong consistency of $\hat{a}_n(x)$ is proved. To do so, the two following lemmas are required. One is owing to Kronecker (Shiryaev, 1996) and the other is owing to Masry (1987):

Lemma 4.3.12 (Kronecker)**.** *If b_n is an increasing real sequence with $b_n \to \infty$, and x_n is a real*

sequence such that $\sum_{n=1}^{\infty} x_n$ *exists, then as* $n \to \infty$

$$\frac{1}{b_n}\sum_{i=1}^{n} b_i x_i \to 0$$

Lemma 4.3.13 (Masry)**.** *Let* $\{X_n\}$ *be an* α*-mixing process*[2] *and* $g_n(\cdot)$ *is a sequence of Borel measurable functions on the real line R. Let* $Z_i = g_i(X_i) - Eg_i(X_i)$*, and put* $S_n = \sum_{i=1}^{n} Z_i$*. If* $\sum_{i=1}^{\infty}[E|Z_i|^r]^{2/r} < \infty$ *and*

$$\sum_{n=1}^{\infty}(\log n)(\log_2 n)^{1+\delta}\alpha_n^{1-2/r}\cdot\sum_{i=n}^{\infty}[E|Z_i|^r]^{2/r} < \infty$$

for some $r > 2$ *and* $\delta > 0$*, then* S_n *converges almost surely to a finite limit as* $n \to \infty$*.*

In addition, we need the extra assumption:

A7 Suppose the bandwidth h_n and the mixing coefficient of the process ρ_n satisfy

i. $\sum_{i=1}^{\infty}[i^2 h_i^{3/2}\triangle]^{-1} < \infty$

i. $\sum_{n=1}^{\infty}(\log n)(\log_2 n)^{1+\delta}\rho_n^{1/2}\cdot\sum_{i=n}^{\infty}[i^2 h_i^{3/2}\triangle]^{-1} < \infty$

We use the expansion

$$\begin{aligned}\hat{a}_n(x) - a(x) &= \frac{\hat{g}_n(x) - a(x)\hat{f}_n(x)}{\hat{f}_n(x)} \\ &= \frac{\hat{g}_n(x) - a(x)\hat{f}_n(x) - E[\hat{g}_n(x) - a(x)\hat{f}_n(x)] + E[\hat{g}_n(x) - a(x)\hat{f}_n(x)]}{\hat{f}_n(x)}\end{aligned}$$

Then it can be proved that

Proposition 4.3.14. *Under* ***A1-A6****,*

$$E[\hat{g}_n(x) - a(x)\hat{f}_n(x)] \to 0$$

Proof. From Proposition 4.3.4 and 4.3.5, it can be seen that

$$E[\hat{g}_n(x) - a(x)\hat{f}_n(x)] = E\hat{g}_n(x) - a(x)E\hat{f}_n(x) = -a'(x)f(x)c_1\beta h_n + O(h_n^2 + \triangle) \to 0$$

□

[2]The reader who wishes a brief review on mixing processes is directed to Appendix C.

Proposition 4.3.15. *Under **A1**-**A7**, we have*

$$\hat{g}_n(x) - E\hat{g}_n(x) \xrightarrow{a.s.} 0$$

Proof. From the definition of $\hat{g}_n(x)$ and Lemma 4.3.3, we have

$$\hat{g}_n(x) - E\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n}[Y_i K_{h_i}(x - X_i) - EY_i K_{h_i}(x - X_i)]$$

Let $W_i = \frac{1}{i}[Y_i K_{h_i}(x - X_i) - EY_i K_{h_i}(x - X_i)]$, then it follows that

$$E|W_i|^4 = \frac{1}{i^4}E[Y_i K_{h_i}(x - X_i) - EY_i K_{h_i}(x - X_i)]^4$$

Now from Lemma 4.3.3, we have

$$\frac{1}{i^4}EY_i^4 K_{h_i}^4(x - X_i) = \frac{1}{i^4}\int_R K_{h_i}^4(x - z)\left[\frac{3b^4(z)}{\triangle^2} + O(1/\triangle)\right]f(z)dz = O\left(\frac{1}{i^4 h_i^3 \triangle^2}\right)$$

and

$$\begin{aligned}
&\frac{1}{i^4}EY_i^3 K_{h_i}^3(x - X_i) \cdot EY_i K_{h_i}(x - X_i) \\
= &\frac{1}{i^4}\int_R K_{h_i}^3(x - z)\left[\frac{3a(z)b^2(z) + \frac{3}{2}b^2(z)(b^2)'(z)}{\triangle} + O(1)\right]f(z)dz \\
&\cdot \int_R K_{h_i}(x - z)[a(z) + O(\triangle)]f(z)dz = O\left(\frac{1}{i^4 h_i^2 \triangle}\right)
\end{aligned}$$

and similarly we can obtain that

$$\begin{aligned}
\frac{1}{i^4}EY_i^2 K_{h_i}^2(x - X_i) \cdot [EY_i K_{h_i}(x - X_i)]^2 &= O\left(\frac{1}{i^4 h_i \triangle}\right) \\
\frac{1}{i^4}[EY_i K_{h_i}(x - X_i)]^4 &= O\left(\frac{1}{i^4}\right).
\end{aligned}$$

Thus by combining the above expressions, it follows that $E|W_i|^4 = O\left(\frac{1}{i^4 h_i^3 \triangle^2}\right)$, which implies that

$$\sum_{i=1}^{\infty}[E|W_i|^4]^{1/2} = \sum_{i=1}^{\infty} O\left(\frac{1}{i^2 h_i^{3/2} \triangle}\right) < \infty.$$

Note that **A7** and Lemma 4.3.13 imply that $\sum_{i=1}^{n} W_i$ converges almost surely to a finite limit.

Also it can be seen that

$$\hat{g}_n(x) - E\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} iW_i.$$

From Lemma 4.3.12 by taking $b_i = i$, it follows that as $n \to \infty$

$$\hat{g}_n(x) - E\hat{g}_n(x) \xrightarrow{a.s.} 0.$$

□

Following the above pattern, we can obtain that $\hat{f}_n(x) - E\hat{f}_n(x) \xrightarrow{a.s.} 0$. Therefore from Proposition 4.3.14 and 4.3.15, we have

Theorem 4.3.16. *Under **A1**-**A7**,*

$$\hat{a}_n(x) \xrightarrow{a.s.} a(x)$$

Proof. As mentioned above, it is noted that

$$\begin{aligned}\hat{a}_n(x) - a(x) &= \frac{\hat{g}_n(x) - a(x)\hat{f}_n(x) - E[\hat{g}_n(x) - a(x)\hat{f}_n(x)] + E[\hat{g}_n(x) - a(x)\hat{f}_n(x)]}{\hat{f}_n(x)} \\ &= \frac{[\hat{g}_n(x) - E\hat{g}_n(x)] - a(x)[\hat{f}_n(x) - E\hat{f}_n(x)]}{\hat{f}_n(x)} + \frac{E[\hat{g}_n(x) - a(x)\hat{f}_n(x)]}{\hat{f}_n(x)}\end{aligned}$$

We have shown that $\hat{g}_n(x) - E\hat{g}_n(x) \xrightarrow{a.s.} 0$, $\hat{f}_n(x) - E\hat{f}_n(x) \xrightarrow{a.s.} 0$ and $E[\hat{g}_n(x) - a(x)\hat{f}_n(x)] \to 0$. Meanwhile in Proposition 4.3.5, $E\hat{f}_n(x) \to f(x)$, thus

$$\hat{f}_n(x) = \hat{f}_n(x) - E\hat{f}_n(x) + E\hat{f}_n(x) \xrightarrow{a.s.} f(x)$$

Therefore it follows that

$$\hat{a}_n(x) \xrightarrow{a.s.} a(x)$$

□

Now to prove strong consistency of $\hat{b}_n^2(x)$, we need the following lemma

Proposition 4.3.17. *Under **A1**-**A7**, we have*

(1) $E[\hat{d}_n(x) - b^2(x)\hat{f}_n(x)] \to 0$

(1) $\hat{d}_n(x) - E\hat{d}_n(x) \xrightarrow{a.s.} 0$

Proof. For (1), it is noted that

$$E[\hat{d}_n(x) - b^2(x)\hat{f}_n(x)] = E\hat{d}_n(x) - b^2(x)E\hat{f}_n(x) = -(b^2)'f(x)c_1\beta h_n + O(h_n^2 + \triangle) \to 0$$

For (2), we have

$$\hat{d}_n(x) - E\hat{d}_n(x) = \frac{1}{n}\sum_{i=1}^{n}[Z_i K_{h_i}(x - X_i) - EZ_i K_{h_i}(x - X_i)] \triangleq \frac{1}{n}\sum_{i=1}^{n} iW_i$$

where

$$E|W_i|^4 = \frac{1}{i^4}E[Z_i K_{h_i}(x - X_i) - EZ_i K_{h_i}(x - X_i)]^4$$

It follows that

$$\begin{aligned}
\frac{1}{i^4}EZ_i^4 K_{h_i}^4(x - X_i) &= O\left(\frac{1}{i^4 h_i^3}\right) \\
\frac{1}{i^4}EZ_i^3 K_{h_i}^3(x - X_i) \cdot EZ_i K_{h_i}(x - X_i) &= O\left(\frac{1}{i^4 h_i^2}\right) \\
\frac{1}{i^4}EZ_i^2 K_{h_i}^2(x - X_i) \cdot [EZ_i K_{h_i}(x - X_i)]^2 &= O\left(\frac{1}{i^4 h_i}\right)
\end{aligned}$$

thus **A7** implies that $\sum_{i=1}^{\infty}[E|W_i|^4]^{1/2} < \infty$. And also from Lemma 4.3.13 and Lemma 4.3.12, as $n \to \infty$, we have $\hat{d}_n(x) - E\hat{d}_n(x) \xrightarrow{a.s.} 0$. □

So the following theorem can be obtained

Theorem 4.3.18. *Under **A1**-**A7**, we have*

$$\hat{b}_n^2(x) \xrightarrow{a.s.} b^2(x)$$

Proof. It can be seen that

$$\hat{b}_n^2(x) - b^2(x) = \frac{[\hat{d}_n(x) - E\hat{d}_n(x)] - a(x)[\hat{f}_n(x) - E\hat{f}_n(x)]}{\hat{f}_n(x)} + \frac{E[\hat{d}_n(x) - a(x)\hat{f}_n(x)]}{\hat{f}_n(x)}$$

By Proposition 4.3.17 and $\hat{f}_n(x) \xrightarrow{a.s.} f(x)$ in Theorem 4.3.16, we can prove that $\hat{b}_n^2(x) \xrightarrow{a.s.} b^2(x)$. □

4.3.4 Asymptotic Normality

Before showing asymptotic normality of $\hat{a}_n(x)$, the following assumption and lemma are useful:

A8 The bandwidth and discretization step size satisfy $nh_n^3\triangle \to 0$ as $k \to \infty$ and $n \to \infty$

Lemma 4.3.19. *Under **A1**-**A6**, we have for any i and j*

$$|Cov[Y_iK_{h_i}(x-X_i), Y_jK_{h_j}(x-X_j)]| = O\left[\left(h_ih_j\triangle^2\right)^{-1/2}\right] \tag{4.18}$$

$$|Cov[Y_iK_{h_i}(x-X_i), K_{h_j}(x-X_j)]| = O\left[\left(h_ih_j\triangle\right)^{-1/2}\right] \tag{4.19}$$

$$|Cov[K_{h_i}(x-X_i), K_{h_j}(x-X_j)]| = O\left[\left(h_ih_j\right)^{-1/2}\right] \tag{4.20}$$

Proof. Here we only prove (4.18) because the same steps can be applied to (4.19) and (4.20). Note that

$$\begin{aligned} & |Cov[Y_iK_{h_i}(x-X_i), Y_jK_{h_j}(x-X_j)]| \\ \leq\ & |EY_iY_jK_{h_i}(x-X_i)K_{h_j}(x-X_j)| + |EY_iK_{h_i}(x-X_i)| \cdot |EY_jK_{h_j}(x-X_j)| \end{aligned}$$

where by the Hölder inequality, Lemma 4.3.3 and Lemma 4.3.1

$$\begin{aligned} & |EY_iY_jK_{h_i}(x-X_i)K_{h_j}(x-X_j)| \leq E|Y_iY_jK_{h_i}(x-X_i)K_{h_j}(x-X_j)| \\ \leq\ & \left[EY_i^2K_{h_i}^2(x-X_i) \cdot EY_j^2K_{h_j}^2(x-X_j)\right]^{1/2} \\ =\ & \left\{h_i^{-1}\int_R h_i^{-1}K^2\left(\frac{x-z}{h_i}\right)[\triangle^{-1}b^2(z)+O(1)]f(z)dz\right\}^{1/2} \\ & \cdot\left\{h_j^{-1}\int_R h_j^{-1}K^2\left(\frac{x-z}{h_j}\right)[\triangle^{-1}b^2(z)+O(1)]f(z)dz\right\}^{1/2} \\ =\ & (h_ih_j\triangle^2)^{-1/2}b^2(x)f(x)\int_R K^2(u)du + o\left[\left(h_ih_j\triangle^2\right)^{-1/2}\right] = O\left[\left(h_ih_j\triangle^2\right)^{-1/2}\right] \end{aligned}$$

and

$$\begin{aligned} & |EY_iK_{h_i}(x-X_i)| \cdot |EY_jK_{h_j}(x-X_j)| \\ =\ & \left|\int_R K_{h_i}(x-z)[a(z)+O(\triangle)]f(z)dz \cdot \int_R K_{h_j}(x-z)[a(z)+O(\triangle)]f(z)dz\right| \\ =\ & a^2(x)f^2(x) + O(h_i+h_j+\triangle) = o\left[\left(h_ih_j\triangle^2\right)^{-1/2}\right] \end{aligned}$$

therefore we have

$$|Cov[Y_iK_{h_i}(x-X_i), Y_jK_{h_j}(x-X_j)]| = O\left[\left(h_ih_j\triangle^2\right)^{-1/2}\right]$$

□

In order to establish asymptotic normality of $\hat{a}_n(x)$, it can be seen that

$$\hat{a}_n - a(x) = \frac{[\hat{g}_n(x) - g(x)] - a(x)[\hat{f}_n(x) - f(x)]}{\hat{f}_n(x)}$$

where $g(x) = a(x)f(x)$. Then in the following propositions, we will show that for $\lambda_1^2 + \lambda_2^2 \neq 0$

$$\sqrt{nh_n\triangle}\{\lambda_1[\hat{g}_n(x) - g(x)] + \lambda_2[\hat{f}_n(x) - f(x)]\} \to \mathcal{N}(0, \iota^2(x))$$

where

$$\iota^2(x) = \lambda_1^2\alpha_1 b^2(x)f(x)\int_R K^2(u)du \tag{4.21}$$

In Proposition 4.3.4 and 4.3.5 and **A8**, we have established that

$$\sqrt{nh_n\triangle}[E\hat{g}_n(x) - g(x)] \to 0 \qquad \sqrt{nh_n\triangle}[E\hat{f}_n(x) - f(x)] \to 0$$

therefore we need to prove that

$$\sqrt{nh_n\triangle}\{\lambda_1[\hat{g}_n(x) - E\hat{g}_n(x)] + \lambda_2[\hat{f}_n(x) - E\hat{f}_n(x)]\} \to \mathcal{N}(0, \iota^2(x))$$

Let

$$U_i = \sqrt{h_n\triangle}\{\lambda_1[Y_iK_{h_i}(x - X_i) - EY_iK_{h_i}(x - X_i)] + \lambda_2[K_{h_i}(x - X_i) - EK_{h_i}(x - X_i)]\}$$

and $S_n = \sum_{i=1}^n U_i$, then it is equivalent to prove $n^{-1/2}S_n \to \mathcal{N}(0, \iota^2(x))$. Here we borrow the "big block" and "small block" technique introduced by Doob (1953) and Masry (1986) to prove asymptotic normality of $n^{-1/2}S_n$. Partition the set $\Xi_n = \{1, 2, \ldots, n\}$ into $2r + 1$ subsets

$$\Xi_n = \bigcup_{i=0}^{r-1}\Xi_{i,n}^{'} + \bigcup_{i=0}^{r-1}\Xi_{i,n}^{''} + \Xi_{r,n}^{'''}$$

where

$$\begin{aligned}
\Xi_{i,n}^{'} &= \{i(p+q)+1, \ldots, i(p+q)+p\} \quad \text{for } i = 0, 1, \ldots, r-1 \\
\Xi_{i,n}^{''} &= \{i(p+q)+p+1, \ldots, (i+1)(p+q)\} \quad \text{for } i = 0, 1, \ldots, r-1 \\
\Xi_{r,n}^{'''} &= \{r(p+q)+1, \ldots, n\}
\end{aligned}$$

In the above, $p = p_n, q = q_n, r = r_n$ depend on n such that

$$\frac{p_n^2}{nh_n\triangle} \to 0 \qquad \frac{q_nr_n}{n} \to 0$$

For $j = 0, \ldots, r-1$

$$\eta_j = \sum_{i \in \Xi'_{j,n}} U_i \qquad \xi_j = \sum_{i \in \Xi''_{j,n}} U_i$$

and

$$\zeta_r = \sum_{i \in \Xi'''_{r,n}} U_i$$

Then

$$S_n = \sum_{i=1}^{n} U_i = \sum_{j=0}^{r-1} \eta_j + \sum_{j=0}^{r-1} \xi_j + \zeta_r \triangleq S'_n + S''_n + S'''_n$$

It is easy to prove that

Lemma 4.3.20. *Under **A1**-**A6**,*

$$Var(U_i) = O(h_n h_i^{-1})$$

Proof. From the definition of U_i, we know that

$$\begin{aligned} Var(U_i) &= h_n \triangle \{\lambda_1^2 Var[Y_i K_{h_i}(x - X_i)] + \lambda_2^2 Var[K_{h_i}(x - X_i)] \\ &\quad + 2\lambda_1 \lambda_2 Cov[Y_i K_{h_i}(x - X_i), K_{h_i}(x - X_i)]\} \end{aligned}$$

By Lemma 4.3.19 for the case $i = j$,

$$\begin{aligned} Var[Y_i K_{h_i}(x - X_i)] &= O\left(h_i^{-1} \triangle^{-1}\right) \\ Var[K_{h_i}(x - X_i)] &= O\left(h_i^{-1}\right) \\ Cov[Y_i K_{h_i}(x - X_i), K_{h_i}(x - X_i)] &= O\left(h_i^{-1} \triangle^{-1/2}\right) \end{aligned}$$

thus

$$Var(U_i) = h_n \triangle O(h_i^{-1} \triangle^{-1}) = O(h_n h_i^{-1})$$

□

Proposition 4.3.21. *Under **A1**-**A6**, we have*

$$n^{-1} E[S'''_n]^2 \to 0$$

Proof. It can be found that

$$\frac{1}{n}E[S_n^{'''}]^2 = \frac{1}{n}Var(S_n^{'''}) = \frac{1}{n}\sum_{i=r(p+q)+1}^{n} Var(U_i) + \frac{2}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} Cov(U_i, U_j)$$

where by Lemma 4.3.20

$$\frac{1}{n}\sum_{i=r(p+q)+1}^{n} Var(U_i) = \frac{1}{n}\sum_{i=r(p+q)+1}^{n} O(h_n h_i^{-1}) = \frac{n-r(p+q)}{n}O(1) = \left(1 - \frac{qr}{n} - \frac{pr}{n}\right)O(1)$$

Note that $qr/n \to 0$ and $pr/n \to 1$, thus

$$\frac{1}{n}\sum_{i=r(p+q)+1}^{n} Var(U_i) \to 0 \tag{4.22}$$

On the other hand, by the Cauchy-Schwartz inequality

$$\begin{aligned}
\frac{1}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} Cov(U_i, U_j) &\leq \frac{1}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} |Cov(U_i, U_j)| \leq \frac{1}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} |\rho(j-i)|\sqrt{Var(U_i)}\sqrt{Var(U_j)} \\
&\leq \frac{1}{n}\left[\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} \rho^2(j-i)\right]^{1/2}\left[\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} Var(U_i)\cdot Var(U_j)\right]^{1/2} \\
&= \frac{1}{n}\left[\sum_{k=1}^{n-r(p+q)} k\rho^2(k)\right]^{1/2}\left[\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} O(h_n^2 h_i^{-1} h_j^{-1})\right]^{1/2} \\
&= \left(1 - \frac{qr}{n} - \frac{pr}{n}\right)O(1)\cdot\left[\sum_{k=1}^{n-r(p+q)} k\rho^2(k)\right]^{1/2}
\end{aligned}$$

Since the ρ-mixing coefficient has exponential decay, we have $\sum_{n=1}^{\infty} n\rho^2(n) < \infty$, implying that $\left[\sum_{k=1}^{n-r(p+q)} k\rho^2(k)\right]^{1/2}$ is bounded by some constant. So

$$\frac{1}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} |Cov(U_i, U_j)| \to 0 \tag{4.23}$$

Through (4.22) and (4.23), it can be found that

$$0 \leq n^{-1}E[S_n^{'''}]^2 \leq \frac{1}{n}\sum_{i=r(p+q)+1}^{n} Var(U_i) + \frac{2}{n}\sum_{\substack{i,j=r(p+q)+1 \\ i<j}}^{n} |Cov(U_i, U_j)| \to 0$$

that is, $n^{-1}E[S_n^{'''}]^2 \to 0$. □

Proposition 4.3.22. *Under **A1-A6**, we have*

$$n^{-1}E[S_n^{''}]^2 \to 0$$

Proof. Note that

$$\frac{1}{n}E[S_n^{''}]^2 = \frac{1}{n}Var(S_n^{''}) = \frac{1}{n}Var\left(\sum_{j=0}^{r-1}\xi_j\right) = \frac{1}{n}\sum_{j=0}^{r-1}Var(\xi_j) + \frac{2}{n}\sum_{\substack{i,j=0\\i<j}}^{r-1}Cov(\xi_i,\xi_j)$$

Let $l_j = j(p+q)+p$, we have

$$\frac{1}{n}\sum_{j=0}^{r-1}Var(\xi_j) = \frac{1}{n}\sum_{j=0}^{r-1}\sum_{i=1}^{q}Var(U_{l_j+i}) + \frac{2}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}Cov(U_{l_j+i},U_{l_j+k})$$

Similar to Proposition 4.3.21, we have

$$\frac{1}{n}\sum_{j=0}^{r-1}\sum_{i=1}^{q}Var(U_{l_j+i}) = \frac{1}{n}\sum_{j=0}^{r-1}\sum_{i=1}^{q}O(h_n h_{l_j+i}^{-1}) = \frac{qr}{n}O(1) = o(1)$$

and

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}Cov(U_{l_j+i},U_{l_j+k}) &\le \frac{1}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}|Cov(U_{l_j+i},U_{l_j+k})|\\ &= \frac{1}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}|\rho(k-i)|\sqrt{Var(U_{l_j+i})}\sqrt{Var(U_{l_j+k})}\end{aligned}$$

by the Cauchy-Schwartz inequality,

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}|Cov(U_{l_j+i},U_{l_j+k})| &\le \frac{1}{n}\left[\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}\rho^2(k-i)\right]^{1/2}\left[\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}Var(U_{l_j+i})Var(U_{l_j+k})\right]^{1/2}\\ &= \frac{1}{n}\left[\sum_{j=0}^{r-1}\sum_{k=1}^{q-1}k\rho^2(k)\right]^{1/2}\left[\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}O(h_n^2h_{l_j+i}^{-1}h_{l_j+k}^{-1})\right]^{1/2}\\ &\le \frac{1}{n}\left[\sum_{j=0}^{r-1}\sum_{k=1}^{\infty}k\rho^2(k)\right]^{1/2}\left[\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\i<k}}^{q}O(h_n^2h_{l_j+i}^{-1}h_{l_j+k}^{-1})\right]^{1/2}\end{aligned}$$

Since the ρ-mixing coefficient has exponential decay, we have $\sum_{k=1}^{\infty} k\rho^2(k) < \infty$, implying that $\sum_{j=0}^{r-1}\sum_{k=1}^{\infty} k\rho^2(k) = O(r)$. Thus

$$\frac{1}{n}\sum_{j=0}^{r-1}\sum_{\substack{i,k=1\\ i<k}}^{q} |Cov(U_{l_j+i}, U_{l_j+k})| \leq \frac{\sqrt{r}\sqrt{q^2 r}}{n} O(1) = \frac{qr}{n} O(1) = o(1)$$

So we have

$$\frac{1}{n}\sum_{j=0}^{r-1} Var(\xi_j) = o(1). \tag{4.24}$$

On the other hand, it can be seen that

$$\begin{aligned}\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} Cov(\xi_i, \xi_j) &\leq \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} |Cov(\xi_i, \xi_j)| \leq \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1\in\Xi''_{i,n}}\sum_{k_2\in\Xi''_{j,n}} |Cov(U_{k_1}, U_{k_2})| \\ &= \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} |Cov(U_{l_i+k_1}, U_{l_j+k_2})|\end{aligned}$$

that is,

$$\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} |Cov(\xi_i, \xi_j)| \leq \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} |\rho(l_j + k_2 - l_i - k_1)| \sqrt{Var(U_{l_i+k_1})}\sqrt{Var(U_{l_j+k_2})}$$

by the Cauchy-Schwartz inequality,

$$\begin{aligned}\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} |Cov(\xi_i, \xi_j)| &\leq \frac{1}{n}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} \rho^2(l_j + k_2 - l_i - k_1)\right]^{1/2}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} Var(U_{l_i+k_1})Var(U_{l_j+k_2})\right]^{1/2} \\ &= \frac{1}{n}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} \rho^2(l_j + k_2 - l_i - k_1)\right]^{1/2}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} O(h_n^2 h_{l_i+k_1}^{-1} h_{l_j+k_2}^{-1})\right]^{1/2} \\ &= \frac{qr}{n} O(1)\cdot\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} \rho^2(l_j + k_2 - l_i - k_1)\right]^{1/2}.\end{aligned}$$

Note that the difference of indices $l_j + k_2 - l_i - k_1$ is at least p such that

$$\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=1}^{q}\sum_{k_2=1}^{q} \rho^2(l_j + k_2 - l_i - k_1) \leq \sum_{i=1}^{n-p}\sum_{j=i+p}^{n} \rho^2(j-i) = \sum_{k=p}^{n-1} k\rho^2(k) \leq \sum_{k=p}^{\infty} k\rho^2(k).$$

As $p \to \infty$, it can be seen that $\sum_{k=p}^{\infty} k\rho^2(k) \to 0$. So we have

$$\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} |Cov(\xi_i, \xi_j)| = o(1). \tag{4.25}$$

Through (4.24) and (4.25) it follows that $n^{-1}E[S_n^{''}]^2 \to 0$. □

Proposition 4.3.23. *Under **A1**-**A6**, then*

$$\frac{1}{n}\sum_{j=0}^{r-1} Var(\eta_j) \to \iota^2(x)$$

where $\iota^2(x)$ is defined in (4.21).

Proof. Note that

$$\frac{1}{n}E[S_n^{'}]^2 = \frac{1}{n}Var(S_n^{'}) = \frac{1}{n}Var\left(\sum_{j=0}^{r-1}\eta_j\right) = \frac{1}{n}\sum_{j=0}^{r-1} Var(\eta_j) + \frac{2}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} Cov(\eta_i, \eta_j)$$

We have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} Cov(\eta_i, \eta_j)\right| \le \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} |Cov(\eta_i, \eta_j)| \le \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1\in\Xi_{i,n}^{'}}\sum_{k_2\in\Xi_{j,n}^{'}} |Cov(U_{k_1}, U_{k_2})| \\
= &\ \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} |Cov(U_{k_1}, U_{k_2})| \\
= &\ \frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} |\rho(k_2-k_1)|\sqrt{Var(U_{k_1})}\sqrt{Var(U_{k_2})} \\
\le &\ \frac{1}{n}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} \rho^2(k_2-k_1)\right]^{1/2}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} Var(U_{k_1})Var(U_{k_2})\right]^{1/2} \\
= &\ \frac{1}{n}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} \rho^2(k_2-k_1)\right]^{1/2}\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} O(h_n^2 h_{k_2}^{-1} h_{k_1}^{-1})\right]^{1/2} \\
= &\ \frac{pr}{n}O(1)\cdot\left[\sum_{\substack{i,j=0\\ i<j}}^{r-1}\sum_{k_1=i(p+q)+1}^{i(p+q)+p}\sum_{k_2=j(p+q)+1}^{j(p+q)+p} \rho^2(k_2-k_1)\right]^{1/2} \le \frac{pr}{n}O(1)\cdot\left[\sum_{i=1}^{n-q}\sum_{j=i+q}^{n}\rho^2(j-i)\right]^{1/2} \\
= &\ \frac{pr}{n}O(1)\cdot\left[\sum_{k=q}^{n-1} k\rho^2(k)\right]^{1/2} \le \frac{pr}{n}O(1)\cdot\left[\sum_{k=q}^{\infty} k\rho^2(k)\right]^{1/2}
\end{aligned}$$

Because $pr/n \to 1$ and $\sum_{k=q}^{\infty} k\rho^2(k) \to 0$, we have

$$\frac{1}{n}\sum_{\substack{i,j=0\\ i<j}}^{r-1} Cov(\eta_i, \eta_j) \to 0 \tag{4.26}$$

thus $n^{-1}\sum_{j=0}^{r-1} Var(\eta_j)$ has the same limit as $n^{-1}E[S_n^{'}]^2$. From the definition of S_n,

$$n^{-1}ES_n^2 = n^{-1}Var(S_n) = nh_n\triangle\{\lambda_1^2 Var[\hat{g}_n(x)] + 2\lambda_1\lambda_2 Cov[\hat{g}_n(x), \hat{f}_n(x)] + \lambda_2^2 Var[\hat{f}_n(x)]\}$$

The proof of Proposition 4.3.8 indicates that $nh_n\triangle \cdot Var[\hat{g}_n(x)] \to \alpha_1 b^2(x)f(x)\int_R K^2(u)du$ and $nh_n\triangle \cdot Cov[\hat{g}_n(x), \hat{f}_n(x)] = O(\triangle^{1/2})$ and $nh_n\triangle \cdot Var[\hat{f}_n(x)] = O(\triangle)$. Hence as $\triangle \to 0$ and $nh_n\triangle \to \infty$ in **A5** and **A6**, we have

$$n^{-1}ES_n^2 \to \lambda_1^2\alpha_1 b^2(x)f(x)\int_R K^2(u)du = \iota^2(x)$$

In addition, through Proposition 4.3.21 and 4.3.22 as well as the relationship $S_n = S_n^{'} + S_n^{''} + S_n^{'''}$, it can be found that

$$\begin{aligned}
n^{-1}\left|Cov(S_n^{'}, S_n^{''})\right| &\leq \sqrt{n^{-1}Var(S_n^{'})}\sqrt{n^{-1}Var(S_n^{''})} \to 0\\
n^{-1}\left|Cov(S_n^{'}, S_n^{'''})\right| &\leq |\sqrt{n^{-1}Var(S_n^{'})}\sqrt{n^{-1}Var(S_n^{'''})} \to 0\\
n^{-1}\left|Cov(S_n^{''}, S_n^{'''})\right| &\leq \sqrt{n^{-1}Var(S_n^{''})}\sqrt{n^{-1}Var(S_n^{'''})} \to 0
\end{aligned}$$

so $n^{-1}E[S_n^{'}]^2$ has the same limit as $n^{-1}ES_n^2$, thus $n^{-1}\sum_{j=0}^{r-1} Var(\eta_j) \to \iota^2(x)$. □

Before proving asymptotic normality of S_n, the following lemma from Volkonskii and Rozanov (1959) is useful where $\alpha(c)$ is the α-mixing coefficient as defined and described in Appendix C:

Lemma 4.3.24 (Volkonskii and Rozanov)**.** *Let random variables $Z_1, Z_2, \ldots, Z_r$ be measurable with respect to $\mathscr{F}_{k_1}^{l_1}, \mathscr{F}_{k_2}^{l_2}, \cdots, \mathscr{F}_{k_m}^{l_m}$ respectively for $1 \leq k_1 < l_1 < k_2 < l_2 < \ldots < k_r < l_r \leq n$, $k_{i+1} - l_i \geq c \geq 1$ and $|Z_i| \leq 1, i = 1, 2, \ldots, r$. Then we have*

$$\left|E\prod_{i=1}^{r} Z_i - \prod_{i=1}^{r} EZ_i\right| \leq 16(r-1)\alpha(c)$$

Proposition 4.3.25. *Under **A1**-**A6**, it can be obtained that*

$$n^{-1/2}S_n \to \mathcal{N}(0, \iota^2(x))$$

where $\iota^2(x)$ is defined in (4.21).

Proof. By the above propositions, it has been shown that $n^{-1}E[S_n^{''}]^2 \to 0$ and $n^{-1}E[S_n^{'''}]^2 \to 0$. Thus in order to prove asymptotic normality of S_n, we only need to show that $n^{-1/2}S_n^{'} \to \mathcal{N}(0, \iota^2(x))$. As pointed by Bradley (1983a), we can always find the independent random variables $\eta_j^{'}$ for $j = 0, 1, \ldots, r-1$ such that $\eta_j^{'}$ and $n^{-1/2}\eta_j$ have the same distribution, thus $E\eta_j^{'} = 0$. Let $\Phi_j(t)$ be the characteristic function of η_j, then $\prod_{j=0}^{r-1}\Phi_j(n^{-1/2}t)$ is the characteristic function of $\sum_{j=0}^{r-1}\eta_j^{'}$. By Lemma 4.3.24,

$$\begin{aligned}\left|Ee^{itn^{-1/2}S_n^{'}} - \prod_{j=0}^{r-1} Ee^{itn^{-1/2}\eta_j}\right| &= \left|E\prod_{j=0}^{r-1} e^{itn^{-1/2}\eta_j} - \prod_{j=0}^{r-1} Ee^{itn^{-1/2}\eta_j}\right| \\ &\leq 16(r-1)\alpha(q) \leq 4(r-1)\rho(q) \to 0\end{aligned}$$

So it suffices to show that $\prod_{j=0}^{r-1} Ee^{itn^{-1/2}\eta_j} = \prod_{j=0}^{r-1}\Phi_j(n^{-1/2}t)$ converges to the characteristic function of $\mathcal{N}(0, \iota^2(x))$. To this end, set $\eta_j^{''} = \eta_j^{'}/s_n$ where by Proposition 4.3.23

$$s_n^2 = \sum_{j=0}^{r-1} Var(\eta_j^{'}) = \frac{1}{n}\sum_{j=0}^{r-1} Var(\eta_j) \to \iota^2(x)$$

thus it is easy to see that for the independent random variables $\eta_j^{''}$, we have $E\eta_j^{''} = 0$ and $\sum_{j=0}^{r-1} Var(\eta_j^{''}) = 1$. Then we check the Lindeberg-Feller condition for $\eta_j^{''}$, that is, for all $\varepsilon > 0$,

$$\sum_{j=0}^{r-1} E\left[(\eta_j^{''})^2 I(|\eta_j^{''}| > \varepsilon)\right] \to 0$$

By the Hölder inequality, we can find that

$$\sum_{j=0}^{r-1} E\left[(\eta_j^{''})^2 I(|\eta_j^{''}| > \varepsilon)\right] = \sum_{j=0}^{r-1} E\left[\frac{(\eta_j^{'})^2}{s_n^2} I(|\eta_j^{'}| > \varepsilon s_n)\right] \leq \frac{\|\eta_j^{'}\|_\infty^2}{s_n^2}\sum_{j=0}^{r-1} P(|\eta_j^{'}| > \varepsilon s_n)$$

Note that

$$\begin{aligned}\|\eta_j^{'}\|_\infty^2 &= \|n^{-1/2}\eta_j\|_\infty^2 = n^{-1}\|\eta_j\|_\infty^2 = n^{-1}\left\|\sum_{i=j(p+q)+1}^{j(p+q)+p} U_i\right\|_\infty^2 \\ &= n^{-1}\left\|\sqrt{h_n\triangle}\sum_{i=j(p+q)+1}^{j(p+q)+p}\{\lambda_1[Y_iK_{h_i}(x-X_i) - EY_iK_{h_i}(x-X_i)]\right. \\ &\quad \left. +\lambda_2[K_{h_i}(x-X_i) - EK_{h_i}(x-X_i)]\}\right\|_\infty^2 \\ &= n^{-1}h_n\triangle\left\|\sum_{i=j(p+q)+1}^{j(p+q)+p} O\left(\frac{1}{h_i\triangle}\right)\right\|_\infty^2 = O\left(\frac{p^2}{nh_n\triangle}\right) = o(1)\end{aligned}$$

so by the Markov inequality

$$\sum_{j=0}^{r-1} E\left[(\eta_j^{''})^2 I(|\eta_j^{''}| > \varepsilon)\right] \leq \frac{1}{s_n^2}\sum_{j=0}^{r-1}\frac{E(\eta_j^{'})^2}{\varepsilon^2 s_n^2}o(1) = \frac{1}{\varepsilon^2 s_n^2}o(1) = o(1)$$

This completes the proof of $n^{-1/2}S_n \to \mathcal{N}(0, \iota^2(x))$. □

Now we proceed with the following theorem to characterize asymptotic normality of $\hat{a}_n(x)$ by using the above propositions

Theorem 4.3.26. *Under **A1**-**A6** and **A8**, we have*

$$\sqrt{nh_n\triangle}[\hat{a}_n(x) - a(x)] \xrightarrow{d} \mathcal{N}\left(0, b^2(x)f^{-1}(x)\int_R K^2(u)du\right)$$

Proof. By above propositions, we have proved that

$$\sqrt{nh_n\triangle}\{\lambda_1[\hat{g}_n(x) - g(x)] + \lambda_2[\hat{f}_n(x) - f(x)]\} \to \mathcal{N}(0, \iota^2(x))$$

therefore, by Wold's device and **A8**, we can obtain

$$\sqrt{nh_n\triangle}[\hat{a}_n(x) - a(x)] \xrightarrow{d} \mathcal{N}\left(0, b^2(x)f^{-1}(x)\int_R K^2(u)du\right)$$

□

Similarly we present the sketch of the proof of asymptotic normality of $\hat{b}^2(x)$. In fact we can prove that

Lemma 4.3.27. *Under **A1**-**A6**, it can be calculated that*

(1) For any i and j,

(i) $|Cov[Z_iK_{h_i}(x - X_i), Z_jK_{h_j}(x - X_j)]| = O\left[(h_ih_j)^{-1/2}\right]$

(i) $|Cov[Z_iK_{h_i}(x - X_i), K_{h_j}(x - X_j)]| = O\left[(h_ih_j)^{-1/2}\right]$

(i) $Var(U_i) = O(h_nh_i^{-1})$

(1) Let $U_i = \sqrt{h_n}\{\lambda_1[Z_iK_{h_i}(x - X_i) - EZ_iK_{h_i}(x - X_i)] + \lambda_2[K_{h_i}(x - X_i) - EK_{h_i}(x - X_i)]\}$ *and* $S_n \triangleq S_n^{'} + S_n^{''} + S_n^{'''}$ *on the partition of* $\Xi_n = \{1, 2, \ldots, n\} = \bigcup_{i=0}^{n-1}\Xi_{i,n}^{'} + \bigcup_{i=0}^{n-1}\Xi_{i,n}^{''} + \Xi_{i,n}^{'''}$ *respectively, then*

(i) $n^{-1}E[S_n^{'''}]^2 \to 0$

(i) $n^{-1}E[S_n^{''}]^2 \to 0$

$$(i)\ \ n^{-1}E[S_n^{'}]^2 \to \mathcal{N}\left(0, \left[3\lambda_1^2 b^4(x) + 2\lambda_1\lambda_2 b^2(x) + \lambda_2^2\right]\alpha_1 f(x)\int_R K^2(u)du\right)$$

Proof. For (1), we have

$$\begin{aligned}
& |Cov[Z_iK_{h_i}(x-X_i), Z_jK_{h_j}(x-X_j)]| \\
\leq\ & |EZ_iZ_jK_{h_i}(x-X_i)K_{h_j}(x-X_j)| + |EZ_iK_{h_i}(x-X_i)|\cdot|EZ_jK_{h_j}(x-X_j)| \\
\leq\ & \left[EZ_i^2K_{h_i}^2(x-X_i)\cdot EZ_j^2K_{h_j}^2(x-X_j)\right]^{1/2} + |EY_iK_{h_i}(x-X_i)|\cdot|EY_jK_{h_j}(x-X_j)| \\
=\ & O\left[(h_ih_j)^{-1/2}\right]
\end{aligned}$$

similarly it follows that $|Cov[Z_iK_{h_i}(x-X_i), K_{h_j}(x-X_j)]| = O\left[(h_ih_j)^{-1/2}\right]$. We can also obtain that

$$\begin{aligned}
Var(U_i) \ =\ & h_n\{\lambda_1^2 Var[Z_iK_{h_i}(x-X_i)] + \lambda_2^2 Var[K_{h_i}(x-X_i)] \\
& +2\lambda_1\lambda_2 Cov[Z_iK_{h_i}(x-X_i), K_{h_i}(x-X_i)]\} = O(h_nh_i^{-1})
\end{aligned}$$

For (2), note that the proofs of Proposition 4.3.21 and 4.3.22 use the properties of $Var(U_i) = O(h_nh_i^{-1})$ and the mixing coefficients, and $Var(U_i) = O(h_nh_i^{-1})$ is applied to cases of both $\hat{a}_n(x)$ and $\hat{b}_n^2(x)$. So we can use the same arguments to prove (i) and (ii). Thus here it suffices to show (iii). From Lemma 4.3.10, it can be seen that

$$n^{-1}ES_n^2 = nh_n\{\lambda_1^2 Var[\hat{d}_n(x)] + 2\lambda_1\lambda_2 Cov[\hat{d}_n(x), \hat{f}_n(x)] + \lambda_2^2 Var[\hat{f}_n(x)]\}$$

where

$$\begin{aligned}
nh_nVar[\hat{d}_n(x)] \ &\to\ 3\alpha_1 b^4(x)f(x)\int_R K^2(u)du \\
nh_nCov[\hat{d}_n(x), \hat{f}_n(x)] \ &\to\ \alpha_1 b^2(x)f(x)\int_R K^2(u)du \\
nh_nVar[\hat{f}_n(x)] \ &\to\ \alpha_1 f(x)\int_R K^2(u)du
\end{aligned}$$

so

$$n^{-1}ES_n^2 \to \left[3\lambda_1^2 b^4(x) + 2\lambda_1\lambda_2 b^2(x) + \lambda_2^2\right]\alpha_1 f(x)\int_R K^2(u)du$$

implying

$$n^{-1}E[S_n^{'}]^2 \to \left[3\lambda_1^2 b^4(x) + 2\lambda_1\lambda_2 b^2(x) + \lambda_2^2\right]\alpha_1 f(x)\int_R K^2(u)du.$$

We can use the proof of Proposition 4.3.25 to prove that

$$n^{-1/2}S_n^{'} \to \mathcal{N}\left(0, \left[3\lambda_1^2 b^4(x) + 2\lambda_1\lambda_2 b^2(x) + \lambda_2^2\right]\alpha_1 f(x)\int_R K^2(u)du\right).$$

□

Therefore by Wold's device, it results in asymptotic normality of $\hat{b}^2(x)$ given as below

Theorem 4.3.28. *Under **A1**-**A6** and **A8**, we have*

$$\sqrt{nh_n}\left[\frac{\hat{b}_n^2(x)}{b^2(x)} - 1\right] \xrightarrow{d} \mathcal{N}\left(0, 4f^{-1}(x)\int_R K^2(u)du\right).$$

4.4 Concluding Remark

In this chapter, we propose the online kernel estimators of the drift and diffusion for the time-homogeneous stationary diffusion model. Quadratic convergence, strong consistency and asymptotic normality are also established. These results can be used for further inference, e.g. constructing confidence intervals through asymptotic normality.

In fact, by using the method from (Stanton, 1997), we can propose the online estimators with other orders. For example, second-order estimators of $a(x)$ and $b^2(x)$ in (Stanton, 1997) are given by

$$\begin{aligned}
\hat{a}(x) &= \frac{1}{2\triangle}\{4E[X_{t+\triangle} - X_t|X_t = x] - E[X_{t+2\triangle} - X_t|X_t = x]\} + O(\triangle^2) \\
\hat{b}^2(x) &= \frac{1}{2\triangle}\left\{4E[(X_{t+\triangle} - X_t)^2|X_t = x] - E[(X_{t+2\triangle} - X_t)^2|X_t = x]\right\} + O(\triangle^2)
\end{aligned}$$

So let

$$\begin{aligned}
Y_n &= \frac{4(X_{n-1} - X_{n-2}) - (X_n - X_{n-2})}{2\triangle} \\
Z_n &= \frac{4(X_{n-1} - X_{n-2})^2 - (X_n - X_{n-2})^2}{2\triangle}
\end{aligned}$$

We can propose the estimators as the form in (4.4) and (4.6), then higher-order estimators are obtained. However in this case there is a tradeoff between the temporal and spatial complexity. In other words, to achieve faster convergence, two-step observations X_{n-2}, X_{n-1} and X_n must be used.

Additionally, we use the offline estimator with a small number of observations to provide the initial value for the online method. Note that the offline method can provide a reliable estimate because all historical data are used. For example, given the current observations $X_0, X_1, \ldots, X_n$, suppose now we are concerned with the drift and diffusion at x, then we will use the offline method to derive $\hat{a}_n(x)$ and $\hat{b}_n^2(x)$ and put them into (4.4) and (4.6) to update the estimates when there are new data. Also note that we do not need to consider the starting value for the offline nonparametric method. Recall that the offline method uses all the past

observations and is not computed iteratively, so knowing yesterday's offline estimate does not speed the process of determining today's offline estimate. The effect of the initial value on the performance of the online estimator will be discussed in the next chapter.

Chapter 5

Simulation and Case Study of Online Estimators

In this chapter, numerical simulation and a case study are used to evaluate the performance of the online estimators developed in Chapter 4 for the time-homogeneous stationary diffusion models. For numerical examples, the comparison with the offline method and sensitivity to parameters are investigated. For the case study, we give the estimates of the drift and diffusion for US 3-month treasury bill yields and then apply these estimates to calculate Value-at-Risk and Expected Shortfall for market risk management.

5.1 Simulation

Numerical simulation can help us evaluate effectiveness and efficiency of the procedure. We choose the following Vasicek model by Aït-Sahalia (1999) and the CIR model by Nicolau (2003) using monthly Federal funds data from January 1963 to December 1998:

$$dX_t = 0.261(0.0717 - X_t)dt + 0.02237dW_t \tag{5.1}$$

$$dX_t = 0.219(0.0721 - X_t)dt + 0.06665\sqrt{X_t}dW_t \tag{5.2}$$

For simplicity, the Euler scheme as described in Appendix A is used to sample the paths of (5.1) and (5.2). Note that the Euler scheme could simulate a negative value for (5.2), which results in the failure of taking the square root of the value. In this case, we re-simulate the value until the positive one is obtained. For example, Figure 5.1 demonstrates a sample path of (5.1) and (5.2) where the parameters $X_0 = 0.07, \triangle = 1/260$ and $T = 20$. That is, 260 trading days are assumed per year and the period of 20 years is considered in total[1]. In this case, it means there are $n = T/\triangle = 5200$ observations in the sample path. From Figure 5.1, we find that both

[1]The actual / 260 day count convention is used in this thesis. There are 52 weekends in one year, namely $52 \times 2 = 104$ days plus January 1, so 105 days are excluded in total.

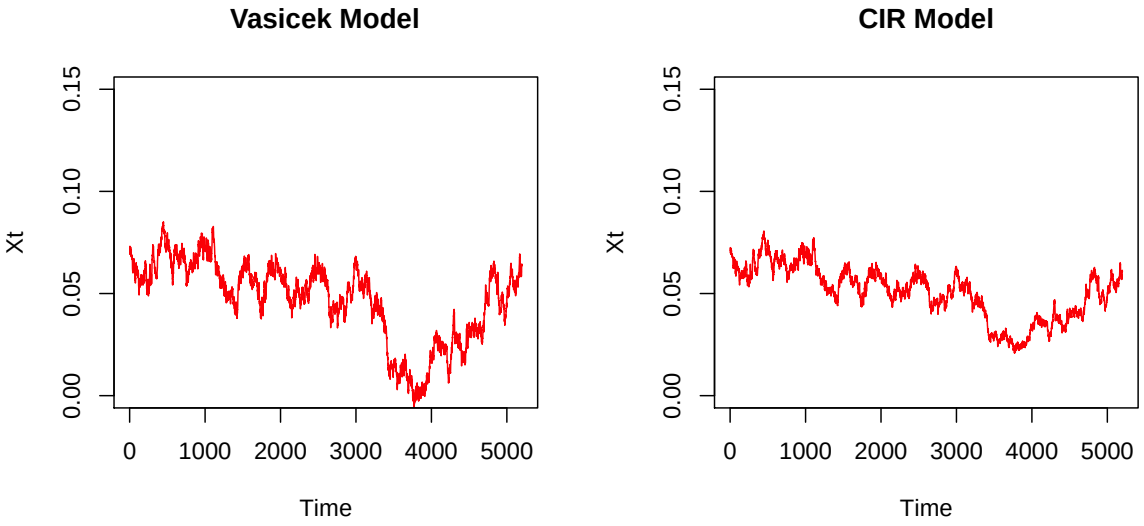

Figure 5.1: The sample path of (5.1) and (5.2) with the same random seed where $\triangle = 1/260$ and $T = 20$.

models have similar sampling paths, but the difference is that the Vasicek model allows for the possibility of negative X_t.

We use the mean integrated squared error (MISE) as a measure of the quality of the estimators. Given sequential observations $x_1, x_2, \ldots, x_i$, suppose that the points for consideration $a = z_1 < z_2 < \ldots < z_N = b$ are chosen uniformly and evenly spaced, and there are M replications of sample paths where the estimator $\hat{\theta}_i(x)$, i.e. $\hat{a}_i(x)$ or $\hat{b}_i^2(x)$ in this thesis, for each replication is denoted by $\hat{\theta}_{ij}(x)$, then the MISE is defined as

$$MISE_i = MISE(\hat{\theta}_i) = \frac{\triangle}{MN} \sum_{j=1}^{M} \sum_{k=1}^{N} [\hat{\theta}_{ij}(z_k) - \theta(z_k)]^2 \tag{5.3}$$

Additionally, given $x_1, x_2, \ldots, x_n$ of each replication, in order to initialize online estimators, we use offline estimators for $x_1, x_2, \ldots, x_m$ where $m < n$ to trigger the online procedure.

5.1.1 Comparison with offline estimators

We compare the offline and online methods for effectiveness and efficiency, where the goodness of fit is evaluated for effectiveness and the running time is considered for efficiency.

Suppose x_i is observed at this moment, let $\hat{a}_{i,\text{off}}(x), \hat{b}^2_{i,\text{off}}(x)$ and $\hat{a}_{i,\text{on}}(x), \hat{b}^2_{i,\text{on}}(x)$ denote the corresponding offline and online estimators of $a(x)$ and $b^2(x)$ respectively, where in this thesis the Nadaraya-Watson estimator is taken as the offline method for comparison. This means that $\hat{a}_{i,\text{off}}(x)$ and $\hat{b}^2_{i,\text{off}}(x)$ are obtained by using $x_1, x_2, \ldots, x_i$, while $\hat{a}_{i-1,\text{on}}(x)$ and $\hat{b}^2_{i-1,\text{on}}(x)$ are updated to $\hat{a}_{i,\text{on}}(x)$ and $\hat{b}^2_{i,\text{on}}(x)$ by using x_i. Additionally, in this thesis, the standard Gaussian

kernel is selected and the empirical bandwidth is chosen such that the fitting curve is smooth, i.e. $h_i = \hat{\sigma}_i \times i^{-0.02}$ for the drift and $h_i = \hat{\sigma}_i \times i^{-0.01}$ for the diffusion where $\hat{\sigma}_i$ is the sample standard deviation of $x_1, x_2, \ldots, x_i$. The common parameters can be seen in Table 5.1 and 1000 replications are generated for comparison.

X_0	T	$\triangle$	m
0.07	20	1 / 260	1040

Table 5.1: Common parameters in simulation.

Goodness of Fit

First of all, we demonstrate how MISE changes as observations are available sequentially. Note that as mentioned above, the offline method is used to initialize the online version for the first m observations so that they have the same values of MISE, thus we only show the x-axis starting from $m + 1$, namely 1041 in this case.

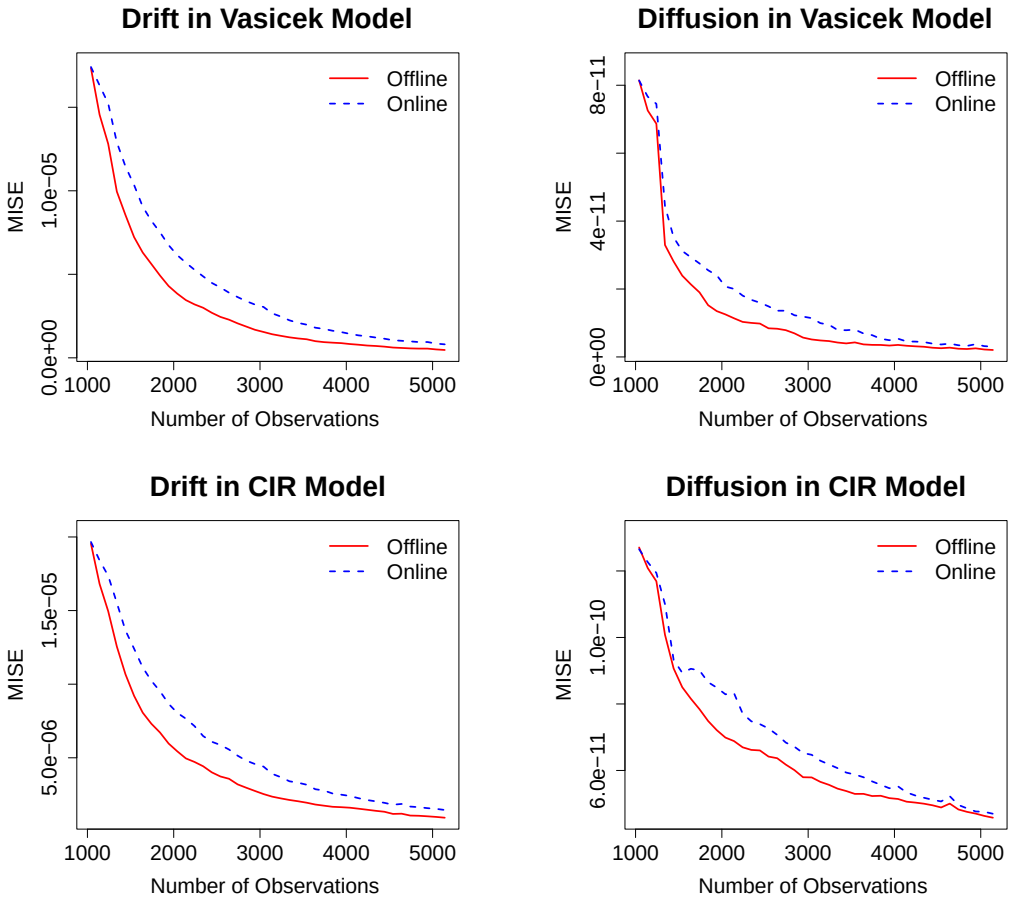

Figure 5.2: Demonstration of MISE for sequential observations, parameters as in Table 5.1.

From Figure 5.2, we can find that as more observations are available, both online and offine

methods can better estimate the drift and diffusion in the Vasicek model and CIR model based on MISE. Comparatively, the online method has slightly worse MISE than the offline counterpart. This is because on average, online estimators use a bigger bandwidth which depends on the number of samples. This indicates that online estimators have a larger bias. Thus the offline method is better than the online method. However note that when the sample size is large enough, the difference of their performances tends to diminish. In addition, it seems that the drift and diffusion in the Vasicek model are easier to be estimated than in the CIR model by both offline and online estimators. This might be due to the simpler form of the Vasicek model.

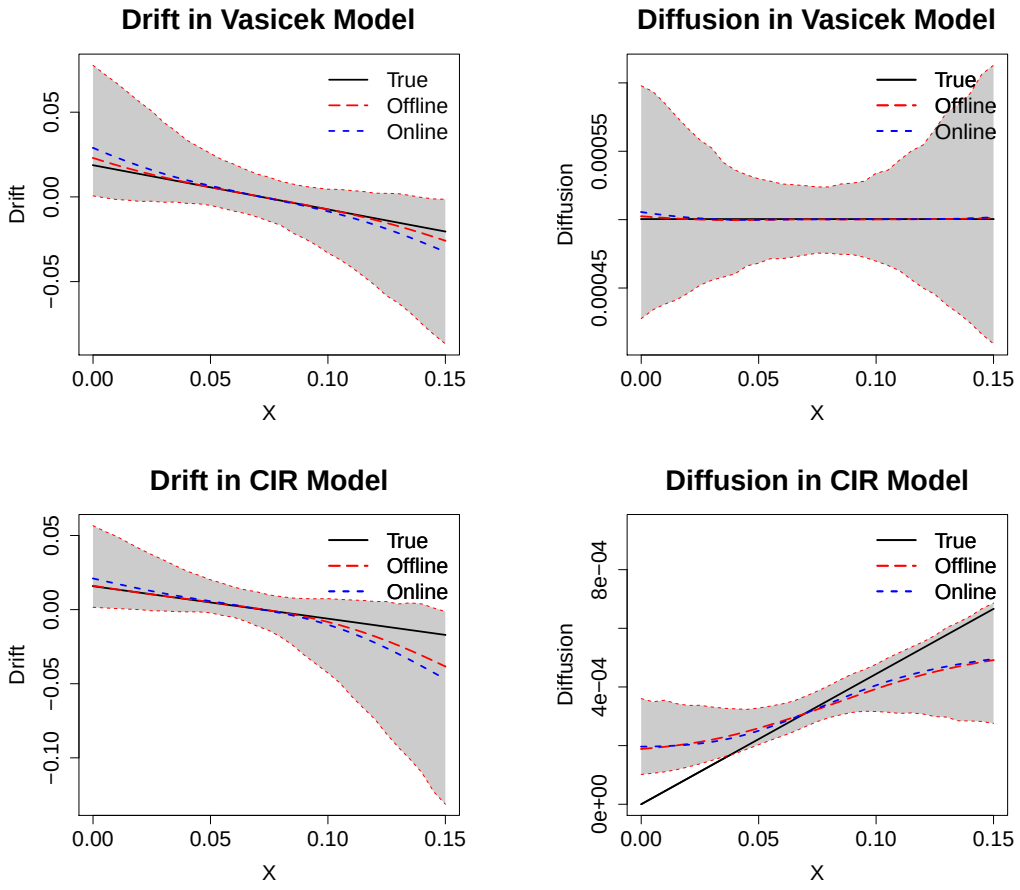

Figure 5.3: Fitting values by offline and online estimators which are averaged on 1000 replications, and the 95% confidence band of the online estimators, parameters as in Table 5.1.

We also check fitting values by offline and online estimators (Figure 5.3). We can find that both offline and online methods can roughly describe the general trend of the true drift and diffusion. Especially we note that we have better fit for the diffusion than for the drift. But the 95% confidence band of the online estimators shows that there is estimation bias on the boundary, namely boundary effect of the fitting lines. One reason is owing to lack of data at the boundary, as a result extrapolation to the boundary is not reliable. Another reason is that locally constant smoothing is used in both the offline and online methods, and as we have mentioned

in Chapter 2, locally constant smoothing cannot correct the boundary effect. In addition, it is interesting to note that the estimators of the drift and diffusion present surprising nonlinearity even though true drift and diffusion are linear functions.

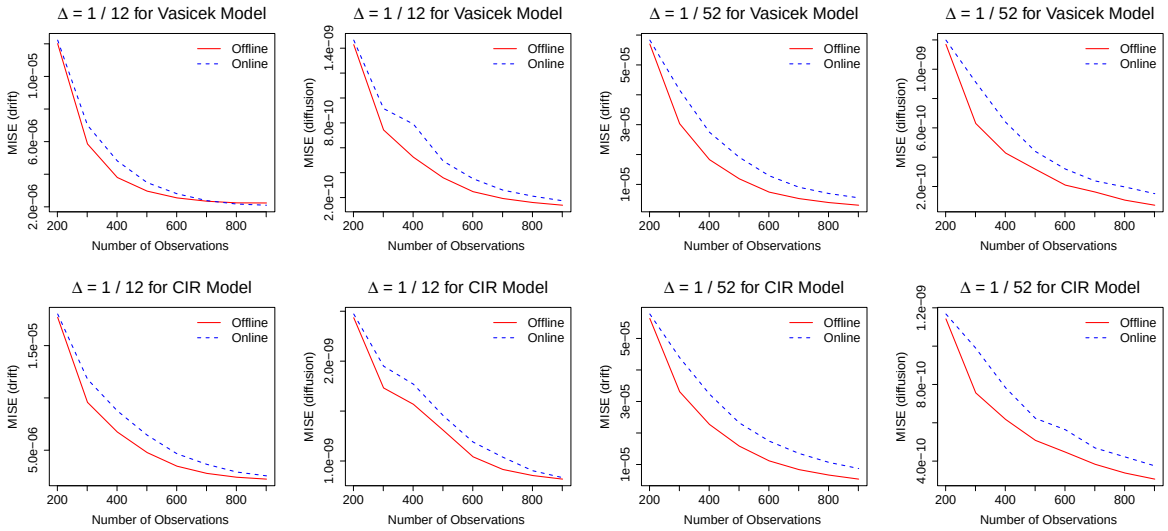

Figure 5.4: MISE for both offline and online estimators when $\triangle = 1/12$ and $1/52$ given $n = 1000$ and $m = 0.2n$.

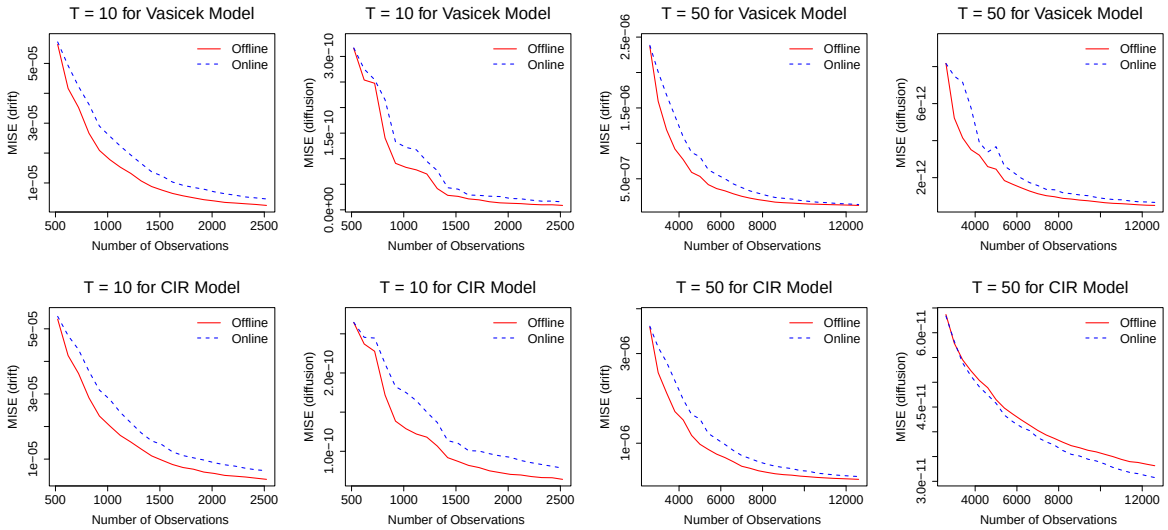

Figure 5.5: MISE for both offline and online estimators when $T = 10$ and 50 given $\triangle = 1/260$ and $m = 0.2n$.

In addition, we compare the MISE of offline and online estimators for different values of $\triangle$ and T. From Figure 5.4 and Figure 5.5, it is sure that the offline method is slightly better but this advantage tends to diminish as more observations are available, i.e. $\triangle$ increases for fixed n or T increases as well. In addition, it is interesting to note that the online estimator of the diffusion has smaller MISE for the CIR model when $T = 50$ than when $T = 10$.

Running Time

As shown above, the online method has larger MISE than the offline counterpart. However this disadvantage in estimation can be offset by the running time. In this part, we run a series of simulations to compare the running time of the online and offline methods, where different values of $\triangle$, T and m are considered[2].

unit (seconds)	drift		diffusion	
$\triangle = 1/12$	offline / online	ratio	offline / online	ratio
Vasicek model	7.96 / 0.46	17.30	7.90 / 0.44	17.95
CIR model	8.18 / 0.46	17.78	8.18 / 0.44	18.59
$\triangle = 1/52$	offline / online	ratio	offline / online	ratio
Vasicek model	8.46 / 0.46	18.39	8.44 / 0.46	18.35
CIR model	8.74 / 0.42	20.81	8.74 / 0.48	18.21
$\triangle = 1/260$	offline / online	ratio	offline / online	ratio
Vasicek model	7.80 / 0.44	17.73	7.80 / 0.44	17.73
CIR model	8.22 / 0.48	17.13	8.24 / 0.46	17.91

Table 5.2: Running time for $\triangle = 1/12, 1/52$ and $1/260$ where $n = 1000$ and $m = 0.2n$. That is, the time period $T = n\triangle = 1000/12, 1000/52$ and $1000/260$ for each case.

From Table 5.2, we see that for fixed n, both methods are irrelevant to $\triangle$ and the online method runs around 17 times faster. This is because when the sample size is fixed, $\triangle$ mainly determines the error rate of the estimators. This is illustrated by the theoretical results in Chapter 4.

unit (seconds)	drift		diffusion	
$T = 10$	offline / online	ratio	offline / online	ratio
Vasicek model	47.70 / 1.14	41.84	48.76 / 1.16	42.03
CIR model	50.98 / 1.18	43.20	51.80 / 1.16	44.66
$T = 20$	offline / online	ratio	offline / online	ratio
Vasicek model	164.28 / 2.42	67.88	168.00 / 2.34	71.79
CIR model	181.08 / 2.40	75.45	184.96 / 2.40	77.07
$T = 30$	offline / online	ratio	offline / online	ratio
Vasicek model	346.32 / 4.54	76.28	443.20 / 4.54	97.62
CIR model	447.08 / 3.60	124.19	401.00 / 3.58	112.01

Table 5.3: Running time for $T = 10, 20$ and 30 given $\triangle = 1/260$ and $m = 0.2n$.

Furthermore, we can find in Table 5.3 that as T increases (i.e. n increases for fixed $\triangle$), the difference in running time is more obvious and matters in real application. It can be seen that when $T = 30$, the online estimator can be more than 100 times faster. One can imagine

[2]The current computational environment is: (1) OS: Windows 7; (2) CPU: Intel Core i5-3317U 1.7GHz; (3) RAM: 8GB; (4) Programming Language: R 3.1.1.

that for large financial institutions, one or two hours could be taken by the online estimator for thousands of portfolios, but at this moment the offline estimator could take up 5 days or more to generate desirable results. Running for several days is inappropriate for practical purposes, for example calculating daily VaR limits. Thus in real applications our online method is more suitable for the demand of real-time computation such as managing risks. This is also a tradeoff between effectiveness and efficiency.

unit (seconds)	drift		diffusion	
$m = 0.1n$	offline / online	ratio	offline / online	ratio
Vasicek model	213.78 / 3.38	63.25	227.92 / 3.36	67.83
CIR model	235.64 / 2.68	87.93	221.16 / 2.66	83.14
$m = 0.3n$	offline / online	ratio	offline / online	ratio
Vasicek model	195.72 / 2.52	77.67	205.36 / 2.62	78.38
CIR model	213.10 / 2.38	89.54	208.76 / 2.50	83.50
$m = 0.5n$	offline / online	ratio	offline / online	ratio
Vasicek odel	158.44 / 1.92	82.52	164.02 / 1.88	87.24
CIR model	175.58 / 1.76	99.76	173.44 / 1.84	94.26
$m = 0.7n$	offline / online	ratio	offline / online	ratio
Vasicek odel	101.42 / 1.20	84.52	104.38 / 1.26	82.84
CIR model	114.72 / 1.2	95.60	117.72 / 1.22	96.49

Table 5.4: Running time for $m = 0.1n, 0.3n, 0.5n$ and $0.7n$ where $\triangle = 1/260$ and $n = 5200$.

In addition, we compare the running time for different values of m (see in Table 5.4). It can be also found that the online estimator is consistently faster than the offline estimator.

5.1.2 Sensitivity to parameters

In the sequel we focus on the online estimator and consider the sensitivity to different settings of parameters. In the online estimator, parameters include the discretization step size $\triangle$, the span of period T and the initial steps for estimators m. The study of sensitivity is helpful to determine the pros and cons of the estimator.

Discretization step size $\triangle$

In the proof of quadratic convergence and asymptotic normality of the estimator, it is assumed that $\triangle \to 0$ and $n\triangle \to \infty$. For demonstration, we fix $n = 1000$ and let $\triangle$ take 1 / 12 (for monthly data), 1 / 52 (for weekly data) and 1 / 260 (for daily data). The pointwise 95% confidence band is reported in order to check the quality of estimation.

Figure 5.6 shows the sensitivity of the online estimator to different values of $\triangle$. From the figure, we can find that when the sample size n is fixed, the 95% confidence band becomes thinner as $\triangle$ increases, which means the performance of the estimator can be improved in this

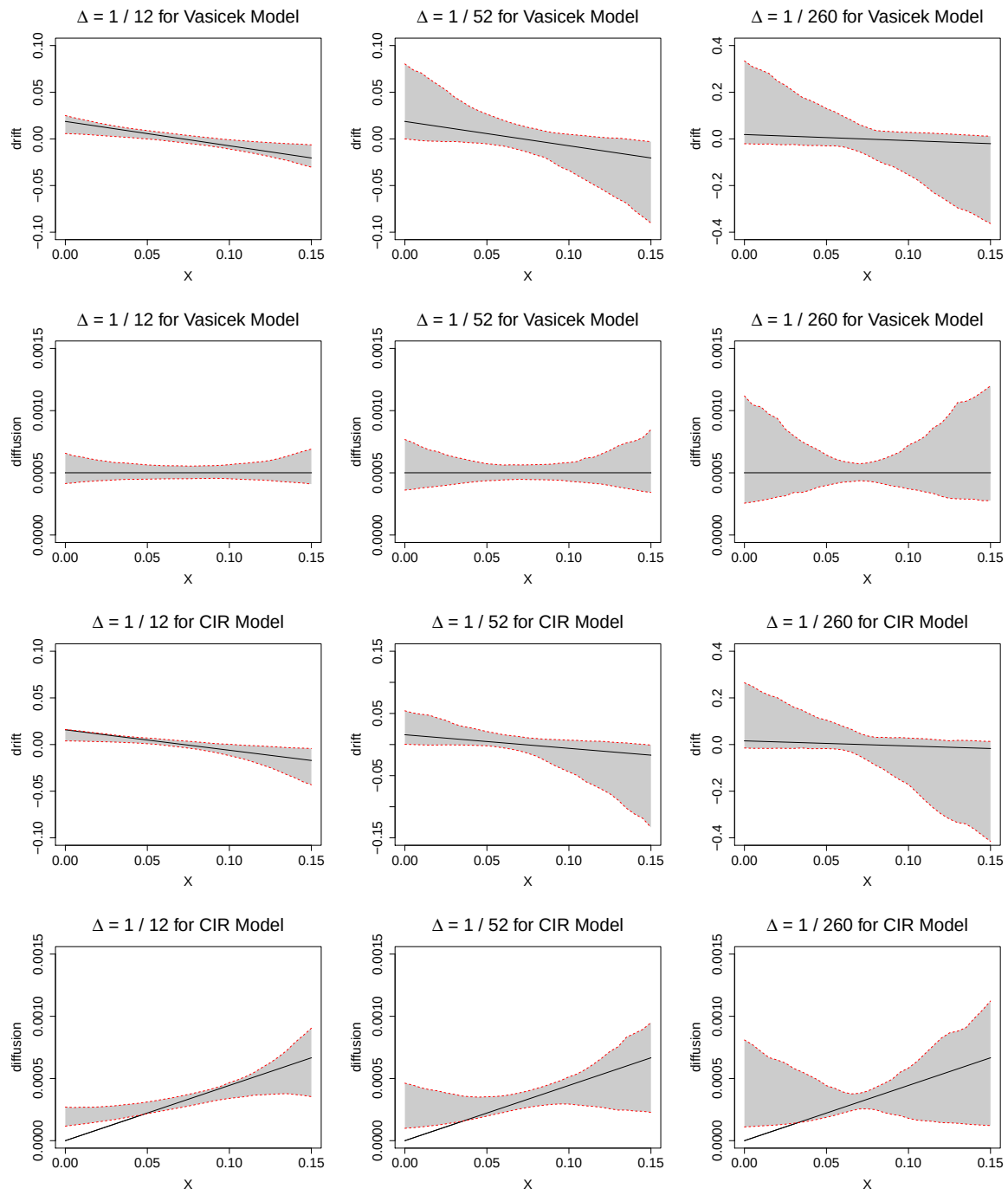

Figure 5.6: Sensitivity to $\triangle = 1/12, 1/52$ and $1/260$ where $n = 1000$ and $m = 0.2n$. The solid line represents the true value and the shadow area is the 95% confidence band.

case. Recall that in our proof, we assume that $\triangle \to 0$ and $n\triangle \to \infty$, that is, to consider the case of high-frequency data over a growing time window. The simulation is complementary to our theoretical analysis in that accurate estimates can be obtained even for low-frequency data. We also observe that the 95% confidence band becomes narrower in the middle part and gets wider toward two ends of the boundary. This is caused by two reasons, one of which is

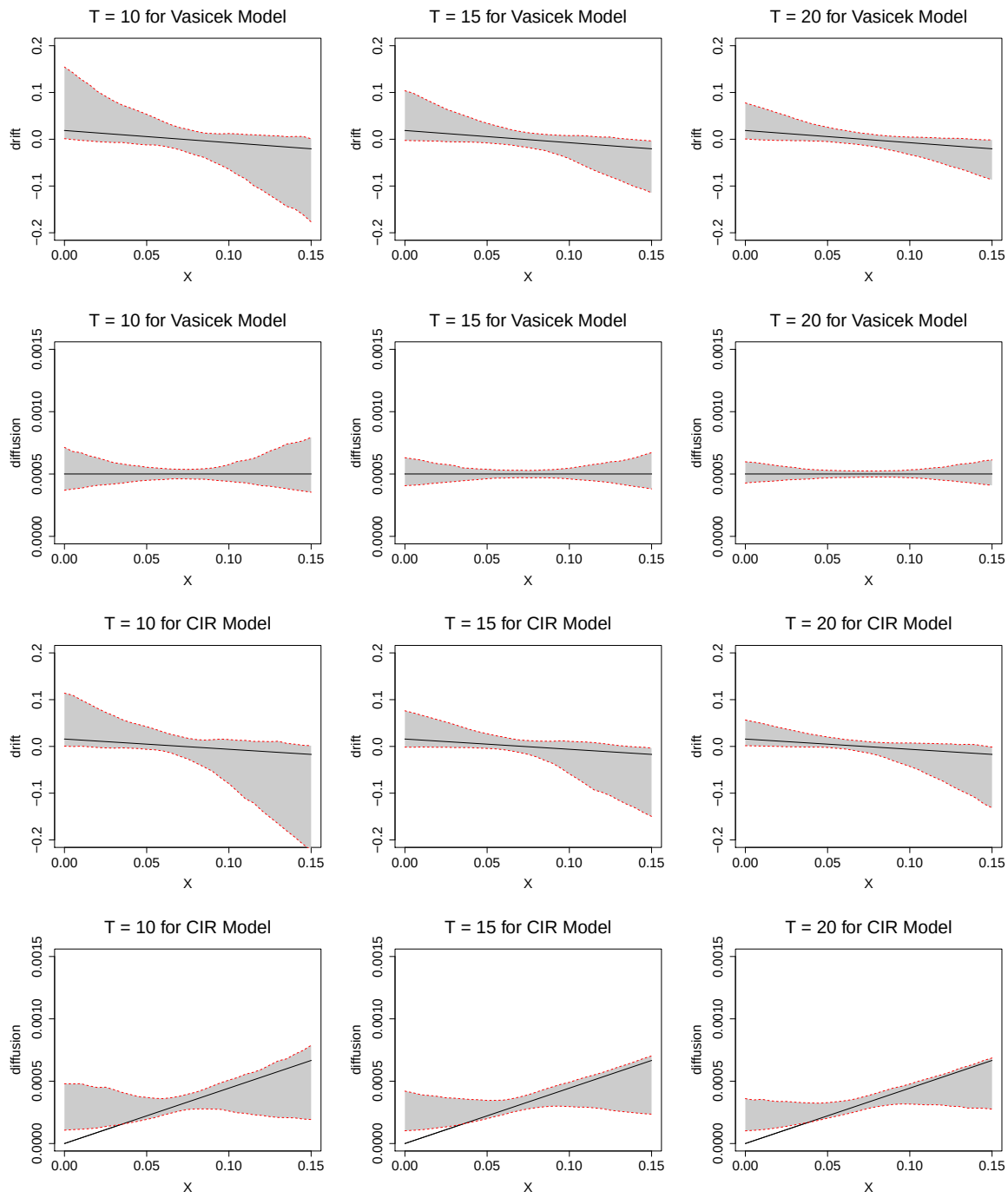

Figure 5.7: Sensitivity to $T = 10, 15$ and 20 where $\triangle = 1/260$ and $m = 0.2n$. The solid line represents the true value and the shadow area is the 95% confidence band.

the boundary effect of the kernel method and the other is the relative scarcity of observations at the boundary. In addition, it is interesting to note that in the CIR model, the confidence band of the estimator cannot cover the lower bound of true values. Thus it seems not reliable to use the online estimator to extrapolate the values to the lower bound for the CIR model.

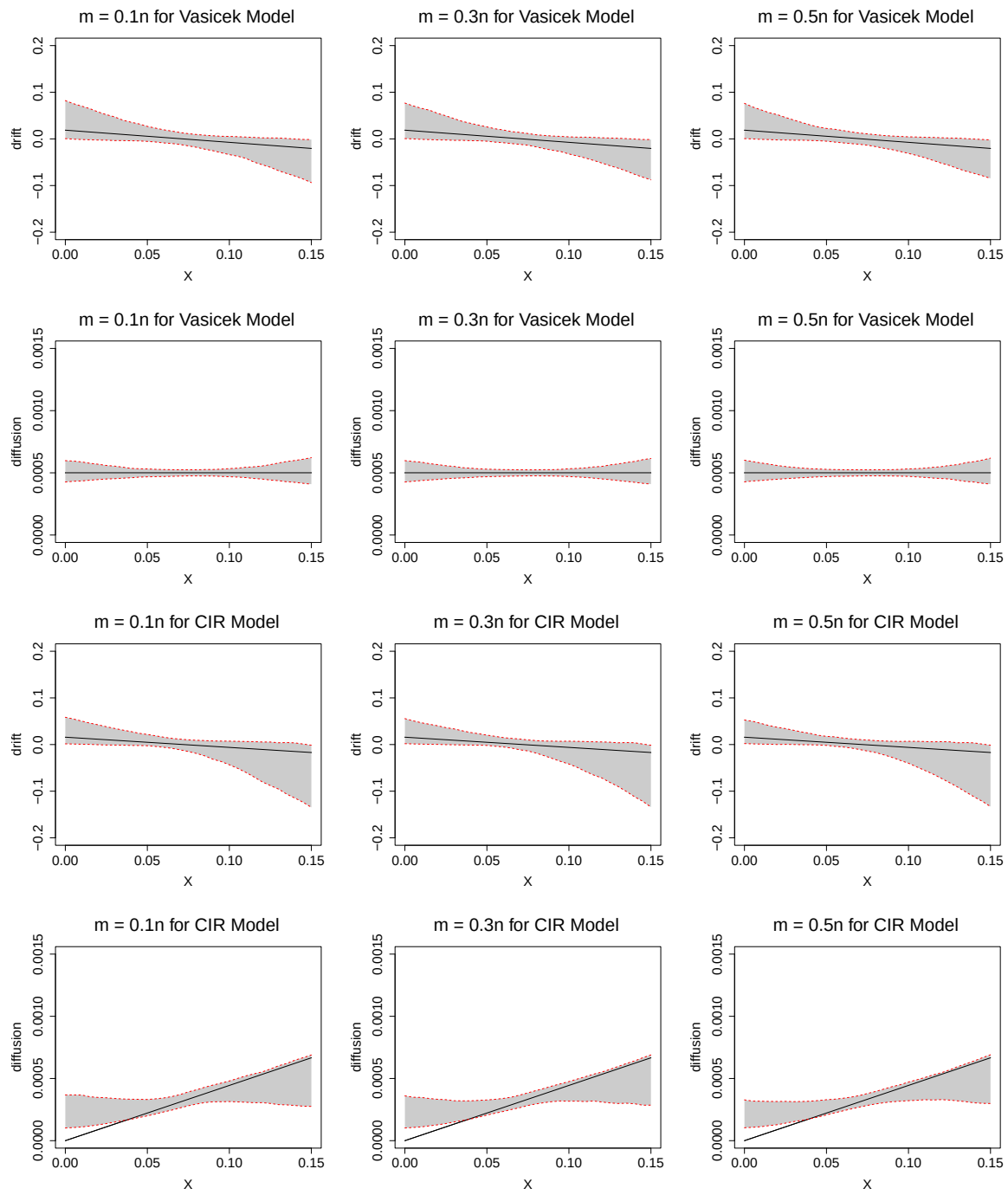

Figure 5.8: Sensitivity to $m = 0.1n, 0.3n$ and $0.5n$ given $\triangle = 1/260$ and $n = 5200$, where the solid line represents the true value and the shadow area is the 95% confidence band.

Span of periods T

According to the statements in Chapter 4, it is known that $n \to \infty$ and $T = n\triangle \to \infty$ are sufficient to prove the asymptotic properties of the estimator. In simulation, similarly when $\triangle$ is fixed to be 1 / 260, we study the performance of the estimator as T varies. In Figure 5.7, we can find that as T increases, the 95% confidence band becomes consistently narrower for

the estimators of the drift and diffusion of the Vasicek model and CIR model. This indicates that more accurate estimates can be obtained if there are more sequential observations, which verifies our theoretical analysis.

The number of initial steps m

We also consider how the number of initial steps m affects the performance of the estimator (see in Figure 5.8). It is of interest to find that when n is fixed, m is not related to the behaviors of the estimator. This implies that even though only 10 percent of observations are used to initialize the online estimator, we can still have good estimates of the drift and diffusion. So the online estimator provides flexibilities to use and is able to achieve desirable performances. In this way, when the data are not big, we can use the offline estimator first and then switch to the online estimator if the computational time becomes a big issue.

5.2 Case Study: US 3-Month Treasury Bill Rates

One application of stochastic differential equations is to model the short-term interest rate for managing risks of portfolios. In this section, we apply our online method in a case study by using the daily annualized US 3-month treasury bill rates from the Federal Reserve Bank of St. Louis. The data cover the period from May 8, 1978 to November 14, 2014. For the research purpose, when observations are missing, the corresponding time stamps are also removed from the dataset so that no extra errors are introduced. So the current data set contain 9121 observations. The interest rate time series and its shocks are plotted in Figure 5.9 and Figure 5.10.

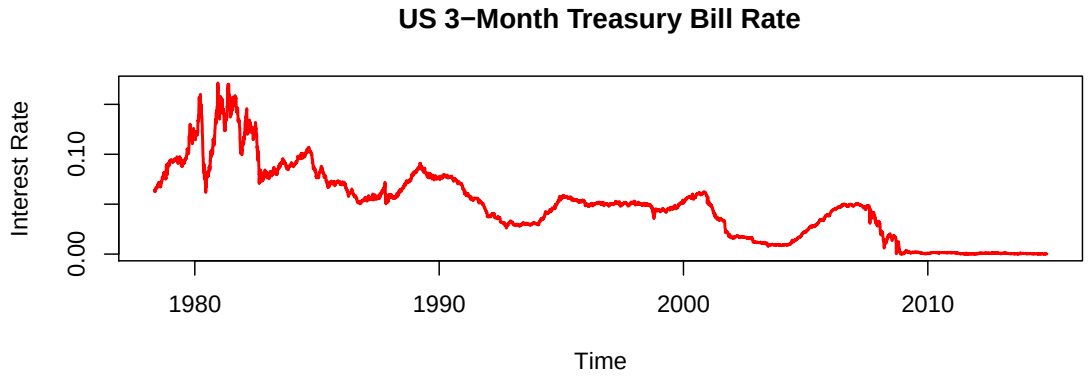

Figure 5.9: The daily US 3-month treasury bill rates from May 8, 1978 to November 14, 2014.

Before estimation, we provide summary statistics of the data. Table 5.5 presents the mean,

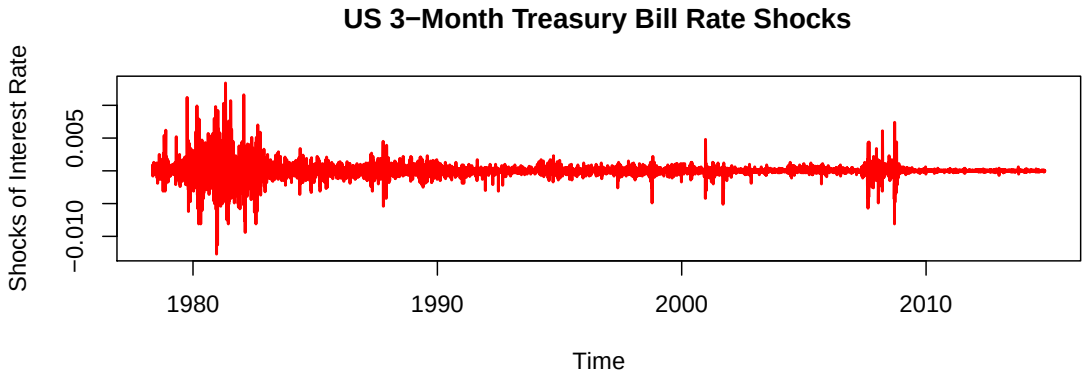

Figure 5.10: The absolute shocks of the daily US 3-month treasury bill rates.

Summary of Statistics					
Statistics	Mean	Median	SD	Kurtosis	Skewness
X_t	4.902%	4.960%	3.614%	3.224	0.625
Statistics	ρ_1	ρ_5	ρ_{10}	ρ_{15}	ρ_{20}
X_t	0.999	0.997	0.994	0.990	0.986

Table 5.5: Summary statistics of US 3-month treasury bill rates between May 8, 1978 and November 14, 2014.

median, standard deviation, kurtosis[3], skewness[4], and the first 20 lags of autocorrelation of the data. Table 5.5 shows the distribution of X_t is slightly leptokurtic and right-skewed, and the autocorrelation of X_t decays slowly. The statistical properties are also reported in Figure 5.11.

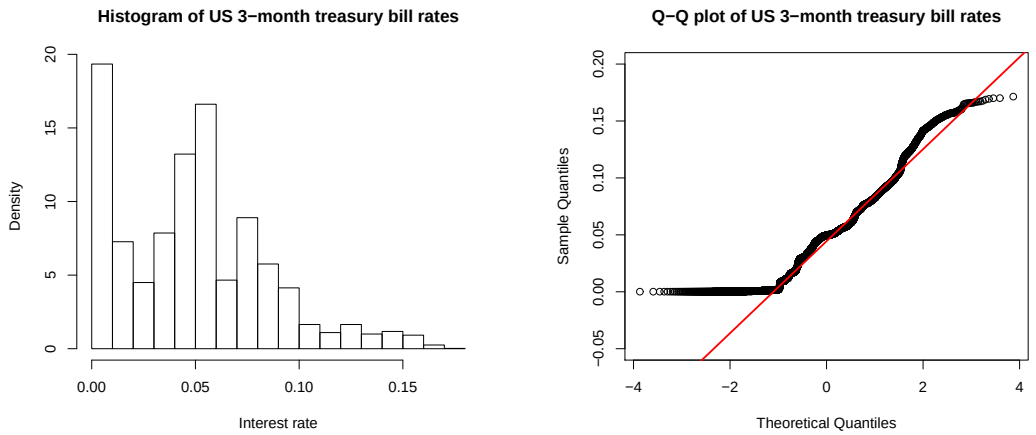

Figure 5.11: Histogram and QQ-plot of the US 3-month treasury bill rates between May 8, 1978 and November 14, 2014.

[3]Kurtosis is a measure of the weight of the tails of a distribution. The normal distribution has a kurtosis of 3, and distributions with higher values (heavier tails) are leptokurtic, while those with lower values are platykurtic.

[4]Skewness is a measure of the asymmetry of a distribution function. If the skewness is negative, the distribution is left-skewed, whereas if the skewness is positive, the distribution is right-skewed.

Furthermore, several hypotheses are tested to characterize different aspects of the data. Table 5.6 shows the results of the Augmented Dickey-Fuller stationarity test (Said and Dickey, 1984), Ljung-Box independence test (Ljung and Box, 1978) and Jarque-Bera normality test (Jarque and Bera, 1980). The null hypotheses of non-stationarity, independence and normality are rejected at the 5% significance level.

Augmented Dickey-Fuller stationarity test		
H_0	Test statistic	p value
Non-stationarity	-4.1653	0.01
Ljung-Box independence test		
H_0	Test statistic	p value
Independence	9114.838	$< 2.2 \times 10^{-16}$
Jarque-Bera normality test		
H_0	Test statistic	p value
Normality	612.2911	$< 2.2 \times 10^{-16}$

Table 5.6: Hypothesis tests of the data for the stationarity, independence and normality.

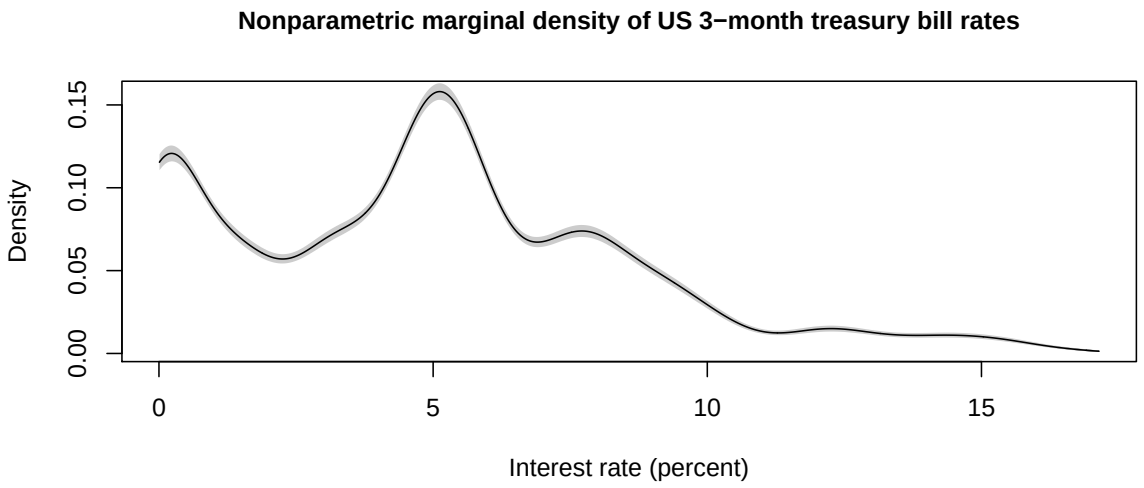

Figure 5.12: Nonparametric marginal density of the data, where the solid line represents true values and the shadow area is the 95% confidence band.

The marginal density of the observations by nonparametric density estimators as mentioned in Chapter 2, with the 95% confidence band by the sample quantiles can be seen in Figure 5.12. The Gaussian kernel is used and the empirical bandwidth $\hat{h}_n = \hat{\sigma}_n \times n^{-0.2}$ is chosen. From the figure, it is validated that the density appears non-normal, asymmetric and right-skewed. Also the noticeable features of the density are the upper fat tail and several modes. The largest modes are centered on 0.05% and 5%, where the mode of 0.05% happened after 2008 and the mode of 5% occurred between 1996 and 1998 and between 2006 and 2007. Most of the low

interest rates could be due to the resulting financial crisis observed in the few years beginning in 2007. The high interest rates are recorded between 1979 and 1982, which was caused by the monetary policy by the US Federal Reserve to curtail rising price inflation (Friedman, 1984).

5.2.1 Estimation Results

In this section, we apply the online method developed in the last chapter to estimate the drift and diffusion of US 3-month treasury bill rates by assuming the observations are available sequentially. As a reference, the parameters in the Vasicek model and the CIR model are calibrated first. Here we use linear regression to calibrate the parameters. For the Vasicek model, the following discrete autoregressive process approximates the true model $dX_t = a(b - X_t)dt + cdW_t$ as the discretization step size $\triangle \to 0$

$$X_i = X_{i-1}e^{-a\triangle} + b\left(1 - e^{-a\triangle}\right) + c\sqrt{\frac{1 - e^{-2a\triangle}}{2a}}\varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$. Note that the above equation can be written as the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i^{'} \tag{5.4}$$

where $Y_i = X_{i+1}$ and

$$\beta_0 = b\left(1 - e^{-a\triangle}\right) \qquad \beta_1 = e^{-a\triangle} \qquad \varepsilon_i^{'} \sim \mathcal{N}\left(0, c^2\frac{1 - e^{-2a\triangle}}{2a}\right)$$

Thus the parameters in the Vasicek model are given by

$$\hat{a} = -\frac{\ln\hat{\beta}_1}{\triangle} \qquad \hat{b} = \frac{\hat{\beta}_0}{1 - \hat{\beta}_1} \qquad \hat{c} = \hat{\sigma}_{\varepsilon_i^{'}}\sqrt{\frac{-2\ln\hat{\beta}_1}{\triangle(1 - \hat{\beta_1}^2)}}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from (5.4), and $\hat{\sigma}_{\varepsilon'}$ is the squared root of mean squared errors. For the CIR model, the Euler scheme gives the discretized form of the true model $dX_t = a(b - X_t)dt + c\sqrt{X_t}dW_t$ as $X_i - X_{i-1} = a(b - X_{i-1})\triangle + c\sqrt{X_{i-1}\triangle}\varepsilon_i$, that is,

$$\frac{X_i - X_{i-1}}{\sqrt{X_{i-1}}} = \frac{ab\triangle}{\sqrt{X_{i-1}}} - a\sqrt{X_{i-1}}\triangle + c\sqrt{\triangle}\varepsilon_i$$

The related linear regression of the above equation is expressed as

$$Y_i = \beta_0\frac{1}{\sqrt{X_i}} + \beta_1\sqrt{X_i} + \varepsilon_i^{'}$$

where $Y_i = \frac{X_{i+1}-X_i}{\sqrt{X_i}}$ and

$$\beta_0 = ab\triangle \qquad \beta_1 = -a\triangle \qquad \varepsilon_i^{'} \sim \mathcal{N}(0, c^2\triangle)$$

so the parameters in the CIR model are given by

$$\hat{a} = -\frac{\hat{\beta}_1}{\triangle} \qquad \hat{b} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} \qquad \hat{c} = \frac{\sigma_{\varepsilon_i^{'}}}{\sqrt{\triangle}}$$

For the case of US 3-month treasury bill rates, if the data are assumed to follow the Vasicek or CIR model, then by letting $\triangle = 1/260$, the calibrated parameters are reported in Table 5.7. From the table, it can be seen that the mean levels to which the Vasicek or CIR processes revert are 0.026 and 0.036 respectively, and the speeds of reversion are 0.078 and 0.141 respectively.

Estimate	Vasicek model	CIR model
$\hat{a}$	0.078	0.141
$\hat{b}$	0.026	0.036
$\hat{c}$	0.016	0.080

Table 5.7: Calibration of parameters in the Vasicek and CIR model.

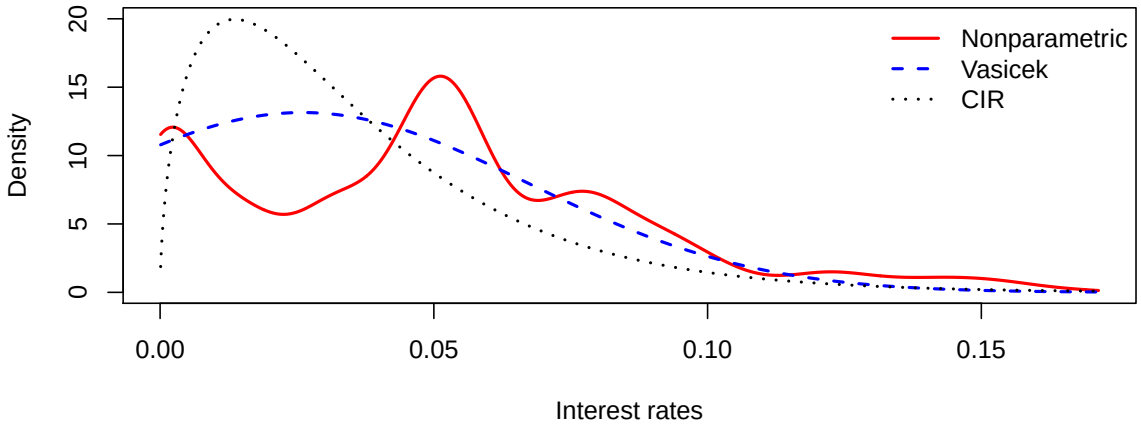

Figure 5.13: Comparison between nonparametric marginal density function and those of the Vasicek and CIR model.

Based on the parameters calibrated above, we also plot the density of the Vasicek and CIR model[5] and compare them with the nonparametric marginal density given by Figure 5.12. The

[5]It is known that the Vasicek model has the stationary density $\mathcal{N}(b, c^2/2a)$, but the interest rates are positive, thus the truncated normal distribution is used here. For the CIR model, X_t converges in distribution to $c^2/4a$ times the stationary density χ^2_d where $d = 4ab/c^2$ (Glasserman, 2003). In other words, let $Y \sim \chi^2_d$ and its density function $f_Y(y)$, then based on univariate Jacobian transformation, the density of X_t should be $f_{X_t}(x) = 4af_Y(4ax/c^2)/c^2$.

result is reported in Figure 5.13. It is obvious that all density functions are centered on lower observations and have longer tails on higher observations.

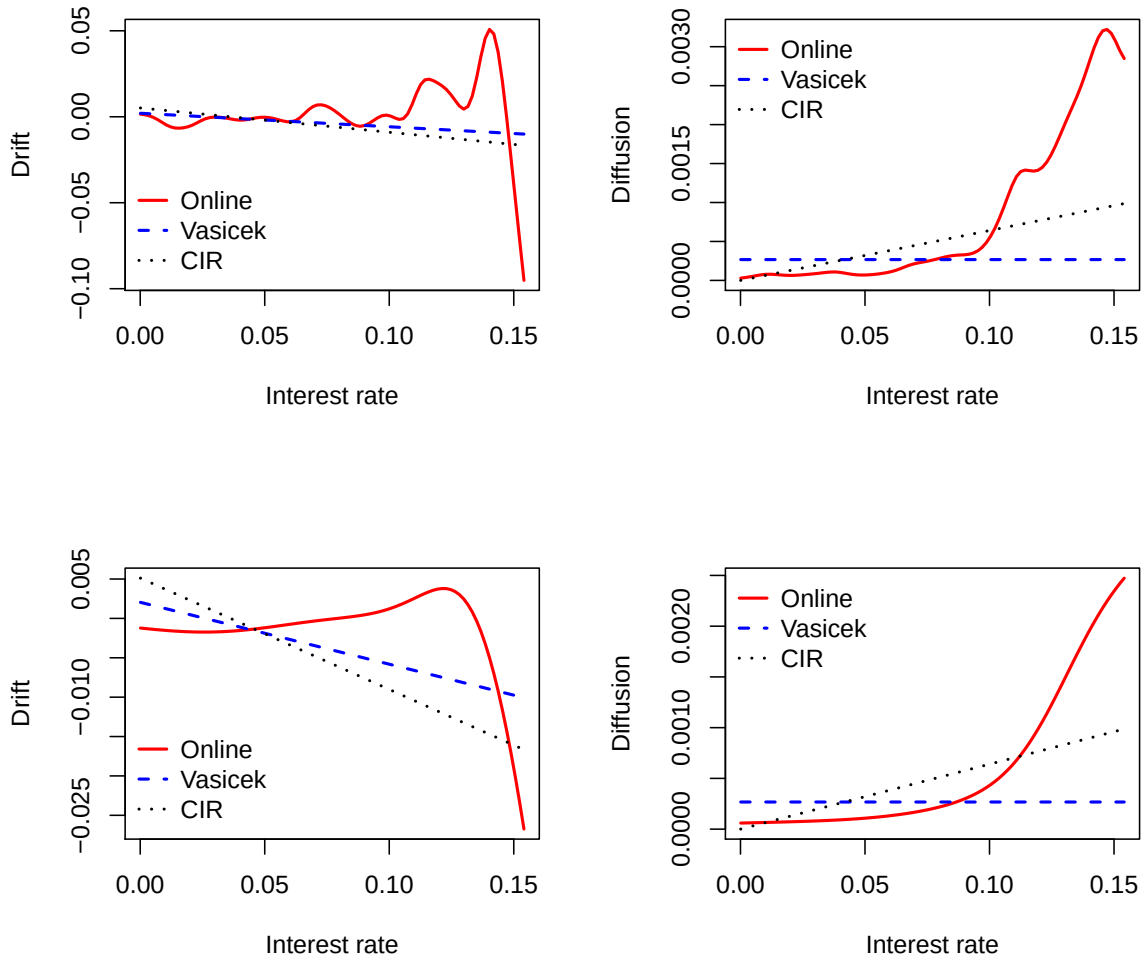

Figure 5.14: Estimation of the drift and diffusion by the calibrated Vasicek and CIR model, and the online method with different bandwidths $h_i = \hat{\sigma}_i \times i^{-0.2}$ (the top) and $h_i = \hat{\sigma}_i \times i^{-0.02}$ (the bottom).

Figure 5.14 shows the estimated drift and diffusion functions by the calibrated Vasicek and CIR model as well as the online method with different bandwidths. The online method uses 20% of the observations for initialization and the discretization step size $\triangle = 1/260$. There are three noticeable features. First, the three methods have similar drift functions for the low interest rates, but for the high observations, the drift by the online method is more negative. One reason is due to the boundary bias of the online method, and the other reason could be misspecification of the Vasicek and CIR model. But it is hard to identify which reason is dominant. Second, the online method presents the nonlinear drift and diffusion function, which is different from the Vasicek and CIR model. Although it is observed that the drift is

more likely to be nonlinear, which is also claimed by some authors (Jiang and Knight, 1997; Aït-Sahalia, 1996), we must admit that whether the drift has such form is still unknown. A similar statement could be applied to the diffusion. But both the online method and CIR model show that the diffusion is higher as the interest rate increases. This follows our intuition that the added volatility is associated with high interest rates for compensating the risk. Third, given different bandwidths, the online estimator presents a similar trend for the drift and diffusion. The difference is that when the bandwidth is bigger ($h_i = \hat{\sigma}_i \times i^{-0.02}$), the fitted curve looks smoother, i.e. smaller variance. Thus the drift and diffusion revealed by the online method seem reasonable. Therefore, if the parametric form should be specified for US 3-month treasury bill yields, the CIR model could be alternative with more possibilities than the Vasicek model. In addition to the advantage of the running speed, our online estimator is complementary to the parametric method.

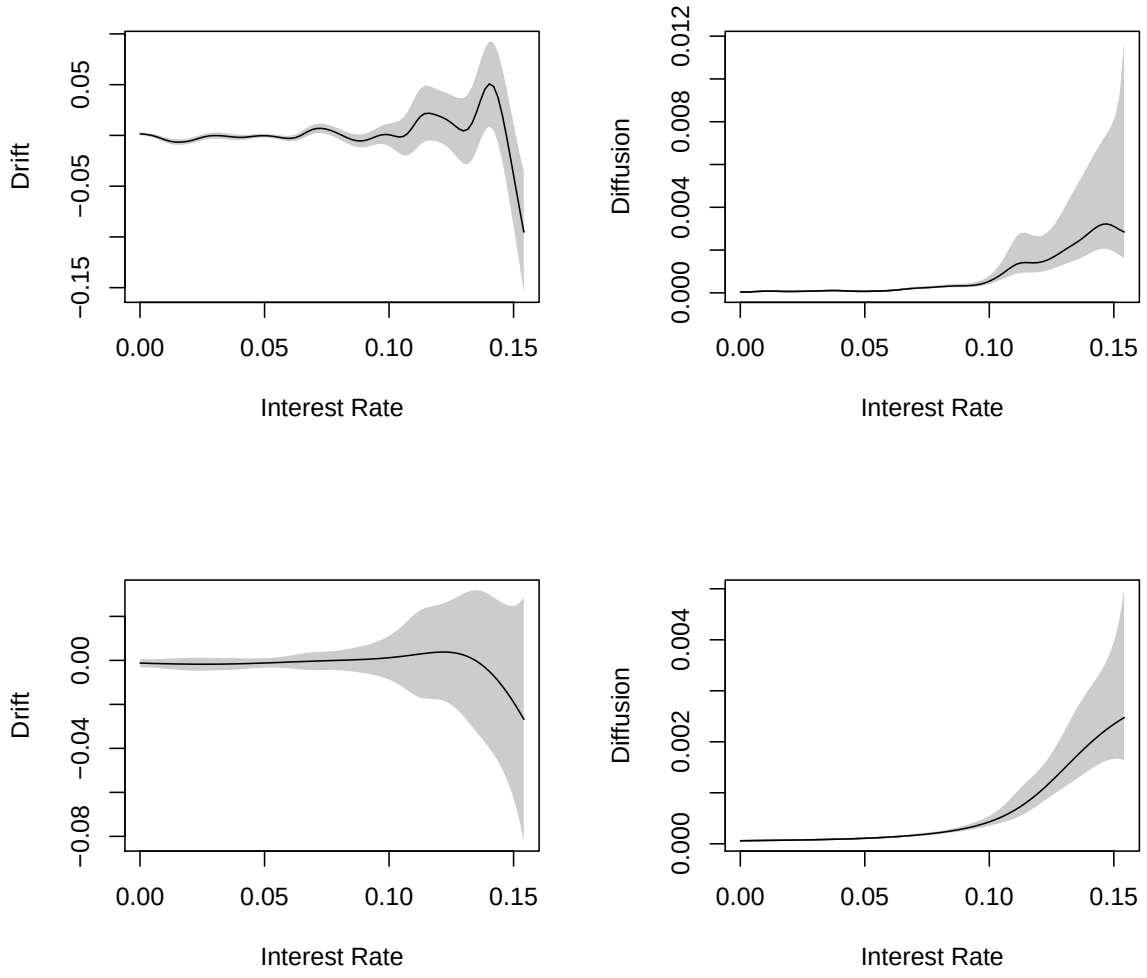

Figure 5.15: Estimation of the drift and diffusion as well as 90% pointwise confidence band by the online method for the bandwidths $h_i = \hat{\sigma}_i \times i^{-0.2}$ (the top) and $h_i = \hat{\sigma}_i \times i^{-0.02}$ (the bottom).

Figure 5.15 presents a pointwise 90% confidence band by the online method for different bandwidths, where the confidence band is constructed based on Theorem 4.3.26 and Theorem 4.3.28. We can find that the 90% confidence band tends to widen as the interest rate increases for the lack of enough observations around high interest rates, which can be seen through the nonparametric marginal density in Figure 5.12 as well.

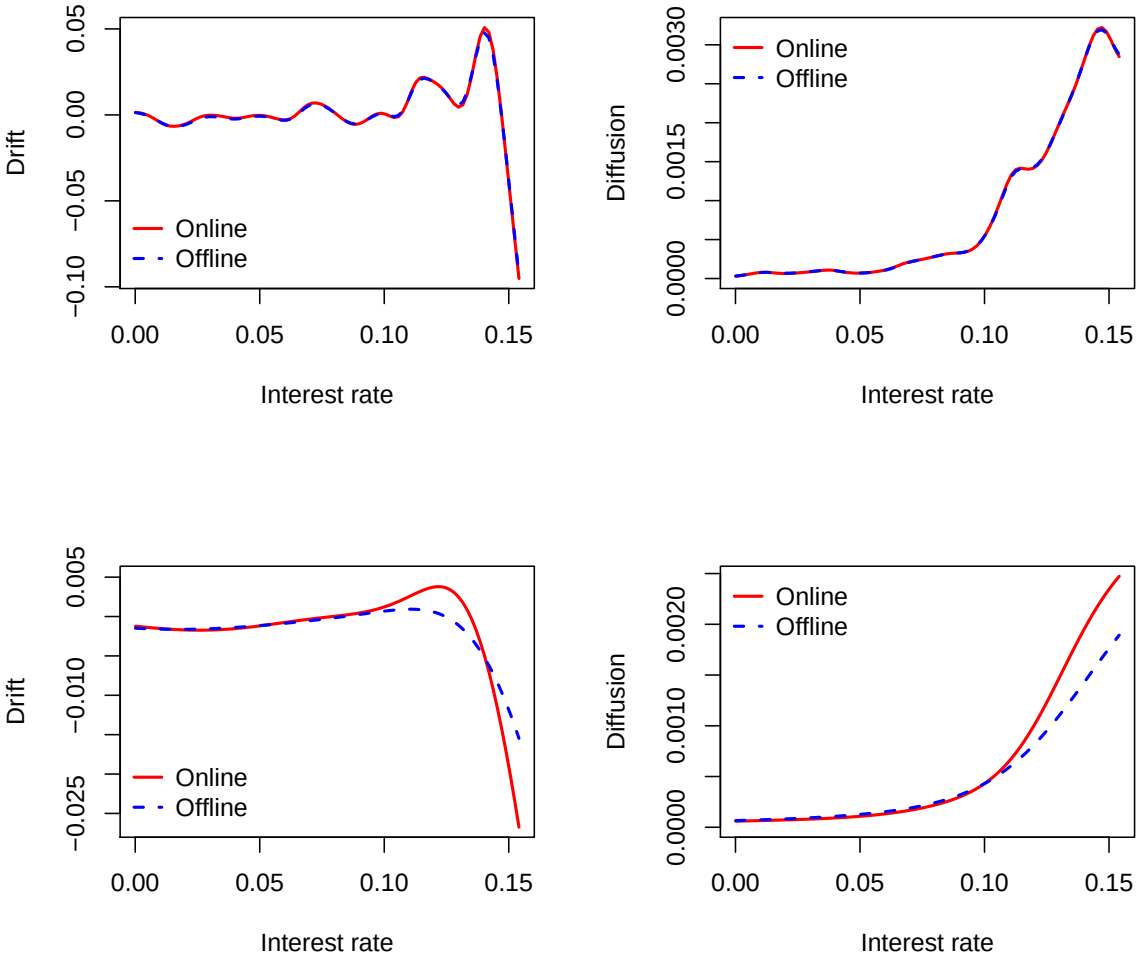

Figure 5.16: Comparison of the drift and diffusion specification by the offline and online methods for the bandwidths $h_i = \hat{\sigma}_i \times i^{-0.2}$ (the top) and $h_i = \hat{\sigma}_i \times i^{-0.02}$ (the bottom).

In addition, we compare the performance of the offline and online methods for the case study. The offline method is used for all historical data since its running time for sequential data is unbearable. Figure 5.16 shows estimation of the drift and diffusion by the offline and online methods with different bandwidths. It can seen that when the bandwidth takes a smaller value, there is no difference between the two methods. As the bandwidth increases, both methods can give more smooth estimates, but the drift and diffusion by the offline method have smaller

absolute values for the high interest rates.

5.2.2 Application in Risk Management

As is shown in Figure 5.9, the spot short-term interest rate is changing over time, as a result financial institutions have to manage their risks of a financial product depending on the interest rate. In this part, we apply our online method to market risk management of US 3-month treasury bill rates and also compare the performance of the online method with historical simulation and the parametric Monte Carlo method (i.e. the Vasicek model and the CIR model). From the perspective of calculating VaR and ES, our online technique can be considered as a semiparametric method, compared to historical simulation and parametric Monte Carlo method. This is because a diffusion model is assumed to describe the process of the risk factor, but calibration of the drift and diffusion is nonparametric. In the application, we focus on calculation of daily 99% VaR and daily 97.5% ES. For historical simulation and parametric Monte Carlo method, one-year time horizon is considered. For parametric Monte Carlo method and the online method, 100,000 PnL scenarios are simulated for each time step. In addition, the Gaussian kernel function is chosen and the bandwidth is $h_i = \hat{\sigma}_i \times i^{-0.02}$ for the online method.

Calculation of VaR and ES

First of all, Table 5.8 gives the results of calculating VaR and ES by historical simulation, the Monte Carlo method by the Vasicek and CIR model, and the online method. The values in the table are reported over the whole period. It can be found that the online method has the fewest breaches of VaR and ES, which is followed by historical simulation. The Monte Carlo method by the CIR model has the most breaches. This can be shown by the basic statistics in the table. Note that the minimum and maximum of VaR and ES by the online method are larger but having smaller standard deviation, so more conservative prediction of shocks is given by the online method. We can also find that the number of 97.5% ES breaches is smaller than that of 99% VaR breaches by all methods.

In order to evaluate these methods in detail, Figure 5.17 and Figure 5.18 report the time series of VaR and ES. From the figures, we observe that the online method has two red zones. One of them happens during the period of monetary policy by the US Federal Reserve, and the other is during the recent financial crisis. Both periods have severe fluctuation of shocks. Comparatively, historical simulation is more responsive to the recent changes and has fewer breaches in both stressed periods. The difference of the performance in stressed periods is that historical simulation has a one year time horizon, however essentially the online method is based on all previous observations although the most recent observation is used to update estimates. So the online method is less responsive to the recent changes.

	VaR					
	Breaches	Percentage	Min	Max	Mean	S.D.
HS*	110	1.24%	0.0002	0.0081	0.0021	0.0020
VAS**	168	1.90%	0.0001	0.0078	0.0017	0.0017
CIR***	203	2.29%	-0.0008	0.0082	0.0016	0.0016
Online	93	1.05%	0.0010	0.0094	0.0022	0.0015
	ES					
	Breaches	Percentage	Min	Max	Mean	S.D.
HS*	103	1.16%	0.0002	0.0084	0.0021	0.0020
VAS**	166	1.87%	0.00015	0.0078	0.0017	0.0017
CIR***	199	2.25%	-0.0008	0.0082	0.0016	0.0016
Online	92	1.03%	0.0010	0.0094	0.0023	0.0015

Table 5.8: Comparison of VaR and ES

* HS stands for historical simulation
** VAS stands for Monte Carlo method by the Vasicek model
*** CIR stands for Monte Carlo method by the CIR model

Confidence Interval of VaR and ES

It is noted from the definition of VaR and ES (2.1) and (2.2) that both VaR and ES are random variables, thus given scenarios, we can quantify the performance of the method by considering the confidence interval of VaR and ES. However according to the original definition (2.1) and (2.2), we need the distribution function of the risk factor to construct the confidence interval of VaR and ES but it is usually unknown in real applications. On the other hand, not much research work considers the construction of their confidence intervals. So in this part, the confidence interval of VaR and ES is derived first which is based on the practical definition (2.3) and (2.4), then we evaluate the performance of the method based on the derived confidence interval.

Given a distribution function $F(x)$ for the i.i.d. shocks, according to the asymptotic results for order statistics by Mosteller (1946), there is the asymptotic normality result

$$\sqrt{n}\left(X_{(\lceil np \rceil)} - F^{-1}(p)\right) \xrightarrow{d} N\left(0, \frac{p(1-p)}{[f(F^{-1}(p))]^2}\right)$$

In other words, we can construct the following asymptotic confidence interval with $1-\tau$ level for VaR_α based on the definition (2.3)

$$-\hat{F}^{-1}(1-\alpha) \pm Z_{1-\frac{\tau}{2}} \sqrt{\frac{\alpha(1-\alpha)}{n[\hat{f}(\hat{F}^{-1}(1-\alpha))]^2}} \tag{5.5}$$

where $\hat{F}(x)$ and $\hat{f}(x)$ are the empirical distribution function and density function respectively because the true functions $F(x)$ and $f(x)$ are usually unknown. Furthermore given the realization of ordered shocks $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, we approximate $\hat{F}^{-1}(1-\alpha)$ by $x_{\lceil n(1-\alpha) \rceil}$. For $\hat{f}(x)$, the

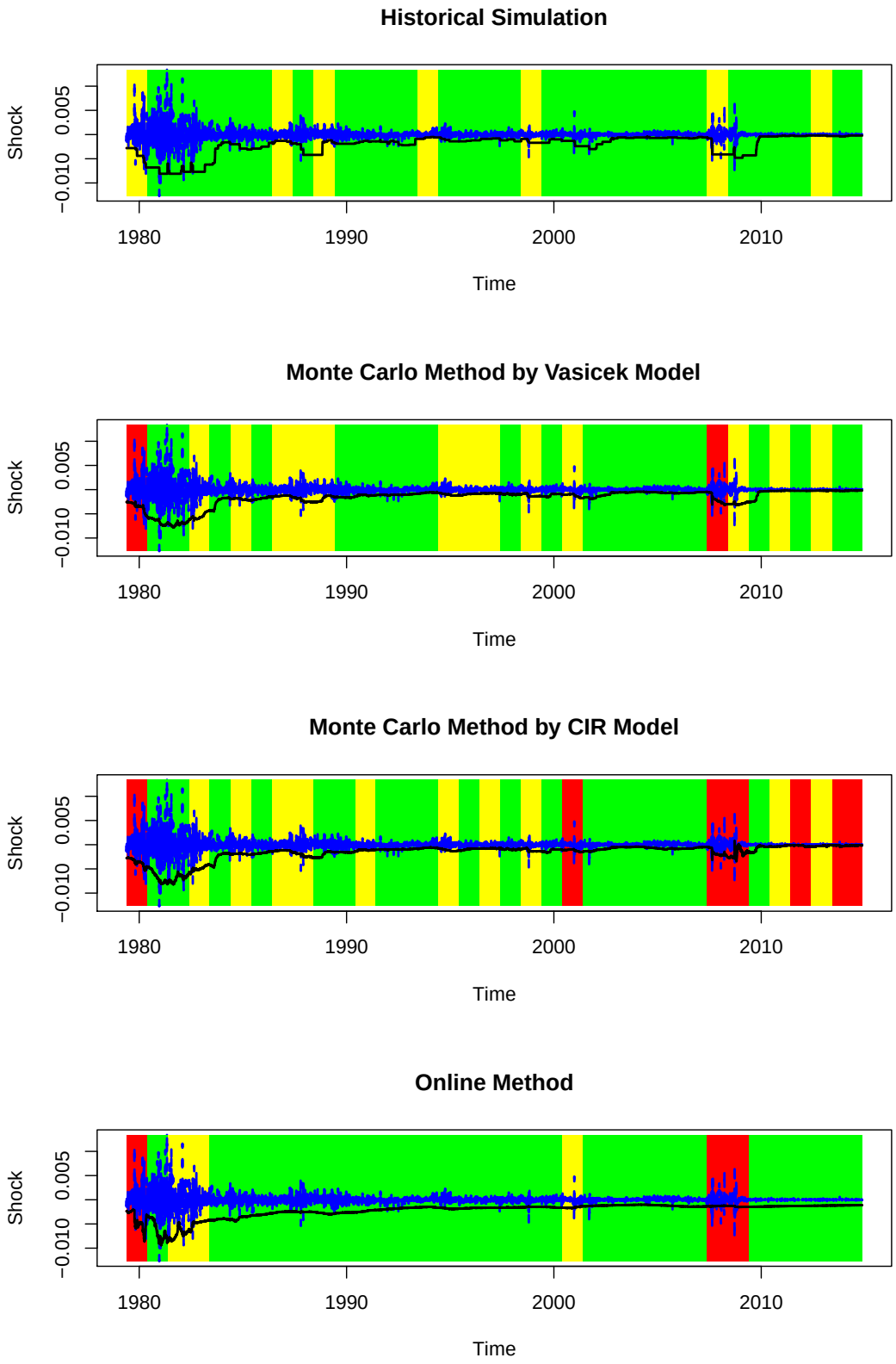

Figure 5.17: Calculation of -VaR by historical simulation, Monte Carlo method by the Vasicek and CIR model, and online method, where the dashed line is shocks and the solid line is -VaR.

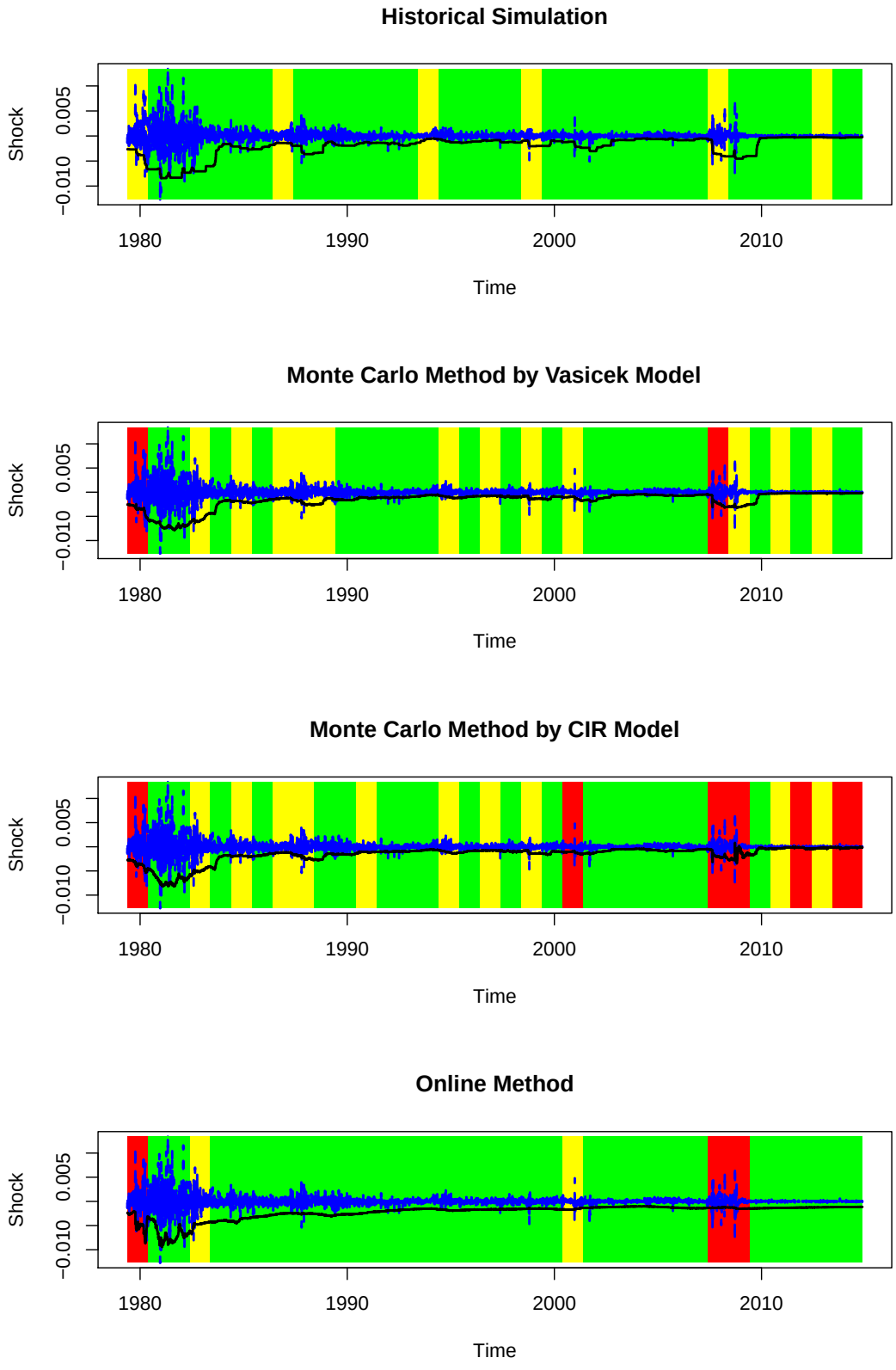

Figure 5.18: Calculation of -ES by historical simulation, Monte Carlo method by the Vasicek and CIR model, and online method, where the dashed line is shocks and the solid line is -ES.

kernel density estimator is chosen. Then (5.5) is rewritten as

$$-x_{\lceil n(1-\alpha)\rceil} \pm Z_{1-\frac{\tau}{2}} \sqrt{\frac{\alpha(1-\alpha)}{n\hat{f}^2(x_{\lceil n(1-\alpha)\rceil})}} \tag{5.6}$$

To derive the asymptotic confidence interval of ES, note that (2.4) is the special case of the trimmed mean which is given by

$$S_n = \frac{1}{\lceil \beta n \rceil - \lceil \alpha n \rceil} \sum_{i=\lceil \alpha n \rceil + 1}^{\lceil \beta n \rceil} X_{(i)}$$

where $X_{(i)}$ is the i-th order statistics of $X_1, X_2, \ldots, X_n$ with the distribution function $F(x)$. By letting $p_1 = F^{-1}(\alpha) - F^{-1}(\alpha-)$ and $p_2 = F^{-1}(\beta) - F^{-1}(\beta-)$ and defining

$$G(x) = \frac{F(x) - \alpha}{\beta - \alpha} I[F^{-1}(\alpha) \le x < F^{-1}(\beta-)] + I[x \ge F^{-1}(\beta-)]$$

and also letting

$$\mu_G = \int_{-\infty}^{\infty} x dG(x) \qquad \sigma_G^2 = \int_{-\infty}^{\infty} x^2 dG(x) - \mu_G^2$$

Stigler (1973) derived the following theorem

Theorem 5.2.1. *As $n \to \infty$, then $\sqrt{n}(S_n - \mu_G) \xrightarrow{d} W$ where W can be expressed as*

$$W = \frac{1}{\beta - \alpha} \left\{ Y + \left[F^{-1}(\alpha) - \mu_G \right] Y_1 + \left[F^{-1}(\beta) - \mu_G \right] Y_2 - p_1 \max(0, Y_1) + p_2 \max(0, Y_2) \right\}$$

where $Y \sim N(0, (\beta - \alpha)\sigma_G^2)$ which is independent of the random vector $(Y_1, Y_2) \sim N(\mathbf{0}, \boldsymbol{C})$ where

$$\boldsymbol{C} = \begin{pmatrix} \alpha(1-\alpha) & -\alpha(1-\beta) \\ -\alpha(1-\beta) & \beta(1-\beta) \end{pmatrix}$$

For our calculation, we can simply set $\alpha = 0$ and $\beta = 0.025$ for 97.5% ES. In addition, we assume that $X_1, X_2, \ldots, X_n$ are continuous random variables, so $p_1 = 0$ and $p_2 = 0$. Furthermore, we have

$$G(x) = \frac{F(x)}{\beta} I[x < F^{-1}(\beta)] + I[x \ge F^{-1}(\beta)]$$

so that $\sqrt{n}(-ES - \mu_G) \xrightarrow{d} W$ where

$$W = \frac{1}{\beta} \left\{ Y + \left[F^{-1}(\beta) - \mu_G \right] Y_2 \right\}$$

where $Y \sim N(0, \beta\sigma_G^2)$ and $Y_2 \sim N(0, \beta(1-\beta))$. Therefore we have

$$\sqrt{n}(ES + \mu_G) \xrightarrow{d} N\left(0, \frac{1}{\beta}\left\{\sigma_G^2 + [F^{-1}(\beta) - \mu_G]^2(1-\beta)\right\}\right)$$

In other words, the β-level ES has the asymptotic distribution

$$\sqrt{n}(ES_\beta + \mu_G) \xrightarrow{d} N\left(0, \frac{1}{1-\beta}\left\{\sigma_G^2 + [F^{-1}(1-\beta) - \mu_G]^2\beta\right\}\right)$$

Thus the confidence interval with $1-\tau$ level is

$$-\mu_G \pm \frac{Z_{1-\frac{\tau}{2}}}{\sqrt{n(1-\beta)}}\sqrt{\sigma_G^2 + [F^{-1}(1-\beta) - \mu_G]^2\beta} \tag{5.7}$$

For example, if we need to calculate a 95% confidence interval of 97.5% ES for 259 shocks, then the above expression is written as

$$-\mu_G \pm 0.77\sqrt{\sigma_G^2 + 0.975[F^{-1}(0.025) - \mu_G]^2}$$

But for (5.7), it is still required to calculate μ_G, σ_G^2 and $F^{-1}(1-\beta)$. Given the realization of ordered shocks $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, we can use $x_{\lceil n(1-\beta)\rceil}$ to approximate $F^{-1}(1-\beta)$. Moreover,

$$\begin{aligned}
\mu_G &= \frac{1}{1-\beta}\int_{-\infty}^{x_{\lceil n(1-\beta)\rceil}} x d\hat{F}(x) = \frac{1}{\lceil n(1-\beta)\rceil}\sum_{i=1}^{\lceil n(1-\beta)\rceil} x_{(i)} \\
\sigma_G^2 &= \frac{1}{1-\beta}\int_{-\infty}^{x_{\lceil n(1-\beta)\rceil}} x^2 d\hat{F}(x) - \mu_G^2 = \frac{1}{\lceil n(1-\beta)\rceil}\sum_{i=1}^{\lceil n(1-\beta)\rceil} x_{(i)}^2 - \mu_G^2
\end{aligned}$$

Hence for the above example, if we need to construct the 95% confidence interval of 97.5% ES for 259 shocks, then we have $\hat{F}^{-1}(0.025) = x_{(7)}$ and

$$\begin{aligned}
\mu_G &= \sum_{i=1}^{7} x_{(i)}/7 \\
\sigma_G^2 &= \sum_{i=1}^{7} x_{(i)}^2/7 - \mu_G^2
\end{aligned}$$

so the confidence interval is

$$-\mu_G \pm 0.77\sqrt{\sigma_G^2 + 0.975(x_{(7)} - \mu_G)^2}$$

Now we use (5.5) and (5.7) to derive the 95% confidence band of daily 99% VaR and daily 97.5% ES by historical simulation and the online method for the case of US 3-month treasury

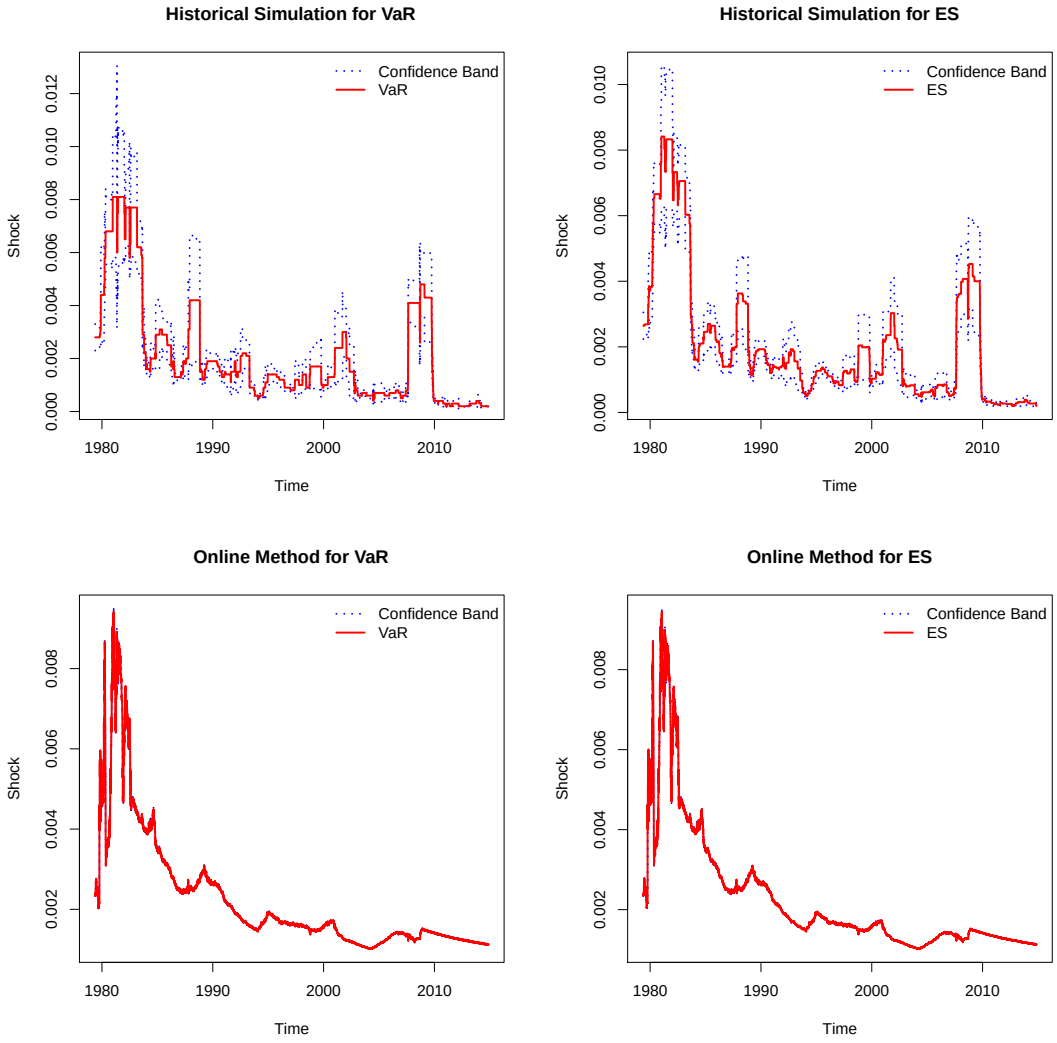

Figure 5.19: 95% Confidence band for daily 99% VaR and 97.5% ES by historical simulation and the online method.

bill yields. From Figure 5.19, note that there is a wide confidence band in the stressed period (that is, 1979-1982 and 2007-2009) by historical simulation but a narrow confidence band can be seen in the regular time. The online method consistently gives a narrow confidence band by (5.5) and (5.7).

5.3 Concluding Remark

In this chapter, we evaluate the performance of the online method by numerical examples and a case study. For numerical examples, we compare the online method with the offline counterpart for the Vasicek model and the CIR model, and it can be found that the online method runs much faster. Sensitivity to parameters in the method is studied as well. For the case study, we compare the online method with historical simulation and the parametric Monte

Carlo method by the Vasicek model and the CIR model, and we find that the online technique has the fewest breaches in all methods but is less responsive to the most recent changes of shocks. Additionally, we also derive the asymptotic confidence band for VaR and ES.

There are two things worth mentioning about comparison of the online method with its offline counterpart. The first one is a tradeoff between effectiveness and efficiency. As is shown above, the offline method has slightly better MISE but with much longer running time. In real application, the error rate with more than one decimal could not be reliable, however the big difference of the computational cost between the offline and online methods makes us believe that our online technique is more applicable. In addition, we do not need to store all past observations like in the offline method now that our online estimators have good performance in effectiveness and efficiency. Sometimes storage limitations of the data are another restriction on using the offline method.

In the case study, we calculate daily VaR and ES as market risk measures. In fact, in addition to the recommendation of using ES, the consultative document "Fundamental review of the trading book" by the Basel Committee also suggests different liquidity horizons for different risk factors. This is because it is realized that the liquidity risk partly contributed to the financial crisis since 2007. For example, the interest rate has 20-day liquidity horizon, the high yield credit spread has 120-day liquidity horizon, and the energy price has 20-day liquidity horizon (more details can be seen in the document). We have to mention that the 20-day VaR and ES by historical simulation have more red zones and breaches than the daily VaR and ES (see in Figure 5.20). Therefore in order to remedy the breaches and get the regulator's approval, one practical way is to multiply a factor to the shocks, however it means that more reserve capital are charged for financial institutions.

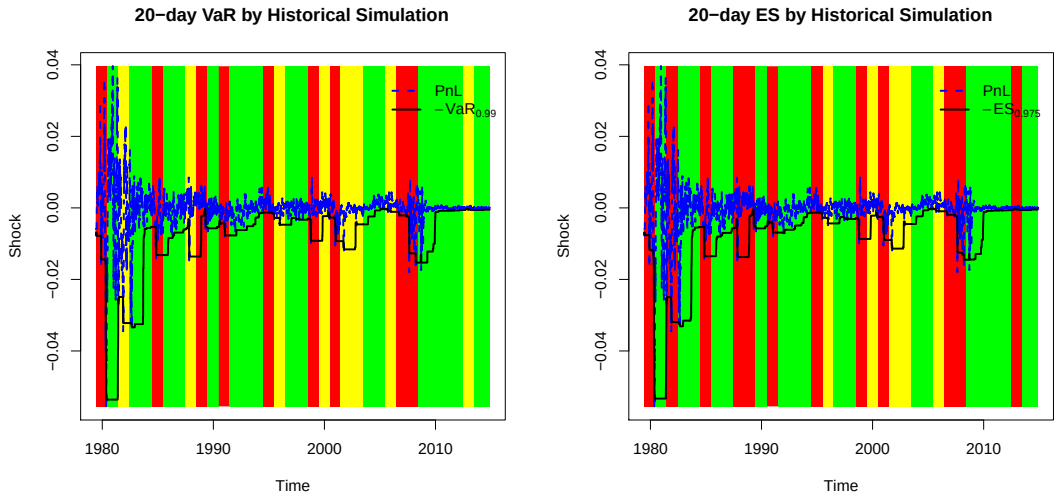

Figure 5.20: 20-day 99% VaR and 97.5% ES by historical simulation.

Note that in theoretical analysis and simulation studies, we use the same bandwidth for

the numerator and denominator of (4.4) and (4.6) for convenience. However in order to obtain more flexibility, we could choose different bandwidths for them. In addition, the optimal empirical bandwidth for the drift and diffusion will be investigated in the further work.

Chapter 6

Online Kernel Estimators for Second-Order Diffusion Models

We proposed the online kernel estimator for the stationary time-homogeneous diffusion model in Chapter 4 and studied its applications to the interest rate data in Chapter 5. However, some risk factors are not best described by a stationary model, e.g. the stock index, exchange rate and commodity price, so it is necessary to put forward the online estimator for the non-stationary time-homogeneous process.

In this chapter, we follow Nicolau (2007)'s idea that a non-stationary process $\{Y_t\}$ can be modelled by a second-order diffusion equation given by

$$\begin{cases} dY_t = X_t dt \\ dX_t = a(X_t)dt + b(X_t)dW_t \end{cases} \tag{6.1}$$

where $\{X_t\}$ is a latent stationary process. Note that the above equations can be written as $d\,(df(Y_t)/dt) = a(X_t)dt + b(X_t)dW_t$, so this is called a second-order equation.

This chapter is organized as follows. In Section 6.1, the online estimator is proposed, then Section 6.2 proves its weak consistency. Section 6.3 evaluates the methods by numerical simulation and real applications. A concluding remark is given in Section 6.4.

6.1 Method

Given discrete-time sequential observations $\{Y_i\}_{i\geq 0}$ at time points t_i in (6.1), the Euler scheme to approximate X_i is given by

$$\tilde{X}_i = \frac{Y_i - Y_{i-1}}{\triangle_i}$$

where $\triangle = \triangle_i = t_i - t_{i-1}$ and $\tilde{X}_i$ is the proxy of X_i because X_i is usually non-observable in practice. For example, the stock prices or indices are often available but their returns cannot

be observed directly. So we need to use these proxies to estimate the drift and diffusion in the second equation of (6.1).

Similar to (4.4) and (4.6), the online kernel estimator of the drift $a(x)$ is given by

$$\hat{g}_n(x) = \hat{g}_{n-1}(x) + \frac{1}{n}[U_n K_{h_n}(x - \tilde{X}_n) - \hat{g}_{n-1}(x)] \tag{6.2}$$

$$\hat{f}_n(x) = \hat{f}_{n-1}(x) + \frac{1}{n}[K_{h_n}(x - \tilde{X}_n) - \hat{f}_{n-1}(x)] \tag{6.3}$$

$$\hat{a}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)} \tag{6.4}$$

where $U_n = \frac{\tilde{X}_{n+1} - \tilde{X}_n}{\triangle}$ and the online kernel estimator of the diffusion $b^2(x)$ is given by

$$\hat{d}_n(x) = \hat{d}_{n-1}(x) + \frac{1}{n}[V_n K_{h_n}(x - \tilde{X}_n) - \hat{d}_{n-1}(x)] \tag{6.5}$$

$$\hat{b}^2_n(x) = \frac{\hat{d}_n(x)}{\hat{f}_n(x)} \tag{6.6}$$

where $V_n = \frac{\frac{3}{2}(\tilde{X}_{n+1} - \tilde{X}_n)^2}{\triangle}$.

From the above methods, it can be seen that instead of modeling the non-stationary process $\{Y_t\}$, we could assume $\{Y_t\}$ is built on a stationary process $\{X_t\}$. Then by differentiating $\{Y_t\}$, the proxy of $\{X_t\}$ is used to estimate the drift and diffusion by the methods that are applicable to the stationary process as is described in previous chapters. Thus this way helps us circumvent the problem of modeling the non-stationary process directly. From the practical perspective, it means that we can model the returns of a risk factor by a stationary diffusion model first and then integrate these returns as the estimate of the risk factor.

6.2 Theoretical Analysis

As mentioned above, the latent stationary process $\{X_t\}$ is non-observable in practice, so the proof given in Chapter 4 cannot simply be used here without any change. It is fortunate that we can apply the similar steps to prove weak consistency of the estimators by the proxy $\{\tilde{X}_t\}$. Here we give a sketch of the proof.

6.2.1 Assumptions and a Preliminary Lemma

In addition to **A1**-**A6** and the lemmas given in Chapter 4, we still need the following assumptions:

A7 For the kernel function $K(\cdot)$, $E[|K'(\psi_i)|^\alpha/h]$ is bounded for all i's where $\alpha = 2$ or 4 and $\psi_i = \lambda[(x - X_i)/h] + (1 - \lambda)[(x - \tilde{X}_i)/h]$ for $\lambda \in [0, 1]$

A8 The discretization step size $\triangle$ and the bandwidth h_n satisfy $\triangle/h_n^5 \to 0$ as $n \to \infty$

A9 The following relationship is required:

$$\alpha_l = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{h_n}{h_i}\right)^l < \infty \text{ for } l = 7/2$$

In fact, as pointed by Nicolau (2007), the stationarity of $\{X_i\}$ implies that $\{\tilde{X}_i\}$ is stationary. Both $\{X_i\}$ and $\{\tilde{X}_i\}$ are mixing with the same value of mixing coefficients (see Appendix C). So according to assumption 3, $\{\tilde{X}_i\}$ is also ρ-mixing with exponential decay. In addition, we assume that the item in the expression $O(\triangle^k)$ or $o(\triangle^k)$ is uniform in its domain for any $k \geq 0$.

Next, the following lemma is useful:

Lemma 6.2.1 (Nicolau, 2007)**.** *For (6.1), one can have*

$$\begin{aligned} E[(\tilde{X}_{t+\triangle} - \tilde{X}_t)|X_t] &= a(X_t)\triangle + O(\triangle^2) \\ E[(\tilde{X}_{t+\triangle} - \tilde{X}_t)^2|X_t] &= \frac{2}{3}b^2(X_t)\triangle + O(\triangle^2) \\ E[(\tilde{X}_t - X_t)^4|X_t] &= \frac{1}{3}b^4(X_t)\triangle^2 + O(\triangle^3) \end{aligned}$$

6.2.2 Weak Consistency

We have established quadratic convergence of the estimators for the stationary process $\{X_i\}$, and quadratic convergence implies weak consistency. Now we will show that the statement we have obtained for $\{X_i\}$ is also applicable to the proxy $\{\tilde{X}_i\}$.

Similarly we can write (6.3) as $\hat{f}_n(x) = \sum_{i=1}^n K_{h_i}(x - \tilde{X}_i)/n$, then we have the following proposition:

Proposition 6.2.2. *Under **A1**-**A8**, we have the following statement*

$$\hat{f}_n(x) \xrightarrow{p} f(x)$$

Proof. Let $\tilde{f}_n(x) = \sum_{i=1}^n K_{h_i}(x - X_i)/n$, and in Proposition 4.3.5, we have proved that $E|\tilde{f}_n(x) - f(x)| \to 0$ and the triangular inequality implies that

$$E|\hat{f}_n(x) - f(x)| \leq E|\tilde{f}_n(x) - f(x)| + E|\hat{f}_n(x) - \tilde{f}_n(x)|$$

Now it suffices to show $E|\hat{f}_n(x) - \tilde{f}_n(x)| \to 0$. Note that

$$E|\hat{f}_n(x) - \tilde{f}_n(x)| = E\left|\frac{1}{n}\sum_{i=1}^{n}\left[K_{h_i}\left(\frac{x-\tilde{X}_i}{h_i}\right) - K_{h_i}\left(\frac{x-X_i}{h_i}\right)\right]\right| \leq \frac{1}{n}\sum_{i=1}^{n} E\left|K_{h_i}\left(\frac{x-\tilde{X}_i}{h_i}\right) - K_{h_i}\left(\frac{x-X_i}{h_i}\right)\right|$$

The mean value theorem implies that

$$K\left(\frac{x-\tilde{X}_i}{h_i}\right) = K\left(\frac{x-X_i}{h_i}\right) + K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}$$

so from the Hölder inequality,

$$E|\hat{f}_n(x) - \tilde{f}_n(x)| \le \frac{1}{n}\sum_{i=1}^{n} E\left|\frac{K'(\psi_i)(X_i-\tilde{X}_i)}{h_i^2}\right| \le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^{3/2}}E^{1/2}[|K'(\psi_i)|^2/h_i]E^{1/2}(X_i-\tilde{X}_i)^2$$

By the Lyapunov inequality, note that $E^{1/2}(X_i-\tilde{X}_i)^2 \le E^{1/4}(X_i-\tilde{X}_i)^4$, then by **A7** and **A8**

$$\begin{aligned} E|\hat{f}_n(x) - \tilde{f}_n(x)| &\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^{1/3}}E^{1/2}[|K'(\psi_i)|^2/h_i]\cdot E^{1/4}(X_i-\tilde{X}_i)^4 \\ &\le C\left[\frac{1}{n}\sum_{i=1}^{n}\frac{h_n^{3/2}}{h_i^{3/2}}\right]\frac{\left[\frac{1}{3}\triangle^2 E[b^4(X_0)] + O(\triangle^3)\right]^{1/4}}{h_n^{3/2}} \to 0 \end{aligned}$$

Therefore it is obtained that $E|\hat{f}_n(x) - f(x)| \to 0$. Then from the Markov inequality,

$$P[|\hat{f}_n(x) - f(x)| > \varepsilon] \le \frac{E|\hat{f}_n(x) - f(x)|}{\varepsilon} \to 0$$

that is,

$$\hat{f}_n(x) \xrightarrow{p} f(x)$$

□

Proposition 6.2.3. *Under **A1**-**A9**, it can be obtained that*

$$\hat{g}_n(x) \xrightarrow{p} a(x)f(x)$$

Proof. For $\hat{g}_n(x) = \sum_{i=1}^{n}\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}K_{h_i}(x-\tilde{X}_i)/n$, denote $\tilde{g}_n(x) = \sum_{i=1}^{n}\frac{X_{i+1}-X_i}{\triangle}K_{h_i}(x-X_i)/n$, then it suffices to show

$$\hat{g}_n(x) - \tilde{g}_n(x) \xrightarrow{p} 0$$

To this end, let

$$\begin{aligned} \delta_{1,n} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\left[K_{h_i}(x-\tilde{X}_i) - K_{h_i}(x-X_i)\right] \\ \delta_{2,n} &= \frac{1}{n}\sum_{i=1}^{n}K_{h_i}(x-X_i)\left[\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle} - \frac{X_{i+1}-X_i}{\triangle}\right] \end{aligned}$$

so it is sufficient to prove that $\delta_{1,n} \xrightarrow{p} 0$ and $\delta_{2,n} \xrightarrow{p} 0$.

First, by the mean value theorem

$$\delta_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}$$

Then

$$\begin{aligned} E\delta_{1,n} &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}E\left[\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\middle| X_i\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}\left[a(X_i)+O(\triangle)\right]\right\} \end{aligned}$$

By the Hölder inequality, Lyapunov inequality and stationarity

$$|E\delta_{1,n}| \leq \frac{1}{n}\sum_{i=1}^{n}\frac{h_n^{3/2}}{h_i^{3/2}}E^{1/2}[|K'(\psi_i)|^2/h_i]\frac{\left[\frac{1}{3}\triangle^2 E[b^4(X_0)]+O(\triangle^3)\right]^{1/4}}{h_n^{3/2}}E^{1/4}\left[a(X_0)+O(\triangle)\right]^4 \to 0$$

Furthermore, we have

$$Var[\delta_{1,n}] = \frac{1}{n\triangle}Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{\triangle}\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}\right] \triangleq \frac{1}{n\triangle}Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_i\right]$$

where

$$r_i = \sqrt{\triangle}\frac{\tilde{X}_{i+1}-\tilde{X}_i}{\triangle}\frac{1}{h_i}K'(\psi_i)\frac{X_i-\tilde{X}_i}{h_i}$$

Also denote $Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_i\right] \triangleq I_n + R_n$ where

$$\begin{aligned} I_n &= \frac{1}{n}\sum_{i=1}^{n} Var[r_i] \\ R_n &= \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov[r_i, r_j] \end{aligned}$$

Similar to the proof in Chapter 4, we have

$$\begin{aligned} E[r_i^2] &= E\left\{\frac{1}{h_i^2}|K'(\psi_i)|^2\frac{(X_i-\tilde{X}_i)^2}{h_i^2}E\left[\frac{(\tilde{X}_{i+1}-\tilde{X}_i)^2}{\triangle}\middle| X_i\right]\right\} \\ &= E\left\{\frac{1}{h_i^2}|K'(\psi_i)|^2\frac{(X_i-\tilde{X}_i)^2}{h_i^2}\left[\frac{2}{3}b^2(X_0)+O(\triangle)\right]\right\} \end{aligned}$$

then by the Hölder inequality and Lyapunov inequality

$$E[r_i^2] \leq \frac{1}{h_i^{3/2}} E^{1/2}[|K'(\psi_i)|^4/h_i] \frac{\left[\frac{1}{3}\triangle^2 E[b^4(X_0)] + O(\triangle^3)\right]^{1/4}}{h_i^2} E^{1/4}\left[\frac{2}{3}b^2(X_0) + O(\triangle)\right]^4$$

thus under **A8** and **A9**

$$I_n \leq \frac{1}{n}\sum_{i=1}^{n} E[r_i^2] = O\left(\frac{1}{n}\sum_{i=1}^{n} \frac{h_n^{7/2}}{h_i^{7/2}} \frac{\triangle^{1/2}}{h_n^{5/2}} \frac{1}{h_n}\right) = o(1/h_n)$$

On the other hand,

$$\begin{aligned} R_n &\leq \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho(j-i) E^{1/2}[r_i^2] E^{1/2}[r_j^2] = \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho(j-i) O\left(\frac{\triangle^{1/4}}{h_j^{7/4}} \frac{\triangle^{1/4}}{h_i^{7/4}}\right) \\ &= \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \rho(j-i) O\left(\frac{\triangle^{1/2}}{h_n^{7/2}} \frac{h_n^{7/2}}{h_i^{7/4} h_j^{7/4}}\right) = O\left(\frac{\triangle^{1/2}}{h_n^{7/2}}\right) = o(1/h_n) \end{aligned}$$

Therefore

$$Var[\delta_{1,n}] = \frac{1}{n\triangle}(I_n + R_n) = o\left(\frac{1}{nh_n\triangle}\right) = o(1)$$

So by the Chebyshev inequality we can obtain that $\delta_{1,n} \xrightarrow{p} 0$.

Similarly, we can prove that

$$E\delta_{2,n} = \frac{1}{n}\sum_{i=1}^{n} E\left\{K_{h_i}(x - X_i) E\left[\left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle} - \frac{X_{i+1} - X_i}{\triangle}\right)\middle| X_i\right]\right\} = 0$$

furthermore,

$$Var[\delta_{2,n}] = \frac{1}{n\triangle} Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} K_{h_i}(x - X_i)\sqrt{\triangle}\left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle} - \frac{X_{i+1} - X_i}{\triangle}\right)\right] \triangleq \frac{1}{n\triangle} Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right]$$

where

$$s_i = K_{h_i}(x - X_i)\sqrt{\triangle}\left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle} - \frac{X_{i+1} - X_i}{\triangle}\right)$$

We have

$$\begin{aligned} E[s_i^2] &= E\left[K_{h_i}^2(x - X_i)\triangle\left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle} - \frac{X_{i+1} - X_i}{\triangle}\right)^2\right] \\ &= E\left\{K_{h_i}^2(x - X_i) E\left[\triangle\left(\frac{\tilde{X}_{i+1} - \tilde{X}_i}{\triangle} - \frac{X_{i+1} - X_i}{\triangle}\right)^2 \middle| X_i\right]\right\} \end{aligned}$$

By the inequality $(a-b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned} E[s_i^2] &\le 2E\left\{K_{h_i}^2(x-X_i)\left[E\left(\frac{(\tilde{X}_{i+1}-\tilde{X}_i)^2}{\triangle}\middle|X_i\right)+E\left(\frac{(X_{i+1}-X_i)^2}{\triangle}\middle|X_i\right)\right]\right\} \\ &= \frac{10}{3}E\left\{K_{h_i}^2(x-X_i)[b^2(X_i)+O(\triangle)]\right\} = \frac{10}{3h_i}\int_R \frac{1}{h_i}K^2\left(\frac{x-z}{h_i}\right)[b^2(z)+O(\triangle)]f(z)dz \\ &= O(1/h_i) \end{aligned}$$

so

$$I_n' = \frac{1}{n}\sum_{i=1}^{n} Var[s_i] \le \frac{1}{n}\sum_{i=1}^{n} E[s_i^2] \le \frac{1}{n}\sum_{i=1}^{n} O(1/h_i) = O(1/h_n)$$

and

$$R_n' \le \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho(j-i)E^{1/2}[s_i^2]E^{1/2}[s_j^2] \le \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho(j-i)O\left(\frac{1}{h_n}\frac{h_n}{h_i^{1/2}h_j^{1/2}}\right) = O(1/h_n)$$

therefore

$$Var[\delta_{2,n}] = \frac{1}{n\triangle}(I_n' + R_n') \le O\left(\frac{1}{nh_n\triangle}\right) = o(1)$$

implying that $\delta_{2,n} \xrightarrow{p} 0$ by the Chebyshev inequality.

Finally, from the statement $\delta_{1,n} \xrightarrow{p} 0$ and $\delta_{2,n} \xrightarrow{p} 0$, we can obtain that

$$P(|\hat{g}_n(x)-\tilde{g}_n(x)|>\varepsilon) = P(|\delta_{1,n}+\delta_{2,n}|>\varepsilon) \le P(|\delta_{1,n}|>\varepsilon/2)+P(|\delta_{2,n}|>\varepsilon/2) \to 0$$

so $\hat{g}_n(x)-\tilde{g}_n(x) \xrightarrow{p} 0$. And in Proposition 4.3.4 we have proved that $E\tilde{g}_n(x) \to a(x)f(x)$ implying $\tilde{g}_n(x) \xrightarrow{p} a(x)f(x)$, therefore we can conclude that

$$\hat{g}_n(x) \xrightarrow{p} a(x)f(x)$$

□

By combining Proposition 6.2.2 and Proposition 6.2.3, the following theorem holds true

Theorem 6.2.4. *Under **A1**-**A9**, it can be obtained that*

$$\hat{a}_n(x) \xrightarrow{p} a(x)$$

Proof. Because $\hat{g}_n(x) \xrightarrow{p} a(x)f(x)$ and $\hat{f}_n(x) \xrightarrow{p} f(x)$ in the above propositions, we have

$$\hat{a}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)} \xrightarrow{p} \frac{a(x)f(x)}{f(x)} = a(x)$$

□

Similarly, we can prove that

Theorem 6.2.5. *Under **A1**-**A9**, it can be obtained that*

$$\hat{b}_n^2(x) \xrightarrow{p} b^2(x)$$

6.3 Examples

This part will evaluate the performance of online kernel estimators (6.4) and (6.6) by numerical examples and real data of the stock indices, FX rates and commodity prices.

6.3.1 Numerical Simulation

We borrow an example from (Nicolau, 2007) where the first equation in (6.1) is written as the integrated form $Y_t = 10 + \int_0^t X_s ds$ and its differentiated process is given by

$$dX_t = -10X_t dt + \sqrt{0.1 + 0.1X_t^2} dW_t \tag{6.7}$$

defined in the interval $t \in [0, 10]$, and the Euler scheme is used to generate the sequential path with $\triangle = 1/100$ and $x_0 = 0$ (see in Figure 6.1). From the figure, we can see the integrated

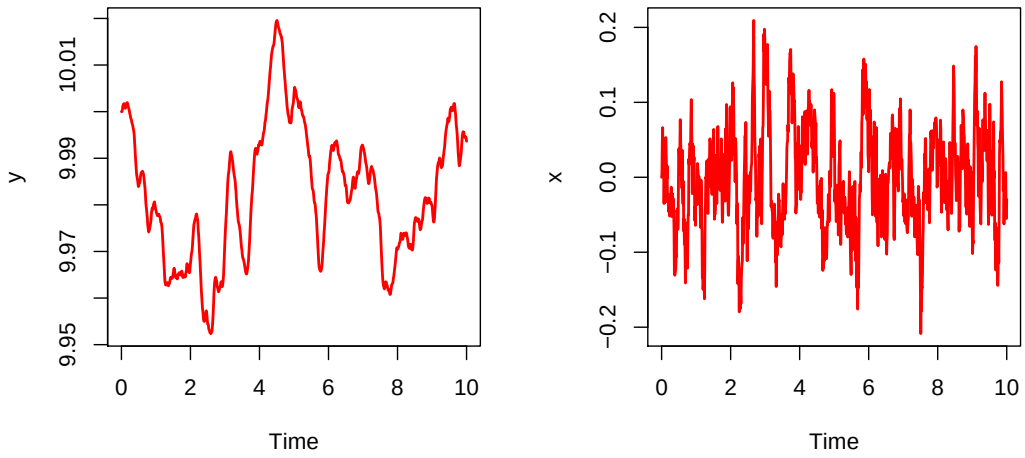

Figure 6.1: The sample path of (6.7) as well as the integrated process

process (the left) is the cumulation of all past changes (the right).

Next, we apply the online kernel estimator to fit the drift $a(x)$ and diffusion $b^2(x)$, where 20% of observations are used to initiate the procedures. Also, 1000 replicas of $\{X_t\}$ and $\{Y_t\}$ are generated for evaluation. MISE as mentioned in Chapter 5 is used to measure the dynamics

of the performance. The standard Gaussian kernel function is selected and the empirical bandwidth is chosen as $h_i = \hat{\sigma}_i \times i^{-0.02}$ for the drift and $h_i = \hat{\sigma}_i \times i^{-0.01}$ for the diffusion[1] such that the fitting curve is smooth. In addition, we use sample quantiles to calculate the confidence band of the drift and diffusion in order to check the fitting quality of the method.

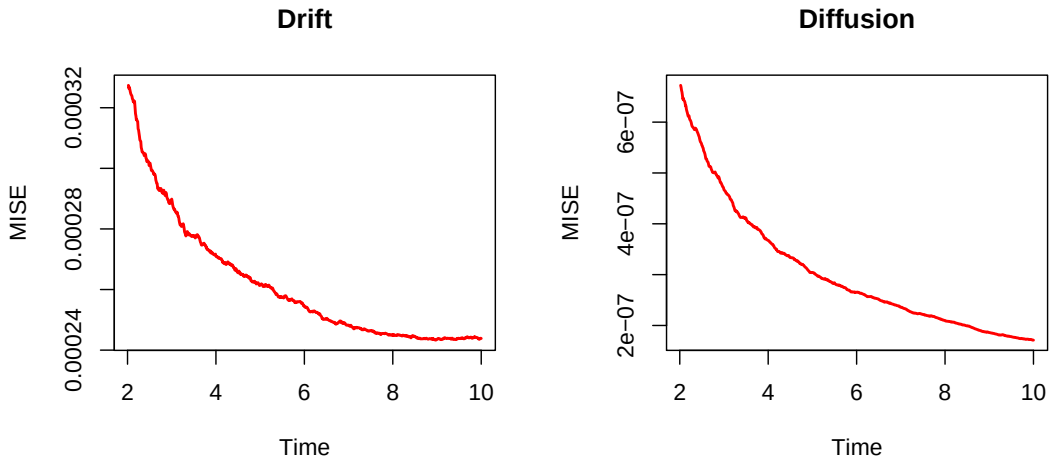

Figure 6.2: MISE behaviors of (6.4) and (6.6) for sequential observations

From Fig 6.2, we can find that as more observations are available, MISE decays to a low level gradually. Thus this example indicates that more information is helpful to get more accurate estimation.

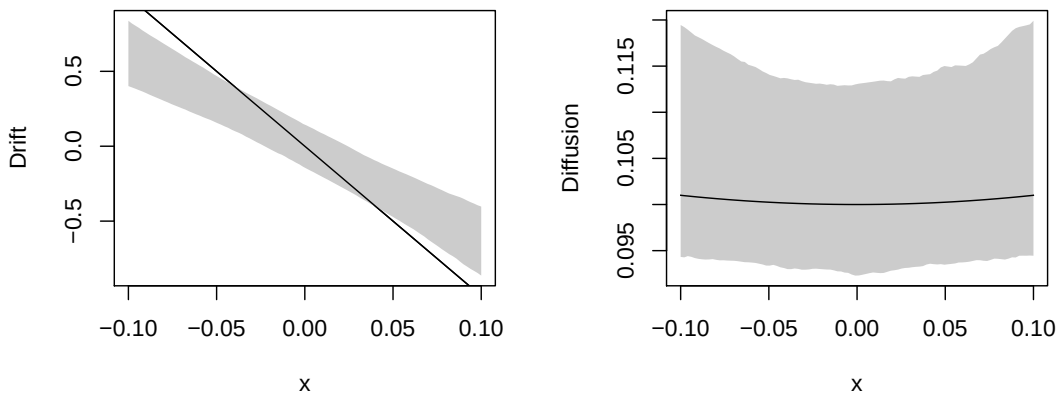

Figure 6.3: The 95% confidence band of the estimators (6.4) and (6.6). The solid line is the true value.

We also check the fitting of the online estimators of the true drift and diffusion (Figure 6.3). It is noted that the estimator can fit the diffusion better than the drift.

[1]The bandwidth converges to 0.058 for the drift and 0.062 for the diffusion in this example.

6.3.2 Real Application

In this section, we model the stock indices, FX rates and commodity prices by a second-order SDE and then apply the online estimators to update estimation of the drift and diffusion. Similar to (Nicolau, 2008), we assume that

$$\begin{cases} d\log Y_t = X_t dt \\ dX_t = a(X_t)dt + b(X_t)dW_t \end{cases} \tag{6.8}$$

where $\{Y_t\}$ is the integrated process for the risk factor (i.e. the stock indices, FX rates and commodity prices), and $\{X_t\}$ is the latent process for the log-returns. In other words, the risk factor is assumed to have the logarithm form. So the proxy of the latent process is given by

$$\tilde{X}_i = \frac{\log Y_{i+1} - \log Y_i}{\triangle} \tag{6.9}$$

Stock Index

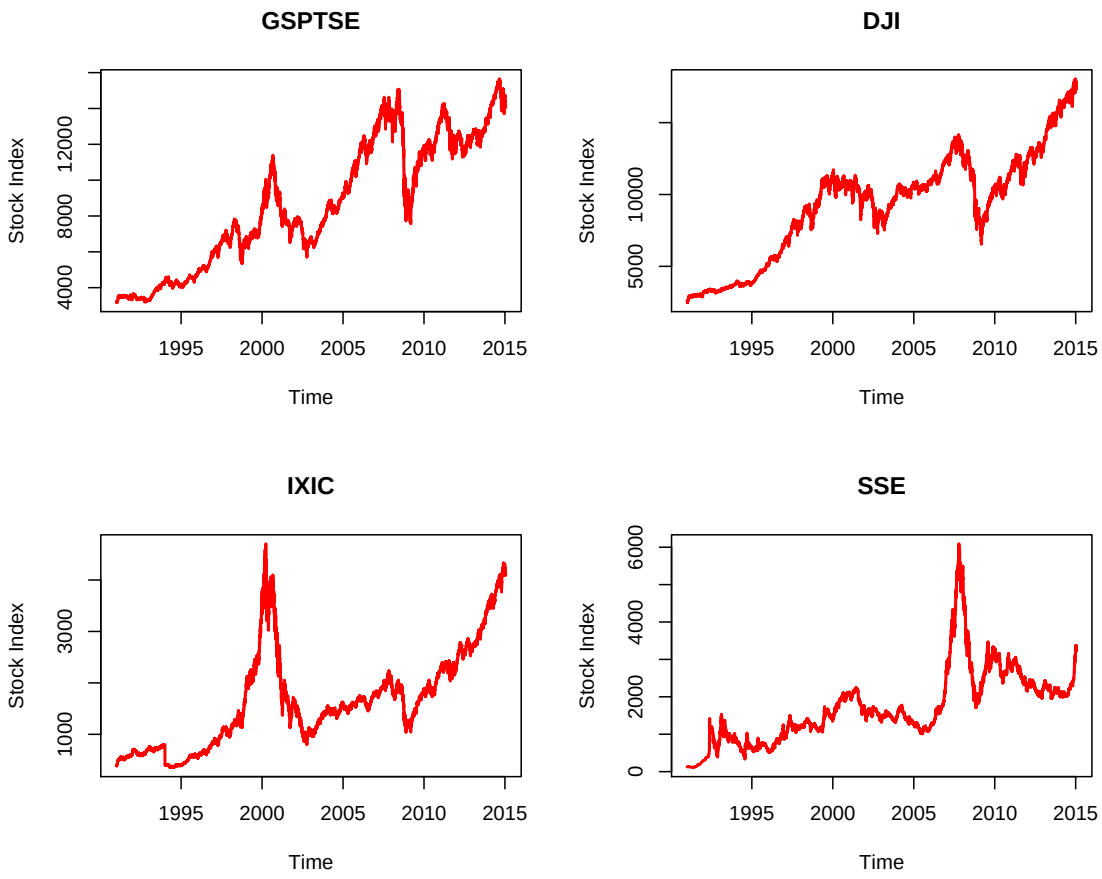

Figure 6.4: The time series of the stock index GSPTSE, DJI, IXIC and SSE from Jan 2, 1991 to Jan 16, 2015.

We apply our online method to estimate the drift and diffusion for four different stock indices from around the world, i.e. S&P/TSX Composite Index (GSPTSE), Dow Jones Industrial Average (DJI), NASDAQ Composite (IXIC) and Shanghai Composite Index (SSE), where GSPTSE can be seen in Chapter 2. Three other stock indices are cited from Yahoo! Finance. All stock indices cover the period from Jan 2, 1991 to Jan 16, 2015, where there are 6065 observations for GSPTSE, 6060 observations for DJI and IXIC, and 6161 observations for SSE. The time series of four stock indices are plotted in Figure 6.4 and the proxy of the latent process is seen in Figure 6.5.

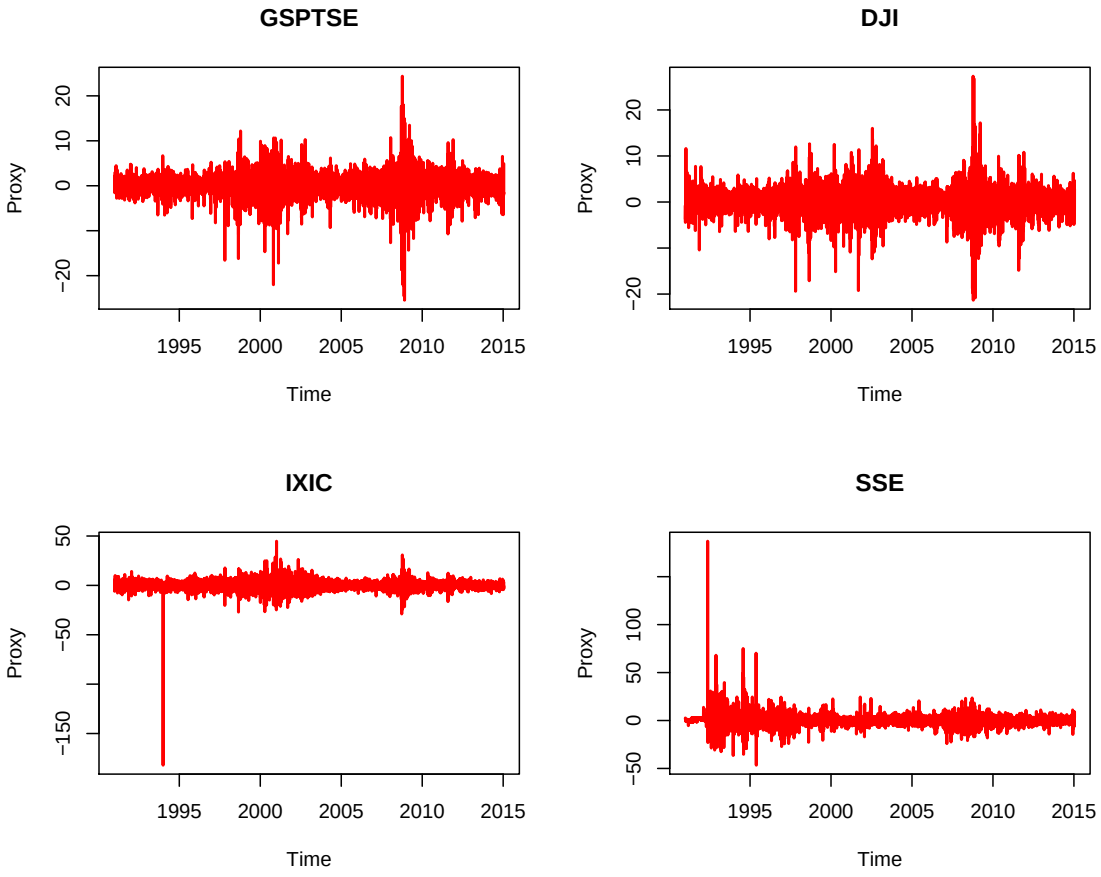

Figure 6.5: The proxy of the stock index GSPTSE, DJI, IXIC and SSE (daily data from Jan 2, 1991 to Jan 16, 2015).

Through the Augmented Dickey-Fuller stationarity test (Table 6.1), we see that the null hypothesis of non-stationarity is accepted at the 5% significance level for the stock index but is rejected for the proxy of the corresponding latent process.

In addition, if GBM is used, the parameters calibrated by (A.13) in Appendix A are given by Table 6.2. From the table, we can see that SSE has higher volatility but with more expected returns, whereas GSPTSE has lower volatility and fewer expected returns.

Now we will use the online kernel estimators (6.4) and (6.6) to find the drift and diffusion

	Augmented Dickey-Fuller stationarity test	
	Test statistic	p value
GSPTSE	-2.7651	0.2543
Proxy	-17.7712	< 0.01
DJI	-1.8889	0.6254
Proxy	-18.5911	< 0.01
IXIC	-1.3237	0.8648
Proxy	-17.3208	< 0.01
SSE	-2.8485	0.219
Proxy	-16.443	< 0.01

Table 6.1: Augmented Dickey-Fuller stationarity test of stock indices GSPTSE, DJI, IXIC and SSE as well as their proxies.

	Calibrated Parameters	
	$\hat{\mu}$	$\hat{\sigma}$
GSPTSE	0.078	0.165
DJI	0.098	0.176
IXIC	0.152	0.322
SSE	0.209	0.377

Table 6.2: The calibrated parameters if GBM is assumed to model the stock index.

in the latent process by the proxy (6.9). Similar to the numerical simulation, 20% of the observations are used to initiate the procedure. The standard Gaussian kernel function is chosen and the bandwidth is $h_i = \hat{\sigma}_i \times i^{-0.02}$ for the drift and $h_i = \hat{\sigma}_i \times i^{-0.01}$ for the diffusion, seen in Table 6.3.

	Drift	Diffusion
GSPTSE	2.108	2.309
DJI	2.378	2.594
IXIC	3.894	4.249
SSE	5.101	5.566

Table 6.3: The convergent bandwidth in online estimators of the drift and diffusion.

Figure 6.6 shows online estimation of the drift and diffusion in the latent process. It can be found that all results show the linear drift for the latent process, but different patterns for the diffusion. Meanwhile, GSPTSE and DJI exhibit the similar patterns for the diffusion. In addition, note that in the figure, the estimated drift is shown to be a decreasing function. This indicates the higher daily log-returns correspond to the lower drift in the latent process.

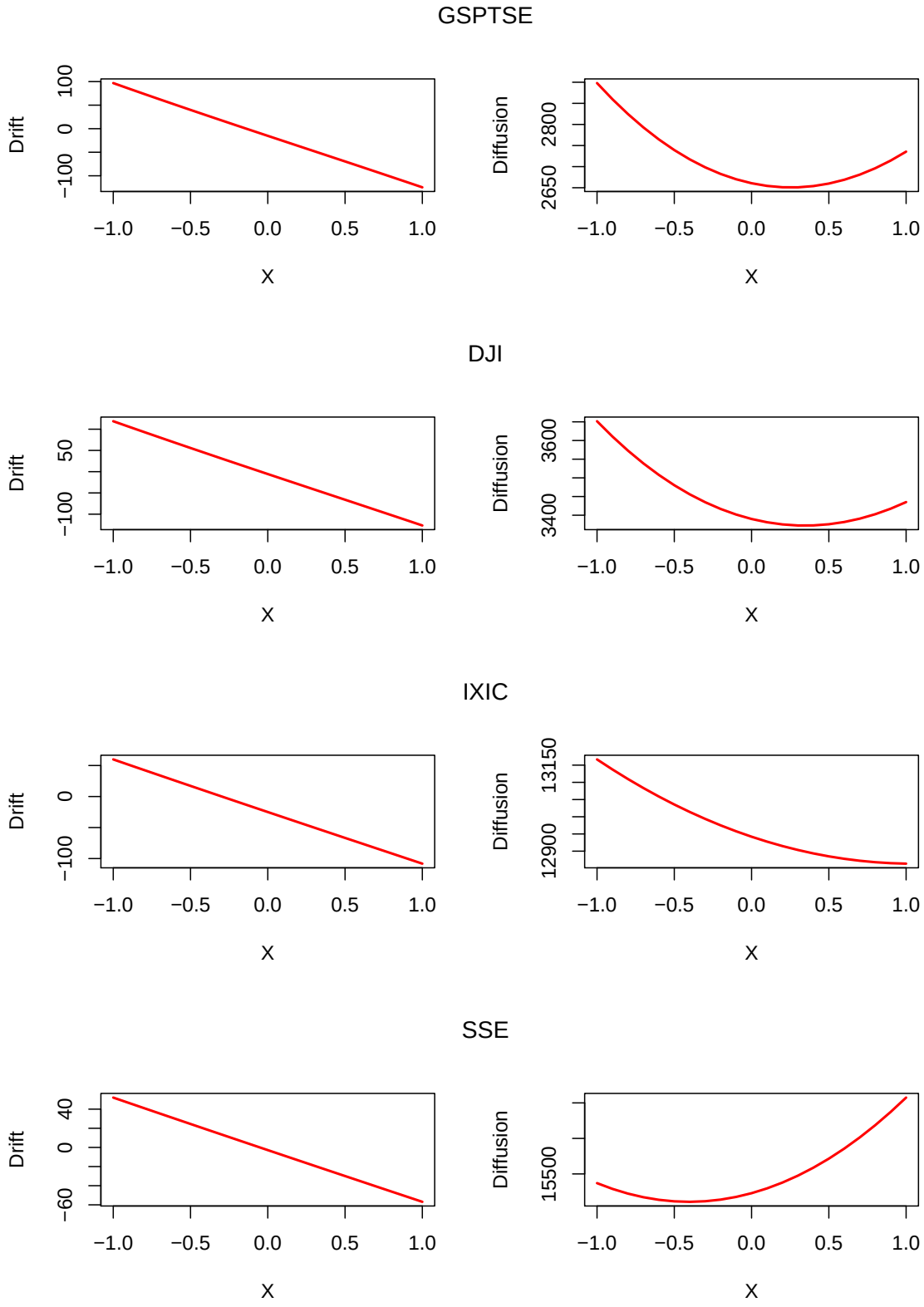

Figure 6.6: Online estimation of the drift and diffusion in the latent process $\{\tilde{X}_i\}$ for the stock index GSPTSE, DJI, IXIC and SSE.

FX Rate

It is also known that FX rate is non-stationary and it is often modeled by GBM. In this part, we apply the online estimator to FX rate, where five currencies, i.e. Canadian Dollar (CAD), US Dollar (USD), Chinese Yuan (CNY), British Pound (GBP) and Euro (EUR), and four FX rates, i.e. CAD/CNY, CAD/USD, CAD/GBP and CAD/EUR are considered. The weekly data are cited from OANDA, covering the period of five years from Jan 31, 2010 to Jan 16, 2015.

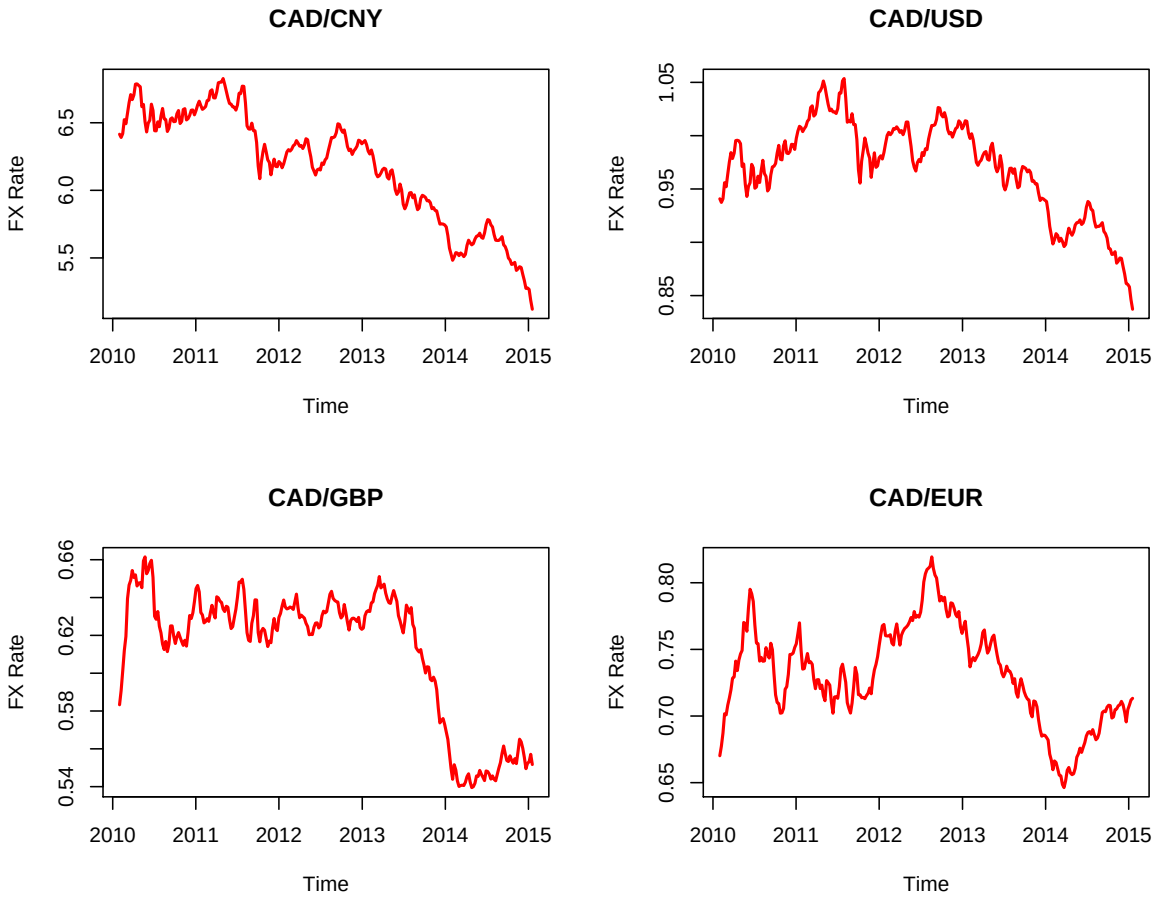

Figure 6.7: The FX rate of CAD/CNY, CAD/USD, CAD/GBP and CAD/EUR from Jan 31, 2010 to Jan 16, 2015.

Figure 6.7 shows that the FX rate of CAD/CNY and CAD/USD is declining since the middle of 2014 due to the decrease of the crude oil price. Generally speaking, CAD becomes weaker because it is vulnerable to the commodity price. Similarly, we use (6.8) to model the FX rate and (6.9) to obtain the proxy of the latent process, where $\triangle = 1/52$ corresponding to the weekly data. The proxy of the latent process is seen in Figure 6.8.

The parameters calibrated by (A.13) in Appendix A are given in Table 6.4 if GBM is used to model the FX rate. The results in the table confirm the statement that Canadian dollar has recently weakened because of the negative drifts for all rows.

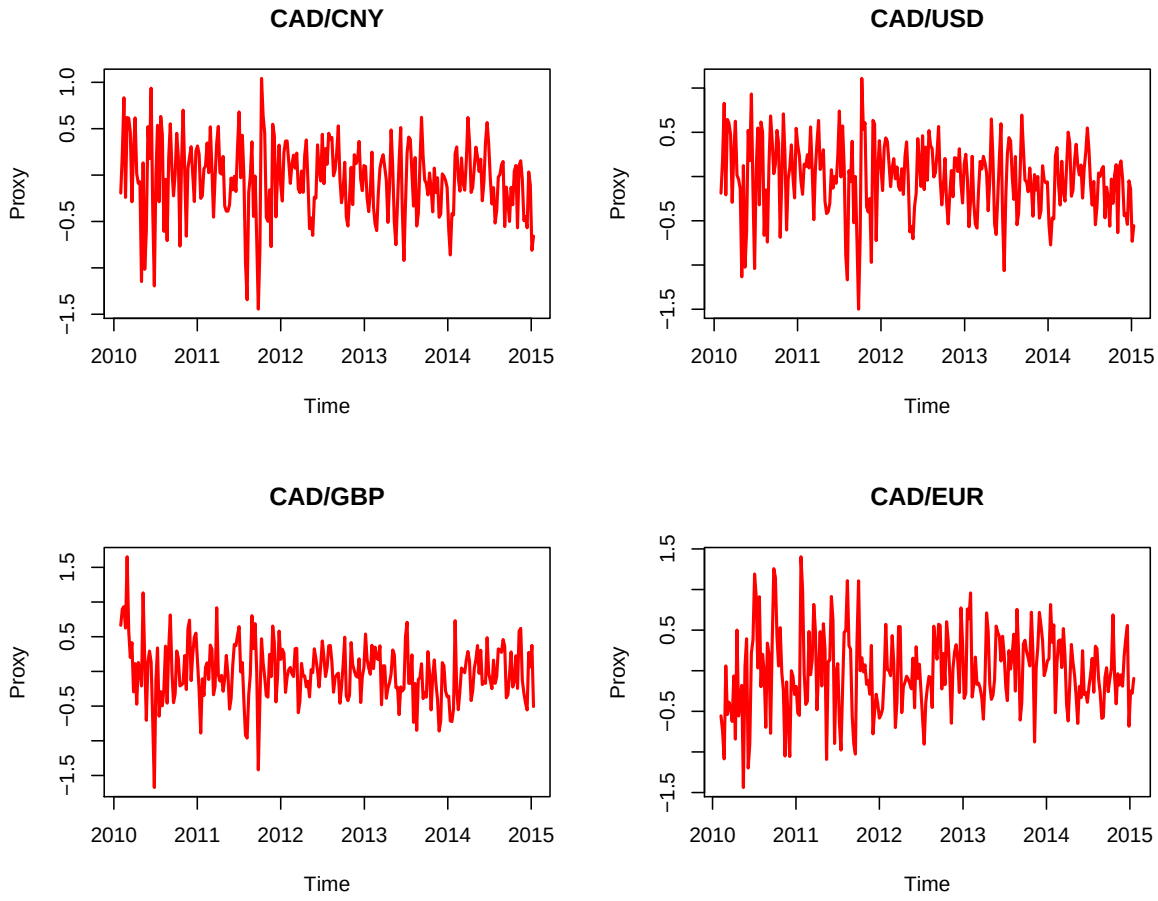

Figure 6.8: The proxy of CAD/CNY, CAD/USD, CAD/GBP and CAD/EUR.

	Calibrated Parameters	
	$\hat{\mu}$	$\hat{\sigma}$
CAD/CNY	-0.044	0.055
CAD/USD	-0.022	0.057
CAD/GBP	-0.010	0.057
CAD/EUR	-0.010	0.068

Table 6.4: The calibrated parameters if GBM is assumed to model the FX rate.

Next the online estimator is used to estimate the drift and diffusion. The choices of the kernel function and bandwidth are the same as those used for the stock index. From Figure 6.9, we can find that the latent processes of different FX rates display similar patterns, i.e. the linear drift and quadratic diffusion. Additionally, it can be seen that the drift in the latent process is estimated to be decreasing and the diffusion reaches the minimum near the naught.

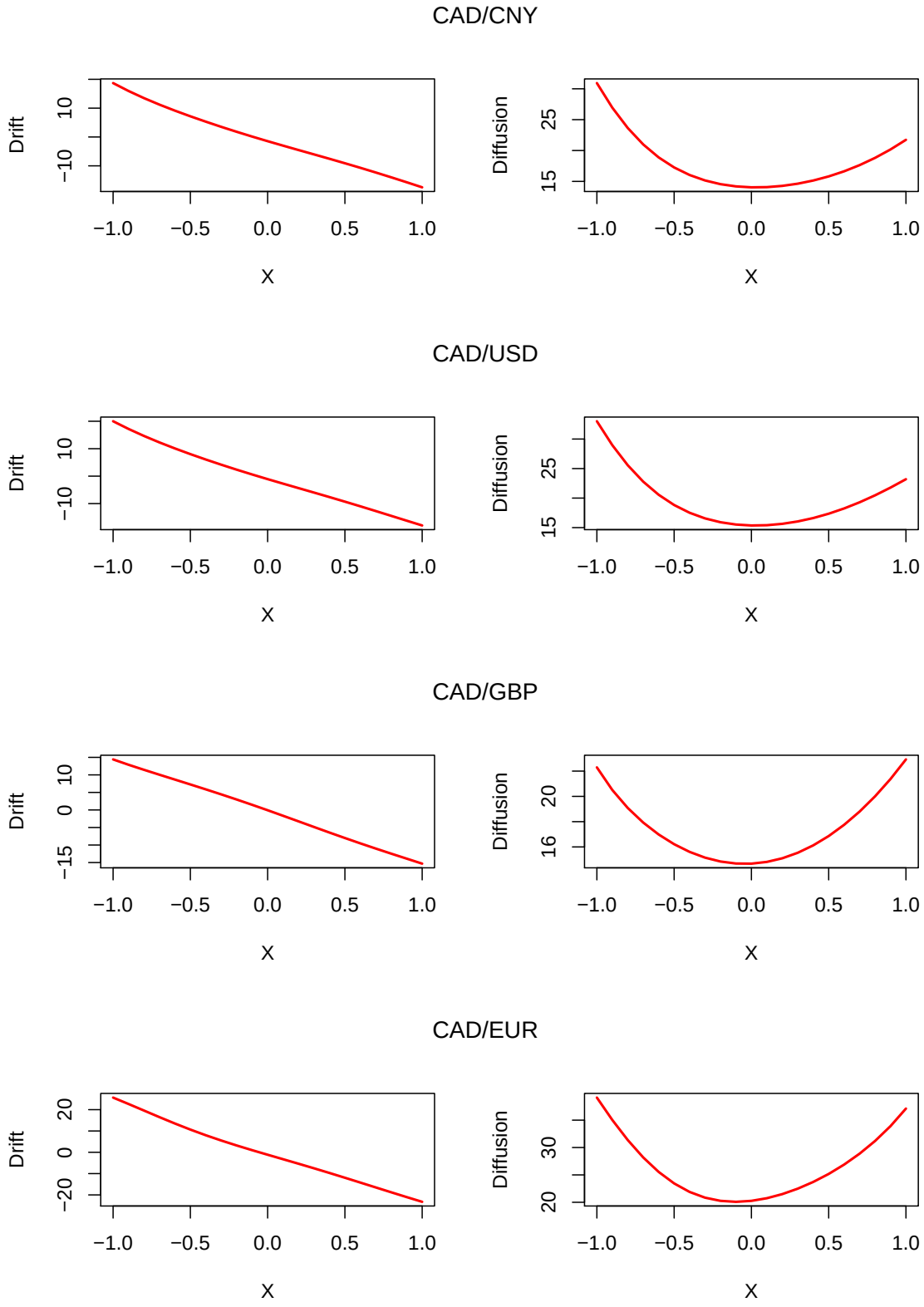

Figure 6.9: Online estimation of the drift and diffusion in the latent process $\{\tilde{X}_i\}$ for the FX rate CAD/CNY, CAD/USD, CAD/GBP and CAD/EUR.

Commodity Price

We also consider the application of the online estimator to crude oil and gold prices. Both price series are obtained from the Federal Reserve Bank of St. Louis, where the daily data of the crude oil price is from Jan 12, 2005 to Jan 12, 2015 and the daily data of the gold price is from Jan 17, 2005 to Jan 16, 2015. The unit is USD/barrel and USD/Troy Ounce, and the price and the proxy of the latent process can be seen in Figure 6.10. From the figure, it can be found that the decline of the crude oil price coincides with that of the FX rate CAD/CNY and CAD/USD. And since 2012, the gold price has decreased. It is for this reason that the Canadian dollar is often called a "commodity currency".

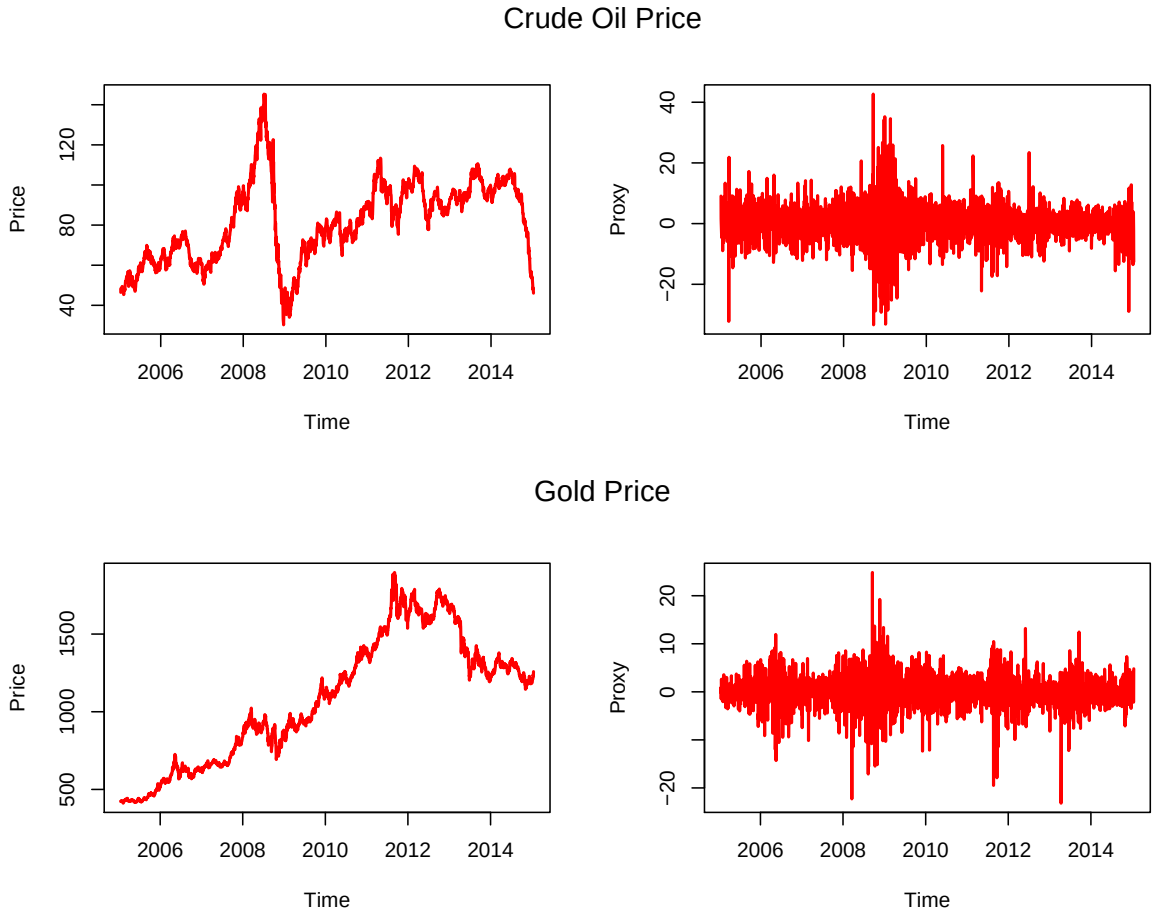

Figure 6.10: The time series of the crude oil prices and gold prices as well as their proxies.

After that, we apply the online estimator with the same settings as above to estimate the drift and diffusion in the latent process by the proxy (6.9). Figure 6.11 shows the estimation results. The drift is estimated to be decreasing for both the crude oil and gold price, but they have different shapes of the diffusion coefficient. For the crude oil price, the diffusion of the latent process is monotonically decreasing in the domain, whereas for the gold price, the diffusion has the minimum between 0.0 and 0.5.

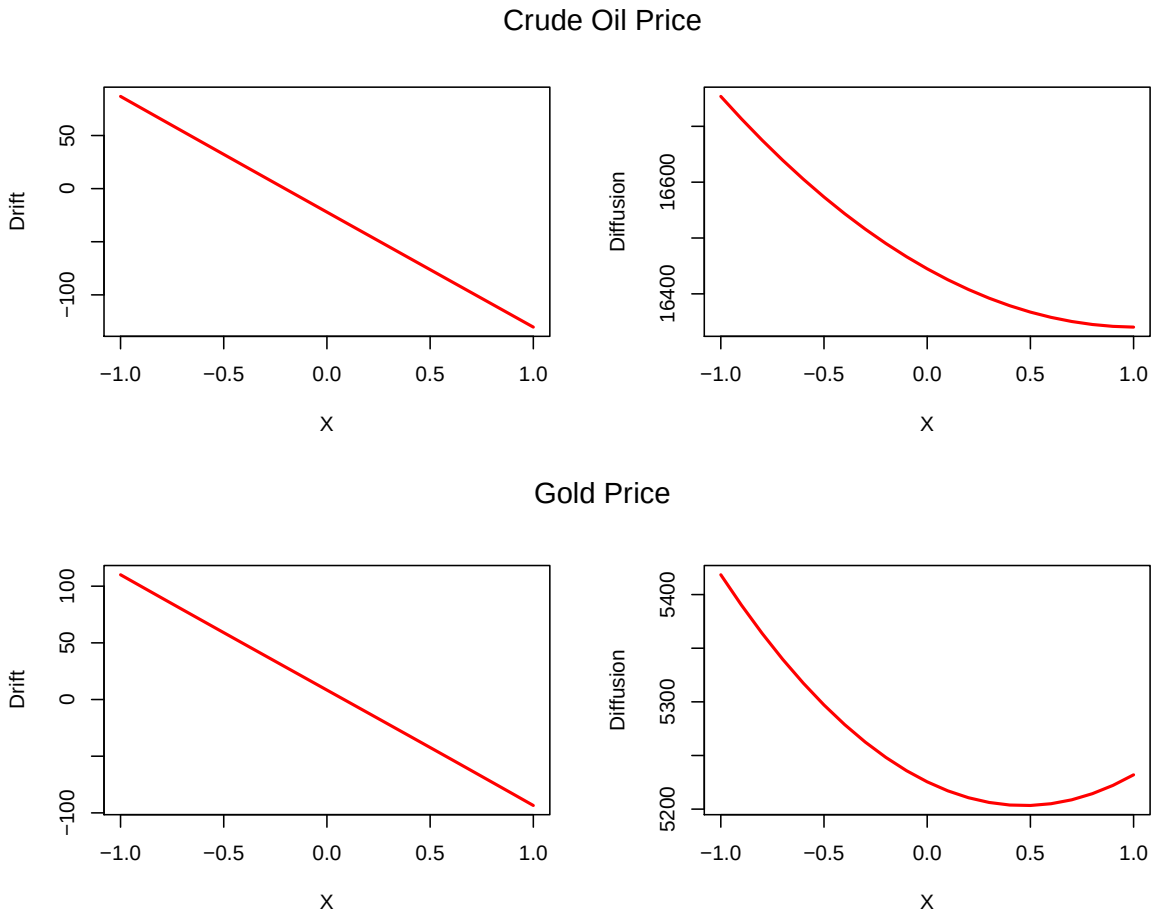

Figure 6.11: Online estimation of the drift and diffusion in the latent process $\{\tilde{X}_i\}$ for the crude oil prices and gold prices.

6.4 Concluding Remark

In this chapter, we use a second-order stochastic differential equation to model the non-stationary diffusion process. By proposing the online estimators of the drift and diffusion in the latent process, we prove weak consistency of the estimators under a series of assumptions. Then numerical examples are used to evaluate our method. We also apply the online technique to estimate the drift and diffusion for the stock index, FX rate and commodity price.

Chapter 7

Conclusion and Future Work

With the development of modern techniques, the need to deal with big data has become more urgent in business and finance. In this thesis, we proposed an online procedure for a large number of sequential observations. Because many risk factors can be described by stochastic diffusion models, we applied our online kernel method to estimate the drift and diffusion of the models. It was found that they outperform the traditional methods in computational speed but also in memory requirements. Therefore, the online kernel method is applicable to financial practices.

7.1 Contributions

There are three main contributions in this thesis as follows:

1. *Online kernel method for stochastic diffusion models*

Stochastic diffusion models have been widely used to model financial systems and estimation of the drift and diffusion in models plays an important role in their applications in finance. By the infinitesimal generator, the traditional kernel method can be used to fulfill this task. However, the traditional method includes all past history in the procedure, which means that if new observations are available, all past information is used in each calculation. Therefore, we proposed the online methods to tackle this problem and used them to estimate the drift and diffusion in stochastic diffusion models.

For the stationary case, we derived the estimators based on the infinitesimal generator directly. For the non-stationary case, we adopted Nicolau (2007)'s method to model the process by a second-order stochastic differential equation, and used the proxy of the latent process to obtain the estimators.

The main merit of our method over the traditional one is the computational speed. By numerical examples, it is found that the online method is at least 17 times faster. Thus the method is more practical for the financial use in the era of big data.

2. *Asymptotic properties of estimators in the stationary and non-stationary case*

Previous theoretical work concentrated on asymptotic properties of either the offline estimator or the online estimator in other application scenarios. This thesis bridged the gap by providing large-sample properties of the online estimators of the drift and diffusion in stochastic differential equations, especially in stochastic diffusion models.

For the stationary process, we proved quadratic convergence, strong consistency and asymptotic normality of the estimators. In the proof, we assumed the stationary process satisfies some mixing conditions so that we can quantify the autocovariances in the terms. By borrowing Masry (1986)'s block technique, we illustrated the asymptotic properties of the terms on "small blocks" and "big blocks" so that asymptotic normality can be proved. It is found that with similar assumptions, the online method can reach the same convergence rate as the offline counterpart.

For the non-stationary process, weak consistency of the estimators has been proved under stronger assumptions. Because the latent stationary process is assumed to be non-observable, we needed to estimate the drift and diffusion by the proxy of the latent process. Thus the proof is not straightforward. With more efforts and the results on the stationary process we have obtained, weak consistency of the estimators in the second-order stochastic diffusion model is established.

3. *Practical applications of the online method in market risk management*

The purpose of this thesis is not only to serve as an analytic method in statistics, but also to provide a professional tool in market risk management. To this end, we follow the most recent document in 2013 by the Basel Committee about the suggestion of replacing VaR by ES and adding the liquidity horizon for different risk factors.

For the stationary case, we applied our method to calculate VaR and ES for the short-term interest rates and compare with historical simulation and the parametric Monte Carlo method. For the non-stationary case, the online estimator was used for the stock index, exchange rate and commodity price.

In addition, this thesis derived the confidence band of VaR and ES. In fact, to our knowledge, little attention has previously been paid to construction of their confidence bands. Based on their empirical definitions and previous results on order statistics, we obtained the confidence bands of VaR and ES and applied them to characterize the quality of the estimators.

7.2 Future Work

Due to the advantages of the online estimator, the following work could be investigated in the future:

1. Time-Inhomogeneous Diffusion Models

The time-homogeneous model cannot characterize the time variation in the model because it has been found that the market condition is time-related (Cheng and Wang, 2007). Many parametric models have been proposed to include the time inhomogeneity (Ho and Lee, 1986; Hull and White, 1990; Black et al., 1990; Black and Karasinski, 1991).

We can generalize our method to the time-inhomogeneous diffusion models. Fan et al. (2003) gave a special class of time-inhomogeneous diffusion models which can include most parametric diffusion models. Therefore we can start from this special class and then extend to the general case.

The estimator for the time-inhomogeneous diffusion models could require a larger sample size. This is because the time-related part of the drift and diffusion needs to be identified as well, which is different from the time-homogeneous case. Therefore, stronger restriction could be imposed on their applications in practice.

2. Jump-Diffusion Models

Recently many efforts have been made on the jump-diffusion models because it is believed that unpredicted events can lead to discontinuities and jumps of the path. Such events could include panic in the market or other sudden mass incidents, and the announcement of monetary policies. For example, the collapse of Bretton Woods system in 1971 led to a significant impact on the global market.

We can propose the online estimator for stochastic differential equations with the jump component which accounts for extreme events. The similar idea can be applied to the second-order jump-diffusion models for the non-stationary case. Additionally, online estimation for the time-inhomogeneous jump-diffusion models could be listed in the further work.

In fact, the regime switching models in time series have been proposed to describe the abrupt changes in the market (Lindgren, 1978; Dueker, 1997). Similar ideas could be borrowed to model the drastic changes by the stochastic differential equations and propose the online procedure to mine the underlying process.

Bibliography

K.M. Abadir and S. Lawford. Optimal asymmetric kernels. *Economics Letters*, 83:61–68, 2004.

Y. Aït-Sahalia. Nonparametric pricing of interest rate derivative securities. *Econometrica*, 64: 527–560, 1996.

Y. Aït-Sahalia. Transition densities for interest rate and other nonlinear diffusions. *Journal of Finance*, 54:1361–1395, 1999.

E. Allen. *Modeling with Itô stochastic differential equations*. Springer, 2007.

A. Amiri, C. Cambes, and B. Thiam. Recursive estimation of nonparametric regression with functional covariate. *Computational Statistics & Data Analysis*, 69:154–172, 2014.

T.G. Andersen and J. Lund. Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics*, 77:343–377, 1997.

D.W.K. Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21 (4):930–934, 1984.

A. Ang and G. Bekaert. Short rate nonlinearities and regime switches. *Journal of Economic Dynamic and Control*, 26:1243–1274, 2002.

M. Arapis and J.T. Gao. Empirical comparisons in short-term interest rate models using non-parametric methods. *Journal of Financial Economics*, 4 (2):310–345, 2006.

M. Arfi. Nonparametric variance estimation from ergodic samples. *Scandinavian Journal of Statistics*, 25 (1):225–234, 1998.

M. Arfi. Drift estimation from $\tilde{\rho}$-mixing sequences. *International Journal of Open Problems in Computer Science and Mathematics*, 1 (1):80–93, 2008.

K.B. Athreya and S.G. Pantula. A note on strong mixing of ARMA processes. *Statistics Probability Letters*, 4:187–190, 1986.

F. Bandi and T.H. Nguyen. On the functional estimation of jump-diffusion models. *Journal of Econometrics*, 116:293–328, 2003.

F.M. Bandi and P.C.B. Phillips. Fully nonparametric estimation of scalar diffusion models. *Econometrica*, 71 (1):241–283, 2003.

G. Barone-Adesi, K. Giannopoulos, and L. Vosper. VaR with correlations for portfolios of derivatives securities. *Journal of Futures Markets*, 19:583–602, 1999.

P. Billingsley. *Probability and measure*. Wiley, 1995.

F. Black and P. Karasinski. Bond and option pricing when short rates are log-normal. *Financial Analysis Journal*, 47:52–59, 1991.

F. Black, E. Derman, and W. Toy. A one-factor model of interest rates and its application to treasury bond options. *Financial Analysis Journal*, 46:33–39, 1990.

T. Bollerslev and H. Zhou. Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics*, 109:33–65, 2002.

D. Bosq. *Nonparametric statistics for stochastic processes*. Spring, 1998.

J. Boudoukh, M. Richardson, and R. Whitelaw. The best of both worlds. *Risk*, 11:64–67, 1998.

A.W. Bowman. An alternative method of cross-validation for the smooth of density estimates. *Biometrika*, 71:353–360, 1984.

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: Forecasting and control*. Prentice Hall, 1994.

R. Bradley. A central limit theorem for stationary ρ-mixing sequences with infinite variance. *The Annals of Statistics*, 16:313–332, 1988.

R.C. Bradley. Approximation theorems for strongly mixing random variables. *The Michigan Mathematical Journal*, 30 (1):69–81, 1983a.

R.C. Bradley. On the ψ-mixing condition for stationary random sequences. *Transactions of the American Mathematical Society*, 276 (1):55–66, 1983b.

R.C. Bradley. Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, 1986.

Z.W. Cai. Weighted Nadaraya-Watson regression estimation. *Statistics and Probability Letters*, 51:307–318, 2001.

Z.W. Cai. Regression quantiles for time series. *Econometric Theory*, 18:169–192, 2002.

Z.W. Cai and Y.M. Hong. *Some recent developments in nonparametric finance*. Emerald Group Publishing Limited, 2009.

J.Y. Campbell, A.W. Lo, and A.C. MacKinlay. *The econometrics of financial markets*. Princeton University Press, 1988.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, 2001.

K. Chan, F. Karolyi, F. Longstaff, and A. Sanders. An empirical comparison of alternative model of the short-term interest rate. *Journal of Finance*, 47:1209–1227, 1992.

D. Chapman and N. Pearson. Is the short rate drift actually nonlinear? *Journal of Finance*, 55: 355–388, 2000.

D. Chapman, J. Long, and N. Pearson. Using proxies for the short rate: When are three months like an instant. *Review of Financial Studies*, 12:763–807, 1999.

H.F. Chen. *Stochastic approximation and its applications*. Springer, 2002a.

H.F. Chen and L. Guo. *Identification and stochastic adaptive control*. Springer, 1991.

S.X. Chen. Local linear smoothers using asymmetric kernels. *Annals of the Institute of Statistical Mathematics*, 54 (2):312–323, 2002b.

X.M. Chen and C. Gao. Recursive local linear regression estimation and its applications. *Control Theory & Applications*, 30 (4):482–491, 2013.

P. Cheng and J.D. Wang. Wavelet estimation of the diffusion coefficient in time dependent diffusion models. *Science in China Series A: Mathematics*, 50 (11):1597–1610, 2007.

M.R. Chernick. A limit theorem for the maximum of autoregressive processes with uniform marginal distributions. *The Annals of Probability*, 9 (1):145–149, 1981.

M. Chesney, R.J. Elliott, D. Madan, and H.L. Yang. Diffusion coefficient estimation and asset pricing when risk period and sensitivities are time varying. *Mathematical Finance*, 3 (2): 85–99, 1993.

G. Collomb and W. Härdle. Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and their Applications*, 23 (1):77–89, 1986.

T.G. Conley, L.P. Hansen, E.G.J. Luttmer, and J.A. Scheinkman. Short-term interest rates as subordinated diffusions. *Review of Financial Studies*, 10:525–577, 1997.

J. Cox, E. Ingersoll, and S. Ross. An intertemporal general equilibrium model of asset price. *Econometrica*, 53:363–384, 1985.

J. Detemple, R. Garcia, and M. Rindisbacher. Asymptotic properties of Monte Carlo estimators of diffusion processes. *Journal of Econometrics*, 134 (1):1–68, 2006.

J.L. Doob. *Stochastic Processes*. Wiley, 1953.

P. Doukhan. *Mixing: Properties and examples*. Springer, 1994.

M. Dueker. Markov switching in GARCH processes and mean-reverting stock-market volatility. *Journal of Business and Economic Statistics*, 15:26–34, 1997.

R.C. Elandt-Johnson and N.L. Johnson. *Survival models and data analysis*. Wiley, 1980.

J. Fan and I. Gijbels. *Local polynomial model and its applications*. Chapman & Hall, 1996.

J.F. Fan and Q.W. Yao. *Nonlinear time series: Nonparametric and parametric methods*. Springer, 2013.

J.Q. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of S*, 21(1):196–216, 1993.

J.Q. Fan and C.M. Zhang. A reexamination of diffusion estimators with applications to financial model validation. *Journal of the American Statistical Association*, 98 (461):118–134, 2003.

J.Q. Fan, J.C. Jiang, C.M. Zhang, and Z.W. Zhou. Time-dependent diffusion models for term structure dynamics. *Statistica Sinica*, 13:965–992, 2003.

D. Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability*, 30 (4):790–804, 1993.

C. Francq and J.M. Zakoian. *GARCH models: Structure, statistical inference and financial applications*. Wiley, 2010.

B.M. Friedman. Lessons from the 1979-1982 monetary policy experiment. *American Economic Review*, 74:382–387, 1984.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

P. Glasserman. *Monte Carlo methods in financial engineering*. Springer, 2003.

E. Gobet, M. Hoffmann, and M. Reiß. Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32 (5):2223–2253, 2004.

N. Gospodinov and M. Hirukawa. Nonparametric estimation of scalar diffusion models of interest rates using asymmetric kernels. *Journal of Empirical Finance*, 19:595–609, 2012.

H.L. Gray and W.R. Schucany. *The generalized Jackknife statistic*. Dekker, 1972.

P. Hall and T.E. Wehrly. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86: 665–672, 1991.

P. Hall, R.C.L. Wolff, and Q.W. Yao. Method for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94 (445):154–163, 1999.

P. Hall, L. Peng, and Q.W. Yao. Prediction and nonparametric estimation for time series with heavy tails. *Journal of Time Series Analysis*, 23 (3):251–275, 2002.

M. Hanif. Local linear estimation of jump-diffusion models by using asymmetric kernels. *Stochastic Analysis and Applications*, 31 (6):956–974, 2013.

B.E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748, 2008.

L.P. Hansen and T.J. Sargent. The dimensionality of the aliasing problem. *Econometrica*, 51: 377–388, 1983.

W. Härdle. *Applied nonparametric regression*. Cambridge University Press, 1990.

T.S.Y. Ho and S.B. Lee. Term structure movements and pricing interest rate contingent claims. *Journal of Finance*, 41:1011–1029, 1986.

Y.X. Huang. *Online inference for time series and series estimation under dependence*. PhD thesis, The University of Chicago, 2011.

Y.X. Huang, X.H. Chen, and W.B. Wu. Recursive nonparametric estimation for time series. *IEEE Transactions on Information Theory*, 60 (2):1301–1312, 2014.

J. Hull. *Risk management and financial institutions*. Wiley, 2012.

J. Hull and H. White. Pricing interest-rate derivative securities. *Review of Financial Studies*, 3:573–592, 1990.

I.A. Ibragimov and Y.V. Linnik. *Independent and stationary sequences of random variables*. Wolters-Noordhoff, 1971.

J. Jacod and P. Protter. Asymptotic error distributions for the Euler method for stochastic differential equations. *Annals of Probability*, 26 (1):267–307, 1998.

C.M. Jarque and A.K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6 (3):255–259, 1980.

G. Jiang and J. Knight. Parametric versus nonparametric estimation of diffusion processes: A Monte Carlo comparison. *Journal of Computational Finance*, 2:5–38, 1999.

G.J. Jiang and J. Knight. A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory*, 13:615–645, 1997.

M.S. Johannes. The economic and statistical role of jumps to interest rates. *Journal of Finance*, 59:227–260, 2004.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23 (3):462–466, 1952.

H. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

Q. Li and J.S. Racine. *Nonparametric econometrics: Theory and practice*. Princeton University Press, 2006.

G. Lindgren. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 5:81–91, 1978.

J. Liu, F.A. Longstaff, and J. Pan. Dynamic asset allocation with event risk. *Journal of Finance*, 58:231–259, 2002.

G.M. Ljung and G.E.P. Box. On a measure of lack of fit in time series models. *Biometrica*, 65: 297–303, 1978.

B.J. Lobo. Jump risk in the U.S. stock market: Evidence using political information. *Review of Financial Economics*, 8:149–163, 1999.

M. Loève. *Probability theory*. Springer, 1977.

M. Mackenzie and A.K. Tieu. Asymmetric kernel regression. *IEEE Transactions on Neural Networks*, 5 (2):276–282, 2004.

E. Masry. Recursive probability density estimation for weakly dependent stationary processes. *IEEE Transactions on Information Theory*, 32 (2):254–267, 1986.

E. Masry. Almost sure convergence of recursive density estimators for stationary mixing processes. *Statistics and Probability Letters*, 5:249–254, 1987.

E. Masry. Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications*, 65:81–101, 1996.

E. Masry and J.Q. Fan. Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, 24 (2):165–179, 1997.

E. Masry and L. Gyorfi. Strong consistency and rates for recursive probability density estimators of stationary processes. *Journal of Multivariate Analysis*, 22:79–93, 1987.

D.L. McLeish. A maximal inequality and dependent strong saw. *Annals of Probability*, 3: 829–839, 1975.

A. Mehta, M. Neukirchen, S. Pfetsch, and T. Poppensieker. Managing market risk: Today and tomorrow. Technical report, McKinsey Working Papers on Risk, Number 32, 2012.

J.M. Mendel and K.S. Fu, editors. *Adaptive, learning, and pattern recognition systems: Theory and applications*. Academic Press, 1970.

P. Michels. Asymmetric kernel function in nonparametric regression analysis and prediction. *The Statistician*, 41:439–454, 1992.

F. Mosteller. On some useful "inefficient" statistics. *The Annals of Mathematical Statistics*, 17 (4):377–408, 1946.

H.G. Müller and U. Stadtmüller. Variable bandwidth kernel estimators of regression curves. *Annals of Statistics*, 15:182–201, 1987.

E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9 (1): 141–142, 1964.

M.H. Ngerng. Recursive nonparametric estimation of local first derivative under dependence conditions. *Communications in Statistics*, 40:1159–1168, 2011.

J. Nicolau. Bias reduction in nonparametric diffusion coefficient estimation. *Econometric Theory*, 19 (5):754–777, 2003.

J. Nicolau. Nonparametric estimation of second-order stochastic differential equations. *Econometric Theory*, 23 (5):880–898, 2007.

J. Nicolau. Modeling financial time series through second order stochastic differential equations. *Statistics and Probability Letters*, 78 (16):2700–2704, 2008.

B. Øksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, 2003.

M.F.M. Osborne. Brownian motion in the stock market. *Operations Research*, 7 (2):145–173, 1959.

O. Papaspiliopoulos, Y. Pokern, G.O. Roberts, and A.M. Stuart. Nonparametric estimation of diffusions: A differential equations approach. *Biometrika*, 99 (3):511–531, 2012.

L. Peng and Q.W. Yao. Nonparametric regression under dependent errors with infinite variance. *Annals of the Institute of Statistical Mathematics*, 56:73–86, 2004.

P.C.B. Phillips. The problem of identification in finite parameter continuous-time model. *Journal of Econometrics*, 1:351–362, 1973.

B. Rao. *Statistical inference for diffusion type processes*. Oxford University Press, 1999.

P.B.L.S. Rao. Estimation of the drift for diffusion processes. *Statistics*, 16:263–275, 1985.

R. Révész. How to apply the method of stochastic approximation in the nonparametric estimation of a regression function. *Series Statistics*, 8 (1):119–126, 1977.

J.A. Rice. Boundary modification for kernel regression. *Communications in Statistics*, 13: 839–900, 1984.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22 (3):400–407, 1951.

M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42:43–47, 1956.

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

S.E. Said and D.A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrica*, 71:599–607, 1984.

A. Schied. Risk mmeasure and robust optimization problems. *Stochastic Models*, 22 (4): 753–831, 2006.

E. Schmisser. Nonparametric adaptive estimation of the drift for a jump-diffusion process. *Stochastic Processes and Their Applications*, 124 (1):883–914, 2014.

W. Schucany. Adaptive bandwidth choice for kernel regression. *Journal of the American Statistical Association*, 90 (430):535–540, 1995.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88 (422):486–494, 1993.

A.N. Shiryaev. *Probability*. Springer, 1996.

B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall, 1986.

K. Sobczyk. *Stochastic differential equations: With applications to physics and engineering*. Springer, 2001.

V. Solo and X. Kong. *Adaptive signal processing algorithms: Stability and performance*. Prentice Hall, 1994.

Y.P. Song and Z.Y. Lin. Empirical likelihood inference for the second-order jump-diffusion model. *Statistics and Probability Letters*, 83:184–195, 2013.

Y.P. Song, Z.Y. Lin, and H.C. Wang. Re-weighted Nadaraya-Watson estimation of second-order jump-diffusion model. *Journal of Statistical Planning and Inference*, 143 (4):730–744, 2013.

P. Soulier. Nonparametric estimation of the diffusion coefficient of a diffusion process. *Stochastic Analysis and Applications*, 16 (1):185–200, 1998.

V.G. Spokoiny. Adaptive drift estimation for nonparametric diffusion model. *Annals of Statistics*, 28 (3):815–836, 2000.

S. Stanton. A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance*, 52 (5):1973C2002, 1997.

S.M. Stigler. The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1 (3): 472–477, 1973.

C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, 12:1285–1297, 1984.

L.T. Tran. Recursive density estimation under dependence. *IEEE Transactions on Information Theory*, 35 (5):1103–1108, 1989.

R.S. Tsay. *Analysis of financial time series*. Wiley, 2005.

P.D. Tuan. Nonparametric estimation of the drift coefficient in the diffusion equation. *Statistics: A Journal of Theoretical and Applied Statistics*, 12:61–73, 1981.

O. Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5:177–188, 1977.

J.M. Vilar and J.A. Vilar. Recursive local polynomial regression under dependence conditions. *TEST*, 9 (1):209–232, 2000.

V.A. Volkonskii and Y.A. Rozanov. Some limit theorems for random functions I. *Theory of Probability and Its Applications*, 4:178–197, 1959.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

M. Wand and M. Jones. *Kernel smoothing*. Chapman & Hall, 1995.

H.C. Wang and Z.Y. Lin. Local linear estimation of second-order diffusion models. *Communications in Statistics: Theory and Methods*, 40 (3):394–407, 2011.

J.X. Wang and Q.X. Xiao. Local linear estimations of time-varying parameters for time-inhomogeneous diffusion models. *Chinese Journal of Applied Probability and Statistics*, 29 (4):392–404, 2013.

Y.Y. Wang, L.X. Zhang, and M.T. Tang. Re-weighted functional estimation of second-order diffusion processes. *Metrika*, 75:1129–1151, 2012.

G.S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 24 (6):359–372, 1964.

R.L. Wheeden and A. Zygmund. *Measure and integral: An introduction to real analysis*. Marcel Dekker, 1977.

W.B. Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (40):14150–14151, 2005.

Z. Xiao, O.B. Linton, R.J. Carroll, and E. Mammen. More efficient local polynomial estimation in nonparametric regression with autocorrelation errors. *Journal of the American Statistical Association*, 98:980–992, 2003.

H.Y. Xu. *Directional kernel regression and point process modeling in wildfire hazard assessment*. PhD thesis, University of California, Los Angeles, 2008.

K.L. Xu. Reweighted function estimation of diffusion models. *Econometric Theory*, 26:541–563, 2010.

K. Yu and M. Jones. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99:139–144, 2004.

Y. Yu, K.M. Yu, H. Wang, and M. Li. Semiparametric estimation for a class of time-inhomogeneous diffusion processes. *Statistica Sinica*, 19:843–867, 2009.

P. Zhang. Model selection via multifold cross validation. *Annals of Statistics*, 21 (1):299–313, 1993.

F.A. Ziegelmann. Nonparametric estimation of volatility functions: The local exponential estimator. *Econometric Theory*, 18:985–991, 2002.

Appendix A

Stochastic differential Equations

In finance, continuous-time models are often used to describe the stochastic behaviors of underlying variables such as the stock price, interest rate and exchange rate. This thesis concentrates on the Itô diffusion process $\{X_t\}$, satisfying the stochastic differential equation of the form[1]

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t \tag{A.1}$$

with the initial value X_0, the corresponding integral form is:

$$X_t = X_0 + \int_0^t a(s, X_s)ds + \int_0^t b(s, X_s)dW_s \tag{A.2}$$

Here $\{W_t, t \geq 0\}$ is a standard Brownian motion or a Wiener process, defined as:

(1) $W_0 = 0$ almost surely;

(2) W_t is continuous almost surely;

(3) The process has independent increments, that is, for any $0 = t_0 \leq t_1 < t_2 < \ldots < t_n = T$, the increments $W_{t_1}, W_{t_2} - W_{t_1}, W_{t_3} - W_{t_2}, \ldots, W_{t_n} - W_{t_{n-1}}$ are independent;

(4) $W_t - W_s \sim N(0, t - s)$.

The famous examples of (A.1) in finance include GBM for the stock price

$$dX_t = \mu X_t dt + \sigma X_t dW_t$$

The Vasicek's model, CIR model and CKLS model for the interest rate

$$\begin{aligned} \text{(VAS)} \quad dX_t &= a(b - X_t)dt + cdW_t \\ \text{(CIR)} \quad dX_t &= a(b - X_t)dt + c\sqrt{X_t}dW_t \\ \text{(CKLS)} \quad dX_t &= a(b - X_t)dt + cX_t^{\gamma}dW_t \end{aligned}$$

[1] In this thesis, we use capital letters X_t for the random variable and the corresponding lowercase letter x_t for its realization.

As is well known, the Brownian motion $\{W_t, t \geq 0\}$ has unbounded variation and is nowhere differentiable, so Riemann-Stieltjes integral cannot be applied to the third term of the right-hand side in (A.2), i.e. the integral with respect to the Brownian motion. Thus Itô gave his definition for the stochastic integral $\int_0^T f(X_t)dW_t$ by approximating sums of the form

$$\sum_{i=1}^{n} f(X_{t_{i-1}})[W_{t_i} - W_{t_{i-1}}]$$

where $0 = t_0 < t_1 < \ldots < t_n = T$ is a partition that becomes finer and finer as $n \to \infty$. In Riemann-Stieltjes integral, we know that the chain rule allows us to calculate the integral without referring to its original definition. For example, suppose $f(t)$ is monotonically increasing and continuous on $[a, b]$, then $\int_a^b f(t)df(t) = f^2(t)\big|_a^b - \int_a^b f(t)df(t)$ implying $\int_a^b f(t)df(t) = [f^2(b) - f^2(a)]/2$. Comparatively, the Itô lemma plays the same role in Itô integral, which is given by

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + a(t, X_t)\frac{\partial f}{\partial x} + \frac{1}{2}b^2(t, X_t)\frac{\partial^2 f}{\partial x^2}\right)dt + b(t, X_t)\frac{\partial f}{\partial x}dW_t \tag{A.3}$$

where $f(t, x)$ is a bivariate function such that $\frac{\partial f}{\partial t}, \frac{\partial f}{\partial x}$ and $\frac{\partial^2 f}{\partial x^2}$ are continuous. For example, consider GBM (see Figure A.1), let $f(x) = \ln x$, then

$$\frac{\partial f}{\partial t} = 0 \quad \frac{\partial f}{\partial x} = \frac{1}{x} \quad \frac{\partial^2 f}{\partial x^2} = -\frac{1}{x^2}$$

Thus based on the Itô lemma (A.3)

$$df(X_t) = (\mu - \sigma^2/2)dt + \sigma dW_t$$

the solution to GBM is

$$X_t = X_0 \exp\left((\mu - \sigma^2/2)t + \sigma W_t\right) \tag{A.4}$$

To ensure the existence and uniqueness of the solution satisfying (A.2), in almost surely sense the following conditions for $a(t, X_t)$ and $b(t, X_t)$ are always assumed, where $t \in [0, T]$ (see (Øksendal, 2003; Jiang and Knight, 1997)):

(1) $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are continuously differentiable and Borel measurable functions, satisfying

$$\int_0^T |a(t, X_t)|dt < \infty \quad \text{and} \quad \int_0^T |b(t, X_t)|dt < \infty$$

(2) $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ satisfy the linear growth condition, that is, $\exists C > 0$ such that

$$|a(t, x)| \leq C(1 + x^2)^{1/2} \quad \text{and} \quad b(t, x) \leq C(1 + x^2)^{1/2}$$

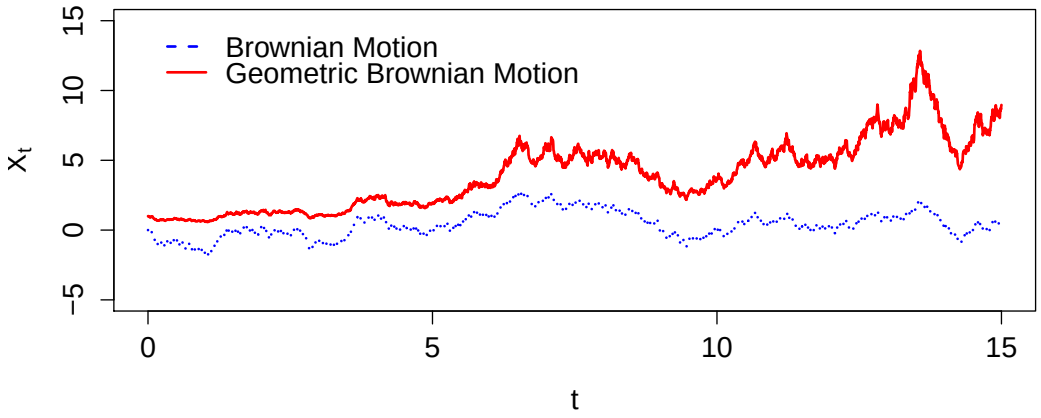

Figure A.1: Trajectories of GBM and Brownian motion, where GBM is $dX_t = 0.2X_t dt + 0.375X_t dW_t$.

(3) $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the uniform Lipschitz condition, that is, for any $x, y \in R$ and some constant $D > 0$ such that

$$|a(t,x) - a(t,y)| \le D|x-y| \quad \text{and} \quad |b(t,x) - b(t,y)| \le D|x-y|$$

If another condition is satisfied for the time homogeneous case $dX_t = a(X_t)dt + b(X_t)dW_t$, i.e.

(4) The transition probability $P(X_{t+h}|X_t)$ satisfies the Chapman-Kolmogorov equation

$$P(X_{t+h}|X_t) = \int_{X_\tau} P(X_{t+h}|X_\tau)P(dX_\tau|X_t) \qquad t < \tau < t+h$$

and $\lim_{h\to 0} \frac{1}{h} P(|X_{t+h} - X_t| \ge \epsilon | X_t(\omega) = X_t) = 0$ a.s. for every $\epsilon > 0$ and every path, then the following Kolmogorov forward equation (A.5) and backward equation (A.6) for the transition probability density function $p(X_t = x|X_{t_0} = x_0)$ hold true.

$$-\frac{\partial p(X_t = x|X_{t_0} = x_0)}{\partial t} = \frac{\partial}{\partial x}[a(x)p(X_t = x|X_{t_0} = x_0)] - \frac{\partial^2[b^2(x)p(X_t = x|X_{t_0} = x_0)]}{2\partial x^2} \quad \text{(A.5)}$$

$$-\frac{\partial p(X_t = x|X_{t_0} = x_0)}{\partial t} = a(x_0)\frac{\partial p(X_t = x|X_{t_0} = x_0)}{\partial x} + \frac{1}{2}b^2(x_0)\frac{\partial^2 p(X_t = x|X_{t_0} = x_0)}{\partial x^2} \quad \text{(A.6)}$$

Usually, it is hard to find the closed form for a solution of (A.2), so numerical methods for stochastic differential equations are essential. In practice, there are two popular methods:

(1) *Euler scheme*

Consider a partition of $[0,T]$, $0 = t_0 < t_1 < \cdots < t_n = T$, and denote $\triangle_i = t_{i+1} - t_i$ and $\triangle_i W = W_{t_{i+1}} - W_{t_i}$. The idea of the Euler scheme is that X_t on $t \in (t_{i-1}, t_i)$ can be obtained by simply linear interpolation of the data $(t_{i-1}, X_{t_{i-1}})$ and (t_i, X_{t_i}). Here for convenience, denote X_i

for X_{t_i}, then the Euler scheme to approximate the solution of (A.1) is given by

$$X_{i+1} = X_i + a(t_i, X_i)\triangle_i + b(t_i, X_i)\triangle_i W \tag{A.7}$$

Based on the definition of the Brownian motion, $\triangle_i W = W_{t_{i+1}} - W_{t_i} \sim N(0, \triangle_i)$, let $\varepsilon_i \sim N(0, 1)$, then $\triangle_i W \overset{d}{=} \sqrt{\triangle_i}\epsilon_i$ where "$\overset{d}{=}$" means equivalence in distribution, so (A.7) can also be written as

$$X_{i+1} = X_i + a(t_i, X_i)\triangle_i + b(t_i, X_i)\sqrt{\triangle_i}\varepsilon_i \tag{A.8}$$

For example, the GBM $dX_t = \mu X_t dt + \sigma X_t dW_t$ has the discretized form given by the Euler scheme

$$X_{i+1} = X_i + \mu X_i \triangle_i + \sigma X_i \sqrt{\triangle_i}\varepsilon_i \tag{A.9}$$

and the CIR model $dX_t = a(b - X_t)dt + c\sqrt{X_t}dW_t$ has the form

$$X_{i+1} = X_i + a(b - X_i)\triangle_i + c\sqrt{X_i}\sqrt{\triangle_i}\varepsilon_i \tag{A.10}$$

(2) *Milstein scheme*

Following the above notation, first we can integrate both sides of (A.1) on the interval (t_{i-1}, t_i)

$$X_i = X_{i-1} + \int_{t_{i-1}}^{t_i} a(s, X_s)ds + \int_{t_{i-1}}^{t_i} b(s, X_s)dW_s \tag{A.11}$$

Let the operators $\mathcal{L} = \frac{\partial}{\partial t} + a\frac{\partial}{\partial x} + \frac{1}{2}b^2\frac{\partial^2}{\partial x^2}$ and $\mathcal{M} = b\frac{\partial}{\partial x}$, and note that $a(s, X_s)$ and $b(s, X_s)$ are functions of s and X_s, so we can apply the Itô lemma and integrate both sides in (A.3) to obtain

$$\begin{aligned} a(s, X_s) &= a(t_{i-1}, X_{i-1}) + \int_{t_{i-1}}^{s} \mathcal{L}adv + \int_{t_{i-1}}^{s} \mathcal{M}adW_v \\ b(s, X_s) &= b(t_{i-1}, X_{i-1}) + \int_{t_{i-1}}^{s} \mathcal{L}bdv + \int_{t_{i-1}}^{s} \mathcal{M}bdW_v \end{aligned}$$

then put the above expressions into (A.11) to get the approximation formula. For the time-homogeneous case, the Milstein scheme is given by

$$X_{i+1} = X_i + a(X_i)\triangle_i + b(X_i)\sqrt{\triangle_i}\varepsilon_i + \frac{1}{2}b(X_i)b'(X_i)\triangle_i(\varepsilon_i^2 - 1) \tag{A.12}$$

For example, the GBM has the discretized form given by the Milstein scheme

$$X_{i+1} = X_i + \mu X_i \triangle_i + \sigma X_i \sqrt{\triangle_i}\varepsilon_i + \frac{1}{2}\sigma^2 X_i \triangle_i(\varepsilon_i^2 - 1)$$

As pointed by Jacod and Protter (1998), when the sample size is n and the discretization step size $\triangle_i = 1/n$, the approximation error rate incurred by the Euler scheme converges weakly

at the rate $1/\sqrt{n}$. But for Milstein scheme, the error rate is $1/n$ (Detemple et al., 2006). This indicates that the error for the Milstein scheme diminishes faster than the Euler scheme, so the Milstein scheme improves the performance of the standard Euler method. However in real applications, the Euler scheme is more favored. This is because the Euler scheme is faster to implement and from (A.8), it can be found that X_i has the conditionally normal distribution. On the contrary, one has to compute the derivative when using the Milstein scheme and X_i in (A.12) has a mixture of the normal and χ^2 distribution. This will result in complexity for further inference.

When we use SDEs in finance, the parametric method is to presume the functional relationships for the drift and diffusion such that (A.1) is written as

$$dX_t = a_\theta(t, X_t)dt + b_\theta(t, X_t)dW_t$$

where θ is the parameter in the model, e.g. GBM has two parameters μ and σ. This is often regarded as a kind of parametric methods as we have described in Chapter 2. Therefore given the presumed form, people need to estimate the parameter θ based on the observations available and then apply the estimated model into practice. There are two main methods to estimate the parameter (Casella and Berger, 2001), the method of moments and the maximum likelihood method. For example, suppose that the stock prices are $S_0, S_1, \ldots, S_n$ for the equispaced sampling interval $\triangle_i = \triangle$ and let the log-return $R_i = \log(S_{i+1}) - \log(S_i)$ for $i = 0, 1, \ldots, n-1$. Then based on (A.4), it is clear that

$$R_i = (\mu - \sigma^2/2)\triangle + \sigma\sqrt{\triangle}\varepsilon_i \sim \mathcal{N}((\mu - \sigma^2/2)\triangle, \sigma^2\triangle)$$

Let $m = n^{-1}\sum_{i=0}^{n-1} R_i$ and $s^2 = (n-1)^{-1}\sum_{i=0}^{n-1}(R_i - m)^2$, then by matching moments we have

$$\begin{aligned} m &= (\mu - \sigma^2/2)\triangle \\ s^2 &= \sigma^2\triangle \end{aligned}$$

so it follows that

$$\hat{\mu} = \frac{m + s^2/2}{\triangle} \quad \text{and} \quad \hat{\sigma} = \frac{s}{\sqrt{\triangle}} \tag{A.13}$$

For the example in Figure A.1, the above procedure gives the result that $\hat{\mu} = 0.22$ and $\hat{\sigma} = 0.38$. Similarly for the stock prices of "Big Five" in Canada, i.e. Royal Bank of Canada (RBC), Toronto-Dominion (TD) Bank, Bank of Montreal (BMO), Scotiabank, Canadian Imperial Bank of Commerce (CIBC) (see Figure A.2), we calculate $\hat{\mu}$ and $\hat{\sigma}$ for these banks if GBM is used to model the stock prices. From Table A.1, the estimates tell us that TD Bank has the biggest $\hat{\mu}$ and smallest $\hat{\sigma}$, which means that the stock of TD Bank has the highest expected return with smallest volatility. In other words, if GBM is plausible to fit the data and the estimates

	$\hat{\mu}$	$\hat{\sigma}$
RBC	0.125	0.370
TD Bank	0.148	0.296
BMO	0.121	0.301
Scotiabank	0.125	0.306
CIBC	0.112	0.325

Table A.1: Estimation of the drift and diffusion for the stock prices of RBC, TD Bank, BMO, Scotiabank and CIBC from Jan 3, 2007 to Nov 6, 2014.

are sufficiently accurate, then the stock of TD Bank is the best choice, that is, having highest return but lowest risk.

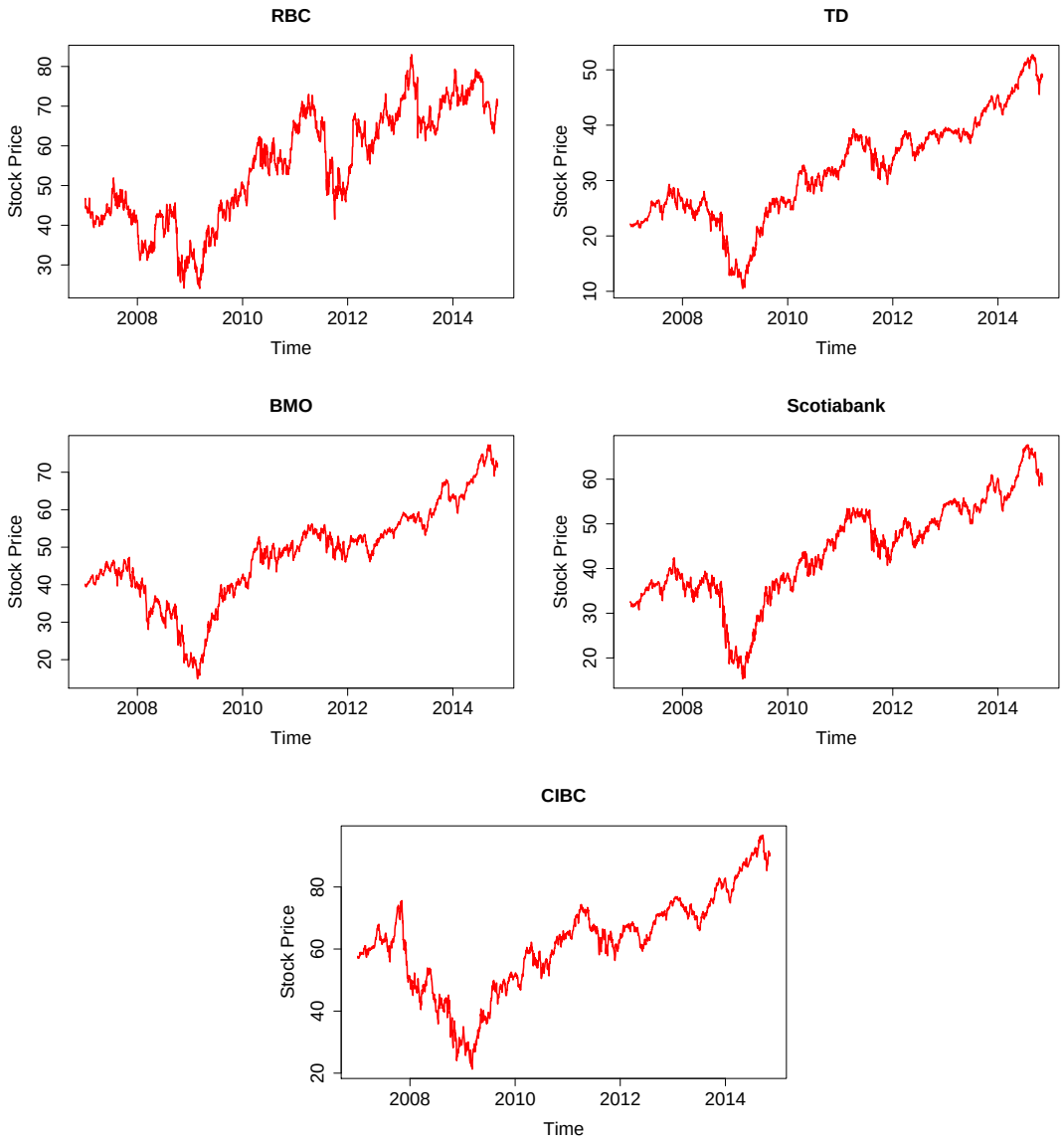

Figure A.2: Stock prices of RBC, TD Bank, BMO, Scotiabank and CIBC where data are from Yahoo! Finance and the time interval is from Jan 3, 2007 to Nov 6, 2014.

Appendix B

Stochastic Approximation

This thesis is partly inspired by the idea of stochastic approximation. Stochastic approximation is a class of recursive methods to find the zero root or the optimum of a function via noisy observations. In last decades, stochastic approximation has been widely applied in many areas, such as signal processing (Solo and Kong, 1994), control theory (Chen and Guo, 1991) and pattern recognition (Mendel and Fu, 1970). Chen (2002a) and Kushner and Yin (2003) provide a detailed and systematic study of stochastic approximation.

In practice, we often come across root-seeking problems such as calculating yield to maturity for fixed income securities or implied volatility for derivatives. If one desires to find the zero root x_0 of some known function $f(x)$ such that $f(x_0) = 0$, then the Newton-Raphson method is an alternative given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{B.1}$$

However, in some situation the form of $f(x)$ is unknown and only noisy observations of $f(x)$ are available via

$$y_n = f(x_n) + \varepsilon_n$$

where ϵ_n's are observational errors. Then the question is how to find the zero root of $f(x)$ via the observations $\{y_n\}$. Robbins and Monro (1951) proposed the recursive procedure (also known as the Robbins-Monro algorithm) to fulfill this work (see in Figure B.1)

$$x_{n+1} = x_n + a_n y_n \tag{B.2}$$

where the step size $a_n > 0$ satisfies $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n^2 < \infty$. For example, one can take $a_n = 1/n$ to satisfy both conditions. Comparing (B.1) and (B.2), it is apparent that the Robbins-Monro algorithm can be regarded as a kind of nonparametric methods which do not leave the form of the function $f(x)$ to be specified. Inspired by Robbins and Monro (1951)'s work, Kiefer and Wolfowitz (1952) proposed a recursive method to find the maximum of a

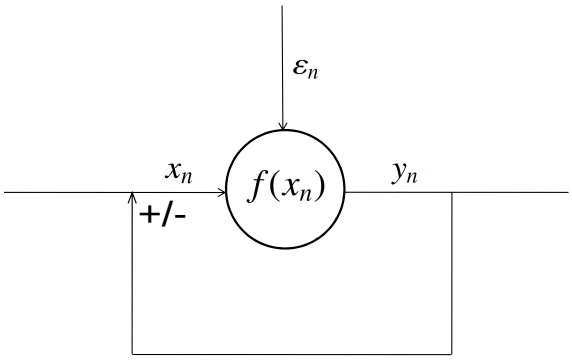

Figure B.1: Demonstration of the Robbins-Monro algorithm.

function. Note that the continuous function $f(x)$ reaches the maximum at the point x_0 when $f'(x) = 0$, so the optimization problem is also equivalent to find the root of $f'(x) = 0$. The authors proposed the recursive procedure to find the maximum of $f(x)$ in the case that there are only observations of $f'(x)$ available.

Further, it can be shown that the problem of root seeking is closely related to nonparametric estimation. In order to estimate $g(x)$ in (2.6), let $r(y) = f(x)y - f(x)g(x)$ where $f(x)$ is the density function of X, then one can estimate $g(x)$ by finding the root y_0 such that $r(y_0) = 0$. Given a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, Révész (1977) proposed a recursive method to estimate $g(x)$ given by

$$g_n(x) = g_{n-1}(x) + \frac{1}{n} K_{h_n}(X_n - x)[Y_n - g_{n-1}(x)] \tag{B.3}$$

where $g_0(x) = 0$. Equation (B.3) could be called a totally recursive procedure because only one formula is used to estimate $g(x)$. Comparing to total recursion, a semi-recursive procedure approximates the numerator and denominator of the Nadaraya-Watson estimator separately, which is given by

$$N_n(x) = N_{n-1}(x) - \frac{1}{n}\left[N_{n-1}(x) - Y_n K_{h_n}(X_n - x)\right] \tag{B.4}$$

$$D_n(x) = D_{n-1}(x) - \frac{1}{n}\left[D_{n-1}(x) - K_{h_n}(X_n - x)\right] \tag{B.5}$$

then

$$g_n(x) = \frac{N_n(x)}{D_n(x)} \tag{B.6}$$

If $X_1, X_2, \ldots, X_n$ are dependent with the specific settings, then Masry (1986) proposed two recursive procedures for density estimation

$$\hat{f}_n(x) = \frac{n-1}{n}\hat{f}_{n-1}(x) + \frac{1}{n} K_{h_n}(x - X_n)$$

and

$$\tilde{f}_n(x) = \frac{n-1}{n}\left(\frac{h_{n-1}}{h_n}\right)^{1/2} \tilde{f}_{n-1}(x) + \frac{1}{n}K_{h_n}(x - X_n).$$

The author presented the quadratic convergence rate and proved asymptotic normality by the method of "big block" and "small block". The results of the almost sure convergence rate are given by Masry and Gyorfi (1987) and the uniform almost sure convergence rate can be found in (Tran, 1989). For a dependent sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, recursive procedures have been used to estimate the regression function. Vilar and Vilar (2000) considered semi-recursive estimation for dependent data. Let $\boldsymbol{x}_n = [1, (X_n - x), \ldots, (X_n - x)^p]^T$, $\tilde{h}_n = h_n^\eta / \sum_{i=1}^{n-1} h_i^\eta$ and $\omega_{n,i} = h_i^\eta K_{h_i}(X_i - x) / \sum_{i=1}^{n} h_i^\eta$ for $\eta \in [0, 1]$, then by the inverse of the partitioned matrix, they obtain their recursive local polynomial estimators given by

$$\begin{aligned} P_n &= (1 + \tilde{h}_n)\left(P_{n-1} - \frac{\tilde{h}_n K_{h_n}(X_n - x) P_{n-1}\boldsymbol{x}_n \boldsymbol{x}_n^T P_{n-1}}{1 + \tilde{h}_n K_{h_n}(X_n - x)\boldsymbol{x}_n^T P_{n-1}\boldsymbol{x}_n}\right) \\ \hat{\boldsymbol{\beta}}_n &= \hat{\boldsymbol{\beta}}_{n-1} + \omega_{n,n}\left(Y_n - \boldsymbol{x}_n^T \hat{\boldsymbol{\beta}}_{n-1}\right) P_n \boldsymbol{x}_n \end{aligned}$$

Through the results of McLeish (1975) and Masry (1987), strong consistency of the estimators has been proven. These authors found that, although the same convergence rate is retained, the recursive algorithm has larger MSE than the nonrecursive counterpart. Chen and Gao (2013) considered the high-dimensional dependent data and proposed the multivariate local linear estimator based on the existing recursive weighted least-square estimator given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n &= \hat{\boldsymbol{\beta}}_{n-1} + a_n P_{n-1}\boldsymbol{z}_n (Y_n - \boldsymbol{z}_n^T \hat{\boldsymbol{\beta}}_{n-1}) \\ P_n &= P_{n-1} - a_n P_{n-1}\boldsymbol{z}_n \boldsymbol{z}_n^T P_{n-1} \\ a_n &= [K_{h_n}^{-1}(X_n - x) + \boldsymbol{z}_n^T P_{n-1}\boldsymbol{z}_n]^{-1} \end{aligned}$$

where $\boldsymbol{z}_n = [1 \;\; (\boldsymbol{X}_n - \boldsymbol{x})]^T$ and $\hat{\boldsymbol{\beta}}_n$ is the parameter to be estimated. Under regularity conditions for the kernel, bandwidth and regression function, they proved that the estimator has strong consistency. Amiri et al. (2014) studied a class of semi-recursive kernel estimators for dependent data with a bandwidth depending on h_n. They were able to show the mean squared error and strong consistency for these estimators, and establish their asymptotic normality. In addition to estimation of the regression function itself, Ngerng (2011) proposed a semi-recursive kernel method similar to (B.4) and (B.5) to estimate the local partial first derivative of the regression function given some regularly conditions. The author claimed that the estimator can be used to calculate the financial "Greeks" since keeping a delta-hedged position of a portfolio calls for frequent updating of first derivative estimates. Similarly, weak/strong consistency and asymptotic normality are proved in the paper.

In addition, Wu (2005) introduced physical and predictive dependence measures to quantify dependent data in a nonlinear system and supplied a method for a limit theory for stationary

processes. Based on the measures, Huang et al. (2014) developed a recursive nonparametric estimation method for time series. Their method is based on the Nadaraya-Watson estimator and is made up of two formulas

$$\begin{aligned} f_n(x) &= f_{n-1}(x) - H_n^{-1} h_n [f_{n-1}(x) - K_{h_n}(X_n - x)/h_n] \\ g_n(x) &= g_{n-1}(x) - \gamma_n [g_{n-1}(x) - Y_n] \end{aligned}$$

where $H_n = \sum_{i=1}^{n} h_n$ and $\gamma_n = (f_n(x) H_n)^{-1} K_{h_n}(X_n - x)$. Note that $g_n(x)$ is the estimator of $g(x)$ and depends on $f_n(x)$. They provided the proof of the laws of the iterated logarithms to characterize strong consistency of the estimator and the corresponding central limit theorems for the normality.

Appendix C

Mixing Processes

Mixing processes are a class of stationary stochastic processes[1] which are widely used to investigate asymptotic properties of dependent data. They are an important tool for our theoretical studies. In this section we supply intuition about of mixing processes.

Classical probability theory mainly studies a group of independent random variables and their properties such as Kolmogorov's strong law of large numbers and Lindeberg-Feller's central limit theorem (Billingsley, 1995). However in practice the assumption of independence may not be realistic for the data. For example, financial observations exhibit temporal dependence at successive time points and spatial dependence among countries. Therefore models for correlation or dependence in data are required.

Given two events B and C in some probability space $(\Omega, \mathscr{F}, P)$, we know that the independence of B and C is defined to satisfy $P(B \cap C) - P(B)P(C) = 0$. It is also known that a σ-field is a collection of events which are closed under complement and countable union. So one can define that two σ-fields[2] $\mathscr{B}$ and $\mathscr{C}$ are independent if $P(B \cap C) - P(B)P(C) = 0$ for any $B \in \mathscr{B}$ and $C \in \mathscr{C}$. One may wonder whether a similar definition can be given to measure the dependence of $\mathscr{B}$ and $\mathscr{C}$. Rosenblatt (1956) introduced the following α-mixing coefficient (also called the strong mixing coefficient) for this purpose

$$\alpha(\mathscr{B}, \mathscr{C}) = \sup_{\substack{B \in \mathscr{B} \\ C \in \mathscr{C}}} |P(B \cap C) - P(B)P(C)| \tag{C.1}$$

From the above definition, one can immediately see that if $\mathscr{B}$ and $\mathscr{C}$ are independent, then

[1] Mixing is a stronger concept than ergodicity (Billingsley, 1995)

[2] Here $\mathscr{B}$ and $\mathscr{C}$ are sub σ-fields of $\mathscr{F}$, i.e. $\mathscr{B}, \mathscr{C} \subset \mathscr{F}$.

$\alpha(\mathscr{B}, \mathscr{C}) = 0$. Additionally, note that

$$\begin{aligned} |P(B \cap C) - P(B)P(C)| &= |P(B \cap C) - P(B \cap C)P(C) - P(B \cap C^c)P(C)| \\ &= |P(B \cap C)P(C^c) - P(B \cap C^c)P(C)| \\ &\leq \max\{P(B \cap C)P(C^c), P(B \cap C^c)P(C)\} \\ &\leq P(C)P(C^c) \leq \frac{1}{4} \end{aligned}$$

so $0 \leq \alpha(\mathscr{B}, \mathscr{C}) \leq \frac{1}{4}$. In addition to independence, we know that correlation is another measure to characterize the dependence of the data. Two random variables X and Y are said to be uncorrelated if their correlation coefficient $Corr(X, Y) = 0$. Thus one can define correlation of $\mathscr{B}$ and $\mathscr{C}$ by the following ρ-mixing coefficient (Bradley, 1988)

$$\rho(\mathscr{B}, \mathscr{C}) = \sup_{\substack{X \in L^2(\mathscr{B}) \\ Y \in L^2(\mathscr{C})}} |Corr(X, Y)| \tag{C.2}$$

where X and Y are square-integrable random variables in $\mathscr{B}$ and $\mathscr{C}$ respectively[3]. It is easy to verify that $0 \leq \rho(\mathscr{B}, \mathscr{C}) \leq 1$. Bradley (1986) established the inequality

$$\alpha(\mathscr{B}, \mathscr{C}) \leq 4\rho(\mathscr{B}, \mathscr{C}) \tag{C.3}$$

Now let $\{X_t\}$ be a stochastic process and the available information between time t and $t + k$ be denoted by $\mathscr{F}_t^{t+k} = \sigma(X_s, t \leq s \leq t + k) \subset \mathscr{F}$. Then the measures for σ-fields as mentioned above can be generalized for the stochastic process. The process is called α-mixing if, as $k \to \infty$,

$$\alpha_k = \sup_t \alpha\left(\mathscr{F}_{-\infty}^t, \mathscr{F}_{t+k}^{\infty}\right) \to 0. \tag{C.4}$$

It is noted that α_k is non-increasing as k becomes larger. This is because, for $k_1 < k_2$, it can be verified that $\mathscr{F}_{t+k_1}^{\infty} \supset \mathscr{F}_{t+k_2}^{\infty}$. If $\{X_t\}$ is strictly stationary, then the above definition of α-mixing can be written by omitting the "sup", i.e.

$$\alpha_k = \alpha\left(\mathscr{F}_{-\infty}^0, \mathscr{F}_k^{\infty}\right) \to 0. \tag{C.5}$$

Similarly the process is called ρ-mixing if, as $k \to \infty$,

$$\rho_k = \sup_t \rho\left(\mathscr{F}_{-\infty}^t, \mathscr{F}_{t+k}^{\infty}\right) \to 0 \tag{C.6}$$

[3]If X is a square-integrable random variable in $\mathscr{B}$, it means that the σ-field generated by X is the subset of $\mathscr{B}$ and $\int_\Omega X^2 dP < \infty$.

and if $\{X_t\}$ is strictly stationary, the above definition can be written as

$$\rho_k = \rho\left(\mathscr{F}_{-\infty}^{0}, \mathscr{F}_{k}^{\infty}\right) \to 0 \tag{C.7}$$

In addition, many researchers have devoted effort to developing other measures to characterize dependence such as β-mixing, ϕ-mixing and ψ-mixing. For example, a strictly stationary Markov chain is ψ-mixing (Bradley, 1983b). More details can be found in the book by Doukhan (1994).

Through the above definitions, it can be found that an α-mixing process is asymptotically independent and a ρ-mixing process is asymptotically uncorrelated. Probability theory tells us that independent random variables X and Y are uncorrelated, but in contrast, uncorrelated random variables might not be independent. For example, let $X \sim \mathcal{N}(0, \sigma^2)$ and $Y = X^2$, thus $Cov(X, Y) = EXY - EXEY = EX^3 = 0$ implying X and Y are uncorrelated, however clearly they are not independent because Y has a functional relationship with X. Through the inequality (C.3), if the process is ρ-mixing, then it is α-mixing. In other words, if the process is asymptotically uncorrelated, then it is asymptotically independent but the opposite may not be true.

As an example, independent processes are trivially special cases of mixing processes. If $\{X_t\}$ is a p-dependent process, that is,

$$X_t = \sum_{i=0}^{p} a_i \varepsilon_{t-i}$$

where the ε_t's are independent, then it can be found that $\alpha_q = 0$ for $q > p$. Thus the process is α-mixing. In addition, Bosq (1998) gave an example of the linear process

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}$$

where $a_j = O(e^{-rj})$ and ε_t's are independent and identically distributed (i.i.d.) satisfying $E\varepsilon_t = 0$ and $Var(\varepsilon_t) < \infty$. Then $\{X_t\}$ is ρ-mixing. Based on the relationship $\alpha_k \leq 4\rho_k$ as mentioned above, X_t is also α-mixing. Certain stationary ARMA processes are α-mixing (Athreya and Pantula, 1986), and ARCH and GARCH processes are α-mixing under some conditions (Francq and Zakoian, 2010). However some popular processes may not be α-mixing such as AR(1) processes (Ibragimov and Linnik, 1971; Chernick, 1981; Andrews, 1984).

In addition, many results of nonparametric estimators relied on independence assumptions, e.g. rates of convergence and asymptotic normality, can be extended to the mixing settings. Masry (1996) presented the uniform convergence rate and asymptotic normality of kernel estimators of density functions for dependent data. Masry and Fan (1997) established asymptotic

normality for the local polynomial estimator when the process is α-mixing or ρ-mixing. Hansen (2008) generalized the existing literature by considering the kernel estimators of both density functions and regression functions for α-mixing processes. The author derived the probability convergence and uniform almost sure convergence rates, where the estimators allow for the kernels with unbounded support. Xiao et al. (2003) studied the regression model with autocorrelated errors and proposed the local linear estimator of the regression function. They showed the estimator has asymptotic normality by assuming that the data are α-mixing with some decay rate. Peng and Yao (2004) investigated nonparametric regression under dependent errors with infinite variance. In addition, prediction problems for dependent data by nonparametric estimators have been considered extensively (Collomb and Härdle, 1986; Bosq, 1998; Hall et al., 2002).

Curriculum Vitae

Name:	Xin Wang
Post-Secondary Education and Degrees:	The University of Western Ontario London, Ontario, Canada 2013 - 2015 Ph.D.
	The University of Western Ontario London, Ontario, Canada 2012 - 2013 M.Sc.
	Academy of Mathematics & Systems Science, Chinese Academy of Sciences Beijing, China 2007 - 2012 Ph.D.
	Beijing Institute of Technology Beijing, China 2003 - 2007 B.Sc.
Honours and Awards:	National Post-Graduate Mathematic Contest in Modeling (GMCM) Second Prize, 2008
	National Undergraduate CSEE Mathematic Contest in Modeling Second Prize, 2005
Related Work Experience:	Risk Analyst, TD Bank May 2014 - Present
	Teaching Assistant, University of Western Ontario 2012 - 2015
	Research Assistant, Chinese Academy of Sciences 2007 - 2012

Publications:

Hol F, Wang X, Keymer JE (2012) *Population Structure Increases the Evolvability of Genetic Algorithms*. Complexity, 17 (5): 58-64. (co-first author)

Han J, Han HW, Wang X (2012) *"Knowing More is Less" in Combinatorial Games*. Proceedings of the 10th world congress on intelligent control and automation (WCICA2012), Beijing, China: 3526-3532.

Wang X, Mu YF, Han J (2012) *What Strategies can Induce Cooperation between Heterogeneous Players?*. Proceedings of the 31th Chinese Control Conference (CCC2012), Hefei, China: 1153-1157.

Wang X, Han J, Han HW (2011) *Special Agents can Promote Cooperation in the Population*. PLoS ONE, 6 (12): e29182.

Wang X, Han J (2011) *Special Agents can Promote Cooperation*. Proceedings of the 30th Chinese Control Conference (CCC2011), Yantai, China: 5764-5768.

Li DN, Wang X, Meng Y (2007) *A Destruction-resistent Routing Algorithm in Low Earth Orbit Satellite Networks*. The 3rd IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCom2007), Shanghai, China: 1841-1844.

Li DN, Wang X, Zhang DK (2007) *Stability-guaranteed Clustering in Satellite Networks*. The 3rd IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCom2007), Shanghai, China: 1817-1820.