Stockholms universitet

Nonparametric Volatility Density Estimation

Xiaofen Huang

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2013:9
Matematisk statistik
December 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2013:9**
http://www.math.su.se

Nonparametric Volatility Density Estimation

Xiaofen Huang*

December 2013

Abstract

Stochastic volatility modelling of financial processes has become popular and most models contain a stationary volatility process. For volatility density estimation Van Es et al.(2003) introduced a deconvolution procedure; in this thesis we instead propose another nonparametric method. It is a two-step procedure, where we first apply some nonparametric regression technique to generate the process estimates, based on which we then use the ordinary kernel density estimator. To find the method parameters, we also suggest automatic parameter selectors using theories from the Nadaraya-Watson estimator and continuous-time kernel density estimation. To evaluate performance of the proposed method in comparison with the deconvolution approach, we apply both methods on data simulated from Heston model and real data. For simulated data, we divide it into two sets; high frequency(hourly) and low frequency(daily). We find that the proposed method slightly outperforms the deconvolution approach in terms of mean integrated squared error(MISE) for high frequency data. However, for low frequency data, the deconvolution procedure obtains far less MISE than the proposed method. Unfortunately, their performances on the real data are hardly comparable.
Keywords: Volatility Density Estimation, Deconvolution, Bandwidth Selection, Nadaraya- Watson Estimator, Continuous-time Kernel Estimation, Heston Model.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: h.xiaofen@gmail.com. Supervisor: Martin Sköld.

Abstract

Stochastic volatility modelling of financial processes has become popular and most models contain a stationary volatility process. For volatility density estimation Van Es et al.(2003) introduced a deconvolution procedure; in this thesis we instead propose another nonparametric method. It is a two-step procedure, where we first apply some nonparametric regression technique to generate the process estimates, based on which we then use the ordinary kernel density estimator. To find the method parameters, we also suggest automatic parameter selectors using theories from the Nadaraya-Watson estimator and continuous-time kernel density estimation. To evaluate performance of the proposed method in comparison with the deconvolution approach, we apply both methods on simulated data from Heston model and real data. For simulated data, we divide it into two sets; high frequency(hourly) and low frequency(daily). We find that the proposed method slightly outperforms the deconvolution approach in terms of mean integrated squared error(MISE) for high frequency data. However, for low frequency data, the deconvolution procedure obtains far less MISE than the proposed method. Unfortunately, their performances on the real data are hardly comparable.

Keywords: Volatility Density Estimation, Deconvolution, Bandwidth Selection, Nadaraya-Watson Estimator, Continuous-time Kernel Estimation, Heston Model.

Acknowledgements

I am grateful to my supervisor of this thesis, Assistant Professor Ph.D. Martin Sköld for his valuable comments and guidance; and for his assistance and insightful reviews.

This work would not have been possible without the help of my family, friends and colleagues. Thanks for their encouragements and unconditional supports.

Stockholm, Nov. 24^{th}, 2013.

Xiaofen Huang

Contents

List of Tables

List of Figures

1 Introduction

Volatility as a measure of variation of prices of some financial asset, is of great importance in many financial applications like risk managements and option pricing. In the widely used Black-Sholes Model, it is assumed constant which unfortunately cannot explain some long-observed features such as the volatility smiles. To solve such a shortcoming, stochastic volatility models(continuous time or discrete time) have been proposed to model the volatility as a random process, often called the volatility process. In this thesis, we will discuss some nonparametric methods for estimating volatility density given observations of the price process of some asset.

Assuming we have discretely observed price data at a regular time instant, denote X_t as its standardized demeaned and de-trended log-return process. To describe the behavior of this type of data, we consider a stochastic model of the form

$$X_t = \sigma_t Z_t \quad t = 1, \ldots, n \tag{1}$$

where $\{Z_t\}_{t=1}^n$ is a typical sequence of i.i.d Gaussian noises and Z_t is assumed to be independent of σ_t for each time t. Here we model the volatility process σ as a strictly stationary process satisfying the strong mixing condition and ergodic properties. In addition, we assume that univariate marginal distribution of σ admits a density $\pi(v)$ w.r.t the Lebesgue on $(0, \infty)$. In the literature, there have been proposed many stochastic volatility models which imply different marginal distributions of σ. For instance, the Heston model by Heston(1993) displays a gamma distribution of σ, while Ornstein-Uhlenbeck process by Wiggins(1987) suggests a normal distribution of $\log \sigma^2$. Besides, most of the widely-used models implies that the invariant distribution of σ is unimodal. It hardly explains the often-seen volatility clustering phenomenon, which may lead to a bimodal marginal distribution of σ. Therefore, it is sensible to implement some nonparametric method to reveal the shape of the volatility density.

By simply taking the logarithm of the squared equation (1), we transform the model to the convolution form

$$\log X_t^2 = \log \sigma_t^2 + \log Z_t^2, \tag{2}$$

based on which Van Es et al.(2003) proposed a deconvolution procedure for the volatility density estimation using ideas from deconvolution theory. In this thesis, we will instead propose a slightly different nonparametric method called transformed volatility density estimation. The method is a two-step procedure, where we first use some nonparametric regression technique-moving average to estimate $\log \sigma_t^2$ and then based on these estimates we apply the ordinary kernel density estimator to estimate the density

of $\log \sigma^2$. There are two parameters involved: the window size k for moving average estimates and the smoothing parameter u for the kernel density estimator. As they control the smoothness of their corresponding estimate curves, it is crucial to have the appropriate choices. That's why we also propose an automatic parameter selector for each of them. For the window size k, it is chosen using the theory of bandwidth selections for the Nadaraya-Watson estimator. It is because the moving average estimator is approximately equivalent to the Nadaraya-Watson estimator with a uniform kernel. For the smoothing parameter u, we apply some ideas from continuous-time kernel density estimation.

We compare these two methods by applying them on both simulated data from Heston model and real data(Nasdaq index). For simulation data, we divide it into two sets; high frequency(hourly) and low frequency(daily). We find that the transformed approach performs slightly better than the deconvolution method in terms of mean integrated squared error(MISE) for high frequency data. However, for low frequency data the deconvolution has far better fit compared to the transformed method. Furthermore, on the basis of real data the performances of both approaches can be hardly determined. The thesis is organized as follows. In chapter 2 we summarize the theoretical background of various nonparametric methods including kernel density estimations and kernel regression. In chapter 3 we first describe a class of stochastic models based on which we review the deconvolution procedure by Van Es. et al(2003) and then propose the transformed kernel volatility density estimator. Finally, we give an example of such a stochastic model. In chapter 4 we present the numerical results and conclude in chapter 5.

2 Nonparametric Kernel Methods

In this chapter, we outline some general theories of classical nonparametric kernel-type approaches for density estimations and regression estimations. We review the kernel density estimation with an addition of an summary of a continuous-time version of this kernel density estimation. After that, we describe the deconvolution kernel density estimation and finish with a brief discussion of kernel regression.

2.1 Kernel Density Estimation

In this section, we present the basic definition of the kernel density estimator along with its error criterion including mean squared error and mean integrated squared error. Additionally, we introduce the asymptotic approximation of mean square error and mean

integrated squared error based on Wand and Jones(1995).

2.1.1 Definition

Suppose we have an independent and identically distributed random sample $X_1, X_2, \ldots, X_n$ taken from a continuous, univariate density f. We try to estimate the density f with the **kernel density estimator** of f denoted as $\hat{f}(x;h)$, which is given by

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{2.3}$$

where K is called the *kernel*, a real-valued function satisfying $\int K(x)dx = 1$ and h is a positive number called the *bandwidth* or *smoothing parameter*. To have a slightly compact formula, we introduce the *scaled kernel* which is defined as $K_h(u) = h^{-1}K(u/h)$. Usually we choose the kernel to be symmetric unimodal probability density function so that $\hat{f}(x;h)$ itself is also a density.
We can see that the kernel estimate at a point x is constructed by centering a scaled kernel on all the observations $X_1, X_2, \ldots, X_n$ and taking the average of these n kernels. It is found that the choice of the shape of the kernel function is not as important as the choice of bandwidth.(see Wand and Jones(1995)) It is critical to select an appropriate bandwidth since there is a danger of *undersmoothing* or *oversmoothing*. More specifically, a smaller bandwidth leads to a very spiky and variable estimate (which is called undersmoothing) while a larger bandwidth results in a oversmoothed estimate which tends to smooth away some important features of the data.

2.1.2 MSE and MISE Error Criterion

After constructing the kernel density estimator, it is necessary to specify an appropriate error criterion for the analysis of its performance at a single point as well as over the whole real line. Let $\hat{f}(x;h)$ be an estimator of the density function $f(x)$ at some point x. We measure the closeness of $\hat{f}(x;h)$ to $f(x)$ by the size of *mean square error*(MSE), which is given by

$$\begin{aligned} \mathrm{MSE}(\hat{f}(x;h)) &= E[(\hat{f}(x;h) - f(x))^2] \\ &= \mathrm{Var}(\hat{f}(x;h)) + (E[\hat{f}(x;h)] - f(x))^2. \end{aligned} \tag{2.4}$$

The MSE can be decomposed into variance and squared bias. Thus, to compute the MSE $(\hat{f}(x;h))$ it requires the expression of the mean and variance which can be derived from (2.3). Denote X as a random variable having density f. They are given as below

$$E\hat{f}(x;h) = EK_h(x - X) = \int K_h(x - y)f(y)dy = (K_h * f)(x), \tag{2.5}$$

$$\text{Var}\{\hat{f}(x;h)\} = n^{-1}\{({K_h}^2 * f)(x) - (K_h * f)^2(x)\} \tag{2.6}$$

where $(K_h * f)(x)$ is the convolution of K_h and f. Inserting them into (2.4), we have

$$\text{MSE}\{\hat{f}(x;h)\} = n^{-1}\{({K_h}^2 * f)(x) - (K_h * f)^2(x)\} + \{(K_h * f)(x) - f(x)\}^2. \tag{2.7}$$

However, in most cases it is preferable to measure the distance between the functions $\hat{f}(\cdot;h)$ and f over the entire real line. One such error criterion is the *mean integrated squared error*(MISE), given by

$$\text{MISE}\{\hat{f}(\cdot;h)\} = E[\text{ISE}\{\hat{f}(\cdot;h)\}] = E\int\{\hat{f}(x;h) - f(x)\}^2 dx \tag{2.8}$$

where ISE is short for *integrated squared error*, with $\text{ISE}\{\hat{f}(\cdot;h)\} = \int\{\hat{f}(x;h) - f(x)\}^2 dx$. By changing the order of integration we get

$$\text{MISE}\{\hat{f}(\cdot;h)\} = \int E\{\hat{f}(x;h) - f(x)\}^2 dx = \int \text{MSE}\{\hat{f}(x;h)\} dx. \tag{2.9}$$

After plugging in the expression for MSE (2.7) and some manipulations, we get the final expression for MISE

$$\begin{aligned}\text{MISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1}\int K^2(x)dx + (1 - n^{-1})\int (K_h * f)^2(x)dx \\ - 2\int (K_h * f)(x)f(x)dx + \int f(x)^2 dx.\end{aligned} \tag{2.10}$$

For the sake of simplicity, we use the notation $R(g) = \int g(x)^2 dx$ for any integrable function g. This means we can rewrite the first term of MISE as $(nh)^{-1}R(K)$.

2.1.3 Asymptotic MSE and MISE Criterion

To have a better understanding of the dependence of MSE and MISE on the bandwidth, we will study the derivation of large sample approximations for the leading bias and variance terms. They can also be useful for obtaining the rate of convergence of the kernel density estimator. We begin with imposing the following assumptions as in Wand and Jones(1995).

Condition 2.1.

1. The density f has continuous second derivative f'' which is also square integrable and ultimately monotone.
2. The bandwidth is a non-random positive sequence satisfying $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.

3. The kernel function is a bounded probability density function having symmetry about the origin and finite fourth moment.

Note that a function is called ultimately monotone if it is monotone over $(-\infty, -M)$ and (M, ∞) for some $M > 0$. By a change of variable and Taylor expansion of f, we have

$$\begin{aligned}\text{bias}[\hat{f}(x;h)] = E[\hat{f}(x;h)] - f(x) &= \frac{1}{2}h^2 f''(x) \int z^2 K(z)dz + o(h^2) \\ &= \frac{1}{2}h^2 \mu_2(K) f''(x) + o(h^2), \end{aligned} \tag{2.11}$$

where we use the notation $\mu_2(K) = \int z^2 K(z)dz$. So the bias is of order of h^2. For the variance, we will obtain

$$\text{Var}\{\hat{f}(x;h)\} = (nh)^{-1} R(K) f(x) + o((nh)^{-1}) \tag{2.12}$$

which means the variance is of order $(nh)^{-1}$. It is easy to notice from the orders of variance and bias, that larger value of bandwidth leads to a decline in variance but an increase in bias. This is the so-called *variance-bias tradeoff*.

Finally, by the sum of (2.12) and the square of (2.11), we get the mean square error

$$\text{MSE}\{\hat{f}(x;h)\} = (nh)^{-1} R(K) f(x) + \frac{1}{4}h^4 \mu_2(K)^2 f''(x)^2 + o\{(nh)^{-1} + h^4\}. \tag{2.13}$$

Under the integrability condition on f we impose, taking the integral of MSE gives us

$$\text{MISE}\{\hat{f}(\cdot;h)\} = \text{AMISE}\{\hat{f}(\cdot;h)\} + o\{(nh)^{-1} + h^4\}, \tag{2.14}$$

where

$$\text{AMISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1} R(K) + \frac{1}{4}h^4 \mu_2(K)^2 R(f''). \tag{2.15}$$

We refer to the AMISE as the asymptotic MISE, which provides useful asymptotic approximation for the MISE especially when finding the optimal bandwidth. More specifically, by setting the derivative of equation (2.15) with respect to h equal to zero, we will have an expression for the optimal bandwidth which minimizes the AMISE. It is given by

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')n}\right]^{1/5} \tag{2.16}$$

The expression depends on $R(f'')$ i.e. the total curvature of f, apart from the known kernel K and n. It also gives the minimum MISE and the rate of convergence of the MISE for this kernel estimator equal to $n^{4/5}$ under the conditions we impose above.

2.2 Continuous-time Kernel Density Estimation

In this section we consider a (strictly) stationary stochastic process $X = \{X_t\}_{t\geq 0}$ with a marginal density f. Based on a random sample $\{X_t; t \in [0,T]\}$ from this process X, for any $x \in \mathbb{R}$ we define the kernel density estimator for the density function f as

$$\hat{f}_T(x) = \frac{1}{Th}\int_0^T K(\frac{x - X_t}{h})dt = \frac{1}{T}\int_0^T K_h(x - X_t)dt \tag{2.17}$$

where K is a kernel function satisfying $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$ and $\int K(u)du = 1$. h is called the bandwidth or smoothing parameter as in kernel density estimator in discrete time. The advantage of having a continuous-time sample is that we can construct an unbiased estimator in terms of local times and occupation-time density(OTD), whose existence gives us a faster rate of convergence to zero than in discrete time (see Sköld&Hössjer 1999). It is shown that given a discrete-time stationary ergodic process $\{X_i\}_{i=1}^n$ having a marginal density with m continuous derivatives, the optimal rate of convergence of MSE for a kernel density estimator is of order $n^{-2m/(2m+1)}$ (Wahba 1975). In comparison, a continuous-time kernel density estimator can obtain a rate of convergence faster than T^{-1} under some conditions imposed by Castellana and Leadbetter(1986). However, for a smooth process the optimal rate will be of order $(\log T)/T$ due to an infinite variance of its OTD shown by Sköld and Hössjer (1999).

Concerning the nature of the assumed dependence structure of the process, we consider the strong mixing coefficient $\alpha_h(s,x)$ for any point x and $s > 0$,

$$\alpha_h(s,x) = \int\int K_h(u-x))K_h(v-x)(f_{X_0,X_s}(u,v) - f(u)f(v))dudv, \tag{2.18}$$

where f_{X_0,X_s} is denoted as the joint distribution of X_0 and X_s and assumed to exist for $s \neq 0$. It is shown by Castellana and Leadbetter(1986) that the asymptotic variance of $\hat{f}_T(x)$ takes the form

$$T\mathrm{Var}(\hat{f}_T(x)) \to 2\int_0^T (1 - s/T)\alpha_h(s,x)ds = 2\int_0^\infty (f_{X_0,X_s}(x,x) - f^2(x))ds \tag{2.19}$$

where $\alpha_h(s,x) \to f_{X_0,X_s}(x,x) - f^2(x)$, as $h \to 0$. Worth mentioning is that the integral in right-hand side of the equation (2.19) is finite for a process behaving locally as Brownian motion, but not for a differentiable process. Define $Y_s = (X_s - X_0)/s$ and $Y_0 := X_0'$. In order to present the results from Sköld and Hössjer (1999), we request the following conditions to be satisfied.

Condition 2.2.

1. The density f is a continuous and bounded function such that its second derivative is continuous, square integrable and ultimately monotone.
2. The kernel K is a symmetric density function with compact support.
3. The strong mixing coefficient α_h satisfies both $\lim_{h\to 0}\int_\delta^\infty |\alpha_h(s,x)|ds < \infty$ for all $\delta > 0$ and $\lim_{h\to 0}\int\int_\delta^\infty |\alpha_h(s,x)|dsdx < \infty$.
4. There exists a constant $f_{X_0,X_0'}(x,0)$ such that $\lim_{\varepsilon\to 0}\sup_{(u,v,s)\in B_\varepsilon\times[0,\varepsilon)} |f_{X_0,Y_s}(u,v) - f_{X_0,X_0'}(x,0)| = 0$, where $B_\varepsilon = \{(u,v); (u-x)^2 + v^2 < \varepsilon^2\}$.
5. There exists a constant $f_{X_0'}(0)$ such that $\lim_{\varepsilon\to 0}\sup_{(v,s)\in(-\varepsilon,\varepsilon)\times[0,\varepsilon)} |f_{Y_s}(v) - f_{X_0'}(0)| = 0$.

Under these assumptions above, Sköld and Hössjer (1999) gave exact form of the asymptotic variance and its integral which are written respectively as follows,

$$\mathrm{Var}(\hat{f}_T(x)) = 2f_{X_0,X_0'}(x,0)\log(h^{-1})T^{-1} + o(\log(h^{-1})T^{-1}), \tag{2.20}$$

$$\int \mathrm{Var}(\hat{f}_T(x))dx = 2f_{X_0'}(0)\log(h^{-1})T^{-1} + o(\log(h^{-1})T^{-1}). \tag{2.21}$$

as $h \to 0$ and $T \to \infty$. These allow us to derive an explicit expression for the asymptotic bandwidth optimum which minimizes the AMISE:

$$h(T) = \left(\frac{1}{CT}\right)^\alpha = \left[\frac{2f_{X_0'}(0)}{\mu_2(K)^2 R(f'')T}\right]^\alpha, \tag{2.22}$$

where

$$\alpha = 1/4 \text{ and } C = \frac{\mu_2(K)^2 R(f'')}{2f_{X_0'}(0)}. \tag{2.23}$$

This optimal bandwidth results in the rate of convergence of MISE equal to $(\log T)/T$ for all $\alpha \geq 1/4$ regardless of the choice of $C > 0$, since

$$\lim_{T\to\infty} T(\log T)^{-1}\int_{-\infty}^{\infty} E(\hat{f}_T(x; h(T)) - f(x))^2 dx = 2\alpha f_{x_0'}(0). \tag{2.24}$$

2.3 Deconvolution Kernel Density Estimation

According to the previous section, the kernel density estimator requires direct data. However, it is not always able to obtain in practise. One example is that data are measured

with some non-negligible error. In this section, we will discuss the nonparametric density estimation from error contaminated data, which is called deconvolution kernel density estimation. We will present the deconvolution estimator as well as its MISE criterion based on chapter 6.2.4 of Wand and Jones(1995). We will also summarize some insights obtained by Fan(1991) about how difficult the deconvolution is. Further, we will describe the computational procedure using Fast Fourier Transform, based on Silverman(1986).

2.3.1 Deconvoluting Kernel Density Estimator

Assume that X_1,X_2,...,X_n are an unobservable random sample drawn from a common density f_X. Our objective is to estimate this unknown density f_X based on the observed data Y_1,...,Y_n which is defined as

$$Y_i = X_i + Z_i \quad i = 1, \ldots, n \tag{2.25}$$

where the error term Z_i's are i.i.d random variables with known *error density* f_Z and independent of X_i. That is why the density of f_Y is the convolution of f_X and f_Z, i.e.

$$f_Y = f_X * f_Z. \tag{2.26}$$

By a usage of Fourier transform(or characteristic function) properties, we have the Fourier transform of density f_Y

$$\varphi_{f_Y}(t) = E(e^{itY}) = E(e^{it(X+Z)}) = E(e^{itX}e^{itZ}) = \varphi_{f_X}(t)\varphi_{f_Z}(t). \tag{2.27}$$

Then we apply Fourier inversion Theorem and obtain the target density f_X written as

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_{f_X}(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_{f_Y}(t)}{\varphi_{f_Z}(t)}dt, \tag{2.28}$$

provided that $\varphi_{f_Z}(t) \neq 0$. If we replace f_Y by its kernel estimator $\hat{f}_Y(y;h) = \frac{1}{n}\sum_{j=1}^{n} K_h(y - Y_j)$ (according to formula (2.3)), an estimate of f_X can be given by

$$\hat{f}_X(x;h) = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_{\hat{f}_Y(x;h)}(t)}{\varphi_{f_Z}(t)}dt \tag{2.29}$$

which is the so-called the **deconvoluting kernel density estimator**. Note that

$$\begin{aligned}
\varphi_{\hat{f}_Y}(t) &= \int e^{itx}\hat{f_Y}(x)dx \\
&= \frac{1}{n}\sum_{j=1}^{n}\int e^{itx}K_h(x-Y_j)dx \\
&= \frac{1}{n}\sum_{j=1}^{n}\int e^{it(x-Y_j)}K_h(x-Y_j)e^{itY_j}dx \\
&= \varphi_{K_h}(t)\cdot\frac{1}{n}\sum_{j=1}^{n}e^{itY_j} \\
&= \int e^{itx}\frac{1}{h}K(\frac{x}{h})dx\cdot\frac{1}{n}\sum_{j=1}^{n}e^{itY_j} \\
&= \varphi_K(th)\frac{1}{n}\sum_{j=1}^{n}e^{itY_j} \qquad (2.30) \\
&= \varphi_K(th)\varphi_{emp}(t) \qquad (2.31)
\end{aligned}$$

where $\varphi_{emp}(t) = \frac{1}{n}\sum_{j=1}^{n}e^{itY_j}$, which is called the *empirical characteristic function* as in Van Es et al.(2005).

By inserting the equation (2.31) into the equation (2.29), we obtain one expression of the deconvolution estimator,

$$\hat{f}_X(x;h) = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_K(th)\varphi_{emp}(t)}{\varphi_{f_Z}(t)}dt. \qquad (2.32)$$

Or we can plug the equation (2.30) into (2.29)

$$\begin{aligned}
\hat{f}_X(x;h) &= \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_K(th)}{\varphi_{f_Z}(t)}\frac{1}{n}\sum_{j=1}^{n}e^{itY_j}dt \\
&= \frac{1}{n}\frac{1}{2\pi}\sum_{j=1}^{n}\int e^{-it(x-Y_j)}\frac{\varphi_K(th)}{\varphi_{f_Z}(t)}dt \\
&= \frac{1}{nh}\frac{1}{2\pi}\sum_{j=1}^{n}\int e^{-is(\frac{x-Y_j}{h})}\frac{\varphi_K(s)}{\varphi_{f_Z}(s/h)}ds \qquad (2.33)
\end{aligned}$$

and have another another expression of the deconvolution kernel density estimator given by

$$\hat{f}_X(x;h) = \frac{1}{nh}\sum_{j=1}^{n}v_h(\frac{x-Y_j}{h}) \qquad (2.34)$$

where

$$v_h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\varphi_K(s)}{\varphi_{f_Z}(s/h)}e^{-isx}ds. \qquad (2.35)$$

From the equation (2.34) above, it is apparent that the deconvolution is in the same basic form as the ordinary kernel estimator except with different kernel functions. Comparing with the kernel K, v_h is referred to as the "effective" kernel as in Wand and Jones(1995), whose shape is dependent on the choice of the bandwidth.

2.3.2 MISE Criterion

To access the quality of the deconvolution estimator and investigate its asymptotic behavior, we need to specify its MISE criterion. According to Wand and Jones(1995), it is given by

$$\begin{aligned}\text{MISE}\{\hat{f}(\cdot;h)\} = \frac{1}{nh}R(v_h(\cdot;h)) + (1-\frac{1}{n})\int (K_h * f)^2(x)dx \\ - 2\int (K_h * f)(x)f(x)dx + \int f(x)^2 dx. \end{aligned} \tag{2.36}$$

Notice that the MISE is also of similar form as the one for ordinary kernel estimator(see equation (2.10)). The only difference lies in the first term, where it is $(nh)^{-1}R(K)$ in the latter case. It is actually the extra measurement error Z gives rise to the increase in MISE which is

$$(nh)^{-1}\{R(v_h(\cdot;h)) - R(K)\}, \tag{2.37}$$

where

$$R(v_h(\cdot;h)) = (2\pi)^{-1}\int \varphi_K(t)^2 |\varphi_{f_Z}(t/h)|^{-2}dt. \tag{2.38}$$

The size of the integral, or more specifically the size of the reciprocal of the characteristic function of error variable determines the effect of the measurement error on MISE.(see Wand and Jones 1995) Hence, the behavior of the error distribution has an impact on the MISE of the deconvolution estimator.

2.3.3 The Difficulty of The Deconvolution

It is found that the difficulty of the deconvolution depends not only on the smoothness of the target density f_X but also on the smoothness of the distribution: the smoother, the harder (Fan 1991). More specifically, Fan(1991) investigated optimal rates of convergence in terms of mean square errors corresponding to two types of error distributions-ordinary smooth and super smooth distributions. For the sake of clarity, here we present their definitions along with the usual condition Fan(1991) imposed on the unknown density f_X .

Condition 2.3. *f belongs to the set $\mathfrak{L}_{m,\alpha,B}$* with

$$\mathfrak{L}_{m,\alpha,B} = \{f(x) = |f^{(m)}(x) - f^{(m)}(x+\delta)| \leq B\delta^{\alpha}\}$$

for known constants m, B and $0 \leq \alpha < 1$

Definition 2.1. (Super smooth) *The distribution of a random variable Z is called super smooth of order β if its characteristic function $\varphi_{f_Z}(t)$ satisfies*

$$d_0|t|^{\beta_0}\exp(-|t|^{\beta}/\gamma) \leq |\varphi_{f_Z}(t)| \leq d_1|t|^{\beta_1}\exp(-|t|^{\beta}/\gamma) \text{ as } t \to \infty \tag{2.39}$$

where d_0,d_1,β,γ are some positive constants and β_0,β_1 are constants.

Definition 2.2. (Ordinary Smooth) *The distribution of a random variable Z is called ordinary smooth of order β if its characteristic function $\varphi_{f_Z}(t)$ satisfies*

$$d_0|t|^{\beta} \leq |\varphi_{f_Z}(t)| \leq d_1|t|^{\beta} \text{ as } t \to \infty \tag{2.40}$$

where d_0,d_1,β are some positive constants.

In other words, by the smoothness of the error distribution Z, we mean the order of its characteristic function φ_{f_Z} as t tends to infinity. Take examples, normal, mixture normal and Cauchy distribution are super smooth; gamma, symmetric gamma distributions are ordinary smooth. It is shown that for deconvoluting a super smooth error, the faster rate of convergence is only of order of $(\log n)^{-a}$ for some positive number a.(see Fotopoulos 2000) Particularly, for normal deconvolution the optimal rate of convergence is much slower. In contrast, if the error is ordinary smooth, the optimum rate of convergence is of order n^{-b} for some positive number b.(see Fotopoulos 2000) This means that deconvolution problem with a super smooth error is much more difficult to solve than the one with an ordinary smooth error. Additionally, for both smoothness cases, the deconvolution gets harder when the order of the smoothness gets higher. Therefore, in practice one has to be careful when deconvoluting with a super smooth error. While for ordinary smooth case, deconvolution techniques are possibly useful.

2.3.4 Computational Algorithm

The difficulty of a deconvolution problem lies in not only the problem itself but its computation, since its direct computing is highly ineffective. Fortunately, fast Fourier transform(FFT) can be used to perform this time-consuming computation thanks to its discrete structure. Denote IFT as inverse Fourier transform, we could rewrite the

equation (2.32) as

$$\hat{f}_X(x;h) = IFT(\frac{\varphi_K(th)\varphi_{emp}(t)}{\varphi_{f_Z}(t)}). \tag{2.41}$$

Here we assume that the characteristic function of K and f_Z are known. Since the fast Fourier transform(FFT) is the algorithm of computing the discrete Fourier transform(DFT) and Inverse discrete Fourier transform(IDFT), it can be used to find the empirical characteristic function $\varphi_{emp}(t)$ and perform Inverse Fourier transform(IFT) in the equation (2.41). The algorithm can be divided into three steps:
First, we begin with data discretization. Let $[a, b]$ be an interval containing all the data. Set $M = 2^r$ for some integer r. The density estimates will be found on the M points. Define

$$\delta = (b-a)/M$$

$$t_k = a + k\delta$$

for $k = 0, 1, \ldots, (M-1)$. The binning can be done in the following way. If a data point X lies in the interval $[t_k, t_{k+1}]$, it is split into a weight $\frac{1}{n\delta}(t_{k+1} - X)$ at t_k and $\frac{1}{n\delta}(X - t_k)$ at t_{k+1}. We call the sequence of weights w_k, whose sum is equal to $1/\delta$.
The next step is to compute the following sum using FFT,

$$Y_l = \frac{1}{M}\sum_{k=0}^{M-1} w_k e^{\frac{i2\pi kl}{M}} \tag{2.42}$$

for $-\frac{1}{2}M \leq l \leq \frac{1}{2}M$. This sum helps to find the value of the empirical function $u(s_l)$ where $s_l = \frac{2\pi l}{b-a}$. It follows,

$$Y_l = \frac{1}{M}e^{ias_l}\sum_{k=0}^{M-1} w_k e^{it_k s_l} \approx \frac{1}{M}\frac{1}{\delta}e^{ias_l}\frac{1}{n}\sum_j e^{is_l S_j} = e^{ias_l}\frac{u(s_l)}{b-a}. \tag{2.43}$$

Finally, define $\zeta_l{}^* = \frac{\varphi_k(hs_l)}{\varphi_{f_Z}(s_l)}u(s_l)$ and $\zeta_k = IFT(\zeta_l{}^*)$, then

$$\begin{aligned}\zeta_k &= \sum_{l=-M/2}^{M/2} e^{\frac{2\pi ikl}{M}}\zeta_l{}^* \approx \sum_l e^{-is_l t_k}\cdot e^{ias_l}\frac{\varphi_k(hs_l)}{\varphi_{f_Z}(s_l)}Y_l \\ &= \sum_l e^{-is_l t_k}\frac{\varphi_k(hs_l)}{\varphi_{f_Z}(s_l)}\frac{u(s_l)}{b-a} \approx \frac{1}{2\pi}\int e^{-ist_k}\frac{\varphi_k(hs)}{\varphi_{f_Z}(s)}u(s)ds = \hat{f}_X(t_k;h).\end{aligned} \tag{2.44}$$

Therefore, from the last equality above we can see that FFT is a suitable computational tool for deconvolution estimator.

2.4 Kernel Regression

As one of the nonparametric regression approaches, kernel regression is developed by using the same ideas and mathematical skills in the analysis of kernel density estimation. We devote this section to study a class of kernel regression estimators called local polynomial kernel estimators, among which we mainly focus on the Nadaraya-Watson estimator. Then we move on to give some basic definitions for kernel functions required before analyzing some asymptotic properties of some error criterions for the Nadaraya-Watson estimator. Lastly, we present some bandwidth selection methods for the Nadaraya-Watson estimator.

2.4.1 Local Polynomial Kernel Estimators

Based on Wand and Jones(1995), the study of nonparametric regression is divided into two contexts- *fixed design* and *random design*, where in the latter case a bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of random pairs is observed. While in the first case $x_1, \ldots, x_n$ known as the design variables are ordered non-random numbers. In particular, when the $x_{i+1} - x_i$ remains constant for all i or more specially $x_i = i/n$ for $i = 1, \ldots, n$, it is a special case called *equally spaced fixed design* which we will mainly focus on. In an equally spaced fixed design model, we assume the respond variable to be satisfying

$$Y_i = m(x_i) + v^{1/2}(x_i)\varepsilon_i, \ i = 1, \ldots, n \tag{2.45}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with zero mean and unit variance. The function m and v are known as the *regression function* and *variance function* respectively, since $E(Y_i) = m(x_i)$ and $\mathrm{Var}(Y_i) = v(x_i)$. If we assume $v(x_i) = \sigma^2$ for all i, then the model is called homoscedastic, otherwise heteroscedastic.

The main idea behind the *local polynomial kernel estimator* is for a given point x fitting a p-th polynomial locally to the data points (x_i, Y_i) by weighted least square with kernel weights $K_h(x_i - x)$, which is a kernel function scaled by the smoothing parameter h. More specifically, at a point x, the local polynomial kernel estimator is a minimizer of

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(x_i - x) - \ldots - \beta_p(x_i - x)^p\}^2 \cdot K_h(x_i - x). \tag{2.46}$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$. Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as the response vector and $\mathbf{X}$ as an $n \times (p+1)$ design matrix given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{pmatrix}$$

Also $\mathbf{W}$ is denoted as a $n \times n$ diagonal matrix of the form $\mathbf{W} = \text{diag}\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}$. Provided the invertibility of $\mathbf{X}^T\mathbf{W}\mathbf{X}$, the theory of standard weighted least squares gives us

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}, \tag{2.47}$$

which leads to the value of the estimate $\hat{m}(x; p, h)$ equal to the height of $\hat{\beta}_0$, i.e.,

$$\hat{m}(x; p, h) = \mathbf{e}_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}, \tag{2.48}$$

where $\mathbf{e}_1$ is a $(p+1) \times 1$ column vector $(1, 0, \ldots, 0)^T$.

In a special case where $p = 0$, we obtain local constant kernel estimator known as the *Nadaraya-Watson estimator*(Nadaraya(1964) and Watson(1964)) expressed explicitly by

$$\hat{m}_{NW}(x; h) = \frac{\sum_{i=n}^{n} K_h(x_i - x)Y_i}{\sum_{i=1}^{n} K_h(x_i - x)}, \tag{2.49}$$

or by

$$\hat{m}_{NW}(x; h) = \sum_{i=1}^{n} W_i(x; h)Y_i, \tag{2.50}$$

where the weights $\{W_i\}_{i=1}^n$ is called the *effective kernel* at x, given by

$$W_i(x; h) = \frac{K_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)}, \ i = 1, \ldots, n. \tag{2.51}$$

Furthermore, if the kernel K is chosen to be the uniform kernel $\bar{K} = \frac{1}{2}\text{I}_{[-1,1]}(x)$, the estimator can be written as

$$\hat{m}_R(x; h) = \frac{\sum_{i=1}^{n} Y_i \text{I}_{[-1,1]}(\frac{x_i - x}{h})}{\sum_{i=1}^{n} \text{I}_{[-1,1]}(\frac{x_i - x}{h})} = \frac{\sum_{i=1}^{n} Y_i \text{I}_{[x-h,x+h]}(x_i)}{\sum_{i=1}^{n} \text{I}_{[x-h,x+h]}(x_i)}. \tag{2.52}$$

At a point x, this estimator gives a *moving average* estimate by taking a local average of Y_i in a neighborhood centered by x and scaled by bandwidth h.(see Horová et al. (2012))

2.4.2 Kernel Functions

The kernel function is constructed in a way that it is related both to the number of its vanishing moments and to the number of existing derivative for the target curve to be estimated.(see Horová et al. 2012) Define $\mu_j(K) = \int x^j K(x)dx$ as the j-th moment of

the kernel K. Then we introduce the same definition for kernel functions used in Horová et al. (2012).

Definition 2.3. *Let ℓ be a nonnegative even integer and assume $\ell \geq 2$. A real-valued function $K \in Lip[-1,1]$, support$(K) = [-1,1]$, satisfying $K \in S_{0,\ell}$, where*

$$S_{\nu,\ell} = \begin{cases} K(1) = K(-1) = 0 \\ \mu_j(K) = \begin{cases} 0, & 0 \leq j < \ell, j \neq \ell \\ 1, & j = 0 \\ \mu_\ell(K) \neq 0, & j = \ell. \end{cases} \end{cases}$$

is called a kernel of order ℓ.

Here $Lip[-1,1]$ denote a class of functions satisfying $|K(x) - K(y)| \leq L|x-y|$, $\forall x, y \in [-1,1]$ for some constant $L > 0$. In addition, in some cases the smoothness of a kernel is also required for kernel estimates. We say the smoothness of a kernel function of order μ over the interval $[-1,1]$ if $K^{(j)}(-1) = K^{(j)}(1) = 0$ for $j = 1, 2, \ldots, \mu$.(see Horová et al.(2012)). Define $C^\mu[-1,1]$ as a set of such functions having μ-times continuous derivatives on $[-1,1]$. Thus, We can define a set $S^\mu_{0,\ell} = S_{0,\ell} \cap C^\mu[-1,1]$. To illustrate, we list some widely used examples from class $S_{0,2}$ in Table 2.1. As for the commonly used

Table 1: Kernel Functions From Class $S_{0,2}$

$K(x) = \frac{1}{2}\mathrm{I}_{[-1,1]}(x)$	uniform kernel	$\mu = -1$
$K(x) = \frac{3}{4}(1-x^2)\mathrm{I}_{[-1,1]}(x)$	Epanechnikov kernel	$\mu = 0$
$K(x) = (1-\lvert x\rvert)\mathrm{I}_{[-1,1]}(x)$	triangle kernel	$\mu = 0$
$K(x) = \frac{15}{16}(1-x^2)^2\mathrm{I}_{[-1,1]}(x)$	quartic kernel	$\mu = 1$

Gaussian kernel, i.e. $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, it doesn't belong to $S_{0,\ell}$ since it has no bounded support.

2.4.3 Asymptotic Error Criterions

Consider an equally spaced fixed model (2.45) with $x_i = i/n$ and $v(x_i) = \sigma^2$ for all i. We first impose some conditions as in Horová et al. (2012):

Condition 2.4.

1. $K \in S_{0,\ell}$
2. $m \in C^{\ell_0}[0,1]$, $\ell_0 > \ell$
3. the bandwidth $h = h_n$ is a non-random sequence satisfying $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = 0$.
4. the point x at which the estimation is taking place is an inner point of the interval $[0,1]$, which means there exists an index n_0 such that $h < x < 1-h$ for all $n \geq n_0$.

Under these assumptions, Koláček(2005) derived for $\ell \geq 2$ the bias and variance of the Nadaraya-Watson estimator, which are given by

$$\text{bias}\{\hat{m}_{NW}(x;h)\} = \frac{\mu_\ell}{\ell!} m^{(\ell)}(x) h^\ell + o(h^\ell) + O(n^{-1}), \tag{2.53}$$

$$\text{Var}\{\hat{m}_{NW}(x;h)\} = \frac{\sigma^2}{nh} R(K) + o((nh)^{-1}). \tag{2.54}$$

From here on we denote $\hat{m}_{NW}(x;h)$ shortly by $\hat{m}(x;h)$ for simplicity. By the usage of the results above and the decomposition property of MSE, some calculations leads us to a global criterion- MISE given by

$$\begin{aligned} \text{MISE}\{\hat{m}(\cdot;h)\} &= E\int_0^1 \{\hat{m}(x;h) - m(x)\}^2 dx = \int_0^1 \text{MSE}\{\hat{m}(x;h)\} dx \\ &= \frac{\sigma^2}{nh} R(K) + \left(\frac{\mu_\ell}{\ell!}\right) R(m^{(\ell)}) h^{2\ell} + o\{h^{2\ell} + (nh)^{-1}\}, \end{aligned} \tag{2.55}$$

$$\tag{2.56}$$

as in Theorem 6.1. in Horová et al.(2012). In order to have the mathematical tractability, we employ the AMISE which is written as

$$\text{AMISE}\{\hat{m}(\cdot;h)\} = \frac{\sigma^2}{nh} R(K) + \left(\frac{\mu_\ell}{\ell!}\right) R(m^{(\ell)}) h^{2\ell}. \tag{2.57}$$

By minimizing the AMISE as in kernel density estimation, i.e.,

$$h_{opt,o,\ell} = \underset{h\in H_n}{\arg\min}\, \text{AMISE}\{\hat{m}(\cdot,h)\} \tag{2.58}$$

with $H_n = [an^{-\frac{1}{2\ell+1}}, bn^{-\frac{1}{2\ell+1}}]$ and $0 < a < b < \infty$, we can obtain the optimal bandwidth given by

$$h_{opt,0,\ell} = \left(\frac{\sigma^2 R(K)(\ell!)^2}{2\ell n \mu_\ell^2(K) R(m^{(\ell)})}\right)^{\frac{1}{2\ell+1}}, \tag{2.59}$$

To avoid the numerical integration in practice, one can instead utilize the *average mean square error* or shortly AMSE, which is defined as

$$\text{AMSE}\{\hat{m}(\cdot, h)\} = \frac{1}{n} E \sum_{i=1}^{n} \{m(x_i) - \hat{m}(x_i; h)\}^2. \tag{2.60}$$

It can be simply estimated by *residual sum of squares*(RSS) given by

$$\text{RSS}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \hat{m}(x_i; h)\}^2, \tag{2.61}$$

Unfortunately this is a biased estimate. For its explanations, interested readers can refer to Theorem 6.2. in Horová et al.(2012).

2.4.4 Bandwidth Selection Methods

Now we can have the discussion about methods for bandwidth selecting. Obviously, the choice of the smoothing parameter in kernel regression plays the same role as in kernel density estimation by affecting the feature of the estimated curve. Although one can choose the bandwidth manually according to subjective judgements, in many situations it is useful to have automatic bandwidth selectors. Here we will mainly focus on a type of plug-in method using Fourier transformation after a brief summary of two other bandwidth selecting approaches using the error function RSS.

Mallow's Method

As we mentioned $\text{RSS}_n(h)$ is a biased estimate of AMSE, in fact it is because its expected value can be written as (see Theorem 6.2 Horová et al. (2012))

$$E(\text{RSS}_n(h)) = \text{AMSE}\{\hat{m}(\cdot; h)\} + \sigma^2 - \frac{2\sigma^2}{n} \sum_{i=1}^{n} W_i(x_i; h). \tag{2.62}$$

By correcting this bias, we consider the error function given by

$$\hat{M}_n(h) = \text{RSS}_n(h) - \hat{\sigma}^2 + \frac{2\hat{\sigma}^2}{n} \sum_{i=1}^{n} W_i(x_i; h), \tag{2.63}$$

where we estimate σ^2 by $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{2n-1} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2. \tag{2.64}$$

Then the estimate $\hat{h}_M$ of the optimal bandwidth is a minimizer of this error function $\hat{M}_n(h)$,

$$\hat{h}_M = \underset{h \in H_n}{\arg\min}\, \hat{M}_n(h). \tag{2.65}$$

Cross-validation Method

As one of the most popular bandwidth selectors, this method is developed by carrying over the main idea of the cross-validation method in kernel density estimation. It is also known as "leave-one-out" method, where we leave out one, say i-th observation, in the regression expression in (2.50):

$$\hat{m}_{-i}(x_i; h) = \sum_{\substack{j=1 \\ j \neq i}}^{n} W_j(x_i; h) Y_j. \tag{2.66}$$

Then by this modified regression estimator $\hat{m}_{-i}(x_i; h)$ we replace $\hat{m}(x_i; h)$ in $\mathrm{RSS}_n(h)$ and obtain the "cross-validation" function given by

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \{\hat{m}_{-i}(x_i) - Y_i\}^2. \tag{2.67}$$

Similarly, the optimal bandwidth is estimated by the minimization of CV(h), i.e.

$$\hat{h}_{\mathrm{CV}} = \underset{h \in H_n}{\arg\min}\, \mathrm{CV}(h). \tag{2.68}$$

Note that $\mathrm{CV}(h)$ is still a biased estimate of AMSE, with $E(CV(h)) = \mathrm{AMSE}\{\hat{m}(\cdot; h)\} + \sigma^2$ (see Theorem 6.3 in Horová et al. (2012)) and in most cases $\hat{h}_{\mathrm{CV}}$ tends to be less than the optimal bandwidth.

Plug-in Method

It is observed by Chiu(1990) that the classical methods based on error function RSS(e.g. Mallows' method) are subject to large sample variation and also give smaller values more frequently than predicted by asymptotic theorems. To overcome this difficulty, Chiu(1990) suggested a procedure which stabilizes RSS by modifying the periodogram of the observations. By applying this procedure, Koláček(2008) propose a type of plug-in method, which produces much more stable bandwidth estimates. Here we will give a brief description of this plug-in method without without delving into details.

To begin with, we suppose a *cyclic design*, i.e., m is assumed to be a smooth periodic function and the estimates are based on the extended series $\tilde{Y}_i = Y_{j+ln}$, for $j = 1, \ldots, n$ and $l = -1, 0, 1$. Similarly, $x_i = i/n$, $i = -n+1, -n+2, \ldots, 2n$. Then the regression estimator can be expressed as

$$\hat{m}(x; h) = \sum_{i=-n+1}^{2n} W_i(x; h) \tilde{Y}_i. \tag{2.69}$$

Denote $\mathbf{Y} = (Y_i, \ldots, Y_n)^T$ as the vectors of observations. We define the periodogram $\mathbf{I}_Y = (I_{Y_1}, \ldots, I_{Y_n})$ of $\mathbf{Y}$ as $I_{Y_j} = |Y_j^-|^2/2\pi n$ for $j = 1, \ldots, n$ where

$$Y_j^- = \sum_{s=0}^{n-1} Y_{s+1} e^{-\frac{i2\pi s(j-1)}{n}} \tag{2.70}$$

is the discrete Fourier transform of the vector $\mathbf{Y}$ and thus can be denoted by $\mathbf{Y}^- = DFT^-(\mathbf{Y})$. Again we denote a vector $w_j^- = (w_1, w_2 \ldots, w_n)$ where

$$w_j = W_1(x_j - 1; h) + W_1(x_j; h) + W_1(x_j + 1; h) \text{ for } j = 1, \ldots, n. \tag{2.71}$$

The application of Parseval's formula leads us to another expression of $\mathrm{RSS}_n(h)$:

$$\mathrm{RSS}_n(h) = \frac{4\pi}{n} \sum_{j=2}^{N} I_{Y_j} \{1 - w_j^-\}^2, \tag{2.72}$$

where $N = [\frac{n}{2}]$ indicates the greatest integer less or equal to $\frac{n}{2}$; $w_j^- = \sum_{s=-n+1}^{n-1} W_1(x_s; h) e^{-\frac{i2\pi s(j-1)}{n}}$ is the discrete Fourier transform of w_j. Define $\tilde{I}_{Y_j}$ as

$$\tilde{I}_{Y_j} = \begin{cases} I_{Y_j}, & 2 \le j < J \\ \hat{\sigma}^2/\pi, & J_1 \le j \le N. \end{cases}$$

where J_1 is the least index such that $I_{Y_{J_1}} < \hat{\sigma}^2/\pi$. Then we can substitute I_{Y_j} in (2.72) with $\tilde{I}_{Y_j}$ and obtain the modified residual sum of squares:

$$\mathrm{MRSS}_n(h) = \frac{4\pi}{n} \sum_{j=2}^{N} \tilde{I}_{Y_j} \{1 - w_j^-\}^2, \tag{2.73}$$

based on which we propose a selector given by

$$\begin{aligned} \tilde{M}_n(h) &= \mathrm{MRSS}_n(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_1 \\ &= \frac{\hat{\sigma}^2}{n} \sum_{j=1}^{n} (w_j^-)^2 + \frac{4\pi}{n} \sum_{j=2}^{J_1 - 1} \{I_{Y_j} - \frac{\hat{\sigma}^2}{2\pi}\} \{1 - w_j^-\}^2. \end{aligned} \tag{2.74}$$

The main idea of the plug-in method is estimating the unknown terms σ^2 and $R(m^{(\ell)})$ in the expression of AMISE((2.57)). We can estimate σ^2 by simply using the formula (2.64). But for $R(m^{(\ell)})$, we need to use the result above. Let J_2 be the last index from $\{1, 2, \ldots, n\}$ for which

$$J_2 \le \frac{\sqrt[\ell+1]{\varepsilon(\ell+1)!}}{2\pi h} \tag{2.75}$$

with $\varepsilon > 0$ and $h \in (0, 1)$. h is a starting approximation of the optimal bandwidth. Usually, taking $\varepsilon = 10^{-3}$ and $h = \ell/n$ yields good results.(see Horová et al.(2012)) Further, we define

$$J = \min\{J_1, J_2 + 1\} \tag{2.76}$$

provided that both conditions for indices J_1 and J_2 are satisfied simultaneously. Thus, by a replacement of the AMISE with the selector (2.74) and some derivations, we obtain an estimate of $R(m^{(\ell)})$ expressed by

$$\widehat{R(m^{(\ell)})} = \frac{4\pi}{n}\sum_{j=1}^{J-2}(2\pi j)^{2\ell}\{I_{Y_{j+1}} - \frac{\hat{\sigma}^2}{2\pi}\}. \tag{2.77}$$

As a result, by plugging this estimate above into (2.59) we will have the plug-in estimator for $h_{opt,0,\ell}$

$$\hat{h}_{PI} = \left(\frac{\sigma^2 R(K)(\ell!)^2}{2\ell n\mu_\ell^2(K)\widehat{R(m^{(\ell)})}}\right)^{\frac{1}{2\ell+1}} \tag{2.78}$$

According to Koláček(2008), the plug-in method could have preferable features to the classical ones since it does not involve any minimization problem of any error function. Besides, computationally it needs far less sample size than classical methods. On the other hand, it has one minor disadvantage due to the requirement of assigning a starting approximation of the unknown smoothing parameter h. Also, plug-in method is limited in a sense that it is only developed for the cyclic design case.

3 Kernel Volatility Density Estimation

In this chapter, we will first describe a class of stochastic volatility models we will be considering. Then we will study the deconvolution procedure for volatility density estimation including its asymptotics. Furthermore, we will propose a relatively different approach called transformed kernel density estimator. Lastly, we will give an example of the same class of stochastic volatility models.

3.1 Stochastic Volatility Models

Denote S_t as the logarithm of the price process for some asset on the financial market. It is common that we model the evolution of log-price as a solution of the following stochastic differential equation,

$$dS_t = b_t dt + \sigma_t dW_t, \quad S_0 = 0 \tag{3.1}$$

where b_t is the drift, W_t is a standard Brownian motion. σ_t is called the volatility process which is independent of W_t by assumption. We model it as a strictly stationary positive diffusion satisfying the mixing condition and the ergodic properties. In addition, we assume that one-dimensional marginal distribution of σ admits a density $\pi(v)$ w.r.t the

Lebesgue on $(0, \infty)$, which is actually the typical case in the literature. So we model the volatility process σ_t by another stochastic differential equation in term of V_t with $V_t = \sigma_t^2$ (or it can also be in $\log \sigma_t^2$). It is given by

$$dV_t = b(V_t)dt + a(V_t)dB_t, \quad V_0 = \eta \tag{3.2}$$

where a and b are real-valued continuous function on $\mathbb{R}$ and η is a positive number. B_t is a standard Brownian motion and independent of W_t.

In summary, in this thesis we will consider a simplified version of this two-dimensional diffusion (S_t, V_t) (refraining the drift term b_t from equation (3.1)). It is given by

$$\begin{cases} dS_t = \sigma_t dW_t, & S_0 = 0 \\ dV_t = b(V_t)dt + a(V_t)dB_t, & V_0 = \eta \end{cases} \tag{3.3}$$

3.2 Deconvolution Kernel Volatility Density Estimation

In this section, we will first construct the deconvolution kernel volatility density estimator using the theory of classical deconvolution kernel estimator. Then we will investigate the asymptotic behavior of this estimator under a couple of mixing conditions on the volatility process.

3.2.1 Construction of The Estimator

Assume that log-price S_t is observed discretely at regular time instant $0, \Delta, 2\Delta, \ldots, n\Delta$ such that $\Delta \to 0$ and $n\Delta \to \infty$. In other words, it is assumed that as the number of observation n tends to infinity, the interval Δ tends to zero and the total length of the observation time tends to infinity. For $i = 1, 2, \ldots, n$, we work with normalized increments as in Genon-Catalot et al.(2000)

$$X_i^\Delta = \frac{1}{\sqrt{\Delta}}(S_{i\Delta} - S_{(i-1)\Delta}). \tag{3.4}$$

We could see that X_i^Δ is the normalized log-return of the stock price. For small Δ, roughly we have the approximation

$$\begin{aligned} X_i^\Delta &= \frac{1}{\sqrt{\Delta}} \int_{(i-1)\Delta}^{i\Delta} \sigma_t dW_t \\ &\approx \sigma_{(i-1)\Delta} \frac{1}{\sqrt{\Delta}} (W_{i\Delta} - W_{(i-1)\Delta}) \\ &= \sigma_{(i-1)\Delta} Z_i^\Delta \end{aligned} \tag{3.5}$$

where

$$Z_i^{\Delta} = \frac{1}{\sqrt{\Delta}}(W_{i\Delta} - W_{(i-1)\Delta}), \text{ for each } i.$$

Since W_t is standard Brownian Motion, Z_1^{Δ},Z_2^{Δ},..., Z_n^{Δ} are i.i.d. standard normal random variables. Moreover, the sequence of Z_i^{Δ} are assumed to be independent of the volatility process. By taking the logarithm of the square of equation X_i^{Δ}, we have the desirable convolution structure

$$\log(X_i^{\Delta})^2 \approx \log \sigma^2_{(i-1)\Delta} + \log(Z_i^{\Delta})^2. \tag{3.6}$$

We assume that the approximation is accurate enough that we can use the approximate structure to estimate the density of $\log \sigma^2_{(i-1)\Delta}$ based on the observations of $\log(X_i^{\Delta})^2$. Denote g as the density of error variable $\log(Z_i^{\Delta})^2$ and φ_g as its characteristic function. Since for each i $Z_i^{\Delta} \sim N(0,1)$, according to Van ES et al.(2009) the density function g is given by

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x} e^{-\frac{1}{2}e^x} \tag{3.7}$$

and the characteristic function by

$$\varphi_g = \frac{1}{\sqrt{\pi}} 2^{it} \Gamma\left(\frac{1}{2} + it\right), \tag{3.8}$$

where $\Gamma(\cdot)$ is the gamma function for a complex number. We plot the density function g

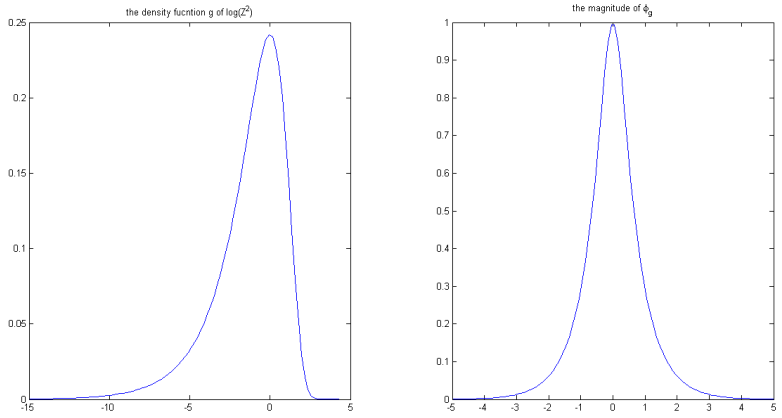

Figure 1: Left: density function g; Right: modulus of φ_g

along with the modulus of its characteristic function φ_g in Figure 1. Using the classical deconvolution theory which we reviewed in section 2.3, we can write the deconvolution kernel volatility density estimator as

$$\hat{f}(x;h) = \frac{1}{nh} \sum_{j=1}^{n} v_h\left(\frac{x - \log(X_j^{\Delta})^2}{h}\right), \tag{3.9}$$

where

$$v_h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\varphi_K(s)}{\varphi_g(s/h)}e^{-isx}ds \tag{3.10}$$

and φ_K denoted as the characteristic function of the kernel K. It can be verified that v_h and therefore $\hat{f}(x;h)$ is a real-valued function.

3.2.2 Asymptotics of The Estimator

It is proven in Lemma 4.1 of Van Es et al.(2003) that the expectation of the deconvolution kernel estimator is the same as that of the ordinary kernel estimator, thus their biases are also the same. Also it is known that the variance of the deconvolution estimator depends heavily on the smoothness of the error distribution g.(see Van Es et al. 2001) About its characteristic function φ_g, we have the following equality from Van Es et al.(2009)

$$|\varphi_g(t)| = \sqrt{2}e^{-\frac{1}{2}\pi|t|}\left(1+O(\tfrac{1}{|t|})\right), \quad \text{as } t\to\infty. \tag{3.11}$$

This means the error distribution is super smooth according to the definition (2.2). Besides, the tail of the characteristic function φ_g is similar to the tail of a Cauchy characteristic function. Thus, we can expect the same logarithmic rate of convergence as in a deconvolution problem with a Cauchy error from i.i.d observations. However, this argument is not directly applicable in our case, since the i.i.d assumption has been relaxed. In our deconvolution problem

$$\log(X_i^{\Delta})^2 \approx \log\sigma^2_{(i-1)\Delta} + \log(Z_i^{\Delta})^2,$$

$\log\sigma^2_{(i-1)\Delta}$ is a strictly stationary sequence satisfying the strong mixing condition and independent of the i.i.d noise sequence $\log(Z_i^{\Delta})^2$. Our task here is to expand the bias and the variance of the deconvolution estimator under a couple of mixing conditions.
To begin with, we list the definitions and properties of the mixing conditions we will use according to Bradley(2005). In a probability space$(\Omega,\mathcal{F},P)$, $\mathcal{A},\mathcal{B}$ are two σ-field included in $\mathcal{F}$. The mixing coefficients α,β,ϕ and ρ are defined by

$$\alpha(\mathcal{A},\mathcal{B}) = \sup|P(A\cap B)-P(A)P(B)| \quad A\in\mathcal{A},B\in\mathcal{B} \tag{3.12}$$

$$\beta(\mathcal{A},\mathcal{B}) = \sup\tfrac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}|P(A_i\cap B_j)-P(A_i)P(B_j)| \quad A\in\mathcal{A},B\in\mathcal{B} \tag{3.13}$$

$$\phi(\mathcal{A},\mathcal{B}) = \sup|P(A|B)-P(A)| \quad A\in\mathcal{A},B\in\mathcal{B},P(B)>0 \tag{3.14}$$

$$\rho(\mathcal{A},\mathcal{B}) = \sup|corr(X,Y)|, \quad X\in\mathcal{L}^2_{\text{real}}(\mathcal{A}),Y\in\mathcal{L}^2_{\text{real}}(\mathcal{B}) \tag{3.15}$$

where the second supremum is taken over all pairs of partitions $\{A_1,\ldots,A_I\}$ and $\{B_1,\ldots,B_J\}$ of Ω such that $A_i\in\mathcal{A}$ for each i and $B_j\in\mathcal{B}$ for each j. In the last supremum, $\mathcal{L}^2_{\text{real}}(\mathcal{A})$

is denoted as the space of square-integrable, $\mathcal{A}$-measurable and real-valued random variables. Also we have the following inequalities hold,

$$2\alpha(\mathcal{A},\mathcal{B}) \leq \beta(\mathcal{A},\mathcal{B}) \leq \phi(\mathcal{A},\mathcal{B}) \leq 1,$$

$$4\alpha(\mathcal{A},\mathcal{B}) \leq \rho(\mathcal{A},\mathcal{B}) \leq 1. \tag{3.16}$$

Suppose $X = (X_t, t \in \mathbb{R}^+, \text{or } t \in \mathbb{N})$ is a strictly stationary stochastic process, for $-\infty \leq J \leq L \leq \infty$ we define the σ-field

$$\mathcal{F}_J^L = \sigma(X_t, J \leq t \leq L). \tag{3.17}$$

Then $\alpha(t), \beta(t), \phi(t), \rho(t)$ can be defined as

$$c(t) = \sup c(\mathcal{F}_{-\infty}^0, \mathcal{F}_t^\infty) \tag{3.18}$$

with $c = \alpha, \beta, \phi$ or ρ. A process is called c-mixing if $c(t) \to 0$ as $t \to \infty$. Moveover, from the two relations (3.16), we have the fact that β-mixing, ϕ-mixing and ρ-mixing all imply strong mixing or α mixing.

Now we turn to our target. First we impose the following conditions on the continuity and the mixing condition of the volatility process σ_t^2 as well as on the kernel function K and its characteristic function φ_K.

Condition 3.1.
1. σ_t^2 is L^1-Hölder continuous of order one-half, i.e. $E|\sigma_t^2 - \sigma_0^2| = O(t^{\frac{1}{2}})$ for $t \to 0$.
2. σ_t^2 is strongly mixing with coefficient $\alpha(t)$ satisfying

$$\int_0^\infty \alpha(t)^q dt < \infty \quad \text{for some } 0 < q < 1 \tag{3.19}$$

Condition 3.2.
1. K is a real symmetric function satisfying

$$\int_{-\infty}^\infty |K(u)|du < \infty, \quad \int_{-\infty}^\infty K(u)du = 1, \quad \int_{-\infty}^\infty u^2|K(u)|du < \infty \tag{3.20}$$

2. φ_K is a real-valued symmetric function with bounded support $[-1,1]$ and satisfies $\varphi_K(1-t) = At^\xi + o(t^\xi)$ for some $\xi > 0$,$A \in \mathbb{R}$

One example of such kernel function is taken from Wand(1998), which is

$$K(x) = \frac{48x(x^2-15)\cos(x) - 144(2x^2-5)\sin(x)}{\pi x^7}, \tag{3.21}$$

with its characteristic function given by

$$\varphi_K(t) = (1-t^2)^3, \quad |t| \leq 1. \tag{3.22}$$

We will present the following theorem by Van Es et al.(2009) which concerns the mean square error of the estimator at a fixed point x. Note that although the result is based on the simplified model we are considering, it also applies to the original model which contains the drift term b_t. For its proof and deriving, interested readers can refer to Van Es(2003).

Theorem 3.3. *Assume $E[b_t^2]$ is bounded. Suppose the process σ satisfy Condition 3.1 and the kernel function K satisfy the condition 3.2. Moveover, the density f of $\log \sigma_t^2$ is assumed to be twice continuously differentiable with a bounded second derivation. In addition, let the density of σ_t^2 to be bounded in a neighborhood of zero. Assume that $\Delta = n^{-\delta}$ for given $0 < \delta < 1$ and set $h = \gamma\pi/\log n$, where $\gamma > 4/\delta$. Then the bias of the estimator* (3.9) *satisfies*

$$E\hat{f}(x;h) - f(x) = \frac{1}{2}h^2 f''(x) \int u^2 K(u)du + o(h^2), \tag{3.23}$$

whereas the variance of the estimator satisfies the order bounds

$$\mathrm{Var}\hat{f}(x;h) = O(\frac{1}{n}h^{2\xi}e^{\pi/h}) + O(\frac{1}{nh^{1+q}\Delta}). \tag{3.24}$$

Remark 3.4 Choose $\Delta = n^{-\delta}$ with $0 < \delta < 1$ and $h = \gamma\pi/\log n$ with $\gamma > 4/\delta$. By some elementary computations, we have the order bounds of the variance

$$\mathrm{Var}\hat{f}(x;h) = O(n^{-1+\frac{1}{\gamma}}(\log n)^{-2\xi}) + O(n^{-1+\delta}(\log n)^{1+q}). \tag{3.25}$$

Since $\gamma > 4/\delta$ results in $\frac{1}{\gamma} < \delta/4 < \delta$, the second term in the above dominates the first term. This means the variance is of order $n^{-1+\delta}(\log n)^{1+q}$. Obviously, the bias is of order $(\log n)^{-2}$, which dominates the variance. Therefore, the mean square error is of the same order as the squared bias, which is $(\log n)^{-4}$.

Remark 3.5 We can obtain a better bound for the variance under a stronger mixing condition, say uniform mixing with mixing coefficient ϕ as defined previously. As we know, uniform mixing implies strong mixing and we have the relation $\alpha \leq \frac{1}{2}\phi(t)$. If we instead assume σ_t^2 is uniform with coefficient $\phi(t)$ satisfying $\int_0^\infty \phi(t)^{\frac{1}{2}}dt < \infty$ in the Theorem 3.3, we will have the following variance bound given by

$$\mathrm{Var}\hat{f}(x;h) = O(\frac{1}{n}h^{2\xi}e^{\pi/h}) + O(\frac{1}{nh\Delta}). \tag{3.26}$$

However, it turns out that the variance bound cannot be improved by this stronger assumption on σ_t^2. Hence, the order of MSE stays unchanged.

Remark 3.6 There are many examples of such a stochastic volatility model which

belongs to the same class of the model (3.3) and satisfy a mixing condition. They are Ornstein-Uhlenbeck process proposed by Wiggins(1987), GARCH(1,1)-M by Nelson(1990) and Heston model suggested by Heston 1993. By proper choices of model parameters, these continuous-time stochastic models ensure an ergodic stationary solution satisfying ρ-mixing for the volatility process(see Genon-Catalot 2000). Note that in these models the assumption on $\alpha(t)$ in Theorem 3.3 still holds.

3.3 Transformed Kernel Volatility Density Estimation

In this section, we will propose a transformed density estimator along with parameter selectors for its parameters using theory of continuous-time kernel density estimation and kernel regression. After that, we will apply it to estimate the volatility density.

3.3.1 Transformed Kernel Density Estimator

Let us consider the convolution structure

$$Y_i = X_i + Z_i \quad i = 1, 2, \ldots, n. \tag{3.27}$$

where $\{X_i\}_{i=1}^n$ is a discrete random sample with a regular time interval Δ from a continuous-time stationary ergodic process having a marginal density f. The noises $\{Z_i\}_{i=1}^n$ are assumed to be identically and independently distributed with a zero mean and constant variance. In addition, for each i, Z_i is independent of X_i. If the error terms have non-zero mean, we can easily correct it by subtract both sides of (3.27) by $E(Z_i)$. The transformed kernel estimator consists of two steps.

On the first step, we estimate X_i by $\hat{X}_i$ in the same style as a non-parametric regression technique called moving average (see Takezawa(2006)). It is the most typical technique for smoothing one-dimensional equally spaced data and commonly used to determine a rough trend of the time-series behavior of the target variable. Thus, we refer to $\hat{X}_i$ as **moving average estimator**. Suppose that the width of the window upon which we take the average is $2\tau + 1$ for a nonnegative integer τ. That is, the estimate of X_i is the average of $(2\tau+1)$ data points $\{Y_j\}((i-\tau) \leq j \leq (i+\tau))$. Thus, the resulting estimator $\hat{X}_i$ is given by

$$\hat{X}_i = \frac{1}{2\tau + 1} \sum_{j=i-\tau}^{i+\tau} Y_j. \tag{3.28}$$

or by

$$\hat{X}_i = \sum_{j=1-\tau}^{n+\tau} w_{ij} Y_j, \tag{3.29}$$

where

$$w_{ij} = \begin{cases} \frac{1}{2\tau+1} & \text{if } -\tau \leq (i-j) \leq \tau \\ 0 & \text{otherwise.} \end{cases}$$

For the estimates close to two ends, we adopt the reflection boundary condition, which is

$$Y_0 = Y_1, \quad Y_{-1} = Y_2, \quad \ldots, \quad Y_{-\tau+1} = Y_\tau; \tag{3.30}$$

$$Y_{n+1} = Y_n, \quad Y_{n+2} = Y_{n-1}, \quad \ldots, \quad Y_{n+\tau} = Y_{n-\tau+1}. \tag{3.31}$$

Now let us take a look at the conditional mean and variance of this estimator $\hat{X}_i$. Denote the window size $k = 2\tau + 1$ and $X = \{X_i\}_{i=1}^n$. For a fixed i,

$$\begin{aligned} E\left(\hat{X}_i|X\right) &= E\left(\sum_j Y_j/k|X\right) = E\left(\sum_j (X_j + Z_j)/k|X\right) \\ &= \frac{1}{k}E\left(\sum_j X_j|X\right) + \frac{1}{k}E\left(\sum_j Z_j|X\right) \\ &= \frac{1}{k}\sum_j X_j + \frac{1}{k}E\left(\sum_j Z_j\right) \\ &= \frac{1}{k}\sum_j X_j + E\left(Z_i\right) \end{aligned}$$

$$\text{bias}(\hat{X}_i|X) = E\left(\hat{X}_i|X\right) - X_i = \frac{1}{k}\sum_j X_j - X_i + E\left(Z_i\right) \tag{3.32}$$

$$\begin{aligned} \text{Var}(\hat{X}_i|X) &= \text{Var}\left(\frac{1}{k}\sum_j Y_j|X\right) = \text{Var}\left(\frac{1}{k}\sum_j (X_j + Z_j)|X\right) \\ &= \frac{1}{k^2}\text{Var}\left(\sum_j Z_j\right) = \frac{k}{k^2}\text{Var}(Z_i) = \frac{\text{Var}(Z_i)}{k} \end{aligned} \tag{3.33}$$

That is to say, the conditional variance of $\hat{X}_i$ is equal to the variance of Z_i divided by k and the conditional bias equal to the sum of $\frac{1}{k}\sum_j X_j - X_i$ and the mean of Z_i. This indicates that conditioned on X, a rise in window size k reduces the variance of $\hat{X}_i$ but increases its bias. Thus, the selection of window size is a *bias-variance trade off*.

Denote $\hat{X} = \{\hat{X}_i\}_{i=1}^n$. The second step is to utilize the ordinary kernel density estimator on the estimate $\hat{X}$ from step one. By applying the formula (2.3), we would have the **transformed kernel density estimator**

$$\hat{f}_X(x;u) = \frac{1}{n}\sum_{i=1}^n K_u(x - \hat{X}_i) = \frac{1}{n}\sum_{i=1}^n K_u(x - \sum_{j=i-\tau}^{i+\tau} Y_j), \tag{3.34}$$

where u is the smoothing parameter. For this estimator, the choices of the parameters τ and u are of great importance. Thus, we will investigate how to select them in the following.

3.3.2 Parameter Selectors

We will also analyze how to choose the parameters in two steps, where we first choose the parameter τ for the moving average estimator. It can be shown that the moving average estimator above is approximately equivalent to a Nadaraya-Watson estimator with the uniform kernel $\bar{K}(x)$ based on observations $\{(i, Y_i)\}_{i=1}^n$. We first transform the observations to $\{(i/n, Y_i)\}_{i=1}^n$ in order to have a equally spaced fixed design over $[0, 1]$. Then according to the formula (2.52), we have the Nadaraya-Watson estimator written as

$$\hat{m}_R(\frac{i}{n}; h) = \frac{\sum_{j=1}^{n} Y_j \mathrm{I}_{[\frac{i}{n}-h \le \frac{j}{n} \le \frac{i}{n}+h]}}{\sum_{j=1}^{n} \mathrm{I}_{[\frac{i}{n}-h \le \frac{j}{n} \le \frac{i}{n}+h]}} = \frac{\sum_{j=1}^{n} Y_j \mathrm{I}_{[i-nh \le j \le i+nh]}}{\sum_{j=1}^{n} \mathrm{I}_{[i-nh \le j \le i+nh]}} \tag{3.35}$$

$$\approx \frac{1}{2[nh]+1} \sum_{j=i-[nh]}^{i+[nh]} Y_j = \hat{X}_i \tag{3.36}$$

where $[nh]$ indicates the greatest integer less or equal to nh. As we mentioned previously, the uniform kernel belongs to $S_{0,2}^{-1}$, which satisfies the first assumption in Condition 2.4. In addition we assume that its second assumption is also fulfilled. So if we choose the Fourier-type plug-in method for its computational advantages, we can estimate the parameter τ by $[n\hat{h}_{PI}]$, where $\hat{h}_{PI}$ is the optimal bandwidth estimate generated using the plug-in method (2.78) with $\ell = 2$. Therefore, we have one estimate of the window size k given by

$$\hat{k} = 2[n\hat{h}_{PI}] + 1 \tag{3.37}$$

In the second step, as $\hat{X} = \{\hat{X}_i\}_{i=1}^n$ is a series of moving average estimates, it is sensible for us to consider it as a continuous-time sample $\{\hat{X}_t, t \in [0, T]\}$ with $T = n\Delta$. That is to say,

$$\hat{f}_X(x; u) = \frac{1}{n} \sum_{i=1}^{n} K_u(x - \hat{X}_i) \approx \frac{1}{T} \int_0^T K_u(\hat{X}_t - x) dx. \tag{3.38}$$

If we assume that the kernel function K and the target density f satisfy all the assumptions in Condition 2.2., then we can estimate the bandwidth u by the continuous-time bandwidth selector $h(T)$ in (2.22):

$$h(T) = \left[\frac{2 f_{X_0'}(0)}{\mu_2(K)^2 R(f'') T} \right]^{1/4} \tag{3.39}$$

If we choose the uniform kernel $\bar{K}$, we can obtain $\mu_2(K) = 1/3$ by some simple calculations. To estimate the total curvature measure $R(f'')$, we assume f is normal and use the formula(in Wand and Jones(1995)) which is given by

$$R(f^{(s)}) = (-1)^s \int f^{(2s)}(x) f(x) dx. \tag{3.40}$$

Define $\psi_r = \int f^{(r)}(x) f(x) dx$. According to Wand and Jones(1995), given a normal density f with variance σ^2, for r even we have

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}. \tag{3.41}$$

Thus, some elementary computation yields $R(f'') = 3/(8\hat{\sigma}^5\pi^{1/2})$ with $\hat{\sigma}$ equal to the corrected sample standard deviation of $\hat{X}$ given by

$$\text{Std}(\hat{X}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{X}_i - (\sum_{i=1}^{n}\hat{X}_i)/n)^2} \tag{3.42}$$

which is an estimate of σ. As for $f_{X_0'}(0)$, i.e. the value of marginal density of X_0' at 0, we first approximate the derivatives by $\hat{X}_i' = \frac{\hat{X}_{i+1} - \hat{X}_i}{\Delta}$. Denote $\hat{X}' = \{\hat{X}_i'\}_{i=1}^n$. Then we estimate $f_{X_0'}$ by $\frac{1}{\hat{\sigma}'}\phi(0)$, where $\hat{\sigma}' = \text{Std}(\hat{X}')$ and ϕ is the normal probability density function. To sum up, we have the final expression of the bandwidth estimator written as

$$\hat{u} = \left[\frac{16\pi^{1/2}\phi(0)\hat{\sigma}^5}{\hat{\sigma}'\Delta n}\right] \tag{3.43}$$

3.3.3 Transformed Kernel Volatility Density Estimator

Now we turn to our problem

$$\log(X_i^\Delta)^2 \approx \log\sigma^2_{(i-1)\Delta} + \log(Z_i^\Delta)^2.$$

For notation simplicity, we write $\log\sigma^2_{i\Delta}$ instead of $\log\sigma^2_{(i-1)\Delta}$. Before following the two-step estimator procedure described previously, we first examine if the mean of the error terms $\{\log(Z_i^\Delta)^2\}_{i=1}^n$ is equal to zero or not. We find numerically the mean is approximately equal to -1.27. So by applying the formula (3.34), we obtain the transformed kernel volatility density estimator expressed as

$$\hat{f}(x;u) = \frac{1}{n}\sum_{i=1}^{n} K_u(x - \widehat{\log\sigma^2_{i\Delta}}) \tag{3.44}$$

where

$$\widehat{\log\sigma^2_{i\Delta}} = \frac{1}{k}\sum_{j=i-\tau}^{i+\tau}\log(X_j^\Delta)^2 + 1.27 \tag{3.45}$$

Assume that all the assumptions are satisfied in the Condition 2.2. and the Condition 2.4. Also suppose the normality assumption for the target density is appropriate. Then we can use the parameter selectors (3.37) and (3.43) in the previous section to estimate the k and u here.

3.4 An Example of The Stochastic Volatility Model

In this section, we will outline the basic dynamics of Heston Model and summarize a few facts about the model parameters and the process involved. We also discuss the monte carlo simulation of the model mainly about its discretization schemes.

3.4.1 The Model

The Heston model proposed by Heston(1993) is defined by the following two stochastic differential equations:

$$dS_t = \sigma_t dW_t^1, \quad S_0 = 0 \tag{3.46}$$

$$dV_t = \kappa(\theta - V_t)dt + \lambda\sqrt{V_t}dW_t^2, \quad V_0 = \eta \tag{3.47}$$

$$dW_t^1 dW_t^2 = \rho dt, \quad \rho \in [-1, 1]$$

where S_t is the logarithm of the price process and $V_t = \sigma_t^2$ is the volatility process. κ, θ, λ and η are strictly positive constants and W^1, W^2 are scalar Brownian Motion in some probability measure. It is easy to see from the equations that Heston model belong to the model class (3.3). The Heston model is related to the square root process, which was initially used to study the term structure of interest rate by Cox, Ingersol and Ross(1985). A square root process is appealing for modeling interest rates and volatility since it can never become negative given an initial nonnegative value, see e.g. Feller(1951). The Heston Model is now used as an extension of Black-Scholes Model to incorporate stochastic volatility and often for derivative pricing e.g. exotic options. Its increasing popularity is mainly because of its main features:

- The volatility process is mean-reverting given $\kappa > 0$;
- The model can reproduce a smile-like implied volatility curve, similar to the market one;
- The existence of semi-analytical solution for European Option.

To know the Heston model better, let us look at the parameters in the variance process

- κ- mean-reverting speed ($\kappa > 0$)
- θ- The long-term level of variance V_t
- λ- the volatility of V_t, often referred to volatility of volatility
- ρ- The correlation between the Brownian Motion driving the stock price process S_t and the Brownian Motion driving the volatility process V_t
- V_0- The starting value of the variance process

These parameters follow the proposition below: see e.g. Andersen(2007)

Proposition *Assume that $V_0 = 0$. If $2\kappa\theta \geq \lambda^2$ then the process for V_t can never reach zero. If $2\kappa\theta < \lambda^2$, the origin is accessible and strongly reflecting.*

In typical cases, $2\kappa\theta$ is often much less than λ^2. That is to say, it is quite likely for V_t to touch zero. By the origin being accessible and strongly reflecting, we mean the variance can touch zero but leave immediately. Also noteworthy is the fact that the process of the stock price and volatility process are negatively correlated, i.e. market volatility increases when stock prices go down. Therefore, the parameter ρ tends to be negative. However, in this work we study the special case when ρ is equal to zero.
According to Genon-Catalot et al.(2000), there is another important fact about the volatility process. Set $a = 2\kappa\theta/\lambda^2$, $\mu = 2\kappa/\sigma^2$. If $\mu > 0$ and $a \geq 1$ (i.e.$\kappa > 0$, $2\kappa\theta > \lambda^2$), the stationary distribution π for V has density

$$\pi(v) = \frac{\mu^a}{\Gamma(a)} v^{a-1} e^{-\mu v}, \quad v > 0 \tag{3.48}$$

This is gamma distribution with parameter (a, μ), where a is the scale parameter, μ is the rate parameter.

3.4.2 Monte Carlo Simulation

Now our task is to find a suitable discretization scheme for Heston Model. There are plenty of ways biased or exact, among which Euler scheme is the most simple and straightforward method to implement. It takes the following form

$$\begin{cases} S_{t+\Delta} = S_t + \sqrt{V_t \Delta} W_1 \\ V_{t+\Delta} = V_t + \kappa(\theta - V_t)dt + \lambda\sqrt{V_t \Delta} W_2 \end{cases} \tag{3.49}$$

where we replace σ_t by $\sqrt{V_t}$, $W_1 \sim N(0,1)$, $W_2 = \rho W_1 + \sqrt{1-\rho^2}W_3$ and $W_3 \sim N(0,1)$. Unfortunately, it gives rise to numerical problems. While the variance process itself is guaranteed to be nonnegative, the discretization is not. Specifically, given $V_t > 0$, for any choice of time step the probability of the variance becoming a negative value at the next time step is strictly greater than zero. That is to say,

$$P(V_{t+\Delta} < 0) = \Phi(\frac{-V_t - \kappa(\theta - V_t)\Delta}{\lambda\sqrt{V_t\Delta}}) > 0 \tag{3.50}$$

where Φ is the standard normal cumulative distribution function. Moreover, negative variance makes the computation of $\sqrt{V_t}$ fail. Therefore, when using Euler Scheme, one has to carefully think about how to fix the negative variance.

Practitioners have often opted for a quick 'fix' by either setting the process equal to zero when it attains a negative value or reflecting it in the origin and continuing from there on. These fixes are referred to as absorption and reflection by Gatheral(2006). In comparison, Lord et al.(2008) proposed a slightly different approach to fix the problem which is so-called full truncation scheme. It is given by

$$\begin{cases} S_{t+\Delta} &= S_t + \sqrt{V_t^+\Delta}W_1 \\ V_{t+\Delta} &= V_t + \kappa(\theta - V_t^+)dt + \lambda\sqrt{V_t^+\Delta}W_2 \end{cases} \tag{3.51}$$

where $V_t^+ = \max(V_t, 0)$, $W_1 \sim N(0,1)$, $W_2 = \rho W_1 + \sqrt{1-\rho^2}W_3$ and $W_3 \sim N(0,1)$.

According to Andersen(2007), the main feature of full truncation scheme is that the process V is allowed to go below zero, at which point the process V become deterministic with an upward drift of $\kappa\theta$. Additionally, Lord et al.(2008) found that the full truncation scheme outperformed most biased scheme like absorption and reflection Euler schemes in terms of bias and root-mean-square error(RMSE). Moveover, when the volatility of the volatility(λ) is not too high, it has relatively smaller bias and RMSE error even than some exact schemes(e.g. quasi-second by Kahl and Jäckel(2006)) given a certain computational budget. Thus, we choose the full truncation scheme as our discretization scheme for the simulation.

4 Numerical Results

This chapter will be devoted to apply deconvolution kernel volatility estimator and transformed kernel volatility estimator on both simulated data from Heston model and real data. We will also compare their performance based on each data set.

4.1 Simulations

Given two different time intervals, we will generate two sets of data from Heston model which we refer to high frequency and low frequency data. Then we will utilize both methods on these data.

4.1.1 High Frequency Data

We choose the size of the simulation n equal to 1.8×10^5 and Δ equal to $1/(250 \times 24)$, which corresponds to a sample of 30yrs of hourly data. Suppose that the Δ is small enough for the approximation (3.6) to hold. We consider the parameters of Heston model listed in Table 2, which are fitted parameters for S&P 500 with annualized value. They satisfy the conditions $\kappa > 0$ and $2\kappa\theta \geq \lambda^2$ such that the process σ^2 has a Gamma

Table 2: The Parameter Set

	κ	θ	λ	ρ	V_0
parameter set	3.46	0.008	0.14	0	0.007569

Note: this parameter set is taken from Table 7 of Broadie and Kaya(2006) except for setting $\rho = 0$.

distribution $\Gamma(a, \mu)$, where $a = 2.8245$ and $\mu = 353.0612$. By changing the coordinate we can have the density of $\log \sigma^2$, which can be written as

$$f_{\log \sigma^2} = e^x \pi(e^x) \tag{4.1}$$

where $\pi(\cdot)$ is the density function of the gamma distribution, which is given in equation (3.48). This would be assumed as the true density of $\log \sigma^2$ to compare with. Then we use the full truncation Euler scheme(equation (3.51)) to simulate a sample path $\{S_i\}_{i=1}^n$ of log-price S and transform it to a sample path of $\log(X_i^\Delta)^2$, where

$$\log(X_i^\Delta)^2 = \log(\frac{1}{\sqrt{\Delta}}(S_{i\Delta} - S_{(i-1)\Delta}))^2 \tag{4.2}$$

After that, we first apply the transformed kernel density estimator on this transformed sample path $\{\log(X_i^\Delta)^2\}_{i=1}^{n-1}$. To investigate the impact of the choice of window size k, we fix the bandwidth u equal to 0.3 and vary the window size $k = 481, 721, 961, 1081$, which gives us the result in Figure 2. Similarly, we also set $k = 721$ while varying the bandwidth $u = 0.2, 0.25, 0.28, 0.3$, which renders to the resulting Figure 3. Note that in these figures and all the figures based on simulations below the dashed line stands for the corresponding true density. From these two figures we can see that the larger the window size, the higher the peak of the estimated density. In contrast, the bigger the bandwidth, the lower the peak.

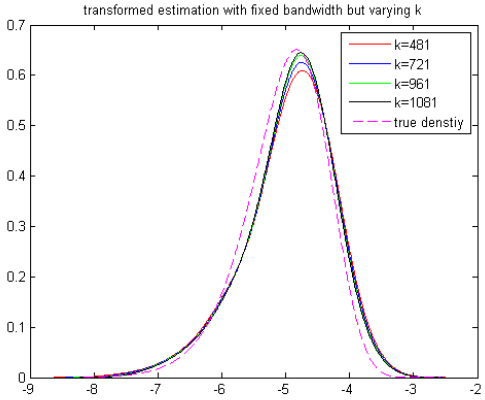

Figure 2: Transformed kernel estimation with fixed $u = 0.3$ and varying k(high)

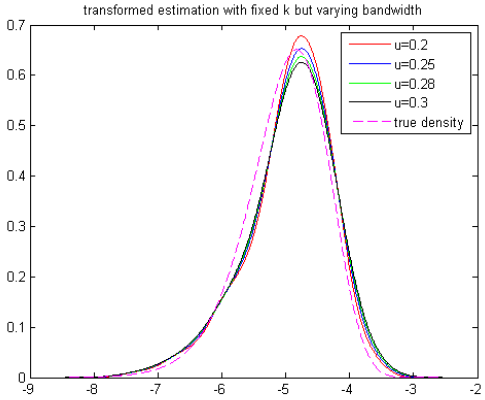

Figure 3: Transformed kernel estimation with $k = 721$ and varying u(high)

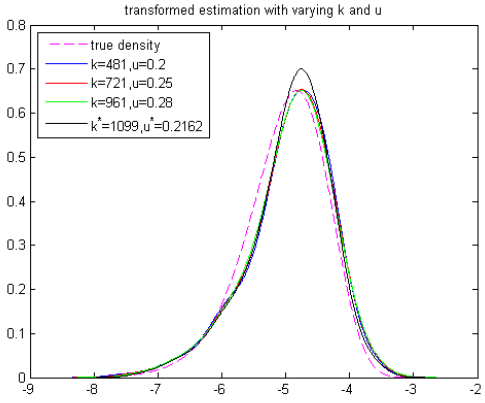

Figure 4: Transformed kernel estimation with varying k and u(high)

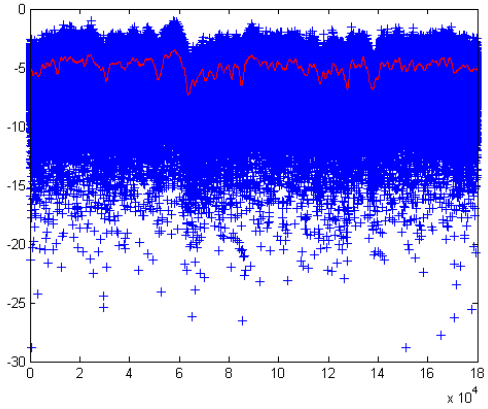

Figure 5: Moving average estimate curve(high)

According to the impacts of both parameters, we change k and u manually according to its fit to the true density and find three approximately optimal estimated densities. Their corresponding values of parameters are $k = 481, u = 0.2$; $k = 721, u = 0.25$; $k = 961, u = 0.28$. On the other hand, we also apply the automatic parameter selectors $\hat{k} = 2[n\hat{h}_{PI}] + 1$ and $\hat{u}$ we described previously, where we implement the Matlab toolbox attached with Horová et al. (2012) for the computation of $\hat{h}_{PI}$. As a result, we find the optimal window size $k^* = 1099$ and bandwidth $u^* = 0.2047$ respectively. The resulting estimated densities from both manual selections and automatic parameter selectors are all plotted in Figure 4. It is obvious that the three estimated densities using manually picked parameters are all good fits and they are close to each other, meaning the optimal value of the parameter pair is not unique. However, the one with parameter selectors is slightly different with a taller peak. We also plot the moving average estimates $\{\widehat{\log \sigma^2_{i\Delta}}\}_{i=1}^{n-1}$(red line) along with the sample $\{\log(X_i^\Delta)^2\}_{i=1}^{n-1}$(blue $'+'$) in Figure 5.

For deconvolution method, we choose (3.21) as our kernel function and utilize the fast Fourier transform in section 2.3.4 for the computation. Concerning its bandwidth selection, Theorem 3.3 suggests a theoretical optimal bandwidth $h_T = \gamma\pi/\log n$ with $\gamma > 4/\delta$ and $\Delta = n^{-\delta}$. This results in the inequalities $-\log\Delta = \delta\log n$ and $h_T > 4\pi/(\delta\log n) = 4\pi/(-\log\Delta)$. If we plug in $\Delta = 1.6667 \times 10^{-4}$ in our case, we will have $h_T > 1.445$. Of course, this implied value of bandwidth cannot be useful here, unless we have an extremely small time interval. To illustrate, in order to have $h_T > 0.5$ we at least need the time interval $\Delta = 1.2162 \times 10^{-11}$. However, such a data sample is hardly obtainable in practice. Thus, instead we select the bandwidth manually by looking at several densities over a range of different values of bandwidth and choosing the most appropriate one by subjective judgements. More specifically, we start with a larger bandwidth and gradually decrease the amount of smoothing until the density estimate appear not too wiggly. Here we try different values of the bandwidth

$h = 0.2, 0.25, 0.27, 0.3$ and we pick the optimal $h^* = 0.2$ shown in Figure 6.

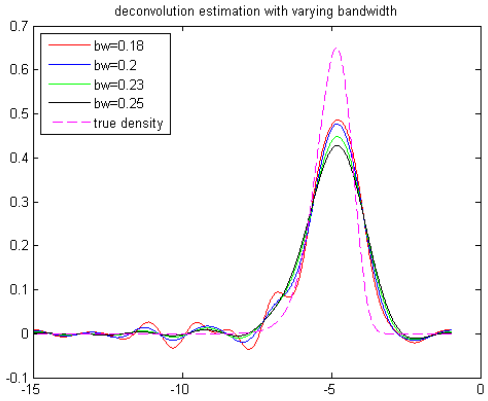

Figure 6: Deconvolution with different bandwidths(high)

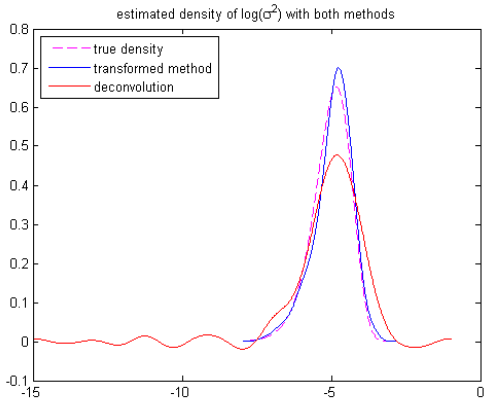

Figure 7: Estimated densities with both methods(high)

In addition, we plot the optimal estimated density from the deconvolution estimator along with the optimum one using the transformed estimator in Figure 7. Clearly, from this figure we can say that the transformed method performs better than the deconvolution estimation. To confirm our argument above, we approximate the MSE for a given point x_0 and MISE for both methods by

$$\hat{\text{MSE}}(x_0) = \frac{1}{m}\sum_{j=1}^{m}\Big(\hat{f}_j(x_0) - f(x_0)\Big)^2, \tag{4.3}$$

$$\hat{\text{MISE}} = \frac{1}{2m}\sum_{j=1}^{m}\sum_{i=1}^{\tilde{n}}\Big(\hat{f}_j(x_i) - f(x_i)\Big)^2, \tag{4.4}$$

where $\hat{f}_j$ is a density estimate based on j-th simulated data. For the transformed estimator, we have its parameters chosen using the parameter selectors for each replication. As for deconvolution method, we estimate the density over several choices of bandwidth $h = 0.2, 0.22, 0.24, 0.25, 0.26$. m is the number of replications and $-10 = x_1 < x_2 < \ldots < x_{\tilde{n}} = 0$ is an equally spaced grid over $[-10, 0]$, on which we will estimate the MISE. Here we take $x_0 = -5$, $m = 50$ and $\tilde{n} = 20$ and the results are listed in Table 3. Focusing on the global criterion, we can see that the transformed estimator obtain slightly smaller MISE than the deconvolution, which has its least MSE and MISE when bandwidth equal to 0.2.

Table 3: MSE & MISE(high frequency data)

	transformed	deconvolution with bandwidths				
		0.2	0.22	0.24	0.25	0.26
MSE	0.0183	**0.0176**	0.0243	0.0314	0.0351	0.0389
MISE	0.0227	**0.0279**	0.0339	0.0413	0.0452	0.0493

4.1.2 Low Frequency Data

To obtain a sample of low frequency data, we draw every 24^{th} data from the high frequency data $\{S_i\}_{i=1}^{n}$ above and form a new data set of size $n_l = 7500$. This means the time interval Δ_l is equal to $1/250$ and the new sample corresponds to 30yrs of daily data. In the same way as in the previous section, we transform the new sample path of log-prices $\{S_j\}_{j=1}^{n_l}$ into a sample of $\{\log(X_j^{\Delta_l})^2\}_{j=1}^{n_l-1}$, on which we apply both methods. As for the transformed estimator, the parameters are chosen both manually and automatically. The optimums are $k = 53, u = 0.2; k = 71, u = 0.27; k = 81, u = 0.29; k^* = 513, u^* = 0.2162$ and corresponding density estimates are presented in Figure 8. Obviously, the one with automatically selected parameters fits poorly compared to the rest. It is mainly because the moving average estimator produced over-smoothed estimates with a far too big value of k. This is also consistent with Figure 9, where the blue $'+'$ stands for the sample $\{\log(X_j^{\Delta_l})^2\}_{j=1}^{n_l-1}$ and the red line is the corresponding moving average estimates $\{\widehat{\log\sigma^2_{j\Delta_l}}\}_{j=1}^{n_l-1}$. For the deconvolution estimator, we select the bandwidth manually and find the optimal $h^* = 0.25$ (see Figure 10). Finally, we plot the optimal results from both estimators in Figure 11 (with the transformed estimator using k^* and u^*). It is difficult to tell from this figure which estimator performs better. That's why we also approximate the MSE(-5) and MISE in Table 4, where the deconvolution has far less MSE and MISE comparing to the transformed estimator.

Table 4: MSE & MISE(low frequency data)

	transformed	deconvolution with bandwidths				
		0.2	0.22	0.24	0.25	0.26
MSE	0.4646	0.0177	0.0233	0.0303	0.0339	0.0377
MISE	0.3107	0.0547	0.0402	0.0432	0.0458	0.0490

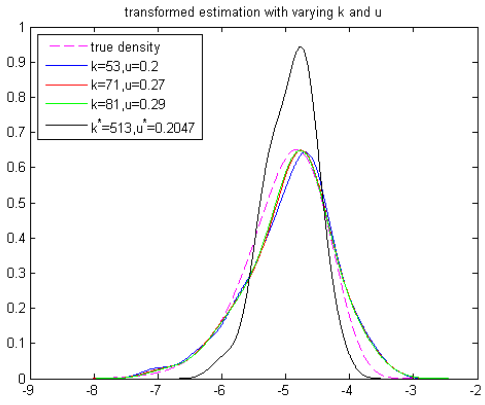

Figure 8: Transformed kernel estimation with varying k and h(low)

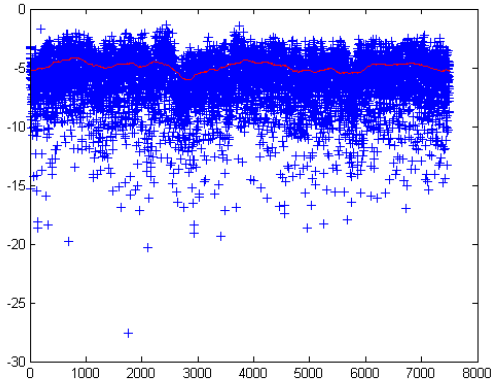

Figure 9: Moving average estimate curve(low)

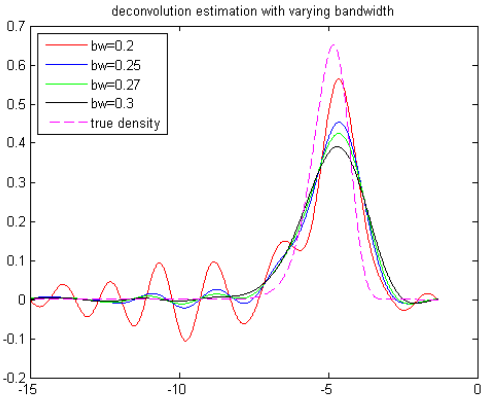

Figure 10: Deconvolution with different bandwidth(low)

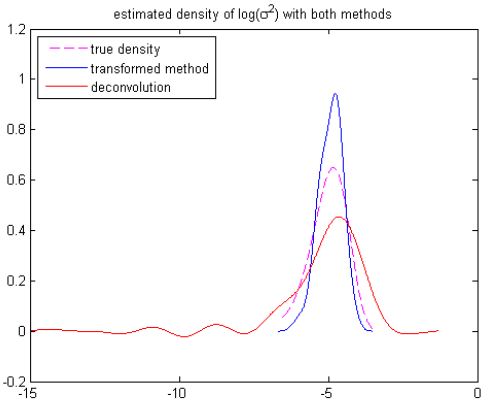

Figure 11: Estimated densities with both methods(low)

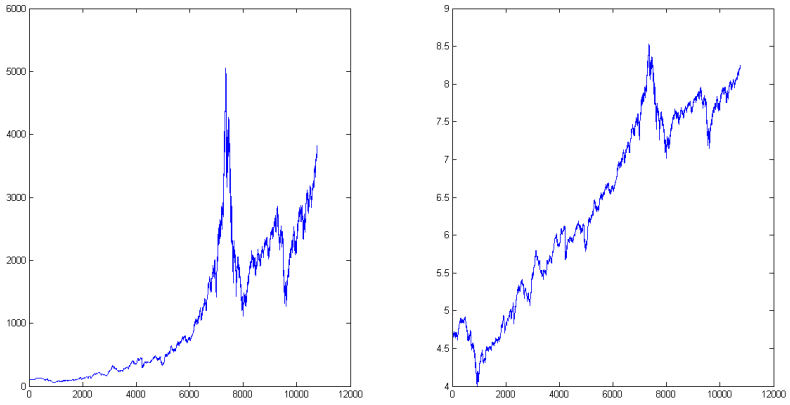

Figure 12: Left: daily closing prices; Right: log of daily prices

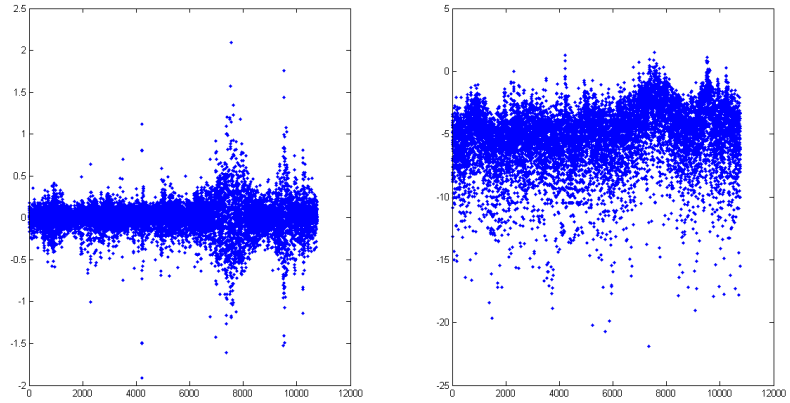

Figure 13: Left:demended and de-trended log return X_t; Right: the series of $\log(X_t^2)$

4.2 Nasdaq Index

We will estimate the volatility density based on 10760 daily closing prices of NASDAQ index from 1971/2/5 until 2013/10/2, which is plotted along with the log-prices in Figure 12. After transforming the log-prices into normalized log-returns, we demean and de-trend the log-returns and plot the resulting series of X_t as well as the series of $\log(X_t^2)$ in Figure 13. We find the optimal window size $k^* = 373$ and smoothing parameter $u^* = 0.532$ for the transformed estimator and plot the corresponding moving average estimates in figure 14. Also the optimal bandwidth for deconvolution method is found to be $h^* = 0.28$, which is shown in Figure 15. Lastly, we compare the both estimators through the plot in Figure 16.

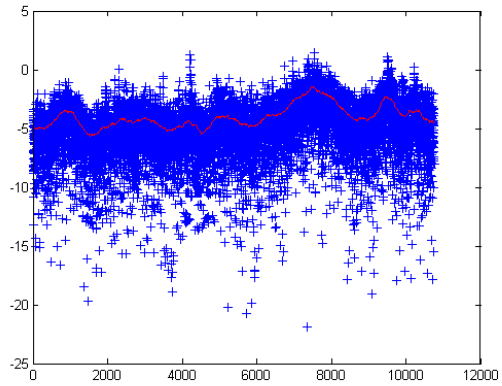

Figure 14: Moving average estimate curve(real)

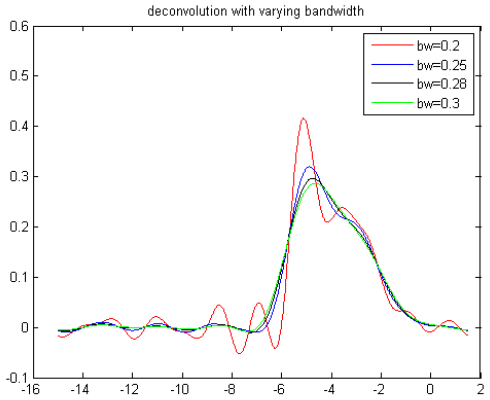

Figure 15: Deconvolution with varying bandwidth(real)

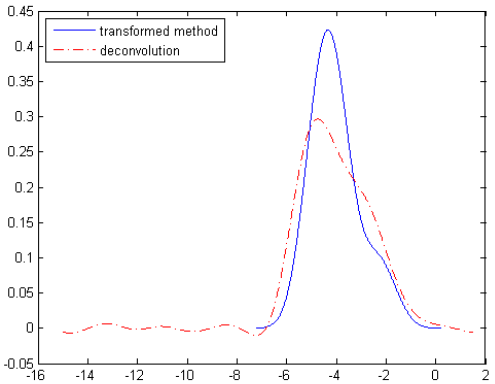

Figure 16: Estimated densities with both methods(real)

5 Conclusions

To investigate the shape of the volatility density, we propose a new nonparametric method called transformed kernel volatility density estimator. It is a two-step procedure, where we first generate the moving average estimates of $\log \sigma_t^2$ based on which we then apply the ordinary kernel density estimator. Moreover, we also suggest automatic selectors for its parameters k and u. Since the moving average estimator is approximately equivalent to the Nadaraya-Watson estimator with a uniform kernel, we estimate the window size parameter k by the bandwidth selector, more specifically the Fourier-type plug-in method for the Nadaraya-Watson estimator. As for the bandwidth u, we utilize the explicit formula for the asymptotic optimum of bandwidth for continuous-time kernel density estimation derived by Sköld and Hössjer(1999). It is because the moving average estimates are considered as a continuous-time sample and the kernel density estimator in the second step can be approximated by the continuous-time kernel density estimator.

We compare the transformed method with the deconvolution procedure by applying them on both simulated data and real data. We generate two sets of simulated data-high frequency and low frequency data from Heston model with different time intervals using full truncation scheme. The bandwidth for the deconvolution is selected manually, while the parameters in transformed method are chosen both by hand and by the automatic parameter selectors. Since without knowing the true density of the volatility it is impossible to choose the parameters manually for the transformed approach, we use the results with automatically selected parameters to compare with the deconvolution method. We find that for high frequency data the transformed approach performs better than the deconvolution method, which is also in accordance with the transformed estimator giving slightly smaller mean integrated squared error. However, for the low frequency data, the deconvolution has a much better performance in terms of MISE because the moving average estimates in the transformed method are over-smoothed. Thus, in order to have satisfying estimates from the transformed method, one had better use high frequency data if possible. Moreover, if we look at those estimates from the transformed approach with manually picked parameters alone, they all produce good estimations of the density. This implies that there is more than one optimal pair of the parameters, probably because each of the parameters has an opposite impact on the estimate curve. More specifically, the bigger the window size the taller the peak, while the larger the smoothing parameter the lower the peak. Also we can say that it is possible for the transformed method to achieve a good fit by appropriate choices of its parameters and the parameter selectors we propose are worth improving. Lastly, for the real data, the performance of both methods are incomparable although the result from

the deconvolution displays some sign of bimodal feature.

On a critical note, there is one limitation concerning the stochastic model we consider, since we make the assumption that the volatility process is independent of the Brownian motion driving the price process. As a matter of fact, it is observed that stock prices are negatively correlated with the volatility. Moreover, when proposing the parameter selectors for the transformed method, it is debatable for us to assume the conditions 2.2. and the conditions 2.4 to be satisfied. In particular, when estimating the smoothing parameter normality assumption we impose on target density is worth questioning. Thus, one can investigate some bandwidth selector without assuming any normality for further research. Besides, as the Fourier-type plug-in method we used to estimate the window size parameter is limited to the cyclic design case, one can try the classical plug-in method for comparison.

References

[1] Andersen, Leif B. G. (2007). Efficient Simulation of the Heston Stochastic Volatility Model. Available at SSRN: `http://ssrn.com/abstract=946405` or `http://dx.doi.org/10.2139/ssrn.946405`

[2] Bradley, R.C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* 2, 107-144.

[3] Broadie, M. and Kaya, O. (2006). Exact Simulation of Stochastic Volatility and other Affine Jump Diffusion Processes. *Operations Research* 54, 217-231.

[4] Castellana, J.V, Leadbetter, M.R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Processes and their Applications* 21(2), 179-193.

[5] Chiu, S.(1990). Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. *Biometrika* 77(1), 222-226.

[6] Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica* 53, 385-407.

[7] Fan, J. (1991). On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *The Annals of Statistics* 19(3), 1257-1272.

[8] Feller, W. (1951). Two singular diffusion problems. *Annals of Mathematics* 54, 173-182.

[9] Fotopoulosa, S.B. (2000). Invariance principles for deconvolving kernel density estimation for stationary sequences of random variables. *Journal of Statistical Planning and Inference* 86(1), 31-50.

[10] Gatheral, J. (2006). The Volatility Surface: A Practitioner's Guide. Wiley.

[11] Genon-Catalot, V., Jeantheau, T. and Larédo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* 6, 1051-1079.

[12] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to Bond and Currency options. *The Review of Financial Studies* 6(2), 327-343.

[13] Horová, I., Koláček, J. and Zelinka, J. (2012). Kernel smoothing in MATLAB. *World Scientific.*

[14] Iakezawa, K. (2006). Introduction to Nonparametric Regression. John Wiley & Sons Inc., Hoboken, New Jersey.

[15] Kahl, C. and Jäckel, P. (2006). Fast Strong Approximation Monte-Carlo Schemes for Stochastic Volatility Models. *Quantitative Finance* 6(6), 513-536.

[16] Koláček, J. (2005). Kernel Estimation of the Regression Function (in Czech), Ph.D. thesis, Masaryk University, Brno.

[17] Koláček, J. (2008). An improved estimator for removing boundary bias in kernel cumulative distribution function estimation (in Czech), *Proceedings in Computational Statistics COMPSTAT'08* 549-556.

[18] Lord, R., Koekkoek, R. and Van Dijk, Dick J.C. (2008) A Comparison of Biased Simulation Schemes for Stochastic Volatility Models. Tinbergen Institute Discussion Paper No. 06-046/4. Available at SSRN: `http://ssrn.com/abstract=903116` or `http://dx.doi.org/10.2139/ssrn.903116`

[19] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* 9(1), 141-142.

[20] Nelson, D.B. (1990). ARCH models as diffusion approximations. *J. Econometrics* 45, 7-38.

[21] Silverman, B.W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

[22] Sköld, M. and Hössjer, O. (1999). On the asymptotic variance of the continuous-time kernel density estimator. *Statistics & Probability Letters* 44(1), 97-106.

[23] Van Es, B., Spreij, P. and Van Zanten, H. (2001). URL `http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=E7EA729EA38DA6025EC55CAC6121EFFA?doi=10.1.1.20.9315&rep=rep1&type=pdf`

[24] Van Es, B., Spreij, P. and Van Zanten, H. (2003). Nonparametric volatility density estimation. *Bernoulli* 9(3), 451-465.

[25] Van Es, B., Spreij, P. and Van Zanten, H. (2005). Nonparametric volatility density estimation for discrete time models. *Journal of Nonparametric Statistics* 17(2), 237-249.

[26] Van Es, S. and Spreij, P. (2009) Nonparametric methods for volatility density estimation. *ArXiv e-prints* URL `http://arxiv.org/pdf/0910.5185v1.pdf`

[27] Wahba, G. (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statistics* 3(1), 15-29.

[28] Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.

[29] Watson, G. S. (1964). Smooth regression analysis.*Sankhya: The Indian Journal of Statistics, Series A* 26(4), 359-372.

[30] Wiggins, J.B. (1987). Option valuation under stochastic volatility. *Journal of Financial Economics* 19, 351-372.