Nonparametric Econometrics: Theory and Applications[1]

ZONGWU CAI

Department of Economics, University of Kansas
E-mail: caiz@ku.edu

November 3, 2025

[1] This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes.

Preface

This is the advanced level of nonparametric econometrics with theory and applications. Here, the focus is on both the theory and the skills of analyzing real data using nonparametric econometric techniques and statistical softwares such as R or Python. This is along the line with the spirit "STRONG THEORETICAL FOUNDATION and SKILL EXCELLENCE". In other words, this course covers the advanced topics in analysis of economic and financial data using nonparametric techniques, particularly in nonlinear time series models and some models related to economic and financial applications. The topics covered start from classical approaches to modern modeling techniques even up to the research frontiers. The difference between this course and others is that you will learn not only the theory but also step by step how to build a model based on data (or so-called *let data speak themselves*) through real data examples using statistical softwares or how to explore the real data using what you have learned. Therefore, there is no a single book serviced as a textbook for this course so that materials from some books and articles will be provided. However, some necessary handouts, including computer codes like R codes, will be provided with your help (You might be asked to print out the materials by yourself).

Several projects (two or three), including the heavy computer works, are assigned throughout the term. The purpose of doing projects is to train students to understand the theoretical concepts and to know how to apply the methodologies learned in class to real problems. The group discussion is allowed to do the projects, particularly writing the computer codes. But, writing the final report to each project must be in your own language. Copying each other will be regarded as a cheating. If you use the R language, similar to SPLUS, you can download it from the public web site at http://www.r-project.org/ and install it into your own computer. You are STRONGLY encouraged to use (but not limited to) the package R or Python since it is a very convenient programming language for doing statistical analysis and Monte Carol simulations as well as various applications in quantitative economics and finance. Of course, you are welcome to use any one of other packages such as SAS, Python, GAUSS, STATA, SPSS and EVIEW. But, I might not be able to give you a help if doing so.

Why do we need to study nonparametric econometrics? Here is a motivated example. For example, let us go back to review the classical sample selection (Heckman) model. That is, model setting is given by

$$y_t^0 = g_1(X_t) + u_t, \quad \text{and} \quad z_t^0 = \gamma^\top W_t + v_t.$$

We only observe data $\{y_t, X_t, W_t, z_t\}_{t=1}^n$, where $y_t = y_t^0$ if $z_t = 1$ with $z_t = I(z_t^0 > 0)$. Then, without normality assumption, we have

$$\begin{aligned} E(y_t\,|X_t, W_t) &= E(y_t^0|\,X_t, W_t, z_t = 1) = g_1(x_t) + E(u_t|\,X_t, W_t, v_t > -\gamma^\top W_t) \\ &= g_1(X_t) + E(u_t\,W_t, v_v > -\gamma^\top W_t) = g_1(X_t) + g_2(W_t), \end{aligned}$$

which is an additive model, where $g_2(W_t)$ denotes the second term on the right hand side in the above equation. Therefore, this is a generalization of the Heckman model. To estimate $g_1(x)$, we need to learn nonparametric methods.

Contents

List of Tables

List of Figures

Chapter 1

Density, Distribution & Quantile Estimations

1.1 Time Series Structure

Since most of economic and financial data are time series, we discuss our methodologies and theory under the framework of time series. For linear models, the time series structure can be often assumed to have some well known forms such as an autoregressive moving average (ARMA) model. However, under nonparametric setting, this assumption might not be valid. Therefore, we can assume a more general time series dependence, which is commonly used in the literature, described as follows.

1.1.1 Mixing Conditions

Mixing dependence is commonly used to characterize the dependent structure and it is often referred to as short range dependence or weak dependence, which means that the distance between two observations goes farther and farther, the dependence becomes weaker and weaker very faster. It is well known that α-mixing (strong mixing) includes many time series models as a special case. In fact, under very mild assumptions, linear processes, including linear autoregressive models and more generally bilinear time series models are α-mixing with mixing coefficients decaying exponentially. Many nonlinear time series models, such as functional coefficient autoregressive processes with/without exogenous variables, nonlinear additive autoregressive models with/without exogenous variables, ARCH and GARCH type processes, stochastic volatility models, and many continuous time diffusion models (including the Black-Scholes type models) are strong mixing under some mild conditions. See Genon-Caralot, Jeantheau and Laredo (2000), Cai and Masry (2000), Cai (2002), Carrasco and

Chen (2002), and Chen and Tang (2005) for more details.

To simplify the notation, we only introduce mixing conditions for strictly stationary processes (in spite of the fact that a mixing process is not necessarily stationary). The idea is to define mixing coefficients to measure the strength (in different ways) of dependence for the two segments of a time series which are apart from each other in time. Let $\{X_t\}$ be a strictly stationary time series. For $n \geq 1$, define

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}} |P(A)P(B) - P(AB)|,$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_t; i \leq t \leq j\}$. Note that $\mathcal{F}_n^\infty \downarrow$. If $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty, \{X_t\}$ is called α-mixing or strong mixing. There are several other mixing conditions such as ρ-mixing, β-mixing, ϕ-mixing, and ψ-mixing; see the books by Hall and Heyde (1980) and Fan and Yao (2003, page 68) for details. Indeed,

$$\begin{aligned} \beta(n) &= E\left\{ \sup_{A \in \mathcal{F}_n^\infty} \mid P(A) - P\left(A \mid X_t, t \leq 0\right) \right\}, \\ \rho(n) &= \sup_{X \in \mathcal{F}_{-\infty}^{0}; Y \in \mathcal{F}_{n}^{\infty}} \operatorname{Corr}(X, Y)|, \\ \phi(n) &= \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}, P(A)>0} |P(B) - P(B \mid A)|, \end{aligned}$$

and

$$\psi(n) = \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}, P(A)P(B)>0} |1 - P(B \mid A)/P(B)|.$$

It is well known that the relationships among the mixing conditions are

$$\alpha(n) \leq \frac{1}{4}\rho(n) \leq \frac{1}{2}\phi(n),$$

so that ψ-mixing $\Longrightarrow$ ϕ-mixing $\Longrightarrow$ ρ-mixing $\Longrightarrow$ α-mixing as well as β-mixing $\Longrightarrow$ α-mixing. Note that all our theoretical results are derived under mixing conditions. The following inequalities are very useful in applications, which can be found in the book by Hall and Heyde (1980, pp. 277-280).

Lemma 1.1: (Davydov's inequality) (i) If $E\left|X_i\right|^p + E\left|X_j\right|^q < \infty$ for some $p \geq 1$ and $q \geq 1$ and $1/p + 1/q < 1$, it holds that

$$|\operatorname{Cov}(X_i, X_j)| \leq 8\alpha^{1/r}(|j-i|)||X_i\|_p||X_j\|_q,$$

where $r = (1 - 1/p - 1/q)^{-1}$.
(ii) If $P\left(|X_i| \le C_1\right) = 1$ and $P\left(|X_j| \le C_2\right) = 1$ for some constants C_1 and C_2, it holds that

$$|\mathrm{Cov}\left(X_i, X_j\right)| \le 4\alpha(|j - i|)C_1C_2$$

Note that if we allow X_i and X_j to be complex-valued random variables, (ii) still holds with the coefficient "4" on the RHS of the inequality replaced by "16".
(iii) If $P\left(|X_i| \le C_1\right) = 1$ and $E\left|X_j\right|^p < \infty$ for some constants C_1 and $p > 1$, then,

$$|\mathrm{Cov}\left(X_i, X_j\right)| \le 6C_1 \left\|X_j\right\|_p \alpha^{1-p^{-1}}(|j - i|).$$

Lemma 1.2: If $E\left|X_i\right|^p + E\left|X_j\right|^q < \infty$ for some $p \ge 1$ and $q \ge 1$ and $1/p + 1/q = 1$, it holds that

$$|\mathrm{Cov}\left(X_i, X_j\right)| \le 2\phi^{1/p}(|j - i|)||X_i\,\|_p||X_j\|_q\,.$$

1.1.2 Martingale and Mixingale

Martingale is very useful in applications. Here is the definition. Let $\{X_n, n \in \mathcal{N}\}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_n, n \in \mathcal{N}\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. Suppose that the sequence $\{X_n, n \in \mathcal{N}\}$ satisfies
(i) X_n is measurable with respect to $\mathcal{F}_n$,
(ii) $E\left|X_n\right| < \infty$,
(iii) $E\left[X_n \mid \mathcal{F}_m\right] = X_m$ for all $m < n, n \in \mathcal{N}$.
Then, the sequence $\{X_n, n \in \mathcal{N}\}$ is said to be a martingale with respect to $\{\mathcal{F}_n, n \in \mathcal{N}\}$. We write that $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is a martingale. If (i) and (ii) are retained and (iii) is replaced by the inequality $E\left[X_n \mid \mathcal{F}_m\right] \ge X_m$ ($E\left[X_n \mid \mathcal{F}_m\right] \le X_m$), then $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a sub-martingale (super-martingale). Define $Y_n = X_n - X_{n-1}$. Then $\{Y_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a martingale difference (MD) if $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a martingale. Clearly, $E\left[Y_n \mid \mathcal{F}_{n-1}\right] = 0$, which means that a MD is not predicable based on the past information. In a finance language, a stock market is *efficient*. Equivalently, it is a MD.

Another type of dependent structure is called mixingale, which is the so-called asymptotic martingale. The concept of mixingale, introduced by McLeish (1975), is defined as follows. Let $\{X_n, n \ge 1\}$ be a sequence of square-integrable random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_n, -\infty < n < \infty\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. Then, $\{X_n, \mathcal{F}_n\}$ is called a L_r-mixingale (difference) sequence for $r \ge 1$ if, for some sequences of nonnegative constants c_n and ψ_m, where $\psi_m \to 0$ as $m \to \infty$, we have

$$(i) \quad \|E(X_n \mid \mathcal{F}_{n-m})\|_r \le \psi_m c_n, \quad \text{and} \quad (ii) \quad \|X_n - E(X_n \mid \mathcal{F}_{n-m})\|_r \le \psi_{m+1} c_n,$$

for all $n \ge 1$ and $m \ge 0$. The idea of mixingale is to try to build a bridge between martingale and mixing. The following examples give the idea of the scope of L_2-mixingales.

Examples:

1. A square-integrable martingale is a mixingale with $c_n = \|X_n\|$ and $\psi_0 = 1$ and $\psi_m = 0$ for $m \ge 1$.
2. A linear process is given by $X_n = \sum_{i=-\infty}^{\infty} \alpha_{i-n}\xi_i$ with $\{\xi_i\}$ iid mean zero and variance σ^2 and $\sum_{i=-\infty}^{\infty} \alpha_i^2 < \infty$. Then, $\{X_n, \mathcal{F}_n\}$ is a mixingale with all $c_n = \sigma$ and $\psi_m^2 = \sum_{|i|\ge m} \alpha_i^2$.
3. If $\{X_n\}$ is a square-integrable sequence of ϕ-mixing, then it is a mixingale with $c_n = 2\|X_n\|_2$ and $\psi_m = \phi^{1/2}(m)$, where $\phi(m)$ is the ϕ-mixing coefficient.
4. If $\{X_n\}$ is a sequence of α-mixing with $\|X_n\|_p < \infty$ for some $p > 2$, then it is a mixingale with $c_n = 2(\sqrt{2}+1)\|X_n\|_2$ and $\psi_m = \alpha^{1/2-1/p}(m)$, where $\alpha(m)$ is the α-mixing coefficient.

Note that Examples 3 and 4 can be derived form the following inequality, due to McLeish (1975).

Lemma 1.3: (McLeish's inequality) Suppose that X is a random variable measurable with respect to $\mathcal{A}$, and $\|X\|_r < \infty$ for some $1 \le p \le r \le \infty$. Then

$$\|E(X \mid \mathcal{F}) - E(X)\|_p \le \begin{cases} 2[\phi(\mathcal{F},\mathcal{A})]^{1-1/r}\|X\|_r, & \text{for } \phi\text{-mixing}, \\ 2\left(2^{1/p}+1\right)[\alpha(\mathcal{F},\mathcal{A})]^{1/p-1/r}\|X\|_r, & \text{for } \alpha\text{-mixing}. \end{cases}$$

1.2 Nonparametric Density Estimate

Let $\{X_i\}$ be a random sample with a (unknown) marginal distribution $F(\cdot)$ (CDF) and its probability density function (PDF) $f(\cdot)$. The question is how to estimate $f(\cdot)$ and $F(\cdot)$. Since

$$F(x) = P(X_i \le x) = E[I(X_i \le x)] = \int_{-\infty}^{x} f(u)\,du,$$

and

$$f(x) = \lim_{h \downarrow 0} \frac{F(x+h) - F(x-h)}{2h} \approx \frac{F(x+h) - F(x-h)}{2h}$$

if h is very small, by the method of moment estimation (MME), $F(x)$ can be estimated by

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right),$$

which is called the empirical cumulative distribution function (ecdf), so that $f(x)$ can be estimated by

$$f_n(x)=\frac{F_n(x+h)-F_n(x-h)}{2h}=\frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i-x\right),$$

where $K(u)=I(|u|\leq 1)/2$ and $K_h(u)=K(u/h)/h$. Indeed, the kernel function $K(u)$ can be taken to be any **symmetric** density function. Here, h is called the bandwidth. Initially, $f_n(x)$ was proposed by Rosenblatt (1956) and Parzen (1962) explored its properties in detail. Therefore, it is called the Rosenblatt-Parzen density estimate.

Remark 1.1: *Let $R(h)=f(x)-[F(x+h)-F(x-h)/2h$ so that $f(x)=[F(x+h)-F(x-h)/2h+R(h)$. Then, $R(h)=O\left(h^2\right)$ is the second order approximation of $f(x)$ if h is small and the second derivative of $f(x)$ is continuous. Therefore, $f_n(x)$ is not the unbiased estimate due to the approximation error.*

Exercise: Please show that $F_n(x)$ is an unbiased estimate of $F(x)$ but $f_n(x)$ is a biased estimate of $f(x)$. **Think about intuitively**

(1) why $f_n(x)$ is biased

(2) where the bias comes from

(3) why $K(\cdot)$ should be symmetric.

1.2.1 Asymptotic Properties

A. Asymptotic Properties for ECDF

If $\{X_i\}$ is stationary, then, $E\left[F_n(x)\right]=F(x)$ and

$$\begin{aligned} n\mathrm{Var}\left(F_n(x)\right)=&\mathrm{Var}\left(I\left(X_i\leq x\right)\right)+2\sum_{i=2}^{n}\left(1-\frac{i-1}{n}\right)\mathrm{Cov}\left(I\left(X_1\leq x\right), I\left(X_i\leq x\right)\right)\\ &=\underbrace{F(x)[1-F(x)]+2\sum_{i=2}^{n}\mathrm{Cov}\left(I\left(X_1\leq x\right), I\left(X_i\leq x\right)\right)}_{\rightarrow\sigma_F^2(x)\quad \text{by assuming that } \sigma_F^2(x)<\infty}\end{aligned}$$

$$\underbrace{-2\sum_{i=2}^{n}\frac{i-1}{n}\text{Cov}\left(I\left(X_1\leq x\right),I\left(X_i\leq x\right)\right)}_{\to 0 \text{ by Kronecker Lemma}}$$

$$\to\ \sigma_F^2(x)\equiv F(x)[1-F(x)]+2\underbrace{\sum_{i=2}^{\infty}\text{Cov}\left(I\left(X_1\leq x\right),I\left(X_i\leq x\right)\right)}_{\text{This term is called } A_d(x)}.$$

Therefore,

$$n\text{Var}\left(F_n(x)\right)\ \to\ \sigma_F^2(x). \tag{1.1}$$

One can show based on the mixing theory that

$$\sqrt{n}\left[F_n(x)-F(x)\right]\ \to\ N\left(0,\sigma_F^2(x),\right). \tag{1.2}$$

which can be derived in the same way as in the proof of Theorem 2.2 in Section 2.4; see Section 2.4.7 for details. It is clear that $A_d(x)=0$ if $\{X_i\}$ are independent so that $\sigma_F^2(x)=F(x)[1-F(x)]$. If $A_d(x)\neq 0$, the question is how to estimate it. For each given x, one can use the HC estimator by White (1980) or the HAC estimator by Newey and West (1987) or the kernel method by Andrews (1991).

The results in (1.2) can used to construct a test statistic to test the null hypothesis

$$H_0: F(x)=F_0(x) \quad \text{versus} \quad H_a: F(x)\neq(>)(<)F_0(x).$$

This test statistic is the well-known Kolmogorov-Smirnov test, defined as

$$D_n=\sup_{-\infty<x<\infty}|F_n(x)-F_0(x)|$$

for the two-sided test. One can show, see, for example, Serfling (1980, p.62) or Billinsley (1999, p.103), that under some regularity conditions, which include that the data are iid,

$$P\left(\sqrt{n}D_n\leq d\right)\ \to\ 1-2\sum_{j=1}^{\infty}(-1)^{j+1}\exp\left(-2j^2d^2\right)$$

and

$$P\left(\sqrt{n}D_n^+\leq d\right)=P\left(\sqrt{n}D_n^-\geq -d\right)\ \to\ 1-\exp\left(-2d^2\right),$$

where $D_n^+=\sup_{-\infty<x<\infty}[F_n(x)-F_0(x)]$ and $D_n^-=\sup_{-\infty<x<\infty}[F_0(x)-F_n(x)]$ for one-sided tests. In R, there is a built-in command for the Kolmogorov-Smirnov test, which is **ks.test()**.

Exercise: What are the most important assumptions on the Kolmogorov-Smirnov test? Please think about the question "Does the Kolmogorov-Smirnov test hold for time series?" If not, please conduct a simulation to verify your conjecture.

B. Asymptotic Properties for Density Estimation

Next, we derive the asymptotic variance for $f_n(x)$. First, define $Z_i = K_h(X_i - x)$. Then,

$$\begin{aligned} E[Z_1 Z_i] &= \iint K_h(u-x)K_h(v-x)f_{1,i}(u,v)dudv \\ &= \iint K(u)K(v)f_{1,i}(x+uh, x+vh)dudv \\ &\to \ f_{1,i}(x,x), \end{aligned}$$

where $f_{1,i}(u,v)$ is the joint density of (X_1, X_i), so that

$$\mathrm{Cov}(Z_1, Z_i) \ \to \ f_{1,i}(x,x) - f^2(x).$$

It is easy to show that

$$h\mathrm{Var}(Z_1) \ \to \ \nu_0(K)f(x),$$

where $\nu_j(K) = \int u^j K^2(u)du$. Therefore,

$$\begin{aligned} n\,h\,\mathrm{Var}(f_n(x)) &= h\mathrm{Var}(Z_1) + \underbrace{2h\sum_{i=2}^{n}\left(1 - \frac{i-1}{n}\right)\mathrm{Cov}(Z_1, Z_i)}_{\equiv A_f \ \to \ 0 \quad \text{under some assumptions}} \\ &\to \ \nu_0(K)f(x) \end{aligned}$$

To show that $A_f \to 0$, let $d_n \to \infty$ and $d_n h \to 0$. Then,

$$|A_f| \le h\sum_{i=2}^{d_n}|\mathrm{Cov}(Z_1, Z_i)| + h\sum_{i=d_n+1}^{n}|\mathrm{Cov}(Z_1, Z_i)|.$$

For the first term, if $f_{1,i}(u,v) \le M_1$, then, it is bounded by $h\,d_n = o(1)$. For the second term, we apply the Davydov's inequality (see Lemma 1.1) to obtain

$$h\sum_{i=d_n+1}^{n}|\mathrm{Cov}(Z_1, Z_i)| \le M_2\sum_{i=d_n+1}^{n}\alpha(i)/h = O\left(d_n^{-\beta+1}h^{-1}\right)$$

if $\alpha(n) = O\left(n^{-\beta}\right)$ for some $\beta > 2$. If $d_n = O\left(h^{-2/\beta}\right)$, then, the second term is dominated by $O\left(h^{1-2/\beta}\right)$ which goes to 0 as $n \to \infty$. Hence,

$$n\,h\,\mathrm{Var}(f_n(x)) \ \to \ \nu_0(K)f(x). \tag{1.3}$$

By a comparison of (1.1) and (1.3), one can see clearly that there is an infinity term involved in $\sigma_F^2(x)$ due to the dependence but the asymptotic variance in (1.3) is the same as that for the iid case (without the infinity term). We can establish the following asymptotic normality for $f_n(x)$ but the proof will be discussed later.

Theorem 1.1: *Under some regularity conditions, we have*

$$\sqrt{n\,h}\left[f_n(x)-f(x)-\frac{h^2}{2}\mu_2(K)f''(x)+o_p\left(h^2\right)\right] \;\to\; N\left(0,\nu_0(K)f(x)\right),$$

where the term $\frac{h^2}{2}\mu_2(K)f''(x)$ is called the asymptotic bias and $\mu_2(K)=\int u^2K(u)du$.

Remark 1.2: *Note that Theorem 1.1 can be proved by using the Linderburg-Feller or Lyapunov central limit theorem (CLT)[1] for triangular arrays, if $\{X_t\}$ are independent. But, for time series cases, the proof is different and is similar to that for Theorem 2.2 (see Section 2.4 later) so that you can follow the idea in Section 2.4 to establish Theorem 1.1 for time series cases. Also, according to Theorem 1.1, $f_n(x)\to f(x)$ for each x as $n\to\infty$ so that $f_n(x)=O_p(1)$, when $\{X_t\}$ is stationary. However, when $\{X_t\}$ is a nonstationary process like a random walk (integrated process), then, one can show that $f_n(x)=O_p(1/\sqrt{n})$; see, for example, the papers by Phillips and Park (1998) and Cai, Li and Park (2009) for details. Thus, the order of magnitude of the density estimate $f_n(x)$ in the integrated case is smaller than in the stationary case when $n\to\infty$. This is explained by the fact that an integrated process like X_t eventually (as $t\to\infty$) has a bigger probability of being away from a given point x than a stationary process and the kernel function $K(\cdot)$ assigns smaller values to the more distant points. This has important implications for kernel regression with nonstationary time series. In effect, this reduces the rate of convergence of the kernel estimate of the density function. Indeed, the asymptotic distribution of $f_n(x)$ for nonstationary $\{X_t\}$ is totally different from that in Theorem 1.1 for stationary case, which is given by*

$$\sqrt{n}\,f_n(x) \;\to\; \xi$$

in probability, where ξ is **non-normal random variable** *(a local time of a Brownian motion), and the rate of convergence of the kernel estimate of the density function is much slower than $\sqrt{n\,h}$ for the stationary case. See Theorem 3.1 in Phillips and Park (1998) and Lemma B.1 in Cai, Li and Park (2009) for details. Indeed, from Lemma B.1 in Cai, Li and Park (2009), one can see that*

$$\xi=\begin{cases} L(1,0)/\sigma_u, & \text{if } \; x \; \text{ is fixed,}\\ L(1,a)/\sigma_u, & \text{if } \; x=a\sqrt{n} \; \text{ for any fixed } a,\end{cases}$$

[1] The Lyapunov CLT says that if triangular arrays $\{Z_{nt}\}_{t=1}^n$ are independent, and the Lyapunov condition $\sum_{t=1}^n E\left(|Z_{nt}|^{2+\delta}\right)/s_n^{2+\delta}\to 0$ for some $\delta>0$ holds, where $s_n^2=\sum_{t=1}^n \mathrm{Var}\left(X_{nt}\right), \sum_{t=1}^n\left[Z_{nt}-E\left(Z_{nt}\right)\right]/s_n$ converges to the standard normal. See, for example, Serfling (1980, p.32) for details.

where σ_u^2 is the variance of $u_t = X_t - X_{t-1}$, and $L(t,x)$ is the local time t of the standard Brownian motion at x, given by

$$L(t,x) = \lim_{\epsilon \downarrow 0} \frac{1}{2\,\epsilon} \int_0^t I(|W_u(s) - x| \le \epsilon) ds \tag{1.4}$$

with $W_u(t)$ being the standard Brownian motion generated by $\{u_t\}$. For the definition and its properties, see, for example, the book by Marcus and Rosen (2006) for details.

Exercise: First, by comparing (1.1) with (1.3), what can you observe? Second, if $\{X_t\}$ is a sequence of nonstationary (say, unit root) random variable, what does $f_n(x)$ estimate? Please think about this problem. You will be asked to do a simulation to see what you can observe in your next homework assignment.

Example 1.1: Let us examine how importance the choice of bandwidth is. The data $\{X_i\}_{i=1}^n$ are generated from $N(0,1)$ (iid) and $n = 300$. The grid points are taken to be $[-4,4]$ with an increment $\Delta = 0.1$. Bandwidth is taken to be 0.25 (red line), 0.5 (green line) and 1.0 (blue), and h_{opt} (cyan line), respectively, and the kernel can be the Epanechnikov kernel $K(u) = 0.75\,(1-u^2)\,I(|u| \le 1)$ or the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$. Comparisons are given in Figure 1.1 (the left panel) for different choices of h. Note that the comparison between two kernels: Gaussian (black line) and Epanechnikov (red line) is displaced in the right panel of Figure 1.1. This simulation shows that the choice of bandwidth h is critical but the choice of $K(u)$ is not so sensitive.

Example 1.2: Next, we apply the kernel density estimation to estimate the density of the weekly 3-month Treasury bill (Secondary Market Rate, Discount Basis) from January 8, 1954 to September 23, 2022.[2] Figure 1.2 displays the ACF and PACF plots for the original data (top panel), denoted by X_t, and the first difference (middle panel), denoted by $r_t = X_t - X_{t-1}$, and the estimated density of the differencing series r_t together with the true standard normal density: the bottom left panel is for $f_n(x)$ (**black** solid line) by using the built-in function **density()** and the bottom right panel is for the own code, respectively, together with the with the density curve of the standard normal (red dashed line). From the top penal in Figure 1.2 first, one can conclude clearly that X_t is nonstationary (possible unit root) so that the differencing of X_t is is needed. Then, define $r_t = X_t - X_{t-1}$ and the ACF and PACF

[2] The dataset can be updated to today and can be downloaded from Federal Bank of St. Louis at https://fred.stlouisfed.org/series/DTB3.

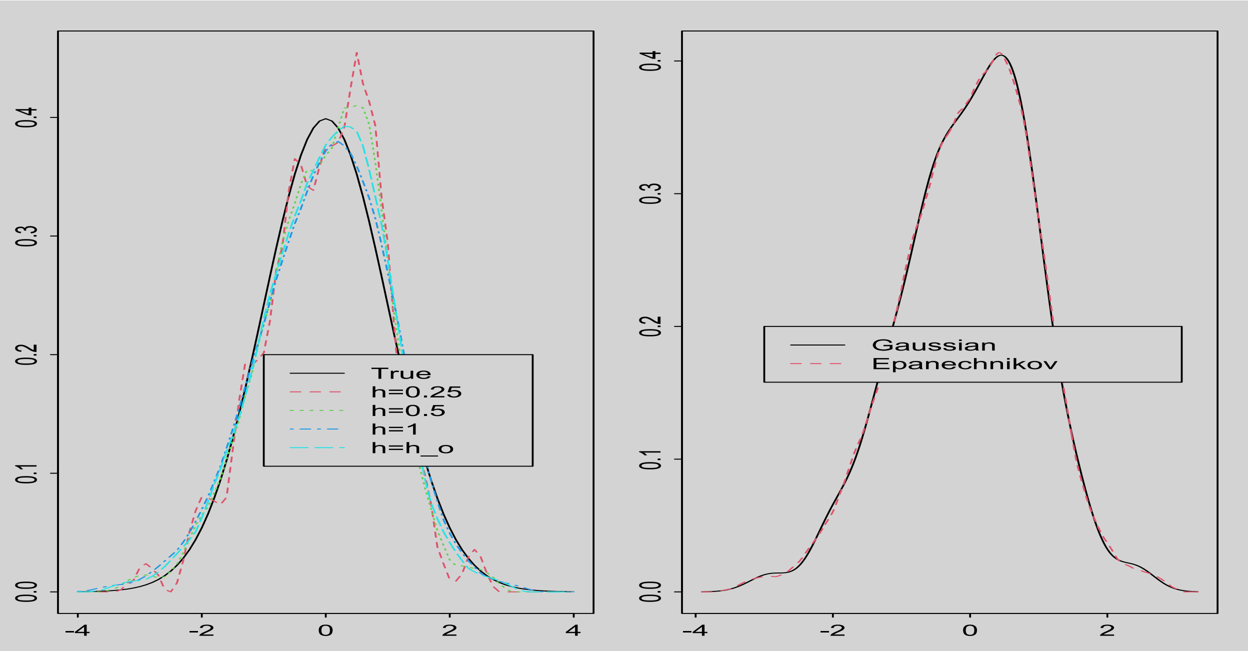


Figure 1.1: Left panel: Together with the true density (**black line**), bandwidth is taken to be 0.25 (red line), 0.5 (green line), 1.0 (blue line) and the optimal one (cyan line line, see later) with the Epanechnikov kernel. Right panel: the kernel density estimates for two different kernel functions: Gaussian (**lack line**) and Epanechnikov (red line).

plots of $\{r_t\}$ are given in the middle panel of Figure 1.2 from which, one can see that r_t is autocorrelated. Finally, the bottom panel concludes that the distribution of r_t is not normal although its distribution looks symmetric and uni-mod. But, at 0, there is a high peak and two tails are heavy, which support the stylized factors about the distribution of the return. Also, one can see that there is no difference between computings based on **density()** and the own code.

Example 1.3: In this example, we consider the case that $\{X_t\}$ is nonstationary and investigate the asymptotic properties of both $f_n(x)$ and $\sqrt{n} f_n(x)$. The data generating process is

$$X_t = (1 - \delta/n)X_{t-1} + u_t$$

with $X_0 = 0$ for some $\delta \geq 0$. Here, we consider two cases $\delta = 0$ (random walk) and $\delta = 1$ (nearly integrated process or nearly random walk or nearly unit root), where $t = 1, \cdots, n$, and $u_t \overset{i.i.d}{\sim} N(0, 1)$. Bandwidth $h = d\, n^{-1/10}$ with d taken to be 0.5, 1 and 2, respectively. The sample size is taken to be 200, 1000 and 5000, respectively. For each setting, the simulation is repeated 10,000 times, and $f_n(x)$ calculated for x being fixed with -5 (magenta), -2.5

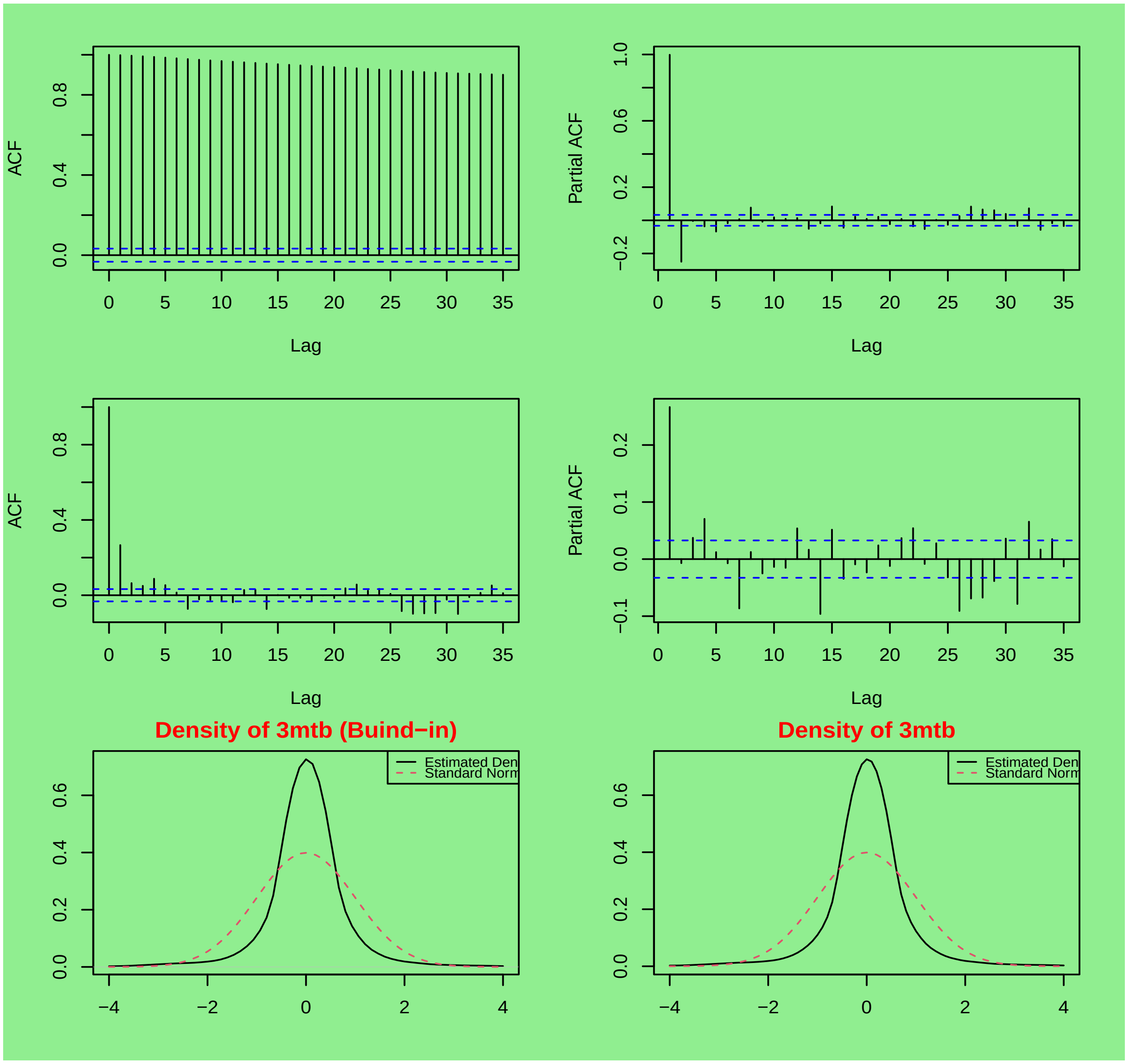


Figure 1.2: The ACF and PACF plots for the original data (top panel), denoted by X_t, and the first difference (middle panel), denoted by $r_t = X_t - X_{t-1}$, which can be regarded as the simple return. The bottom left panel is for $f_n(x)$ (black solid line) by using the built-in function **density()** and the bottom right panel is for the own code, respectively, together with the density curve of the standard normal (red dashed line).

(red), 0 (orange), 2.5 (blue) and 5 (green), respectively. The simulation results are given in Figure 1.3 for boxplots and Table 1.1 for reporting the median and standard devision of the 10,000 values of $f_n(x)$ for each x, each sample size, and each d value, respectively. It is clear from both Figure 1.3 and Table 1.1, for each setting, $f_n(x)$ is closer to 0 as the sample size gets larger, which verifies the theory in Remark 1.2 that $f_n(x)$ converges to 0 as n goes to infinity.

Table 1.1: The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 0$ (random walk case).

The kernel density estimator for a random walk

	$d = 0.5$					$d = 1$					$d = 2$				
n	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$
200	0.017	0.034	0.042	0.034	0.025	0.021	0.034	0.047	0.034	0.025	0.025	0.038	0.045	0.038	0.025
	(0.040)	(0.043)	(0.044)	(0.042)	(0.041)	(0.038)	(0.040)	(0.041)	(0.040)	(0.038)	(0.036)	(0.038)	(0.038)	(0.038)	(0.036)
1000	0.016	0.018	0.020	0.018	0.016	0.016	0.018	0.021	0.018	0.016	0.016	0.019	0.020	0.018	0.016
	(0.019)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
5000	0.008	0.009	0.009	0.009	0.008	0.009	0.009	0.010	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)

Next, for $h = 0.5\,n^{-1/10}$, $\sqrt{n}\,f_n(x)$ is calculated for x being fixed with -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green) respectively. For each setting, the simulation is repeated 10,000 times. The estimated density curves are plotted in Figure 1.4, from which, one can observe that the estimated curves get closer to each other as the sample size gets larger. Therefore, $\sqrt{n}\,f_n(x)$ can be used to approximate the distribution of the local time $L(1, 0)$ of a Brownian motion.

Finally, repeat the above procedures with $x = a\sqrt{n}$ with a taken to be -0.5 (magenta), -0.25 (red), 0 (orange), 0.25 (blue) and 0.5 (green), respectively. The estimated density curves are displayed in Figure 1.5, from which, one can see that the estimated curves approximate different distributions as the sample size increases, which is dependent on the value of a. Therefore, $\sqrt{n}\,f_n(a\sqrt{n})$ can be used to approximate the distribution of the local time $L(1, a)$ for any $a \neq 0$.

Note that we also consider the case that $\delta = 1$ (nearly integrated case). The simulation results are presented in Table 1.2 for each setting and the same conclusions to those for a random walk scenario can be made. The figures similar to Figures 1.3 - 1.5 can re-produced in the same way, but, to save the space, they are not depicted here.

Note that the computer code in **R** for the above two examples can be found in Section 1.5. **R** has a built-in function **density()** for computing the nonparametric density estimation. Also, you can use the command **plot(density())** to plot the estimated density. Further, **R** has a built-in function **ecdf()** for computing the empirical cumulative distribution function

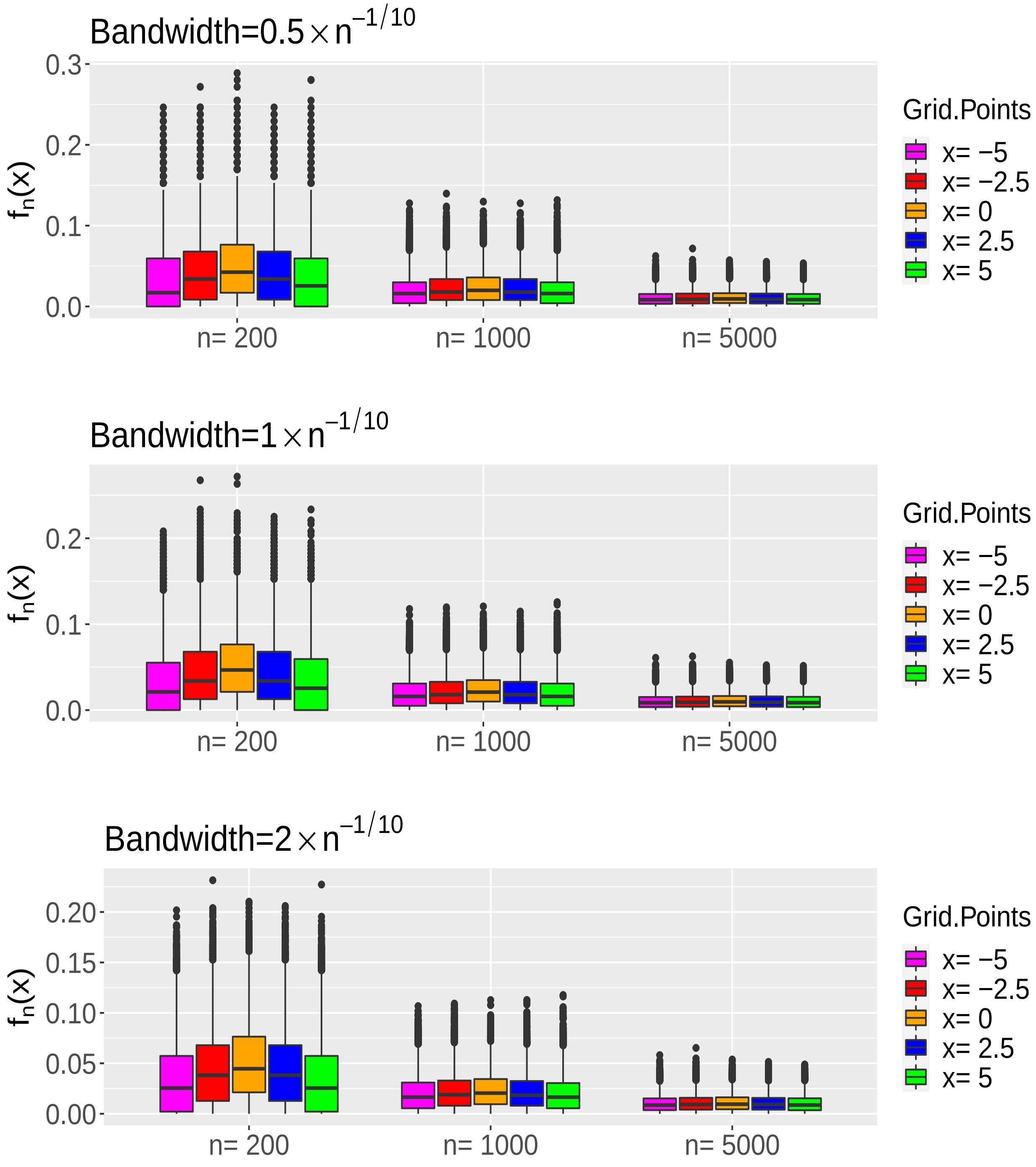


Figure 1.3: The kernel density estimator for the random walk case ($\delta = 0$). Top panel: Boxplots for $f_n(x)$ with bandwidth, $d = 0.5$; Middle panel: Boxplots for $f_n(x)$ with bandwidth, $d = 1$; and Bottom panel: Boxplots for $f_n(x)$ with bandwidth, $d = 2$. In all panels, n is taken to be 200, 1000 and 5000, and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green).

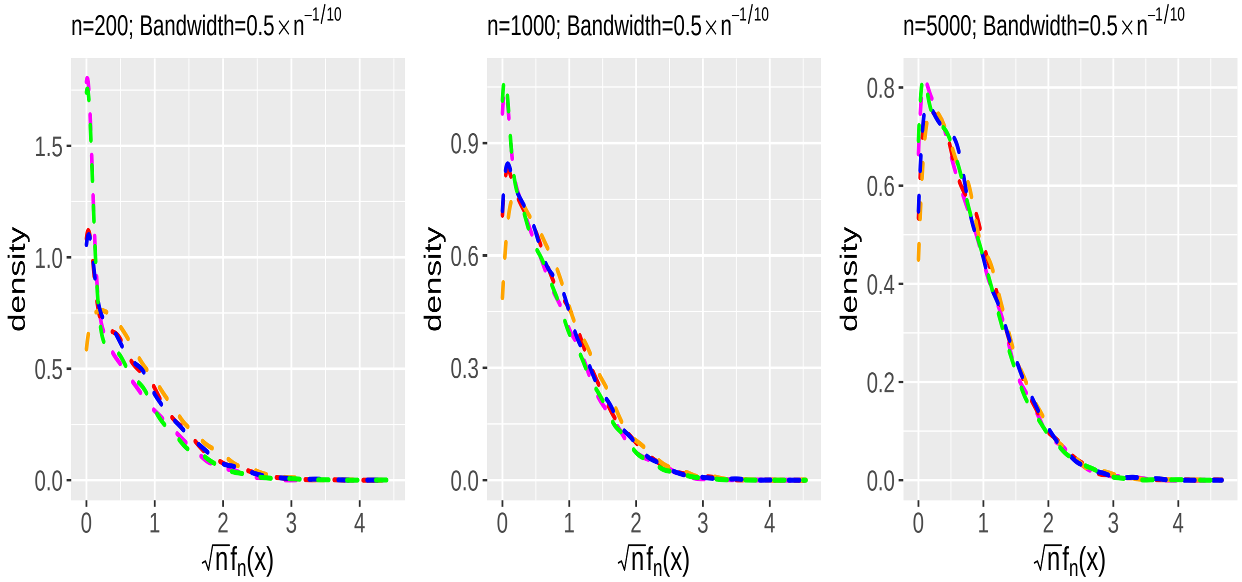


Figure 1.4: The kernel density estimator for the random walk case ($\delta = 0$). From the left to the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green).

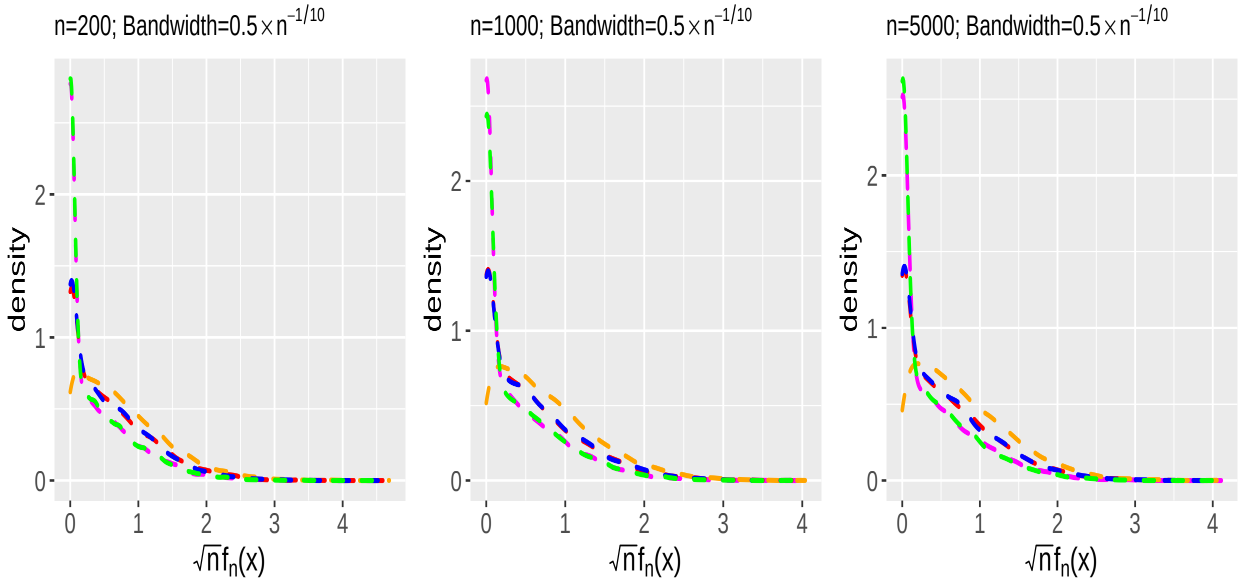


Figure 1.5: The kernel density estimator for the random walk case ($\delta = 0$). From the left and the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and $x = a\sqrt{n}$ with a taken to be -0.5 (magenta), -0.25 (red), 0 (orange), 0.25 (blue) and 0.5 (green).

estimation and **plot(ecdf())** for plotting the step function.

Table 1.2: The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 1$ (nearly integrated case).

The kernel density estimator for a nearly random walk

	$d = 0.5$					$d = 1$					$d = 2$				
n	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$	$x = -5$	$x = -2.5$	$x = 0$	$x = 2.5$	$x = 5$
200	0.025	0.042	0.059	0.042	0.025	0.030	0.047	0.055	0.047	0.030	0.032	0.049	0.057	0.049	0.034
	(0.042)	(0.045)	(0.046)	(0.045)	(0.043)	(0.040)	(0.042)	(0.042)	(0.042)	(0.040)	(0.038)	(0.039)	(0.039)	(0.039)	(0.038)
1000	0.020	0.024	0.026	0.024	0.020	0.021	0.024	0.026	0.024	0.021	0.021	0.024	0.026	0.024	0.021
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
5000	0.011	0.012	0.012	0.011	0.011	0.011	0.011	0.012	0.011	0.011	0.011	0.012	0.012	0.011	0.011
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)

1.2.2 Optimality

As we already have shown that

$$E\left(f_n(x)\right) = f(x) + \frac{h^2}{2}\mu_2(K)f''(x) + o\left(h^2\right), \quad \text{and} \quad \operatorname{Var}\left(f_n(x)\right) = \frac{\nu_0(K)f(x)}{nh} + o\left((nh)^{-1}\right),$$

so that the asymptotic mean integrated squares error (AMISE) is

$$\text{AMISE} = \frac{h^4}{4}\mu_2^2(K)\int \left[f''(x)\right]^2 + \frac{\nu_0(K)}{nh}.$$

Minimizing the AMISE gives the

$$h_{\text{opt}} = C_1(K)\left\|f''\right\|_2^{-2/5} n^{-1/5} = d\, n^{-1/5}, \tag{1.5}$$

where $C_1(K) = \left[\nu_0(K)/\mu_2^2(K)\right]^{1/5}$ and $d = C_1(K)\left\|f''\right\|_2^{-2/5}$ depending on $K(\cdot)$ and $f(\cdot)$. With this asymptotically optimal bandwidth, the optimal AMISE is given by

$$\text{AMISE}_{\text{opt}} = \frac{5}{4}C_2(K)\left\|f''\right\|_2^{2/5} n^{-4/5},$$

where $C_2(K) = \left[\nu_0^2(K)\mu_2(K)\right]^{2/5}$. To choose the best kernel, it suffices to choose one to minimize $C_2(K)$.

Proposition 1: *The nonnegative probability density function K minimizing $C_2(K)$ is a re-scaling of the Epanechnikov kernel:*

$$K_{opt}(u) = \frac{3}{4a}\left(1 - u^2/a^2\right)_+$$

for any $a > 0$.

Proof: First of all, we note that $C_2(K_h) = C_2(K)$ for any $h > 0$. Let K_0 be the Epanechnikov kernel. For any other nonnegative K, by re-scaling if necessary, we assume that $\mu_2(K) = \mu_2(K_0)$. Thus, we need only to show that $\nu_0(K_0) \le \nu_0(K)$. Let $G = K - K_0$. Then,

$$\int G(u)\,du = 0 \text{ and } \int u^2 G(u)\,du = 0,$$

which implies that

$$\int \left(1-u^2\right) G(u)\,du = 0.$$

Using this and the fact that $K_0(\cdot)$ has the support $[-1,1]$, we have

$$\begin{aligned}\int G(u)K_0(u)\,du =& \frac{3}{4}\int_{|u|\le 1} G(u)\left(1-u^2\right)\,du \\ =& -\frac{3}{4}\int_{|u|>1} G(u)\left(1-u^2\right)\,du = \frac{3}{4}\int_{|u|>1} K(u)\left(u^2-1\right)\,du.\end{aligned}$$

Since $K(\cdot)$ is nonnegative, so is the last term. Therefore,

$$\int K^2(u)\,du = \int K_0^2(u)\,du + 2\int K_0(u)G(u)\,du + \int G^2(u)\,du \ge \int K_0^2(u)\,du,$$

which proves that $K_0(\cdot)$ is the optimal kernel. □

Remark 1.3: *This proposition implies that the Epanechnikov kernel with $a = 1$ should be used in practice. Clearly, the Epanechnikov kernel is symmetric and has a finite support as well as is differentiable within its support. The difference between the Epanechnikov and Gaussian kernels can be evidenced from Figure 1.6. As seen in Figure 1.1b, the difference of using two kernels to estimate $f(x)$ is not distinguishable.*

1.2.3 Data-Driven Bandwidth Selection Methods

A. Simple Bandwidth Selectors

I. Normal Reference

The optimal bandwidth (1.5) is not directly usable since it depends on the unknown parameter $\|f''\|_2$. When $f(x)$ is a Gaussian density with standard deviation σ, it is easy to see from (1.5) that $\|f''\|_2^2 = 3/[8\sqrt{\pi}\sigma^5]$ so that

$$h_{opt} = (8\sqrt{\pi}/3)^{1/5} C_1(K)\sigma n^{-1/5},$$

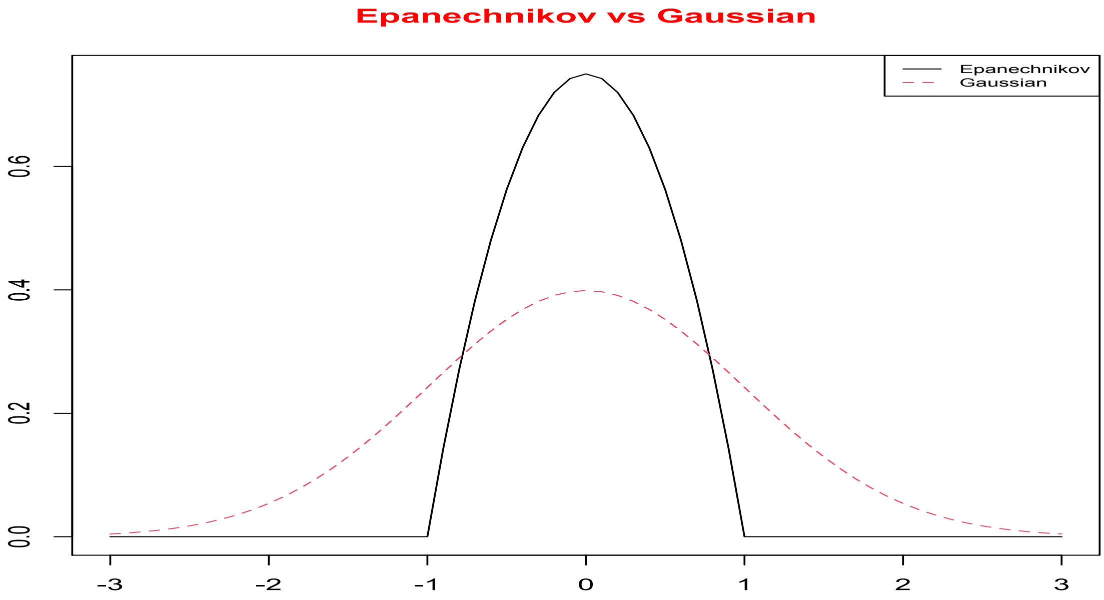


Figure 1.6: The Epanechnikov and Gaussian kernels.

which is called the normal reference bandwidth selector in literature, obtained by replacing the unknown parameter σ in the above equation by its sample standard deviation s. In particular, after calculating the constant $C_1(K)$ numerically, we have the following normal reference bandwidth selector

$$\widehat{h}_{opt,n} = \begin{cases} 1.06\, s\, n^{-1/5} & \text{for the Gaussian kernel} \\ 2.34\, s\, n^{-1/5} & \text{for the Epanechnikov kernel.} \end{cases}$$

Clearly, if the true density of X_t is close to normal, then, the normal reference bandwidth selector should work well and it is often used in practice due to its simplicity. Of course, the true density of X_t is not close to normal, say, having multiple modes, then, the normal reference bandwidth selector should not be used.

If $f(x)$ has a unique mode, one might use Laplace (saddle-point) approximation to $f(x)$ as

$$f(x) \approx f(x_m)\sqrt{2\,\pi}\sigma_m\, \phi((x - x_m)/\sigma_m),$$

where x_m is the mode of $f(x)$, $\phi(x)$ is the density of the standard normal, and $\sigma^2 = -1/h''(x_m)$ with $h(x) = f'(x)/f(x)$. Then, one can use the normal reference bandwidth selector multiply a constance, which might need an estimate.

II. Edgeworth Expansion

Hjort and Jones (1996b) proposed an improved rule obtained by using an Edgeworth expansion for $f(x)$ around the Gaussian density. Such a rule is given by

$$\widehat{h}_{opt,e} = \widehat{h}_{opt,n}\left(1 + \frac{35}{48}\widehat{\gamma}_4 + \frac{35}{32}\widehat{\gamma}_3^2 + \frac{385}{1024}\widehat{\gamma}_4^2\right)^{-1/5},$$

where $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are respectively the sample skewness and kurtosis. For details about the Edgeworth expansion, please see the book by Hall (1992).

III. Plug-in Method

Note that the normal reference bandwidth selector is only a simple rule of thumb. It is a good selector when the data are nearly Gaussian distributed, and is often reasonable in many applications. However, it can lead to over-smooth when the underlying distribution is asymmetric or multi-modal. In that case, one can either subjectively tune the bandwidth, or select the bandwidth by more sophisticated bandwidth selectors. One can also transform data first to make their distribution closer to normal, then estimate the density using the normal reference bandwidth selector and apply the inverse transform to obtain an estimated density for the original data. Such a method is called the transformation method. There are quite a few important techniques for selecting the bandwidth such as cross-validation (CV) and plug-in bandwidth selectors. A conceptually simple technique, with theoretical justification and good empirical performance, is the plug-in technique. This technique relies on finding an estimate of the functional $\|f''\|_2$, which can be obtained by using a pilot bandwidth. An implementation of this approach is proposed by Sheather and Jones (1991) and an overview on the progress of bandwidth selection can be found in Jones, Marron and Sheather (1996).

Function **dpik()** in the package **KernSmooth** in **R** selects a bandwidth for estimating the kernel density estimation using the plug-in method.

IV. Cross-Validation Method

The integrated squared error (ISE) of $f_n(x)$ is defined by

$$\text{ISE}(h) = \int [f_n(x) - f(x)]^2\, dx.$$

A commonly used measure of discrepancy between $f_n(x)$ and $f(x)$ is the mean integrated squared error (MISE) $\text{MSE}(h) = E[\text{ISE}(h)]$. It can be shown easily (or see Chiu, 1991) that

MISE(h) $\approx$ AMISE(h). The optimal bandwidth minimizing the AMISE is given in (1.5). The least squares cross-validation (LSCV) method proposed by Rudemo (1982) and Bowman (1984) is a popular method to estimate the optimal bandwidth h_{opt} . Cross-validation is very useful for assessing the performance of an estimator via estimating its prediction error. The basic idea is to set one of the data point aside for validation of a model and use the remaining data to build the model. The main idea is to choose h to minimize ISE(h). Since

$$\text{ISE}(h)=\int f_n^2(x)dx-2\int f(x)f_n(x)dx+\int f^2(x)dx,$$

the question is how to estimate the second term on the right hand side. Well, let us consider the simplest case when $\{X_t\}$ are iid. Re-express $f_n(x)$ as

$$f_n(x)=\frac{n-1}{n}f_n^{(-s)}(x)+\frac{1}{n}K_h\left(X_s-x\right)$$

for any $1\leq s\leq n$, where

$$f_n^{(-s)}(x)=\frac{1}{n-1}\sum_{t\neq s}^{n}K_h\left(X_t-x\right),$$

which is the kernel density estimate without the sth observation, commonly called the **jackknife** estimate or leave-one-out estimate. It is easy to see that for any $1\leq s\leq n$,

$$f_n(x)\approx f_n^{(-s)}(x).$$

Let $\mathcal{D}_{-s}=\{X_1,\cdots,X_{s-1},X_{s+1},\cdots,X_n\}$. Then,

$$E\left[f_n^{(-s)}\left(X_s\right)\mid\mathcal{D}_{-s}\right]\equiv^{3}\int f_n^{(-s)}(x)f(x)dx\approx\int f_n(x)f(x)dx, \tag{1.6}$$

if $\{X_i\}$ are iid, which, by using the method of moment, can be estimated by $\frac{1}{n}\sum_{s=1}^{n}f_n^{(-s)}\left(X_s\right)$. Therefore, the cross-validation is

$$\text{CV}(h)=\int f_n^2(x)dx-\frac{2}{n}\sum_{s=1}^{n}f_n^{(-s)}\left(X_s\right)=\frac{1}{n^2}\sum_{s,t}K_h^*\left(X_s-X_t\right)-\frac{2}{n(n-1)}\sum_{t\neq s}^{n}K_h\left(X_s-X_t\right),$$

where $K_h^*(\cdot)$ is the convolution of $K_h(\cdot)$ and $K_h(\cdot)$ as

$$K_h^*(u)=\int K_h(v)K_h(u-v)dv.$$

[3] This equality holds only for iid data but not for dependent data.

Let $\widehat{h}_{cv}$ be the minimizer of $\text{CV}(h)$. Then, it is called the optimal bandwidth based on the cross-validation. Stone (1984) showed that $\widehat{h}_{cv}$ is a consistent estimate of the optimal bandwidth h_{opt} in the sense that $\widehat{h}_{cv}/h_{\text{opt}}$ converges to 1 in probability.

Function **lscv()** in the package **locfit** in **R** selects a bandwidth for estimating the kernel density estimation using the least squares cross-validation method.

Remark 1.4: *Note that the above cross-validation method does not work well for time seres cases since (1.6) holds only for the iid data. Indeed, the leave-one-out cross-validation method was challenged by Shao (1993), which claimed that the popular leave-one-out cross-validation method, which is asymptotically equivalent to many other model selection methods such as the Akaike Information Criterion (AIC), the C_p, and the Bootstrap, is asymptotically inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$ and also, Shao (1993) showed that the inconsistency of the leave-one-out cross-validation can be rectified by using a leave-n_ν-out cross-validation with n_ν (block-wise cross-validation), the number of observations reserved for validation, satisfying $n_\nu/n \to 0$ and $n_\nu \to \infty$ as $n \to \infty$. The reader is referred to the paper by Shao (1993) for details.*

Finally, to pay attention to the structure of stationary time series data, one can use other data-driven methods to choose $\widehat{h}$ such as the nonparametric AIC[4] approach proposed in Cai and Tiwari (2000); see details in Section 2.3.5 or the modified multi-fold cross-validation criterion as in Cai, Fan and Yao (2000); see Section 2.4.3 for details.

1.2.4 Boundary Problems

In many applications, the density $f(\cdot)$ has a bounded support. For example, the interest rate can not be less than zero and the income is always nonnegative. It is reasonable to assume that the interest rate has support $[0, 1]$. However, because a kernel density estimator spreads smoothly point masses around the observed data points, some of those near the boundary of the support are distributed outside the support of the density. Therefore, the kernel density estimator under estimates the density in the boundary regions. The problem is more severe for large bandwidth and for the left boundary where the density is high. Therefore, some adjustments are needed. To gain some further insights, let us assume without loss

[4] For the detailed information, please see my lecture notes on **"A Summary of Model Selection Methods"**.

of generality that the density function $f(\cdot)$ has a bounded support $[0,1]$ and we deal with the density estimate at the left boundary. For simplicity, suppose that $K(\cdot)$ has a support $[-1,1]$. For the left boundary point $x = ch(0 \leq c < 1)$, it can easily be seen that as $h \to 0$

$$\begin{aligned} E\left(f_n(ch)\right) = \int_{-c}^{1/h-c} f(ch+hu)K(u)\,du &= f(0+)\mu_{0,c}(K) + hf'(0+)\left[c\mu_{0,c}(K) + \mu_{1,c}(K)\right] \\ &+ o(h), \end{aligned} \tag{1.7}$$

where $f(0+) = \lim_{x\downarrow 0} f(x)$,

$$\mu_{j,c} = \int_{-c}^{\infty} u^j K(u)du, \quad \text{and} \quad \nu_{j,c}(K) = \int_{-c}^{\infty} u^j K^2(u)du.$$

Also, we can show that $\text{Var}\left(f_n(ch)\right) = O(1/nh)$. Therefore,

$$f_n(ch) = f(0+)\mu_{0,c}(K) + hf'(0+)\left[c\mu_{0,c}(K) + \mu_{1,c}(K)\right] + o_p(h).$$

Particularly, if $c = 0$ and $K(\cdot)$ is symmetric, then $E\left(f_n(0)\right) = f(0)/2 + o(1)$.

There are several methods to deal with the density estimation at boundary points. Possible approaches include the boundary kernel (see Gasser and Müller (1979) and Müller (1993)), reflection (see Schuster (1985) and Hall and Wehrly (1991)), transformation (see Wand, Marron and Ruppert (1991) and Marron and Ruppert (1994)) and local polynomial fitting (see Hjort and Jones (1996a) and Loader (1996)), and others.

A. Boundary Kernel

One way of choosing a boundary kernel is

$$K_{(c)}(u) = \frac{12}{(1+c)^4}(1+u)\left\{(1-2c)u + \frac{3c^2-2c+1}{2}\right\} I_{[-1,c]}.$$

Note $K_{(1)}(t) = K(t)$, the Epanechnikov kernel as defined above. Moreover, Zhang and Karunamuni (1998) showed that this kernel is optimal in the sense of minimizing the MSE in the class of all kernels in order $(0,2)$ with exactly one change of sign in their support. The downside to the boundary kernel is that it is not necessarily non-negative, as will be seen on densities where $f(0) = 0$. Actually, this boundary kernel is commonly used in applied research.

B. Reflection

The reflection method is to construct the kernel density estimate based on the synthetic data $\{\pm X_t; 1 \leq t \leq n\}$ where "reflected" data are $\{-X_t; 1 \leq t \leq n\}$ and the original data a re $\{X_t; 1 \leq t \leq n\}$. This results in the estimate

$$f_n(x) = \frac{1}{n}\left\{\sum_{t=1}^{n} K_h\left(X_t - x\right) + \sum_{t=1}^{n} K_h\left(-X_t - x\right)\right\}, \qquad \text{for } x \geq 0.$$

Note that when x is away from the boundary, the second term in the above is practically negligible. Hence, it only corrects the estimate in the boundary region. This estimator is twice the kernel density estimate based on the synthetic data $\{\pm X_t; 1 \leq t \leq n\}$. See Schuster (1985) and Hall and Wehrly (1991).

C. Transformation

The transformation method is to first transform the data by $Y_t = g\left(X_t\right)$, where $g(\cdot)$ is a given monotone increasing function, ranging from $-\infty$ to ∞. Now apply the kernel density estimator to this transformed data set to obtain the estimate $f_n(y)$ for Y and apply the inverse transform to obtain the density of X. Therefore,

$$f_n(x) = |g'(x)| \frac{1}{n} \sum_{t=1}^{n} K_h\left(g\left(X_t\right) - g(x)\right).$$

The density at $x = 0$ corresponds to the tail density of the transformed data since $\log(0) = -\infty$, which can not usually be estimated well due to lack of the data at tails. Except at this point, the transformation method does a fairly good job. If $g(\cdot)$ is unknown in applications, similar to the Box-Cox transformation, Karunamuni and Alberts (2005) suggested a parametric form and then estimated the parameter by the profile likelihood estimation. Also, Karunamuni and Alberts (2005) considered other types of transformations.

D. Local Likelihood Fitting

The main idea is to consider the approximation $\log\left(f\left(X_t\right)\right) \approx P\left(X_t - x\right)$, where $P(u-x) = \sum_{j=0}^{p} a_j(u-x)^j$ with the localized version of log-likelihood

$$\sum_{t=1}^{n} \log\left(f\left(X_t\right)\right) K_h\left(X_t - x\right) - n \int K_h(u - x) f(u) du.$$

With this approximation, the local likelihood proposed in Tibshirani and Hastie (1987) is employed here to estimate $f(x)$, as

$$\mathcal{L}\left(a_0, \cdots, d_p\right) = \sum_{t=1}^{n} P\left(X_t - x\right) K_h\left(X_t - x\right) - n \int K_h(u-x) \exp(P(u-x)) du.$$

Let $\{\widehat{a}_j\}$ be the maximizer of the above local likelihood $\mathcal{L}\left(a_0, \cdots, d_p\right)$. Then, the local likelihood density estimate is

$$f_n(x) = \exp\left(\widehat{a}_0\right).$$

If the maximizer does not exist, then, $f_n(x) = 0$. See Loader (1996) and Hjort and Jones (1996a) for more details. If **R** is used for the local fit for density estimation, please use the function **density.lf()** in the package **locfit**.

Exercise: Please conduct a Monte Carol simulation to see what the boundary effects are and how the correction methods work. For example, you can consider some distribution densities with a finite support such as beta-distribution.

Remark 1.5: *From Cai (2011), it is very interesting to know that the boundary problem does not exist for the Rosenblatt-Parzen estimator when X_t is nonstationary since X_t has a higher probability of taking very large values. Indeed, as argued by Cai (2011), for any fixed a, one has*

$$P(|X_t| \geq a\sqrt{n}) = P(|X_t|/\sigma_u\sqrt{n} \geq a/\sigma_u) \approx P(|W_u(r)| \geq a/\sigma_u) = 2[1 - \Phi(a/\sqrt{r}\sigma_u)] > 0,$$

where $t = rn$ for $0 < r < 1$, $W_u(\cdot)$ is the standard Brownian motion generated by $\{u_t\}$, and $\Phi(\cdot)$ is the CDF for the standard normal.

1.2.5 Curse of Dimensionality

As we discussed earlier, the kernel density or distribution estimation is basically one-dimensional. For the multivariate case, the kernel density estimate is given by

$$f_n(x) = \frac{1}{n} \sum_{t=1}^{n} K_H\left(X_t - x\right), \tag{1.8}$$

where $K_H(u) = K\left(H^{-1}u\right) / \det(H), K(u)$ is a multivariate kernel function, and H is the bandwidth matrix such as for all $1 \leq i, j \leq p, n\, h_{ij} \to \infty$ and $h_{ij} \to 0$ where h_{ij} is the (i, j)

th element of H. The bandwidth matrix is introduced to capture the dependent structure in the independent variables. Particularly, if H is a diagonal matrix and $K(u) = \prod_{j=1}^{p} K_j\left(u_j\right)$ where $K_j(\cdot)$ is a univariate kernel function, then, $f_n(x)$ becomes

$$f_n(x) = \frac{1}{n} \sum_{t=1}^{n} \prod_{j=1}^{p} K_{h_j}\left(X_{jt} - x_j\right),$$

which is called the product kernel density estimation. This case is commonly used in practice. Similar to the univariate case, it is easy to derive the theoretical results for the multivariate case, which is left as an exercise. See Wand and Jones (1995) for details.

For the product kernel estimate with $h_j = h$, we can show easily that

$$E\left(f_n(x)\right) = f(x) + \frac{h^2}{2} \operatorname{tr}\left(\mu_2(K) f''(x)\right) + o\left(h^2\right),$$

where $\mu_2(K) = \int u\, u^T K(u) du$, and

$$\operatorname{Var}\left(f_n(x)\right) = \frac{\nu_0(K) f(x)}{nh^p} + o\left((nh^p)^{-1}\right),$$

so that the AMSE is given by

$$\text{AMSE} = \frac{\nu_0(K) f(x)}{nh^p} + \frac{h^4}{4} B(x),$$

where $B(x) = \left(\operatorname{tr}\left(\mu_2(K) f''(x)\right)\right)^2$. By minimizing the AMSE, we obtain the optimal bandwidth

$$h_{\text{opt}} = \left(\frac{p\nu_0(K) f(x)}{B(x)}\right)^{1/(p+4)} n^{-1/(p+4)},$$

which leads to the optimal rate of convergence for MSE which is $O\left(n^{-4/(4+p)}\right)$ by trading off the rates between the bias and variance. When p is large, the so called *curse of dimensionality* exists. To understand this problem quantitatively, let us look at the rate of convergence. To have a comparable performance with one-dimensional nonparametric regression with n_1 data points, for p-dimensional nonparametric regression, we need the number of data points n_p,

$$O\left(n_p^{-4/(4+p)}\right) = O\left(n_1^{-4/5}\right),$$

or $n_p = O\left(n_1^{(p+4)/5}\right)$. Note that here we only emphasize on the rate of convergence for MSE by ignoring the constant part. Table 1.3 shows the result with $n_1 = 100$. The increase of required sample sizes for higher dimension is in a polynomial rate.

Table 1.3: Sample sizes required for p-dimensional nonparametric estimate to have comparable performance with that of 1-dimensional nonparametric estimate using size $n_1 = 100$.

dimension (p)	2	3	4	5	6	7	8	9	10
sample size (n_p)	252	631	1,585	3,982	10,000	25,119	63,096	158,490	398,108

Exercise: Please derive the asymptotic results given in (1.8) for the general multivariate case.

In **R**, the built-in function **density()** is only for univariate case. For multivariate situations, there are two packages **ks** and **KernSmooth**. Function **kde()** in **ks** can compute the multivariate density estimate for 2 to 6 dimensional data and Function **bkde2D()** in **KernSmooth** computes the 2D kernel density estimate. Also, **ks** provides some functions for some bandwidth matrix selection such as **Hbcv()** and **Hscv()** for 2D case and **Hlscv()** and **Hpi()**.

1.2.6 Reading Materials

Applications in Finance: Please read the papers by Aït-Sahalia and Lo (1998, 2000), Pritsker (1998) and Hong and Li (2005) on how to apply the kernel density estimation to the nonparametric estimation of the state-price densities (SPD) or risk neutral densities and nonparametric risk estimation based on the state-price density. Please download the data from http://finance.yahoo.com/ (say, S&P500 index) to estimate the SPD.

1.3 Semiparametric Estimation of Density Function

To overcome the so-called *curse of dimensionality* as described in Section 1.2.5, Gallant and Nychka (1987) proposed the so-called semiparametric estimation to density function, termed as SNP estimation. The SNP estimator is based on the class of densities

$$\mathbb{F}_K = \left\{ f_K : f_K(x, \boldsymbol{\theta}) = \left[\sum_{i=0}^{K} \theta_i x^i \right]^2 \exp(-x^2/2) + \epsilon_0 \phi(x), \boldsymbol{\theta} \in \Theta_K \right\},$$

where $\Theta_K = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K), \int f_K(x, \boldsymbol{\theta}) dx = 1 \right\}$, ϕ denotes the standard normal density, ϵ_0 is a small positive number, and $K = 0, 1, ldots$. Estimation is by quasi-maximum likelihood

$$\widehat{f}_K = \text{argmax}_{f \in \mathbb{F}_K} \sum_{i=1}^{n} \log \left[\frac{1}{\sigma} f \left(\frac{x_i - \mu}{\sigma} \right) \right],$$

where $K = K_n \to \infty$ in some way. It is convenient to rewrite the SNP density in terms of normalized Hermite polynomials

$$H_{e_i}(x) = (\sqrt{2\pi} i!)^{-1/2} \sum_{m=0}^{\lfloor i/2 \rfloor} (-1)^m \frac{i!}{m! 2^m (i-2m)!} x^{i-2m},$$

which is the so-called Hermite series expansion estimator of density function. Now, the above $\mathbb{F}_K$ becomes to the following

$$\mathbb{F}_n = \left\{ f_n : f_n(x, \boldsymbol{\theta}) = \left[\sum_{i=0}^{K_n} \theta_i H_{e_i}(x) \right]^2 \exp(-x^2/2) + \epsilon_0 \phi(x), \boldsymbol{\theta} \in \Theta_K \right\}$$

with $\Theta_n = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{K_n}), \sum_{i=0}^{K_n} \theta_i^2 + \epsilon_0 = 1 \right\}$. Clearly, the choice of K_n is very crucial, similar to that for h in Section 1.2.3. In practice, one can use the cross-validation described in Section 1.2.3 to choose K_n empirically. For details, please see the papers by Gallant and Nychka (1987) and Coppejansa and Gallant (2002) for the asymptotic theory and practical issues.

1.4 Applications

1.4.1 Distribution Estimation

A. Smoothed Distribution Estimation

The question is how to obtain a smoothed estimate of CDF $F(x)$. Well, one way of doing so is to integrate the estimated PDF $f_n(x)$, given by

$$\widehat{F}_n(x) = \int_{-\infty}^{x} f_n(u) du = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h} \right),$$

where $\mathcal{K}(x) = \int_{-\infty}^{x} K(u) du$; the distribution of $K(\cdot)$. Why do we need this smoothed estimate of CDF? To answer this question, we need to consider the mean squares error.

First, we derive the asymptotic bias. By the integration by parts, we have

$$\begin{aligned} E\left[\widehat{F}_n(x)\right] &= E\left[\mathcal{K}\left(\frac{x - X_i}{h}\right)\right] = \int F(x - hu) K(u) du \\ &= F(x) + \frac{h^2}{2} \mu_2(K) f'(x) + o\left(h^2\right) \end{aligned}$$

Next, we derive the asymptotic variance.

$$E\left[\mathcal{K}^2\left(\frac{x - X_i}{h}\right)\right] = \int F(x - hu) b(u) du = F(x) - h f(x) \theta + o(h),$$

where $b(u) = 2K(u)\mathcal{K}(u)$ and $\theta = \int ub(u)du$. Then,

$$\operatorname{Var}\left[\mathcal{K}\left(\frac{x - X_i}{h}\right)\right] = F(x)[1 - F(x)] - hf(x)\theta + o(h).$$

Define $I_j(x) = \operatorname{Cov}\left(I\left(X_1 \leq x\right), I\left(X_{j+1} \leq t\right)\right) = F_j(x,x) - F^2(x)$ and

$$I_{nj}(x) = \operatorname{Cov}\left(\mathcal{K}\left(\frac{x - X_1}{h}\right), \mathcal{K}\left(\frac{x - X_{j+1}}{h}\right)\right).$$

By means of Lemma 2 in Lehmann (1966), the covariance $I_{nj}(x)$ may be written as follows

$$\begin{aligned} I_{nj}(t) = \int \Bigg\{ & P\left[\mathcal{K}\left(\frac{x - X_1}{h}\right) > u, \mathcal{K}\left(\frac{x - X_{j+1}}{h}\right) > v\right] \\ & - P\left[\mathcal{K}\left(\frac{x - X_1}{h}\right) > u\right] P\left[\mathcal{K}\left(\frac{x - X_{j+1}}{h}\right) > v\right]\Bigg\} dudv. \end{aligned}$$

Inverting the CDF, $\mathcal{K}(\cdot)$ and making two changes of variables, the above relation becomes

$$I_{nj}(x) = \int \left[F_j(x - hu, x - hv) - F(x - hu)F(x - hv)\right] K(u)K(v)dudv.$$

Expanding the right-hand side of the above equation according to Taylor's formula, we obtain

$$|I_{nj}(x) - I_j(x)| \leq C\, h^2.$$

By the Davydov's inequality (see Lemma1.1), we have

$$|I_{nj}(x) - I_j(x)| \leq C\, \alpha(j),$$

so that for any $1/2 < \tau < 1$,

$$|I_{nj}(x) - I_j(x)| \leq C\, h^{2\tau}\, \alpha^{1-\tau}(j).$$

Therefore,

$$\frac{1}{n}\sum_{j=1}^{n-1}(n - j)\left|I_{nj}(x) - I_j(x)\right| \leq \sum_{j=1}^{n-1}\left|I_{nj}(x) - I_j(x)\right| \leq C\, h^{2\tau}\sum_{j=1}^{\infty}\alpha^{1-\tau}(j) = O\left(h^{2\tau}\right)$$

provided that $\sum_{j=1}^{\infty}\alpha^{1-\tau}(j) < \infty$ for some $1/2 < \tau < 1$. Indeed, this assumption is satisfied if $\alpha(n) = O\left(n^{-\beta}\right)$ for some $\beta > 2$. By the stationarity, it is clear that

$$n\operatorname{Var}\left(\widehat{F}_n(x)\right) = \operatorname{Var}\left(\mathcal{K}\left(\frac{x - X_1}{h}\right)\right) + \frac{2}{n}\sum_{j=1}^{n-1}(n - j)I_{nj}(x).$$

Therefore,

$$
\begin{aligned}
n\text{Var}\left(\widehat{F}_n(x)\right) &= F(x)[1-F(x)] - hf(x)\theta + o(h) + 2\sum_{j=1}^{\infty} I_j(x) + O\left(h^{2\tau}\right) \\
&= \sigma_F^2(x) - hf(x)\theta + o(h).
\end{aligned}
$$

We can establish the following asymptotic normality for $\widehat{F}_n(x)$ but the proof will be discussed later.

Theorem 1.2: *Under some regularity conditions, we have*

$$\sqrt{n}\left[\widehat{F}_n(x) - F(x) - \frac{h^2}{2}\mu_2(K)f'(x) + o_p\left(h^2\right)\right] \quad \to \quad N\left(0, \sigma_F^2(x)\right).$$

Similarly, we have

$$n\,AMSE\left(\widehat{F}_n(x)\right) = \frac{nh^4}{4}\mu_2^2(K)\left[f'(x)\right]^2 + \sigma_F^2(x) - hf(x)\theta.$$

If $\theta > 0$, minimizing the AMSE gives the

$$h_{opt} = \left(\frac{\theta f(x)}{\mu_2^2(K)\left[f'(x)\right]^2}\right)^{1/3} n^{-1/3},$$

and with this asymptotically optimal bandwidth, the optimal AMSE is given by

$$n\,AMSE_{opt}\left(\widehat{F}_n(x)\right) = \sigma_F^2(x) - \frac{3}{4}\left(\frac{\theta^2 f^2(x)}{\mu_2(K)f'(x)}\right)^{2/3} n^{-1/3}.$$

Remark 1.6: *From the aforementioned equation, we can see that if $\theta > 0$, the AMSE of $\widehat{F}_n(x)$ can be smaller than that for $F_n(x)$ in the second order. Also, it is easy to that if $K(\cdot)$ is the Epanechnikov kernel, $\theta > 0$.*

B. Relative Efficiency and Deficiency

To measure the relative efficiency and deficiency of $\widehat{F}_n(x)$ over $F_n(x)$, we define

$$i(n) = \min\left\{k \in \{1, 2, \ldots\};\, \text{MSE}\left(F_k(x)\right) \le \text{MSE}\left(\widehat{F}_n(x)\right)\right\}$$

We have the following results without the detailed proofs which can be found in Cai and Roussas (1998).

Proposition 2: *(i) Under some regularity conditions,*

$$\frac{i(n)}{n} \;\to\; 1, \quad \textit{if and only if} \quad n h_n^4 \;\to\; 0.$$

(ii) Under some regularity conditions,

$$\frac{i(n)-n}{n\,h} \;\to\; \theta(x), \quad \textit{if and only if} \quad h_n^3 \;\to\; 0,$$

where $\theta(x)=f(x)\theta/\sigma_F^2(x)$.

Remark 1.7: *It is clear that the quantity $\theta(x)$ may be looked upon as a way of measuring the performance of the estimate $\widehat{F}_n(x)$. Suppose that the kernel $K(\cdot)$ is chosen, so that $\theta>0$, which is equivalent to $\theta(x)>0$. Then, for sufficiently large n, $i(n)>n+nh_n(\theta(x)-\varepsilon)$. Thus, $i(n)$ is substantially larger than n, and, indeed, $i(n)-n$ tends to ∞. Actually, Reiss (1981) and Falk (1983) posed the question of determining the exact value of the superiority of θ over a certain class of kernels. More specifically, let $\mathcal{K}_m$ be the class of kernels $\mathcal{K}:[-1,1]\to\Re$ which are absolutely continuous and satisfy the requirements: $\mathcal{K}(-1)=0, \mathcal{K}(1)=1$, and $\int_{-1}^1 u^\mu K(u)du=0, \mu=1,\cdots,m$, for some $m=0,1,\cdots$ (where the moment condition is vacuous for $m=0$). Set $\Psi_m=\sup\{\theta;\mathcal{K}\in\mathcal{K}_m\}$. Then, Mammitzsch (1984) answered the question posed by showing in an elegant manner. See Cai and Roussas (1998) for more details and simulation results.*

Exercise: Please conduct a Monte Carol simulation to see what the differences are for smoothed and non-smoothed distribution estimations.

1.4.2 Quantile Estimation

Let $X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$ denote the order statistics of $\{X_t\}_{t=1}^n$. Define the inverse of $F(x)$ as $F^{-1}(p)=\inf\{x\in\Re; F(x)\ge p\}$, where $\Re$ is the real line. The traditional estimate of $F(x)$ has been the empirical distribution function $F_n(x)$ based on $X_1,\dots,X_n$, while the estimate of the p-th quantile $\xi_p=F^{-1}(p), 0<p<1$, is the sample quantile function $\xi_{pn}=F_n^{-1}(p)=X_{([np])}$, where $[x]$ denotes the integer part of x. It is a consistent estimator of ξ_p for α-mixing data (Yoshihara, 1995). However, as stated in Falk (1983), $F_n(x)$ does not take into account the smoothness of $F(x)$; i.e., the existence of a probability density function $f(x)$. In order to incorporate this characteristic, investigators proposed several smoothed quantile estimates, one of which is based on $\widehat{F}_n(x)$ obtained as a convolution between $F_n(x)$

and a properly scaled kernel function; see the previous section. Finally, note that **R** has a command **quantile()** which can be used for computing ξ_{pn}, the nonparametric estimate of quantile.

1.4.3 Value-at-Risk and Expected Shortfall

Value at Risk (VaR) is a popular measure of market risk associated with an asset or a portfolio of assets. It has been chosen by the Basel Committee on Banking Supervision as a benchmark risk measure and has been used by financial institutions for asset management and minimization of risk. Let $\{X_t\}_{t=1}^n$ be the market value of an asset over n periods of $t = 1$ a time unit, and let $Y_t = -\log(X_t/X_{t-1})$ be the negative log-returns (loss). Suppose $\{Y_t\}_{j=1}^n$ is a strictly stationary dependent process with marginal distribution function $F(y)$. Given a positive value p close to zero, the $1 - p$ level VaR is

$$\nu_p = \inf\{u : F(u) \geq 1 - p\} = F^{-1}(1 - p),$$

which specifies the smallest amount of loss such that the probability of the loss in market value being larger than ν_p is less than p. Comprehensive discussions on VaR are available in Duffie and Pan (1997) and Jorion (2001), and references therein. Therefore, VaR can be regarded as a special case of quantile. **R** has a built-in package called **cvar** for a set of methods for calculation of VaR, particularly, for some parametric models such as the General Pareto Distribution (GPD). But the restrict parametric specifications might be misspecified.

A more general form for the generalized Pareto distribution with shape parameter $k \neq 0$, scale parameter σ, and threshold parameter θ, is

$$f(x) = \frac{1}{\sigma}\left(1 + k\frac{x-\theta}{\sigma}\right)^{-1/k-1}, \quad \text{and} \quad F(x) = 1 - \left(1 + k\frac{x-\theta}{\sigma}\right)^{-1/k}$$

for $\theta < x$, when $k > 0$. In the limit for $k = 0$, the density is $f(x) = \frac{1}{\sigma}\exp(-(x-\theta)/\sigma)$ for $\theta < x$. If $k = 0$ and $\theta = 0$, the generalized Pareto distribution is equivalent to the exponential distribution. If $k > 0$ and $\theta = \sigma$, the generalized Pareto distribution is equivalent to the Pareto distribution.

Another popular risk measure is the expected shortfall (ES) which is the expected loss, given that the loss is at least as large as some given quantile of the loss distribution (e.g., VaR), defined as

$$\mu_p = E\left(Y_t \mid Y_t > \nu_p\right) = \int_{\nu_p}^{\infty} y f(y) dy/p.$$

It is well known from Artzner, Delbaen, Eber and Heath (1999) that ES is a coherent risk measure such as it satisfies the four axioms: homogeneity (increasing the size of a portfolio by a factor should scale its risk measure by the same factor), monotonicity (a portfolio must have greater risk if it has systematically lower values than another), risk-free condition or translation invariance (adding some amount of cash to a portfolio should reduce its risk by the same amount), and subadditivity (the risk of a portfolio must be less than the sum of separate risks or merging portfolios cannot increase risk). VaR satisfies homogeneity, monotonicity, and risk-free condition but is not sub-additive. See Artzner, et al. (1999) for details.

1.4.4 Smoothed Quantile Estimation

The smoothed sample quantile estimate of $\xi_p, \widehat{\xi}_p$, based on $\widehat{F}_n(x)$, is defined by:

$$\widehat{\xi}_p = \widehat{F}_n^{-1}(1-p) = \inf \left\{ x \in \Re; \widehat{F}_n(x) \geq 1-p \right\}.$$

$\widehat{\xi}_p$ is referred to in literature as the perturbed (smoothed) sample quantile. Asymptotic properties of $\widehat{\xi}_p$, both under independence as well as under certain modes of dependence, have been investigated extensively in literature; see Cai and Roussas (1997) and Chen and Tang (2005).

By the differentiability of $\widehat{F}_n(x)$, we use the Taylor expansion and ignore the higher terms to obtain

$$\widehat{F}_n\left(\widehat{\xi}_p\right) = 1-p \approx \widehat{F}_n\left(\xi_p\right) - f_n\left(\xi_p\right)\left(\widehat{\xi}_p - \xi_p\right), \tag{1.9}$$

then,

$$\widehat{\xi}_p - \xi_p \approx \left[\widehat{F}_n\left(\xi_p\right) - (1-p)\right] / f_n\left(\xi_p\right) \approx \left[\widehat{F}_n\left(\xi_p\right) - (1-p)\right] / f\left(\xi_p\right),$$

since $f_n(x)$ is a consistent estimator of $f(x)$. As an application of Theorem 1.2, we can establish the following theorem for the asymptotic normality of $\widehat{\xi}_p$ but the proof is omitted since it is similar to that for Theorem 1.2.

Theorem 1.3: *Under some regularity conditions, we have*

$$\sqrt{n}\left[\widehat{\xi}_p - \xi_p - \frac{h^2}{2}\mu_2(K) f'\left(\xi_p\right) / f\left(\xi_p\right) + o_p\left(h^2\right)\right] \quad \rightarrow \quad N\left(0, \sigma_F^2\left(\xi_p\right) / f^2\left(\xi_p\right)\right).$$

Next, let us examine the AMSE. To this effect, from Theorem 1.3, it is easy to derive the asymptotic bias and variance, which are $h^2\mu_2(K)f'(\xi_p)/[2\,f(\xi_p)]$ and $\sigma_F^2\,(\xi_p)\,/f^2(\xi_p) - h\theta/f(\xi_p)$, respectively, so that the AMSE is given by

$$n\text{AMSE}\left(\widehat{\xi}_p\right) = \frac{nh^4}{4}\mu_2^2(K)\left[f'\left(\xi_p\right)/f\left(\xi_p\right)\right]^2 + \sigma_F^2\left(\xi_p\right)/f^2\left(\xi_p\right) - h\theta/f\left(\xi_p\right).$$

If $\theta > 0$, minimizing the AMSE gives the

$$h_{opt} = \left(\frac{\theta f\left(\xi_p\right)}{\mu_2^2(K)\left[f'\left(\xi_p\right)\right]^2}\right)^{1/3} n^{-1/3},$$

and with this asymptotically optimal bandwidth, the optimal AMSE is given by

$$n\text{AMSE}_{\text{opt}}\left(\widehat{\xi}_p\right) = \sigma_F^2\left(\xi_p\right)/f^2\left(\xi_p\right) - \frac{3}{4}\left(\frac{\theta^2}{\mu_2(K)f'\left(\xi_p\right)f\left(\xi_p\right)}\right)^{2/3} n^{-1/3},$$

which indicates a reduction to the AMSE of the second order. Chen and Tang (2005) conducted an intensive study on simulations to demonstrate the advantages of nonparametric estimation $\widehat{\xi}_p$ over the sample quantile ξ_{pn} under the VaR setting. We refer to the paper by Chen and Tang (2005) for simulation results and empirical examples.

Exercise: Please use the above procedures to estimate nonparametrically the ES and discuss its properties as well as conduct simulation studies and empirical applications.

1.5 Computer Code

```
###############
# Example 1.1
##############

###############################################
 # Define the Epanechnikov kernel function
  kernel<-function(x){0.75*(1-x^2)*(abs(x)<=1)}
    # Define the kernel density estimator
  kernden=function(x,z,h,ker){
    # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
```

```
   nz<-length(z)
   nx<-length(x)
   x0=rep(1,nx*nz)
   dim(x0)=c(nx,nz)
   x1=t(x0)
   x0=x*x0
   x1=z*x1
   x0=x0-t(x1)
   if(ker==1){x1=kernel(x0/h)}         # Epanechnikov kernel
   if(ker==0){x1=dnorm(x0/h)}          # normal kernel
   f1=apply(x1,2,mean)/h
   return(f1)
}

# Simulation for different bandwidths and different kernels
 n=300                                   # n=300

 ker=1                                    # ker=1 => Epan; ker=0 => Gaussian
 h0=c(0.25,0.5,1)                         # set initial bandwidths
 z=seq(-4,4,by=0.1)                       # grid points
 nz=length(z)                             # number of grid points
 x=rnorm(n)                               # simulate x ~ N(0, 1)
 if(ker==1){h_o=2.34*n^{-0.2}}  # bandwidth for Epanechnikov kernel
 if(ker==0){h_o=1.06*n^{-0.2}}  # bandwidth for normal kernel
 f1=kernden(x,z,h0[1],ker)
 f2=kernden(x,z,h0[2],ker)
 f3=kernden(x,z,h0[3],ker)
 f4=kernden(x,z,h_o,ker)
 text1=c("True","h=0.25","h=0.5","h=1","h=h_o")
 data=cbind(dnorm(z),f1,f2,f3,f4)        # combine them as a matrix

 quartz()
```

```
  matplot(z,data,type="l",lty=1:5,col=1:5,xlab="",ylab="")
  legend(-1,0.2,text1,lty=1:5,col=1:5)

  f5=density(x, kernel=c("gaussian"))$y
  z1=density(x, kernel=c("gaussian"))$x
  f6=density(x, kernel=c("epanechnikov"))$y
  data1=cbind(f5,f6)
  text2=c("Gaussian","Epanechnikov")

  quartz()
  matplot(z1,data1,type="l",lty=1:2,col=1:2,xlab="",ylab="")
  legend(-1,0.2,text2,lty=1:2,col=1:2)

  quartz()
  par(mfrow=c(1,2),mex=0.4,bg="light grey")
  matplot(z,data,type="l",lty=1:5,col=1:5,xlab="",ylab="")
  legend(-1,0.2,text1,lty=1:5,col=1:5)
  text1=c("Gauassian","Epanechnikov")
  matplot(z1,data1,type="l",lty=1:2,col=1:2,xlab="",ylab="")
  legend(-3,0.2,text2,lty=1:2,col=1:2)
###################################################

 ##################
 # Example 1.2
 ##################

##################################################
  z1=read.table(file="/NP_lecture_note/data/ex3-2.txt", header=F)
   # dada: weekly 3-month Treasury bill from 1954 to 2022
  x=z1[,4]/100                      # decimal
  n=length(x)
  y=diff(x)                         # Delta x_t=x_t-x_{t-1}=change rate
  x=x[1:(n-1)]
  n=n-1
  x_star=(x-mean(x))/sqrt(var(x))  # standardized
  den_3mtb=density(x_star,bw=0.30,kernel=c("epanechnikov"),
  from=-3,to=3,n=61)
  den_est=den_3mtb$y                # estimated density values
  z_star=seq(-3,3,by=0.1)
  text1=c("Estimated Density","Standard Norm")

  win.graph()  # for Windows
  # quartz()   # for macOS
  par(bg="light green")
  plot(den_3mtb,main="Density of 3mtb (Buind-in)",ylab="",xlab="",
  col.main="red")
  points(z_star,dnorm(z_star),type="l",lty=2,col=2,ylab="",xlab="")
  legend(0,0.45,text1,lty=c(1,2),col=c(1,2),cex=0.7)

  h_den=0.5
  f_hat=kernden(x_star,z_star,h_den,1)
  ff=cbind(f_hat,dnorm(z_star))

  win.graph()
  par(bg="light blue")
  matplot(z_star,ff,type="l",lty=c(1,2),col=c(1,2),ylab="",xlab="")
  title(main="Density of 3mtb",col.main="red")
  legend(0,0.55,text1,lty=c(1,2),col=c(1,2),cex=0.7)
####################################################


##############################
 # Example 1.3 (delta=0)
##############################
```

```
################################################################
 # Load needed packages
 library(ggplot2)
 library(tidyverse)
 library(ggpubr)
 set.seed(1)                          # to create reproducible results
 cols <- c("magenta", "red", "orange","blue","green")
#####################################################################

################################################################
 # Define the Rosenblatt-Parzen density estimator
 RP_dens_est<-function(x,h,z){
    # parameters: x=observed variable; h=bandwidth; z=grid point;
    nz<-length(z)
    nx<-length(x)
    x0=rep(1,nx*nz)
    dim(x0)=c(nx,nz)
    x1=t(x0)
    x0=x*x0
    x1=z*x1
    x0=x0-t(x1)
    x1=0.5*(abs(x0/h)<=1)                   # the uniform kernel
    f1=apply(x1,2,mean)/h
    return(f1)                              # return fn(z)
 }
#####################################################################

#############################################################################
 #  The Kernel Density Estimator for a Random Walk
 # Simulation for different bandiwidths, sample sizes and values of fixed x.
 rm(list = c())                           # clean the previous variables
```

```
x<-seq(-5,5,length.out=5)                # take 5 values of fixed x
nrep=1e4                                # repeat the simulation nrep times
ns= c(200,1000,5000)                      # sample size
delta=1
ds=c(0.5,1,2)
quest1<-list(NULL,NULL,NULL)              # fn(x)
for (n in ns) {                           # sample size
  for (i in 1:nrep) {
      Xt<-cumsum(rnorm(n))               # generate data from a random walk
      for (h in 1:length(ds)) {
             d<-ds[h]                       # bandwidth= d*n^(-1/10)
             quest1[[h]]<-c(quest1[[h]],    #compute fn(x)
                            RP_dens_est(Xt,h=d*n^(-1/10),x))
     }
   }
 }
tabmed<-list()                               # save median  of fn(x)
tabsd<-list()                                 # save sd of fn(x)
fig1<-list()                                   # save box-plots
for (h in 1:length(ds)) {
  d<-ds[h]
  Quest1<-data.frame(quest=quest1[[h]],       # rearrange simulated data
                     Grid.Points=factor(rep(paste("x=",x),nrep*length(ns)),
                                        levels = paste("x=",x)),
                     n=factor(rep(paste("n=",ns),each=nrep*length(x)),
                              levels = paste("n=",ns)))
  tabmed[[h]]<-(with(Quest1,
                    tapply(quest, list( n=n,Grid.Points=Grid.Points),median))
                %>%as.data.frame())
  tabsd[[h]]<-(with(Quest1,
                    tapply(quest, list( n=n,Grid.Points=Grid.Points),sd))
                %>%as.data.frame())
   fig1[[h]]<-Quest1%>%
     ggplot(aes(y=quest,x=n,fill=Grid.Points))+
     geom_boxplot()+
     scale_fill_manual(values = cols)+
     xlab("")+
     ylab(expression(paste(f[n],'(x)')))+
     labs(title=bquote(paste('Bandwidth=',.(d[1])%*%n^{-1/10})))+
     theme(axis.title = element_text(size=19),
           plot.title = element_text(size=21),
           axis.text= element_text(size=17),
           legend.text= element_text(size=17),
           legend.title= element_text(size=17))
    }
    write.csv(tabmed,"median_of_densityest_rw.csv")
    write.csv(tabsd,"sd_of_densityest_rw.csv")
    ggsave("rwbarplots.pdf", plot = do.call(ggarrange, c(fig1,ncol=1,nrow=3)),
           width = 24, height = 25, dpi =1500, bg = "white",units = "cm")
##################################################################

 #######################
 # Example 1.3 (delta=1)
 #######################
 ############################################################################
 #  The Kernel Density Estimator for a Nearly Random Walk
 # Simulation for different bandiwidths, sample sizes and values of fixed x.
 rm(list = c())
 x<-seq(-5,5,length.out=5)               # take 5 values of fixed x
 nrep=1e4                               # repeat the simulation nrep times
 ns= c(200,1000,5000)                    # sample size
 delta=1
 ds=c(0.5,1,2)
 quest1<-list(NULL,NULL,NULL)            # fn(x)
```

```
for (n in ns) {                         # sample size
   phi<-1-delta/n                         # coefficient for AR(1)
   Phi<-diag(1,ncol=n,nrow=n)
   for (j in 1:n) {
     Phi[j,-(1:j)]<-phi^(1:(n-j))
  }
 for (i in 1:nrep) {
   u<-matrix(rnorm(n),ncol=1)             # error term
   Xt<-as.numeric(Phi%*%u)              # generate data from a nearly random walk
   for (h in 1:length(ds)) {
     d<-ds[h]                           # bandwidth= d*n^(-1/10)
     quest1[[h]]<-c(quest1[[h]],           #compute fn(x)
                    RP_dens_est(Xt,h=d*n^(-1/10),x))
  }
 }
}
tabmed<-list()                            # save median  of fn(x)
tabsd<-list()                               # save sd of fn(x)
fig1<-list()                            # save box-plots
for (h in 1:length(ds)) {
  d<-ds[h]
  Quest1<-data.frame(quest=quest1[[h]],      # rearrange simulated data
                     Grid.Points=factor(rep(paste("x=",x),nrep*length(ns)),
                                           levels = paste("x=",x)),
   n=factor(rep(paste("n=",ns),each=nrep*length(x)),
            levels = paste("n=",ns)))
  tabmed[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),median))
                 %>%as.data.frame())
  tabsd[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),sd))
                 %>%as.data.frame())
  fig1[[h]]<-Quest1%>%
    ggplot(aes(y=quest,x=n,fill=Grid.Points))+
    geom_boxplot()+
    scale_fill_manual(values = cols)+
    xlab("")+
    ylab(expression(paste(f[n],'(x)')))+
    labs(title=bquote(paste('Bandwidth=',.(d[1])%*%n^{-1/10})))+
    theme(axis.title = element_text(size=19),
          plot.title = element_text(size=21),
          axis.text= element_text(size=17),
          legend.text= element_text(size=17),
          legend.title= element_text(size=17))
 }
 write.csv(tabmed,"median_of_densityest_nearrw.csv")
 write.csv(tabsd,"sd_of_densityest_nearrw.csv")
 ggsave("nearrwbarplots.pdf", plot = do.call(ggarrange, c(fig1,ncol=1,nrow=3)),
        width = 24, height = 25, dpi =1500, bg = "white",units = "cm")
##################################################################
```

1.6 References

Aït-Sahalia, Y. and Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance*, **53**(2), 499-547.

Aït-Sahalia, Y. and Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, **94**(1-2), 9-51.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**(3), 817-58.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, **9**(3), 203-228.

Billingsley, P. (1999). *Convergeence of Probability Measures.* Wiley, New York.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**(2), 353-360.

Cai, Z. (2002). Regression quantiles for time series. *Econometric theory*, **18**(1), 169-192.

Cai, Z. (2011). Nonparametric regression models with integrated covariates. In *Nonparametric Statistical Methods and Related Topics (Eds: J. Jiang, G.G. Roussas and F.J. Samaniego): A Festschrift in Honor of Professor P.K. Bhattacharya on his 80th Birthday*, pp.257-275.

Cai, Z., Li, Q. and Park J. (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, **148**(1), 101-113.

Cai, Z. and Masry, E. (2000). Nonparametric estimation of additive nonlinear ARX time series: Local linear fitting and projection. *Econometric Theory*, **16**(4), 465-501.

Cai, Z. and Roussas, G. G. (1997). Smooth estimate of quantiles under association. *Statistics and Probability Letters*, **36**(3), 275-287.

Cai, Z. and Roussas, G. G. (1998). Efficient estimation of a distribution function under quadrant dependence. *Scandinavian Journal of Statistics*, **25**(1), 211-224.

Cai, Z. and Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**(3), 341-350.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, **18**(1), 17-39.

Chen, S. X. and Tang, C. Y. (2005). Nonparametric inference of value-at-risk for dependent financial returns. *Journal of Financial Econometrics*, **3**(2), 227-255.

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *Annals of Statistics*, **19**(4), 1883-1905.

Coppejansa, M. and Gallant, R.A. (2002). Cross-validated SNP density estimates. *Journal of Econometrics*, **110**(1), 27-65.

Duffie, D. and Pan, J. (1997). An overview of value at risk. *Journal of Derivatives*, **4**(3), 7-49.

Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica*, **37**(2), 73-83.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: Nonparametric and parametric methods.* Springer-Verlag, New York.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (pp. 23-68). Springer-Verlag.

Gallant, A.R. and Nychka, D.W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica*, **55**(2), 363–390.

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

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hall, P. and Heyde, C. C. (2014). *Martingale Limit Theory and Its Application.* Academic Press, New York.

Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, **86**(415), 665-672.

Hjort, N. L. and Jones, M. C. (1996a). Locally parametric nonparametric density estimation. *Annals of Statistics*, **24**(4), 1619-1647.

Hjort, N. L. and Jones, M. C. (1996b). Better rules of thumb for choosing bandwidth in density estimation. Working Paper, Department of Mathematics, University of Oslo, Norway.

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, **18**(1), 37-84.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, **91**(433), 401-407.

Jorion, P. (2001). *Value at Risk*, 2nd edition. McGraw-Hill, New York.

Karunamuni, R. J. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, **2**(3), 191-212.

Lehmann, E. L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, **37**(5), 1137-1153.

Loader, C. R. (1996). Local likelihood density estimation. *Annals of Statistics*, **24**(4), 1602-1618.

Mammitzsch, V. (1984). On the asymptotically optimal solution within a certain class of kernel type estimators. *Statistics & Risk Modeling*, **2**(3-4), 247-256.

Marcus, M. and Rosen, J. (2006). *Markov Processes, Gaussian Processes, and Local Times*, 1st edition. Cambridge University Press, New York.

Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society: Series B*, **56**(4), 653-671.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *The Annals of probability*, **3**(5), 829-839.

Müller, H.-G. (1993). On the boundary kernel method for non-parametric curve estimation near endpoints. *Scandinavian Journal of Statistics*, **20**(4), 313-328.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**(3), 703-708.

Parzen, E. (1962). On estimation of a probability of density function and mode. *Annals of Mathematical Statistics*, **33**(3), 1065-1076.

Phillips, P.C.B. and Park, J. (1998). Nonstationary density and kernel autoregression. *Cowles Foundation Discussion Paper No. 1181*, Department of Economics, Yale University.

Pritsker, M. (1998). Nonparametric density estimation and tests of continuous time interest rate models. *Review of Financial Studies*, **11**(3), 449-487.

Reiss, R.D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavia Journal of Statistics*, **8**(2), 116-119.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**(3), 832-837.

Rudemo, M . (1982). Empirical choice of histograms and kernel density estimators. *Scandinavia Journal of Statistics*, **9**(2), 65-78 .

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and methods*, **14**(5), 1123-1136.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics.* John Wiley & Sons, New York.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, **88**(422), 486-494.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B*, **5**(3), 683-690.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**(4), 1285-1297.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**(398), 559-567.

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing.* Chapman and Hall, London.

Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation (with discussion). *Journal of the American Statistical Association*, **86**(414), 343-353.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**(4), 817-838.

Yoshihara, K.-i. (1995). The bahadour representation of sample quantiles for sequences of strongly mixing random variables. *Statistics & Probability Letters*, **24**(4), 299-304.

Zhang, S., and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *Journal of Statistical Planning and Inference*, **70**(2), 301-316.

Chapter 2

Nonparametric Regression Models

2.1 Prediction and Regression Functions

Suppose that we have the information set I_t at time t and we want to forecast the future value, say Y_{t+1} (one step-ahead forecast, or Y_{t+s}, s-step ahead). There are several forecasting criteria available in the literature. The general form is

$$m\left(I_t\right)=\min_a E\left[\rho\left(Y_{t+1}-a\right) \mid I_t\right],$$

where $\rho(\cdot)$ is an objective (loss) function. Here are three major directions.

(1) If $\rho(z)=z^2$ is the quadratic function, then, $m\left(I_t\right)=E\left(Y_{t+1} \mid I_t\right)$, called the mean regression function. Implicitly, it requires that the distribution of Y_t should be symmetric. If the distribution of Y_t is skewed, then this is not a good criterion.

(2) If $\rho_\tau(y)=y\left(\tau-I_{\{y<0\}}\right)$ called the *check function*, where $\tau \in(0,1)$ and I_A is the indicator function of any set A, then, $m\left(I_t\right)$ satisfies

$$\int_{-\infty}^{m(I_t)} f\left(y \mid I_t\right) du=F\left(m\left(I_t\right) \mid I_t\right)=\tau,$$

where $f(y \mid I_t)$ and $F(m(I_t) \mid I_t)$ are the conditional PDF and CDF of Y_{t+1} given I_t, respectively. This $m\left(I_t\right)$ becomes the conditional quantile or quantile regression, dented by $q_\tau\left(I_t\right)$, proposed by Koenker and Bassett (1978,1982). Particularly, if $\tau=1/2$, then, $m\left(I_t\right)$ is the well known least absolute deviation (LAD) regression which is robust. If $q_\tau\left(I_t\right)$ is a linear function of regressors like $\boldsymbol{\beta}_\tau^T \mathbf{X}_t$ as in Koenker and Bassett (1978,1982), Koenker (2005) developed the R module **quantreg** to make statistical inferences on the linear quantile regression model.

To fit a linear quantile regression using R, one can use the command **rq()** in the package **quantreg**. For a nonlinear parametric model, the command is **nlrq()**. For a nonparametric quantile model for univariate case, one can use the command **lprq()** for implementing the local polynomial estimation. For an additive quantile regression, one can use the commands **rqss()** and **qss()**.

(3) If $\rho(x) = \frac{1}{2}x^2 I_{|x|\le M} + M(|x| - M/2) I_{|x|>M}$, the so called Huber function in literature, then it is the Huber robust regression. We will not discuss this topic. If you have an interest, please read the paper by Huber (1964) and the book by Rousseeuw and Leroy (1987). In R, the library **MASS** has the function **rlm()** for robust linear model. Also, the library **lqs** contains functions for bounded-influence regression.

To see differences among above three cases for the loss function $\rho(\cdot)$, please look at the plot of loss functions given in Figure 2.1.

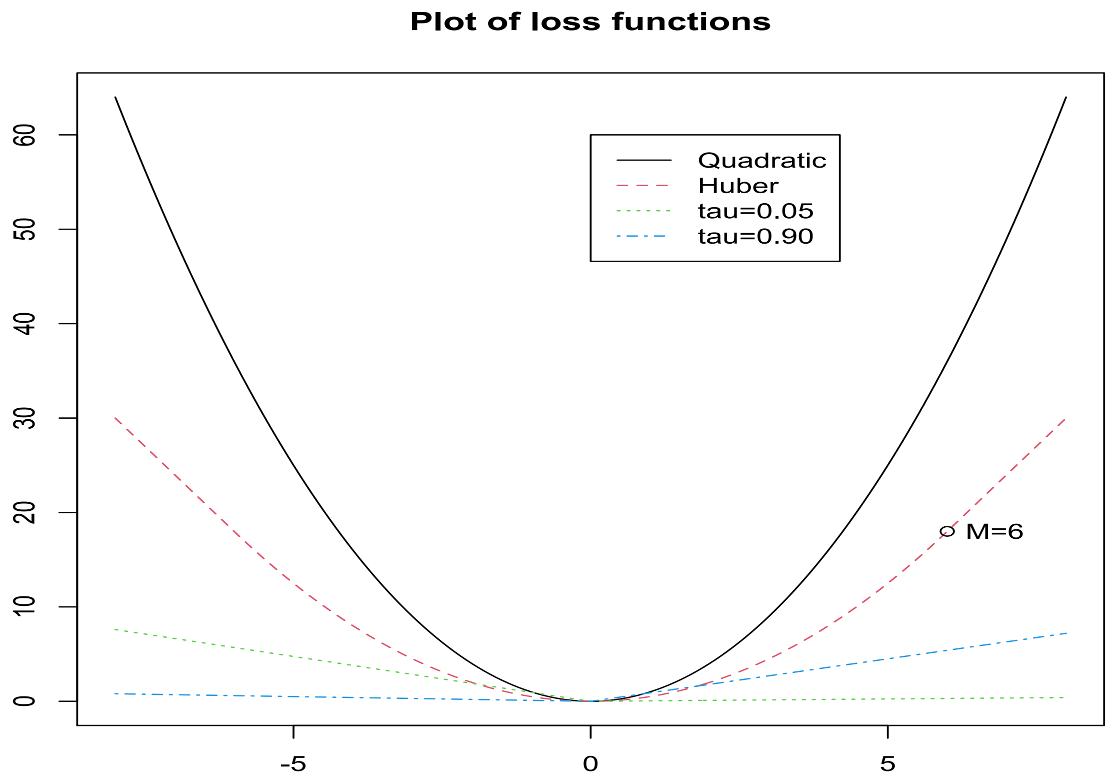


Figure 2.1: The plot of three loss functions: quadratic loss (black solid line), Huber loss (red dashed line) with $M = 6$, the check function ($\tau = 0.05$, green dotted line), and the check function ($\tau = 0.90$, blue dashed-dotted line).

Remark 2.1: *Note that for the second and third cases, the regression functions usually do not have a close form of expression. Since the information set I_t contains too many variables (high dimension), it is often to approximate I_t by some finite numbers of variables, say $X_t = (X_{t1}, \ldots, X_{tp})^T$ $(p \geq 1)$, including the lagged variables and exogenous variables. First, our focus is on the mean regression $m(X_t)$. Of course, by the same token, we can consider the nonparametric estimation of the conditional variance $\sigma^2(x) = Var(Y_t \mid X_t = x)$. Why do we need to consider nonlinear (nonparametric) models in economic practice? To find the answer, please read the paper by Engle, Granger, Rice and Weiss (1986) and some examples in economics and finance in the book by Granger and Teräsvirta (1993).*

Remark 2.2: *Note that throughout this chapter, it is assumed that all regressors are continuous. For the case with discrete and/or partially discrete regressors, the reader is referred to the papers by Li and Racine (2003), Hall, Racine and Li (2004) or the book by Li and Racine (2007) for details.*

2.2 Kernel Estimation

How to estimate $m(x)$ nonparametrically? Let us look at the Nadaraya-Watson estimate of the mean regression $m(x)$. The main idea is as follows:

$$m(x) = \int y f(y \mid x) dy = \frac{\int y f(x, y) dy}{\int f(x, y) dy},$$

where $f(x, y)$ is the joint PDF of X_t and Y_t. To estimate $m(x)$, we can apply the plug-in method. That is, plug the nonparametric kernel density estimate $f_n(x, y)$ (product kernel method) into the right hand side of the above equation to obtain

$$\widehat{m}_{nw}(x) = \frac{\int y f_n(x, y) dy}{\int f_n(x, y) dy} = \cdots = \frac{1}{n} \sum_{t=1}^{n} Y_t K_h(X_t - x) / f_n(x) = \sum_{t=1}^{n} W_t Y_t,$$

where $f_n(x)$ is the kernel density estimation of $f(x)$, defined in Chapter 1 , and

$$W_t = K_h(X_t - x) / \sum_{t=1}^{n} K_h(X_t - x).$$

$\widehat{m}_{nw}(x)$ is the well known Nadaraya-Watson (NW) estimator, proposed by Nadaraya (1964) and Watson (1964). Note that the weights $\{W_t\}$ do not depend on $\{Y_t\}$. Therefore, $\widehat{m}_{nw}(x)$ is called a linear estimator, similar to the least squares estimate (LSE).

Let us look at the NW estimator from a different angle. $\widehat{m}_{nw}(x)$ can be re-expressed as the minimizer of the locally weighted least squares; that is,

$$\widehat{m}_{nw}(x)=\operatorname{argmin}_a \sum_{t=1}^n\left(Y_t-a\right)^2 K_h\left(X_t-x\right).$$

This means that when X_t is in a neighborhood of $x, m\left(X_t\right)$ is approximated by a constant a (local approximation). Indeed, we consider the following working model

$$Y_t=m\left(X_t\right)+\varepsilon_t \approx a+\varepsilon_t$$

with the weights $\left\{K_h\left(X_t-x\right)\right\}$, where $\varepsilon_t=Y_t-E\left(Y_t \mid X_t\right)$. Therefore, the Nadaraya-Watson estimator is also called the local constant estimator.

In the implementation, for each grid point x, we can fit the following transformed linear model

$$Y_t^*=\beta_1 X_t^*+\varepsilon_t,$$

where $Y_t^*=\sqrt{K_h\left(X_t-x\right)} Y_t$ and $X_t^*=\sqrt{K_h\left(X_t-x\right)}$. In R, we can use functions **lm()** or **glm()** with weights $\left\{K_h\left(X_t-x\right)\right\}$ to fit a weighted least squares or generalized linear model. Or, you can use the weighted least squares theory (matrix multiplication); see Section 2.9.

2.2.1 Asymptotic Properties

We derive the asymptotic properties of the nonparametric estimator for the time series situations. Note that the mathematical derivations are different for the iid case and time series situations since the key equality $E\left[Y_t \mid X_1, \cdots, X_n\right]=E\left[Y_t \mid X_t\right]=m\left(X_t\right)$ holds only for the iid case. To ease notation, we consider only the simple case when $p=1$. A simple algebra leads to

$$\widehat{m}_{nw}(x) f_n(x)=\underbrace{\frac{1}{n} \sum_{t=1}^n m\left(X_t\right) K_h\left(X_t-x\right)}_{I_1}+\underbrace{\frac{1}{n} \sum_{t=1}^n K_h\left(X_t-x\right) \varepsilon_t}_{I_2},$$

where $f_n(x)=\sum_{t=1}^n K_h\left(X_t-x\right)/n$. We will show that I_1 contributes only the asymptotic bias and I_2 gives the asymptotic normality. First, we derive the asymptotic bias for the interior boundary points. By the Taylor's expansion, when X_t is in $(x-h, x+h)$, we have

$$m\left(X_t\right)=m(x)+m^{\prime}(x)\left(X_t-x\right)+\frac{1}{2} m^{\prime\prime}\left(x_t\right)\left(X_t-x\right)^2,$$

where $x_t = x + \theta\,(X_t - x)$ with $-1 < \theta < 1$. Then,

$$I_{11} \equiv \frac{1}{n}\sum_{t=1}^{n} m\,(X_t)\,K_h\,(X_t - x) = m(x)f_n(x) + m'(x)\underbrace{\frac{1}{n}\sum_{t=1}^{n}(X_t - x)\,K_h\,(X_t - x)}_{J_1(x)}$$

$$+\frac{1}{2}\underbrace{\frac{1}{n}\sum_{t=1}^{n} m''\,(x_t)\,(X_t - x)^2\,K_h\,(X_t - x)}_{J_2(x)}.$$

Then,

$$E\,[J_1(x)] = E\,[(X_t - x)\,K_h\,(X_t - x)] = \int (u - x)K_h(u - x)f(u)du$$
$$= h\int uK(u)f(x + hu)du = h^2 f'(x)\mu_2(K) + o\left(h^2\right).$$

Similar to the derivation of the variance of $f_n(x)$ in (1.3), we can show that under some conditions,

$$nh\mathrm{Var}\,(J_1(x)) = O(1).$$

Therefore, $J_1(x) = h^2 f'(x)\mu_2(K) + o_p\,(h^2)$. By the same token, we have

$$E\,[J_2(x)] = E\left[m''\,(x_t)\,(X_t - x)^2\,K_h\,(X_t - x)\right]$$
$$= h^2\int m''(x + \theta hu)u^2 K(u)f(x + hu)du = h^2 m''(x)\mu_2(K)f(x) + o\left(h^2\right),$$

and $\mathrm{Var}\,(J_2(x)) = O(1/nh)$. Therefore, $J_2(x) = h^2 m''(x)\mu_2(K)f(x) + o_p\,(h^2)$. Hence,

$$I_1 = m(x)f(x) + m'(x)J_1(x) + \frac{1}{2}J_2(x)$$
$$= m(x)f(x) + \frac{h^2}{2}\mu_2(K)\,[m''(x) + 2m'(x)f'(x)/f(x)]\,f(x) + o_p\left(h^2\right)$$

by the fact that $f_n(x) = f(x) + o_p(1)$. The term $I_1 \approx f(x)\,[m(x) + B_{nw}(x)]$, where

$$B_{nw}(x) = \frac{h^2}{2}\mu_2(K)\left[m''(x) + \boxed{2m'(x)f'(x)/f(x)}\right] \tag{2.1}$$

is regarded as the asymptotic bias. The bias term involves not only curvatures of $m(x)$ ($m''(x)$) but also the unknown density function $f(x)$ and its derivative $f'(x)$ so that the design can not be adaptive.

Under some regularity conditions, similar to (1.3), we can show that for the given grid point x, an interior grid point,

$$nh\mathrm{Var}(I_2) \;\;\to\;\; \nu_0(K)\sigma_\varepsilon^2(x)f(x) \equiv \sigma_m^2(x)f^2(x),$$

where $\sigma_\varepsilon^2(x) = \operatorname{Var}(\varepsilon_t \mid X_t = x)$ and $\sigma_m^2(x) = \nu_0(K)\sigma_\varepsilon^2(x)/f(x)$. Further, by the fact that $f_n(x) = f(x) + o_p(1)$ and the Slutsky theorem, we can establish the asymptotic normality (the proof is provided later)

$$\sqrt{nh}\left[\widehat{m}_{nw}(x) - m(x) - B_{nw}(x) + o_p\left(h^2\right)\right] \quad \rightarrow \quad N\left\{0, \sigma_m^2(x)\right\},$$

where $B_{nw}(x)$ is given in (2.1).

2.2.2 Boundary Behavior

For expositional purpose, in what follows, we only consider the case when $p = 1$. As for the boundary behavior for the NW estimator, we can follow Fan and Gijbels (1996). Without loss of generality, we consider the left boundary point $x = ch, 0 < c < 1$. From Fan and Gijbels (1996), we take $K(\cdot)$ to have support $[-1, 1]$ and $m(\cdot)$ to have support $[0, 1]$. Similar to (1.7), it is easy to see that if $x = ch$,

$$\begin{aligned} E\left[J_1(ch)\right] = E\left[\left(X_t - ch\right) K_h\left(X_t - ch\right)\right] &= \int_0^1 (u - ch) K_h(u - ch) f(u) du \\ &= h \int_{-c}^{1/h - c} u K(u) f(h(u + c)) du \\ &= h f(0+) \mu_{1,c}(K) + h^2 f'(0+)\left[\mu_{2,c}(K) + c\mu_{1,c}(K)\right] + o\left(h^2\right), \end{aligned}$$

and

$$\begin{aligned} E\left[J_2(ch)\right] &= E\left[m''\left(x_t\right)\left(X_t - ch\right)^2 K_h\left(X_t - ch\right)\right] \\ &= h^2 \int_{-c}^{1/h - c} m''(h(c + \theta u)) u^2 K(u) f(h(u + c)) du \\ &= h^2 m''(0+) \mu_{2,c}(K) f(0+) + o\left(h^2\right). \end{aligned}$$

Also, we can see that

$$\operatorname{Var}\left(J_1(ch)\right) = O(1/nh) \quad \text{and} \quad \operatorname{Var}\left(J_2(ch)\right) = O(1/nh),$$

which imply that

$$J_1(ch) = h f(0+) \mu_{1,c}(K) + o_p(h) \quad \text{and} \quad J_2(ch) = h^2 m''(0+) \mu_{2,c}(K) f(0+) + o\left(h^2\right).$$

This, in conjunction with (1.7), gives

$$I_1 - m(ch) = m'(ch) J_1(ch)/f_n(ch) + \frac{1}{2} J_2(ch)/f_n(ch) = a(c, K) h + b(c, K) h^2 + o_p\left(h^2\right),$$

where

$$a(c,K)=\frac{m'(0+)\mu_{1,c}(K)}{\mu_{0,c}(K)},$$

and

$$b(c,K)=\frac{\mu_{2,c}(K)m''(0+)}{2\mu_{0,c}(K)}+\frac{f'(0+)m'(0+)\left[\mu_{2,c}(K)\mu_{0,c}(K)-\mu_{1,c}^2(K)\right]}{f(0+)\mu_{0,c}^2(K)}.$$

Here, $a(c,K)h+b(c,K)h^2$ serves as the asymptotic bias term, which has the order $O(h)$. Also, we can show that at the boundary point, the asymptotic variance has the following form

$$nh\text{Var}\left(\widehat{m}_{nw}(x)\right) \quad\to\quad \nu_{0,c}(K)\sigma_m^2(0+)/\left[\mu_{0,c}(K)f(0+)\right],$$

which has the same order as that for the interior point although the scaling constant is different.

2.3 Local Polynomial Estimate

To overcome the above shortcomings of local constant estimate, we can use the local polynomial fitting scheme; see Fan and Gijbels (1996) for details. The main idea is described as follows.

2.3.1 Formulation

Assume that the regression function $m(x)$ has $(q+1)$ th order continuous derivative. For ease notation, assume that $p=1$. When $X_t\in(x-h,x+h)$, then,

$$m\left(X_t\right)\approx\sum_{j=0}^{q}\frac{m^{(j)}(x)}{j!}\left(X_t-x\right)^j=\sum_{j=0}^{q}\beta_j\left(X_t-x\right)^j,$$

where $\beta_j=m^{(j)}(x)/j\,!$. Therefore, when $X_t\in(x-h,x+h)$, the model becomes

$$Y_t\approx\sum_{j=0}^{q}\beta_j\left(X_t-x\right)^j+\varepsilon_t.$$

Hence, we can apply the weighted least squares method. The locally weighted least squares becomes

$$\sum_{t=1}^{n}\left(Y_t-\sum_{j=0}^{q}\beta_j\left(X_t-x\right)^j\right)^2K_h\left(X_t-x\right). \tag{2.2}$$

Minimizing the above with respect to $\beta = (\beta_0, \ldots, \beta_q)^T$ to obtain the local polynomial estimate $\widehat{\beta}$;

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} Y, \tag{2.3}$$

where $\mathbf{W} = \text{diag}\left\{K_h\left(X_1 - x\right), \cdots, K_h\left(X_n - x\right)\right\}$,

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^q \\ 1 & (X_2 - x) & \cdots & (X_2 - x)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^q \end{pmatrix}, \quad \text{and} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Therefore, for $1 \le j \le q$,

$$\widehat{m}^{(j)}(x) = j!\widehat{\beta}_j.$$

This means that the local polynomial method estimates not only the regression function itself but also derivatives of regression.

2.3.2 Implementation in R and A Real Example

There are several ways of implementing the local polynomial estimator. One way you can do so is to write your own code by using matrix multiplication as in 2.3 or employing function **lm()** or **glm()** with weights $\{K_h\left(X_t - x\right)\}$. Recently, in **R**, there are some build-in packages for implementing the local polynomial estimate. For example, the package **KernSmooth** contains several functions. Function **bkde()** computes the kernel density estimate and Function **bkde2D()** computes the 2D kernel density estimate as well as Function **bkfe()** computes the kernel functional (derivative) density estimate. Function **dpik()** selects a bandwidth for estimating the kernel density estimation using the plug-in method and Function **dpill()** chooses a bandwidth for the local linear ($q = 1$) regression estimation using the plug-in approach. Finally, Function **locpoly()** is for the local polynomial fitting including a local polynomial estimate of the density of X (or its derivative) if the dependent variable is omitted.

Example 2.1: We apply the kernel regression estimation and local polynomial fitting methods to estimate the drift and diffusion of the weekly 3-month Treasury bill from January 2, 1970 to December 26, 1997[1]. Let x_t denote the weekly 3-month Treasury bill. It is often to

[1] Similar to Example 1.2, the data set can be updated to today and it covers a longer period.

model X_t by assuming that it satisfies the continuous-time stochastic differential equation (Black-Scholes model)

$$dX_t = \mu\left(X_t\right) dt + \sigma\left(X_t\right) dW_t,$$

where W_t is a Wiener process, $\mu\left(X_t\right)$ is called the drift function and $\sigma\left(X_t\right)$ is called the diffusion function. Our interest is to identify $\mu\left(X_t\right)$ and $\sigma\left(X_t\right)$. Assume a time series sequence $\{X_{t\Delta}, 1 \leq t \leq n\}$ is observed at **equally spaced** time points. Using the **infinitesimal generator**, see, for example, Øksendal (1985), the first-order approximations of moments of X_t, a discretized version of the Ito's process, are given by Stanton (1997) (see Fan and Zhang (2003) for the higher orders)

$$y(t) = \Delta X_t = \mu\left(X_t\right) \Delta + \sigma\left(X_t\right) \varepsilon\sqrt{\Delta},$$

where $\Delta X_t = X_{t+\Delta} - X_t, \varepsilon \sim N(0,1)$, and x_t and ε_t are independent. Therefore,

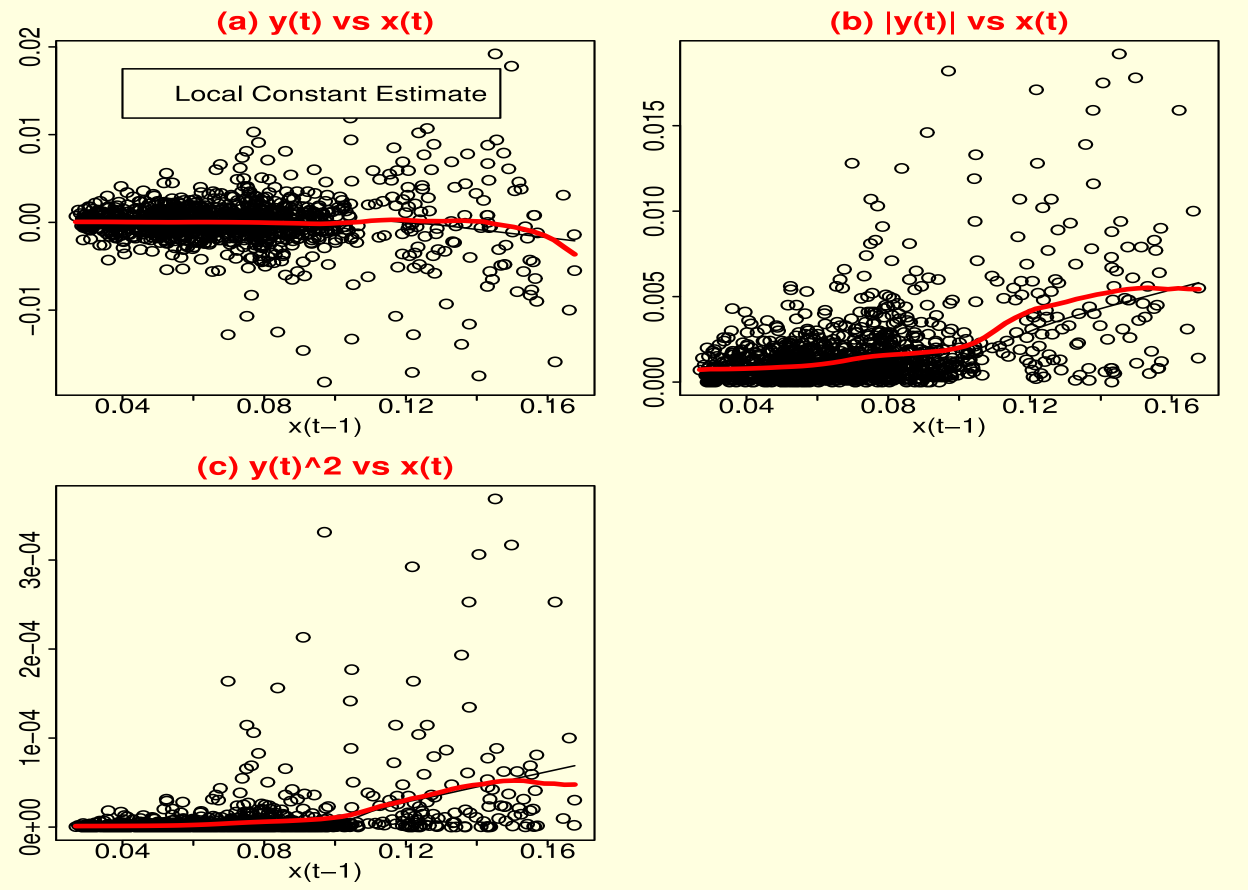


Figure 2.2: Scatterplots of $\Delta X_t, |\Delta X_t|$, and $\left(\Delta X_t\right)^2$ versus $x(t) = X_t$ with the smoothed curves computed using **scatter.smooth()** and the local constant estimation.

$$\mu\left(X_t\right)=\lim_{\Delta\to 0} E\left[\Delta X_t \mid X_t\right]/\Delta \quad \text{and} \quad \sigma^2\left(X_t\right)=\lim_{\Delta\to 0} E\left[\left(\Delta X_t\right)^2 \mid X_t\right]/\Delta.$$

Hence, estimating $\mu(x)$ and $\sigma^2(x)$ becomes a nonparametric regression problem. We can use both local constant and local polynomial method to estimate $\mu(x)$ and $\sigma^2(x)$. As a result, the local constant estimators (red line) together with the **lowess()** smoothers (black line) and the scatterplots of Δx_t in (a), $|\Delta x_t|$ in (b), and $(\Delta x_t)^2$ in (c) versus x_t are presented in Figure 2.2 and the local linear estimators (red line) together with the **lowess()** smoothers (black line) and the scatterplots of ΔX_t in (a), $|\Delta X_t|$ in (b), and $\left(\Delta X_t\right)^2$ in (c) versus X_t are displaced in Figure 2.3. An alternative approach can be found in Aït-Sahalia (1996) to estimate $\mu(x)$ due to the domination of $\sigma\left(X_t\right)\varepsilon\sqrt{\Delta}$ over $\mu\left(X_t\right)\Delta$; see, for example, the paper by Cai and Hong (2009) for details on this regard.

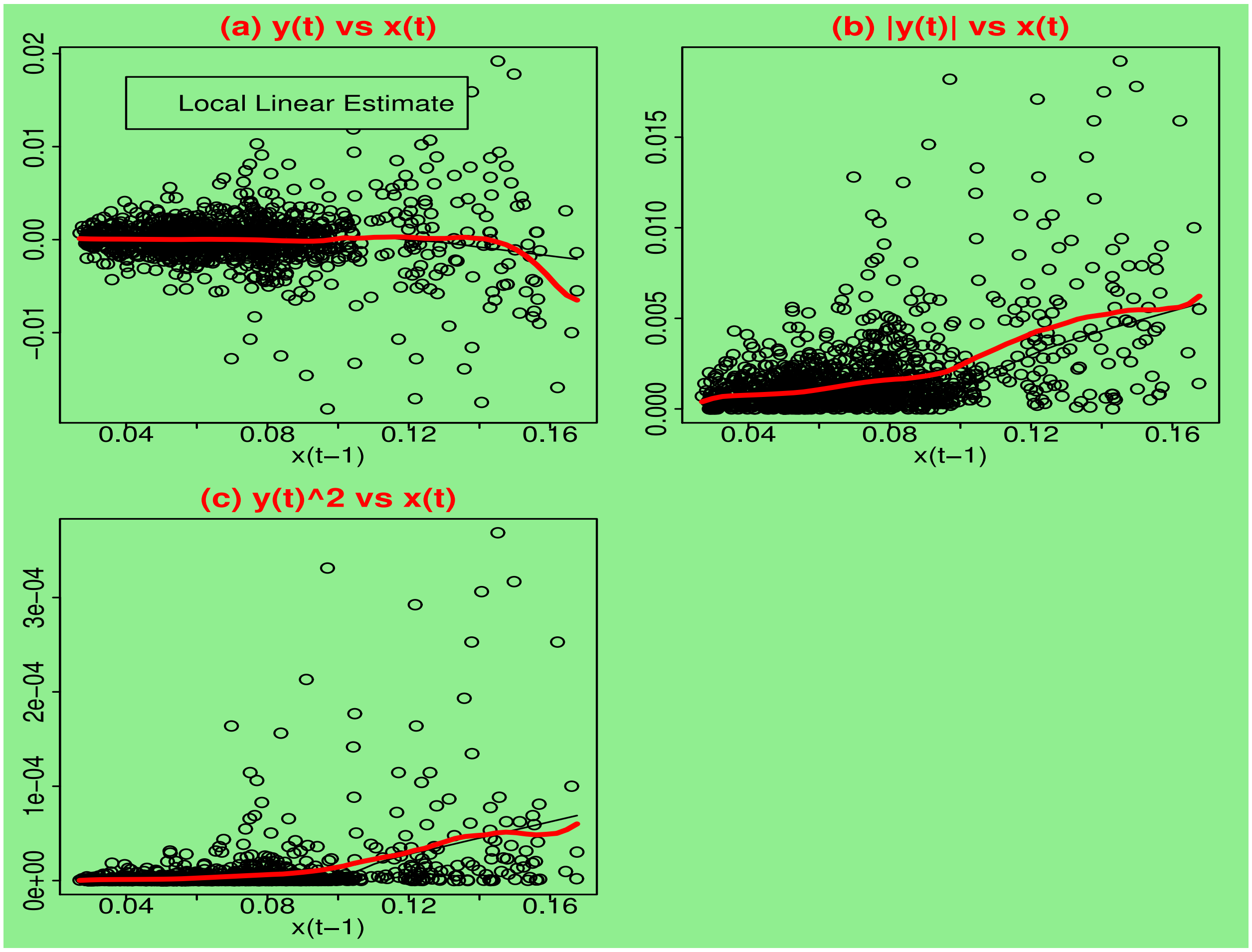


Figure 2.3: Scatterplots of $\Delta X_t, |\Delta X_t|$, and $\left(\Delta X_t\right)^2$ versus $x(t)$ with the smoothed curves computed using **scatter.smooth()** and the local linear estimation.

2.3.3 Complexity of Local Polynomial Estimator

To implement the local polynomial estimator, we have to choose the order of the polynomial q, the bandwidth h and the kernel function $K(\cdot)$. These parameters are of course confounded each other. Clearly, when $h = \infty$, the local polynomial fitting becomes a global polynomial fitting and the order q determines the model complexity. Unlike in the parametric models, the complexity of local polynomial fitting is primarily controlled by the bandwidth, as shown in Fan and Gijbels (1996) and Fan and Yao (2003). Hence q is usually small and the issue of choosing q becomes less critical. We discuss those issues in detail as follows.

(1) If the objective is to estimate $m^{(j)}(\cdot)(j \geq 0)$, the local polynomial fitting corrects automatically the boundary bias when $q - j$ is is odd. Further, when $q - j$ is odd, comparing with the order $q - 1$ fit (so that $q - j - 1$ is even), the order q fit contains one extra parameter without increasing the variance for estimating $m^{(j)}(\cdot)$. But this extra parameter creates opportunities for bias reduction, particularly in the boundary regions; see the next section and the books by Fan and Gijbels (1996) and Ruppert and Wand(1994). For these reasons, the odd order fits (the order q is chosen so that $q - j$ is odd) outperforms the even order fits [the order $(q - 1)$ fit so that q is even]. Based on theoretical and practical considerations, the order $q = j + 1$ is recommended in Fan and Gijbels (1996). If the primary objective is to estimate the regression function, one uses local linear fit and if the target function is the first order derivative, one uses the local quadratic fit and so on.

(2) It is well known that the choice of the bandwidth h plays an important role in any kernel smoothing, including the local polynomial fitting. A too large bandwidth causes over-smoothing (reducing variance), creating excessive modeling bias, while a too small bandwidth results in under-smoothing (reducing bias but increasing variance), obtaining wiggly estimates. The bandwidth can be subjectively chosen by users via visually inspecting resulting estimates, or automatically chosen by data via minimizing an estimated theoretical risk (discussed later). Since the choice of bandwidth is not easy task, it is often attacked by people who do not know well nonparametric techniques.

(3) Since the estimate is based on the local regression (2.2), it is reasonable to require a non-negative weight function $K(\cdot)$. It can be shown (see Fan and Gijbels (1996)) that for all choices of q and j, the optimal weight function is $K(z) = 3/4\,(1 - z^2)_+$, the Epanechnikov kernel, based on minimizing the asymptotic variance of the local polynomial estimator.

Thus, it is a universal weighting scheme and provides a useful benchmark for other kernels to compare with. As shown in Fan and Gijbels (1996) and Fan and Yao (2003), other kernels have nearly the same efficiency for practical use of q and j. Hence, the choice of the kernel function is not critical.

The local polynomial estimator compares favorably with other estimators, including the Nadaraya-Watson (local constant) estimator and other linear estimators such as the Gasser and Müller estimator as in Gasser and Müller (1979) and the Priestley and Chao estimator as in Priestley and Chao (1972). Indeed, it was shown by Fan (1993) that the local linear fitting is asymptotically minimax based on the quadratic loss function among all linear estimators and is nearly minimax among all possible linear estimators. This minimax property is extended by Fan, Gasser, Gijbels, Brockmann and Engel (1995) to more general local polynomial fitting. For the detailed comparisons of the above four estimators, see Fan and Gijbels (1996).

Note that the Gasser and Müller estimator and the Priestley and Chao estimator are particularly for the fixed design. That is, $X_t = t$. Let $s_t = (2t+1)/2 (t = 1, \cdots, n-1)$ with $s_0 = -\infty$ and $s_n = \infty$. The Gasser and Müller estimator is

$$\widehat{m}_{gm}(t_0) = \sum_{t=1}^{n} \int_{s_{t-1}}^{s_t} K_h(u - t_0)\, du Y_t.$$

Unlike the local constant estimator, no denominator is needed since the total weight

$$\sum_{t=1}^{n} \int_{s_{t-1}}^{s_t} K_h(u - t_0)\, du = 1.$$

Indeed, the Gasser and Müller estimator is an improved version of the Priestley and Chao estimator, which is defined as

$$\widehat{m}_{pc}(t_0) = \sum_{t=1}^{n} K_h(t - t_0)\, Y_t.$$

Note that the Priestley and Chao estimator is only applicable for the equi-space setting.

2.3.4 Properties of Local Polynomial Estimator

Define, for $0 \le j \le q$,

$$s_{n,j}(x) = \sum_{t=1}^{n} (X_t - x)^j K_h(X_t - x),$$

and $S_n(x) = \mathbf{X}^T\mathbf{W}\mathbf{X}$. Then, the $(i+1, j+1)$ th element of $S_n(x)$ is $s_{n,i+j}(x)$. Similar to the evaluation of I_{11}, we can show easily that

$$s_{n,j}(x) = nh^j \mu_j(K) f(x) \left\{1 + o_p(1)\right\}.$$

Define, $H = \operatorname{diag}\left\{1, h, \cdots, h^q\right\}$ and $S = \left(\mu_{i+j}(K)\right)_{0 \le i,j \le q}$. Then, it is not difficult to show that $S_n(x) = nf(x)HSH\left\{1 + o_p(1)\right\}$.

First of all, for $0 \le j \le q$, let e_j be a $(q+1) \times 1$ vector with $(j+1)$ th element being one and zero otherwise. Then, $\widehat{\beta}_j$ can be re-expressed as

$$\widehat{\beta}_j = e_j^T \widehat{\beta} = \sum_{t=1}^n W_{j,n,h}\left(X_t - x\right) Y_t,$$

where $W_{j,n,h}\left(X_t - x\right)$ is called the effective kernel in Fan and Gijbels (1996) and Fan and Yao (2003), given by

$$W_{j,n,h}\left(X_t - x\right) = e_j^T S_n(x)^{-1} \left(1, \left(X_t - x\right), \cdots, \left(X_t - x\right)^q\right)^T K_h\left(X_t - x\right).$$

It is not difficult to show (based on the least square theory) that $W_{j,n,h}\left(X_t - x\right)$ satisfies the following the so-called discrete moment conditions

$$\sum_{t=1}^n \left(X_t - x\right)^l W_{j,n,h}\left(X_t - x\right) = \begin{cases} 1 & \text{if } l = j \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Note that the local constant estimator does not have this property; see $J_1(x)$ in Section 2.2.1. This property implies that the local polynomial estimator is unbiased for estimating β_j, when the true regression function $m(x)$ is a polynomial of order q.

To gain more insights about the local polynomial estimator, define the equivalent kernel as in Fan and Gijbels (1996))

$$W_j(u) = e_j^T S^{-1} \left(1, u, \cdots, u^q\right)^T K(u).$$

Then, it can be shown, see, for example, Fan and Gijbels (1996), that

$$W_{j,n,h}\left(X_t - x\right) = \frac{1}{nh^{j+1} f(x)} W_j\left(\left(X_t - x\right)/h\right)\left\{1 + o_p(1)\right\}$$

and

$$\int u^l W_j(u) du = \begin{cases} 1 & \text{if } l = j \\ 0 & \text{otherwise.} \end{cases}$$

The implications of these results are summarized as follows.

As pointed out by Fan and Yao (2003), the local polynomial estimator works like a kernel regression estimation with a known design density $f(x)$. This explains why the local polynomial fit adapts to various design densities. In contrast, the kernel regression estimator has large bias at the region where the derivative of $f(x)$ is large, namely it can not adapt to highly-skewed designs. To see that, imagine the true regression function has large slope in this region. Since the derivative of design density is large, for a given x, there are more points on one side of x than the other. When the local average is taken, the Nadaraya-Watson estimate is biased towards the side with more local data points because the local data are asymmetrically distributed. This issue is more pronounced at the boundary regions, since the local data are even more asymmetric. On the other hand, the local polynomial fit creates asymmetric weights, if needed, to compensate for this kind of design bias. Hence, it is adaptive to various design densities and to the boundary regions.

We next derive the asymptotic bias and variance expression for local polynomial estimators. For independent data, we can obtain the bias and variance expression via conditioning on the design matrix $\mathbf{X}$. However, for time series data, conditioning on $\mathbf{X}$ would mean conditioning on nearly the entire series. Hence, we derive the asymptotic bias and variance using the asymptotic normality rather than conditional expectation. As explained in Chapter 1 localizing in the state domain weakens the dependent structure for the local data. Hence, one would expect that the result for the independent data continues to hold for the stationary process with certain mixing conditions. The mixing condition and the bandwidth should be related, which can be seen later.

Set $B_n(x) = (b_1(x), \cdots, b_n(x))^T$, where, for $0 \le j \le q$,

$$b_{j+1}(x) = \sum_{t=1}^{n} \left[m\left(X_t\right) - \sum_{j=0}^{q} \frac{m^{(j)}(x)}{j!} \left(X_t - x\right)^j \right] \left(X_t - x\right)^j K_h\left(X_t - x\right).$$

Then,

$$\widehat{\beta} - \beta = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} B_n(x) + \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$. It is easy to show that if q is odd,

$$B_n(x) = n h^{q+1} H f(x) \frac{m^{(q+1)}(x)}{(q+1)!} c_{1,q} \left\{1 + o_p(1)\right\},$$

where, for $1 \leq k \leq 3, c_{k,q} = (\mu_{q+k}(K), \cdots, \mu_{2q+k}(K))^T$. If q is even,

$$B_n(x) = nh^{q+2} H f(x) \left[c_{2,q} \frac{m^{(q+1)}(x) f'(x)}{f(x)(q+1)!} + c_{3,q} \frac{m^{(q+2)}(x)}{(q+2)!} \right] \{1 + o_p(1)\}.$$

Note that $f'(x)/f(x)$ does not appear in the right hand side of $B_n(x)$ when q is odd. In either case, we can show that

$$nh \text{Var}[H(\widehat{\beta} - \beta)] \quad \to \quad \sigma^2(x) S^{-1} S^* S^{-1} / f(x) = \Sigma(x),$$

where S^* is a $(q+1) \times (q+1)$ matrix with the (i, j) th element being $\nu_{i+j-2}(K)$.

This shows that the leading conditional bias term depends on whether q is odd or even. By a Taylor series expansion argument, we know that when considering $|X_t - x| < h$, the remainder term from a q th order polynomial expansion should be of order $O\left(h^{q+1}\right)$, so the result for odd q is quite easy to understand. When q is even, $(q+1)$ is odd hence the term h^{q+1} is associated with $\int u^l K(u) du$ for l odd, and this term is zero because $K(u)$ is a even function. Therefore, the h^{q+1} term disappears, while the remainder term becomes $O\left(h^{q+2}\right)$. Since q is either odd or even, then we see that the bias term is an even power of h. This is similar to the case where one uses higher order kernel functions based upon a symmetric kernel function (an even function), where the bias is always an even power of h.

Finally, we can show that when q is odd,

$$\sqrt{nh}[H(\widehat{\beta} - \beta) - B(x)] \quad \to \quad N(0, \Sigma(x)),$$

the asymptotic bias term for the local polynomial estimator is

$$B(x) = \frac{h^{q+1}}{(q+1)!} m^{(q+1)}(x) S^{-1} c_{1,q} \{1 + o_p(1)\}.$$

Or

$$\sqrt{nh^{2j+1}} \left[\widehat{m}^{(j)}(x) - m^{(j)}(x) - B_j(x) \right] \quad \to \quad N\left(0, \sigma_{jj}(x)\right),$$

where the asymptotic bias and variance for the local polynomial estimator of $m^{(j)}(x)$ are

$$B_j(x) = \frac{j! h^{q+1-j}}{(q+1)!} m^{(q+1)}(x) \int u^{q+1} W_j(u) du \{1 + o_p(1)\}$$

and

$$\sigma_{jj}(x) = \frac{(j!)^2 \sigma^2(x)}{f(x)} \int W_j^2(u) du.$$

Similarly, we can derive the asymptotic bias and variance at boundary points if the regression function has a finite support. For details, see the books by Fan and Gijbels (1996), Fan and Yao (2003), and Ruppert and Wand (1994) for details. Indeed, define S_c, S_c^*, and $c_{k,q,c}$ similarly to S, S^* and $c_{k,q}$ with $\mu_j(K)$ and $\nu_j(K)$ replaced by $\mu_{j,c}(K)$ and $\nu_{j,c}(K)$ respectively. We can show that

$$\sqrt{nh}\left[H(\widehat{\beta}(ch)-\beta(ch))-B_c(0)\right] \quad\to\quad N\left(0,\Sigma_c(0)\right), \tag{2.5}$$

where the asymptotic bias term for the local polynomial estimator at the left boundary point is

$$B_c(0)=\frac{h^{q+1}}{(q+1)!}m^{(q+1)}(0)S_c^{-1}c_{1,q,c}\left\{1+o_p(1)\right\},$$

and the asymptotic variance is $\Sigma_c(0)=\sigma^2(0)S_c^{-1}S_c^*S_c^{-1}/f(0)$. Or,

$$\sqrt{nh^{2j+1}}\left[\widehat{m}^{(j)}(ch)-m^{(j)}(ch)-B_{j,c}(0)\right] \quad\to\quad N\left(0,\sigma_{jj,c}(0)\right),$$

where with $W_{j,c}(u)=e_j^TS_c^{-1}\left(1,u,\cdots,u^q\right)^TK(u)$,

$$B_{j,c}(0)=\frac{j!h^{q+1-j}}{(q+1)!}m^{(q+1)}(0)\int_{-c}^{\infty}u^{q+1}W_{j,c}(u)du\left\{1+o_p(1)\right\}$$

and

$$\sigma_{jj,c}(0)=\frac{(j!)^2\sigma^2(0)}{f(0)}\int_{-c}^{\infty}W_{j,c}^2(u)du.$$

Exercise: Please derive the asymptotic properties for the local polynomial estimator. That is to prove (2.5).

The above conclusions show that when $q-j$ is odd, the bias at the boundary is of the same order as that for points on the interior. Hence, the local polynomial fit does not create excessive boundary bias when $q-j$ is odd. Thus, the appealing boundary behavior for local polynomial mean estimation extends to derivative estimation. However, when $q-j$ is even, the bias at the boundary is larger than in the interior, and the bias can also be large at points where $f(x)$ is discontinuous. This is referred to as boundary effect. For these reasons (and the minimax efficiency arguments), it is recommended that one strictly set $q-j$ to be odd when estimating $m^{(j)}(x)$. It is indeed an odd world!

2.3.5 Bandwidth Selection

As seen in previous sections, for stationary sequences of data under certain mixing conditions, the local polynomial estimator performs very much like that for independent data, because windowing reduces dependency among local data. Partially because of this, there are not many studies on bandwidth selection for these problems. However, it is reasonable to expect the bandwidth selectors for independent data continue to work for dependent data with certain mixing conditions. Below, we summarize a few of useful approaches. When data do not have strong enough mixing, the general strategy is to increase bandwidth in order to reduce the variance.

A. Cross-Validation Type Approaches

As what we had already seen for the nonparametric density estimation, the cross-validation method is very useful for assessing the performance of an estimator via estimating its prediction error. The basic idea is to set one of the data point aside for validation of a model and use the remaining data to build the model. It is defined as

$$\mathrm{CV}(h) = \sum_{s=1}^{n} \left[Y_s - \widehat{m}_{-s}\left(X_s\right)\right]^2,$$

where $\widehat{m}_{-s}\left(X_s\right)$ is the local polynomial estimator with $j = 0$ and bandwidth h, but without using the sth observation. The above summand is indeed a squared-prediction error of the sth data point using the training set $\{(X_t, Y_t) : t \neq s\}$. This idea of the cross-validation method is simple but is computationally intensive. An improved version, in terms of computation, is the generalized cross-validation (GCV), proposed by Wahba (1977) and Craven and Wahba (1979). This criterion can be described as follows. The fitted values $\widehat{Y} = \left(\widehat{m}\left(X_1\right), \cdots, \widehat{m}\left(X_n\right)\right)^T$ can be expressed as $\widehat{Y} = H(h)Y$, where $H(h)$ is an $n \times n$ hat matrix, depending on the $\mathbf{X}$-variate and bandwidth h, and it is also called a smoothing matrix. Then the GCV approach selects the bandwidth h that minimizes

$$\mathrm{GCV}(h) = \left[n^{-1}\mathrm{tr}(I - H(h))\right]^{-2} \mathrm{MASE}(h),$$

where $\mathrm{MASE}(h) = \sum_{t=1}^{n}\left(Y_t - \widehat{m}\left(X_t\right)\right)^2/n$ is the average of squared residuals.

A drawback of the cross-validation type method is its inherited variability, see, for example, Hall and Johnstone (1992). Further, it can not be directly applied to select bandwidths for estimating derivative curves. As pointed out by Fan, Heckman and Wand (1995), the cross-validation type method performs poorly due to its large sample variation, even worse for dependent data; see, for example, Shao (1993). Plug-in methods avoid these problems. The basic idea is to find a bandwidth h minimizing estimated mean integrated square error (MISE). See Ruppert, Sheather and Wand (1995) and Fan and Gijbels (1995) for details.

B. Nonparametric AIC Selector

Inspired by the nonparametric version of the Akaike final prediction error criterion proposed by Tjøstheim and Auestad (1994b) for the lag selection in nonparametric setting, Cai and Tiwari (2000) proposed a simple and quick method to select bandwidth for the foregoing estimation procedures, which can be regarded as a nonparametric version of the AIC to be attentive to the structure of time series data and the overfitting or under-fitting tendency. Note that the idea is also motivated by its analogue of Cai and Tiwari (2000). The basic idea is described as follows.

By recalling the classical AIC for linear models under the likelihood setting

$$-2(\text{ maximized log likelihood}) \; + 2 \text{ (number of estimated parameters)},$$

Cai and Tiwari (2000) proposed the following nonparametric AIC to select h minimizing

$$\text{AIC}(h) = \log\{\text{MASE}\} + \psi(\text{tr}(H(h)), n), \tag{2.6}$$

where $\psi(\text{tr}(H(h)), n)$ is chosen particularly to be the form of the bias-corrected version of the AIC, due to Hurvich and Tsai (1989),

$$\psi(\text{tr}(H(h)), n) = 2\{\text{tr}(H(h)) + 1\}/[n - \{\text{tr}(H(h)) + 2\}], \tag{2.7}$$

and $\text{tr}(H(h))$ is the trace of the smoothing matrix $H(h)$, regarded as the nonparametric version of degrees of freedom, called the effective number of parameters, denoted by df. See the book by Hastie and Tibshirani (1990, Section 3.5) for the detailed discussion on this aspect for nonparametric models.[2] Note that actually, (2.6) is a generalization of the AIC for the parametric regression and autoregressive time series contexts, in which $\text{tr}(H(h))$ is the number of regression (autoregressive) parameters in the fitting model. In view of (2.7), when $\psi(\text{tr}(H(h)), n) = -2\log(1 - \text{tr}(H(h))/n)$, then, (2.6) becomes the generalized

[2] Indeed, the df can be fined as either df $= \text{tr}\left(H(h)H(h)^T\right)$ or $\text{tr}\left(2H(h) - H(h)H(h)^T\right)$ or the average of the aforementioned two since $\text{tr}(H(h)) \le \text{tr}\left(H(h)H(h)^T\right) \le \text{tr}\left(2H(h) - H(h)H(h)^T\right)$. See Hastie and Tibshirani (1990, p. 54).

cross-validation (GCV) criterion, commonly used to select the bandwidth in the time series literature even in the iid setting, when $\psi(\text{tr}(H(h)), n) = 2\text{tr}(H(h))/n$, then, (2.6) is the classical AIC discussed in Engle, Granger, Rice, and Weiss (1986) for time series data, and when $\psi(\text{tr}(H(h)), n) = -\log(1 - 2\text{tr}(H(h))/n)$, (2.6) is the T-criterion, proposed and studied by Rice (1984) for iid samples. It is clear that when $\text{tr}(H(h))/n \to 0$, then the nonparametric AIC, the GCV and the T-criterion are asymptotically equivalent. However, the T-criterion requires $\text{tr}(H(h))/n < 1/2$, and, when $\text{tr}(H(h))/n$ is large, the GCV has relatively weak penalty. This is especially true for the nonparametric setting. Therefore, the criterion proposed here counteracts the over-fitting tendency of the GCV. Note that Hurvich, Simonoff, and Tsai (1998) gave the detailed derivation of the nonparametric AIC for the nonparametric regression problems under the iid Gaussian error setting and they argued that the nonparametric AIC performs reasonably well and better than some existing methods in the literature.

2.4 Functional Coefficient Model

2.4.1 Model and Its Properties

As mentioned earlier, when p is large, there exists the so called curse of dimensionality. To overcome this shortcoming, one way to do so is to consider the functional coefficient model as studied in Cai, Fan and Yao (2000) and the additive model discussed in Section 2.5. First, we study the functional coefficient model. To use the notation from Cai, Fan and Yao (2000), we change the notation from the previous sections.

Let $\{\mathbf{U}_i, \mathbf{X}_i, Y_i\}_{i=-\infty}^{\infty}$ be jointly strictly stationary processes with $\mathbf{U}_i$ taking values in $\Re^k$ and $\mathbf{X}_i$ taking values in $\Re^p$. Typically, k is small. Let $E\left(Y_1^2\right) < \infty$. We define the multivariate regression function

$$m(\mathbf{u}, \mathbf{x}) = E(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}), \tag{2.8}$$

where $(\mathbf{U}, \mathbf{X}, Y)$ has the same distribution as $(\mathbf{U}_i, \mathbf{X}_i, Y_i)$. In a pure time series context, both $\mathbf{U}_i$ and $\mathbf{X}_i$ consist of some lagged values of Y_i. The functional-coefficient regression model has the form

$$m(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{p} a_j(\mathbf{u}) x_j, \tag{2.9}$$

where the functions $\{a_j(\cdot)\}$ are measurable from $\Re^k$ to $\Re^1$ and $\mathbf{x} = (x_1, \ldots, x_p)^T$. This model has been studied extensively in the literature; see Cai, Fan and Yao (2000) for the detailed discussions.

For simplicity, in what follows, we consider only the case $k = 1$ in (2.9). Extension to the case $k > 1$ involves no fundamentally new ideas. Note that models with large k are often not practically useful due to the "curse of dimensionality". If k is large, to overcome the problem, one way to do so is to consider an index functional coefficient model proposed by Fan, Yao and Cai (2003)

$$m(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{p} a_j \left(\boldsymbol{\beta}^T \mathbf{u}\right) x_j, \tag{2.10}$$

where $\beta_1 = 1$, and Fan, Yao and Cai (2003) studied the estimation procedures, bandwidth selection and applications. Furthermore, Cai, Juhl and Yang (2015) considered the model in (2.10) on how to select $\boldsymbol{\beta}$ and $\{a_j(\mathbf{u})\}$ by using the least absolute shrinkage and selection operator (LASSO) type method.

As elaborated by Cai et al. (2006) and Cai (2010), functional coefficient models are appropriate and flexible enough for many applications, in particular when additive separability of covariates is unsuitable for the problem at hand. For ease of notation, we assume here that $p = 1$ and $k = 1$. Indeed, by assuming that $m(x, u)$ has a higher order partial derivative with respect to x and applying Taylor expansion to $m(x, u)$, one obtains

$$m(x, u) = \sum_{j=1}^{\infty} \frac{\partial^j m(0, u)}{\partial x^j} \frac{x^j}{j!} \approx \sum_{j=0}^{p} a_j(u) x_j \tag{2.11}$$

for some p (large), where $a_j(u) = (j!)^{-1} \partial^j m(0, u)/\partial x^j$ and $x_j = x^j$. Equation (2.11) implies that a functional coefficient model in (2.9) might be a good approximation to a general nonparametric model in (2.8).

More importantly, as argued in Cai (2010), the functional coefficient model in (2.9) has an ability to capture heteroscedasticity. To get insights about this, it is easy to see that

$$\operatorname{Var}\left(Y_i \mid \mathbf{U}_t\right) = \mathbf{a}\left(\mathbf{U}_i\right)^{\top} \operatorname{Var}\left(\mathbf{X}_i \mid \mathbf{U}_i\right) \mathbf{a}\left(\mathbf{U}_i\right) + \sigma_{\varepsilon}^2\left(\mathbf{U}_i\right),$$

where $\sigma_{\varepsilon}^2\left(\mathbf{U}_i\right) = \operatorname{Var}\left(\varepsilon_i \mid \mathbf{U}_i\right)$. Therefore, the first term in the above expression behaves as an ARCH type model. Furthermore, the functional coefficient approach allows appreciable flexibility on the structure of fitted models without suffering from the *curse of dimensionality* since the nonparametric estimation is conducted in $\Re^k$ instead of $\Re^{p+k}$.

Finally, functional coefficient model can be used as a tool to study covariate adjusted regression for situations where both predictors and response in a regression model are not directly observable, but are contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate (confounding variable); see S entürk and Müller (2005) and Cai and Xu (2008) for more details. For more advantages for the model in (2.9), the reader is referred to the paper by Cai (2010), in particular, about applying functional coefficient model to analyze economic and financial data. Actually, Hong and Lee (2003) considered the applications of model (2.10) to the exchange rates, Juhl (2005) studied the unit root behavior of nonlinear time series models, Li, Huang, Li and Fu (2002) modeled the production frontier using China's manufactural industry data, S entürk and Müller (2006) modeled the nonparametric correlation between two variables using a functional coefficient model as in (2.10), and Cai et al. (2006) considered the nonparametric two-stage instrumental variable estimators for returns to education.

2.4.2 Local Linear Estimation

As recommended by Fan and Gijbels (1996), we estimate the coefficient functions $\{a_j(\cdot)\}$ using the local linear regression method from observations $\{U_i, \mathbf{X}_i, Y_i\}_{i=1}^n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$. We assume throughout that $a_j(\cdot)$ has a continuous second derivative. Note that we may approximate $a_j(\cdot)$ locally at u_0 by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$. The local linear estimator is defined as $\widehat{a}_j\,(u_0) = \widehat{a}_j$, where $\left\{\left(\widehat{a}_j, \widehat{b}_j\right)\right\}$ minimize the sum of weighted squares

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\{a_j + b_j\,(U_i - u_0)\}\,X_{ij}\right]^2 K_h\,(U_i - u_0)\,, \tag{2.12}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h), K(\cdot)$ is a kernel function on $\Re^1$ and $h > 0$ is a bandwidth. It follows from the least squares theory that

$$\widehat{a}_j\,(u_0) = \sum_{k=1}^{n} K_{n,j}\,(U_k - u_0, \mathbf{X}_k)\,Y_k, \tag{2.13}$$

where

$$K_{n,j}(u, \mathbf{x}) = \mathbf{e}_{j,2p}^T\left(\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}}\right)^{-1}\begin{pmatrix}\mathbf{x}\\ u\mathbf{x}\end{pmatrix} K_h(u), \tag{2.14}$$

$\mathbf{e}_{j,2p}$ is the $2p \times 1$ unit vector with 1 at the j th position, $\widetilde{\mathbf{X}}$ denotes an $n \times 2p$ matrix with $\left(\mathbf{X}_i^T, \mathbf{X}_i^T\,(U_i - u_0)\right)$ as its i th row, and $\mathbf{W} = \text{diag}\,\{K_h\,(U_1 - u_0)\,, \ldots, K_h\,(U_n - u_0)\}$.

2.4.3 Bandwidth Selection: Multi-Fold Cross-Validation Criterion

Various existing bandwidth selection techniques for nonparametric regression can be adapted for the foregoing estimation; see, e.g., Fan, Yao, and Cai (2003) and the nonparametric AIC as discussed in Section 2.3.5. Also, Fan and Gijbels (1996) and Ruppert, Sheather, and Wand (1995) developed data-driven bandwidth selection schemes based on asymptotic formulas for the optimal bandwidths, which are less variable and more effective than the conventional data-driven bandwidth selectors such as the cross-validation bandwidth rule. Similar algorithms can be developed for the estimation of functional-coefficient models based on (2.24); however, this will be a future research topic.

Indeed, Cai, Fan and Yao (2000) proposed a simple and quick method for selecting bandwidth h. It can be regarded as a modified multi-fold cross-validation criterion that is attentive to the structure of stationary time series data. Let m and Q be two given positive integers and $n > mQ$. The basic idea is first to use Q sub-series of lengths $n - qm(q = 1, , \cdots, Q)$ to estimate the unknown coefficient functions and then compute the one-step forecasting errors of the next section of the time series of length m based on the estimated models. More precisely, we choose h that minimizes the average mean squared (AMS) error

$$\text{AMS}(h) = \sum_{q=1}^{Q} \text{AMS}_q(h), \tag{2.15}$$

where for $q = 1, \cdots, Q$,

$$\text{AMS}_q(h) = \frac{1}{m} \sum_{i=n-qm+1}^{n-qm+m} \left\{ Y_i - \sum_{j=1}^{p} \widehat{a}_{j,q}\left(U_i\right) X_{i,j} \right\}^2 ,$$

and $\{\widehat{a}_{j,q}(\cdot)\}$ are computed from the sample $\{(U_i, \mathbf{X}_i, Y_i), 1 \le i \le n - qm\}$ with bandwidth equal $h[n/(n-qm)]^{1/5}$. Note that we re-scale bandwidth h for different sample sizes according to its optimal rate, i.e. $h \propto n^{-1/5}$. In practical implementations, we may use $m = [0.1n]$ and $Q = 4$. The selected bandwidth does not depend critically on the choice of m and Q, as long as mQ is reasonably large so that the evaluation of prediction errors is stable. A weighted version of AMS(h) can be used, if one wishes to down-weight the prediction errors at an earlier time. We believe that this bandwidth should be good for modeling and forecasting for time series.

2.4.4 Smoothing Variable Selection

Of importance is to choose an appropriate smoothing variable U in applying functional coefficient regression models if U is a lagged variable. Knowledge on physical background of the data may be very helpful, as Cai, Fan and Yao (2000) discussed in modeling the lynx data. Without any prior information, it is pertinent to choose U in terms of some data-driven methods such as the Akaike information criterion and its variants, cross-validation, and other criteria. Ideally, we would choose U as a linear function of given explanatory variables according to some optimal criterion, which can be fully explored in the work by Fan, Yao and Cai (2003). Nevertheless, we propose here a simple and practical approach: let U be one of the given explanatory variables such that AMS defined in (2.15) obtains its minimum value. Obviously, this idea can be also extended to select p (number of lags) as well.

2.4.5 Goodness-of-Fit Test

To test whether model (2.9) holds with a specified parametric form which is popular in economic and financial applications, such as the threshold autoregressive (TAR) models

$$a_j(u) = \begin{cases} a_{j1}, & \text{if } u \le \eta \\ a_{j2}, & \text{if } u > \eta, \end{cases}$$

or generalized exponential autoregressive (EXPAR) models[3]

$$a_j(u) = \alpha_j + (\beta_j + \gamma_j u)\exp\left(-\theta_j u^2\right),$$

or smooth transition autoregressive (STAR) models

$$\begin{aligned} a_j(u) &= \left[1 - \exp\left(-\theta_j u\right)\right]^{-1} \quad (\text{ logistic}), \\ &\text{or} \\ a_j(u) &= 1 - \exp\left(-\theta_j u^2\right) \quad (\text{exponential}), \\ &\text{or} \\ a_j(u) &= \left[1 - \exp\left(-\theta_j |u|\right)\right]^{-1} \quad (\text{absolute}), \end{aligned}$$

we propose a goodness-of-fit test based on the comparison of the residual sum of squares (RSS) from both parametric and nonparametric fittings. This method is closely related

[3] For more discussions on those models, please see the survey paper by van Dijk, Teräsvirta and Franses (2002).

to the sieve likelihood method proposed by Fan, Zhang and Zhang (2001). Those authors demonstrated the optimality of this kind of procedures for independent samples.

Consider the null hypothesis

$$H_0 : a_j(u) = \alpha_j(u, \boldsymbol{\theta}), \quad 1 \leq j \leq p, \tag{2.16}$$

where $\alpha_j(\cdot, \boldsymbol{\theta})$ is a given family of functions indexed by unknown parameter vector $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$. The RSS under the null hypothesis is

$$\mathrm{RSS}_0 = n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \alpha_1 \left(U_i, \widehat{\boldsymbol{\theta}} \right) X_{i1} - \cdots - \alpha_p \left(U_i, \widehat{\boldsymbol{\theta}} \right) X_{ip} \right\}^2 .$$

Analogously, the RSS corresponding to model (2.9) is

$$\mathrm{RSS}_1 = n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \widehat{a}_1 \left(U_i \right) X_{i1} - \cdots - \widehat{a}_p \left(U_i \right) X_{ip} \right\}^2 .$$

The test statistic is defined as

$$T_n = \left(\mathrm{RSS}_0 - \mathrm{RSS}_1 \right) / \mathrm{RSS}_1 = \mathrm{RSS}_0 / \mathrm{RSS}_1 - 1,$$

and we reject the null hypothesis (2.16) for large value of T_n. Clearly, T_n can be re-expressed as

$$n \left(T_n + 1 \right) \approx n \ln \left(\mathrm{RSS}_0 / \mathrm{RSS}_1 \right) = -2 \log \text{ likelihood ratio}$$

if $\varepsilon_i \sim N\left(0, \sigma^2\right)$. Therefore, T_n is termed as a generalized likelihood ratio (GLR) test in Cai, Fan and Yao (2000) and a generalized F-test in Cai and Tiwari (2000), which can be used to do testing when regressors are even persistent; see the paper by Zhu, Liu, Ling and Cai (2023)

Since there is no asymptotic theory for the proposed test statistic T_n, we suggest using the following nonparametric Bootstrap approach to evaluate the p value of the test:

1. Generate the Bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ from the empirical distribution of the centered residuals $\{\widehat{\varepsilon}_i - \bar{\varepsilon}\}_{i=1}^n$, where

$$\widehat{\varepsilon}_i = Y_i - \widehat{a}_1 \left(U_i \right) X_{i1} - \cdots - \widehat{a}_p \left(U_i \right) X_{ip}, \quad \bar{\widehat{\varepsilon}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i,$$

and define

$$Y_i^* = \alpha_1 \left(U_i, \widehat{\boldsymbol{\theta}} \right) X_{i1} + \cdots + \alpha_p \left(U_i, \widehat{\boldsymbol{\theta}} \right) X_{ip} + \varepsilon_i^*$$

2. Calculate the Bootstrap test statistic T_n^* based on the sample $\{U_i, \mathbf{X}_i, Y_i^*\}_{i=1}^n$.

3. Reject the null hypothesis H_0 when T_n is greater than the upper-α point of the conditional distribution of T_n^* given $\{U_i, \mathbf{X}_i, Y_i\}_{i=1}^n$.

The p-value of the test is simply the relative frequency of the event $\{T_n^* \geq T_n\}$ in the replications of the Bootstrap sampling. For the sake of simplicity, we use the same bandwidth in calculating T_n^* as that in T_n. Note that we Bootstrap the centralized residuals from the nonparametric fit instead of the parametric fit, because the nonparametric estimate of residuals is always consistent, no matter whether the null or the alternative hypothesis is correct. The method should provide a consistent estimator of the null distribution even when the null hypothesis does not hold. Actually, Kreiss, Neumann and Yao (2009) considered nonparametric Bootstrap tests in a general nonparametric regression setting. They proved that, asymptotically, the conditional distribution of the Bootstrap test statistic is indeed the distribution of the test statistic under the null hypothesis. It may be proven that the similar result holds here as long as $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ at the rate $n^{-1/2}$. Note that the above nonparametric Bootstrap does not work when the heterogeneity exists. If so, Cai (2007) suggested using the wild Bootstrap instead of the aforementioned nonparametric Bootstrap, see the paper by Cai (2007) for details.

Finally, note that it is a great challenge to derive the asymptotic property of the testing statistic T_n under time series context and some necessary assumptions. That is to show that

$$b_n\left[T_n - \lambda_n\right] \quad \rightarrow \quad N\left(0, \sigma^2\right)$$

for some normalization constants b_n and λ_n, which is a great project for future research. Note that Fan, Zhang and Zhang (2001) derived the above result for the iid sample.

2.4.6 Asymptotic Results

We first present a result on mean squared convergence that serves as a building block for our main result and is also of independent interest. We now introduce some notation. Let

$$\mathbf{S}_n = \mathbf{S}_n\left(u_0\right) = \begin{pmatrix} \mathbf{S}_{n,0} & \mathbf{S}_{n,1} \\ \mathbf{S}_{n,1} & \mathbf{S}_{n,2} \end{pmatrix}$$

and

$$\mathbf{T}_n = \mathbf{T}_n\left(u_0\right) = \begin{pmatrix} \mathbf{T}_{n,0}\left(u_0\right) \\ \mathbf{T}_{n,1}\left(u_0\right) \end{pmatrix}$$

with

$$\mathbf{S}_{n,j} = \mathbf{S}_{n,j}\left(u_0\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\left(\frac{U_i-u_0}{h}\right)^j K_h\left(U_i-u_0\right)$$

and

$$\mathbf{T}_{n,j}\left(u_0\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\frac{U_i-u_0}{h}\right)^j K_h\left(U_i-u_0\right)Y_i. \tag{2.17}$$

Then, the solution to (2.12) can be expressed as

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{-1}\mathbf{S}_n^{-1}\mathbf{T}_n, \tag{2.18}$$

where $\mathbf{H} = \text{diag}(1,\ldots,1,h,\ldots,h)$ with p-diagonal elements 1's and p diagonal elements h 's. To facilitate the notation, we denote

$$\boldsymbol{\Omega} = \left(\omega_{l,m}\right)_{p\times p} = E\left(\mathbf{X}\mathbf{X}^T \mid U = u_0\right) \tag{2.19}$$

Also, let $f(u,\mathbf{x})$ denote the joint density of $(U,\mathbf{X})$ and $f_u(u)$ be the marginal density of U. We use the following convention: if $U = X_{j_0}$ for some $1 \le j_0 \le p$, then $f(u,\mathbf{x})$ becomes $f(\mathbf{x})$ the joint density of $\mathbf{X}$.

Theorem 2.1: *Let Condition A.1 hold and $f(u,\mathbf{x})$ be continuous at the point u_0. Let $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$. Then it holds that*

$$E\left(\mathbf{S}_{n,j}\left(u_0\right)\right) \quad\to\quad f_u\left(u_0\right)\boldsymbol{\Omega}\left(u_0\right)\mu_j,$$

and

$$nh_n\,Var\Big(\mathbf{S}_{n,j}\left(u_0\right)_{l,m}\Big) \quad\to\quad f_u\left(u_0\right)\nu_{2j}\omega_{l,m}$$

for each $0 \le j \le 3$ and $1 \le l,m \le p$.

As a consequence of Theorem 2.1, we have

$$\mathbf{S}_n \quad\xrightarrow{\mathcal{P}}\quad f_u\left(u_0\right)\mathbf{S}, \qquad \text{and} \qquad \mathbf{S}_{n,3} \quad\xrightarrow{\mathcal{P}}\quad \mu_3 f_u\left(u_0\right)\boldsymbol{\Omega}$$

in the sense that each element converges in probability, where

$$\mathbf{S} = \begin{pmatrix} \boldsymbol{\Omega} & \mu_1\boldsymbol{\Omega} \\ \mu_1\boldsymbol{\Omega} & \mu_2\boldsymbol{\Omega} \end{pmatrix}$$

Put

$$\sigma^2(u,\mathbf{x}) = \text{Var}(Y \mid U = u, \mathbf{X} = \mathbf{x}) \tag{2.20}$$

and

$$\boldsymbol{\Omega}^{*}\left(u_{0}\right)=E\left[\mathbf{X X}^{T} \sigma^{2}(U, \mathbf{X}) \mid U=u_{0}\right]. \tag{2.21}$$

Let $c_0=\mu_2/\left(\mu_2-\mu_1^2\right)$ and $c_1=-\mu_1/\left(\mu_2-\mu_1^2\right)$.

Theorem 2.2: *Let $\sigma^2(u,\mathbf{x})$ and $f(u,\mathbf{x})$ be continuous at the point u_0. Then under Conditions A.1 and A.2,*

$$\sqrt{n h_{n}}\left[\widehat{\mathbf{a}}\left(u_{0}\right)-\mathbf{a}\left(u_{0}\right)-\frac{h^{2}}{2} \frac{\mu_{2}^{2}-\mu_{1} \mu_{3}}{\mu_{2}-\mu_{1}^{2}} \mathbf{a}^{\prime \prime}\left(u_{0}\right)\right] \rightarrow N\left(0, \Theta^{2}\left(u_{0}\right)\right), \tag{2.22}$$

provided that $f_u\left(u_0\right) \neq 0$, where

$$\Theta^{2}\left(u_{0}\right)=\frac{c_{0}^{2} \nu_{0}+2 c_{0} c_{1} \nu_{1}+c_{1}^{2} \nu_{2}}{f_{u}\left(u_{0}\right)} \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \boldsymbol{\Omega}^{*}\left(u_{0}\right) \boldsymbol{\Omega}^{-1}\left(u_{0}\right). \tag{2.23}$$

Theorem 2.2 indicates that the asymptotic bias of $\widehat{a}_j\left(u_0\right)$ is

$$\frac{h^{2}}{2} \frac{\mu_{2}^{2}-\mu_{1} \mu_{3}}{\mu_{2}-\mu_{1}^{2}} a_{j}^{\prime \prime}\left(u_{0}\right)$$

and the asymptotic variance is $\left(n h_n\right)^{-1} \theta_j^2\left(u_0\right)$, where

$$\theta_{j}^{2}\left(u_{0}\right)=\frac{c_{0}^{2} \nu_{0}+2 c_{0} c_{1} \nu_{1}+c_{1}^{2} \nu_{2}}{f_{u}\left(u_{0}\right)} \mathbf{e}_{j, p}^{T} \Omega^{-1}\left(u_{0}\right) \boldsymbol{\Omega}^{*}\left(u_{0}\right) \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \mathbf{e}_{j, p}.$$

When $\mu_1=0$, the bias and variance expressions can be simplified as $h^2\mu_2 a_j^{\prime\prime}\left(u_0\right)/2$ and

$$\theta_{j}^{2}\left(u_{0}\right)=\frac{\nu_{0}}{f_{u}\left(u_{0}\right)} \mathbf{e}_{j, p}^{T} \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \boldsymbol{\Omega}^{*}\left(u_{0}\right) \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \mathbf{e}_{j, p}.$$

The optimal bandwidth for estimating $a_j(\cdot)$ can be defined to be the one that minimizes the squared bias plus variance. The optimal bandwidth is given by

$$h_{j,\, \text{opt}}=\left[\frac{\mu_{2}^{2} \nu_{0}-2 \mu_{1} \mu_{2} \nu_{1}+\mu_{1}^{2} \nu_{2}}{f_{u}\left(u_{0}\right)\left(\mu_{2}^{2}-\mu_{1} \mu_{3}\right)^{2}} \frac{\mathbf{e}_{j, p}^{T} \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \boldsymbol{\Omega}^{*}\left(u_{0}\right) \boldsymbol{\Omega}^{-1}\left(u_{0}\right) \mathbf{e}_{j, p}}{\left\{a_{j}^{\prime \prime}\left(u_{0}\right)\right\}^{2}}\right]^{1 / 5} n^{-1 / 5}. \tag{2.24}$$

2.4.7 Conditions and Proofs

We first impose some conditions on the regression model but they might not be the weakest possible.

Condition A.1

a. The kernel function $K(\cdot)$ is a bounded density with a bounded support $[-1, 1]$.

b. $|f\left(u, v \mid \mathbf{x}_0, \mathbf{x}_1; l\right)| \leq M < \infty$, for all $l \geq 1$, where $f\left(u, v, \mid \mathbf{x}_0, \mathbf{x}_1; l\right)$ is the conditional density of $(U_0, U_l))$ given $(\mathbf{X}_0, \mathbf{X}_l)$, and $f(u \mid \mathbf{x}) \leq M < \infty$, where $f(u \mid \mathbf{x})$ is the conditional density of U given $\mathbf{X} = \mathbf{x}$.

c. The process $\{U_i, \mathbf{X}_i, Y_i\}$ is α-mixing with $\sum k^c[\alpha(k)]^{1-2/\delta} < \infty$ for some $\delta > 2$ and $c > 1 - 2/\delta$.

d. $E|\mathbf{X}|^{2\delta} < \infty$, where δ is given in Condition A.1c.

Condition A.2

a. Assume that

$$E\left\{Y_0^2 + Y_l^2 \mid U_0 = u, \mathbf{X}_0 = \mathbf{x}_0; U_l = v, \mathbf{X}_l = \mathbf{x}_1\right\} \leq M < \infty \qquad (2.25)$$

for all $l \geq 1, \mathbf{x}_0, \mathbf{x}_1 \in \Re^p, u$, and v in a neighborhood of u_0.

b. Assume that $h_n \to$ and $n\, h_n \to \infty$. Further, assume that there exists a sequence of positive integers s_n such that $s_n \to \infty, \quad s_n = o\left(\left(n\, h_n\right)^{1/2}\right)$, and $\left(n/h_n\right)^{1/2} \alpha\left(s_n\right) \to 0$, as $n \to \infty$

c. There exists $\delta^* > \delta$, where δ is given in Condition A.1c, such that

$$E\left\{|Y|^{\delta^*} \mid U = u, \mathbf{X} = \mathbf{x}\right\} \leq M_4 < \infty \qquad (2.26)$$

for all $\mathbf{x} \in \Re^p$ and u in a neighborhood of u_0, and

$$\alpha(n) = O\left(n^{-\theta^*}\right), \qquad (2.27)$$

where $\theta^* \geq \delta\, \delta^* / \left\{2\left(\delta^* - \delta\right)\right\}$

d. $E|\mathbf{X}|^{2\delta^*} < \infty$, and $n^{1/2-\delta/4} h^{\delta/\delta^*-1/2-\delta/4} = O(1)$

Remark 2.3: *We provide a sufficient condition for the mixing coefficient $\alpha(n)$ to satisfy Conditions A.1c and A.2b. Suppose that $h_n = An^{-\rho}(0 < \rho < 1, A > 0)$, $s_n = \left(nh_n/\log n\right)^{1/2}$ and $\alpha(n) = O\left(n^{-d}\right)$ for some $d > 0$. Then Condition A.1c is satisfied for $d > 2(1 -$*

$1/\delta)/(1-2/\delta)$ and Condition A.2b is satisfied if $d > (1+\rho)/(1-\rho)$. Hence both conditions are satisfied if

$$\alpha(n) = O\left(n^{-d}\right), \qquad d > \max\left\{\frac{1+\rho}{1-\rho}, \frac{2(1-1/\delta)}{1-2/\delta}\right\}.$$

Note that this is a trade-off between the order δ of the moment of Y and the rate of decay of the mixing coefficient; the larger the order δ, the weaker the decay rate of $\alpha(n)$.

To study the joint asymptotic normality of $\widehat{\mathbf{a}}\left(u_0\right)$, we need to center the vector $\mathbf{T}_n\left(u_0\right)$ by replacing Y_i with $Y_i - m\left(U_i, \mathbf{X}_i\right)$ in the expression (2.17) of $\mathbf{T}_{n,j}\left(u_0\right)$. Let

$$\mathbf{T}_{n,j}^*\left(u_0\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left(\frac{U_i - u_0}{h}\right)^j K_h\left(U_i - u_0\right)\left[Y_i - m\left(U_i, \mathbf{X}_i\right)\right],$$

and

$$\mathbf{T}_n^* = \begin{pmatrix} \mathbf{T}_{n,0}^* \\ \mathbf{T}_{n,1}^* \end{pmatrix}.$$

Because the coefficient functions $a_j(u)$ are conducted in the neighborhood of $|U_i - u_0| < h$, by Taylor's expansion,

$$m\left(U_i, \mathbf{X}_i\right) = \mathbf{X}_i^T \mathbf{a}\left(u_0\right) + \left(U_i - u_0\right)\mathbf{X}_i^T \mathbf{a}'\left(u_0\right) + \frac{h^2}{2}\left(\frac{U_i - u_0}{h}\right)^2 \mathbf{X}_i^T \mathbf{a}''\left(u_0\right) + o_p\left(h^2\right),$$

where $\mathbf{a}'\left(u_0\right)$ and $\mathbf{a}''\left(u_0\right)$ are the vectors consisting of the first and second derivatives of the functions $a_j(\cdot)$. Then,

$$\mathbf{T}_{n,0} - \mathbf{T}_{n,0}^* = \mathbf{S}_{n,0}\mathbf{a}\left(u_0\right) + h\mathbf{S}_{n,1}\mathbf{a}'\left(u_0\right) + \frac{h^2}{2}\mathbf{S}_{n,2}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right)$$

and

$$\mathbf{T}_{n,1} - \mathbf{T}_{n,1}^* = \mathbf{S}_{n,1}\mathbf{a}\left(u_0\right) + h\mathbf{S}_{n,2}\mathbf{a}'\left(u_0\right) + \frac{h^2}{2}\mathbf{S}_{n,3}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right)$$

so that

$$\mathbf{T}_n - \mathbf{T}_n^* = \mathbf{S}_n\mathbf{H}\boldsymbol{\beta} + \frac{h^2}{2}\begin{pmatrix} \mathbf{S}_{n,2} \\ \mathbf{S}_{n,3} \end{pmatrix}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right) \tag{2.28}$$

where $\boldsymbol{\beta} = \left(\mathbf{a}\left(u_0\right)^T, \mathbf{a}'\left(u_0\right)^T\right)^T$. Thus it follows from (2.18), (2.28), and Theorem 2.1. that

$$\mathbf{H}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = f_u^{-1}\left(u_0\right)\mathbf{S}^{-1}\mathbf{T}_n^* + \frac{h^2}{2}\mathbf{S}^{-1}\begin{pmatrix} \mu_2\boldsymbol{\Omega} \\ \mu_3\boldsymbol{\Omega} \end{pmatrix}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right) \tag{2.29}$$

from which the bias term of $\widehat{\boldsymbol{\beta}}\left(u_0\right)$ is evident. Clearly,

$$\widehat{\mathbf{a}}\left(u_0\right) - \mathbf{a}\left(u_0\right) = \frac{\boldsymbol{\Omega}^{-1}}{f_u\left(u_0\right)\left(\mu_2 - \mu_1^2\right)}\left[\mu_2\mathbf{T}_{n,0}^* - \mu_1\mathbf{T}_{n,1}^*\right] + \frac{h^2}{2}\frac{\mu_2^2 - \mu_1\mu_3}{\mu_2 - \mu_1^2}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right). \tag{2.30}$$

Thus, (2.30) indicates that the asymptotic bias of $\widehat{\mathbf{a}}\left(u_0\right)$ is

$$\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''\left(u_0\right)$$

Let

$$\mathbf{Q}_n=\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i, \tag{2.31}$$

where

$$\mathbf{Z}_i=\mathbf{X}_i\left[c_0+c_1\left(\frac{U_i-u_0}{h}\right)\right]K_h\left(U_i-u_0\right)\left[Y_i-m\left(U_i,\mathbf{X}_i\right)\right] \tag{2.32}$$

with $c_0=\mu_2/\left(\mu_2-\mu_1^2\right)$ and $c_1=-\mu_1/\left(\mu_2-\mu_1^2\right)$. It follows from (2.30) and (2.31) that

$$\sqrt{nh_n}\left[\widehat{\mathbf{a}}\left(u_0\right)-\mathbf{a}\left(u_0\right)-\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''\left(u_0\right)\right]=\frac{\Omega^{-1}}{f_u\left(u_0\right)}\sqrt{nh_n}\mathbf{Q}_n+o_p(1). \tag{2.33}$$

To finish the proof, one needs the following lemma, whose proof is more involved than that for Theorem 2.1. Therefore, we prove only this lemma. Throughout this section, we let C denote a generic constant, which may take different values at different places.

Lemma 2.1: Under Conditions A.1 and A.2 and the assumption that $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$, if $\sigma^2(u,\mathbf{x})$ and $f(u,\mathbf{x})$ are continuous at the point u_0, then we have

(a) $h_n\mathrm{Var}\left(\mathbf{Z}_1\right) \quad \to \quad f_u\left(u_0\right)\Omega^*\left(u_0\right)\left[c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2\right]$

(b) $h_n\sum_{l=1}^{n-1|}\left|\mathrm{Cov}\left(\mathbf{Z}_1,\mathbf{Z}_{l+1}\right)\right|=o(1)$; and

(c) $nh_n\mathrm{Var}\left(\mathbf{Q}_n\right) \quad \to \quad f_u\left(u_0\right)\boldsymbol{\Omega}^*\left(u_0\right)\left[c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2\right]$

Proof: First, by conditioning on $\left(U_1,\mathbf{X}_1\right)$ and using Theorem 1 of Sun (1984), we have

$$\begin{aligned}\mathrm{Var}\left(\mathbf{Z}_1\right)&=E\left[\mathbf{X}_1\mathbf{X}_1^T\sigma^2\left(U_1,\mathbf{X}_1\right)\left\{c_0+c_1\left(\frac{U_1-u_0}{h}\right)\right\}^2K_h^2\left(U_1-u_0\right)\right]\\&=\frac{1}{h}\left[f_u\left(u_0\right)\boldsymbol{\Omega}^*\left(u_0\right)\left\{c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2\right\}+o(1)\right]\end{aligned} \tag{2.34}$$

The result (c) follows in an obvious manner from (a) and (b) along with

$$\mathrm{Var}\left(\mathbf{Q}_n\right)=\frac{1}{n}\mathrm{Var}\left(\mathbf{Z}_1\right)+\frac{2}{n}\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\mathrm{Cov}\left(\mathbf{Z}_1,\mathbf{Z}_{l+1}\right). \tag{2.35}$$

It thus remains to prove part (b). To this end, let $d_n \to \infty$ be a sequence of positive integers such that $d_n\, h_n \to 0$. Define

$$J_1 = \sum_{l=1}^{d_n - 1} |\text{Cov}\,(\mathbf{Z}_1, \mathbf{Z}_{l+1})| \quad \text{and} \quad J_2 = \sum_{l=d_n}^{n-1} |\text{Cov}\,(\mathbf{Z}_1, \mathbf{Z}_{l+1})|$$

It remains to show that $J_1 = o\,(h^{-1})$ and $J_2 = o\,(h^{-1})$.

We remark that because $K(\cdot)$ has a bounded support $[-1, \quad 1]$, $a_j(u)$ is bounded in the neighborhood of $u \in [u_0 - h, \quad u_0 + h]$. Let $B = \max_{1 \le j \le p} \sup_{|u-u_0|<h} |a_j(u)|$ and $g(\mathbf{x}) = \sum_{j=1}^p |x_j|$. Then $\sup_{|u-u_0|<h} |m(u, \mathbf{x})| \le B\, g(\mathbf{x})$. By conditioning on $(U_1, \mathbf{X}_1)$ and $(U_{l+1}, \mathbf{X}_{l+1})$, and using (2.25) and Condition A.1b, we have, for all $l \ge 1$,

$$\begin{aligned}
&|\text{Cov}\,(\mathbf{Z}_1, \mathbf{Z}_{l+1})\,| \\
&\le CE\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^T\right| \{|Y_1| + Bg\,(\mathbf{X}_1)\}\{|Y_{l+1}| + Bg\,(\mathbf{X}_{l+1})\}\, K_h\,(U_1 - u_0)\, K_h\,(U_{l+1} - u_0)\right] \\
&\le CE\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^T\right| \left\{M_2 + B^2g^2\,(\mathbf{X}_1)\right\}^{1/2} \left\{M_2 + B^2g^2\,(\mathbf{X}_{l+1})\right\}^{1/2} K_h\,(U_1 - u_0)\, K_h\,(U_{l+1} - u_0)\right] \\
&\le CE\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^T\right| \{1 + g\,(\mathbf{X}_1)\}\{1 + g\,(\mathbf{X}_{l+1})\}\right] \le C. \qquad (2.36)
\end{aligned}$$

It follows that

$$J_1 \le Cd_n = o\left(h^{-1}\right)$$

by the choice of d_n. We next consider the upper bound of J_2. To this end, using the Davydov's inequality (see Lemma 1.1), we obtain, for all $1 \le j, m \le p$ and $l \ge 1$,

$$|\text{Cov}\,(Z_{1j}, Z_{l+1,m})| \le C[\alpha(l)]^{1-2/\delta} \left[E\,|Z_j|^\delta\right]^{1/\delta} \left[E\,|Z_m|^\delta\right]^{1/\delta}. \qquad (2.37)$$

By conditioning on $(U, \mathbf{X})$ and using Conditions A.1b and A.2c, one has

$$\begin{aligned}
E\left[|Z_j|^\delta\right] &\le CE\left[|X_j|^\delta\, K_h^\delta\,(U - u_0)\left\{|Y|^\delta + B^\delta g^\delta(\mathbf{X})\right\}\right] \\
&\le CE\left[|X_j|^\delta\, K_h^\delta\,(U - u_0)\left\{M_3 + B^\delta g^\delta(\mathbf{X})\right\}\right] \\
&\le Ch^{1-\delta}E\left[|X_j|^\delta\left\{M_3 + B^\delta g^\delta(\mathbf{X})\right\}\right] \le Ch^{1-\delta}. \qquad (2.38)
\end{aligned}$$

A combination of (2.37) and (2.38) leads to

$$J_2 \le C\, h^{2/\delta - 2} \sum_{l=d_n}^{\infty} [\alpha(l)]^{1-2/\delta} \le C\, h^{2/\delta-2} d_n^{-c} \sum_{l=d_n}^{\infty} l^c [\alpha(l)]^{1-2/\delta} = o\left(h^{-1}\right) \qquad (2.39)$$

by choosing d_n such that $h^{1-2/\delta} d_n^c = C$, so the requirement that $d_n\, h_n \to 0$ is satisfied. □

Proof of Theorem 2.2

We use the small-block and large-block technique-namely, partition $\{1, \ldots, n\}$ into $2q_n + 1$ subsets with large block of size $r = r_n$ and small block of size $s = s_n$. Set

$$q = q_n = \left\lfloor \frac{n}{r_n + s_n} \right\rfloor . \tag{2.40}$$

We now use the CramÃ©r-Wold device to derive the asymptotic normality of $\mathbf{Q}_n$. For any unit vector $\mathbf{d} \in \Re^p$, let $Z_{n,i} = \sqrt{h}\mathbf{d}^T \mathbf{Z}_{i+1}, i = 0, \ldots, n-1$. Then,

$$\sqrt{nh}\mathbf{d}^T \mathbf{Q}_n = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} Z_{n,i},$$

and, by Lemma 2.1,

$$\begin{aligned} \operatorname{Var}\left(Z_{n,0}\right) &\approx f_u\left(u_0\right) \mathbf{d}^T \mathbf{\Omega}^*\left(u_0\right) \mathbf{d} \left[c_0^2 \nu_0 + 2 c_0 c_1 \nu_1 + c_1^2 \nu_2\right] \\ &\equiv \theta^2\left(u_0\right) \end{aligned} \tag{2.41}$$

and

$$\sum_{l=0}^{n-1} \left|\operatorname{Cov}\left(Z_{n,0}, Z_{n,l}\right)\right| = o(1). \tag{2.42}$$

Define the random variables, for $0 \leq j \leq q-1$,

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} Z_{n,i}, \quad \xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)} Z_{n,i}, \quad \text{and} \quad \zeta_q = \sum_{i=q(r+s)}^{n-1} Z_{n,i}.$$

Then,

$$\sqrt{n\,h}\mathbf{d}^T \mathbf{Q}_n = \frac{1}{\sqrt{n}} \left\{ \sum_{j=0}^{q-1} \eta_j + \sum_{j=0}^{q-1} \xi_j + \zeta_q \right\} \equiv \frac{1}{\sqrt{n}} \left\{Q_{n,1} + Q_{n,2} + Q_{n,3}\right\}. \tag{2.43}$$

We show that as $n \to \infty$

$$\frac{1}{n} E\left[Q_{n,2}\right]^2 \;\to\; 0, \quad \frac{1}{n} E\left[Q_{n,3}\right]^2 \;\to\; 0, \tag{2.44}$$

$$\left| E\left[\exp\left(itQ_{n,1}\right)\right] - \prod_{j=0}^{q-1} E\left[\exp\left(it\eta_j\right)\right] \right| \;\to\; 0, \tag{2.45}$$

$$\frac{1}{n} \sum_{j=0}^{q-1} E\left(\eta_j^2\right) \;\to\; \theta^2\left(u_0, \right), \tag{2.46}$$

and

$$\frac{1}{n}\sum_{j=0}^{q-1} E\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \quad\to\quad 0 \tag{2.47}$$

for every $\varepsilon > 0$. (2.44) implies that $Q_{n,2}$ and $Q_{n,3}$ are asymptotically negligible in probability, (2.45) shows that the summands η_j in $Q_{n,1}$ are asymptotically independent and (2.46) and (2.47) are the standard Lindeberg-Feller conditions for asymptotic normality of $Q_{n,1}$ for the independent setup.

First, we establish (2.44). For this purpose, we choose the large block size. Condition A.2b implies that there is a sequence of positive constants $\gamma_n \to \infty$ such that $\gamma_n s_n = o\left(\sqrt{nh_n}\right)$ and

$$\gamma_n\left(n/h_n\right)^{1/2}\alpha\left(s_n\right) \quad\to\quad 0. \tag{2.48}$$

Define the large block size r_n by $r_n = \left\lfloor \left(n\,h_n\right)^{1/2}/\gamma_n \right\rfloor$ and the small block size s_n. Then it can easily be shown from (2.48) that as $n \to \infty$,

$$s_n/r_n \quad\to\quad 0, \quad r_n/n \quad\to\quad 0, \quad r_n\left(nh_n\right)^{-1/2} \quad\to\quad 0, \tag{2.49}$$

and

$$\left(n/r_n\right)\alpha\left(s_n\right) \quad\to\quad 0. \tag{2.50}$$

Observe that

$$E\left[Q_{n,2}\right]^2 = \sum_{j=0}^{q-1}\mathrm{Var}\left(\xi_j\right) + 2\sum_{0\leq i<j\leq q-1}\mathrm{Cov}\left(\xi_i, \xi_j\right) \equiv I_1 + I_2. \tag{2.51}$$

It follows from stationarity and Lemma 2.1 that

$$I_1 = q_n\mathrm{Var}\left(\xi_1\right) = q_n\mathrm{Var}\left(\sum_{j=1}^{s_n} Z_{n,j}\right) = q_n s_n\left[\theta^2\left(u_0\right) + o(1)\right]. \tag{2.52}$$

Next, consider the second term I_2 in the right side of (2.51). Let $r_j^* = j\left(r_n + s_n\right)$, then $r_j^* - r_i^* \geq r_n$ for all $j > i$. Thus, we have

$$\begin{aligned}|I_2| &\leq 2\sum_{0\leq i<j\leq q-1}\sum_{j_1=1}^{s_n}\sum_{j_2=1}^{s_n}\left|\mathrm{Cov}\left(Z_{n,r_i^*+r_n+j_1}, Z_{n,r_j^*+r_n+j_2}\right)\right| \\ &\leq 2\sum_{j_1=1}^{n-r_n}\sum_{j_2=j_1+r_n}^{n}\left|\mathrm{Cov}\left(Z_{n,j_1}, Z_{n,j_2}\right)\right|.\end{aligned}$$

By stationarity and Lemma 2.1, one obtains

$$|I_2| \leq 2n \sum_{j=r_n+1}^{n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| = o(n). \tag{2.53}$$

Hence, by (2.49) - (2.53), we have

$$\frac{1}{n} E\left[Q_{n,2}\right]^2 = O\left(q_n s_n n^{-1}\right) + o(1) = o(1). \tag{2.54}$$

It follows from stationarity, (2.49), and Lemma 2.1 that

$$\mathrm{Var}\left[Q_{n,3}\right] = \mathrm{Var}\left(\sum_{j=1}^{n-q_n(r_n+s_n)} Z_{n,j}\right) = O\left(n - q_n\left(r_n + s_n\right)\right) = o(n). \tag{2.55}$$

Combining (2.49), (2.54), and (2.55), we establish (2.44). As for (2.46) by stationarity, (2.49), (2.50), and Lemma 2.1, it is easily seen that

$$\frac{1}{n} \sum_{j=0}^{q_n-1} E\left(\eta_j^2\right) = \frac{q_n}{n} E\left(\eta_1^2\right) = \frac{q_n r_n}{n} \cdot \frac{1}{r_n} \mathrm{Var}\left(\sum_{j=1}^{r_n} Z_{n,j}\right) \quad \rightarrow \quad \theta^2\left(u_0\right).$$

To establish (2.45), we use Lemma 1.1 of Volkonskii and Rozanov (1959) (see also Ibragimov and Linnik 1971, p. 338) to obtain

$$\left| E\left[\exp\left(itQ_{n,1}\right)\right] - \prod_{j=0}^{q_n-1} E\left[\exp\left(it\eta_j\right)\right] \right| \leq 16\left(n/r_n\right) \alpha\left(s_n\right)$$

tending to 0 by (2.50).

It remains to establish (2.47). For this purpose, we use Theorem 4.1 in Shao and Yu (1996) and Condition A.2 to obtain

$$E\left[\eta_1^2 I\left\{|\eta_1| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq C n^{1-\delta/2} E\left(|\eta_1|^{\delta}\right) \leq C n^{1-\delta/2} r_n^{\delta/2}\left\{E\left(|Z_{n,0}|^{\delta^*}\right)\right\}^{\delta/\delta^*} \tag{2.56}$$

As in (2.38),

$$E\left(|Z_{n,0}|^{\delta^*}\right) \leq C h^{1-\delta^*/2}. \tag{2.57}$$

Therefore, by (2.56) and (2.57),

$$E\left[\eta_1^2 I\left\{|\eta_1| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq C n^{1-\delta/2} r_n^{\delta/2} h^{(2-\delta^*)\delta/(2\delta^*)}. \tag{2.58}$$

Thus, by (2.40) and the definition of r_n, and using Conditions A.2c and A.2d, we obtain

$$\frac{1}{n} \sum_{j=0}^{q-1} E\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq C\gamma_n^{1-\delta/2} n^{1/2-\delta/4} h_n^{\delta/\delta^*-1/2-\delta/4} \quad \rightarrow \quad 0, \tag{2.59}$$

because $\gamma_n \rightarrow \infty$. This completes the proof of the theorem.

2.4.8 Applications

1. Applications to Time Series

See Cai, Fan and Yao (2000) for the detailed Monte Carlo simulation results and applications for time series data.

2. Analysis Of Boston Housing Data

A. Description of Data

The well known **Boston house price data** set[4] consists of 14 variables, collected on each of 506 different houses from a variety of locations. The Boston house-price data set was used originally by Harrison and Rubinfeld (1978) and it was re-analyzed in Belsley, Kuh and Welsch (1980) by various transformations in the table on pages 244-261. Variables are, denoted by $X_1, \cdots, X_{13}$ and Y, in order: The dependent variable is Y, the median value of owner-occupied homes in $\$1,000$'s (house price). The major factors possibly affecting the house prices used in the literature are: X_{13} = proportion of population of lower educational

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10,000USD
PTRATIO	pupil-teacher ratio by town
B	1000(Bk − 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	lower status of the population
MEDV	Median value of owner-occupied homes in $1000â s

[4] This dataset can be downloaded from the web site at http://lib.stat.cmu.edu/datasets/boston.

status X_6 = the average number of rooms per house, X_1 = the per capita crime rate, X_{10} = the full property tax rate, and X_{11} = the pupil/teacher ratio. For the complete description of all 14 variables, see Harrison and Rubinfeld (1978) and Gilley and Pace (1996) for corrections.

B. Linear Models

Harrison and Rubinfeld (1978) was the first to analyze this data set using a standard regression model Y versus all 13 variables including some higher order terms or transformations on Y and X_j 's. The purpose of this study is to see whether there are the effects of pollution on housing prices via hedonic pricing methodology. Belsley, Kuh and Welsch (1980) used this data set to illustrate the effects of using robust regression and outlier detection strategies. From these results, we might conclude that the model might not be linear and there might exist outliers. Also, Pace and Gilley (1997) added a geo-referencing idea (spatial statistics) and used a spatial estimation method to consider this data set.

Exercise: Please use all possible methods to explore this dataset to see what is the best linear model you can obtain.

C. Fit a Varying-Coefficient Model

S entürk and Müller (2006) studied the correlation between the house price Y and the crime rate X_1 adjusted by the confounding variable X_{13} through a varying coefficient model and they concluded that the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. Finally, it is surprising that all the existing nonparametric models aforementioned above did not include the crime rate X_1, which may be an important factor affecting the housing price, and did not consider the interaction terms such as X_{13} and X_1. See the paper by Fan and Huang (2005) for fitting a varying coefficient partially linear model to the Boston housing data, which will be discussed in detail in Section **??**.

Exercise: Please fit a a varying coefficient model to the Boston housing data.

3. Functional Coefficient Capital Asset Pricing Model

The model in (2.8) was successfully applied by Cai, Ren and Yang (2015) to study the conditional capital asset pricing model (CAPM) to argue that the β in the conventional CAPM changes over time.

$$r_t = \beta(\mathbf{Z}_t)^\top \mathbf{X}_t + \varepsilon_t, \tag{2.60}$$

where r_t is the return for an asset, $\mathbf{X}_t$ is a vector of factors, say, factors in the three (four, five or six)-factors Fama-French type model, and $\mathbf{Z}_t$ is a variable that drives the β to change over time. If the dimension of $\mathbf{Z}_t$ is large, Cai, Ren and Yang (2015) extended the model in (2.60) into the following functional coefficient index model

$$r_t = \beta(\gamma^\top \mathbf{Z}_t)^\top \mathbf{X}_t + \varepsilon_t, \tag{2.61}$$

which will be discussed further in Section 2.6.3.

2.5 Additive Model

2.5.1 Model

In this section, we use the notation from Cai (2002). Let $\{\mathbf{X}_t, \mathbf{Y}_t, Z_t\}_{t=-\infty}^{\infty}$ be jointly stationary processes, where $\mathbf{X}_t$ and $\mathbf{Y}_t$ take values in $\Re^p$ and $\Re^q$ with $p, q \geq 0$, respectively. The regression surface is defined by

$$m(\mathbf{x}, \mathbf{y}) = E\left\{Z_t \mid \mathbf{X}_t = \mathbf{x}, \mathbf{Y}_t = \mathbf{y}\right\}. \tag{2.62}$$

Here, it is assumed that $E\left|Z_t\right| < \infty$. Note that the regression function $m(\cdot, \cdot)$ defined in (2.62) can identify only the sum

$$m(\mathbf{x}, \mathbf{y}) = \mu + g_1(\mathbf{x}) + g_2(\mathbf{y}). \tag{2.63}$$

Such a decomposition holds, for example, for the following nonlinear additive autoregressive model with exogenous variables (ARX)

$$Y_t = \mu + g_1\left(X_{t-j_1}, \ldots, X_{t-j_p}\right) + g_2\left(Y_{t-i_1}, \ldots, Y_{t-i_q}\right) + \eta_t,$$

and

$$X_{t-j_1} = g_3\left(X_{t-j_2}, \ldots, X_{t-j_p}\right) + \varepsilon_t.$$

For detailed discussions on the ARX model, the reader is referred to the papers by Masry and Tjøstheim (1997) and Cai and Masry (2000). For identifiability, it is assumed that $E\left\{g_1\left(\mathbf{X}_t\right)\right\} = 0$ and $E\left\{g_2\left(\mathbf{Y}_t\right)\right\} = 0$. Then, the projection of $m(\mathbf{x}, \mathbf{y})$ on the $g_1(\mathbf{x})$-direction is defined by

$$E\left\{m\left(\mathbf{x}, \mathbf{Y}_t\right)\right\} = \mu + g_1(\mathbf{x}) + E\left\{g_2\left(\mathbf{Y}_t\right)\right\} = \mu + g_1(\mathbf{x}). \tag{2.64}$$

Clearly, $g_1(\cdot)$ can be identified up to an additive constant and $g_2(\cdot)$ can be retrieved likewise.

A thorough discussion of additive time series models defined in (2.63) can be found in Chen and Tsay (1993). Additive components can be estimated with a one-dimensional nonparametric rate. In most papers, to estimate additive components, several methods have been proposed. For example, Chen and Tsay (1993) used the iterative backfitting procedures, such as the ACE algorithm and the BRUTO approach; see Hastie and Tibshirani (1990) for details. But, their asymptotic properties are not well understood due to the implicit definition of the resulting estimators. To attenuate the drawbacks of iterative procedures, Auestad and Tjøstheim (1991) and Tjøstheim and Auestad (1994a) proposed a direct method based on an average regression surface idea, referred to as projection method in Tjøstheim and Auestad (1994a) for time series data. As pointed out by Cai and Fan (2000), a direct method has some advantages, such as it does not rely on iterations, it can make computation fast, and more importantly, it allows an asymptotic analysis. Finally, the projection method was extended to nonlinear ARX models by Masry and Tjøstheim (1997) using the kernel method and Cai and Masry (2000) coupled with the local polynomial approach. It should be remarked that the projection method, under the name of marginal integration, was proposed independently by Newey (1994) and Linton and Nielsen (1995) for iid samples, and since then, some important progresses have been made by some authors. For example, by combining the marginal integration with one-step backfitting, Linton (1997,2000) presents an efficient estimator, Mammen, Linton, and Nielsen (1999) established rigorously the asymptotic theory of the backfitting, Cai and Fan (2000) considered estimating each component using the weighted projection method coupled with the local linear fitting in an efficient way, and Sperlich, Tjøstheim, and Yang (2002) extended the efficient method to models with simple interactions.

The projection method has some disadvantages although it has the aforementioned merits. The projection method may not be efficient if covariates (endogenous or exogenous variables) are strongly correlated, which is particularly relevant for autoregressive models. The intuitive interpretation is that additive components are not orthogonal. To overcome this shortcoming, two efficient estimation methods have been proposed in the literature. The first one is called weight function procedure, proposed by Fan, Härdle, and Mammen (1998) for iid samples and extended to time series situations by Cai and Fan (2000). With an appropriate choice of the weight function, additive components can be efficiently estimated in the sense that an additive component can be estimated with the same asymptotic bias and variance as if the rest of components were known. The second one is to combine the marginal integration with one-step backfitting, introduced by Linton (1997,2000) for iid samples and extended by Sperlish, Tjøstheim, and Yang (2002) to additive models with single interactions, but this method has not been advocated for time series situations. However, there has not been any attempt to discuss the bandwidth selection for the projection method and its variations in the literature due to their complexity. In practice, one bandwidth is usually used for all components although Cai and Fan (2000) argued that different bandwidths might be used theoretically to deal with the situation that additive components posses the different smoothness. Therefore, the projection method may not be optimal in practice in the sense that one bandwidth is used.

To estimate unknown additive components in (2.63 efficiently, following the spirit of the marginal integration with one-step backfitting proposed by Linton (1997) for iid samples, I use a two-stage method, due to Linton (2000), coupled with the local linear (polynomial) method, which has some attractive properties, such as mathematical efficiency, bias reduction and adaptation of edge effect (see Fan and Gijbels, 1996). The basic idea of the two-stage approach is described as follows. At the first stage, one obtains the initial estimated values for all components. More precisely, the idea for estimating any additive component is first to estimate directly high-dimensional regression surface by the local linear method and then to average the regression surface over the rest of variables to stabilize variance. Such an initial estimate, in general, is under-smoothed so that the bias should be asymptotically negligible. At the second stage, the local linear (polynomial) technique is used again to estimate any additive component by using the initial estimated values of the rest of components. In such a way, it is shown that the estimate at the second stage is not only efficient in the sense of being equivalent to a procedure based on knowing other components, but also making the bandwidth selection much easier. Note that this technique is not novel to this chapter since the two-stage method is first used by Linton (1997,2000) for iid samples, but many details and insights are.

2.5.2 Backfitting Algorithm

The building block of the generalized additive model algorithm is the scatterplot smoother. We will first describe scatterplot smoothing in a simple setting, and then indicate how it is used in generalized additive modeling. Here y is a response or outcome variable, and x is a prognostic factor. We wish to fit a smooth curve $f(x)$ that summarizes the dependence of y on x. If we were to find the curve that simply minimizes $\sum_{i=1}^{n}\left[y_i - f\left(x_i\right)\right]^2$, the result would be an interpolating curve that would not be smooth at all. The cubic spline smoother imposes smoothness on $f(x)$. We seek the function $f(x)$ that minimizes

$$\sum_{i=1}^{n}\left[y_i - f\left(x_i\right)\right]^2 + \lambda \int \left[f''(x)\right]^2 dx. \tag{2.65}$$

Notice that $\int \left[f''(x)\right]^2 dx$ measures the "wiggliness" of the function $f(x)$: linear $f(x)$ s have $\int \left[f''(x)\right]^2 dx = 0$, while non-linear fs produce values bigger than zero. λ is a non-negative smoothing parameter that must be chosen by the data analyst. It governs the tradeoff between the goodness of fit to the data and (as measured by and wiggleness of the function. Larger values of λ force $f(x)$ to be smoother.

For any value of λ, the solution to (2.65) is a cubic spline, i.e., a piecewise cubic polynomial with pieces joined at the unique observed values of x in the dataset. Fast and stable numerical procedures are available for computation of the fitted curve. What value of did we use in practice? In fact it is not a convenient to express the desired smoothness of $f(x)$ in terms of λ, as the meaning of λ depends on the units of the prognostic factor x. Instead, it is possible to define an *effective number of parameters* or *degrees of freedom* of a cubic spline smoother, and then use a numerical search to determine the value of λ to yield this number. In practice, if we chose the effective number of parameters to be 5 , roughly speaking, this means that the complexity of the curve is about the same as a polynomial regression of degrees 4. However, the cubic spline smoother "spreads out" its parameters in a more even manner, and hence is much more flexible than a polynomial regression. Note that the degrees of freedom of a smoother need not be an integer.

The above discussion tells how to fit a curve to a single prognostic factor. With multiple prognostic factors, if x_{ij} denotes the value of the j th prognostic factor for the i th observation, we fit the additive model

$$y_i = \sum_{j=1}^{d} f_j\left(x_{ij}\right) + \varepsilon_i.$$

A criterion like (2.65) can be specified for this problem, and a simple iterative procedure exists for estimating the f_j s. We apply a cubic spline smoother to the outcome $y_i - \sum_{j\neq k}^{d} \widehat{f_j}\left(x_{ij}\right)$ as a function of x_{ik}, for each prognostic factor in turn. The process is continues until the estimates $\widehat{f}_j(x)$ stabilize. These procedure is known as "back-fitting" and the resulting fit is analogous to a multiple regression for linear models.

To fit an additive model or a partially additive model in R, the function is **gam()** in the package **gam**. For details, please look at the help command **help(gam)** after loading the package **gam** "**library(gam)**" . Note that the function **gam()** allows to fit a semiparametric additive model as

$$Y = \boldsymbol{\beta}^T \mathbf{X} + \sum_{j=1}^{p} g_j\left(Z_j\right) + \varepsilon,$$

which can be done by specifying some components without smooth.

2.5.3 Projection Method

This section is devoted to a brief review of the projection method and discusses its merits and disadvantages.

It is assumed that all additive components have continuous second partial derivatives, so that $m(\mathbf{u}, \mathbf{v})$ can be locally approximated by a linear term in a neighborhood of $(\mathbf{x}, \mathbf{y})$, namely, $m(\mathbf{u}, \mathbf{v}) \approx \beta_0 + \boldsymbol{\beta}_1^T(\mathbf{u} - \mathbf{x}) + \boldsymbol{\beta}_2^T(\mathbf{v} - \mathbf{y})$ with $\left\{\boldsymbol{\beta}_j\right\}$ depending on $\mathbf{x}$ and $\mathbf{y}$, where $\boldsymbol{\beta}_1^T$ denotes the transpose of $\boldsymbol{\beta}_1$.

Let $K(\cdot)$ and $L(\cdot)$ be symmetric kernel functions in $\Re^p$ and $\Re^q$, respectively, and $h_{11} = h_{11}(n) > 0$ and $h_{12} = h_{12}(n) > 0$ be bandwidths in the step of estimating the regression surface. Here, to handle various degrees of smoothness, Cai and Fan (2000) propose using h_{11} and h_{12} differently although the implementation may not be easy in practice. The reader is referred to the paper by Cai and Fan (2000) for details. Given observations $\{\mathbf{X}_t, \mathbf{Y}_t, Z_t\}_{t=1}^n$, let $\widehat{\boldsymbol{\beta}}_j$ be the minimizer of the following locally weighted least squares

$$\sum_{t=1}^{n} \left\{Z_t - \beta_0 - \boldsymbol{\beta}_1^T\left(\mathbf{X}_t - \mathbf{x}\right) - \boldsymbol{\beta}_2^T\left(\mathbf{Y}_t - \mathbf{y}\right)\right\}^2 K_{h_{11}}\left(\mathbf{X}_t - \mathbf{x}\right) L_{h_{12}}\left(\mathbf{Y}_t - \mathbf{y}\right),$$

where $K_h(\cdot) = K(\cdot/h)/h^p$ and $L_h(\cdot) = L(\cdot/h)/h^q$. Then, the local linear estimator of the regression surface $m(\mathbf{x}, \mathbf{y})$ is $\widehat{m}(\mathbf{x}, \mathbf{y}) = \widehat{\beta}_0$. By computing the sample average of $\widehat{m}(\cdot, \cdot)$ based on (2.64), the projection estimators of $g_1(\cdot)$ and $g_2(\cdot)$ are defined as, respectively,

$$\widehat{g}_1(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} \widehat{m}\left(\mathbf{x}, \mathbf{Y}_t\right) - \widehat{\mu}, \quad \text{and} \quad \widehat{g}_2(\mathbf{y}) = \frac{1}{n}\sum_{t=1}^{n} \widehat{m}\left(\mathbf{X}_t, \mathbf{y}\right) - \widehat{\mu},$$

where $\widehat{\mu} = n^{-1}\sum_{t=1}^n Z_t$. Under some regularity conditions, by using the same arguments as those employed in the proof of Theorem 3 in Cai and Masry (2000), it can be shown (although not easy and tedious) that the asymptotic bias and asymptotic variance of $\widehat{g}_1(\mathbf{x})$ are, respectively, $h_{11}^2 \operatorname{tr}\left\{\mu_2(K) g_1''(\mathbf{x})\right\}/2$ and $v_1(\mathbf{x}) = \nu_0(K) A(\mathbf{x})$, where

$$A(\mathbf{x}) = \int p_2^2(\mathbf{y}) \sigma^2(\mathbf{x}, \mathbf{y}) p^{-1}(\mathbf{x}, \mathbf{y}) d\mathbf{y} \quad \text{ and } \quad \sigma^2(\mathbf{x}, \mathbf{y}) = \operatorname{Var}\left(Z_t \mid \mathbf{X}_t = \mathbf{x}, \mathbf{Y}_t = \mathbf{y}\right).$$

Here, $p(\mathbf{x}, \mathbf{y})$ stands for the joint density of $\mathbf{X}_t$ and $\mathbf{Y}_t, p_1(\mathbf{x})$ denotes the marginal density of $\mathbf{X}_t, p_2(\mathbf{y})$ is the marginal density of $\mathbf{Y}_t, \nu_0(K) = \int K^2(\mathbf{u}) d\mathbf{u}$, and $\mu_2(K) = \int \mathbf{u}\mathbf{u}^T K(\mathbf{u}) d\mathbf{u}$.

The foregoing method has some advantages, such as it is easy to understand, it can make computation fast, and it allows an asymptotic analysis. However, it can be quite inefficient in an asymptotic sense. To demonstrate this idea, let us consider the ideal situation that $g_2(\cdot)$ and μ are known. In such a case, one can estimate $g_1(\cdot)$ by directly regressing the partial error $\widetilde{Z}_t = Z_t - \mu - g_2\left(\mathbf{Y}_t\right)$ on $\mathbf{X}_t$ and such an ideal estimator is optimal in an asymptotic minimax sense (see, e.g., Fan and Gijbels, 1996). The asymptotic bias for the ideal estimator is $h_{11}^2 \operatorname{tr}\left\{\mu_2(K) g_1''(\mathbf{x})\right\}/2$ and the asymptotic variance is

$$v_0(\mathbf{x}) = \nu_0(K) B(\mathbf{x}) \quad \text{ with } \quad B(\mathbf{x}) = p_1^{-1}(\mathbf{x}) E\left\{\sigma^2\left(\mathbf{X}_t, \mathbf{Y}_t\right) \mid \mathbf{X}_t = \mathbf{x}\right\} \tag{2.66}$$

(see, e.g., Masry and Fan, 1997). It is clear that $v_1(\mathbf{x}) = v_0(\mathbf{x})$ if $\mathbf{X}_t$ and $\mathbf{Y}_t$ are independent. If $\mathbf{X}_t$ and $\mathbf{Y}_t$ are correlated and when $\sigma^2(\mathbf{x}, \mathbf{y})$ is a constant, it follows from the CauchySchwarz inequality that

$$B(\mathbf{x}) = \frac{\sigma^2}{p_1(\mathbf{x})} \int p^{1/2}(\mathbf{y} \mid \mathbf{x}) \frac{p_2(\mathbf{y})}{p^{1/2}(\mathbf{y} \mid \mathbf{x})} d\mathbf{y} \leq \frac{\sigma^2}{p_1(\mathbf{x})} \int \frac{p_2^2(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})} d\mathbf{y} = A(\mathbf{x})$$

which implies that the ideal estimator has always smaller asymptotic variance than the projection method although both have the same bias. This suggests that the projection method could lead to an inefficient estimation of $g_1(\cdot)$ and $g_2(\cdot)$ when $\mathbf{X}_t$ and $\mathbf{Y}_t$ are serially correlated, which is particularly relevant for autoregressive models. To alleviate this shortcoming, I propose the two-stage approach described next.

2.5.4 Two-Stage Procedure

The two-stage method due to Linton (1997,2000) is introduced. The basic idea is to get an initial estimate for $\widehat{g}_2(\cdot)$ using a small bandwidth h_{12}. The initial estimate can be obtained by the projection method and h_{12} can be chosen so small that the bias of estimating $\widehat{g}_2(\cdot)$ can be asymptotically negligible. Then, using the partial residuals $Z_t^* = Z_t - \widehat{\mu} - \widehat{g}_2\left(\mathbf{Y}_t\right)$, we apply the local linear regression technique to the pseudo regression model

$$Z_t^* = g_1\left(\mathbf{X}_t\right) + \varepsilon_t^*$$

to estimate $g_1(\cdot)$. This leads naturally to the weighted least-squares problem

$$\sum_{t=1}^{n} \left\{ Z_t^* - \beta_1 - \boldsymbol{\beta}_2^T \left(\mathbf{X}_t - \mathbf{x}\right) \right\}^2 J_{h_2}\left(\mathbf{X}_t - \mathbf{x}\right), \tag{2.67}$$

where $J(\cdot)$ is the kernel function in $\Re^p$ and $h_2 = h_2(n) > 0$ is the bandwidth at the second stage. The advantage of this is twofold: the bandwidth h_2 can now be selected purposely for estimating $g_1(\cdot)$ only and any bandwidth selection technique for nonparametric regression can be applied here. Maximizing (2.67) with respect to β_1 and $\boldsymbol{\beta}_2$ gives the two-stage estimate of $g_1(\mathbf{x})$, denoted by $\widetilde{g}_1(\mathbf{x}) = \widehat{\beta}_1$, where $\widehat{\beta}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are the minimizer of (2.67).

It is shown in Theorem 2.3, in which follows, that under some regularity conditions, the asymptotic bias and variance of the two-stage estimate $\widetilde{g}_1(\mathbf{x})$ are the same as those for the ideal estimator, provided that the initial bandwidth h_{12} satisfies $h_{12} = o\left(h_2\right)$.

Sampling Properties

To establish the asymptotic normality of the two-stage estimator, it is assumed that the initial estimator satisfies a linear approximation; namely,

$$\widehat{g}_2\left(\mathbf{Y}_t\right) - g_2\left(\mathbf{Y}_t\right) \approx \frac{1}{n}\sum_{i=1}^{n} L_{h_{12}}\left(\mathbf{Y}_i - \mathbf{Y}_t\right)\Gamma\left(\mathbf{X}_i, \mathbf{Y}_t\right)\delta_i + \frac{1}{2}h_{12}^2 \mathrm{tr}\left\{\mu_2(L) g_2''\left(\mathbf{Y}_t\right)\right\}, \tag{2.68}$$

where $\delta_t = Z_t - m\left(\mathbf{X}_t, \mathbf{Y}_t\right)$ and $\Gamma(\mathbf{x}, \mathbf{y}) = p_1(\mathbf{x})/p(\mathbf{x}, \mathbf{y})$. Note that under some regularity conditions, by following the same arguments as in Masry (1996), one might show (although the proof is not easy, quite lengthy, and tedious) that (2.68) holds. Note that this assumption is also imposed in Linton (2000) for iid samples to simplify the proof of the asymptotic results of the two-stage estimator. Now, the asymptotic normality for the two-stage estimator is stated here and its proof can be found in Cai (2002).

Theorem 2.3: *Under* (2.68) *and Assumptions A1-A9 stated in Cai (2002), if bandwidths* h_{12} *and* h_2 *are chosen such that* $h_{12} \to 0, nh_{12}^q \to \infty, h_2 \to 0,$ *and* $nh_2^p \to \infty$ *as* $n \to \infty$, *then*

$$\sqrt{nh_2^p}\left[\widetilde{g}_1(\mathbf{x}) - g_1(\mathbf{x}) - bias(\mathbf{x}) + o_p\left(h_{12}^2 + h_2^2\right)\right] \quad \to \quad N\left\{0, v_0(\mathbf{x})\right\},$$

where the asymptotic bias is

$$bias(\mathbf{x}) = \frac{h_2^2}{2} tr\left\{\mu_2(J) g_1''(\mathbf{x})\right\} - \frac{h_{12}^2}{2} tr\left\{\mu_2(L) E\left(g_2''\left(\mathbf{Y}_t\right) \mid \mathbf{X}_t = \mathbf{x}\right)\right\}$$

and the asymptotic variance is $v_0(\mathbf{x}) = \nu_0(J)B(\mathbf{x})$.

We remark that by Theorem 2.3, the asymptotic variance of the two-stage estimator is independent of the initial bandwidths. Thus, the initial bandwidths should be chosen as small as possible. This is another benefit of using the two-stage procedure: the bandwidth selection problem becomes relatively easy. In particular, when $h_{12} = o(h_2)$, the bias from the initial estimation can be asymptotically negligible. For the ideal situation that $g_2(\cdot)$ is known, Masry and Fan (1997) show that under some regularity conditions, the optimal estimate of $g_1(\mathbf{x})$, denoted by $\widehat{g}_1^*(\mathbf{x})$, by using (2.68) in which the partial residual Z_t^* is replaced by the partial error $\widetilde{Z}_t = \mathbf{Y}_t - \mu - g_2(\mathbf{Y}_t)$, is asymptotically normally distributed,

$$\sqrt{nh_2^p}\left[\widehat{g}_1^*(\mathbf{x}) - g_1(\mathbf{x}) - \frac{h_2^2}{2}\mathrm{tr}\left\{\mu_2(J) g_1''(\mathbf{x})\right\} + o_p\left(h_2^2\right)\right] \quad \rightarrow \quad N\left\{0, v_0(\mathbf{x})\right\}.$$

This, in conjunction with Theorem 2.3, shows that the two-stage estimator and the ideal estimator share the same asymptotic bias and variance if $h_{12} = o(h_2)$.

Finally, note that the reader is referred to the paper by Cai (2002) for the detailed Monte Carlo simulation results and applications. Also, one can see the paper by Mammen, Linton and Nielsen (1999) for some more approaches on additive modeling.

2.5.5 Analysis of the Boston House Price Data via Additive Model

There have been several papers devoted to the analysis of this dataset using some nonparametric methods. For example, Breiman and Friedman (1985), Pace (1993), Chaudhuri, Doksum and Samarov (1997), and Opsomer and Ruppert (1998) used four covariates: X_6, X_{10}, X_{11} and X_{13} or their transformations (including the transformation on Y) to fit the data through a mean additive regression model such as

$$\log(Y) = \mu + g_1(X_6) + g_2(X_{10}) + g_3(X_{11}) + g_4(X_{13}) + \varepsilon, \tag{2.69}$$

where the additive components $\{g_j(\cdot)\}$ are unspecified smooth functions. Pace (1993) and Chaudhuri, Doksum and Samarov (1997) also considered the nonparametric estimation of the first derivative of each additive component which measures how much the response changes as one covariate is perturbed while the other covariates are held fixed; see Chaudhuri, Doksum and Samarov (1997). Let us use model (2.69) to fit the Boston house price data. The results are summarized in Figure 2.4 (the **R** code can be found in Section 2.8.2). Also, we fit a semi-parametric additive model (partially linear model as in (2.71)) as

$$\log(Y) = \mu + g_1(X_6) + \beta_2 X_{10} + \beta_3 X_{11} + \beta_4 X_{13} + \varepsilon. \tag{2.70}$$

The results are summarized in Figure 2.5 (the **R** code can be found in Section 2.8.2).

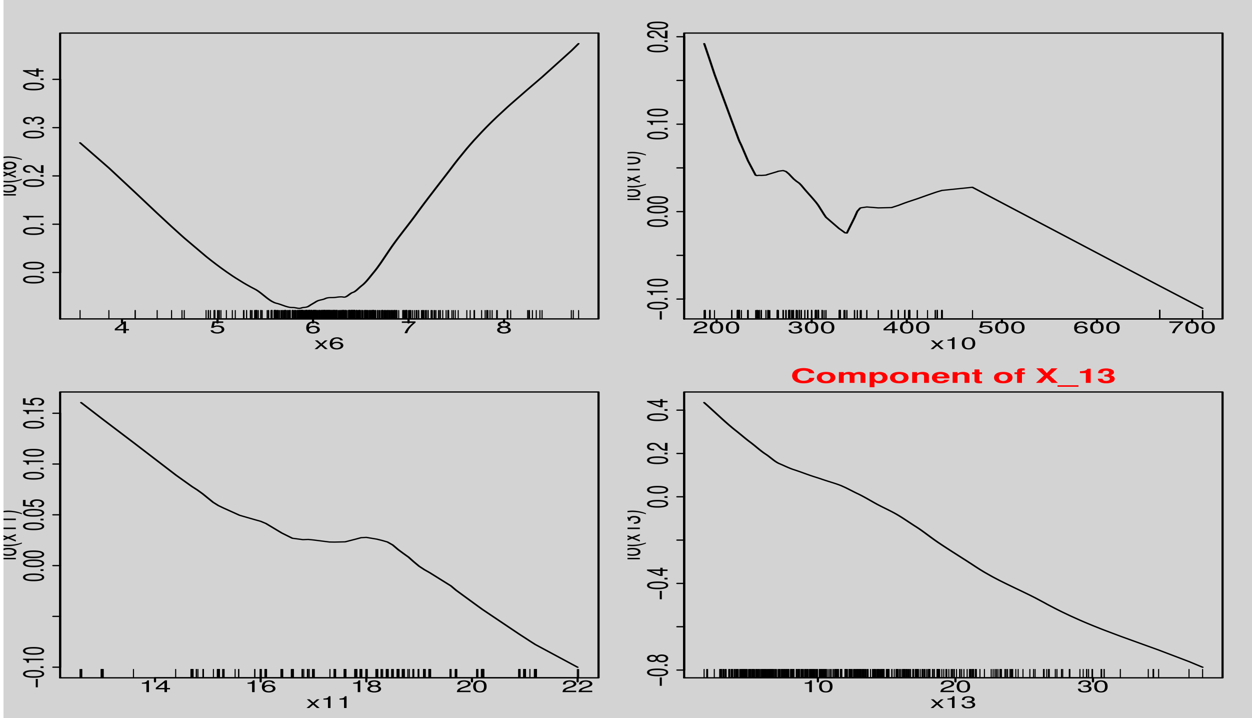


Figure 2.4: The results from model (2.69).

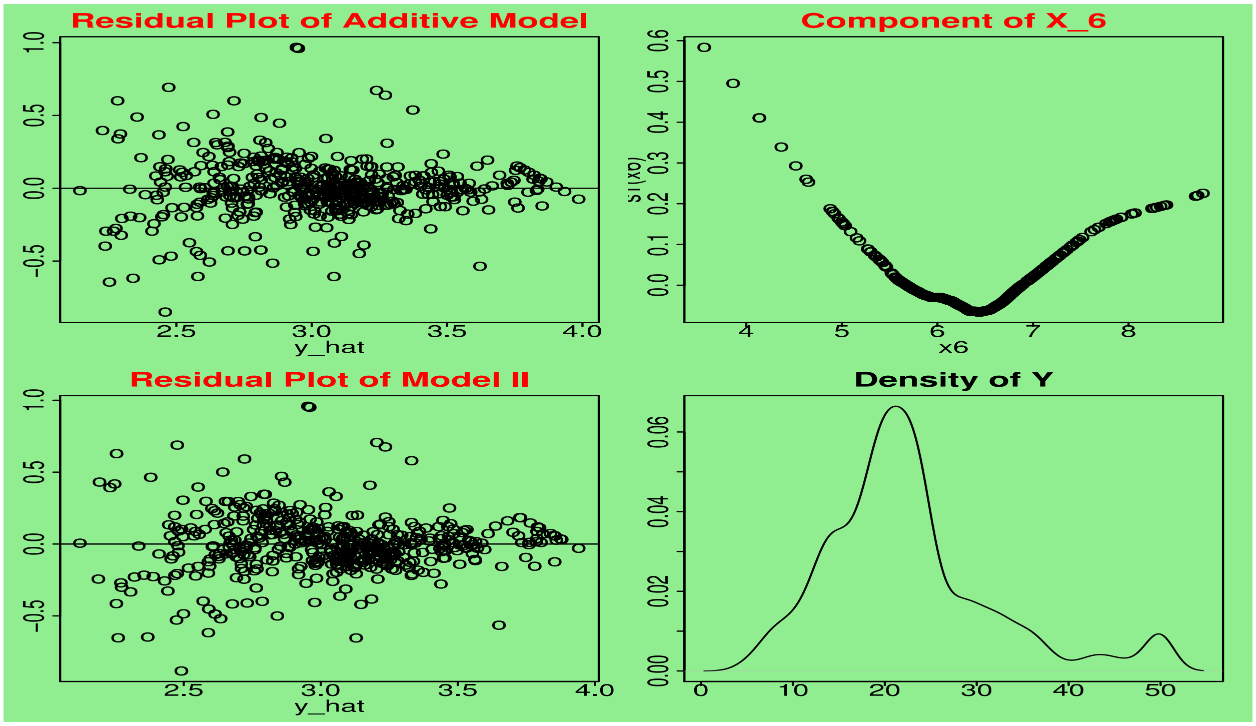


Figure 2.5: (a) Residual plot for model (2.69). (b) Plot of $g_1(x_6)$ versus x_6. (c) Residual plot for model (2.70). (d) Density estimate of Y.

2.6 Semiparametric Models

2.6.1 Partially Linear Models

Initiated by applications in economics as in Shiller (1984) for estimating the U-shaped cost curve from the electric utility industry and Engle et al. (1986) for a nonlinear relationship between electricity sales and temperature, we consider the following partially linear model

$$E\left(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t\right) = \beta^\top \mathbf{X}_t + g(\mathbf{Z}_t), \tag{2.71}$$

where $g(\cdot)$ is an unknown link function. From (2.71), one can obtain

$$E(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t) - E(Y_t|Z_t) = \beta^\top[\mathbf{X}_t - E(\mathbf{X}_t|\mathbf{Z}_t)],$$

which leads to the estimate of β by the method of moment estimation approach, proposed by Robinson (1988), given by

$$\widehat{\beta} = \left(\sum_{t=1}^{n}(\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))(\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))^\top\right)^{-1}\sum_{t=1}^{n}(\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))(Y_t - \widehat{m}_y(\mathbf{Z}_t)),$$

where $m_x(\mathbf{Z}_t) = E(\mathbf{X}_t|Z_t)$ and $m_y(\mathbf{Z}_t) = E(Y_t|\mathbf{Z}_t)$. Here, $\widehat{m}_x(\mathbf{z})$ is a nonparametric estimate of $m_x(\mathbf{z})$ and $\widehat{m}_y(\mathbf{z})$ is a nonparametric estimate of $m_y(\mathbf{z})$. Now, having estimated parameter vector β, it is possible to fit the nonlinear relation between $\mathbf{Z}_t$ and Y_t by simply estimating equation (2.72) presented below nonparametrically

$$\widehat{Y}_t = Y_t - \widehat{\beta}^\top \mathbf{X}_t = g(\mathbf{Z}_t) + u_t. \tag{2.72}$$

Then, the local linear (or polynomial) technique can be applied here to estimate $g(\cdot)$. Robinson (1988) investigated the asymptotic properties of $\widehat{\beta}$ (consistency and asymptotic normality). Especially, Robinson (1988) showed that $\widehat{\beta}$ is efficient in the sense that its asymptotic variance of $\widehat{\beta}$ is the smallest even $g(\cdot)$ is unknown. However, the method by Robinson (1988) needs to estimate both $m_x(\mathbf{Z}_t) = E(\mathbf{X}_t|Z_t)$ and $m_y(\mathbf{Z}_t) = E(Y_t|\mathbf{Z}_t)$ nonparametrically.

To avoid the above disadvantage, one can the so-called profile least squares method proposed by Speckman (1988), described as follows, which becomes profile likelihood estimation when the error is normally distributed; see, for example, Speckman (1988). Suppose that we have a random sample of size n, $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t)\}_{t=1}^n$ from model (2.72). For given β, (2.72) becomes

$$Y_t(\beta) = Y_t - \beta^\top \mathbf{X}_t = g(\mathbf{Z}_t) + u_t, \tag{2.73}$$

where $Y_t(\beta) = Y_t - \beta^\top \mathbf{X}_t$. This transforms the partially linear model in (2.72) into the conventional nonparametric regression model as in Section 2.3. Then, the local linear regression technique outlined in Section 2.3 is applied to estimating the function $g(\cdot)$ in (2.72). Thus, one can express $\widehat{g}(\mathbf{Z}_t)$ as

$$\begin{pmatrix} \widehat{g}(\mathbf{Z}_1) \\ \vdots \\ \widehat{g}(\mathbf{Z}_n) \end{pmatrix} = \mathbf{S}\,\mathbf{Y}(\beta) = \mathbf{S}\,(\mathbf{Y} - \mathbf{X}\beta),$$

where $\mathbf{S}$ is the smooth matrix, and substitute $\widehat{\mathbf{S}\mathbf{Y}}$ into (2.73) to obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{Y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\beta + \mathbf{u}.$$

Applying the least squares method to obtain the ordinary least squares estimate of β, denoted by $\widehat{\beta}_{pls}$, given by

$$\widehat{\beta}_{pls} = [\mathbf{X}^\top(\mathbf{I} - \mathbf{S})^\top(\mathbf{I} - \mathbf{S})\mathbf{X}]^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{S})^\top(\mathbf{I} - \mathbf{S})\mathbf{Y}.$$

Moreover, $(\widehat{g}(\mathbf{Z}_1, \ldots, \widehat{g}(\mathbf{Z}_n))^\top = \mathbf{S}\,(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{pls})$. Speckman (1988) showed that $\widehat{\beta}_{pls}$ is semiparametrically efficient. Clearly, the profile least squares method is better than that by Robinson (1988), since it needs only to estimate a nonparametric regression function of $Y_t(\beta)$ versus $\mathbf{Z}_t$. Furthermore, the model in (2.72) was extended by Fan and Huang (2005) to the following varying-coefficient partially linear model

$$Y_t = \beta^\top \mathbf{X}_t + a(\mathbf{Z}_t)^\top \mathbf{W}_t + v_t, \tag{2.74}$$

Then, Fan and Hung (2005) applied the profile least squares approach to estimate both $a(\cdot)$ and β and derived the asymptotic normality of the profile least-squares estimator. Also, they showed that the profile least squares estimator of β is semi-parametrically efficient and the model in (2.74) was employed by Fan and Huang to study the Boston housing data; see, for instance, Fan and Huang (2005) for details. Add more

2.6.2 Single Index Models

An object of interest such as the conditional density $f(y \mid x)$ or conditional distribution $F(y \mid x)$ or conditional mean $E\,(Y_t \mid X_t = x)$ is a single index model when it only depends on the vector x through a single linear combination of x as $\beta^\top x$. Indeed, most parametric models are single index, including, for example, normal regression, logit, probit, Tobit, and

Poisson regression. In a semiparametric single index model, the object of interest depends on x through the function $g\left(\beta^\top x\right)$, where $\beta \in \mathbb{R}^k$ and $g : \mathbb{R} \to \mathbb{R}$ are unknown. $g(\cdot)$ is sometimes called a link function in generalized line model (GLM) literature. In single index models, there is only one nonparametric dimension. These methods fall in the class of dimension reduction techniques. The semiparametric single index regression model is

$$E\left(Y_t \mid X_t = x\right) = g\left(\beta^\top x\right), \tag{2.75}$$

where $g(\cdot)$ is an unknown link function. The semiparametric single index binary choice model is

$$P\left(Y_t = 1 \mid X_t = x\right) = E\left(Y_t \mid X_t = x\right) = g\left(\beta^\top x\right), \tag{2.76}$$

where $g(\cdot)$ is an unknown distribution function.[5] We use $g(\cdot)$ (rather than, say, $F(\cdot)$) to emphasize the connection with the regression model.

In both contexts, the function $g(\cdot)$ includes any location and level shift, so the vector X_t cannot include an intercept. The level of β is not identified, so some normalization criterion for is needed. It is typically easier to impose this on β than on $g(\cdot)$. One approach is to set $\beta^\top\beta = 1$. A second approach is to set one component of β to equal one. (This second approach requires that this variable correctly has a non-zero coefficient.) The vector X_t must be dimension 2 or larger. If X_t is one-dimensional, then β is simply normalized to one, and the model is the one-dimensional nonparametric regression $E\left(Y_t \mid X_t = x\right) = g(x)$ with no semiparametric component. Identification of β and $g(\cdot)$ also requires that X_t contains at least one continuously distributed variable, and that this variable has a non-zero coefficient. If not, $\beta^\top x$ only takes a discrete set of values, and it would be impossible to identify a continuous function $g(\cdot)$ on this discrete support. Therefore, in what follows, it is assumed that the model in (2.75) is identified without a further mention.

A. Ichumura's Estimator

The semiparametric single index regression model is

$$Y_t = g\left(\beta^\top X_t\right) + e_t$$

[5] If $g(\cdot)$ is a known link function, the model in (2.76) is either logist or probit model, popularly in many applications in economics and finance' see, the books by Hastie and TibShibrani (1990) and Cameron and Trivedi (2005) for details.

with $E\left(e_t \mid X_t\right) = 0$, and it generalizes the linear regression model (which sets $g(\cdot)$ to be linear), and is a restriction of the nonparametric regression model. The gain over full nonparametric setting is that there is only one nonparametric dimension, so the curse of dimensionality is avoided. Suppose that $g(\cdot)$ were known. Then, you could estimate β by a (nonlinear) least-squares (LS) with the LS criterion

$$S_n(\beta, g) = \sum_{t=1}^{n} \left(Y_t - g\left(\beta^\top X_t\right)\right)^2. \tag{2.77}$$

You could think about to replacing $g(\cdot)$ by $\widehat{g}(\cdot)$. But, since $g(\cdot)$ is unknown conditional mean of Y_t given $\beta^\top X_t = z, g(\cdot)$ depends on β, so that Ichimura (1993) suggested a two-step (2LS) estimation procedure as follows. First, a leave-one out Nadaraya-Waston estimation of $g(\cdot)$ is used

$$\widehat{g}_{-t}\left(\beta^\top X_t\right) = \sum_{st/}^{n} Y_s K\left(\beta^\top\left(X_s - X_t\right)/h\right) / \sum_{s\neq t}^{n} K\left(\beta^\top\left(X_s - X_t\right)/h\right), \tag{2.78}$$

and then, Ichimura (1993) suggested replacing $g(\cdot)$ in $S_n(\beta, g)$ by $\widehat{g}_{-t}\left(\beta^t opX_t\right)$,

$$S_n(\beta) = \sum_{t=1}^{n} \left(Y_t - \widehat{g}_{-t}\left(\beta^\top X_t\right)\right)^2 I_t(b),$$

where $I_t(b)$ is a trimming function to make the computation easy. The Ichimura's estimator is $\widehat{\beta}_{\text{2LS}} = \operatorname{argmin} l_n(\beta)$. However, Ichimura (1993) did not discuss on how to choose $I_t(b)$ in $S_n(\beta)$. As pointed out by by Härdle, Hall, and Ichimura (1993), the criterion in the Ichimura's estimator is somewhat similar to cross-validation so that the Ichimura's estimator may not be efficient (optimal). To obtain the efficient estimation of β, Härdle, Hall, and Ichimura (1993) suggested picking β and the bandwidth h simultaneously by minimizing $S_n(\beta)$, denoted by $\widehat{\beta}_{\text{hhi}}$.

Finally, for the asymptotic theory for $\widehat{\beta}_{\text{2LS}}$ or $\widehat{\beta}_{\text{hhi}}$, the reader is referred to the papers by Ichimura (1993) and Härdle, Hall, and Ichimura (1993), respectively.

B. Klein and Spady's Binary Choice Estimator

Klein and Spady (1993) proposed an estimator of the semiparametric single index binary choice model which has strong similarities with Ichimura's estimator. The model is given in (2.76) and can be re-expressed as follows

$$Y_t = I\left(\beta^\top X_t > e_t\right),$$

where e_t is an error, which is a special case of (2.75). If e_t is independent of X_t, and has distribution function $g(\cdot)$, then, the data satisfy the single-index regression as

$$E\left(Y_t \mid X_t\right)=g\left(\beta^{\top} X_t\right),$$

and it follows that Ichimura's estimator can be directly applied to this model. However, different from the Ichimura's estimator, Klein and Spady (1993) suggested a semiparametric likelihood approach. Given $g(\cdot)$, the log likelihood is

$$l_n(\beta, g)=\sum_{t=1}^n\left[Y_t \ln \left(g\left(\beta^{\top} X_t\right)\right)+\left(1-Y_t\right) \ln \left(1-g\left(\beta^{\top} X_t\right)\right]\right. .$$

Since $g(\cdot)$ is unknown, making this substitution of $g(\cdot)$ by $\widehat{g}_{-t}\left(\beta^{\top} X_t\right)$ in (2.78), and adding trimming function, this leads to the feasible likelihood criterion

$$l_n(\beta)=\sum_{t=1}^n\left[Y_t \ln \left(\widehat{g}_{-t}\left(\beta^{\top} X_t\right)\right)+\left(1-Y_t\right) \ln \left(1-\widehat{g}_{-t}\left(\beta^{\top} X_t\right)\right] I_t(b),\right.$$

where, as suggested by Klein and Spady (1993), the trimming indicator can be taken to be

$$I_t(b)=I\left(\widehat{f}_{\tilde{\beta}^{\top} X_t}\left(\tilde{\beta}^{\top} X_t\right)>b\right),$$

where $\tilde{\beta}$ is a preliminary estimator of β and $\widehat{f}(\cdot)$ is an estimation of the density function of $\tilde{\beta}^{\top} X_t$. It can be seen from Klein and Spady (1993) that trimming does not seem to matter in their simulations. Finally, the Klein and Spady estimator for β is the value $\widehat{\beta}$ which maximizes $l_n(\eta)$ and in many respects, the Ichimura and Klein-Spady estimators are quite similar.

C. Average Derivative Estimator

Let $m\left(X_t\right)=E\left(Y_t \mid X_t\right)$ and $m^{\prime}(x)$ denote the first order derive of $m(x)$. Define the weighted derivative as follows

$$\delta=E\left[m^{\prime}\left(X_t\right) w\left(X_t\right)\right],$$

where $w(x)$ is a weight function, which is particularly convenient to set $w(x)=f_x(x)$ with $f_x(x)$ being the marginal density of X_t, suggested by Powell, Stock and Stoker (1989). A simple algebra leads to the following expression

$$\delta=\int m^{\prime}(x) f_x^2(x)=-2 E\left[Y_t f_x^{\prime}\left(X_t\right)\right],$$

where $f'_x(x)$ is the first order derivative of $f_x(x)$, which, clearly, leads to a consistent estimate of δ, given by

$$\widehat{\delta} = -\frac{2}{n-1}\sum_{t=1}^{n} Y_t \widehat{f}'_{x,-t}\left(X_t\right)$$

where $\widehat{f}'_{x,-t}(x)$ is the first derivative of the leave-one-out density estimator of $f_x(x)$. One can see that this is a convenient estimator. There is no denominator messing with uniform convergence. There is only a density estimator, no conditional mean needed. Powell, Stock and Stoker (1989) showed that $\widehat{\delta}$ is $n^{1/2}$-consistent and asymptotically normal, with a convenient covariance matrix.

Now, for the single-index model, it is easy to see $m'(x) = \beta g'\left(\beta^\top x\right)$ so that

$$\delta = c\beta,$$

where $c = E\left[g'\left(\beta^\top X_t\right) f_x\left(X_t\right)\right]$, from which, one can obtain the average derivative estimator for β by

$$\widehat{\beta} = \widehat{\delta}/\widehat{c},$$

where $\widehat{c}$ is a consistent estimate of c. However, the problem goes back to estimating a density function with a possible a high dimension.

D. MAVE Estimator

Due to the fact that the single index model shares a close connection with the central mean subspace in the sufficient dimension reduction, Xia et al. (2002) proposed the (conditional) minimum average variance estimation (MAVE) method for the dimension reduction problem and later, Xia (2006) showed that this method can be applied to the single index model. Therefore, the MAVE method proposed in Xia (2006) is employed in our setting to estimate β and also the penalized MAVE considered in Wang et al. (2013) is utilized for selecting X, described as follows.

Notice that under the least squares loss,

$$\beta = \arg\min_{\tilde{\beta}\in\mathbb{R}^k} E\left[Y - E(Y|\tilde{\beta}^\top X)\right]^2. \tag{2.79}$$

In our setting, the index is estimated by the observed data for the control units, $\{Y_j,\, X_j\}_{j=1}^n$. Motivated by the local linear smoothing technique, the sample analogue of (2.79) can be written as

$$\begin{aligned}\beta &= \arg \min_{\tilde\beta\in\mathbb{R}^k:\tilde\beta^\top\tilde\beta=1} \sum_{j=1}^n \left\{ \min_{a_j,b_j} \sum_{i=1}^n \left[Y_i - a_j - b_j\tilde\beta^\top (X_i - X_j)\right]^2 w_{ij} \right\} \\ &= \arg \min_{\substack{\tilde\beta\in\mathbb{R}^k:\tilde\beta^\top\tilde\beta=1 \\ a_j,b_j}} \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - a_j - b_j\tilde\beta^\top (X_i - X_j)\right]^2 w_{ij}, \end{aligned} \tag{2.80}$$

where $a_j = g(\beta^\top X_j)$, $b_j = \partial g(u)/\partial u|_{u=\beta^\top X_j}$, and $w_{ij} = K_h(\beta^\top(X_i - X_j))$ with $K_h(v) = K(v/h)/h$ and $K(\cdot)$ being a kernel function as well as h being the bandwidth. Xia (2006) proposed the following algorithm for estimating β:

Step 1. Set an initial value $\beta^{(0)}$.

Step 2. For $\ell \geq 1$, calculate

$$\begin{pmatrix} \widehat{a}_j^{\beta^{(\ell-1)}} \\ \widehat{d}_j^{\beta^{(\ell-1)}} h \end{pmatrix} = \left\{ \sum_{j=1}^n K_h\left(\beta^{(\ell-1)^\top} X_{ij}\right) Z_{ij}^{(k-1)} Z_{ij}^{(\ell-1)^\top} \right\}^{-1} \sum_{j=1}^n K_h\left(\beta^{(\ell-1)^\top} X_{ij}\right) Z_{ij}^{(\ell-1)} Y_j,$$

where $Z_{ij}^{(\ell-1)} = \left(1, \beta^{(\ell-1)^\top} X_{ij}/h\right)^\top$ with $X_{ij} = X_i - X_j$, and also, obtain

$$\widehat{f}_{\beta^{(\ell-1)}}(\beta^{(\ell-1)^\top} X_j) = \frac{1}{n}\sum_{i=1}^n K_h(\beta^{(\ell-1)^\top} X_{ij}), \quad \text{and} \quad \widehat{\rho}_j^{\beta^{(\ell-1)^\top}} = \rho_n(\widehat{f}_{\beta^{(\ell-1)}}(\beta^{(\ell-1)^\top} X_j)),$$

where $\rho_n(\cdot)$ is a trimming function for the boundary points. Following the suggestion from Xia (2006), $\rho_n(v)$ is chosen as a bounded function with bounded derivative on $\mathbb{R}$ such that $\rho_n(v) = I(v > 2c_0 n^{-\varepsilon})$, where $I(A)$ is the indicator function of set A.

Step 3. Calculate

$$\begin{aligned}\beta^{(\ell)} = &\left\{ \sum_{i=1}^n \sum_{j=1}^n K_h\left(\beta^{(\ell-1)^\top} X_{ij}\right) \widehat{\rho}_j^{\beta^{(k-1)}} \left(\widehat{d}_j^{\beta^{(\ell-1)}}\right)^2 X_{ij} X_{ij}^\top / \widehat{f}_{\beta^{(\ell-1)}}\left(\beta^{(\ell-1)^\top} X_j\right) \right\}^{-1} \\ &\times \sum_{i=1}^n \sum_{j=1}^n K_h\left(\beta^{(\ell-1)^\top} X_{ij}\right) \widehat{\rho}_j^{\beta^{(\ell-1)}} \widehat{d}_j^{\beta^{(\ell-1)}} X_{ij} \left(Y_i - \widehat{a}_j^{\beta^{(\ell-1)}}\right) / \widehat{f}_{\beta^{(\ell-1)}}\left(\beta^{(\ell-1)^\top} X_j\right).\end{aligned}$$

Step 4. Set $\beta^{(\ell)} = \text{sign}(\beta^{(\ell)})\beta^{(\ell)}/\|\beta^{(\ell)}\|$. Then, repeat Steps 2 and 3 until convergence reaches.

Denote the ultimate estimator for β as $\widehat{\beta}_{\text{MAVE}}$. Theoretically, Xia (2006) derived the asymptotic normality for $\widehat{\beta}_{\text{MAVE}}$ and showed that the asymptotic covariance matrix of $\widehat{\beta}_{\text{MAVE}}$ can achieve the information lower bound in the semiparametric sense. From Xia (2006), one can see that under some regularity conditions, $\widehat{\beta}_{\text{MAVE}}$ has the following asymptotic behavior

$$\sqrt{n}\left[\widehat{\beta}-\beta\right]=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\phi(X_j,Y_j)+o_p(1)\ \rightarrow\ N(0,\Sigma_\beta), \tag{2.81}$$

where $\phi(X_j,Y_j) = W_g^{+}g'(\beta^\top X_j)v_\beta(X_j)\,e_j$, $W_g = E\left\{g'(\beta^\top X)^2 v_\beta(X)v_\beta^\top(X)\right\}$, $W_{m_0}^{+}$ is the Moore-Penrose inverse of W_g, and $v_\beta(x) = E(X|\beta^\top X = \beta^\top x) - x$, while the asymptotic variance is given by

$$\Sigma_\beta = \left[E\{g'(\beta^\top X)^2 W(X)\}\right]^{+} E\left\{g'(\beta^\top X)^2 W_0(X)e^2\right\}\left[E\{g'(\beta^\top X)^2 W(X)\}\right]^{+},$$

where $W(x) = E(XX^\top|\beta^\top X = \beta^\top x) - E(X|\beta^\top X = \beta^\top x)E^\top(X|\beta^\top X = \beta^\top x)$ and $W_0(x) = v_\beta(x)v_\beta^\top(x)$.

From the above discussion, we know that the MAVE estimate of β is obtained by solving the minimization problem (2.80). Generally, to select the relevant variables, we can add a penalty term to the least-squares-form loss function in (2.80):

$$\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i - a_j - b_j\tilde{\beta}^\top(X_i - X_j)]^2 w_{ij} + n\sum_{l=1}^{k}p_{\lambda_n}(|\tilde{\beta}_l|),$$

where $p_\lambda(\cdot)$ denotes a penalty function and λ_n denotes the penalty parameter. Different choices of $p_\lambda(\cdot)$ can lead to different variable selection methods.

The simplest choice is to set $p_{\lambda_n}(|\tilde{\beta}_l|) = \lambda_n|\tilde{\beta}_l|$, which corresponds to the well-known least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996). Indeed, Wang and Yin (2008) adopted this L_1 norm penalty and proposed the sparse MAVE method and Zeng et al. (2012) further explored the idea of combining MAVE and LASSO, and proposed the sim-LASSO method. The sim-LASSO method not only penalizes the L_1 norm of the index parameter β, but also penalizes the terms $\{b_j\}_{j=1}^n$ in (2.80). Since $b_j = \partial g(u)/\partial u|_{u=\beta^\top X_j}$, adding this penalty contributes to excluding the data points with less information on estimating β, which stabilizes and improves the estimation of β. Finally, Wang et al. (2013) proposed the penalized MAVE method, combining the bridge regression with MAVE. In the case of the single-index-model, the penalized MAVE estimator has the oracle property.

It is widely accepted that a good penalty function should lead to an unbiased, sparse and continuous estimator. However, the LASSO estimator is biased for large parameters.

Alternatively, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty. The SCAD penalty is defined via its first derivative as

$$p'_\lambda(\beta_l) = \lambda\{I(\beta_l \leq \lambda) + \frac{(a\lambda - \beta_l)_+}{(a-1)\lambda} I(\tilde{\beta}_l > \lambda)\}.$$

Due to the oracle property of the SCAD penalty justified by Fan and Li (2001), Peng and Huang (2011) explored the idea of introducing the SCAD penalty into the single index model. Given that the dimension of β is a fixed constant, the SCAD estimator has the oracle property. Hence, we can also combine the SCAD penalty with MAVE, and modify the objective function in (2.80) as:

$$\beta = \arg \min_{\substack{\tilde{\beta} \in \mathbb{R}^k : \tilde{\beta}^\top \tilde{\beta} = 1 \\ a_j, b_j}} \left\{ \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - a_j - b_j \tilde{\beta}^\top (X_i - X_j) \right]^2 w_{ij} + n \sum_{l=1}^k p_{\lambda_n}^{\mathrm{SCAD}}(|\tilde{\beta}_l|) \right\} \tag{2.82}$$

Similarly, the optimization problem in (2.82) can be solved alternatively and iteratively and the SCAD-MAVE algorithm can be summarized as follows:

Step 1. Given data $\{Y_j, X_j\}_{j=1}^n$, calculate the initial estimator $\widehat{\beta}_{(0)}$ by the MAVE method. Set $\ell = 1$.

Step 2. Given $\widehat{\beta}_{(\ell-1)}$, calculate the refined weights as

$$w_{ij}^{(\ell-1)} = K_{h_1}\left[\widehat{\beta}_{(\ell-1)}^\top (X_i - X_j)\right] \Big/ \sum_{l=1}^n K_{h_1}\left[\widehat{\beta}_{(\ell-1)}^\top (X_l - X_j)\right].$$

Then, solve the inner optimization problem for $j = 1, \ldots, n$:

$$\min_{a_j, b_j} \sum_{i=1}^n \left[Y_i - a_j - b_j \widehat{\beta}_{(\ell-1)}^\top (X_i - X_j) \right]^2 w_{ij}^{(\ell-1)}$$

Clearly, this problem is analogous to the weighted least squares problem. We can easily derive the analytical solutions and denote them as $\widehat{a}_j^{(\ell-1)}$ and $\widehat{b}_j^{(\ell-1)}$.

Step 3. Given $\widehat{a}_j^{(\ell-1)}$ and $\widehat{b}_j^{(\ell-1)}$, we solve the outer optimization problem:

$$\min_{\tilde{\beta} \in \mathbb{R}^k : \tilde{\beta}^\top \tilde{\beta} = 1} \left\{ \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - \widehat{a}_j^{(\ell-1)} - \widehat{b}_j^{(\ell-1)} \tilde{\beta}^\top (X_i - X_j) \right]^2 w_{ij}^{(\ell-1)} + n \sum_{l=1}^k p_{\lambda_n}^{\mathrm{SCAD}}(|\tilde{\beta}_l|) \right\}$$

Obviously, regardless of the constraint $\tilde{\beta}^\top \tilde{\beta} = 1$, we can rewrite the first part in least squares form, then we can use the *ncvreg* package in R to optimize it and obtain the estimator $\widehat{\beta}_{(\ell)}$. Let $\widehat{\beta}_{(\ell)} = \mathrm{sign}(\widehat{\beta}_{(\ell)})\widehat{\beta}_{(\ell)} / \|\widehat{\beta}_{(\ell)}\|$.

Step 4. Check whether $\|\widehat{\beta}_{(\ell)}-\widehat{\beta}_{(\ell-1)}\|^2 < c$, where c is an arbitrarily small positive constant, if not, set $\ell = \ell + 1$ and go to Step 2. Denote the final estimator as $\widehat{\beta}_{\text{scad-MAVE}}$.

Based on the above discussion, we can use the SCAD-MAVE method to select relevant variables or control units at first, then, construct the index $\widehat{Z}_j = \widehat{\beta}_{\text{scad-MAVE}}^\top X_j$ for $j = 1, \ldots, n$. Finally, from Peng and Huang (2011), one can show that under some regularity conditions, $\widehat{\beta}_{\text{scad-MAVE}}$ satisfies (2.81) under some regularity conditions.

2.6.3 Functional Coefficient Index Models

Fan, Yao and Cai (2003) proposed the functional coefficient index model for the iid data, while Cai, Juhl and Yang (2015) investigated the case for time series data and further considered variable selection based on functional coefficient index models

$$E\left(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t\right) = \beta(\gamma^\top \mathbf{Z}_t)^\top \mathbf{X}_t, \tag{2.83}$$

where $\beta(\cdot)$ is an unknown coefficient function, which was used to a financial application by Cai, Ren and Yang (2015); see (2.61) for details. Add more

2.6.4 Distributional Index Models

In this section, we study the following distributional index model

$$F_y(y|\mathbf{X}_t) = F_y(y|\beta^\top \mathbf{X}_t), \tag{2.84}$$

where $F_g(y|\mathbf{X}_t)$ is the conditional distribution of Y_t given $\mathbf{X}_t$, which is called the distributional index model. Indeed, the model in (2.84) was successfully applied to a financial application by Aït-Shahalia and Brant (2001). Also, the model in (2.84) implies that the conditional quantile of Y_t given $\mathbf{X}_t$ is for any $0 < \tau < 1$,

$$q_\tau(\mathbf{X}_t) = F_y^{-1}(\tau|\mathbf{X}_t) = q_\tau(\beta^\top \mathbf{X}_t), \tag{2.85}$$

which is a special case of index quantile regression model. Also, $F_y(y|\mathbf{x})$ and $q_\tau(\mathbf{x})$ have the following relationship

$$F_y(y|\mathbf{x}) = \int_0^1 I(q_\tau(\mathbf{x}) \le y)d\tau,$$

which can be used to approximate $F_y(y|\mathbf{x})$ if $q_\tau(\mathbf{x})$ is estimable, as follows,

$$F_y(y|\mathbf{x}) \approx \eta + \int_\eta^{1-\eta} I(q_\tau(\mathbf{x}) \le y)d\tau \approx \eta + \sum_{j=2}^{S} (\tau_j - \tau_{j-1}) I(q_{\tau_j}(\mathbf{x}) \le y), \tag{2.86}$$

where the trimming by η (a very small number) avoids estimation of tail quantiles and an equally spaced mesh $\eta = \tau_1 < \ldots < \tau_S = 1 - \eta$ is used for a very large S so that $\tau_j - \tau_{j-1}$ is very small. One of the popular applications is to assume that $q_\tau(\mathbf{x})$ is a linear model as $q_\tau(\mathbf{x}) = \beta_\tau^\top \mathbf{x}$, which provides an easy estimation of β_τ. The idea in (2.86) was used by Chernozhukov, Fernández-Val and Melly (2013) and Cai et al. (2023, 2024) in estimating counterfactual distributions, which can be used for estimating quantile treatment effects (QTE); see, for instance, the papers by Cai et al. (2023, 2024) for detailed the methodology and its theory with applications.

Alternatively, one can use the method proposed by Hall and Yao (2005) by using matching approach. Add more

2.7 Time-Varying Coefficient Models

Now, consider the popular model as follows

$$Y_t = \beta(t)^\top \mathbf{X}_t + u_t, \tag{2.87}$$

where $\beta(t)$ is a vector of unknown functionals, which can be regarded as a special case of model (2.9) with $U_t = t$. Add more

2.8 Computer Codes

2.8.1 Codes fro Example 2.1

```
 # 12-03-2024
 graphics.off() # clean the previous graphs on the screen

 ###############
 # Example 2.1
 ##############

##########################################################################
 z1=read.table(file="/NP_lecture_note/data/ex4-1.txt")
 # dada: weekly 3-month Treasury bill from 1970 to 1997
```

```
x=z1[,4]/100
n=length(x)
y=diff(x)           # Delta x_t=x_t-x_{t-1}
x=x[1:(n-1)]
n=n-1
x_star=(x-mean(x))/sqrt(var(x))
z=seq(min(x),max(x),length=50)

win.graph()
#postscript(file="/NP_lecture_note/figs/fig-4.1.eps",
# horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light blue")
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(a) y(t) vs x(t)",col.main="red")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(b) |y(t)| vs x(t)",col.main="red")
scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(c) y(t)^2 vs x(t)",col.main="red")
#dev.off()
#######################################################################

#########################
# Nonparametric Fitting #
#########################

#########################################################
# Define the Epanechnikov kernel function
kernel<-function(x){0.75*(1-x^2)*(abs(x)<=1)}

###############################################################
# Define the kernel density estimator
kernden=function(x,z,h,ker){
 # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
 nz<-length(z)
 nx<-length(x)
 x0=rep(1,nx*nz)
 dim(x0)=c(nx,nz)
 x1=t(x0)
 x0=x*x0
 x1=z*x1
 x0=x0-t(x1)
 if(ker==1){x1=kernel(x0/h)}        # Epanechnikov kernel
 if(ker==0){x1=dnorm(x0/h)}         # normal kernel
 f1=apply(x1,2,mean)/h
 return(f1)
}
###############################################################
# Define the local constant estimator
local.constant=function(y,x,z,h,ker){
 # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
 nz<-length(z)
 nx<-length(x)
 x0=rep(1,nx*nz)
 dim(x0)=c(nx,nz)
 x1=t(x0)
 x0=x*x0
 x1=z*x1
 x0=x0-t(x1)
 if(ker==1){x1=kernel(x0/h)}        # Epanechnikov kernel
 if(ker==0){x1=dnorm(x0/h)}         # normal kernel
 x2=y*x1
 f1=apply(x1,2,mean)
 f2=apply(x2,2,mean)
 f3=f2/f1
 return(f3)
}

####################################################################
# Define the local linear estimator
local.linear<-function(y,x,z,h){
 # parameters: y=response, x=design matrix; h=bandwidth; z=grid point
 nz<-length(z)
 ny<-length(y)
 beta<-rep(0,nz*2)
 dim(beta)<-c(nz,2)
 for(k in 1:nz){
  x0=x-z[k]
  w0<-kernel(x0/h)
  beta[k,]<-glm(y~x0,weight=w0)$coeff
 }
 return(beta)
}
##############################################################

h=0.02

# Local constant estimate

mu_hat=local.constant(y,x,z,h,1)
sigma_hat=local.constant(abs(y),x,z,h,1)
sigma2_hat=local.constant(y^2,x,z,h,1)

#win.graph()
postscript(file="/NP_lecture_note/figs/fig-2.1.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light yellow")
```

```
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)")
points(z,mu_hat,type="l",lty=1,lwd=3,col=2)
title(main="(a) y(t) vs x(t)",col.main="red")
legend(0.04,0.0175,"Local Constant Estimate")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)")
points(z,sigma_hat,type="l",lty=1,lwd=3,col=2)
title(main="(b) |y(t)| vs x(t)",col.main="red")
scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)")
title(main="(c) y(t)^2 vs x(t)",col.main="red")
points(z,sigma2_hat,type="l",lty=1,lwd=3,col=2)
dev.off()

# Local Linear Estimate

fit2=local.linear(y,x,z,h)
mu_hat=fit2[,1]
fit2=local.linear(abs(y),x,z,h)
sigma_hat=fit2[,1]
fit2=local.linear(y^2,x,z,h)
sigma2_hat=fit2[,1]

#win.graph()
postscript(file="/NP_lecture_note/figs/fig-2.2.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light green")
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)")
points(z,mu_hat,type="l",lty=1,lwd=3,col=2)
title(main="(a) y(t) vs x(t)",col.main="red")
legend(0.04,0.0175,"Local Linear Estimate")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)")
points(z,sigma_hat,type="l",lty=1,lwd=3,col=2)
title(main="(b) |y(t)| vs x(t)",col.main="red")
```

```
 scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)")
 title(main="(c) y(t)^2 vs x(t)",col.main="red")
 points(z,sigma2_hat,type="l",lty=1,lwd=3,col=2)
 dev.off()
 ######################################################################
```

2.8.2 Codes for Additive Modeling Analysis of Boston Data

The following is the R code for making figures in Figures 2.4 and 2.5, respectively.

```
data=read.table("file="/NP_lecture_note/data/ex4-2.txt")
y=data[,14]
x1=data[,1]
x6=data[,6]
x10=data[,10]
x11=data[,11]
x13=data[,13]
y_log=log(y)
library(gam)
fit_gam=gam(y_log~lo(x6)+lo(x10)+lo(x11)+lo(x13))
resid=fit_gam$residuals
y_hat=fit_gam$fitted

postscript(file="/NP_lecture_note/figs/fig-2.3.eps",
horizontal=F,width=6,height=6,bg="light grey")
par(mfrow=c(2,2),mex=0.4)
plot(fit_gam)
title(main="Component of X_13",col.main="red",cex=0.6)
dev.off()
fit_gam1=gam(y_log~lo(x6)+x10+x11+x13)
s1=fit_gam1$smooth[,1]           # obtain the smoothed component
resid1=fit_gam1$residuals
y_hat1=fit_gam1$fitted
```

```
print(summary(fit_gam1))

postscript(file="/NP_lecture_note/figs/fig-2.4.eps",
horizontal=F,width=6,height=6,bg="light green")
par(mfrow=c(2,2),mex=0.4)
plot(y_hat,resid,type="p",pch="o",ylab="",xlab="y_hat")
title(main="Residual Plot of Additive Model",col.main="red",cex=0.6)
abline(0,0)
plot(x6,s1,type="p",pch="o",ylab="s1(x6)",xlab="x6")
title(main="Component of X_6",col.main="red",cex=0.6)
plot(y_hat1,resid1,type="p",pch="o",ylab="",xlab="y_hat")
title(main="Residual Plot of Model II",col.main="red",cex=0.5)
abline(0,0)
plot(density(y),ylab="",xlab="",main="Density of Y")
dev.off()
```

2.9 References

Aït-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica*, **64**(3), 527-560.

Aït-Sahalia, Y. and Brant, M.W. (2001). Variable selection for portfolio choice. *Journal of Finance*, **56**(4), 1297-1351.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostic: Identifying Influential Data and Sources of Collinearity.* Wiley & Sons, New York.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformation for multiple regression and correlation. *Journal of the American statistical Association*, **80**(391), 580-598.

Cai, Z. (2002). A two-stage approach to additive time series models. *Statistica Neerlandica*, **56**(4), 415-433.

Cai, Z. (2007). Trending time varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, **136**(1), 163-188.

Cai, Z. (2010). Functional coefficient models for economic and financial data. In *Oxford Handbook of Functional Data Analysis (Eds: F. Ferraty and Y. Romain).* Oxford University Press, Oxford, UK, pp.166-186.

Cai, Z., Das, M., Xiong, H. and Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, **133**(1), 207â 241.

Cai, Z. and Fan, J. (2000). Average regression surface for dependent data. *Journal of Multivariate Analysis*, **75**(1), 112- 142.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of American Statistical Association*, **95**(451), 941-956.

Cai, Z., Fang, Y., Lin, M. and Zhan, M. (2023). Estimating quantile treatment effects for panel data. Forthcoming in *Scientia Sinica Mathematica*.

Cai, Z., Fang, Y., Lin, M. and Wu, Y. (2024). Estimating counterfactual distribution functions via optimal distribution balancing with applications. https://econpapers.repec.org/paper/kanwpaper/202315.htm.

Cai, Z. and Hong, Y. (2009). Some recent developments in nonparametric finance. *Advances in Econometrics*, **25**, 379-432.

Cai, Z., Juhl, T. and Yang, B. (2015). Functional index coefficient models with variable selection. *Journal of Econometrics*, **189**(2), 272-284.

Cai, Z. and Masry, E. (2000). Nonparametric estimation of additive nonlinear ARX time series: Local linear fitting and projection. *Econometric Theory*, **16**(4), 465-501.

Cai, Z. and Tiwari, R.C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**(3), 341-350.

Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics: Methods and Applications.* Cambridge University Press, New York.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *Annuals of Statistics*, **25**(2), 715-744.

Chen, R. and Tsay, R.S. (1993). Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**(423), 310-320.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, **81**(6), 2205-2268.

Engle, R.F., Granger, C.W., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, **81**(394), 310-320.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics*, **21**(1), 196-216.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, **49**(1), 79-99.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J., Heckman, N. E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**(429), 141-150.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**(6), 1031-1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: Nonparametric and Parametric Methods.* Springer-Verlag, Berlin.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models.*Journal of the Royal Statistical Society: Series B*, **65**(1), 57-80.

Fan, J. and Zhang, C. (2003). A re-examination of diffusion estimators with applications to financial model validation. *Journal of the American Statistical Association*, **98**(461), 118-134.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of statistics*, **29**(1), 153-193.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics*, **757**, 23-28. Springer-Verlag, New York.

Gilley, O. W., Pace, R.K., et al. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, **31**(3), 403-405.

Granger, C. and Teräsvirta, T. (1993). *Modeling Nonlinear Economic Relationships.* Oxford University Press, Oxford, UK.

Hall, P. and Heyde, C.C. (1980). *Martingale limit theory and its application.* Academic press, New York.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society: Series B*, **54**(2), 475-509.

Hall, P., J.S. Racine and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, **99**(468), 1015-1026.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics*, **33**(3), 1404-1421.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**(1), 157-178.

Harrison Jr, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**(1), 81-102.

Hastie, T. and Tibshirani, R. (1990).*Generalized additive models.* Chapman and Hall, London.

Hong, Y. and Lee, T.-H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, **85**(4), 1048-1062.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**(4), 73-101.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, **60**(2), 271-293.

Ichimuara, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

Jiang, G.J. and Knight, J.L. (1997). A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory*, **13**(5), 615-645.

Johannes, M.S. (2004). The statistical and economic role of jumps in continuous-time interest rate models. *Journal of Finance*, **59**(1), 227-260.

Juhl, T. (2005). Functional coefficient models under unit root behavior. *The Econometrics Journal*, **8**(2), 197-213.

Klein, R.W. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, **61**(2), 387-421.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, New York.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, **46**(1), 33-50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**(1), 43-61.

Kreiss, J.-P., Neumann, M.H. and Yao, Q. (1998). Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and Its Interface*, **1**(2), 367-380.

Li, Q., Huang, C.J., Li, D. and Fu, T.-T. (2002). Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics*, **20**(3), 412-422.

Li, Q. and J.S. Racine (2003). Nonparametric estimation of distributions with both categorical and continuous data. *Journal of Multivariate Analysis*, **86**(2), 266-292. MR1997765

Li, Q. and R.S. Racine (2007). *Nonparametric Econometrics, Theory and Practice.* Princeton University Press, New York.

Linton, O. B. (1997). Miscellanea efficient estimation of additive nonparametric regression models. *Biometrika*, **84**(2), 469-473.

Linton, O.B. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, **16**(4), 502-523.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**(1), 93-100.

Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**(5), 1443-1490.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, **24**(2), 165-179.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**(2), 214-252.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**(1), 141-142.

Ø ksendal, B. (1985). *Stochastic Differential Equations: An Introduction with Applications*, 3rd edition. Springer-Verlag, New York.

Opsomer, J.D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, **93**(442), 605-619.

Pace, R.K. (1993). Nonparametric methods with applications to hedonic models. *Journal of Real Estate Finance and Economics*, **7**(3), 185-204.

Pace, R.K. and Gilley, O.W. (1997). Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics*, **14**(3), 333-340.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4), 1362-1379.

Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, **57**, 1403-1430.

Priestley, M.B. and Chao, M. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society: Series B*, **34**(3), 385-392.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, **12**(4), 1215-1230.

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, **56**(4), 931–954.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**(432), 1257-1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**(4), 1346-1370.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection*. John Wiley & sons, New York.

S entürk, D. and Müller, H.-G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Annals of Statistics*, **34**(2), 654-679.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**(422), 486-494.

Shao, Q.-M. and Yu, H. (1996). Weak convergence for weighted empirical processes of dependent sequences. *Annals of Probability*, **24**(4), 2098-2127.

Shiller R.J. (1984). Smoothness priors and nonlinear regression. *Journal of the American Statistical Association*, **72**(376), 420-423.

Speckman, P. (1988). Kernel smoothing partial linear models. *Journal of Royal Statistical Society, Series B*, **50**(3), 413–426.

Sperlich, S., Tjøstheim, D. and Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, **18**(2), 197-251.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *The Journal of Finance*, **52**(5), 1973-2002.

Sun, Z. (1984). Asymptotic unbiased and strong consistency for density function estimator. *Acta Mathematica Sinica*, **27**(4), 769-782.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1), 267-288.

Tjøstheim, D. and Auestad, B.H. (1994a). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, **89**(428), 1398-1409.

Tjøstheim, D. and Auestad, B.H. (1994b). Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association*, **89**(428), 1410-1419.

van Dijk, D., T. Teräsvirta and P.H. Franses (2002). Smooth transition autoregressive models - A survey of recent developments. *Econometric Reviews*, **21**(1), 1-47.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis*, 52(9), 4512-4520.

Wang, T., Xu, P. and Zhu, L. (2013). Penalized minimum average variance estimation. *Statistica Sinica*, **23**(2), 543-569.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā*, **26**(4), 359-372.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6), 1112-1137.

Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B*, 64(2), 363-410.

Zeng, P., He, T. and Zhu, Y. (2012). A LASSO-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, 21(1), 92-109.

Zhu, F., Liu, M., Ling, S. and Cai, Z. (2023). Testing for structural change of predictive regression model to threshold predictive regression model. *Journal of Business & Economic Statistics*, 41(1), 228-240.

Chapter 3

Nonparametric Quantile Models

For details, see the papers by Cai and Xu (2008) and Cai and Xiao (2012). Next we present only a part of the whole paper of Cai and Xu (2008).

3.1 Introduction

Over the last three decades, quantile regression, also called conditional quantile or regression quantile, introduced by Koenker and Bassett (1978), has been used widely in various disciplines, such as finance, economics, medicine, and biology. It is well-known that when the distribution of data is typically skewed or data contains some outliers, the median regression, a special case of quantile regression, is more explicable and robust than the mean regression. Also, regression quantiles can be used to test heteroscedasticity formally or graphically (Koenker and Bassett, 1982; Efron, 1991; Koenker and Zhao, 1996; Koenker and Xiao, 2002). Although some individual quantiles, such as the conditional median, are sometimes of interest in practice, more often one wishes to obtain a collection of conditional quantiles which can characterize the entire conditional distribution. More importantly, another application of conditional quantiles is the construction of prediction intervals for the next value given a small section of the recent past values in a stationary time series (Granger, White, and Kamstra, 1989; Koenker, 1994; Zhou and Portnoy, 1996; Koenker and Zhao, 1996; Taylor and Bunn, 1999). Also, Granger, White, and Kamstra (1989), Koenker and Zhao (1996), and Taylor and Bunn (1999) considered an interval forecasting for parametric autoregressive conditional heteroscedastic (ARCH) type models. For more details about the historical and recent developments of quantile regression with applications for time series data, particularly in finance, see, for example, the papers and books by J.P. Morgan (1995), Duffie and Pan

(1997), Jorin (2000), Koenker (2000), Koenker and Hallock (2001), Tsay (2000, 2002), Khindanova and Rachev (2000), and Bao, Lee and Saltog lu (2006), and the references therein.

Recently, the quantile regression technique has been successfully applied to politics. For example, in the 1992 presidential selection, the Democrats used the yearly Current Population Survey data to show that between 1980 and 1992 there was an increase in the number of people in the high-salary category as well as an increase in the number of people in the low-salary category. This phenomena could be illustrated by using the quantile regression method as follows: computing 90% and 10% quantile regression functions of salary as a function of time. An increasing 90% quantile regression function and a decreasing 10% quantile regression function corresponded to the Democrats' claim that ”the rich got richer and the poor got poorer” during the Republican administrations; see Figure 6.4 in Fan and Gijbels (1996, p. 229).

More importantly, by following the regulations of the Bank for International Settlements, many of financial institutions have begun to use a uniform measure of risk to measure the market risks called Value-at-Risk (VaR), which can be defined as the maximum potential loss of a specific portfolio for a given horizon in finance. In essence, the interest is to compute an estimate of the lower tail quantile (with a small probability) of future portfolio returns, conditional on current information. Therefore, the VaR can be regarded as a special application of the quantile regression. There is a vast amount of literature in this area; see, to name just a few, J.P. Morgan (1995), Duffie and Pan (1997), Engle and Manganelli (2004), Jorion (2000), Tsay (2000, 2002), Khindanova and Rachev (2000), and Bao, Lee and Saltog lu (2006), and references therein.

In this chapter, we assume that $\{\mathbf{X}_t, Y_t\}_{t=-\infty}^{\infty}$ is a stationary sequence. Denote $F(y \mid \mathbf{x})$ the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}_t = (X_{t1}, \ldots, X_{td})'$ with $'$ denoting the transpose of a matrix or vector, is the associated covariate vector in $\Re^d$ with $d \geq 1$, which might be a function of exogenous (covariate) variables or some lagged (endogenous) variables or time t. The regression (conditional) quantile function $q_\tau(\mathbf{x})$ is defined as, for any $0 < \tau < 1$,

$$q_\tau(\mathbf{x}) = \inf\left\{y \in \Re^1 : F(y \mid \mathbf{x}) \geq \tau\right\}, \text{ or } q_\tau(\mathbf{x}) = \text{argmin}_{a \in \Re^1} E\left\{\rho_\tau\left(Y_t - a\right) \mid \mathbf{X}_t = \mathbf{x}\right\}, \tag{3.1}$$

where $\rho_\tau(y) = y\left(\tau - I_{\{y<0\}}\right)$ with $y \in \Re^1$ is called the loss (“check”) function, and I_A is the indicator function of any set A. There are several advantages of using a quantile regression:

- A quantile regression does not require knowing the distribution of the dependent variable.

- It does not require the symmetry of the measurement error.

- It can characterize the heterogeneity.

- It can estimate the mean and variance simultaneously.

- It is a robust procedure.

- There are a lot more.

Having conditioned on the observed characteristics $\mathbf{X}_t = \mathbf{x}$, based on the Skorohod representation[1], Y_t and the quantile function $q_\tau(\mathbf{x})$ have a following relationship as

$$Y_t = q\left(\mathbf{X}_t, U_t\right), \tag{3.2}$$

where $U_t \mid \mathbf{X}_t \sim U(0,1)$. We will refer to U_t as the rank variable, and note that representation (3.2) is essential to what follows. The rank variable U_t is responsible for heterogeneity of outcomes among individuals with the same observed characteristics $\mathbf{X}_t$. It also determines their relative ranking in terms of potential outcomes; hence one may think of rank U_t as representing some unobserved characteristic. This interpretation makes quantile analysis an interesting tool for describing and learning the structure of heterogeneous effects and controlling for unobserved heterogeneity.

Clearly, the simplest form of model (3.1) is $q_\tau(\mathbf{x}) = \boldsymbol{\beta}_\tau' \mathbf{x}$, which is called the linear quantile regression model well studied by many authors. For details, see the papers by Duffie and Pan (1997), Koenker (2000), Tsay (2002), Koenker and Hallock (2001), Khindanova and Rachev (2000), and Bao, Lee and Saltog lu (2006), Engle and Manganelli (2004), and references therein.

In many practical applications, however, the linear quantile regression model might not be "rich" enough to capture the underlying relationship between the quantile of response variable and its covariates. Indeed, some components may be highly nonlinear or some covariates may be interactive. To make the quantile regression model more flexible, there is a swiftly growing literature on nonparametric quantile regression. Various smoothing

[1] For the definition, please see the book by Durret (2019).

techniques, such as kernel methods, splines, and their variants, have been used to estimate the nonparametric quantile regression for both the independent and time series data. For the recent developments and the detailed discussions on theory, methodologies, and applications, see, for example, the papers by He, Ng, and Portony (1998), Yu and Jones (1998), He and Ng (1999), He and Portony (2000), Honda (2000, 2004), Tsay (2000, 2002), Lu, Hui and Zhao (2000), Khindanova and Rachev (2000), Bao, Lee and Saltog lu (2006), Cai (2002a), De Gooijer, and Gannoun (2003), Horowitz and Lee (2005), Yu and Lu (2004), and Li and Racine (2008), and references therein. In particular, for the univariate case, recently, Honda (2000) and Lu, Hui and Zhao (2000) derived the asymptotic properties of the local linear estimator of the quantile regression function under α-mixing condition. For the high dimensional case, however, the aforementioned methods encounter some difficulties such as the so-called "curse of dimensionality" and their implementation in practice is not easy as well as the visual display is not so useful for the exploratory purposes.

To attenuate the above problems, De Gooijer and Zerom (2003), Horowitz and Lee (2005), and Yu and Lu (2004) considered an additive quantile regression model $q_\tau\left(\mathbf{X}_t\right) = \sum_{k=1}^{d} g_k\left(X_{tk}\right)$. To estimate each component, for the time series case, De Gooijer and Zerom (2003) first estimated a high dimensional quantile function by inverting the conditional distribution function estimated by using a weighted Nadaraya-Watson approach, proposed by Cai (2002a), and then used a projection method to estimate each component, as discussed in Cai and Masry (2000), while Yu and Lu (2004) focused on the independent data and used a back-fitting algorithm method to estimate each component. On the other hand, to estimate each additive component for the independent data, Horowitz and Lee (2005) used a two-stage approach consisting of the series estimation at the first step and a local polynomial fitting at the second step. For the independent data, the above model was extended by He, Ng and Portony (1998), He and Ng (1999), and He and Portony (2000) to include interaction terms by using spline methods.

In this chapter, we adapt another dimension reduction modeling method to analyze dynamic time series data, termed as the smooth (functional or varying) coefficient modeling approach. This approach allows appreciable flexibility on the structure of fitted models. It allows for linearity in some continuous or discrete variables which can be exogenous or lagged and nonlinear in other variables in the coefficients. In such a way, the model has the ability of capturing the individual variations. More importantly, it can ease the so-called "curse of dimensionality" and combines both additivity and interactivity. A smooth coefficient quantile regression model for time series data takes the following form

$$q_\tau\left(\mathbf{U}_t, \mathbf{X}_t\right) = \sum_{k=0}^{d} a_k\left(\mathbf{U}_t\right) X_{tk} = \mathbf{X}_t' \mathbf{a}_\tau\left(\mathbf{U}_t\right), \tag{3.3}$$

where $\mathbf{U}_t$ is called the smoothing variable, which might be one part of $X_{t1}, \ldots, X_{td}$ or just time or other exogenous variables or the lagged variables, $\mathbf{X}_t = (X_{t0}, X_{t1}, \ldots, X_{td})'$ with $X_{t0} \equiv 1, \{a_k(\cdot)\}$ are smooth coefficient functions, and $\mathbf{a}_\tau(\cdot) = (a_{0,\tau}(\cdot), \ldots, a_{d,\tau}(\cdot))'$. Here, some of $\{a_{k,\tau}(\cdot)\}$ are allowed to depend on τ. For simplicity, we drop τ from $\{a_{k,\tau}(\cdot)\}$ in what follows. It is our interest here to estimate the coefficient functions $\mathbf{a}(\cdot)$ rather than the quantile regression surface $q_\tau(\cdot, \cdot)$ itself. Note that model (3.3) was studied by Honda (2004) for the independent sample, but our focus here is on the dynamic model for nonlinear time series, which is more appropriate for economic and financial applications.

The general setting in (3.3) covers many familiar quantile regression models, including the quantile autoregressive model (QAR) proposed by Koenker and Xiao (2004) who applied the QAR model for the unit root inference. In particular, it includes a specific class of ARCH models, such as heteroscedastic linear models considered by Koenker and Zhao (1996). Also, if there is no $\mathbf{X}_t$ in the model ($d = 0$), $q_\tau\left(\mathbf{U}_t, \mathbf{X}_t\right)$ becomes $q_\tau\left(\mathbf{U}_t\right)$ so that model (3.3) reduces to the ordinary nonparametric quantile regression model which has been studied extensively. For the recent developments, refer to the papers by He, Ng and Portony (1998), Yu and Jones (1998), He and Ng (1999), He and Portony (2000), Honda (2000), Lu, Hui and Zhao (2000), Cai (2002a), De Gooijer and Zerom (2003), Horowitz and Lee (2005), Yu and Lu (2004), and Li and Racine (2008). If $\mathbf{U}_t$ is just time, then the model is called the timevarying coefficient quantile regression model, which is potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the aforementioned illustrative example for the 1992 presidential election and the analysis of the reference growth data by Cole (1994), Wei, Pere, Koenker and He (2006), and Wei and He (2006), and the references therein. However, if $\mathbf{U}_t$ is time, the observed time series might not be stationary. Therefore, the treatment for non-stationary case would require a different approach so that it is beyond the scope of this chapter and deserves a further investigation. For more applications, see the work in Xu (2005). Finally, note that the smooth coefficient mean regression model is one of the most popular nonlinear time series models in mean regression and has various applications. For more discussions, refer to the papers by Chen and Tsay (1993), Cai, Fan, and Yao (2000), Cai and Tiwari (2000), Cai (2007), Hong and Lee (2003), and Wang (2003), and the book by Tsay (2002), and references therein.

The motivation of this study comes from an analysis of the well known Boston housing price data, consisting of several variables collected on each of 506 different houses from a variety of locations. The interest is to identify the factors affecting the house price in Boston area. As argued by Sentürk and Müller (2006), the correlation between the house price and the crime rate can be adjusted by the confounding variable which is the proportion of population of lower educational status through a varying coefficient model and the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. The interesting features of this dataset are that the response variable is the median price of a home in a given area and the distributions of the price and the major covariate (the confounding variable) are left skewed. Therefore, quantile methods are suitable for the analysis of this dataset. Therefore, such a problem can be tackled by using model (3.3). In another example, one is interested in exploring the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rates such as the Japanese Yen per US dollar. The detailed analysis of these data sets is reported in Section 3.3.

3.2 Modeling Procedures

3.2.1 Local Linear Quantile Estimate

Now, we apply the local polynomial method to the smooth coefficient quantile regression model as follows. For the sake of brevity, we only consider the case where $\mathbf{U}_t$ in (3.3) is one-dimensional, denoted by U_t in what follows. Extension to multivariate $\mathbf{U}_t$ involves fundamentally no new ideas although the theory and procedure continue to hold. Note that the models with high dimension might not be practically useful due to the curse of dimensionality. A local polynomial fitting has several nice properties such as high statistical efficiency in an asymptotic minimax sense, design-adaptation, and automatic edge correction (see, e.g., Fan and Gijbels, 1996).

We estimate the functions $\{a_k(\cdot)\}$ using the local polynomial regression method from observations $\{(U_t, \mathbf{X}_t, Y_t)\}_{t=1}^n$. We assume throughout the chapter that the coefficient functions $\mathbf{a}(\cdot)\}$ have the $(q+1)^{th}$ derivative, so that for any given gird point $u_0, a_k(\cdot)$ can be approximated by a polynomial function in a neighborhood of the given grid point u_0 as $\mathbf{a}\left(U_t\right) \approx \mathbf{a}\left(u_0\right) + \mathbf{a}'\left(u_0\right)\left(U_t - u_0\right) + \cdots + \mathbf{a}^{(q)}\left(u_0\right)\left(U_t - u_0\right)^q / q!$ and

$$q_\tau\left(U_t, \mathbf{X}_t\right) \approx \sum_{j=0}^{q} \mathbf{X}_t' \boldsymbol{\beta}_j \left(U_t - u_0\right)^j,$$

where $\boldsymbol{\beta}_j = \mathbf{a}^{(j)}\left(u_0\right) / j\ !$. Then, the locally weighted loss function is

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=0}^{q} \mathbf{X}_t' \boldsymbol{\beta}_j \left(U_t - u_0\right)^j \right) K_h\left(U_t - u_0\right), \tag{3.4}$$

where $K(\cdot)$ is a kernel function, $K_h(x) = K(x/h)/h$, and $h = h_n$ is a sequence of positive numbers tending to zero, which controls the amount of smoothing used in estimation. Solving the minimization problem in (3.4) gives $\widehat{\mathbf{a}}\left(u_0\right) = \widehat{\boldsymbol{\beta}}_0$, the local polynomial estimate of $\mathbf{a}\left(u_0\right)$, and $\widehat{\mathbf{a}}^{(j)}\left(u_0\right) = j! \widehat{\boldsymbol{\beta}}_j (j \geq 1)$, the local polynomial estimate of the j^{th} derivative $\mathbf{a}^{(j)}\left(u_0\right)$ of $\mathbf{a}\left(u_0\right)$. By moving u_0 along with the real line, one obtains the estimate for the entire curve. For various practical applications, Fan and Gijbels (1996) recommended using the local linear fit ($q = 1$). Therefore, for the expositional purpose, in what follows, we only consider the case $q = 1$ (local linear fitting).

The programming involved in the local (polynomial) linear quantile estimation is relatively simple and can be modified with few efforts from the existing programs for a linear quantile model. For example, for each grid point u_0, the local linear quantile estimation can be implemented in the **R** package **quantreg**, of Koenker (2004) by setting covariates as $\mathbf{X}_t$ and $\mathbf{X}_t\left(U_t - u_0\right)$ and the weight as $K_h\left(U_t - u_0\right)$.

Although some modifications are needed, the method developed here for the local linear quantile estimation is applicable to a general local polynomial quantile estimation. In particular, we note that the local constant (Nadaraya-Watson type) quantile estimation of $\mathbf{a}\left(u_0\right)$, denoted by $\widetilde{\mathbf{a}}\left(u_0\right)$, is $\widetilde{\boldsymbol{\beta}}$ minimizing the following subjective function

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \mathbf{X}_t' \boldsymbol{\beta}\right) K_h\left(U_t - u_0\right), \tag{3.5}$$

which is a special case of (3.4) with $q = 0$. We compare $\widehat{\mathbf{a}}\left(u_0\right)$ and $\widetilde{\mathbf{a}}\left(u_0\right)$ theoretically at the end of Section **??** and empirically in Section 3.1 and the comparison leads to suggest that one should use the local linear approach in practice.

3.2.2 Asymptotic Results

We first give some regularity conditions that are sufficient for the consistency and asymptotic normality of the proposed estimators, although they might not be the weakest possible. We introduce the following notations. Denote

$$\Omega\left(u_0\right) \equiv E\left[\mathbf{X}_t \mathbf{X}_t^{\prime} \mid \mathbf{U}_t=u_0\right] \quad \text{and} \quad \Omega^*\left(u_0\right) \equiv E\left[\mathbf{X}_t \mathbf{X}_t^{\prime} f_{y \mid u, x}\left(q_\tau\left(u_0, \mathbf{X}_t\right)\right) \mid \mathbf{U}_t=u_0\right],$$

where $f_{y|u,x}(y)$ is the conditional density of Y given U and $\mathbf{X}$. Let $f_u(u)$ present the marginal density of U.

Assumptions:

(C1) $\mathbf{a}(u)$ is twice continuously differentiable in a neighborhood of u_0 for any u_0.

(C2) $f_u(u)$ is continuous and $f_u\left(u_0\right)>0$.

(C3) $f_{y|u,x}(y)$ is bounded and satisfies the Lipschitz condition.

(C4) The kernel function $K(\cdot)$ is symmetric and has a compact support, say $[-1,1]$.

(C5) $\{(\mathbf{X}_t, Y_t, \mathbf{U}_t)\}$ is a strictly α-mixing stationary process with mixing coefficient $\alpha(t)$ satisfies $\sum_{t\geq 1}^{\infty} t^l \alpha^{(\delta-2)/\delta}(t)<\infty$ for some positive real number $\delta>2$ and $l>(\delta-2)/\delta$.

(C6) $E\left\|\mathbf{X}_t\right\|^{2\delta^*}<\infty$ with $\delta^*>\delta$.

(C7) $\Omega\left(u_0\right)$ is positive-definite and continuous in a neighborhood of u_0

(C8) $\Omega^*\left(u_0\right)$ is continuous and positive-definite in a neighborhood of u_0.

(C9) The bandwidth h satisfies $h \rightarrow 0$ and $nh \rightarrow \infty$.

(C10) $f\left(u, v \mid \mathbf{x}_0, \mathbf{x}_s ; s\right) \leq M<\infty$ for $s \geq 1$, where $f\left(u, v \mid \mathbf{x}_0, \mathbf{x}_s ; s\right)$ is the conditional density of (U_0, U_s) given $(\mathbf{X}_0=\mathbf{x}_0, \mathbf{X}_s=\mathbf{x}_s)$.

(C11) $n^{1/2-\delta/4} h^{\delta/\delta^*-1/2-\delta/4}=O(1)$.

Remark 3.1: *(Discussion of Conditions) Assumptions (C1)-(C3) include some smoothness conditions on functionals involved. The requirement in (C4) that $K(\cdot)$ be compactly supported is imposed for the sake of brevity of proofs, and can be removed at the cost of lengthier* *arguments. In particular, the Gaussian kernel is allowed. The α-mixing is one of the weakest mixing conditions for weakly dependent stochastic processes. Stationary time series or Markov chains fulfilling certain (mild) conditions are α-mixing with exponentially decaying coefficients; see the discussions in Section 1.1 and Cai (2002a) for more examples. On the other hand, the assumption on the convergence rate of $\alpha(\cdot)$ in (C5) might not be the weakest possible and is imposed to simplify the proof. Further, (C10) is just a technical assumption, which is also imposed by Cai (2002a). (C6) - (C8) require some standard moments. Clearly, (C11) allows the choice of a wide range of smoothing parameter values and is slightly stronger than the usual condition of $nh \rightarrow \infty$. However, for the bandwidths of optimal size (i.e., $h = O\left(n^{-1/5}\right)$), (C11) is automatically satisfied for $\delta \geq 3$ and it is still fulfilled for $2 < \delta < 3$ if δ^* satisfies $\delta < \delta^* \leq 1 + 1/(3-\delta)$, so that we do not concern ourselves with such refinements. Indeed, this assumption is also imposed by Cai, Fan and Yao (2000) for the mean regression. Finally, if there is no $\mathbf{X}_t$ in model* (3.3)*, (C5) can be replaced by* $(\mathrm{C5})'$: $\alpha(t) = O\left(t^{-\delta}\right)$ *for some* $\delta > 2$ *and* (C11) *can be substituted by* $(\mathrm{C11})'$: $nh^{\delta/(\delta-2)} \rightarrow \infty$*; see Cai (2002a) for details.*

Remark 3.2: *(Identification) It is clear from* (3.3) *that*

$$\Omega\left(u_0\right)\mathbf{a}\left(u_0\right) = E\left[q_\tau\left(u_0, \mathbf{X}_t\right)\mathbf{X}_t \mid U_t = u_0\right].$$

Then, $\mathbf{a}\left(u_0\right)$ *is identified (uniquely determined) if and only if* $\Omega\left(u_0\right)$ *is positive definite for any* u_0*. Therefore, Assumption* (C7) *is the necessary and sufficient condition for the model identification.*

To establish the asymptotic normality of the proposed estimator, similar to Chaudhuri (1991), we first derive the local Bahadur representation for the local linear estimator. To this end, our analysis follows the approach of Koenker and Zhao (1996), which can simplify the theoretical proofs. Define, $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K^2(u)du$. Also, set $\psi_\tau(x) = \tau - I_{\{x<0\}}, U_{th} = \left(U_t - u_0\right)/h, \mathbf{X}_t^* = \begin{pmatrix} \mathbf{X}_t \\ U_{th}\mathbf{X}_t \end{pmatrix}, Y_t^* = Y_t - \mathbf{X}_t'\left[\mathbf{a}\left(u_0\right) + \mathbf{a}'\left(u_0\right)\left(U_t - u_0\right)\right]$, and $\boldsymbol{\theta} = \sqrt{nh}\mathbf{H}\begin{pmatrix} \boldsymbol{\beta}_0 - \mathbf{a}\left(u_0\right) \\ \boldsymbol{\beta}_1 - \mathbf{a}'\left(u_0\right) \end{pmatrix}$ with $\mathbf{H} = \mathrm{diag}\{\mathbf{I}, h\mathbf{I}\}$.

Theorem 3.1: *(Local Bahadur Representation) Under Assumptions (C1)- (C9), we have*

$$\widehat{\boldsymbol{\theta}} = \frac{\left[\Omega_1^*\left(u_0\right)\right]^{-1}}{\sqrt{nh}f_u\left(u_0\right)}\sum_{t=1}^{n}\psi_\tau\left(Y_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right) + o_p(1), \tag{3.6}$$

where $\Omega_1^*\left(u_0\right) = diag\left\{\Omega^*\left(u_0\right), \mu_2\Omega^*\left(u_0\right)\right\}$*.*

Remark 3.3: *From Theorem 3.1 and Lemma 3.1 (in Section 3.4), it is easy to see that the local linear estimator* $\widehat{\mathbf{a}}(u_0)$ *is consistent with the optimal nonparametric convergence rate* $\sqrt{nh}$

Theorem 3.2: *(Asymptotic Normality) Under Assumptions (C1)- (C11), we have the following asymptotic normality*

$$\sqrt{nh}\left[\mathbf{H}\begin{pmatrix}\widehat{\mathbf{a}}(u_0)-\mathbf{a}(u_0)\\ \widehat{\mathbf{a}'}(u_0)-\mathbf{a}'(u_0)\end{pmatrix}-\frac{h^2}{2}\begin{pmatrix}\mathbf{a}''(u_0)\mu_2\\ 0\end{pmatrix}+o_p\left(h^2\right)\right] \quad\to\quad N\left\{0,\mathbf{\Sigma}(u_0)\right\}$$

where $\mathbf{\Sigma}(u_0)=diag\{\tau(1-\tau)\nu_0\mathbf{\Sigma}_a(u_0),\tau(1-\tau)\nu_2\mathbf{\Sigma}_a(u_0)\}$ *with*

$$\mathbf{\Sigma}_a(u_0)=\left[\Omega^*(u_0)\right]^{-1}\Omega(u_0)\left[\Omega^*(u_0)\right]^{-1}/f_u(u_0) \tag{3.7}$$

In particular,

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(u_0)-\mathbf{a}(u_0)-\frac{h^2\mu_2}{2}\mathbf{a}''(u_0)+o_p\left(h^2\right)\right] \quad\to\quad N\left\{0,\tau(1-\tau)\nu_0\mathbf{\Sigma}_a(u_0)\right\}$$

Remark 3.4: *From Theorem 3.2, the asymptotic mean squares error (AMSE) of* $\widehat{\mathbf{a}}(u_0)$ *is given by*

$$AMSE=\frac{h^4\mu_2^2}{4}\left\|\mathbf{a}''(u_0)\right\|^2+\frac{\tau(1-\tau)\nu_0}{nhf_u(u_0)}tr\left(\mathbf{\Sigma}_a(u_0)\right),$$

which gives the optimal bandwidth h_{opt} *by minimizing the AMSE*

$$h_{opt}=\left(\frac{\tau(1-\tau)\nu_0 tr\left(\mathbf{\Sigma}_a(u_0)\right)}{f_u(u_0)\left\|\mathbf{a}''(u_0)\right\|^2}\right)^{1/5}n^{-1/5},$$

and the optimal AMSE is

$$AMSE_{opt}=\frac{5}{4}\left(\frac{\tau(1-\tau)\nu_0 tr\left(\mathbf{\Sigma}_a(u_0)\right)}{f_u(u_0)}\right)^{4/5}\left\|\mathbf{a}''(u_0)\right\|^{2/5}n^{-4/5}.$$

Further, notice that the similar results in Theorem 3.2 were obtained by Honda (2004) for the independent data. Finally, it is interesting to note that the asymptotic bias in Theorem 3.2 is the same as that for the mean regression case but the two asymptotic variances are different; see, for example, Cai, Fan and Yao (2000).

If model (3.3) does not have $\mathbf{X}(d=0)$, it becomes the nonparametric quantile regression model $q_\tau(\cdot)$. Then, we have the following asymptotic normality for the local linear estimator of the nonparametric quantile regression function $q_\tau(\cdot)$, which covers the results in Yu and Jones (1998), Honda (2000), Lu, Hui and Zhao (2000), and Cai (2002a) for both the independent and time series data.

Corollary 3.2.1: *If there is no* $\mathbf{X}_t$ *in* (3.3), *then,*

$$\sqrt{nh}\left[\widehat{q}_\tau\left(u_0\right)-q_\tau\left(u_0\right)-\frac{h^2\mu_2}{2}q_\tau''\left(u_0\right)+o_p\left(h^2\right)\right] \quad\rightarrow\quad N\left\{0,\sigma_\tau^2\left(u_0\right)\right\},$$

where $\sigma_\tau^2\left(u_0\right)=\tau(1-\tau)\nu_0 f_u^{-1}\left(u_0\right) f_{y|u}^{-2}\left(q_\tau\left(u_0\right)\right)$.

Now we consider the comparison of the performance of the local linear estimation $\widehat{\mathbf{a}}\left(u_0\right)$ obtained in (3.4) with that of the local constant estimation $\widetilde{\mathbf{a}}\left(u_0\right)$ given in (3.5). To this effect, first, we derive the asymptotic results for the local constant estimator but the proof is omitted since it is along the same line with the proof of Theorems 3.1 and 3.2; see Xu (2005) for details. Under some regularity conditions, it can be shown that

$$\sqrt{nh}\left[\widetilde{\mathbf{a}}\left(u_0\right)-\mathbf{a}\left(u_0\right)-\widetilde{\mathbf{b}}+o_p\left(h^2\right)\right] \quad\rightarrow\quad N\left\{0,\tau(1-\tau)\nu_0\Sigma_a\left(u_0\right)\right\},$$

where

$$\widetilde{\mathbf{b}}=\frac{h^2\mu_2}{2}\left[\mathbf{a}''\left(u_0\right)+2\mathbf{a}'\left(u_0\right)f_u'\left(u_0\right)/f_u\left(u_0\right)+2\left\{\Omega^*\left(u_0\right)\right\}^{-1}\Omega^{*\prime}\left(u_0\right)\mathbf{a}'\left(u_0\right)\right],$$

which implies that the asymptotic bias for $\widetilde{\mathbf{a}}\left(u_0\right)$ is different from that for $\widehat{\mathbf{a}}\left(u_0\right)$ but both have the same asymptotic variance. Therefore, the local constant quantile estimator does not adapt to nonuniform designs: the bias can be large when $f_u'\left(u_0\right)/f_u\left(u_0\right)$ or $\left\{\Omega^*\left(u_0\right)\right\}^{-1}\Omega^{*\prime}\left(u_0\right)$ is large even when the true coefficient functions are linear. It is surprising that to the best of our knowledge, this finding seems to be new for the nonparametric quantile regression setting although it is well documented in literature for the ordinary regression case; see Fan and Gijbels (1996) for details.

Finally, to examine the asymptotic behaviors of the local linear and local constant quantile estimators at the boundaries, we offer Theorem 3.3 below but its proofs are omitted due to their similarity to those for Theorem 3.2 with some modifications and for the ordinary regression setting (Fan and Gijbels, 1996); see Xu (2005) for the detailed proofs. Without loss of generality, we consider only the left boundary point $u_0=ch, 0<c<1$, if U_t takes values only from $[0,1]$. A similar result in Theorem 3.3 holds for the right boundary point $u_0=1-ch$. Define $\mu_{j,c}=\int_{-c}^1 u^j K(u)du$ and $\nu_{j,c}=\int_{-c}^1 u^j K^2(u)du$.

Theorem 3.3: *(Asymptotic Normality) Under the assumptions in Theorem 3.2 , we have the following asymptotic normality of the local linear quantile estimator at the left boundary point,*

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(ch)-\mathbf{a}(ch)-\frac{h^2b_c}{2}\mathbf{a}''(0+)+o_p\left(h^2\right)\right] \quad\rightarrow\quad N\left\{0,\tau(1-\tau)v_c\mathbf{\Sigma}_a(0+)\right\},$$

where

$$b_c = \frac{\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c}}{\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2} \qquad \textit{and} \qquad v_c = \frac{\mu_{2,c}^2\nu_{0,c} - 2\mu_{1,c}\mu_{2,c}\nu_{1,c} + \mu_{1,c}^2\nu_{2,c}}{\left[\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2\right]^2}.$$

Further, we have the following asymptotic normality of the local constant quantile estimator at the left boundary point $u_0 = ch$ *for* $0 < c < 1$*,*

$$\sqrt{nh}\left[\widetilde{\mathbf{a}}(ch) - \mathbf{a}(ch) - \widetilde{\mathbf{b}}_c + o_p\left(h^2\right)\right] \quad \rightarrow \quad N\left\{0, \tau(1-\tau)\nu_{0,c}\boldsymbol{\Sigma}_a(0+)/\mu_{0,c}^2\right\}.$$

where

$$\widetilde{\mathbf{b}}_c = \left[h\mu_{1,c}\mathbf{a}'(0+) + \frac{h^2\mu_{2,c}}{2}\left\{\mathbf{a}''(0+) + \frac{2\mathbf{a}'(0+)f_u'(0+)}{f_u(0+)} + 2\Omega^{*-1}(0+)\Omega^{*\prime}(0+)\mathbf{a}'(0+)\right\}\right]/\mu_{0,c}.$$

Similar results hold for the right boundary point $u_0 = 1 - ch$*.*

Remark 3.5: *We remark that if the point 0 were an interior point, then, Theorem 3.3 would hold with* $c = 1$*, which becomes Theorem 3.2. Also, as* $c \to 1, b_c \to \mu_2$*, and* $v_c \to \nu_0$ *and these limits are exactly the constant factors appearing respectively in the asymptotic bias and variance for an interior point. Therefore, Theorem 3.3 shows that the local linear estimation has the automatic good behavior at boundaries without the need of boundary correction. Further, one can see from Theorem 3.3 that at the boundaries, the asymptotic bias term for the local constant quantile estimate is of the order* h *by comparing to the order* h^2 *for the local linear quantile estimate. This shows that the local linear quantile estimate does not suffer from boundary effects but the local constant quantile estimate does, which is another advantage of the local linear quantile estimator over the local constant quantile estimator. This suggests that one should use the local linear approach in practice.*

As a special case, Theorem 3.3 includes the asymptotic properties for the local constant quantile estimator of the nonparametric quantile function $q_\tau(\cdot)$ at both the interior and boundary points, stated as follows.

Corollary 3.3.1: *If there is no* $\mathbf{X}_t$ *in* (3.3)*, then, the asymptotic normality of the local constant quantile estimator is given by*

$$\sqrt{nh}\left[\widetilde{q}_\tau\left(u_0\right) - q_\tau\left(u_0\right) - \frac{h^2\mu_2}{2}\left\{q_\tau''\left(u_0\right) + 2q_\tau'\left(u_0\right)f_u'\left(u_0\right)/f_u\left(u_0\right)\right\} + o_p\left(h^2\right)\right] \quad \rightarrow \quad N\left\{0, \sigma_\tau^2\left(u_0\right)\right\}.$$

Further, at the left boundary point, we have

$$\sqrt{n\,h}\left[\widetilde{q}_\tau(c\,h) - q_\tau(ch) - \widetilde{b}_c^* + o_p\left(h^2\right)\right] \quad \rightarrow \quad N\left\{0, \sigma_c^2\right\},$$

where

$$\widetilde{b}_c^* = \left[h\mu_{1,c} q_\tau'(0+) + \frac{h^2 \mu_{2,c}}{2} \left\{ q_\tau''(0+) + 2q_\tau'(0+) f_u'(0+)/f_u(0+) \right\} \right] / \mu_{0,c}$$

and $\sigma_c^2 = \tau(1-\tau)\nu_{0,c} f_u^{-1}(0+) f_{y|u}^{-2}\left(q_\tau(0+)\right)/\mu_{0,c}^2$.

3.2.3 Bandwidth Selection

It is well known that the bandwidth plays an essential role in the trade-off between reducing bias and variance. To the best of our knowledge, there has been almost nothing done about selecting the bandwidth in the context of estimating the coefficient functions in the quantile regression even though there is a rich amount of literature on this issue in the mean regression setting; see, for example, Cai, Fan and Yao (2000). In practice, it is desirable to have a quick and easily implemented data-driven fashioned method. Based on this spirit, Yu and Jones (1998) or Yu and Lu (2004) proposed a simple and convenient method for the nonparametric quantile estimation. Their approach assumes that the second derivatives of the quantile function are parallel. However, this assumption might not be valid for many applications in economics and finance due to (nonlinear) heteroscedasticity. Further, the mean regression approach can not directly estimate the variance function. To attenuate these problems, we propose a method of selecting bandwidth for the foregoing estimation procedure, based on the nonparametric version of the Akaike information criterion, which can attend to the structure of time series data and the over-fitting or under-fitting tendency. This idea is motivated by its analogue of Cai and Tiwari (2000) and Cai (2002b) for nonlinear time series models. The basic idea is described below.

By recalling the classical AIC for linear models under the likelihood setting

$$-2(\text{maximized log quasi-likelihood}) + 2(\text{number of estimated parameters}),$$

we propose the following nonparametric version of the bias-corrected AIC, due to Hurvich and Tsai (1989) for parametric models and Hurvich, Simonoff and Tsai (1998) for nonparametric regression models, to select h by minimizing

$$\text{AIC}(h) = \log\left\{\widehat{\sigma}_\tau^2\right\} + 2\left(p_h + 1\right)/\left[n - \left(p_h + 2\right)\right], \tag{3.8}$$

where $\widehat{\sigma}_\tau^2$ and p_h are defined later. This criterion may be interpreted as the AIC for the local quantile smoothing problem and seems to perform well in some limited applications.

Note that similar to (3.8), Koenker, Ng and Portnoy (1994) considered the Schwarz information criterion (SIC) of Schwarz (1978) with the second term on the right-hand side of (3.8) replayed by $2n^{-1}p_h \log n$, where p_h is the number of "active knots" for the smoothing spline quantile setting, and Machado (1993) studied similar criteria for parametric quantile regression models and more general M-estimators of regression.

Now the question is how to define $\widehat{\sigma}_\tau^2$ and p_h in this setting. In the mean regression setting, $\widehat{\sigma}_\tau^2$ is just the estimate of the variance σ^2. In the quantile regression, we define $\widehat{\sigma}_\tau^2$ as $n^{-1}\sum_{t=1}^t \rho_\tau\left(Y_t - \mathbf{X}_t'\widehat{\mathbf{a}}\left(U_t\right)\right)$, which may be interpreted as the mean square error in the least square setting and was also used by Koenker, Ng and Portnoy (1994). In nonparametric models, p_h is the nonparametric version of degrees of freedom, called the effective number of parameters, and it is usually based on the trace of various quasi-projection (hat) matrices in the least square theory (linear estimators); see, for example, Hastie and Tibshirani (1990), Cai and Tiwari (2000), and Cai (2002b) for a cogent discussion for nonparametric regression models and nonlinear time series models. For the quantile smoothing setting, the explicit expression for the quasi-projection matrix does not exist due to its nonlinearity. However, we can use the first order approximation (the local Bahadur representation) given in (3.6) to derive an explicit expression, which may be interpreted as the quasi-projection matrix in this setting. To this end, define

$$\mathbf{S}_n = \mathbf{S}_n\left(u_0\right) = a_n \sum_{t=1}^n \xi_t \mathbf{X}_t^* \mathbf{X}_t^{*\prime} K\left(U_{th}\right),$$

where $\xi_t = I\left(Y_t \le \mathbf{X}_t'\mathbf{a}\left(u_0\right) + a_n\right) - I\left(Y_t \le \mathbf{X}_t'\mathbf{a}\left(u_0\right)\right)$ and $a_n = (n\,h)^{-1/2}$. It is shown in Section 3.5 that

$$\mathbf{S}_n\left(u_0\right) = f_u\left(u_0\right)\Omega_1^*\left(u_0\right) + o_p(1). \tag{3.9}$$

From (3.6), it is easy to verify that $\widehat{\boldsymbol{\theta}} \approx a_n \mathbf{S}_n^{-1}\sum_{t=1}^n \psi_\tau\left(Y_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right)$. Then, we have

$$\widehat{q}_\tau\left(U_t, \mathbf{X}_t\right) - q_\tau\left(U_t, \mathbf{X}_t\right) \approx \frac{1}{n}\sum_{s=1}^n \psi_\tau\left(Y_s^*\left(U_t\right)\right) K_h\left(\left(U_s - U_t\right)/h\right) \mathbf{X}_t^{0\prime}\mathbf{S}_n^{-1}\left(U_t\right)\mathbf{X}_s^*$$

where $\mathbf{X}_t^0 = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{0} \end{pmatrix}$. The coefficient of $\psi_\tau\left(Y_s^*\left(U_s\right)\right)$ on the right-hand side of the above expression is $\gamma_s = a_n^2 K(0)\mathbf{X}_s^{0\prime}\mathbf{S}_n^{-1}\left(U_s\right)\mathbf{X}_s^0$. Now, we have that $p_h = \sum_{s=1}^n \gamma_s$, which can be regarded as an approximation to the trace of the quasi-projection (hat) matrix for linear estimators. In the practical implementation, we need to estimate $\mathbf{a}\left(u_0\right)$ first since $\mathbf{S}_n\left(u_0\right)$

involves $\mathbf{a}\left(u_0\right)$. We recommend using a pilot bandwidth which can be chosen as the one proposed by Yu and Jones (1998). Similar to the least square theory, as expected, the criterion proposed in (3.6) counteracts the over-fitting tendency of the generalized crossvalidation due to its relatively weak penalty and the under-fitting of the SIC of Schwarz (1978) studied by Koenker, Ng and Portnoy (1994) because of the heavy penalty.

3.2.4 Covariance Estimate

For the purpose of statistical inference, we next consider the estimation of the asymptotic covariance matrix to construct the pointwise confidence intervals. In practice, a quick and simple way to estimate the asymptotic covariance matrix is desirable. In view of (3.7), the explicit expression of the asymptotic covariance provides a direct estimator. Therefore, we can use the so-called "sandwich" method. In other words, we need to obtain a consistent estimate for both $\Omega\left(u_0\right)$ and $\Omega^*\left(u_0\right)$. To this effect, define,

$$\widehat{\Omega}_{n,0} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t\mathbf{X}_t' K_h\left(U_t - u_0\right) \quad \text{and} \quad \widehat{\Omega}_{n,1} = \frac{1}{n}\sum_{t=1}^{n} w_t\mathbf{X}_t\mathbf{X}_t' K_h\left(U_t - u_0\right)$$

where $w_t = I\left(\mathbf{X}_t'\widehat{\mathbf{a}}\left(u_0\right) - \delta_n < Y_t \le \mathbf{X}_t'\widehat{\mathbf{a}}\left(u_0\right) + \delta_n\right) / \left(2\delta_n\right)$ for any $\delta_n \to 0$ as $n \to \infty$. It is shown in Section 3.5 that

$$\widehat{\Omega}_{n,0} = f_u\left(u_0\right)\Omega\left(u_0\right) + o_p(1) \quad \text{and} \quad \widehat{\Omega}_{n,1} = f_u\left(u_0\right)\Omega^*\left(u_0\right) + o_p(1). \tag{3.10}$$

Therefore, the consistent estimate of $\mathbf{\Sigma}_a\left(u_0\right)$ is given by

$$\widehat{\Sigma}_a\left(u_0\right) = \left[\widehat{\Omega}_{n,1}\left(u_0\right)\right]^{-1} \widehat{\Omega}_{n,0}\left(u_0\right) \left[\widehat{\Omega}_{n,1}\left(u_0\right)\right]^{-1}.$$

Note that $\widehat{\Omega}_{n,1}\left(u_0\right)$ might be close to singular for some sparse regions. To avoid this computational difficulty, there are two alternative ways to construct a consistent estimate of $f_u\left(u_0\right)\Omega^*\left(u_0\right)$ through estimating the conditional density of $Y, f_{y|u,x}\left(q_\tau(u,\mathbf{x})\right)$. The first method is the Nadaraya-Watson type (or local linear) double kernel method of Fan, Yao and Tong (1996) defined as,

$$\widehat{f}_{y|u,x}\left(q_\tau(u,\mathbf{x})\right) = \sum_{t=1}^{n} K_{h_2}\left(U_t - u, \mathbf{X}_t - \mathbf{x}\right) L_{h_1}\left(Y_t - q_\tau(u,\mathbf{x})\right) / \sum_{t=1}^{n} K_{h_2}\left(U_t - u, \mathbf{X}_t - \mathbf{x}\right),$$

where $L(\cdot)$ is a kernel function, and the second one is the difference quotients method of Koenker and Xiao (2004) such as

$$\widehat{f}_{y|u,x}\left(q_\tau(u,\mathbf{x})\right) = \left(\tau_j - \tau_{j-1}\right) / \left[q_{\tau_j}(u,\mathbf{x}) - q_{\tau_{j-1}}(u,\mathbf{x})\right],$$

for some appropriately chosen sequence of $\{\tau_j\}$; see Koenker and Xiao (2004) for more discussions. Then, in view of the definition of $f_u(u_0)\,\Omega^*(u_0)$, the estimator $\widetilde{\Omega}_{n,1}$ can be constructed as,

$$\widetilde{\Omega}_{n,1} = \frac{1}{n}\sum_{t=1}^{n} \widehat{f}_{y|u,x}\left(\widehat{q}_\tau\left(U_t, \mathbf{X}_t\right)\right)\mathbf{X}_t\mathbf{X}_t' K_h\left(U_t - u_0\right).$$

By an analogue of (3.10), one can show that under some regularity conditions, both estimators are consistent.

3.3 Empirical Examples

In this section we report a Monte Carlo simulation to examine the finite sample property of the proposed estimator and to further explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate of the Japanese Yen per US dollar and to identify the factors affecting the house price in Boston. In our computation, we use the Epanechnikov kernel $K(u) = 0.75\,(1-u^2)\,I(|u| \le 1)$ and construct the pointwise confidence intervals based on the consistent estimate of the asymptotic covariance described in Section **??** without the bias correction. For a predetermined sequence of h 's from a wide range, say from h_a to h_b with an increment h_δ, based on the AIC bandwidth selector described in Section **??**, we compute AIC(h) for each h and choose h_{opt} to minimize AIC(h).

3.3.1 A Simulated Example

Example 3.1: We consider the following data generating process

$$Y_t = a_1(U_t)\,Y_{t-1} + a_2(U_t)\,Y_{t-2} + \sigma(U_t)\,e_t, \quad t = 1, \ldots, n, \tag{3.11}$$

where $a_1(U_t) = \sin\left(\sqrt{2}\pi U_t\right), a_2(U_t) = \cos\left(\sqrt{2}\pi U_t\right)$, and $\sigma(U_t) = 3\exp\left(-4\,(U_t-1)^2\right) + 2\exp\left(-5\,(U_t-2)^2\right)$. U_t is generated from uniform $(0,3)$ independently and $e_t \sim N(0,1)$. The quantile regression is

$$q_\tau(U_t, Y_{t-1}, Y_{t-2}) = a_0(U_t) + a_1(U_t)\,Y_{t-1} + a_2(U_t)\,Y_{t-2},$$

where $a_0(U_t) = \Phi^{-1}(\tau)\sigma(U_t)$ and $\Phi^{-1}(\tau)$ is the τ-th quantile of the standard normal. Therefore, only $a_0(\cdot)$ is a function of τ. Note that $a_0(\cdot) = 0$ when $\tau = 0.5$. To assess the performance of finite samples, we compute the mean absolute deviation errors (MADE) for $\widehat{a}_j(\cdot)$, which is defined as

$$\text{MADE}_j = n_0^{-1} \sum_{k=1}^{n_0} \left| \widehat{a}_j \left(u_k \right) - a_j \left(u_k \right) \right| ,$$

where $\widehat{a}_j(\cdot)$ is either the local linear or local constant quantile estimate of $a_j(\cdot)$ and $\{z_k = 0.1(k-1) + 0.2 : 1 \leq k \leq n_0 = 27\}$ are the grid points. The Monte Carlo simulation is repeated 500 times for each sample size $n = 200, 500$, and 1000 and for each $\tau = 0.05, 0.50$ and 0.95. We compute the optimal bandwidth for each replication, sample size, and τ. We compute the median and standard deviation (in parentheses) of 500MADE values for each scenario and summarize the results in Table 3.1.

Table 3.1: The Median and Standard Deviation of 500 MADE Values

The Local Linear Estimator

	$\tau = 0.05$			$\tau = 0.5$			$\tau = 0.95$		
n	MADE$_0$	MADE$_1$	MADE$_2$	MADE $_0$	MADE$_1$	MADE$_2$	MADE $_0$	MADE$_1$	MADE$_2$
200	0.911	0.186	0.177	0.401	0.092	0.089	0.920	0.187	0.175
	(0.520)	(0.041)	(0.041)	(0.091)	(0.032)	(0.032)	(0.517)	(0.042)	(0.039)
500	0.510	0.085	0.083	0.311	0.055	0.055	0.517	0.085	0.083
	(0.414)	(0.023)	(0.02)	(0.056)	(0.019)	(0.018)	(0.390)	(0.023)	(0.023)
1000	0.419	0.060	0.059	0.311	0.050	0.049	0.416	0.060	0.059
	(0.071)	(0.018)	(0.017)	(0.051)	(0.014)	(0.014)	(0.072)	(0.017)	(0.017)

The Local Linear Estimator

	$\tau = 0.05$			$\tau = 0.5$			$\tau = 0.95$		
n	MADE$_0$	MADE$_1$	MADE$_2$	MADE $_0$	MADE$_1$	MADE$_2$	MADE $_0$	MADE$_1$	MADE$_2$
200	3.753	0.285	0.290	0.501	0.144	0.147	3.763	0.287	0.287
	(2.937)	(0.050)	(0.051)	(0.115)	(0.027)	(0.028)	(3.188)	(0.052)	(0.051)
500	2.201	0.147	0.146	0.355	0.084	0.085	2.223	0.147	0.147
	(3.025)	(0.024)	(0.025)	(0.062)	(0.016)	(0.015)	(3.320)	(0.025)	(0.025)
1000	0.883	0.086	0.086	0.322	0.060	0.061	0.882	0.086	0.087
	(0.462)	(0.015)	(0.014)	(0.054)	(0.012)	(0.011)	(0.427)	(0.015)	(0.015)

From Table 3.1, we can observe that the MADE values for both the local linear and local constant quantile estimates decrease when n increases for all three values of τ and the local linear estimate outperforms the local constant estimate. This is another example to show that the local linear method is superior over the local constant even in the quantile setting. Also, the performance for the median quantile estimate is slightly better than that for two tails ($\tau = 0.05$ and 0.95). This observation is not surprising because of the sparsity of data in the tailed regions. Moreover, another benefit of using the quantile method is that we can obtain the estimate of $a_0(\cdot)$ (conditional standard deviation) simultaneously with the estimation of $a_1(\cdot)$ and $a_2(\cdot)$ (functions in the conditional mean), which, in contrast, avoids a two-stage approach needed to estimate the variance function in the mean regression; see Fan and Yao (1998) for details. However, it is interesting to see that due to the larger variation, the performance for $a_0(\cdot)$, although it is reasonably good, is not as good as that of $a_1(\cdot)$ and $a_2(\cdot)$. This can be further evidenced from Figure 3.1. The results in this simulated

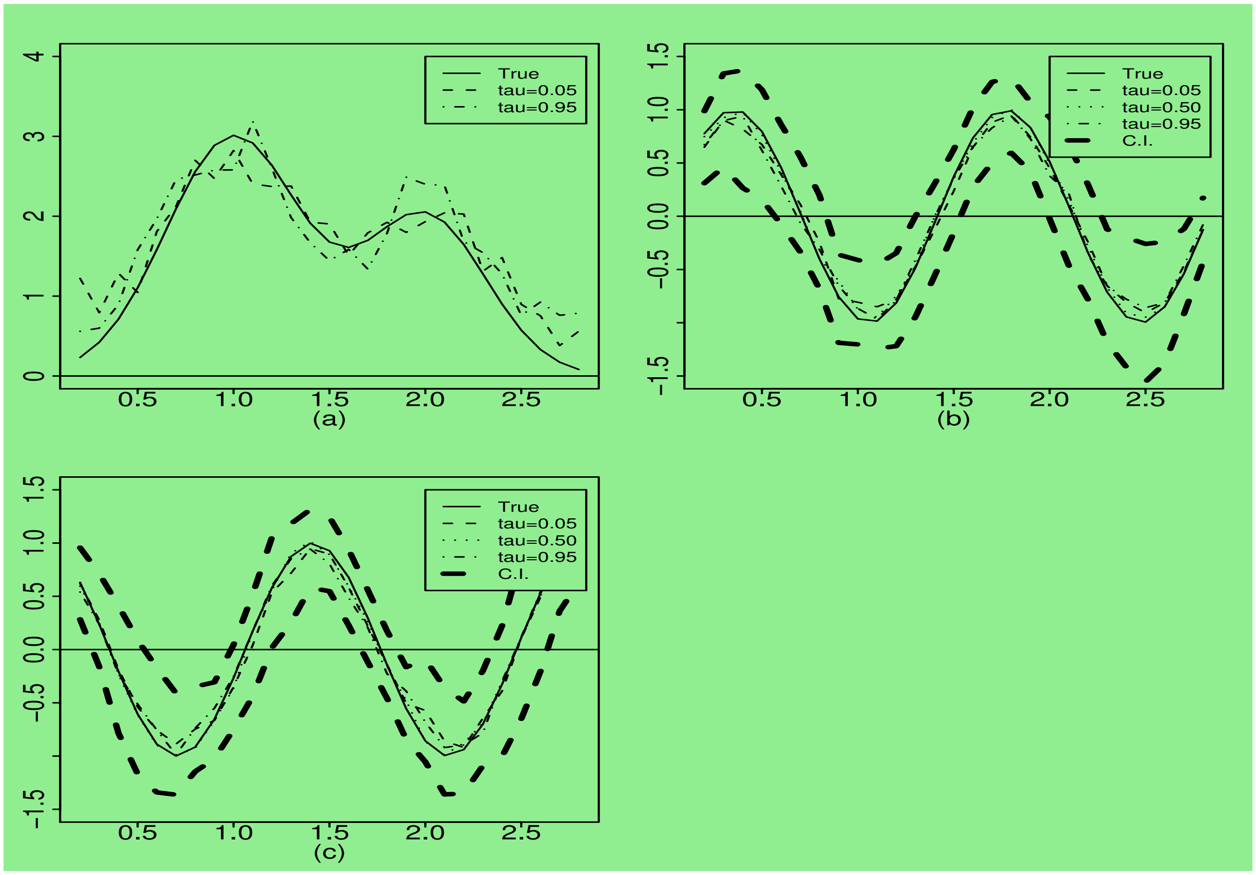


Figure 3.1: *Simulated Example*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (dashed line), $\tau = 0.50$ (dotted line), and $\tau = 0.95$ (dot-dashed line) with their true functions (solid line): $\sigma(u)$ versus u in (a), $a_1(u)$ versus u in (b), and $a_2(u)$ versus u in (c), together with the 95% point-wise confidence interval (thick line) with the bias ignored for the $\tau = 0.5$ quantile estimate.

experiment show that the proposed procedure is reliable and they are along the line of our asymptotic theory.

Finally, Figure 3.1 plots the local linear estimates for all three coefficient functions with their true values (solid line): $\sigma(\cdot)$ in Figure 3.1 (a), $a_1(\cdot)$ in Figure 3.1(b), and $a_2(\cdot)$ in Figure 3.1(c), for three quantiles $\tau = 0.05$ (dashed line), 0.50 (dotted line) and 0.95 (dotted-dashed line), for $n = 500$ based on a typical sample which is chosen based on its MADE value equal to the median of the 500 MADE values. The selected optimal bandwidths are $h_{\text{opt}} = 0.10$ for $\tau = 0.05, 0.075$ for $\tau = 0.50$, and 0.10 for $\tau = 0.95$. Note that the estimate of $\sigma(\cdot)$ for $\tau = 0.50$ can not be recovered from the estimate of $a_0(\cdot) = 0$ and it is not presented in Figure 3.1(a). The 95% point-wise confidence intervals without the bias correction are depicted in Figure 3.1 in thick lines for the $\tau = 0.05$ quantile estimate. By the same token, we can compute the point-wise confidence intervals (not shown here) for the rest. Basically, all confidence intervals cover the true values. Also, we can see that the confidence interval for $\widehat{a}_0(\cdot)$ is wider than that for $\widehat{a}_1(\cdot)$ and $\widehat{a}_2(\cdot)$ due to the larger variation. Similar plots are obtained (not shown here) for the local constant estimates due to the space limitations. Overall, the proposed modeling procedure performs fairly well.

3.3.2 Real Data Examples

Example 3.2: (*Boston House Price Data*) We analyze a subset of the Boston house price data (available at http://lib.stat.cmu.edu/datasets/boston) of Harrison and Rubinfeld (1978). This dataset consists of 14 variables collected on each of 506 different houses from a variety of locations. The dependent variable is Y, the median value of owner-occupied homes in \$1, 000 's (house price); some major factors affecting the house prices used are: proportion of population of lower educational status (i.e. proportion of adults with high school education and proportion of male workers classified as labors), denoted by U, the average number of rooms per house in the area, denoted by X_1, the per capita crime rate by town, denoted by X_2, the full property tax rate per \$10, 000, denoted by X_3, and the pupil/teacher ratio by town school district, denoted by X_4. For the complete description of all 14 variables, see Harrison and Rubinfeld (1978). Gilley and Pace (1996) provided corrections and examined censoring. Recently, there have been several papers devoted to the analysis of this dataset. For example, Breiman and Friedman (1985), Chaudhuri, Doksum and Samarov (1997), and Opsomer and Ruppert (1998) used four covariates: X_1, X_3, X_4 and U or their transformations to fit the data through a mean additive regression model whereas Yu and Lu (2004) employed the additive quantile technique to analyze the data. Further, Pace and Gilley (1997) added the geo-referencing factor to improve estimation by a spatial approach. Recently, Sentürk and Müller (2006) studied the correlation between the house price Y and the crime rate X_2 adjusted by the confounding variable U through a varying coefficient model and they concluded that the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. Some existing analyses (e.g., Breiman and Friedman, 1985; Yu and Lu, 2004) in both mean and quantile regressions concluded that most of the variation seen in housing prices in the restricted data set can be explained by two major variables: X_1 and U. Indeed, the correlation coefficients between Y and U and X_1 are -0.7377 and 0.6954 respectively. The scatter plots of Y versus U and X_1 are displayed in Figures 3.2(a) and 3.2(b), respectively. The interesting features of this data set are that the response variable is the median price of a home in a given area and the

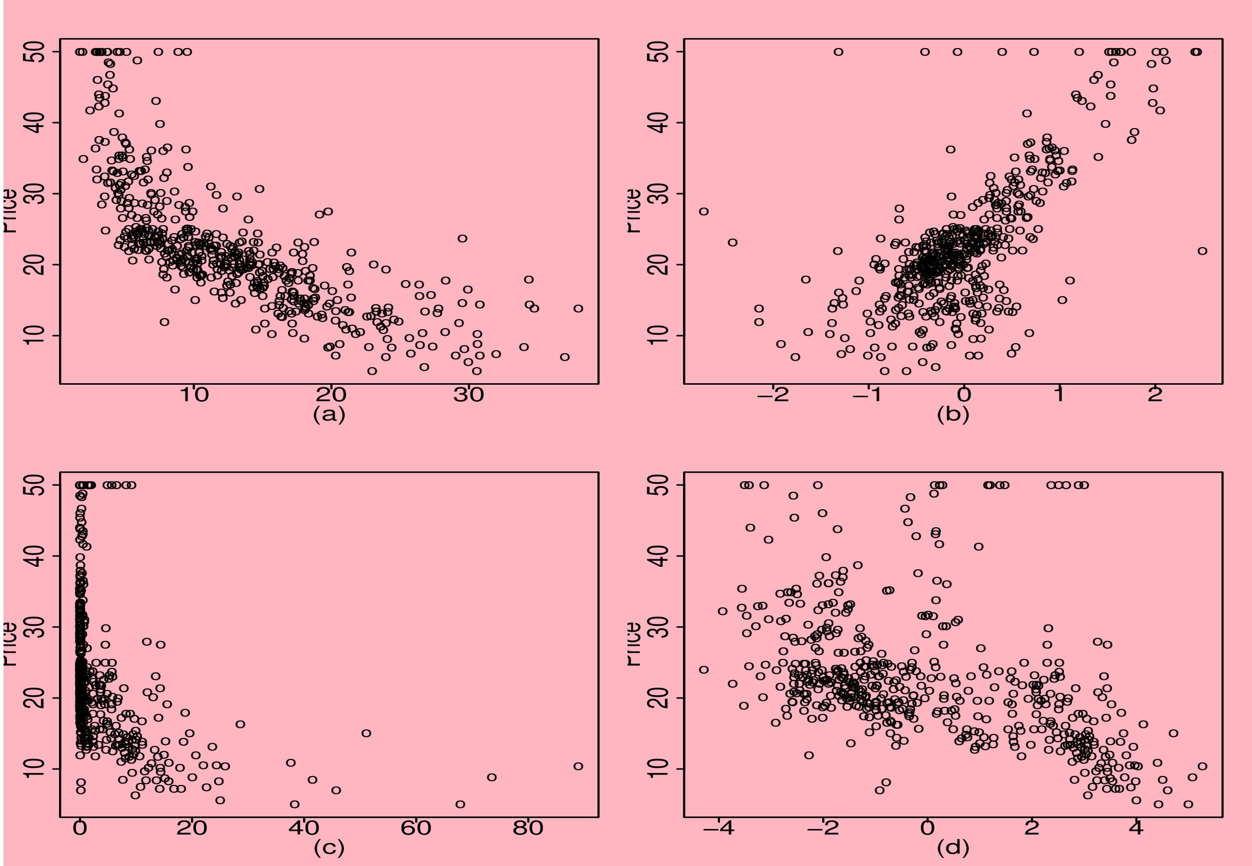


Figure 3.2: *Boston Housing Price Data*: Displayed in (a)-(d) are the scatter plots of the house price versus the covariates U, X_1, X_2 and $\log(X_2)$, respectively.

distributions of Y and the major covariate U are left skewed (the density estimates are not presented). Therefore, quantile methods are particularly well suited to the analysis of this dataset. Finally, it is surprising that all the existing nonparametric models aforementioned above did not include the crime rate X_2, which may be an important factor affecting the housing price, and did not consider the interaction terms such as U and X_2.

Based on the above discussions, it concludes that the model studied in this chapter might be well suitable to the analysis of this dataset. Therefore, we analyze this dataset by the following quantile smooth coefficient mode[2]

$$q_\tau\left(U_t, \mathbf{X}_t\right)=a_{0,\tau}\left(U_t\right)+a_{1,\tau}\left(U_t\right) X_{t1}+a_{2,\tau}\left(U_t\right) X_{t2}^*, \quad 1 \leq t \leq n=506, \tag{3.12}$$

where $X_{t2}^*=\log\left(X_{t2}\right)$. The reason for using the logarithm of X_{t2} in (3.12), instead of X_{t2} itself, is that the correlation between Y_t and X_{t2}^* (the correlation coefficient is -0.4543) is slightly stronger than that for Y_t and $X_{t2}(-0.3883)$, which can be witnessed as well from Figures 3.2(c) and 3.2(d). In the model fitting, covariates X_1 and X_2 are centralized. For the purpose of comparison, we also consider the following functional coefficient model in the mean regression

$$Y_t=a_0\left(U_t\right)+a_1\left(U_t\right) X_{t1}+a_2\left(U_t\right) X_{t2}^*+e_t, \tag{3.13}$$

and we employ the local linear fitting technique to estimate the coefficient functions $\{a_j(\cdot)\}$, denoted by $\{\widehat{a}_j(\cdot)\}$; see Cai, Fan and Yao (2000) for details.

The coefficient functions are estimated through the local linear quantile approach by using the bandwidth selector described in Section **??**. The selected optimal bandwidths are $h_{\text{opt}}=2.0$ for $\tau=0.05, 1.5$ for $\tau=0.50$, and 3.5 for $\tau=0.95$. Figures 3.3(e), 3.3(f) and 3.3(g) present the estimated coefficient functions $\widehat{a}_{0,\tau}(\cdot), \widehat{a}_{1,\tau}(\cdot)$, and $\widehat{a}_{2,\tau}(\cdot)$ respectively, for three quantiles $\tau=0.05$ (solid line), 0.50 (dashed line) and 0.95 (dotted line), together with the estimates $\{\widehat{a}_j(\cdot)\}$ from the mean regression model (dot-dashed line). Also, the 95% point-wise confidence intervals for the median estimate are displayed by the thick dashed lines without the bias correction. First, from these three figures, one can see that the median estimates are quite close to the mean estimates and the estimates based on the mean regression are always within the 95% confidence interval of the median estimates. It can be concluded that the distribution of the measurement error e_t in (3.13) might be symmetric and $\widehat{a}_{j,0.5}(\cdot)$ in

[2] We do not include the other variables such as X_3 and X_4 in model (3.12), since we found that the coefficient functions for these variables seem to be constant. Therefore, a semiparametric model would be appropriate if the model includes these variables. It of course deserves a further investigation.

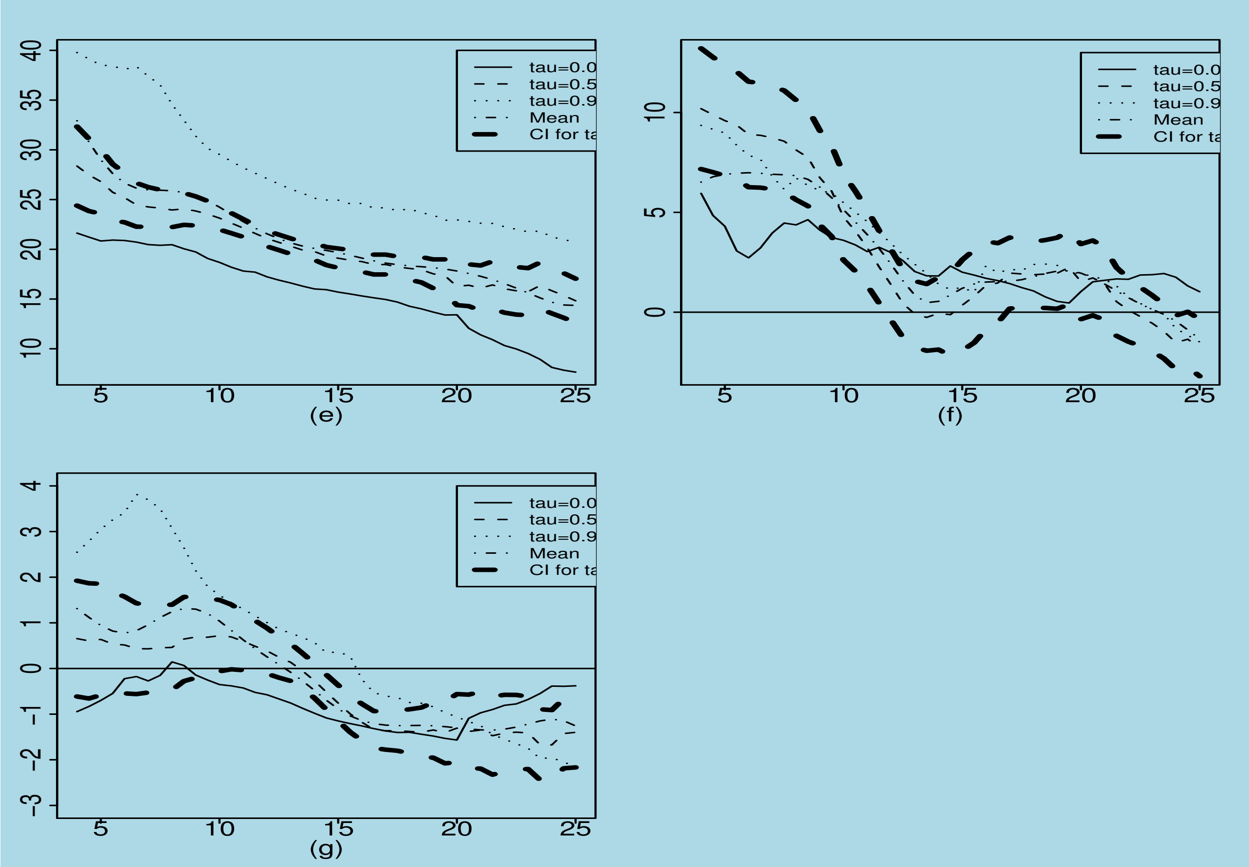


Figure 3.3: *Boston Housing Price Data*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,\tau}(u)$ and $\widehat{a}_0(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored.

(3.12) is almost same as $\widehat{a}_j(\cdot)$ in (3.13). Also, one can observe from Figure 3.3(e) that three quantile curves are parallel, which implies that the intercept in $\widehat{a}_{0,\tau}(\cdot)$ depends on τ, and they decrease exponentially, which can support that the logarithm transformation may be needed as argued in Yu and Lu (2004). More importantly, one can observe from Figures 3.3(f) and 3.3(g) that three quantile estimated coefficient curves are intersect. This reveals that the structure of quantiles is complex and the lower and upper quantiles have different behaviors and the heteroscedasticity might exist. But unfortunately, this phenomenon was not observed in any previous analyses in the aforementioned papers.

From Figure 3.3(f), first, we can observe that $\widehat{a}_{1,0.50}(\cdot)$ and $\widehat{a}_{1,0.95}(\cdot)$ are almost same but $\widehat{a}_{1,0.05}(\cdot)$ is different. Secondly, we can see that the correlation between the house price and the number of rooms per house is almost positive except for houses with the median price and/or higher than ($\tau = 0.50$ and 0.95) in very low educational status neighborhoods ($U > 23$). Thirdly, for the low price houses ($\tau = 0.05$), the correlation is always positive and it deceases when U is between 0 and 14 and then keeps almost constant afterwards. This implies that the expected effect of increasing the number of rooms can make the house price slightly higher in any low educational status neighborhoods but much higher in relatively high educational status neighborhoods. Finally, for the median and/or higher price houses, the correlation deceases when U is between 0 and 14 and then keeps almost constant until U up to 20 and finally deceases again afterwards, and it becomes negative for U larger than 23 . This means that the number of room has a positive effect on the median and/or higher price houses in relatively high and low educational status neighborhoods but increasing the number of rooms might not increase the house price in very low educational status neighborhoods. In other words, it is very difficult to sell high price houses with high number of rooms at a reasonable price in very low educational status neighborhoods.

Finally, from Figure 3.3(g), first, one can conclude that the overall trend for all curves is decreasing with $\widehat{a}_{3,0.95}(\cdot)$ deceasing faster than the others, and that $\widehat{a}_{3,0.05}(\cdot)$ and $\widehat{a}_{3,0.50}(\cdot)$ tend to be constant for U larger than 16. Secondly, the correlation between the housing prices ($\tau = 0.50$ and 0.95) and the crime rate seems to be positive for smaller U values (about $U \leq 13$) and becomes negative afterwards. This positive correlation between the housing prices ($\tau = 0.50$ and 0.95) and the crime rate for relatively high educational status neighborhoods seems against intuitive. However, the reason for this positive correlation is the existence of high educational status neighborhoods close to central Boston where high house prices and crime rate occur simultaneously. Therefore, the expected effect of increasing crime rate on declining house prices for $\tau = 0.50$ and 0.95 seems to be observed only for lower educational status neighborhoods in Boston. Finally, it can be seen that the correlation between the housing prices for $\tau = 0.05$ and the crime rate is almost negative although the degree depends on the value of U. This implies that increasing crime rate slightly decreases relatively the house prices for the cheap houses ($\tau = 0.05$).

In summary, it concludes that there is a nonlinear relationship between the conditional quantiles of the housing price and the affecting factors. It seems that the factors U, X_1 and X_2 do have different effects on the different quantiles of the conditional distribution of the housing price. Overall, the housing price and the proportion of population of lower educational status have a strong negative correlation, and the number of rooms has a mostly positive effect on the housing price whereas the crime rate has the most negative effect on the housing price. In particular, by using the proportion of population of lower educational status U as the confounding variable, we demonstrate the substantial benefits obtained by characterizing the affecting factors X_1 and X_2 on the housing price based on the neighborhoods.

Example 3.3: (*Exchange Rate Data*) This example concerns the closing bid prices of the Japanese Yen (JPY) in terms of US dollar. There is a vast amount of literature devoted to the study of the exchange rate time series; see Sercu and Uppal (2000) and the references therein for details. Here we use the proposed model and its modeling approaches to explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate series. The data is a weekly series from January 1, 1974 to December 31, 2003. The daily noon buying rates in New York City certified by the Federal Reserve Bank of New York for customs and cable transfers purposes were obtained from the Chicago Federal Reserve Board (http://www.chicagofed.org). The weekly series is generated by selecting the Wednesdays series (if a Wednesday is a holiday then the following Thursday is used), which has 1566 observations. The use of weekly data avoids the so-called weekend effect as well as other biases associated with non-trading, bid-ask spread, asynchronous rates and so on, which are often present in higher frequency data. The previous analysis of this "particularly difficult" data set can be found in Gallant, Hsieh and Tauchen (1991), Fan, Yao and Cai (2003), and Hong and Lee (2003), and the references within. We model the return series $Y_t = 100 \log\left(\xi_t/\xi_{t-1}\right)$, plotted in Figure 3.4(a), using the techniques developed in this chapter, where ξ_t is an exchange rate level on the t-th week. Typically the classical financial theory would treat $\{Y_t\}$ as a martingale difference process. Therefore, Y_t would be unpredictable. But this assumption was strongly rejected by Hong and Lee (2003) by examining five major currencies and applying several testing procedures. Note that the return series $\{Y_t\}$ has 1565 observations. Figure 3.4(b) shows that there exists almost no significant autocorrelation in $\{Y_t\}$, which also was confirmed by Tsay (2002) and Hong and Lee (2003) by using several statistical testing procedures.

Based on the evidence from Fan, Yao and Cai (2003) and Hong and Lee (2003), the exchange rate series is predictable by using the functional coefficient autoregressive model

$$Y_t = a_0\left(U_t\right) + \sum_{j=1}^{d} a_j\left(U_t\right) Y_{t-j} + \sigma_t e_t, \tag{3.14}$$

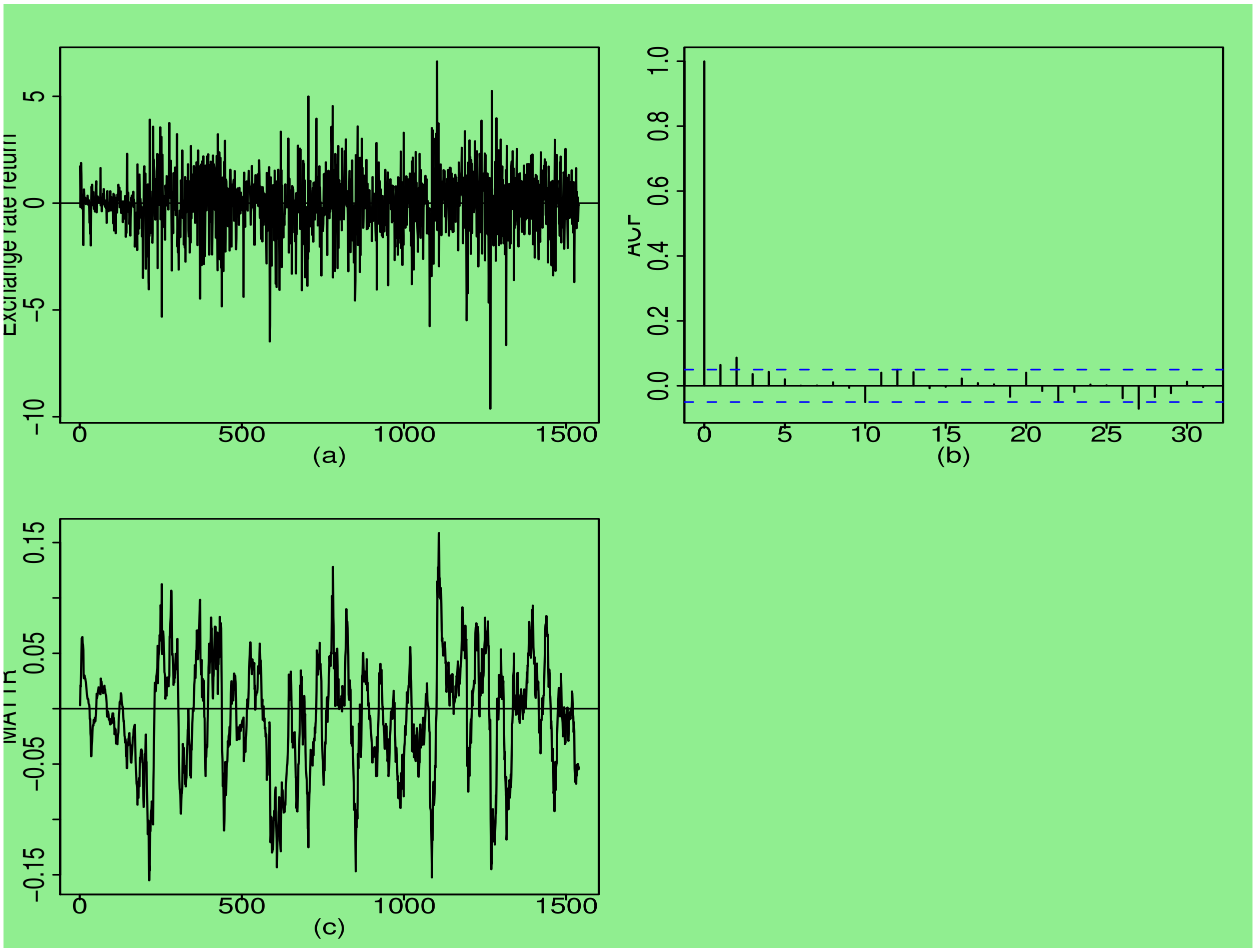


Figure 3.4: *Exchange Rate Series:* (a) Japanese-dollar exchange rate return series $\{Y_t\}$; (b) autocorrelation function of $\{Y_t\}$; (c) moving average trading technique rule.

where U_t is the smooth variable defined later and σ_t is a function of U_t and the lagged variables. If $\{U_t\}$ is observable, $a_j(\cdot)$ can be estimated by a local linear fitting; see Cai, Fan and Yao (2000) for details, denoted by $\widehat{a}_j(\cdot)$. Here, σ_t is the stochastic volatility which may depend on U_t and the lagged variables $\{Y_{t-j}\}$. Now the question is how to choose U_t. Usually, U_t can be chosen based on the knowledge of data or economic theory. However, if no prior information is available, U_t may be chosen as a function of explanatory vector $\{\xi_{t-j}\}$ or through the use of data-driven methods such as AIC or cross-validation. Recently, Fan, Yao and Cai (2003) proposed a data-driven method to the choice of U_t by a linear combination of $\{\xi_{t-j}\}$ and the lagged variables $\{Y_{t-j}\}$. By following the analysis of Fan, Yao and Cai (2003) and Hong and Lee (2003), we choose the smooth variable U_t as a moving average technical trading rule (MATTR) in finance so that the autoregressive coefficients vary with investment positions. U_t is defined as $U_t = \xi_{t-1}/M_t - 1$, where $M_t = \sum_{j=1}^{L} \xi_{t-j}/L$, which is the moving average and can be regarded as a proxy for the trend at the time $t-1$. Similar to Hong and Lee (2003), We choose $L = 26$ (half a year). $U_t + 1$ is the ratio of the exchange rate at the time $t-1$ to the average rate of the most recent L periods of exchange rates at time $t-1$. The time series plot of $\{U_t\}$ is given in Figure 3.4(c). As pointed out by Hong and Lee (2003), U_t is expected to reveal some useful information on the direction of changes. The MATTR signals 1 (the position to buy JPY) when $U_t > 0$ and -1 (the position to sell JPY) when $U_t < 0$. For the detailed discussions of the MATTR, see (for example) the papers by LeBaron (1997, 1999), Hong and Lee (2003), Fan, Yao and Cai (2003), and the reference therein. Note that model (3.12) was studied by Fan, Yao and Cai (2003) for the daily data and Hong and Lee (2003) for the weekly data under the homogenous assumption (assume that $\sigma_t = \sigma$) based on the least square theory. In particular, Hong and Lee (2003) provided some empirical evidences to conclude that model (3.14) outperforms the martingale model and autoregressive models.

We analyze this exchange rate series by using the smooth coefficient model under the quantile regression framework with only two lagged variables[3] as follows

$$q_\tau\left(U_t, Y_{t-1}, Y_{t-2}\right) = a_{0,\tau}\left(U_t\right) + a_{1,\tau}\left(U_t\right) Y_{t-1} + a_{2,\tau}\left(U_t\right) Y_{t-2}. \tag{3.15}$$

The first 1540 observations of $\{Y_t\}$ are used for estimation and the last 25 observations are left for prediction. The coefficient functions $\{a_{j,\tau}(\cdot)\}$ are estimated through the local linear quantile approach, denoted by $\{\widehat{a}_{j,\tau}(\cdot)\}$. The previous analysis of this “particularly difficult” data set can be found in optimal bandwidths are $h_{\text{opt}} = 0.03$ for $\tau = 0.05, 0.025$ for $\tau = 0.50$, and 0.03 for $\tau = 0.95$. Figures 3.5(d) - 3.5(g) depict the estimated coefficient functions $\widehat{a}_{0,\tau}(\cdot), \widehat{a}_{1,\tau}(\cdot)$, and $\widehat{a}_{2,\tau}(\cdot)$ respectively, for three quantiles $\tau = 0.05$ (solid line), 0.50 (dashed line) and 0.95 (dotted line), together with the estimates $\{\widehat{a}_j(\cdot)\}$ (dot-dashed line) from the mean regression model in (3.14). Also, the 95% point-wise confidence intervals for the median estimate are displayed by the thick dashed lines without the bias correction.

First, from Figures 3.5(d), 3.5(f) and 3.5(g), we see clearly that the median estimates $\widehat{a}_{j,0.50}(\cdot)$ in (3.15) are almost parallel with or close to the mean estimates $\widehat{a}_j(\cdot)$ in (3.14) and the mean estimates are almost within the 95% confidence interval of the median estimates. Secondly, $\widehat{a}_{0,0.50}(\cdot)$ in Figure 3.5(d) shows a nonlinear pattern (increasing and then decreasing) and $\widehat{a}_{0,0.05}(\cdot)$ and $\widehat{a}_{0,0.95}(\cdot)$ in Figure 3.5(e) exhibit nonlinearly (slightly U-shape) and

[3] We also considered the models with more than two lagged variables and we found that the conclusions are similar and not reported here.

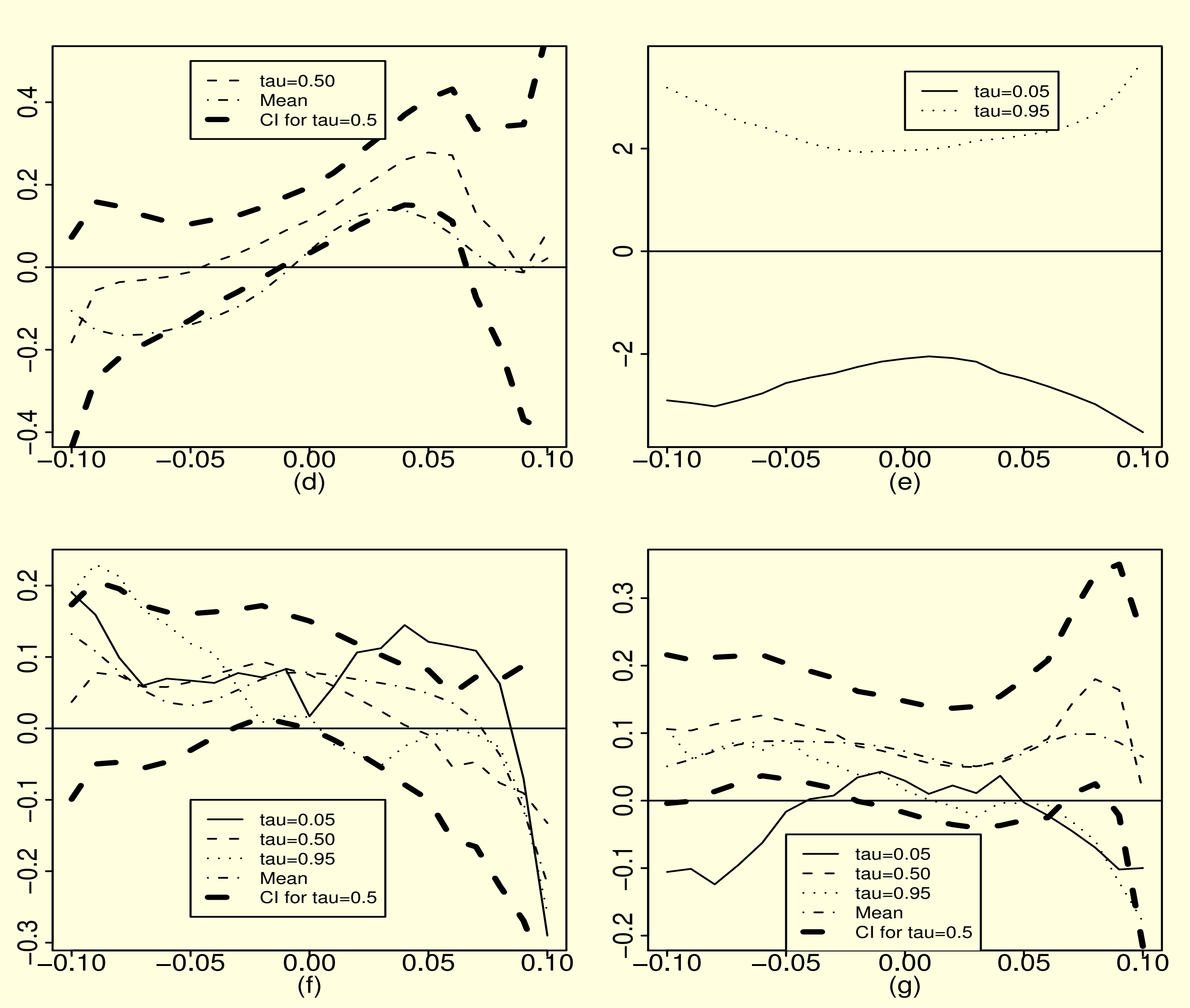


Figure 3.5: *Exchange Rate Series:* The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,0.50}(u)$ and $\widehat{a}_0(u)$ versus u in (d), $\widehat{a}_{0,0.05}(u)$ and $\widehat{a}_{0,0.95}(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored.

symmetrically. More importantly, one can observe from Figures 3.5(f) and 3.5(g) that the lower and upper quantile estimated coefficient curves are intersect and they behave slightly differently. Particularly, from Figure 3.5(g), we observe that $\widehat{a}_{2,0.05}\,(U_t)$ seems to be nonlinear but $\widehat{a}_{2,0.95}\,(U_t)$ looks like constant when $U_t < 0.06$, and both $\widehat{a}_{2,0.05}\,(U_t)$ and $\widehat{a}_{2,0.95}\,(U_t)$ decrease when $U_t > 0.06$. One might conclude that the distribution of the measurement error e_t in (3.14) might not be symmetric about 0 and there exists a nonlinearity in $a_{j,\tau}(\cdot)$. This supports the nonlinearity test of Hong and Lee (2003). Also, our findings lead to the conclusions that the quantile has a complex structure and the heteroscedasticity exists. This observation supports the existing conclusion in literature that the GARCH (generalized ARCH) effects occur in the exchange rate time series; see Engle, Ito and Lin (1990) and Tsay (2002).

Finally, we consider the post-sample forecasting for the last 25 observations based on the local linear quantile estimators which are computed by using the same bandwidths as those used in the model fitting. The 95% nonparametric prediction interval is constructed as $(\widehat{q}_{0.025}(\cdot), \widehat{q}_{0.975}(\cdot))$ and the prediction results are reported in Table 3.2, which shows that 24 out of 25 (96%) predictive intervals contain the corresponding true values. The average length of the intervals is 5.77, which is about 35.5% of the range of the data. Therefore, we can conclude that under the dynamic smooth coefficient quantile regression model assumption, the prediction intervals based on the proposed method work reasonably well.

3.4 Derivations

In this section, we give the derivations of the theorems and present certain lemmas with their detailed proofs relegated to Section 3.5. First, we need the following two lemmas.

Lemma 3.1: Let $V_n(\Delta)$ be a vector function that satisfies
(i) $-\Delta' V_n(\lambda\Delta) \geq -\Delta' V_n(\Delta)$ for $\lambda \geq 1$
and
(ii) $\sup_{\|\Delta\| \leq M} \|V_n(\Delta) + \mathbf{D}\Delta - \mathbf{A}_n\| = o_p(1)$, where $\|\mathbf{A}_n\| = O_p(1), 0 < M < \infty$, and $\mathbf{D}$ is a positive-definite matrix. Suppose that Δ_n is a vector such that $\|V_n(\Delta_n)\| = o_p(1)$, then, we have

$$(1)\, \|\Delta_n\| = O_p(1) \quad \text{and} \quad (2) \Delta_n = \mathbf{D}^{-1}\mathbf{A}_n + o_p(1).$$

Proof : The proof follows from Jurekova (1977) and Koenker and Zhao (1996). □

Lemma 3.2: Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of the function

$$\sum_{t=1}^{n} w_t \rho_\tau \left(y_t - \mathbf{X}_t' \boldsymbol{\beta}\right),$$

where $w_t > 0$. Then,

$$\left\| \Sigma_{t=1}^{n} w_t \mathbf{X}_t \psi_\tau \left(y_t - \mathbf{X}_t' \widehat{\boldsymbol{\beta}}\right) \right\| \leq \dim(\mathbf{X}) \max_{t \leq n} \|w_t \mathbf{X}_t\|.$$

Table 3.2: The Post-Sample Predictive Intervals For Exchange Rate Data

Observation	True Value	Prediction Interval
Y_{1541}	0.3920	$(-2.891, 2.412)$
Y_{1542}	0.5090	$(-3.099, 2.405)$
Y_{1543}	1.5490	$(-2.943, 2.446)$
Y_{1544}	-0.121	$(-2.684, 2.525)$
Y_{1545}	-0.991	$(-2.677, 2.530)$
Y_{1546}	-0.646	$(-3.110, 2.401)$
Y_{1547}	-0.354	$(-3.178, 2.365)$
Y_{1548}	-1.393	$(-3.083, 2.372)$
Y_{1549}	0.9970	$(-3.110, 2.230)$
Y_{1550}	-0.916	$(-3.033, 2.431)$
Y_{1551}	$\mathbf{-3.707}$	$(\mathbf{-3.021}, \mathbf{2.286})$
Y_{1552}	-0.919	$(-3.841, 2.094)$
Y_{1553}	-0.901	$(-3.603, 2.770)$
Y_{1554}	0.0710	$(-3.583, 2.821)$
Y_{1555}	-0.497	$(-3.351, 2.899)$
Y_{1556}	-0.648	$(-3.436, 2.783)$
Y_{1557}	1.6480	$(-3.524, 2.866)$
Y_{1558}	-1.184	$(-3.121, 2.810)$
Y_{1559}	0.5300	$(-3.529, 2.531)$
Y_{1560}	0.1070	$(-3.222, 2.648)$
Y_{1561}	-0.804	$(-3.294, 2.651)$
Y_{1562}	0.2740	$(-3.419, 2.534)$
Y_{1563}	-0.847	$(-3.242, 2.640)$
Y_{1564}	-0.060	$(-3.426, 2.532)$
Y_{1565}	-0.088	$(-3.300, 2.576)$

Proof : The proof follows from Ruppert and Carroll (1980). From the definition of $\boldsymbol{\theta}$, we have

$$\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a}(u_0) \\ \mathbf{a}'(u_0) \end{pmatrix} + a_n \mathbf{H}^{-1} \boldsymbol{\theta},$$

where a_n is defined in (3.10). Then, $Y_t - \sum_{j=0}^{q} \mathbf{X}_t' \boldsymbol{\beta}_j (U_t - u_0)^j = Y_t^* - a_n \boldsymbol{\theta}' \mathbf{X}_t^*$. Therefore,

$$\widehat{\boldsymbol{\theta}} = \text{argmin} \sum_{t=1}^{n} \rho_\tau \left[Y_t^* - a_n \boldsymbol{\theta}' \mathbf{X}_t^* \right] K(U_{th}) \equiv \text{argmin} G(\boldsymbol{\theta}).$$

Now, define $V_n(\boldsymbol{\theta})$ as

$$V_n(\boldsymbol{\theta}) = a_n \sum_{t=1}^{n} \psi_\tau \left[Y_t^* - a_n \boldsymbol{\theta}' \mathbf{X}_t^*\right] \mathbf{X}_t^* K\left(U_{th}\right). \tag{3.16}$$

To establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}$, in the next three lemmas, we show that $V_n(\boldsymbol{\theta})$ satisfies Lemma 3.1 so that we can derive the local Bahadur representation for $\widehat{\boldsymbol{\theta}}$. The results are stated here and their detailed proofs are given in Section 3.5 For the notational convenience define $A_m = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq M\}$ for some $0 < M < \infty$. □

Lemma 3.3: Under the assumptions in Theorem 3.1, we have

$$\sup_{\boldsymbol{\theta} \in A_m} \|V_n(\theta) - V_n(0) - E\left[V_n(\theta) - V_n(0)\right]\| = o_p(1).$$

Lemma 3.4: Under then assumptions in Theorem 3.1, we have

$$\sup_{\boldsymbol{\theta} \in A_m} \|E\left[V_n(\boldsymbol{\theta}) - V_n(0)\right] + f\left(u_0\right) \Omega_1^*\left(u_0\right) \boldsymbol{\theta}\| = o(1).$$

Lemma 3.5: Let $\mathbf{Z}_t = \psi_\tau\left(Y_t^*\right) \mathbf{X}_t^* K\left(U_{th}\right)$. Under the assumptions in Theorem 3.1, we have

$$E\left[\mathbf{Z}_1\right] = \frac{h^3 f\left(u_0\right)}{2} \begin{pmatrix} \mu_2 \Omega^*\left(u_0\right) \mathbf{a}''\left(u_0\right) \\ \mathbf{0} \end{pmatrix} \{1 + o(1)\}$$

and

$$\operatorname{Var}\left[\mathbf{Z}_1\right] = h\tau(1-\tau) f\left(u_0\right) \Omega_1\left(u_0\right) \{1 + o(1)\},$$

where

$$\Omega_1\left(u_0\right) = \begin{pmatrix} \nu_0 \Omega\left(u_0\right) & \mathbf{0} \\ \mathbf{0} & \nu_2 \Omega\left(u_0\right) \end{pmatrix}.$$

Further,

$$\operatorname{Var}\left[V_n(0)\right] \quad \to \quad \tau(1-\tau) f\left(u_0\right) \Omega_1\left(u_0\right).$$

Therefore, $\|V_n(0)\| = O_p(1)$.

Now we can embrace the proofs of the theorems.

Proof of Theorem 3.1. By Lemmas 3.5, 3.3, and 3.4, $V_n(\boldsymbol{\theta})$ satisfies the condition (ii) of Lemma 3.1; that is, $\|\mathbf{A}_n\| = O_p(1)$ and $\sup_{\boldsymbol{\theta} \in A_m} \|V_n(\boldsymbol{\theta}) + \mathbf{D}\boldsymbol{\theta} - \mathbf{A}_n\| = o_p(1)$ with $\mathbf{D} =$

$f_u(u_0)\,\Omega_1^*(u_0)$ and $\mathbf{A}_n = V_n(0)$. It follows Lemma 3.2 that $\left\|V_n(\widehat{\boldsymbol{\theta}})\right\| = o_p(1)$, where $\widehat{\boldsymbol{\theta}}$ is the minimizer of $G(\boldsymbol{\theta})$. Finally, since $\psi_\tau(x)$ is an increasing function of x, then,

$$\begin{aligned}-\boldsymbol{\theta}' V_n(\lambda\boldsymbol{\theta}) &= a_n \sum_{t=1}^{n} (-\boldsymbol{\theta}')\left(\psi_\tau\left(Y_t^* - \lambda a_n \boldsymbol{\theta}'\mathbf{X}_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right)\right. \\ &= a_n \sum_{t=1}^{n} \psi_\tau\left[Y_t^* + \lambda a_n\left(-\boldsymbol{\theta}'\mathbf{X}_t^*\right)\right]\left(-\boldsymbol{\theta}'\mathbf{X}_t^*\right) K\left(U_{th}\right)\end{aligned}$$

is an increasing function of λ. Thus, the condition (i) of Lemma 3.1 is satisfied. Therefore, it follows that

$$\widehat{\boldsymbol{\theta}} = \mathbf{D}^{-1}\mathbf{A}_n + o_p(1) = \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \psi_\tau\left(Y_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right) + o_p(1). \tag{3.17}$$

This proves (3.6). □

Proof of Theorem 3.2. Let $\varepsilon_t = \psi_\tau\left(Y_t - \mathbf{X}_t'\mathbf{a}\left(U_t\right)\right)$. Then, $E\left(\varepsilon_t\right) = 0$ and $\mathrm{Var}\left(\varepsilon_t\right) = \tau(1-\tau)$. From (3.17),

$$\widehat{\boldsymbol{\theta}} \approx \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \left[\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right]\mathbf{X}_t^* K\left(U_{th}\right) + \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \varepsilon_t \mathbf{X}_t^* K\left(U_{th}\right) \equiv \mathbf{B}_n + \boldsymbol{\xi}_n$$

Similar to the proof of Theorem 2 in Cai, Fan and Yao (2000), by using the small-block and large-block technique and the CramÃ©r-Wold device, one can show that

$$\boldsymbol{\xi}_n \quad \to \quad N\left(\mathbf{0}, \boldsymbol{\Sigma}\left(u_0\right)\right). \tag{3.18}$$

By the stationarity and Lemma 3.5,

$$E\left[\mathbf{B}_n\right] = \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} nE\left[\mathbf{Z}_1\right]\{1+o(1)\} = a_n^{-1}\frac{h^2}{2}\begin{pmatrix}\mathbf{a}''\left(u_0\right)\mu_2 \\ \mathbf{0}\end{pmatrix}\{1+o(1)\}. \tag{3.19}$$

Since $\psi_\tau\left(Y_t^*\right) - \varepsilon_t = I\left(Y_t \le \mathbf{X}_t'\mathbf{a}\left(U_t\right)\right) - I\left(Y_t \le \mathbf{X}_t'\left(\mathbf{a}\left(u_0\right) + \mathbf{a}'\left(u_0\right)\left(U_t - u_0\right)\right)\right)$, then,

$$\left[\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right]^2 = I\left(d_{1t} < Y_t \le d_{2t}\right) \tag{3.20}$$

where $d_{1t} = \min\left(c_{1t}, c_{2t}\right)$ and $d_{2t} = \max\left(c_{1t}, c_{2t}\right)$ with $c_{1t} = \mathbf{X}_t'\mathbf{a}\left(U_t\right)$ and $c_{2t} = \mathbf{X}_t'\left[\mathbf{a}\left(u_0\right) + \mathbf{a}'\left(u_0\right)\left(U_t - u_0\right)\right]$. Further,

$$E\left[\left\{\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right\}^2 K^2\left(U_{th}\right)\mathbf{X}_t^*\mathbf{X}_t^{*\prime}\right] = E\left[\left\{F_{y|u,x}\left(d_{2t}\right) - F_{y|u,x}\left(d_{1t}\right)\right\} K^2\left(U_{th}\right)\mathbf{X}_t^*\mathbf{X}_t^{*\prime}\right] = O\left(h^3\right).$$

Thus, $\mathrm{Var}\left(\mathbf{B}_n\right) = o(1)$. This, in conjunction with (3.18) and (3.19) and the Slutsky Theorem, proves the theorem. □

3.5 Proofs of Lemmas

Note that the same notations in Sections 3.2 and 3.4 are used here. Throughout this section, we denote a generic constant by C, which may take different values at different appearances. Let $F_{y|u,x}(y)$ denote the conditional distribution of Y given U and $\mathbf{X}$.

Proof of Lemma 3.3. First, for any $\boldsymbol{\theta} \in A_m$, we consider the following term

$$V_n(\boldsymbol{\theta}) - V_n(0) = a_n \sum_{t=1}^{n} \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t^* K\left(U_{th}\right) \equiv a_n \sum_{i=1}^{n} V_{nt}(\boldsymbol{\theta}),$$

where $Y_{nt}^* = Y_t^* - a_n \boldsymbol{\theta}' \mathbf{X}_t^*$ and $V_{nt}(\boldsymbol{\theta}) = V_{nt} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t^* K\left(U_{th}\right) = \left(V_{nt1}', V_{nt2}'\right)'$ with

$$V_{nt1} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t K\left(U_{th}\right) \quad \text{and} \quad V_{nt2} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t U_{th} K\left(U_{th}\right).$$

Thus,

$$\begin{aligned} &\left\| V_n(\boldsymbol{\theta}) - V_n(0) - E\left[V_n(\boldsymbol{\theta}) - V_n(0)\right] \right\| \\ \leq & a_n \left\| \sum_{t=1}^{n} \left(V_{nt1} - EV_{nt1}\right) \right\| + a_n \left\| \sum_{t=1}^{n} \left(V_{nt2} - EV_{nt2}\right) \right\| \equiv V_n^{(1)} + V_n^{(2)}. \end{aligned}$$

Clearly,

$$V_n^{(1)} \equiv a_n \left\| \sum_{t=1}^{n} \left(V_{nt1} - EV_{nt1}\right) \right\| \leq \sum_{i=0}^{d} \left\| V_n^{(1i)} \right\|,$$

where $V_n^{(1i)} = a_n \sum_{t=1}^{n} \left(V_{nt1}^{(i)} - EV_{nt1}^{(i)}\right)$ and $V_{nt1}^{(i)} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] X_{ti} K\left(U_{th}\right)$, which is the i-th component of V_{nt1}. Then,

$$\begin{aligned} \operatorname{Var}\left(V_n^{(1i)}\right) &= a_n^2 E\left\{ \sum_{t=1}^{n} \left(V_{nt1}^{(i)} - EV_{nt1}^{(i)}\right) \right\}^2 \\ &= a_n^2 \left[\sum_{t=1}^{n} \operatorname{Var}\left(V_{nt1}^{(i)}\right) + 2 \sum_{s=1}^{n-1} \left(1 - \frac{s}{n}\right) \operatorname{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right) \right] \\ &\leq \frac{1}{h} \left[\operatorname{Var}\left(V_{n11}^{(i)}\right) + 2 \sum_{s=1}^{d_n - 1} \left| \operatorname{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right) \right| + 2 \sum_{s=d_n}^{\infty} \left| \operatorname{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right) \right| \right] \\ &\equiv J_1 + J_2 + J_3 \end{aligned}$$

for some $d_n \to \infty$ specified later. For J_3, use the Davydov's inequality (see Lemma 1.1) to obtain

$$\left|\mathrm{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right| \le C\alpha^{1-2/\delta}(s)\left[E\left|V_{n11}^{(i)}\right|^{\delta}\right]^{2/\delta}.$$

Similar to (3.20), for any $k > 0$,

$$\left|\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right|^k = I\left(d_{3t} < Y_t \le d_{4t}\right),$$

where $d_{3t} = \min\left(c_{2t}, c_{2t} + c_{3t}\right)$ and $d_{4t} = \max\left(c_{2t}, c_{2t} + c_{3t}\right)$ with $c_{3t} = a_n\boldsymbol{\theta}'\mathbf{X}_t^*$. Therefore, by Assumption (C3), there exists a $C > 0$ independent of $\boldsymbol{\theta}$ such that

$$E\left\{\left|\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right|^k \mid \mathbf{U}_t, \mathbf{X}_t\right\} = F_{y|u,x}\left(c_{4t}\right) - F_{y|u,x}\left(c_{3t}\right) \le Ca_n\left|\boldsymbol{\theta}'\mathbf{X}_t^*\right|,$$

which implies that

$$\begin{aligned} E\left|V_{n11}^{(i)}\right|^{\delta} &= E\left[\left|\psi_\tau\left(Y_{n1}^*\right) - \psi_\tau\left(Y_1^*\right)\right|^{\delta}\left|X_{1i}\right|^{\delta}K^{\delta}\left(U_{1h}\right)\right] \\ &\le Ca_nE\left[\left|\boldsymbol{\theta}'\mathbf{X}_t^*\right|\left|X_{1i}\right|^{\delta}K^{\delta}\left(U_{1h}\right)\right] \le Ca_nh \end{aligned}$$

uniformly in $\boldsymbol{\theta}$ over A_m by Assumption (C6). Then,

$$J_3 \le Ca_n^{2/\delta}h^{2/\delta-1}\sum_{s=d_n}^{\infty}[\alpha(s)]^{1-2/\delta} \le Ca_n^{2/\delta}h^{2/\delta-1}d_n^{-l}\sum_{s=d_n}^{\infty}s^l[\alpha(s)]^{1-2/\delta} = o\left(a_n^{2/\delta}h^{2/\delta-1}d_n^{-l}\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. As for J_2, we use Assumption (C10) to get

$$\left|\mathrm{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right| \le C\left[E\left\{\left|X_{1i}X_{(s+1)i}\right|K\left(U_{1h}\right)K\left(U_{(s+1)h}\right)\right\} + a_n^2h^2\right] = O\left(h^2\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. It follows that $J_2 = O\left(d_nh\right)$ uniformly in $\boldsymbol{\theta}$ over A_m. Analogously,

$$J_1 = h^{-1}\mathrm{Var}\left(V_{n11}^{(i)}\right) \le h^{-1}E\left(V_{n11}^{(i)}\right)^2 = O\left(a_n\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. By choosing d_n such that $d_n^lh^{1-2/\delta} = c$, then, $d_nh \to 0$ and $\mathrm{Var}\left(V_n^{(1i)}\right) = o(1)$. Therefore, $V_n^{(1i)} = o_p(1)$ so that $V_n^{(1)} = o_p(1)$ uniformly in $\boldsymbol{\theta}$ over A_m. By the same token, we can show that $V_n^{(2)} = o_p(1)$ uniformly in $\boldsymbol{\theta}$ over A_m. This completes the proof of the lemma. □

Proof of Lemma 3.4. It is easy to justify that

$$\begin{aligned} E\left[V_n(\theta) - V_n(0)\right] &= na_nE\left[\left(\psi_\tau\left(Y_t^* - a_n\boldsymbol{\theta}'\mathbf{X}_t^*\right) - \psi_\tau\left(Y_t^*\right)\right)\right]\mathbf{X}_t^*K\left(U_{th}\right)] \\ &= na_nE\left[\left\{F_{y|u,x}\left(c_{2t}\right) - F_{y|u,x}\left(c_{2t} + a_n\boldsymbol{\theta}'\mathbf{X}_t^*\right)\right\}\mathbf{X}_t^*K\left(U_{th}\right)\right] \\ &\approx -\frac{1}{h}E\left[f_{y|u,x}\left(c_{2t}\right)\mathbf{X}_t^*\mathbf{X}_t^{*\prime}K\left(U_{th}\right)\right]\boldsymbol{\theta} \\ &\approx -f_u\left(u_0\right)\Omega_1^*\left(u_0\right)\boldsymbol{\theta} \end{aligned}$$

uniformly in $\boldsymbol{\theta}$ over A_m by Assumption (C3). The proof of the lemma is complete. □

Proof of Lemma 3.5. Observe by Taylor expansions and Assumption (C3) that

$$
\begin{aligned}
E\left[\mathbf{Z}_t\right] &= E\left[\left\{\tau - F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^{*}K\left(U_{th}\right)\right] \\
&\approx E\left[\left\{F_{y|u,x}\left(c_{2t} + \mathbf{X}_t'\mathbf{a}''\left(u_0\right)h^2U_{th}^2/2\right) - F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^{*}K\left(U_{th}\right)\right] \\
&\approx \frac{h^2}{2}E\left[f_{y|u,x}\left(c_{2t}\right)\mathbf{X}_t^{*}\mathbf{X}_t'\mathbf{a}''\left(u_0\right)U_{th}^2K\left(U_{th}\right)\right] \\
&\approx \frac{h^2}{2}E\left[f_{y|u,x}\left(q_\tau\left(u_0,\mathbf{X}_t\right)\right)\mathbf{X}_t^{*}\mathbf{X}_t'\mathbf{a}''\left(u_0\right)U_{th}^2K\left(U_{th}\right)\right] \\
&\approx \frac{h^3f_u\left(u_0\right)}{2}\begin{pmatrix}\mu_2\Omega^{*}\left(u_0\right)\mathbf{a}''\left(u_0\right) \\ \mathbf{0}\end{pmatrix}.
\end{aligned} \tag{3.21}
$$

Also, we have

$$
\begin{aligned}
\operatorname{Var}\left[\mathbf{Z}_t\right] &= E\left[\left\{\tau - I\left(Y_t < c_{2t}\right)\right\}^2\mathbf{X}_t^{*}\mathbf{X}_t^{*\prime}K^2\left(U_{th}\right)\right] \\
&\approx E\left[\left\{\tau^2 - 2\tau F_{y|u,x}\left(c_{2t}\right) + F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^{*}\mathbf{X}_t^{*\prime}K^2\left(U_{th}\right)\right] \\
&\approx \tau(1-\tau)E\left[\mathbf{X}_t^{*}\mathbf{X}_t^{*\prime}K^2\left(U_{th}\right)\right] \\
&\approx \tau(1-\tau)hf_u\left(u_0\right)\Omega_1\left(u_0\right).
\end{aligned} \tag{3.22}
$$

Next, we show that the last part of lemma holds true. Clearly, $V_n(0) = a_n\sum_{t=1}^n \mathbf{Z}_t$. Similar to the proof of Lemma 3.3, we have

$$
\begin{aligned}
\operatorname{Var}\left[V_n(0)\right] &= \frac{1}{h}\operatorname{Var}\left(\mathbf{Z}_1\right) + \frac{2}{h}\sum_{s=1}^{d_n-1}\left(1-\frac{s}{n}\right)\operatorname{Cov}\left(\mathbf{Z}_1,\mathbf{Z}_{s+l}\right) + \frac{2}{h}\sum_{s=d_n}^{n}\left(1-\frac{s}{n}\right)\operatorname{Cov}\left(\mathbf{Z}_1,\mathbf{Z}_{s+l}\right) \\
&\equiv J_4 + J_5 + J_6
\end{aligned}
$$

for some $d_n \to \infty$ specified later. By (3.22),

$$
J_4 \quad \to \quad \tau(1-\tau)f_u\left(u_0\right)\Omega_1\left(u_0\right).
$$

Therefore, it suffices to show that $|J_5| = o(1)$ and $|J_6| = o(1)$. For J_6, using the Davydov's inequality (see, e.g., Lemma 1.1) and the boundedness of $\psi_\tau(\cdot)$ to obtain

$$
\left|\operatorname{Cov}\left(\mathbf{Z}_1,\mathbf{Z}_{s+1}\right)\right| \le C\alpha^{1-2/\delta}(s)\left[E\left|\mathbf{Z}_1\right|^\delta\right]^{2/\delta} \le Ch^{2/\delta}\alpha^{1-2/\delta}(s),
$$

which gives

$$
J_6 \le Ch^{2/\delta-1}\sum_{s=d_n}^{\infty}[\alpha(s)]^{1-2/\delta} \le Ch^{2/\delta-1}d_n^{-l}\sum_{s=d_n}^{\infty}s^l[\alpha(s)]^{1-2/\delta} = o\left(h^{2/\delta-1}d_n^{-l}\right) = o(1)
$$

by choosing d_n to satisfy $d_n^l h^{1-2/\delta} = c$. As for J_5, we use Assumption (C10) and (3.21) to get

$$|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{s+1}\right)| \leq C\left[E\left\{\left|\mathbf{X}_1^* \mathbf{X}_{s+1}^*\right| K\left(U_{1h}\right) K\left(U_{(s+1)h}\right)\right\} + h^6\right] = O\left(h^2\right)$$

so that $J_5 = O\left(d_n h\right) = o(1)$ by the choice of d_n. We finish the proof of this lemma. □

Proof of (3.9) **and** (3.10). By the Taylor expansion,

$$E\left[\xi_t \mid U_t, \mathbf{X}_t\right] = F_{y|u,x}\left(\mathbf{X}_t' \mathbf{a}\left(u_0\right) + a_n\right) - F_{y|u,x}\left(\mathbf{X}_t' \mathbf{a}\left(u_0\right)\right) \approx f_{y|u,x}\left(\mathbf{X}_t' \mathbf{a}\left(u_0\right)\right) a_n.$$

Therefore,

$$E\left[\mathbf{S}_n\right] \approx h^{-1} E\left[f_{y|u,x}\left(\mathbf{X}_t' \mathbf{a}\left(u_0\right)\right) \mathbf{X}_t^* \mathbf{X}_t^{*\prime} K\left(U_{th}\right)\right] \approx f_u\left(u_0\right) \Omega_1^*\left(u_0\right).$$

Similar to the proof of Var $[V_n(0)]$ in Lemma 3.5, one can show that Var $\left(\mathbf{S}_n\right) \rightarrow 0$. Therefore, $\mathbf{S}_n \rightarrow f_u\left(u_0\right) \Omega_1^*\left(u_0\right)$ in probability. This proves (3.9). Clearly,

$$E\left[\widehat{\Omega}_{n,0}\right] = E\left[\mathbf{X}_t \mathbf{X}_t' K_h\left(U_t - u_0\right)\right] = \int \Omega\left(u_0 + hv\right) f_u\left(u_0 + hv\right) K(v) dv \approx f_u\left(u_0\right) \Omega\left(u_0\right).$$

Similarly, one can show that Var $\left(\widehat{\Omega}_{n,0}\right) \rightarrow 0$. This proves the first part of (3.10). By the same token, one can show that $E\left[\widehat{\Omega}_{n,1}\right] \approx f_u\left(u_0\right) \Omega^*\left(u_0\right)$ and Var $\left(\widehat{\Omega}_{n,1}\right) \rightarrow 0$. Thus, $\widehat{\Omega}_{n,1} = f_u\left(u_0\right) \Omega^*\left(u_0\right) + o_p(1)$. We prove (3.10). □

3.6 Computer Codes

Please see the files chapter3-1.r, chapter3-2.r, and chapter3-3.r for making figures. If you want to learn the codes for computation, they are available upon request.

3.7 References

An, H.Z and Chen, S.G. (1997). A Note on the Ergodicity of Nonlinear Autoregressive Models. *Statistics & Probability Letters*, **34**(4), 365-372.

An, H.Z. and Huang, F.C. (1996). The Geometrical Ergodicity of Nonlinear Autoregressive Models. *Statistica Sinica*, **6**, 943-956.

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, **77**(4), 669-687.

Bao, Y., Lee, T.-H. and Saltog lu, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *Journal of Forecasting*, **25**(2), 101-128.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformation for multiple regression and correlation. *Journal of the American Statistical Association*, **80**(391), 580-598.

Cai, Z. (2002a). Regression quantile for time series. *Econometric Theory*, **18**(1), 169-192.

Cai, Z. (2002b). A two-stage approach to additive time series models. *Statistica Neerlandica*, **56**(4), 415-433.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1), 163-188

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**(451), 941-956.

Cai, Z., T. Juhl and B. Yang (2015). Functional index coefficient models with variable selection. *Journal of Econometrics*, **189**(2), 272-284.

Cai, Z. and Masry, E. (2000). Nonparametric estimation in nonlinear ARX time series models: Local linear fitting and projections.*Econometric Theory*, **16**(4), 465-501.

Cai, Z., Y. Ren and B. Yang (2015). A semiparametric conditional capital asset pricing model. *Journal of Banking and Finance*, **61**(1), 117-126.

Cai, Z. and Tiwari, R.C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11** (3), 341-350.

Cai, Z. and Xiao, Z. (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics*, **167**(2), 413-425.

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, **103**(484), 1595-1608.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, **19**(2), 760-777.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *The Annuals of Statistics*, **25**(2), 715-744.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**(421), 298-308.

Cole, T.J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, **13**(23-24), 2477-2492.

De Gooijer, J. G. and Zerom, D. (2003). On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, **98**(461), 135-146.

Duffie, D. and Pan, J. (1997). An overview of value at risk. *Journal of Derivatives*, **4**(3), 7-49.

Durrett, R. (2019). *Probability: Theory and Examples*, Fifth Edition. Cambridge university press, New York.

Engle, R. F., Ito, T. and Lin, W.-L. (1990). Meteor showers or heat waves? Heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica*, **58**(3), 525-542.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of business & economic statistics*, **22**(4), 367-381.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, **1**(1), 93-125.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**(3), 645-660.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B*, **65**(1), 57-80.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**(1), 189-206.

Gallant, A. R., Hsieh, D. A. and Tauchen, G. E. (1991). On fitting a recalcitrant series: The pound/dollar exchange rate, 1974-1983. In *Nonparametric And Semiparametric Methods in Econometrics and Statistics (W.A. Barnett, J. Powell and G.E. Tauchen, eds.)*, pp.199-240. Cambridge University Press, Cambridge.

Gilley, O. W., Pace, R. K., et al. (1996). On the Harrison and Rubinfeld Data. *Journal of Environmental Economics and Management*, **31**(3), 403-405.

Gorodetskii, V. V. (1977). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications*, **22**(2), 411-413.

Granger, C. W. J., White, H. and Kamstra, M. (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, **40**(1), 87-96.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Applications.* Academic Press, New York.

Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and demand for clean air. *Journal of Environmental Economics and Management*, **5**(1), 81-102.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

He, X. and Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, **75**(2), 343-352.

He, X., Ng, P. and Portnoy, S. (1998). Bivariate quantile smoothing splines.*Journal of the Royal Statistical Society: Series B*, **60**(3), 537-550.

He, X. and Portnoy, S. (2000). Some asymptotic results on bivariate quantile splines. *Journal of Statistical Planning and Inference*, **91**(2), 341-349.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for α-mixing processes. *Annals of the Institute of Statistical Mathematics*, **52**(3), 459-470.

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inferences*, **121**(1), 113-125.

Hong, Y. and Lee, T.-H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, **85**(4), 1048-1062.

Horowitz, J.L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, **100**(472), 1238-1249.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, **60**(2), 271-293.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297-307.

Jorion, P. (2000). *Value at Risk*, 2ed. McGraw Hill, New York.

Jurekova, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression model. *Annals of Statistics*, **5**(3), 464-472.

Khindanova, I.N. and Rachev, S.T. (2000). Value at risk: Recent advances. *Handbook on Analytic-Computational Methods in Applied Mathematics*, CRC Press LLC.

Koenker, R. (1994). Confidence intervals for regression quantiles. *In Proceedings of the Fifth Prague Symposium on Asymptotic Statistics* (P. Mandl and M. Huskova, eds.), 349-359. Physica, Heidelberg.

Koenker, R. (2004). *Quantreg: An R package for quantile regression and related methods.* The Comprehensive R Archive Network website.

Koenker R. (2000). Galton, Edgeworth, Frisch and prospects for quantile regression in econometrics. *Journal of Econometrics*, **9 5**, 347-374.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, **46**(1), 33-50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**(1), 43-61.

Koenker, R. and Hallock, K. (2000). Quantile regression: An introduction. *Journal of Economic Perspectives*, **15**, 143-157.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**(4) , 673-680.

Koenker, R. and Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, **70**, 1583-1612.

Koenker, R. and Xiao, Z. (2004). Unit root quantile autoregression inference. *Journal of American Statistical Association*, **99**(467), 775-787.

Koenker, R. and Zhao, Q. (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory*, **12**(5), 793-813.

LeBaron, B. (1997). Technical trading rules and regime shifts in foreign exchange. *Advanced Trading Rules*, 5-40.

LeBaron, B. (1999). Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics*, **49**(1), 125-143.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, **26**(4), 423-434.

Lu, Z. (1998). On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, **8**(4), 1205-1217.

Lu, Z., Hui, Y. and Zhao, Q. (1998). Local linear quantile regression under dependence: Bahadur representation and application. *Working Paper*, Department of Management Sciences, City University of Hong Kong.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series strong convergence and asymptotic normality: Strong convergence and asymptotic normality. *Econometric theory*, **11**(2), 258-289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**(2), 214-252.

Machado, J.A.F. (1993). Robust model selection and M-estimation. *Econometric Theory*, **9**(3), 478-493.

Morgan, J.P. (1995). *Riskmetrics Technical Manual*, 3rd edition.

Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, **93**(442), 605-619.

Pace, R. K. and Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics*, **14**(3), 333-340.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **75**(372), 828-838.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Sentürk, D. and Müller, H.-G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Annals of Statistics*, **34**(2), 654-679.

Sercu, P., Uppal, R., et al. (2006). *Exchange Rate Volatility, Trade, and Capital Flows under Alternative Rate Regimes.* Cambridge University Press, New York.

Taylor, J. W. and Bunn, D. W. (1999). A quantile regression approach to generating prediction intervals. *Management Science*, **45**(2), 225-237.

Tsay, R.S. (2002). *Analysis of Financial Time Series.* John Wiley & Sons, New York.

Wang, K. Q. (2003). Asset pricing with conditioning information: A new test. *Journal of Finance*, **58**(1), 161-196.

Wei, Y. and He, X. (2006). Conditional growth charts (with discussion). *Annals of Statistics*, **34**(5), 2069-2097.

Wei, Y., Pere, A., Koenker, R. and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, **25**(8), 1369-1382.

Withers, C. S. (1981). Conditions for linear processes to be strong-mixing. *Zeitschrift fur Wahrscheinlichkeitstheorie verwandte Gebiete*, **57**(4), 477-480.

Yu, K. and Jones, M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**(411), 228-237.

Yu, K. and Lu, Z. (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics*, **31**(3), 333-346.

Xu, X. (2005). *Semiparametric Quantile Dynamic Time Series Models and Their Applications. Ph.D. Dissertation*, University of North Carolina at Charlotte.

Zhou, K. Q. and Portnoy, S. L. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics*, **24**(1), 287-306.

Chapter 4

Conditional VaR and Expected Shortfall

For details, see the paper by Cai and Wang (2008). If you like to read the whole paper, you can download it from the web site *Journal of Econometrics*.

4.1 Introduction

The value-at-risk (VaR) and expected shortfall (ES) have become two popular measures on market risk associated with an asset or a portfolio of assets during the last decade. In particular, VaR has been chosen by the Basle Committee on Banking Supervision as the benchmark of risk measures for capital requirements and both of them have been used by financial institutions for asset managements and minimization of risk as well as have been developed rapidly as analytic tools to assess riskiness of trading activities. See, to name just a few, Morgan (1996), Duffie and Pan (1997), Jorion (2001, 2003), and Duffie and Singleton (2003) for the financial background, statistical inferences, and various applications. In terms of the formal definition, VaR is simply a quantile of the loss distribution (future portfolio values) over a prescribed holding period (e.g., 2 weeks) at a given confidence level, while ES is the expected loss, given that the loss is at least as large as some given quantile of the loss distribution (e.g., VaR). It is well known from Artzner, Delbaen, Eber and Heath (1999) that ES is a coherent risk measure such as it satisfies the following four axioms:

- homogeneity: increasing the size of a portfolio by a factor should scale its risk measure by the same factor,

- monotonicity: a portfolio must have greater risk if it has systematically lower values than another,

- risk-free condition or translation invariance: adding some amount of cash to a portfolio should reduce its risk by the same amount, and
- subadditivity: the risk of a portfolio must be less than the sum of separate risks or merging portfolios cannot increase risk.

VaR satisfies homogeneity, monotonicity, and risk-free condition but is not sub-additive. See Artzner, et al. (1999) for details. As advocated by Artzner, et al. (1999), ES is preferred due to its better properties although VaR is widely used in applications.

Measures of risk might depend on the state of the economy since economic and market conditions vary from time to time. This requires risk managers should focus on the conditional distributions of profit and loss, which take full account of current information about the investment environment (macroeconomic and financial as well as political) in forecasting future market values, volatilities, and correlations. As pointed out by Duffie and Singleton (2003), not only are the prices of the underlying market indices changing randomly over time, the portfolio itself is changing, as are the volatilities of prices, the credit qualities of counter-parties, and so on. On the other hand, one would expect the VaR to increase as the past returns become very negative, because one bad day makes the probability of the next somewhat greater. Similarly, very good days also increase the VaR, as would be the case for volatility models. Therefore, VaR could depend on the past returns in someway. Hence, an appropriate risk analytical tool or methodology should be allowed to adapt to varying market conditions and to reflect the latest available information in a time series setting rather than the iid framework. Most of the existing risk management literature has concentrated on unconditional distributions and the iid setting although there have been some studies on the conditional distributions and time series data. For more background, see Chernozhukov and Umanstev (2001), Cai (2002), Fan and Gu (2003), Engle and Manganelli (2004), Cai and Xu (2008), Scaillet (2005), and Cosma, Scaillet and von Sachs (2007), and references therein for conditional models, and Duffie and Pan (1997), Artzner, et al. (1999), Rockafellar and Uryasev (2000), Acerbi and Tasche (2002), Frey and McNeil (2002), Scaillet (2004), Chen and Tang (2005), Chen (2008), and among others for unconditional models. Also, most of studies in the literature and applications are limited to parametric models, such as all standard industry models like CreditRisk$^+$, CreditMetrics, CreditPortfolio View and the model proposed by the KMV corporation. See Chernozhukov and Umanstev (2001), Frey and McNeil (2002), Engle and Manganelli (2004), and references therein on parametric models in practice and Fan and Gu (2003) and references therein for semiparametric models.

The main focus of this chapter is on studying the conditional value-at-risk (CVaR) and conditional expected shortfall (CES) and proposing a new nonparametric estimation procedure to estimate CVaR and CES functions where the conditional information is allowed to contain economic and market (exogenous) variables and the past observed returns. Parametric models for CVaR and CES can be most efficient if the underlying functions are correctly specified. See Chernozhukov and Umanstev (2001) for a polynomial type regression model and Engle and Manganelli (2004) for a GARCH type parametric model for CVaR based on regression quantile. However, a misspecification may cause serious bias and model constraints may distort the underlying distributions. A nonparametric modeling is appealing in several aspects. One of the advantages for nonparametric modeling is that little or no restrictive prior information on functionals is needed. Further, it may provide a useful insight for further parametric fitting.

The approach proposed by Cai and Wang (2008) has several advantages. The first one is to propose a new nonparametric approach to estimate CVaR and CES. In essence, our estimator for CVaR is based on inverting a newly proposed estimator of the conditional distribution function for time series data and the estimator for CES is by a plugging-in method based on plugging in the estimated conditional probability density function and the estimated CVaR function. Note that they are analogous to the estimators studied by Scaillet (2005) by using the Nadaraya-Watson (NW) type double kernel (smoothing in both the y and x directions) estimation, and Cai (2002) by utilizing the weighted NadarayaWatson (WNW) kernel type technique to avoid the so-called boundary effects as well as Yu and Jones (1998) by employing the double kernel local linear method. More precisely, our newly proposed estimator combines the WNW method of Cai (2002) and the double kernel local linear technique of Yu and Jones (1998), termed as *weighted double kernel local linear* (WDKLL) estimator.

The second merit is to establish the asymptotic properties for the WDKLL estimators of the conditional probability density function and cumulative distribution function for the α-mixing time series at both boundary and interior points. It is therefore shown that the WDKLL method enjoys the same convergence rates as those of the double kernel local linear estimator of Yu and Jones (1998) and the WNW estimator of Cai (2002). It is also shown that the WDKLL estimators have desired sampling properties at both boundary and interior points of the support of the design density, which seems to be seminal. Finally, we derive the WDKLL estimator of CVaR by inverting the WDKLL conditional distribution estimator and the WDKLL estimator of CES by plugging in the WDKLL estimators of PDF and CVaR. We show that the WDKLL estimator of CVaR exists always due to the WDKLL estimator of CDF being a distribution function itself, and that it inherits all better properties from the WDKLL estimator of CDF; that is, the WDKLL estimator of CDF is a CDF and differentiable, and it possess the asymptotic properties such as design adaption, avoiding boundary effects, and mathematical efficiency. Note that to preserve shape constraints, recently, Cosma, Scaillet and von Sachs (2007) used a wavelet method to estimate conditional probability density and cumulative distribution functions and then to estimate conditional quantiles.

Note that CVaR defined here is essentially the conditional quantile or quantile regression of Koenker and Bassett (1978), based on the conditional distribution, rather than CVaR defined in some risk management literature (see, e.g., Rockafellar and Uryasev, 2000; Jorion, 2001, 2003), which is what we call ES here. Also, note that the ES here is called TailVaR in Artzner, et al. (1999). Moreover, as aforementioned, CVaR can be regarded as a special case of quantile regression. See Cai and Xu (2008) for the state-of-the-art about current research on nonparametric quantile regression, including CVaR. Further, note that both ES and CES have been known for decades among actuary sciences and they are very popular in insurance industry. Indeed, they have been used to assess risk on a portfolio of potential claims, and to design reinsurance treaties. See the book by Embrechts, Kluppelberg, and Mikosch (1997) for the excellent review on this subject and the papers by McNeil (1997), Hürlimann (2003), Scaillet (2005), and Chen (2008). Finally, ES or CES is also closely related to other applied fields such as the mean residual life function in reliability and the biometric function in biostatistics. See Oakes and Dasu (1990) and Cai and Qian (2000) and references therein.

4.2 Setup

Assume that the observed data $\{(X_t, Y_t);1\le t\le n\}, X_t\in\Re^d$, are available and they are observed from a stationary time series model. Here Y_t is the risk or loss variable which can be the negative logarithm of return (log loss) and X_t is allowed to include both economic and market (exogenous) variables and the lagged variables of Y_t and also it can be a vector. But, for the expositional purpose, we consider only the case when X_t is a scalar ($d = 1$). Note that the proposed methodologies and their theory for the univariate case ($d = 1$) continue to hold for multivariate situations ($d > 1$). Extension to the case $d > 1$ involves no fundamentally new ideas. Note that models with large d are often not practically useful due to "curse of dimensionality".

We now turn to considering the nonparametric estimation of the conditional expected shortfall $\mu_p(x)$, which is defined as

$$\mu_p(x) = E\left[Y_t \mid Y_t \geq \nu_p(x), X_t = x\right],$$

where $\nu_p(x)$ is the conditional value-at-risk, which is defined as the solution of

$$P\left(Y_t \geq \nu_p(x) \mid X_t = x\right) = S\left(\nu_p(x) \mid x\right) = p$$

or expressed as $\nu_p(x) = S^{-1}(p \mid x)$, where $S(y \mid x)$ is the conditional survival function of Y_t given $X_t = x$; $S(y \mid x) = 1 - F(y \mid x)$, and $F(y \mid x)$ is the conditional cumulative distribution function. It is easy to see that

$$\mu_p(x) = \int_{\nu_p(x)}^{\infty} y f(y \mid x) dy/p,$$

where $f(y \mid x)$ is the conditional probability density function of Y_t given $X_t = x$. To estimate $\mu_p(x)$, one can use the plugging-in method as

$$\widehat{\mu}_p(x) = \int_{\widehat{\nu}_p(x)}^{\infty} y \widehat{f}(y \mid x) dy/p, \tag{4.1}$$

where $\widehat{\nu}_p(x)$ is a nonparametric estimation of $\nu_p(x)$ and $\widehat{f}(y \mid x)$ is a nonparametric estimation of $f(y \mid x)$. But the bandwidths for $\widehat{\nu}_p(x)$ and $\widehat{f}(y \mid x)$ are not necessary to be same.

Note that Scaillet (2005) used the NW type double kernel method to estimate $f(y \mid x)$ first, due to Roussas (1969), denoted by $\widetilde{f}(y \mid x)$, and then estimated $\nu_p(x)$ by inverting the estimated conditional survival function, denoted by $\widetilde{\nu}_p(x)$, and finally estimated $\mu_p(x)$ by plugging $\widetilde{f}(y \mid x)$ and $\widetilde{\nu}_p(x)$ into (4.1), denoted by $\widetilde{\mu}_p(x)$, where $\widetilde{\nu}_p(x) = \widetilde{S}^{-1}(y \mid x)$ and $\widetilde{S}(y \mid x) = \int_y^{\infty} \widetilde{f}(u \mid x) du$. But, it is well documented (see, e.g., Fan and Gijbels, 1996) that the NW kernel type procedures have serious drawbacks: the asymptotic bias involves the design density so that they can not be adaptive, and boundary effects exist so that they require boundary modifications. In particular, boundary effects might cause a serious problem for estimating $\nu_p(x)$ since it is only concerned with the tail probability. The question is now how to provide a better estimate for $f(y \mid x)$ and $\nu_p(x)$ so that we have a good estimate for $\mu_p(x)$. Therefore, we address this issue in the next section.

4.3 Nonparametric Estimating Procedures

We start with the nonparametric estimators for the conditional density function and its distribution function first and then turn to discussing the nonparametric estimators for the conditional VaR and ES functions.

There are several methods available for estimating $\nu_p(x), f(y \mid x)$, and $F(y \mid x)$ in the literature, such as kernel and nearest-neighbor[1]. To attenuate these drawbacks of the kernel type estimators mentioned in Section 4.2 recently, some new methods have been proposed to estimate conditional quantiles. The first one, a more direct approach, by using the "check" function such as the robustified local linear smoother, was provided by Fan, Hu, and Troung (1994) and further extended by Yu and Jones (1997,1998) for iid data. A more general nonparametric setting was explored by Cai and Xu (2008) for time series data. This modeling idea was initialed by Koenker and Bassett (1978) for linear regression quantiles and Fan, Hu, and Troung (1994) for nonparametric models. See Cai and Xu (2008) and references therein for more discussions on models and applications. An alternative procedure is first to estimate the conditional distribution function by using double kernel local linear technique of Fan, Yao, and Tong (1996) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile or CVaR. Yu and Jones (1997, 1998) compared these two methods theoretically and empirically and suggested that the double kernel local linear would be better.

4.3.1 Estimation of Conditional PDF and CDF

To make a connection between the conditional density (distribution) function and nonparametric regression problem, it is noted by the standard kernel estimation theory (see, e.g.,

[1] To name just a few, see Lejeune and Sarda (1988), Troung (1989), Samanta (1989), and Chaudhuri (1991) for iid errors, Roussas (1969) and Roussas (1991) for Markovian processes, and Troung and Stone (1992) and Boente and Fraiman (1995) for mixing sequences.

Fan and Gijbles, 1996) that for a given symmetric density function $K(\cdot)$,

$$E\left\{K_{h_0}\left(y-Y_t\right) \mid X_t=x\right\}=f(y \mid x)+\frac{h_0^2}{2} \mu_2(K) f^{2,0}(y \mid x)+o\left(h_0^2\right) \approx f(y \mid x), \text{ as } h_0 \quad \rightarrow \quad 0, \tag{4.2}$$

where $K_{h_0}(u)=K\left(u / h_0\right) / h_0, \mu_2(K)=\int_{-\infty}^{\infty} u^2 K(u) d u, f^{2,0}(y \mid x)=\partial^2 / \partial y^2 f(y \mid x)$, and $\approx$ denotes an approximation by ignoring the higher terms. Note that $Y_t^*(y)=K_{h_0}\left(y-Y_t\right)$ can be regarded as an initial estimate of $f(y \mid x)$ smoothing in the y direction. Also, note that this approximation ignores the higher order terms $O\left(h_0^j\right)$ for $j \geq 2$, since they are negligible if $h_0=o(h)$, where h is the bandwidth used in smoothing in the x direction (see (4.3) below). Therefore, the smoothing in the y direction is not important in the context of this subject so that intuitively, it should be under-smoothed. Thus, the left hand side of (4.2) can be regraded as a nonparametric regression of the observed variable $Y_t^*(y)$ versus X_t and the local linear (or polynomial) fitting scheme of Fan and Gijbles (1996) can be applied to here. This leads us to consider the following locally weighted least squares regression problem:

$$\sum_{t=1}^n\left\{Y_t^*(y)-a-b\left(X_t-x\right)\right\}^2 W_h\left(x-X_t\right), \tag{4.3}$$

where $W(\cdot)$ is a kernel function and $h=h(n)>0$ is the bandwidth satisfying $h \rightarrow 0$ and $n h \rightarrow \infty$ as $n \rightarrow \infty$, which controls the amount of smoothing used in the estimation. Note that (4.3) involves two kernels $K(\cdot)$ and $W(\cdot)$. This is the reason of calling "double kernel".

Minimizing the above locally weighted least squares in (4.3) with respect to a and b, we obtain the locally weighted least squares estimator of $f(y \mid x)$, denoted by $\widehat{f}(y \mid x)$, which is $\widehat{a}$. From Fan and Gijbels (1996) or Fan, Yao and Tong (1996), $\widehat{f}(y \mid x)$ can be re-expressed as a linear estimator form as

$$\widehat{f}_{ll}(y \mid x)=\sum_{t=1}^n W_{ll, t}(x, h) Y_t^*(y),$$

where with $S_{n, j}(x)=\sum_{t=1}^n W_h\left(x-X_t\right)\left(X_t-x\right)^j$, the weights $\left\{W_{ll, t}(x, h)\right\}$ are given by

$$W_{ll, t}(x, h)=\frac{\left[S_{n, 2}(x)-\left(x-X_t\right) S_{n, 1}(x)\right] W_h\left(x-X_t\right)}{S_{n, 0}(x) S_{n, 2}(x)-S_{n, 1}^2(x)} .$$

Clearly, $\left\{W_{ll, t}(x, h)\right\}$ satisfy the so-called discrete moments conditions as follows: for $0 \leq j \leq 1$,

$$\sum_{t=1}^n W_{ll, t}(x, h)\left(X_t-x\right)^j=\delta_{0, j}=\begin{cases}1 & \text{if } j=0 \\ 0 & \text{otherwsie}\end{cases} \tag{4.4}$$

based on the least squares theory; see (3.12) of Fan and Gijbels (1996, p.63). Note that the estimator $\widehat{f}_{ll}(y \mid x)$ can range outside $[0, \infty)$. The double kernel local linear estimator of $F(y \mid x)$ is constructed (see (8) of Yu and Jones (1998)) by integrating $\widehat{f}_{ll}(y \mid x)$

$$\widehat{F}_{ll}(y \mid x) = \int_{-\infty}^{y} \widehat{f}_{ll}(y \mid x) dy = \sum_{t=1}^{n} W_{ll,t}(x, h) G_{h_0}(y - Y_t),$$

where $G(\cdot)$ is the distribution function of $K(\cdot)$ and $G_{h_0}(u) = G(u/h_0)$. Clearly, $\widehat{F}_{ll}(y \mid x)$ is continuous and differentiable with respect to y with $\widehat{F}_{ll}(-\infty \mid x) = 0$ and $\widehat{F}_{ll}(\infty \mid x) = 1$. Note that the differentiability of the estimated distribution function can make the asymptotic analysis much easier for the nonparametric estimators of CVaR and CES (see later).

Although Yu and Jones (1998) showed that the double kernel local linear estimator has some attractive properties such as no boundary effects, design adaptation, and mathematical efficiency (see, e.g., Fan and Gijbels, 1996), it has the disadvantage of producing conditional distribution function estimators that are not constrained either to lie between zero and one or to be monotone increasing, which is not good for estimating CVaR if the inverting method is used. In both these respects, the NW method is superior, despite its rather large bias and boundary effects. The properties of positivity and monotonicity are particularly advantageous if the method of inverting conditional distribution estimator is applied to produce the estimator of a conditional quantile or CVaR. To overcome these difficulties, Hall, Wolff, and Yao (1999) and Cai (2002) proposed the WNW estimator based on an empirical likelihood principle, which is designed to possess the superior properties of local linear methods such as bias reduction and no boundary effects, and to preserve the property that the NW estimator is always a distribution function, although it might require more computational efforts since it requires estimating and optimizing additional weights aimed at the bias correction. Cai (2002) discussed the asymptotic properties of the WNW estimator at both interior and boundary points for the mixing time series under some regularity assumptions and showed that the WNW estimator has a better performance than other competitors. See Cai (2002) for details. Recently, Cosma, Scaillet and von Sachs (2007) proposed a shape preserving estimation method to estimate cumulative distribution functions and probability density functions using the wavelet methodology for multivariate dependent data and then to estimate a conditional quantile or CVaR.

The WNW estimator of the conditional distribution $F(y \mid x)$ of Y_t given $X_t = x$ is defined by

$$\widehat{F}_{c1}(y \mid x) = \sum_{t=1}^{n} W_{c,t}(x,h) I\left(Y_t \le y\right), \tag{4.5}$$

where the weights $\{W_{c,t}(x,h)\}$ are given by

$$W_{c,t}(x,h) = \frac{p_t(x) W_h\left(x - X_t\right)}{\sum_{t=1}^{n} p_t(x) W_h\left(x - X_t\right)}, \tag{4.6}$$

and $\{p_t(\mathbf{x})\}$ is chosen to be $p_t(x) = n^{-1}\left\{1 + \lambda\left(X_t - x\right) W_h\left(x - X_t\right)\right\}^{-1} \ge 0$ with λ, a function of data and x, uniquely defined by maximizing the logarithm of the empirical likelihood

$$L_n(\lambda) = -\sum_{t=1}^{n} \log\left\{1 + \lambda\left(X_t - x\right) W_h\left(x - X_t\right)\right\}$$

subject to the constraints $\sum_{t=1}^{n} p_t(x) = 1$ and the discrete moments conditions in (4.4); that is,

$$\sum_{t=1}^{n} W_{c,t}(x,h)\left(X_t - x\right)^j = \delta_{0,j} \tag{4.7}$$

for $0 \le j \le 1$. Also, see Cai (2002) for details on this aspect. In implementation, Cai (2002) recommended using the Newton-Raphson scheme to find the root of equation $L_n'(\lambda) = 0$. Note that $0 \le \widehat{F}_{c1}(y \mid x) \le 1$ and it is monotone in y. But $\widehat{F}_{c1}(y \mid x)$ is not continuous in y and of course, not differentiable in y either. Note that under regression setting, Cai (2001) provided a comparison of the local linear estimator and the WNW estimator and discussed the asymptotic minimax efficiency of the WNW estimator.

To accommodate all nice properties (monotonicity, continuity, differentiability, and lying between zero and one) and the attractive asymptotic properties (design adaption, avoiding boundary effects, and mathematical efficiency, see Cai (2002) for detailed discussions) of both estimators $\widehat{F}_{ll}(y \mid x)$ and $\widehat{F}_{c1}(y \mid x)$ under a unified framework, we propose the following nonparametric estimators for the conditional density function $f(y \mid x)$ and its conditional distribution function $F(y \mid x)$, termed as *weighted double kernel local linear estimation*,

$$\widehat{f}_c(y \mid x) = \sum_{t=1}^{n} W_{c,t}(x,h) Y_t^*(y),$$

where $W_{c,t}(x,h)$ is given in (4.6), and

$$\widehat{F}_c(y \mid x) = \int_{-\infty}^{y} \widehat{f}_c(y \mid x) dy = \sum_{t=1}^{n} W_{c,t}(x,h) G_{h_0}\left(y - Y_t\right). \tag{4.8}$$

Note that if $p_t(x)$ in (4.6) is a constant for all t, or $\lambda = 0$, then $\widehat{f}_c(y \mid x)$ becomes the classical NW type double kernel estimator used by Scaillet (2005). However, Scaillet (2005) adopted a single bandwidth for smoothing in both the y and x directions. Clearly, $\widehat{f}_c(y \mid x)$ is a probability density function so that $\widehat{F}_c(y \mid x)$ is a cumulative distribution function (monotone, $0 \leq \widehat{F}_c(y \mid x) \leq 1, \widehat{F}_c(-\infty \mid x) = 0$, and $\widehat{F}_c(\infty \mid x) = 1\Big)$. Also, $\widehat{F}_c(y \mid x)$ is continuous and differentiable in y. Further, as expected, it will be shown that like $\widehat{F}_{c1}(y \mid x), \widehat{F}_c(y \mid x)$ has the attractive properties such as no boundary effects, design adaptation, and mathematical efficiency.

4.3.2 Estimation of Conditional VaR and ES

We now are ready to formulate the nonparametric estimators for $\nu_p(x)$ and $\mu_p(x)$. To this end, from (4.8), $\nu_p(x)$ is estimated by inverting the estimated conditional survival distribution $\widehat{S}_c(y \mid x) = 1 - \widehat{F}_c(y \mid x)$, denoted by $\widehat{\nu}_p(x)$ and defined as $\widehat{\nu}_p(x) = \widehat{S}_c^{-1}(p \mid x)$. Note that $\widehat{\nu}_p(x)$ always exists since $\widehat{S}_c(p \mid x)$ is a survival function itself. Plugging-in $\widehat{\nu}_p(x)$ and $\widehat{f}_c(y \mid x)$ into (4.1), we obtain the nonparametric estimation of $\mu_p(x)$,

$$\begin{aligned}\widehat{\mu}_p(x) &= p^{-1}\int_{\widehat{\nu}_p(x)}^{\infty} y\widehat{f}_c(y \mid x)dy = p^{-1}\sum_{t=1}^{n} W_{c,t}(x,h)\int_{\widehat{\nu}_p(x)}^{\infty} yK_{h_0}\left(y - Y_t\right)dy \\ &= p^{-1}\sum_{t=1}^{n} W_{c,t}(x,h)\left[Y_t\bar{G}_{h_0}\left(\widehat{\nu}_p(x) - Y_t\right) + h_0 G_{1,h_0}\left(\widehat{\nu}_p(x) - Y_t\right)\right], \qquad (4.9)\end{aligned}$$

where $\bar{G}(u) = 1 - G(u), G_{1,h_0}(u) = G_1\left(u/h_0\right)$, and $G_1(u) = \int_u^{\infty} vK(v)dv$. Note that as mentioned earlier, $\widehat{\nu}_p(x)$ in (4.9) can be an any consistent estimator.

4.4 Distribution Theory

4.4.1 Assumptions

Before we proceed with the asymptotic properties of the proposed nonparametric estimators, we first list all assumptions needed for the asymptotic theory, although some of them might not be the weakest possible. Note that proofs of the asymptotic results presented in this section may be found in Section 4.6 with some lemmas and their detailed proofs relegated to Section 4.6.2. First, we introduce some notation. Let $\alpha(K) = \int_{-\infty}^{\infty} uK(u)\bar{G}(u)du$ and

$\mu_j(W)=\int_{-\infty}^{\infty} u^j W(u)du$. Also, for any $j \geq 0$, write

$$l_j(u \mid v)=E\left[Y_t^j I\left(Y_t \geq u\right) \mid X_t=v\right]=\int_u^{\infty} y^j f(y \mid v)dy, \quad l_j^{a,b}(u \mid v)=\frac{\partial^{ab}}{\partial u^a \partial v^b} l_j(u \mid v),$$

and $l_j^{a,b}\left(\nu_p(x) \mid x\right)=\left. l_j^{a,b}(u \mid v)\right|_{u=\nu_p(x), v=x}$. Clearly, $l_0(u \mid v)=S(u \mid v)$ and $l_1\left(\nu_p(x) \mid x\right)=p\mu_p(x)$. Finally, $l_j^{1,0}(u \mid v)=-u^j f(u \mid v)$ and $l_j^{2,0}(u \mid v)=-\left[u^j f^{1,0}(u \mid v)+j u^{j-1} f(u \mid v)\right]$. We now list the following regularity conditions.

Assumption A:

A1. For fixed y and $x, 0<F(y \mid x)<1, g(x)>0$, the marginal density of X_t, and is continuous at x, and $F(y \mid x)$ has continuous second order derivative with respect to both x and y.

A2. The kernels $K(\cdot)$ and $W(\cdot)$ are symmetric, bounded, and compactly supported density.

A3. $h \rightarrow 0$ and $n\,h \rightarrow \infty$, and $h_0 \rightarrow 0$ and $n\,h_0 \rightarrow \infty$, as $n \rightarrow \infty$.

A4. Let $g_{1,t}(\cdot, \cdot)$ be the joint density of X_1 and X_t for $t \geq 2$. Assume that $\mid g_{1,t}(u, v)-g(u) g(v) \mid \leq M<\infty$ for all u and v.

A5. The process $\{(X_t, Y_t)\}$ is a stationary α-mixing with the mixing coefficient satisfying $\alpha(t)=O\left(t^{-(2+\delta)}\right)$ for some $\delta>0$.

A6. $n\,h^{1+2/\delta} \rightarrow \infty$

A7. $h_0=o(h)$.

Assumption B:

B1. Assume that $E\left(\left|Y_t\right|^\delta \mid X_t=u\right) \leq M_3<\infty$ for some $\delta>2$, in a neighborhood of x.

B2. Assume that $\left|g_{1,t}\left(y_1, y_2 \mid x_1, x_2\right)\right| \leq M_1<\infty$ for all $t \geq 2$, where $g_{1,t}\left(y_1, y_2 \mid x_1, x_2\right)$ be the conditional density of Y_1 and Y_t given $X_1=x_1$ and $X_t=x_2$.

B3. The mixing coefficient of the α-mixing process $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ satisfies $\sum_{t \geq 1} t^a \alpha^{1-2/\delta}(t)$ $<\infty$ for some $a>1-2/\delta$, where δ is given in Assumption B1.

B4. Assume that there exists a sequence of integers $s_n>0$ such that $s_n \rightarrow \infty, s_n=o\left((nh)^{1/2}\right)$, and $(n/h)^{1/2} \alpha\left(s_n\right) \rightarrow 0$, as $n \rightarrow \infty$.

B5. There exists $\delta^* > \delta$ such that $E\left(|Y_t|^{\delta^*} \mid X_t = u\right) \leq M_4 < \infty$ in a neighborhood of $x, \alpha(t) = O\left(t^{-\theta^*}\right)$, where δ is given in Assumption B1, $\theta^* \geq \delta^*\delta/\left\{2\left(\delta^* - \delta\right)\right\}$, and $n^{1/2-\delta/4}h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Remark 4.1: *Note that Assumptions A1 - A5 and B1 - B5 are used commonly in the literature of time series data (see, e.g., Masry and Fan, 1997, Cai, 2001). Note that α-mixing imposed in Assumption A5 is weaker than β-mixing in Hall, Wolff, and Yao (1999) and ρ-mixing in Fan, Yao, and Tong (1996). Because A6 is satisfied by the bandwidths of optimal size (i.e., $h \approx n^{-1/5}$) if $\delta > 1/2$, we do not concern ourselves with such refinements. Indeed, Assumptions A1 - A6 are also required in Cai (2002). Assumption A7 means that the initial step bandwidth should be chosen as small as possible so that the bias from the initial step can be ignored. Since the common technique - truncation approach for time series data is not applicable to our setting (see, e.g., Masry and Fan, 1997), the purpose of Assumption B5 is to use the moment inequality. If $\alpha(t)$ decays geometrically, then Assumptions B4 and B5 are satisfied automatically. Note that Assumptions B3, B4, and B5 are stronger than Assumptions A5 and A6. This is not surprising because the higher moments involved, the faster decaying rate of $\alpha(\cdot)$ is required. Finally, Assumptions B1 - B5 are also imposed in Cai (2001).*

4.4.2 Asymptotic Properties for Conditional PDF and CDF

First, we investigate the asymptotic behaviors of $\widehat{f}_c(y \mid x)$, including the asymptotic normality stated in the following theorem.

Theorem 4.1: *Under Assumptions A1 - A6 with h in A3 and A6 replaced by h_0h, we have*

$$\sqrt{n\,h_0h}\left[\widehat{f}_c(y \mid x) - f(y \mid x) - B_f(y \mid x)\right] \;\rightarrow\; N\left\{0, \sigma_f^2(y \mid x)\right\},$$

where the asymptotic bias is

$$B_f(y \mid x) = \frac{h^2}{2}\mu_2(W)f^{0,2}(y \mid x) + \frac{h_0^2}{2}\mu_2(K)f^{2,0}(y \mid x),$$

and the asymptotic variance is $\sigma_f^2(y \mid x) = \mu_0\left(K^2\right)\mu_0\left(W^2\right)f(y \mid x)/g(x)$.

Remark 4.2: *The asymptotic results for $\widehat{f}_c(y \mid x)$ in Theorem 4.1 are similar to those for $\widehat{f}_{ll}(y \mid x)$ in Fan, Yao, and Tong (1996) for the ρ-mixing sequence, which is stronger than*

α-mixing, but as mentioned earlier, $\widehat{f}_{ll}(y \mid x)$ is not always a probability density function. The asymptotic bias and variance are intuitively expected. The bias comes from the approximations in both x and y directions and the variance is from the local conditional variance in the density estimation setting, which is $f(y \mid x)$.

Next, we study the asymptotic behaviors for $\widehat{S}_c(y \mid x)$ at both interior and boundary points. Similar to Theorem 4.1 for $\widehat{f}_c(y \mid x)$, we have the following asymptotic normality for $\widehat{S}_c(y \mid x)$

Theorem 4.2: *Under Assumptions A1 - A6, we have*

$$\sqrt{nh}\left[\widehat{S}_c(y \mid x) - S(y \mid x) - B_S(y \mid x)\right] \quad \to \quad N\left\{0, \sigma_S^2(y \mid x)\right\},$$

where the asymptotic bias is given by

$$B_S(y \mid x) = \frac{h^2}{2}\mu_2(W)S^{0,2}(y \mid x) - \frac{h_0^2}{2}\mu_2(K)f^{1,0}(y \mid x),$$

and the asymptotic variance is $\sigma_S^2(y \mid x) = \mu_0\left(W^2\right)S(y \mid x)[1 - S(y \mid x)]/g(x)$. In particular, if Assumption A7 holds true, then,

$$\sqrt{nh}\left[\widehat{S}_c(y \mid x) - S(y \mid x) - \frac{h^2}{2}\mu_2(W)S^{0,2}(y \mid x)\right] \quad \to \quad N\left\{0, \sigma_S^2(y \mid x)\right\}.$$

Remark 4.3: *Note that the asymptotic results for $\widehat{S}_c(y \mid x)$ in Theorem 4.2 are analogous to those for $\widehat{S}_{ll}(y \mid x) = 1 - \widehat{F}_{ll}(y \mid x)$ in Yu and Jones (1998) for iid data, but as mentioned previously, $\widehat{F}_{ll}(y \mid x)$ is not always a distribution function. A comparison of $B_s(y \mid x)$ with the asymptotic bias for $\widehat{S}_{c1}(y \mid x)$ (see Theorem 1 in Cai (2002)), it reveals that there is an extra term $\frac{h_0^2}{2}f^{1,0}(y \mid x)\mu_2(K)$ in the asymptotic bias expression $B_s(y \mid x)$ due to the vertical smoothing in the y direction. Also, there is an extra term in the asymptotic variance (see* (4.20)*). These extra terms are carried over from the initial estimate but they can be ignored if the bandwidth at the initial step is taken to be a higher order than the bandwidth at the smoothing step.*

Remark 4.4: *It is important to examine the performance of $\widehat{S}_c(y \mid x)$ by considering the asymptotic mean squared error (AMSE). Theorem 4.2 concludes that the AMSE of $\widehat{S}_c(y \mid x)$ is*

$$\begin{aligned} AMSE\left(\widehat{S}_c(y \mid x)\right) =& \frac{\left\{h^2\mu_2(W)S^{0,2}(y \mid x) - h_0^2\mu_2(K)f^{1,0}(y \mid x)\right\}^2}{4} \\ &+ \frac{1}{nh}\frac{\mu_0\left(W^2\right)S(y \mid x)[1 - S(y \mid x)]}{g(x)}. \end{aligned} \tag{4.10}$$

By minimizing AMSE in (4.10) *and taking* $h_0 = o(h)$*, therefore, we obtain the optimal bandwidth given by*

$$h_{opt\ ,S}(y \mid x) = \left[\frac{\mu_0\left(W^2\right) S(y \mid x)[1 - S(y \mid x)]}{\left\{\mu_2(W) S^{0,2}(y \mid x)\right\}^2 g(x)} \right]^{1/5} n^{-1/5}.$$

Therefore, the optimal rate of the AMSE of $\widehat{S}_c(y \mid x)$ *is* $n^{-4/5}$.

As for the boundary behavior of the WDKLL estimator, we can follow Cai (2002) to establish a similar result for $\widehat{S}_c(y \mid x)$ like Theorem 2 in Cai (2002). Without loss of generality, we consider the left boundary point $x = ch, 0 < c < 1$. From Fan, Hu, and Troung (1994), we take $W(\cdot)$ to have support $[-1, 1]$ and $g(\cdot)$ to have support $[0, 1]$. Then, under Assumptions A1 - A7, by following the same proof as that for Theorem 4.2 and using the second assertion in Lemma 4.1, although not straightforward, we can show that

$$\sqrt{nh}\left[\widehat{S}_c(y \mid ch) - S_c(y \mid ch) - B_{S,c}(y)\right] \quad \rightarrow \quad N\left(0, \sigma^2_{S,c}(y)\right), \tag{4.11}$$

where the asymptotic bias term is given by $B_{S,c}(y) = h^2 \beta_0(c) S^{0,2}(y \mid 0+) / \left[2\beta_1(c)\right]$ and the asymptotic variance is $\sigma^2_{S,c}(y) = \beta_2(0) S(y \mid 0+)[1 - S(y \mid 0+)] / \left[\beta_1^2(c) g(0+)\right]$ with $g(0+) = \lim_{z \downarrow 0} g(z)$

$$\beta_0(c) = \int_{-1}^{c} \frac{u^2 W(u)}{1 - \lambda_c u W(u)} du, \quad \beta_j(c) = \int_{-1}^{c} \frac{W^j(u)}{\{1 - \lambda_c u W(u)\}^j} du, \quad 1 \leq j \leq 2,$$

and λ_c being the root of equation $L_c(\lambda) = 0$

$$L_c(\lambda) = \int_{-1}^{c} \frac{u W(u)}{1 - \lambda u W(u)} du.$$

Note that the proof of (4.11) is similar to that for Theorem 2 in Cai (2002) and omitted. Theorem 4.2 and (4.11) reflect two of the major advantages of the WKDLL estimator: (a) the asymptotic bias does not depend on the design density $g(x)$, and indeed it is dependent only on the simple conditional distribution curvature $S^{0,2}(y \mid x)$ and conditional density curvature $f^{1,0}(y \mid x)$; and (b) it has an automatic good behavior at boundaries. See Cai (2002) for the detailed discussions.

Finally, we remark that if the point 0 were an interior point, then, (4.11) would hold with $c = 1$, which becomes Theorem 4.2. Therefore, Theorem 4.2 shows that the WKDLL estimation has the automatic good behavior at boundaries without the need of the boundary correction.

4.4.3 Asymptotic Theory for CVaR and CES

By the differentiability of $\widehat{S}_c\left(\widehat{\nu}_p(x) \mid x\right)$, we use the Taylor expansion and ignore the higher terms to obtain

$$\widehat{S}_c\left(\widehat{\nu}_p(x) \mid x\right) = p \approx \widehat{S}_c\left(\nu_p(x) \mid x\right) - \widehat{f}_c\left(\nu_p(x) \mid x\right)\left(\widehat{\nu}_p(x) - \nu_p(x)\right), \tag{4.12}$$

then, by Theorem 4.1,

$$\widehat{\nu}_p(x) - \nu_p(x) \approx \left[\widehat{S}_c\left(\nu_p(x) \mid x\right) - p\right] / \widehat{f}_c\left(\nu_p(x) \mid x\right) \approx \left[\widehat{S}_c\left(\nu_p(x) \mid x\right) - p\right] / f\left(\nu_p(x) \mid x\right).$$

As an application of Theorem 4.2, we can establish the following theorem for the asymptotic normality of $\widehat{\nu}_p(x)$ but the proof is omitted since it is similar to that for Theorem 4.2.

Theorem 4.3: *Under Assumptions A1 - A6, we have*

$$\sqrt{nh}\left[\widehat{\nu}_p(x) - \nu_p(x) - B_\nu(x)\right] \quad \rightarrow \quad N\left\{0, \sigma_\nu^2(x)\right\},$$

where the asymptotic bias is $B_\nu(x) = B_S\left(\nu_p(x) \mid x\right) / f\left(\nu_p(x) \mid x\right)$ and the asymptotic variance is $\sigma_\nu^2(x) = \mu_0\left(W^2\right) p(1-p) / \left[g(x) f^2\left(\nu_p(x) \mid x\right)\right]$. In particular, if Assumption A7 holds, then,

$$\sqrt{nh}\left[\widehat{\nu}_p(x) - \nu_p(x) - \frac{h^2}{2}\frac{S^{0,2}\left(\nu_p(x) \mid x\right)}{f\left(\nu_p(x) \mid x\right)}\mu_2(W)\right] \quad \rightarrow \quad N\left\{0, \sigma_\nu^2(x)\right\}.$$

Remark 4.5: *First, as a consequence of Theorem 4.3, $\widehat{\nu}_p(x) - \nu_p(x) = O_p\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ so that $\widehat{\nu}_p(x)$ is a consistent estimator of $\nu_p(x)$ with a convergence rate. Also, note that the asymptotic results for $\widehat{\nu}_p(x)$ in Theorem 4.3 are akin to those for $\widehat{\nu}_{l,p}(x) = \widehat{S}_{ll}^{-1}(p \mid x)$ in Yu and Jones (1998) for iid data. But in the bias term of Theorem 4.3, the quantity $S^{0,2}\left(\nu_p(x) \mid x\right) / f\left(\nu_p(x) \mid x\right)$, involving the second derivative of the conditional distribution function with respect to x, replaces $\nu_p''(x)$, the second derivative of the conditional VaR function itself, which is in the bias term of the "check" function type local linear estimator in Yu and Jones (1998) for iid data and Cai and Xu (2008) for time series. See Cai and Xu (2008) for details. This is not surprising since the bias comes only from the approximation. The former utilizes the approximation of the conditional distribution function but the later uses the approximation of the conditional VaR function. Finally, Theorems 4.2 and 4.3 imply that if the initial bandwidth h_0 is chosen small as possible such as $h_0 = o(h)$, the final estimates of $S(y \mid x)$ and $\nu_p(x)$ are not sensitive to the choice of h_0 as long as it satisfies Assumption A7. This makes the selection of bandwidths much easier in practice, which will be elaborated later (see Section 4.5.1).*

Remark 4.6: *Similar to Remark 4.5 , we can derive the asymptotic mean squared error for $\widehat{\nu}_p(x)$. By following Yu and Jones (1998), Theorem 4.3, and* (4.20) *(given in Section 4.6 imply that the AMSE of $\widehat{\nu}_p(x)$ is given by*

$$AMSE\left(\widehat{\nu}_p(x)\right) = \frac{\left\{h^2 S^{0,2}\left(\nu_p(x) \mid x\right) \mu_2(W) - h_0^2 f^{1,0}\left(\nu_p(x) \mid x\right) \mu_2(K)\right\}^2}{4 f^2\left(\nu_p(x) \mid x\right)} + \frac{1}{nh} \frac{\mu_0\left(W^2\right)\left[p(1-p) + 2 h_0 f\left(\nu_p(x) \mid x\right) \alpha(K)\right]}{f^2\left(\nu_p(x) \mid x\right) g(x)}. \tag{4.13}$$

Note that the above result is similar to that in Theorem 1 in Yu and Jones (1998) for the double kernel local linear conditional quantile estimator. But, a comparison of (4.13) *with Theorem 3 in Cai (2002) for the WNW estimator reveals that* (4.13) *has two extra terms (negligible if Assumption A7 is satisfied) due to the vertical smoothing in the y direction, as mentioned previously. By minimizing AMSE in* (4.13) *and taking $h_0 = o(h)$, therefore, we obtain the optimal bandwidth given by*

$$h_{opt\ ,\nu}(x) = \left[\frac{\mu_0\left(W^2\right) p(1-p)}{\left\{\mu_2(W) S^{0,2}\left(\nu_p(x) \mid x\right)\right\}^2 g(x)}\right]^{1/5} n^{-1/5}.$$

Therefore, the optimal rate of the AMSE of $\widehat{\nu}_p(x)$ is $n^{-4/5}$. By comparing $h_{opt,\nu}(x)$ with $h_{opt\ ,S}(y \mid x)$, it turns out that $h_{opt,\nu}(x)$ is $h_{opt,\nu}(y \mid x)$ evaluated at $y = \nu_p(x)$. Therefore, the best choice of the bandwidth for estimating $S_c(y \mid x)$ can be used for estimating $\nu_p(x)$.

Remark 4.7: *Similar to* (4.11)*, one can establish the asymptotic result at boundaries for $\nu_p(x)$ as follows, one can show that under Assumption A7,*

$$\sqrt{nh}\left[\widehat{\nu}_p(ch) - \nu_p(ch) - B_{\nu,c}\right] \quad \rightarrow \quad N\left(0, \sigma_{\nu,c}^2\right),$$

where the asymptotic bias is $B_{\nu,c} = h^2 \beta_2(c) S^{0,2}\left(\nu_p(0+) \mid 0+\right) / \left[2\beta_1(c) f\left(\nu_p(0+) \mid 0+\right)\right]$ and the asymptotic variance is $\sigma_{\nu,c}^2 = \beta_0(0) p[1-p] / \left[\beta_1^2(c) f^2\left(\nu_p(0+) \mid 0+\right) g(0+)\right]$. Clearly, $\widehat{\nu}_p(x)$ inherits all good properties from the WDKLL estimator of $S_c(y \mid x)$. Note that the above result can be established by using the second assertion in Lemma 4.1 and following the same lines along with those used in the proof of Theorem 4.2 and omitted.

Finally, we examine the asymptotic behavior for $\widehat{\mu}_p(x)$ at both interior and boundary points. First, we establish the following theorem for the asymptotic normality for $\widehat{\mu}_p(x)$ when x is an interior point.

Theorem 4.4: *Under Assumptions A1 - A4 and B2 - B5, we have*

$$\sqrt{n\,h}\,[\widehat{\mu}_p(x) - \mu_p(x) - B_\mu(x)] \quad \to \quad N\left\{0, \sigma_\mu^2(x)\right\},$$

where the asymptotic bias is $B_\mu(x) = B_{\mu,0}(x) + \frac{h_0^2}{2}\mu_2(K)p^{-1}f\left(\nu_p(x) \mid x\right)$ *with*

$$B_{\mu,0}(x) = \frac{h^2}{2}\mu_2(W)p^{-1}\left[l_1^{0,2}\left(\nu_p(x) \mid x\right) - \nu_p(x)S^{0,2}\left(\nu_p(x) \mid x\right)\right],$$

and the asymptotic variance is

$$\sigma_\mu^2(x) = \frac{\mu_0\left(W^2\right)}{p\,g(x)}\left[p^{-1}l_2\left(\nu_p(x) \mid x\right) - p\mu_p^2(x) + (1-p)\nu_p(x)\left\{\nu_p(x) - 2\mu_p(x)\right\}\right].$$

In particular, if Assumption A7 holds true, then,

$$\sqrt{nh}\,[\widehat{\mu}_p(x) - \mu_p(x) - B_{\mu,0}(x)] \quad \to \quad N\left\{0, \sigma_\mu^2(x)\right\}.$$

Remark 4.8: *First, Theorem 4.4 concludes that* $\widehat{\mu}_p(x) - \mu_p(x) = O_p\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ *so that* $\widehat{\mu}_p(x)$ *is a consistent estimator of* $\mu_p(x)$ *with a convergence rate. Also, note that the asymptotic results in Theorem 4.4 imply that* $\widehat{\mu}_p(x)$ *is a consistent estimator for* $\mu_p(x)$ *with a convergence rate* $(nh)^{-1/2}$*. Further, note that although the asymptotic variance* $\sigma_\mu^2(x)$ *is the same as that in Scaillet (2005) for* $\widetilde{\mu}_p(x)$*, Scaillet (2005) did not provide an expression for the asymptotic bias term like* $B_\mu(x)$ *in the first result or* $B_{\mu,0}(x)$ *in the second conclusion in Theorem 4.4. Clearly, the second term in the asymptotic bias expression is carried over from the y direction smoothing at the initial step and it is negligible if Assumption A7 is satisfied. Clearly, Assumption A7 implies that* $B_\mu(x)$ *becomes* $B_{\mu,0}(x)$*.*

Remark 4.9: *Like Remark 4.5, the AMSE for* $\widehat{\mu}_p(x)$ *can be derived in the same manner. It follows from Theorem 4.4 that the AMSE of* $\widehat{\mu}_p(x)$ *is given by*

$$AMSE(\widehat{\mu}_p(x)) = \frac{1}{nh}\sigma_\mu^2(x) + \left\{B_{\mu,0}(x) + \frac{h_0^2}{2}\mu_2(K)p^{-1}f\left(\nu_p(x) \mid x\right)\right\}^2. \tag{4.14}$$

Under Assumption A7, minimizing AMSE in (4.14) *with respect to h yields the optimal bandwidth given by*

$$h_{opt\,,\mu}(x) = \left[\frac{\sigma_\mu(x)}{\mu_2(W)p^{-1}\left\{l_1^{0,2}\left(\nu_p(x) \mid x\right) - \nu_p(x)S^{0,2}\left(\nu_p(x) \mid x\right)\right\}}\right]^{2/5} n^{-1/5}.$$

Therefore, as expected, the optimal rate of the AMSE of $\widehat{\mu}_p(x)$ *is* $n^{-4/5}$*.*

Finally, we offer the asymptotic results for $\widehat{\mu}_p(x)$ at the left boundary point $x = ch$. By the same fashion, one can show that under Assumption A7,

$$\sqrt{nh}\,[\widehat{\mu}_p(ch) - \mu_p(ch) - B_{\mu,c}] \quad \to \quad N\left(0, \sigma^2_{\mu,c}\right),$$

where the asymptotic bias is

$$B_{\mu,c} = h^2\beta_2(c)p^{-1}\left[l_1^{0,2}\left(\nu_p(0+) \mid 0+\right) - \nu_p(0+)S^{0,2}\left(\nu_p(0+) \mid 0+\right)\right] / \left[2\beta_1(c)\right],$$

and the asymptotic variance is

$$\sigma^2_{\mu,c} = \frac{\beta_0(0)}{p\beta_1^2(c)g(0+)}\left[p^{-1}l_2\left(\nu_p(0+) \mid 0+\right) - p\mu_p^2(0+) + (1-p)\nu_p(0+)\left\{\nu_p(0+) - 2\mu_p(0+)\right\}\right].$$

Note that the proof of the above result can be carried over by using the second assertion in Lemma 4.1 and following the same lines along with those used in the proof of Theorem 4.4 and omitted. Next, we consider the comparison of the performance of the WDKLL estimation $\widehat{\mu}_p(x)$ with the NW type kernel estimator $\widetilde{\mu}_p(x)$ as in Scaillet (2005). To this effect, it is not very difficult to derive the asymptotic results for the NW type kernel estimator but the proof is omitted since it is along the same line with the proof of Theorem 4.2. See Scaillet (2005) for the results at the interior point. Under some regularity conditions, it can be shown although tediously (see Cai (2002) for details) that at the left boundary $x = ch$, the asymptotic bias term for the NW type kernel estimator $\widetilde{\mu}_p(x)$ is of the order h by comparing to the order h^2 for the WDKLL estimate (see $B_{\mu,c}$ above). This shows that the WDKLL estimate does not suffer from boundary effects but the NW type kernel estimator estimate does. This is another advantage of the WDKLL estimator over the WW type kernel estimator $\widetilde{\mu}_p(x)$.

4.5 Empirical Examples

To illustrate the proposed methods, we consider two simulated examples and two real data examples on stock index returns and security returns. Throughout this section, the Epanechnikov kernel $K(u) = 0.75\left(1 - u^2\right)_+$is used and bandwidths are selected as described in the next section.

4.5.1 Bandwidth Selection

With the basic model at hand, one must address the important bandwidth selection issue, as the quality of the curve estimates depends sensitively on the choice of the bandwidth. For practitioners, it is desirable to have a convenient and effective data-driven rule. However, almost nothing has been done so far about this problem in the context of estimating $\nu_p(x)$ and $\mu_p(x)$ although there are some results available in the literature in other contexts for some specific purposes.

As indicated earlier, the choice of the initial bandwidth h_0 is not very sensitive to the final estimation but it needs to be specified. First, we use a very simple idea to choose h_0. As mentioned previously, the WNW method involves only one bandwidth in estimating the conditional distribution and VaR. Because the WNW estimate is a linear smoother (see (4.5)), we recommend using the optimal bandwidth selector, the so-called nonparametric AIC proposed by Cai and Tiwari (2000), to select the bandwidth, called $\widetilde{h}$. Then we take $0.1 \times \widetilde{h}$ or smaller as the initial bandwidth h_0. For the given h_0, we can select h as follows. According to (4.8), $\widehat{F}_c(\cdot \mid \cdot)$ is a linear estimator so that the nonparametric AIC selector of Cai and Tiwari (2000) can be applied here to select the optimal bandwidth for $\widehat{F}_c(\cdot \cdot \cdot)$, denoted by h_S. As mentioned at the end of Remark 6 , the bandwidth for $\widehat{\nu}_p(x)$ is the same as that for $\widehat{F}_c(\cdot \mid \cdot)$ so that it is simply to take h_S as h_ν. From (4.9), $\widehat{\mu}_p(x)$ is a linear estimator too for given $\widehat{\nu}_p(x)$. Therefore, by the same token, the nonparametric AIC selector is applied to selecting h_μ for $\widehat{\mu}_p(x)$. This simple approach is used in our implementation in the next sections.

4.5.2 Simulated Examples

In the simulated examples, we demonstrate the finite sample performance of the estimators in terms of the mean absolute deviation error. For example, the MADE for $\widehat{\mu}_p(x)$ is defined as

$$\mathcal{E}_{\mu_p} = \frac{1}{n_0} \sum_{k=1}^{n_0} \left| \widehat{\mu}_p\left(x_k\right) - \mu_p\left(x_k\right) \right|,$$

where $\{x_k\}_{k=1}^{n_0}$ are the pre-determined regular grid points. Similarly, we can define the MADE for $\widehat{\nu}_p(x)$, denoted by $\mathcal{E}_{\nu_p}$.

Example 4.1: We consider an ARCH type model with $X_t = Y_{t-1}$,

$$Y_t = 0.9 \sin\left(2.5 X_t\right) + \sigma\left(X_t\right) \varepsilon_t,$$

where $\sigma^2(x) = 0.8\sqrt{1.2 + x^2}$ and $\{\varepsilon_t\}$ are iid standard normal random variables. We consider three sample sizes: $n = 250$ and 500, and 1000 and the experiment is repeated 500 times for each sample size. The mean absolute deviation errors are computed for each sample size and each replication.

The 5% WDKLL and NW estimations are summarized in Figure 4.1 for CVaR and in Figure 4.2 for CES. For each n, the Box-plots of $500\mathcal{E}_{\nu_p}$-values of the WDKLL and NW estimations are plotted in Figure 4.1(d) for CVaR and in Figure 4.2 (d) for CES.

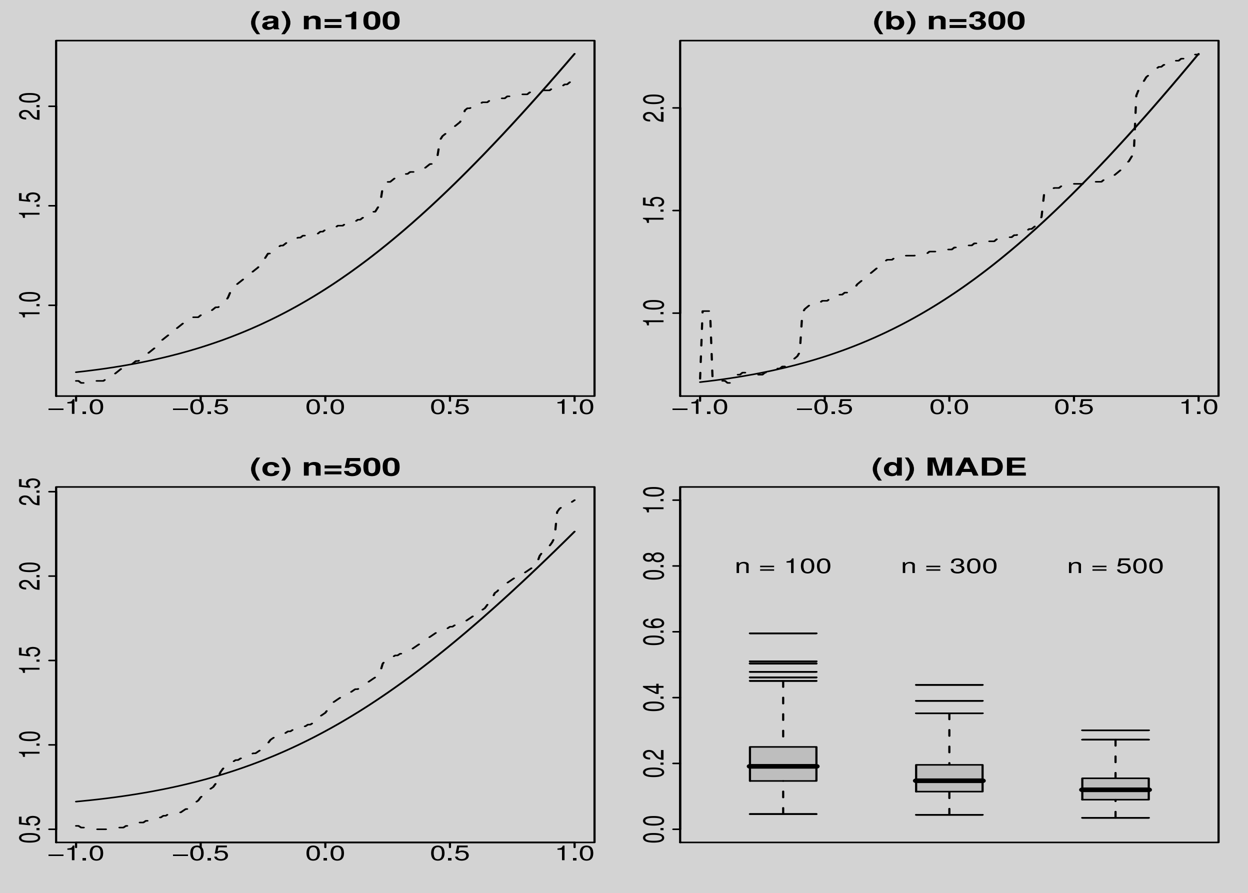


Figure 4.1: Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CVaR are plotted in (d).

From Figures 4.1(d) and 4.2(d), we can observe that the estimation becomes stable as the sample size increases for both the WDKLL and NW estimators. This is in line with our asymptotic theory that the proposed estimators are consistent. Further, it is obvious that the MADEs of the WDKLL estimator are smaller than those for the NW estimator. This indicates that our WDKLL estimator has smaller bias than that for the NW estimator. This implies that the overall performance of the WDKLL estimator should be better than that for the NW estimator.

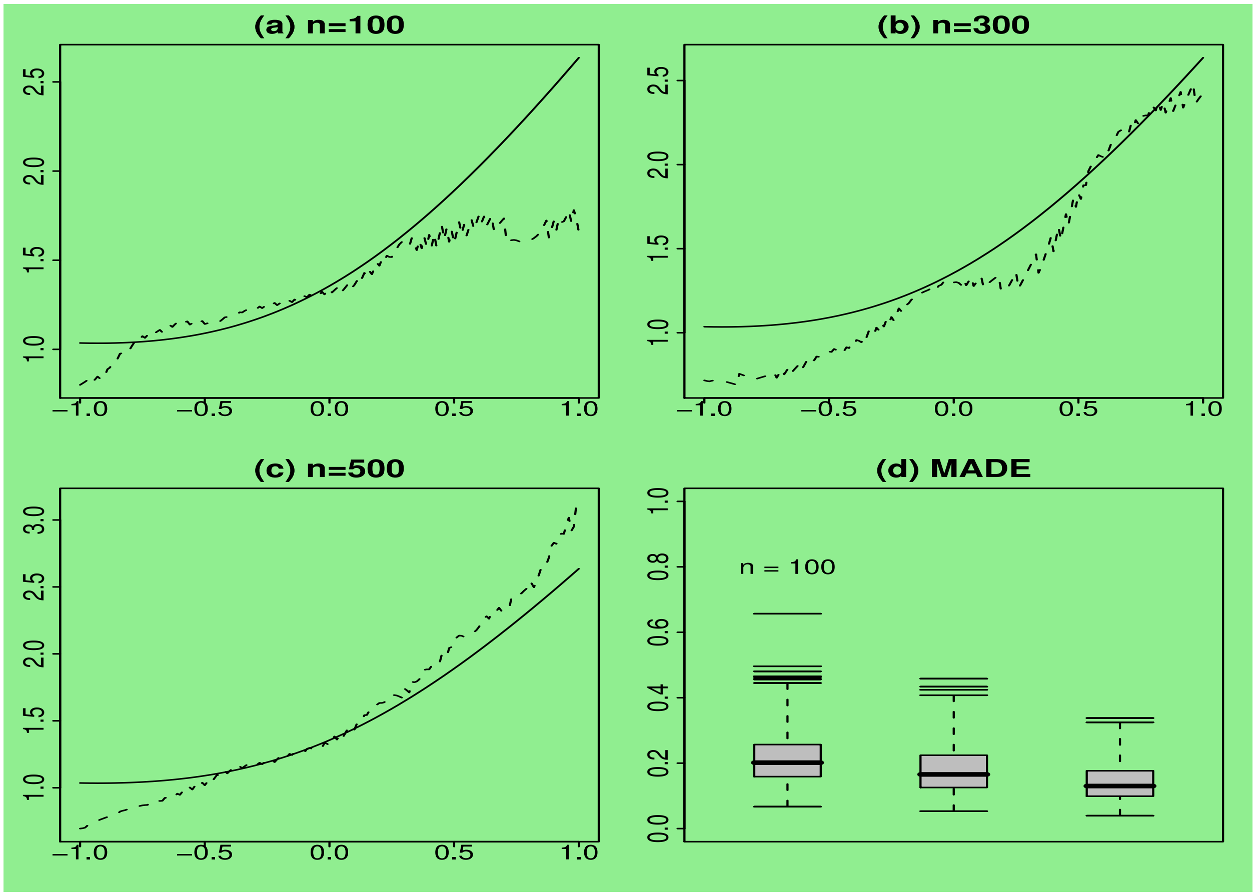


Figure 4.2: Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CES are plotted in (d).

Figures 4.1$(a) - (c)$ for $n = 250, 500$ and 1000, respectively, display the true CVaR function (solid line) $\nu_p(x) = 0.9\sin(2.5x) + \sigma(x)\Phi^{-1}(1-p)$, where $\Phi(\cdot)$ is the standard normal distribution function, together with the dashed and dotted lines representing the proposed WDKLL (dashed) and NW (dotted) estimates of CVaR, respectively, which are computed based on a typical sample. The typical sample is selected in such a way that its $\mathcal{E}_{\nu_p}$ value is equal to the median in the 500 replications. From Figures 4.1$(a) - (c)$, we can observe that both the estimated curves are closer to the true curve as n increases and the performance of the WDKLL estimator is better than that for the NW estimator, especially at boundaries.

In Figures 4.2(a)-(c), the true CES function $\mu_p(x) = 0.9\sin(2.5x)p + \sigma(x)\mu_1\left(\Phi^{-1}(1-p)\right)$ is displayed by the solid line, where $\mu_1(t) = \int_t^\infty u\phi(u)du$ and $\phi(\cdot)$ is the standard normal distribution density function, and the dashed and dotted lines present the proposed WDKLL (dashed) and NW (dotted) estimates of CES, respectively, from a typical sample. The typical sample is selected in such a way that its $\mathcal{E}_{\mu_p}$-value is equal to the median in the 500 replications. We can conclude from Figures $4.2(a)-(c)$ that the CES estimator has a similar performance as that for the CVaR estimator.

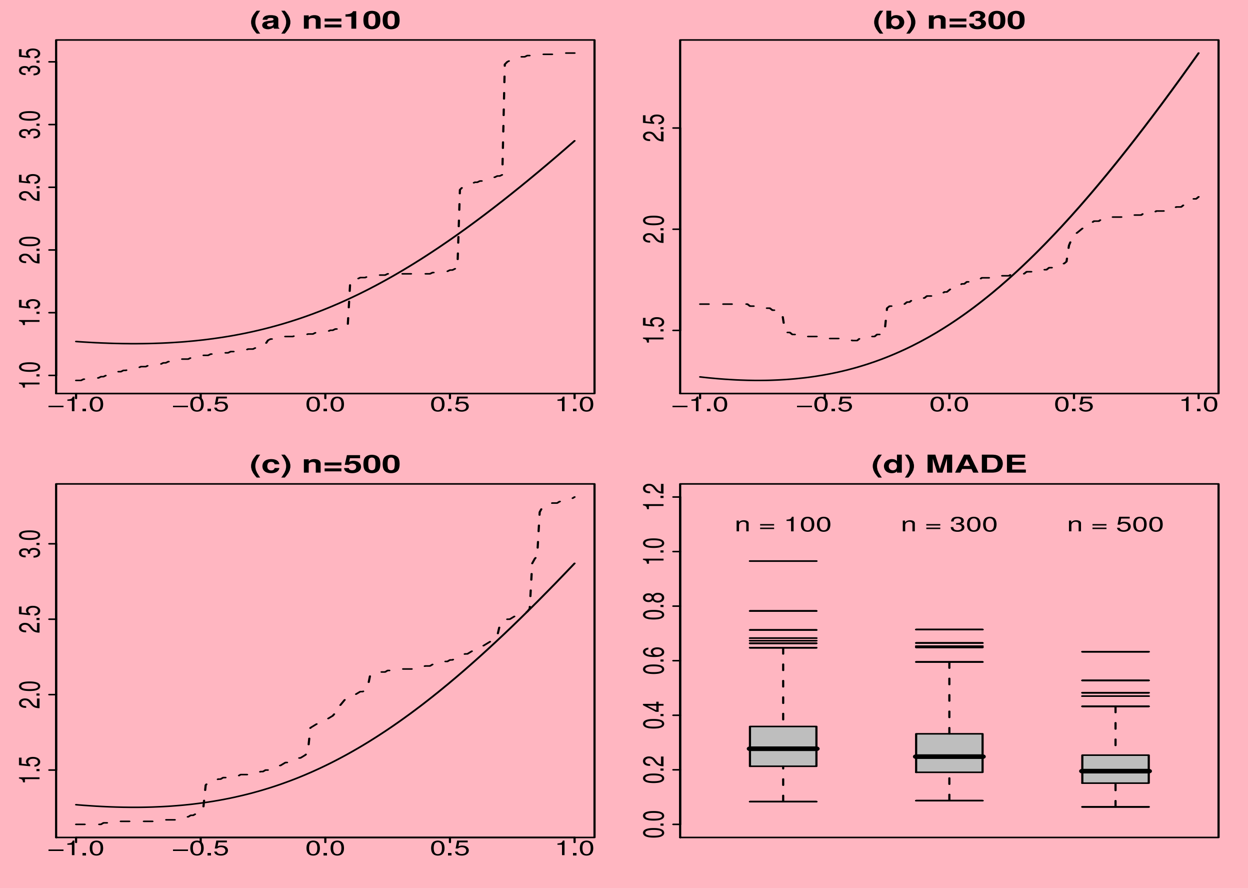


Figure 4.3: Simulation results for Example 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000 , respectively. Box-plots of the 500 MADE values for both WDKLL and NW estimation of the conditional VaR are plotted in (d).

The 1% WDKLL and NW estimates of CVaR and CES are computed under the same setting and they are displayed in Figures 4.3 and 4.4, respectively. Similar conclusions to those for the 5% estimates can be observed. But it is not surprising to see that the performance of the 1% CVaR and CES estimates is not good as that for the 5% estimates due to the sparsity of data.

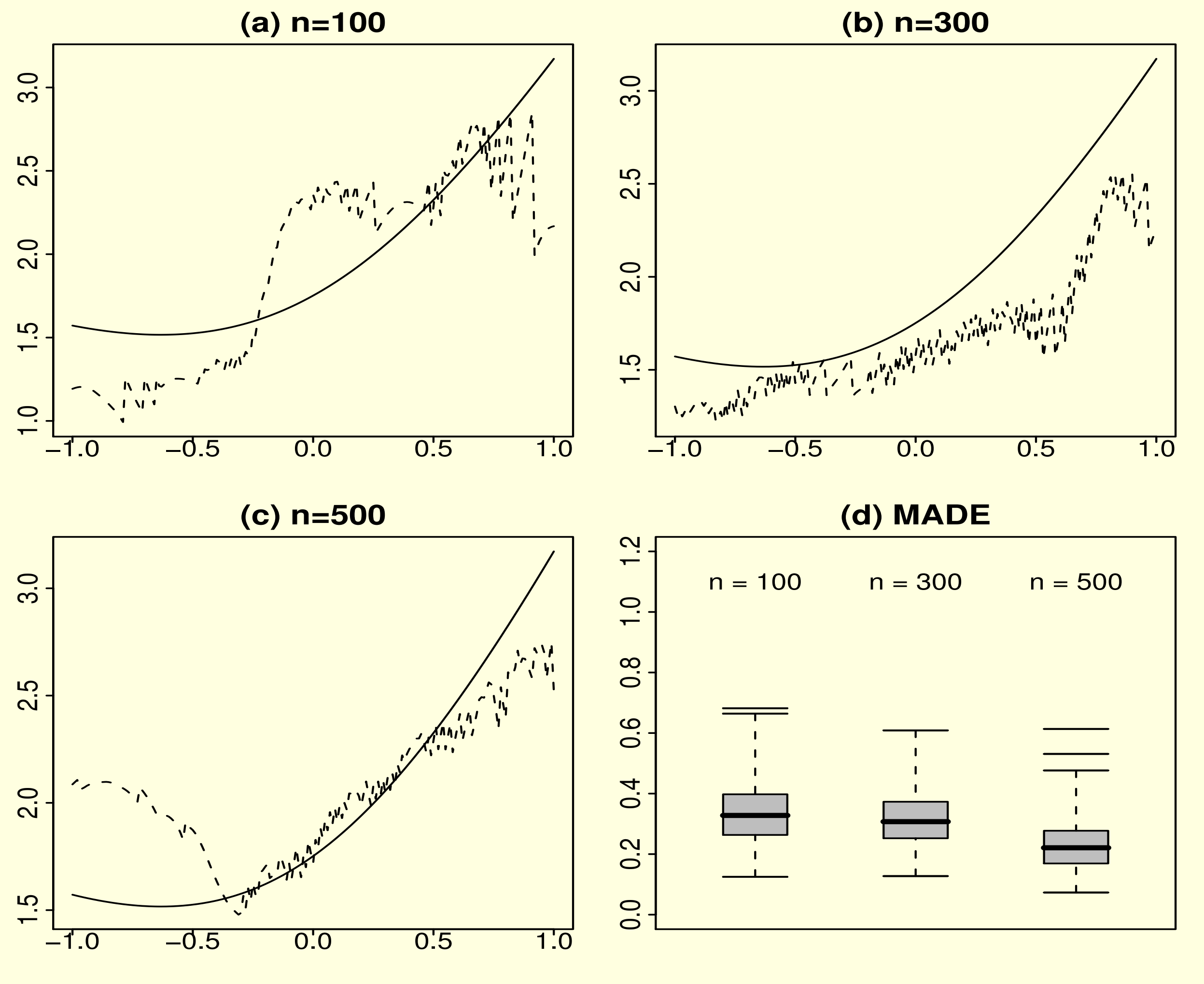


Figure 4.4: Simulation results for 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Boxplots of the 500 MADE values for both the WDKLL and NW estimations of CVaR are plotted in (d).

Example 4.2: In the above example, we consider only the case when X_t is one-dimensional. In this example, we consider the multivariate situation, i.e. X_t consists of two lagged variables: $X_{t1} = Y_{t-1}$ and $X_{t2} = Y_{t-2}$. The data generating model is given below:

$$Y_t = m\left(X_t\right) + \sigma\left(X_t\right)\varepsilon_t,$$

where $m(x) = 0.63x_1 - 0.47x_2, \sigma^2(x) = 0.5 + 0.23x_1^2 + 0.3x_2^2$, and $\{\varepsilon_t\}$ are iid generated from $N(0,1)$. Three sample sizes: $n = 200, 400$, and 600 , are considered here. For each sample size, we replicate the design 500 times. Here we present only the Box-plots of the 500 MADEs for the CVaR and CES estimates in Figure 4.5. Figure 4.5(a) displays the Box-plots of the $500 \mathcal{E}_{\nu_p}$-values of the WDKLL and NW estimates of CVaR and the Box-plots of the $500 \mathcal{E}_{\mu_p}$-values of the WDKLL and NW estimates of CES are given in Figure 4.5(b). From Figures 4.5(a) and (b), it is visually verified that both WDKLL and NW estimations become stable as the sample size increases and the performance of the WDKLL estimator is better than that for the NW estimator.

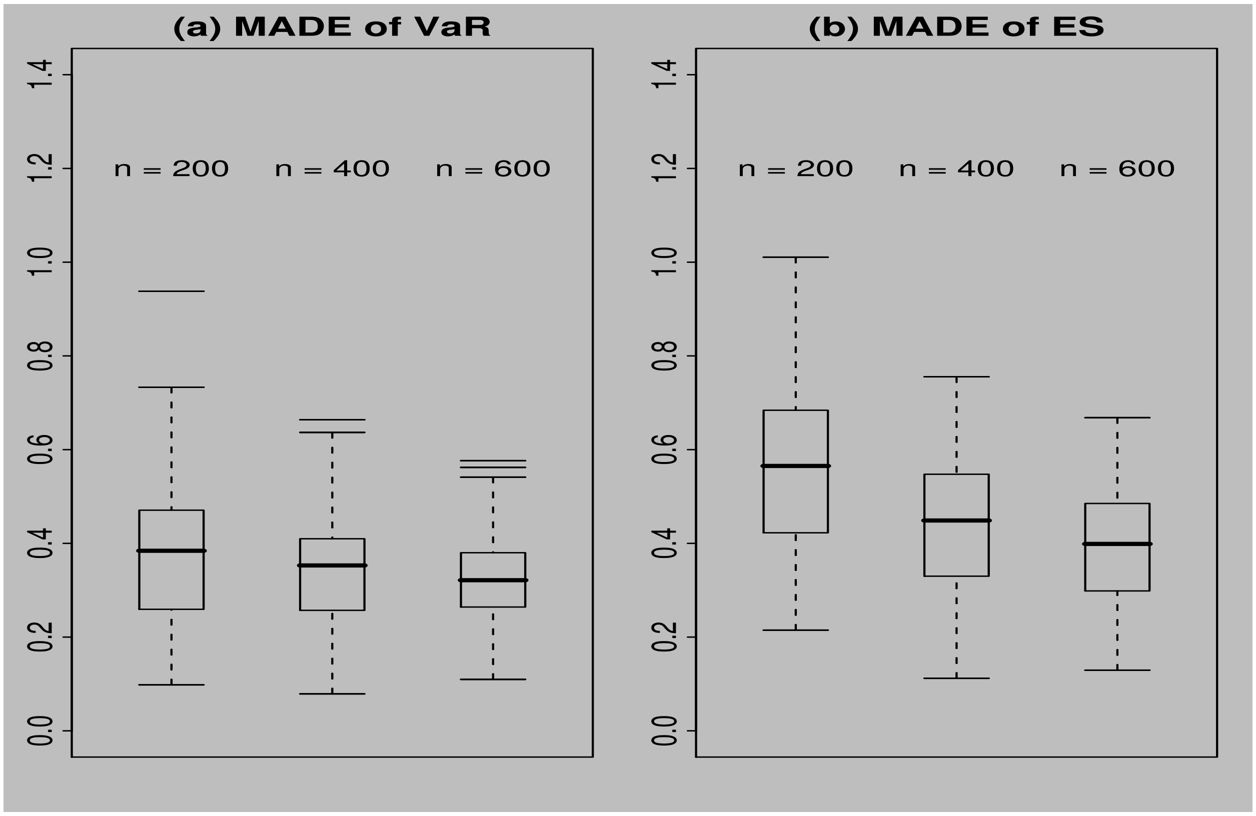


Figure 4.5: Simulation results for Example 4.2 when $p = 0.05$. (a) Box-plots of MADEs for both the WDKLL and NW estimates for CVaR. (b) Box-plots of MADEs for Both the WDKLL and NW estimates for CES.

4.5.3 Real Examples

Example 4.3: Now, we illustrate our proposed methodology by considering a real data set on Dow Jones Industrials (DJI) index returns. We took a sample of 1801 daily prices from

DJI index, from November 3, 1998 to January 3, 2006, and computed the daily returns as 100 times the difference of the log of prices. Let Y_t be the daily negative log return (log loss) of DJI and X_t be the first lagged variable of Y_t. The estimators proposed in this chapter are used to estimate the 5% CVaR and CES functions. The estimation results are shown in Figure 4.6 for the 5% CVaR estimate in Figure 4.6(a) and the 5% CES estimate in Figure 4.6(b). Both CVaR and CES estimates exhibit a U-shape, which corresponds to the so-called ”volatility smile”. Therefore, the risk tends to be lower when the lagged log loss of DJI is close to the empirical average and larger otherwise. We can also observe that the curves are asymmetric. This may indicate that the DJI is more likely to fall down if there was a loss within the last day than there was a same amount positive return.

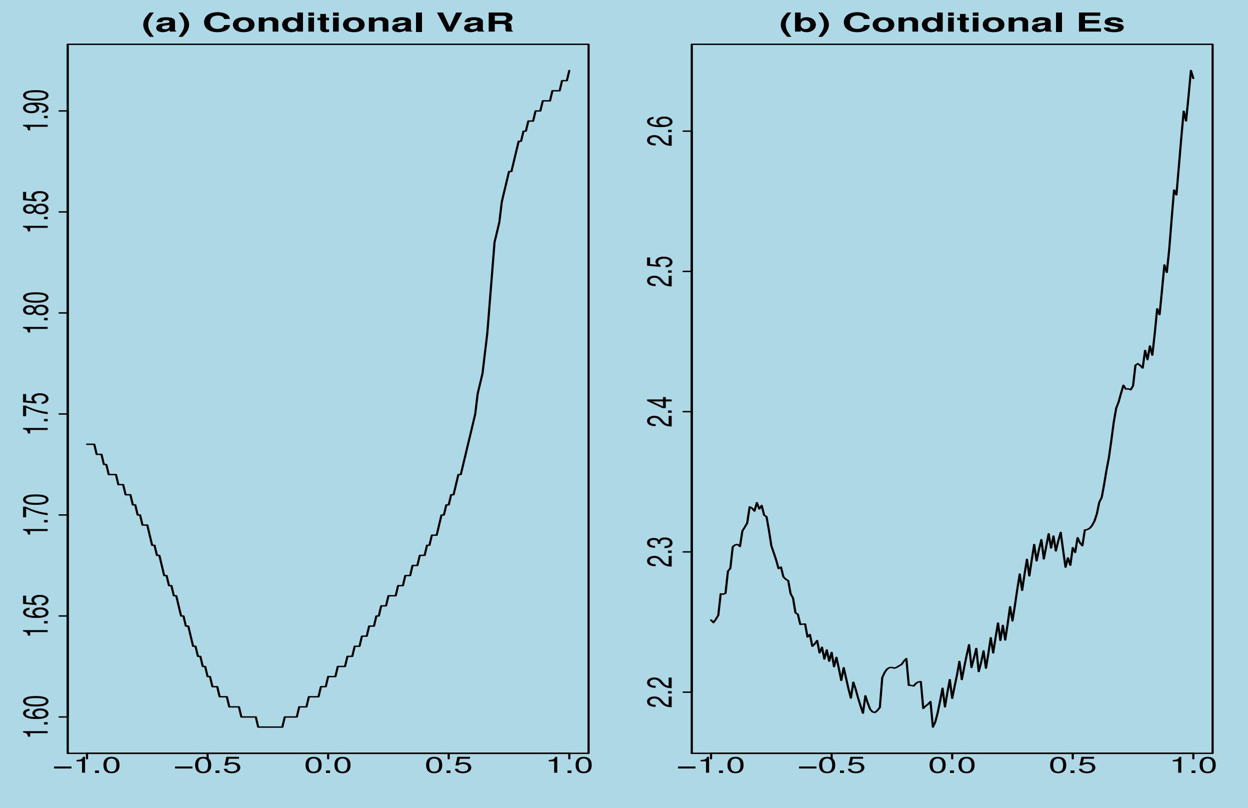


Figure 4.6: (a) 5% CVaR estimate for DJI index. (b) 5% CES estimate for DJI index.

Example 4.4: We apply the proposed methods to estimate the conditional value-at-risk and expected shortfall of the International Business Machine Co. (NYSE: IBM) security returns. The data are daily prices recorded from March 1, 1996 to April 6, 2005. We use the same method to calculate the daily returns as in Example 4.3. In order to estimate the value-at-risk of a stock return, generally, the information set X_t may contain a market index of corresponding capitalization and type, the industry index, and the lagged values of stock return. For this example, Y_t is the log loss of IBM stock returns and only two variables are chosen as information set for the sake of simplicity. Let X_{t1} be the first lagged variable of Y_t and X_{t2} denote the first lagged daily log loss of Dow Jones Industrials (DJI) index. Our main results from the estimation of the model are summarized in Figure 4.7. The surfaces of the estimators of IBM returns are given in Figure 4.7(a) for CVaR and in Figure 4.7(b) for CES. For visual convenience, Figures 4.7(c) and (e) depict the estimated CVaR and CES curves (as function of X_{t2}) for three different values of $X_{t1} = (-0.275, -0.025, 0.325)$ and Figures 4.7(d) and (f) display the estimated CVaR and CES curves (as function of X_{t1}) for three different values of $X_{t2} = (-0.225, 0.025, 0.425)$.

From Figures 4.7(c) - (f), we can observe that most of these curves are U-shaped. This is consistent with the results observed in Example 4.3. Also, we can see that these three curves in each figure are not parallel. This implies that the effects of lagged IBM and lagged DJI variables on the risk of IBM are different and complex. To be concrete, let us examine Figure 4.7(d). Three curves are close to each other when the lagged IBM log loss is around -0.2 and far away otherwise. This implies that DJI has fewer effects (less information) on CVaR around this value. Otherwise, DJI has more effects when the lagged IBM log loss is far from this value.

4.6 Proofs

4.6.1 Proofs of Theorems

In this section, we present the proofs of Theorems 4.1 - 4.4. First, we list two lemmas. The proof of Lemma 4.1 can be found in Cai (2002) and the proof of Lemma 4.2 is relegated to Section 4.6.2.

Lemma 4.1: Under Assumptions A1 - A5, we have

$$\lambda = -h\lambda_0 \{1 + o_p(1)\} \quad \text{ and } \quad p_t(x) = n^{-1} b_t(x) \{1 + o_p(1)\},$$

where $\lambda_0 = \mu_2(W) g'(x) / \left[2\mu_2\left(W^2\right) g(x)\right]$ and $b_t(x) = \left[1 - h\lambda_0 \left(X_t - x\right) W_h \left(x - X_t\right)\right]^{-1}$. Further, we have

$$p_t(ch) = n^{-1} b_t^c(ch) \{1 + o_p(1)\},$$

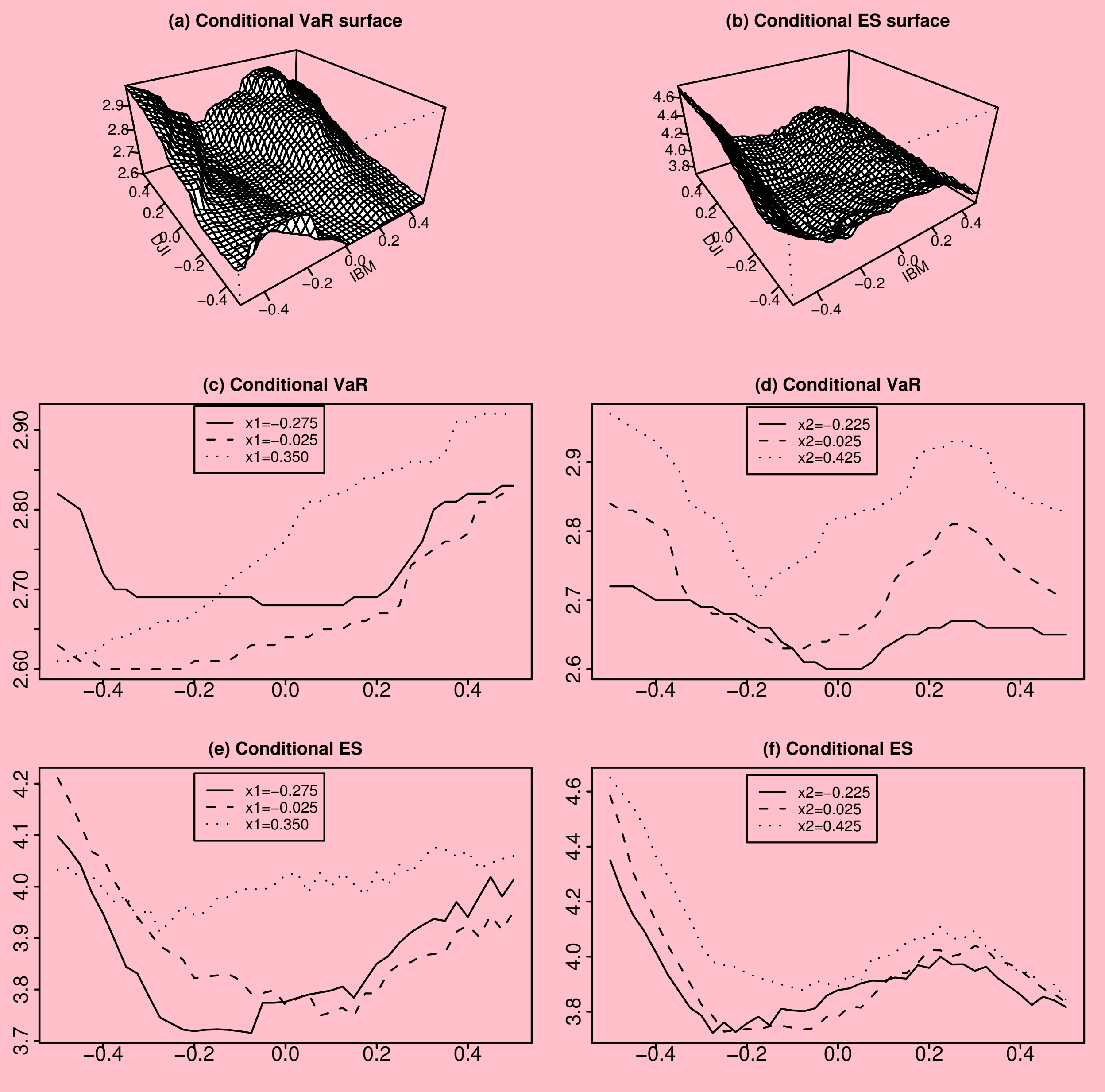


Figure 4.7: (a) 5% CVaR estimates for IBM stock returns. (b) 5% CES estimates for IBM stock returns index. (c) 5% CVaR estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (d) 5% CVaR estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$. (e) 5% CES estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (f) 5% CES estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$.

where $b_t^c(x) = \left[1 + \lambda_c \left(X_t - x\right) K_h \left(x - X_t\right)\right]^{-1}$.

Lemma 4.2: Under Assumptions A1 - A5, we have, for any $j \geq 0$,

$$J_j = n^{-1}\sum_{t=1}^{n} c_t(x)\left(\frac{X_t - x}{h}\right)^j = g(x)\mu_j(W) + O_p\left(h^2\right),$$

where $c_t(x) = b_t(x)W_h\left(x - X_t\right)$.

Before we start to provide the main steps for proofs of theorems. First, it follows from Lemmas 4.1 and 4.2 that

$$W_{c,t}(x,h) \approx \frac{b_t(x)W_h\left(x - X_t\right)}{\sum_{t=1}^{n} b_t(x)W_h\left(x - X_t\right)} \approx n^{-1}g^{-1}(x)b_t(x)W_h\left(x - X_t\right) = \frac{c_t(x)}{ng(x)}. \qquad (4.15)$$

Now we embark on the proofs of the theorems.

Proof of Theorem 4.1. By (4.7), we decompose $\widehat{f}_c(y \mid x) - f(y \mid x)$ into three parts as follows

$$\widehat{f}_c(y \mid x) - f(y \mid x) \equiv I_1 + I_2 + I_3, \qquad (4.16)$$

where with $\varepsilon_{t,1} = Y_t^*(y) - E\left(Y_t^*(y) \mid X_t\right)$,

$$I_1 = \sum_{t=1}^{n} \varepsilon_{t,1}W_{c,t}(x,h), \quad I_2 = \sum_{t=1}^{n}\left[E\left(Y_t^*(y) \mid X_t\right) - f\left(y \mid X_t\right)\right]W_{c,t}(x,h),$$

and

$$I_3 = \sum_{t=1}^{n}\left[f\left(y \mid X_t\right) - f(y \mid x)\right]W_{c,t}(x,h).$$

An application of the Taylor expansion, (4.7), (4.15), and Lemmas 4.1 and 4.2 gives

$$\begin{aligned} I_3 &= \sum_{t=1}^{n} \frac{1}{2} f^{0,2}(y \mid x)W_{c,t}(x,h)\left(X_t - x\right)^2 + o_p\left(h^2\right) \\ &= \frac{1}{2}g^{-1}(x)f^{0,2}(y \mid x)n^{-1}\sum_{t=1}^{n} c_t(x)\left(X_t - x\right)^2 + o_p\left(h^2\right) \\ &= \frac{h^2}{2}\mu_2(W)f^{0,2}(y \mid x) + o_p\left(h^2\right). \end{aligned}$$

By (4.2) and following the same steps as in the proof of Lemma 4.2, we have

$$I_2 = \frac{h_0^2\mu_2(K)}{2g(x)}n^{-1}\sum_{t=1}^{n} f^{2,0}\left(y \mid X_t\right)c_t(x) + o_p\left(h_0^2 + h^2\right) = \frac{h_0^2}{2}\mu_2(K)f^{2,0}(y \mid x) + o_p\left(h_0^2 + h^2\right).$$

Therefore,

$$I_2 + I_3 = \frac{h^2}{2}\mu_2(W)f^{0,2}(y \mid x) + \frac{h_0^2}{2}\mu_2(K)f^{2,0}(y \mid x) + o_p\left(h^2 + h_0^2\right) = B_f(y \mid x) + o_p\left(h^2 + h_0^2\right).$$

Thus, (4.16) becomes

$$\sqrt{nh_0h}\left[\widehat{f}_c(y \mid x) - f(y \mid x) - B_f(y \mid x) + o_p\left(h^2 + h_0^2\right)\right] = \sqrt{nh_0h}I_1$$
$$=g^{-1}(x)I_4\left\{1 + o_p(1)\right\} \quad \to \quad N\left\{0, \sigma_f^2(y \mid x)\right\}$$

where $I_4 = \sqrt{h_0h/n}\sum_{t=1}^n \varepsilon_{t,1}c_t(x)$. This, together with Lemma 4.3 in Section 4.6.2, therefore, proves the theorem. □

Proof of Theorem 4.2. Similar to (4.16), we have

$$\widehat{S}_c(y \mid x) - S(y \mid x) \equiv I_5 + I_6 + I_7, \tag{4.17}$$

where with $\varepsilon_{t,2} = \bar{G}_{h_0}(y - Y_t) - E\left(\bar{G}_{h_0}(y - Y_t) \mid X_t\right)$,

$$I_5 = \sum_{t=1}^n \varepsilon_{t,2}W_{c,t}(x,h), \quad I_6 = \sum_{t=1}^n \left[E\left\{\bar{G}_{h_0}(y - Y_t) \mid X_t\right\} - S(y \mid X_t)\right]W_{c,t}(x,h),$$

and

$$I_7 = \sum_{t=1}^n \left[S(y \mid X_t) - S(y \mid x)\right]W_{c,t}(x,h).$$

Similar to the analysis of I_2, by the Taylor expansion, (4.7), and Lemmas Lemmas 4.1 and 4.2, we have

$$\begin{aligned} I_7 &= \sum_{t=1}^n \frac{1}{2}S^{0,2}(y \mid x)W_{c,t}(x,h)\left(X_t - x\right)^2 + o_p\left(h^2\right) \\ &= \frac{1}{2}S^{0,2}(y \mid x)g^{-1}(x)n^{-1}\sum_{t=1}^n c_t(x)\left(X_t - x\right)^2 + o_p\left(h^2\right) \\ &= \frac{h^2}{2}\mu_2(W)S^{0,2}(y \mid x) + o_p\left(h^2\right). \end{aligned}$$

To evaluate I_6, first, we consider the following

$$E\left[\bar{G}_{h_0}(y - Y_t) \mid X_t = x\right] = \int_{-\infty}^{\infty} K(u)S(y - h_0u \mid x)\,du$$
$$=S(y \mid x) + \frac{h_0^2}{2}\mu_2(K)S^{2,0}(y \mid x) + o\left(h_0^2\right)$$
$$=S(y \mid x) - \frac{h_0^2}{2}\mu_2(K)f^{1,0}(y \mid x) + o\left(h_0^2\right). \tag{4.18}$$

By (4.18) and following the same arguments as in the proof of Lemma 4.2, we have

$$I_6 = -\frac{h_0^2\mu_2(K)}{2g(x)}n^{-1}\sum_{t=1}^n f^{1,0}(y \mid X_t)\,c_t(x) + o_p\left(h_0^2 + h^2\right) = -\frac{h_0^2}{2}\mu_2(K)f^{1,0}(y \mid x) + o_p\left(h_0^2 + h^2\right).$$

Therefore,

$$I_6+I_7=\frac{h^2}{2}\mu_2(W)S^{0,2}(y\mid x)-\frac{h_0^2}{2}\mu_2(K)f^{1,0}(y\mid x)+o_p\left(h^2+h_0^2\right)=B_S(y\mid x)+o_p\left(h^2+h_0^2\right),$$

so that by (4.17),

$$\sqrt{nh}\left[\widehat{S}_c(y\mid x)-S(y\mid x)-B_S(y\mid x)+o_p\left(h^2+h_0^2\right)\right]=\sqrt{nh}I_5.$$

Clearly, to accomplish the proof of theorem, it suffices to establish the asymptotic normality of $\sqrt{nh}I_5$. To this end, first, we compute $\operatorname{Var}\left(\varepsilon_{t,2}\mid X_t=x\right)$. Note that

$$\begin{aligned}E\left[\bar{G}_{h_0}^2\left(y-Y_t\right)\mid X_t=x\right]&=\int_{-\infty}^{\infty}\bar{G}_{h_0}^2(y-u)f(u\mid x)du\\=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K\left(u_1\right)K\left(u_2\right)S\left(\max\left(y-h_0u_1,y-h_0u_2\right)\mid x\right)du_1du_2\\=&S(y\mid x)+2h_0\alpha(K)f(y\mid x)+O\left(h_0^2\right),\end{aligned}\tag{4.19}$$

which, in conjunction with (4.18), implies that

$$\operatorname{Var}\left(\varepsilon_{t,2}\mid X_t=x\right)=S(y\mid x)[1-S(y\mid x)]+2h_0\alpha(K)f(y\mid x)+o\left(h_0\right).$$

This, together with the fact that

$$\operatorname{Var}\left(\varepsilon_{t,2}c_t(x)\right)=E\left[c_t^2(x)E\left\{\varepsilon_{t,2}^2\mid X_t\right\}\right]=E\left[c_t^2(x)\operatorname{Var}\left(\varepsilon_{t,2}\mid X_t\right)\right],$$

leads to

$$h\operatorname{Var}\left\{\varepsilon_{t,2}c_t(x)\right\}=\mu_0\left(W^2\right)g(x)\left[S(y\mid x)\{1-S(y\mid x)\}+2h_0\alpha(K)f(y\mid x)\right]+o\left(h_0\right).$$

Now, since $|\varepsilon_{t,2}|\leq 1$, by following the same arguments as those used in the proofs of Lemmas 4.2 and 4.3 in Section 4.6.2 (or Lemma 1 and Theorem 1 in Cai (2002)), we can show although tediously that

$$\operatorname{Var}\left(I_8\right)=\sigma_S^2(y\mid x)g^2(x)+2\mu_0\left(W^2\right)h_0\alpha(K)f(y\mid x)g(x)+o\left(h_0\right),\tag{4.20}$$

where $I_8=\sqrt{h/n}\sum_{t=1}^n\varepsilon_{t,2}c_t(x)$, and

$$\sqrt{nh}I_5=g^{-1}(x)I_8\left\{1+o_p(1)\right\}\quad\rightarrow\quad N\left\{0,\sigma_S^2(y\mid x)\right\}$$

This completes the proof of Theorem 4.2. □

Proof of Theorem 4.4. Similar to (4.12), we use the Taylor expansion and ignore the higher terms to obtain

$$\begin{aligned}&\int_{\widehat{\nu}_p(x)}^{\infty} yK_{h_0}\left(y-Y_t\right)dy \approx \int_{\nu_p(x)}^{\infty} yK_{h_0}\left(y-Y_t\right)dy - \nu_p(x)K_{h_0}\left(\nu_p(x)-Y_t\right)\left[\widehat{\nu}_p(x)-\nu_p(x)\right]\\ &= Y_t\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right) - \nu_p(x)K_{h_0}\left(\nu_p(x)-Y_t\right)\left[\widehat{\nu}_p(x)-\nu_p(x)\right] + h_0G_{1,h_0}\left(\nu_p(x)-Y_t\right).\end{aligned}$$

Plugging the above into (4.9) leads to

$$p\widehat{\mu}_p(x) \approx \widehat{\mu}_{p,1}(x) + I_9, \tag{4.21}$$

where

$$\widehat{\mu}_{p,1}(x) = \sum_{t=1}^{n} W_{c,t}(x,h)Y_t\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right) - \nu_p(x)\widehat{f}_c\left(\nu_p(x)\mid x\right)\left[\widehat{\nu}_p(x)-\nu_p(x)\right],$$

which will be shown later to be the source of both the asymptotic bias and variance, and

$$I_9 = h_0\sum_{t=1}^{n} W_{c,t}(x,h)G_{1,h_0}\left(\nu_p(x)-Y_t\right),$$

which will be shown to contribute only the asymptotic bias (see Lemma 4.4 in Section (4.7). From (4.12) and (4.8),

$$\widehat{f}_c\left(\nu_p(x)\mid x\right)\left[\widehat{\nu}_p(x)-\nu_p(x)\right] \approx \sum_{t=1}^{n} W_{c,t}(x,h)\left\{\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right)-p\right\}.$$

Therefore, by (4.15),

$$\begin{aligned}\widehat{\mu}_{p,1}(x) &= \sum_{t=1}^{n} W_{c,t}(x,h)\left[\left\{Y_t-\nu_p(x)\right\}\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right)-p\nu_p(x)\right]\\ &= \sum_{t=1}^{n} W_{c,t}(x,h)\varepsilon_{t,3} + \sum_{t=1}^{n} W_{c,t}(x,h)E\left\{\zeta_t(x)\mid X_t\right\}\\ &\approx g^{-1}(x)n^{-1}\sum_{t=1}^{n}\varepsilon_{t,3}c_t(x) + \sum_{t=1}^{n} W_{c,t}(x,h)E\left\{\zeta_t(x)\mid X_t\right\}\\ &\equiv \widehat{\mu}_{p,2}(x)+\widehat{\mu}_{p,3}(x),\end{aligned}$$

where $\zeta_t(x) = \left[Y_t-\nu_p(x)\right]\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right)+p\nu_p(x)$ and $\varepsilon_{t,3} = \zeta_t(x) - E\left\{\zeta_t(x)\mid X_t\right\}$. Next, we derive the asymptotic bias and variance for $\widehat{\mu}_{p,1}(x)$. Indeed, we will show that asymptotic bias of $\widehat{\mu}_p(x)$ comes from both $\widehat{\mu}_{p,3}(x)$ and I_9, and the asymptotic variance for $\widehat{\mu}_{p,1}(x)$ is only from $\widehat{\mu}_{p,2}(x)$. First, we consider $\widehat{\mu}_{p,3}(x)$. Now, it is easy to see by the Taylor expansion that

$$\begin{aligned} E\left[Y_t\bar{G}_{h_0}\left(\nu_p(x)-Y_t\right)\mid X_t=v\right] &= \int_{-\infty}^{\infty} K(u)du\int_{\nu_p(x)-h_0u}^{\infty} yf(y\mid v)dy \\ =&\int_{-\infty}^{\infty} l_1\left(\nu_p(x)-h_0u\mid v\right)K(u)du = l_1\left(\nu_p(x)\mid v\right)+\frac{h_0^2}{2}\mu_2(K)l_1^{2,0}\left(\nu_p(x)\mid v\right)+o\left(h_0^2\right) \\ =&l_1\left(\nu_p(x)\mid v\right)-\frac{h_0^2}{2}\mu_2(K)\left[\nu_p(x)f^{1,0}\left(\nu_p(x)\mid v\right)+f\left(\nu_p(x)\mid x\right)\right]+o\left(h_0^2\right), \end{aligned}$$

which, in conjunction with (4.18), leads to

$$\zeta(v)=E\left[\zeta_t(x)\mid X_t=v\right]=A\left(\nu_p(x)\mid v\right)-\frac{h_0^2}{2}\mu_2(K)f\left(\nu_p(x)\mid v\right)+o\left(h_0^2\right), \tag{4.22}$$

where $A\left(\nu_p(x)\mid v\right)=l_1\left(\nu_p(x)\mid v\right)-\nu_p(x)\left[S\left(\nu_p(x)\mid v\right)-p\right]$. It is easy to verify that $A\left(\nu_p(x)\mid v\right)=E\left[\{Y_t-\nu_p(x)\}I\left(Y_t\geq\nu_p(x)\right)\mid X_t=v\right]+p\nu_p(x)$, $A\left(\nu_p(x)\mid x\right)=p\mu_p(x)$, and $A^{0,2}\left(\nu_p(x)\mid x\right)=l_1^{0,2}\left(\nu_p(x)\mid x\right)-\nu_p(x)S^{0,2}\left(\nu_p(x)\mid x\right)$. Therefore, by (4.22), the Taylor expansion, and (4.7), $\widehat{\mu}_{p,3}(x)$ becomes

$$\widehat{\mu}_{p,3}(x)=\sum_{t=1}^{n}W_{c,t}(x,h)\zeta\left(X_t\right)=\zeta(x)+\frac{1}{2}\zeta''(x)\sum_{t=1}^{n}W_{c,t}(x,h)\left(X_t-x\right)^2+o_p\left(h^2\right).$$

Further, by Lemmas 4.1 and 4.2,

$$\begin{aligned} \widehat{\mu}_{p,3}(x)&=\zeta(x)+\frac{h^2}{2}\mu_2(W)\zeta''(x)+o_p\left(h^2\right) \\ &=p\mu_p(x)+\frac{h^2}{2}\mu_2(W)A^{0,2}\left(\nu_p(x)\mid x\right)-\frac{h_0^2}{2}\mu_2(K)f\left(\nu_p(x)\mid x\right)+o_p\left(h_0^2\right). \end{aligned}$$

This, in conjunction with Lemma 4.4 in Section 4.7 concludes that

$$\widehat{\mu}_{p,3}(x)+I_9=p\left[\mu_p(x)+B_\mu(x)\right]+o_p\left(h^2+h_0^2\right),$$

so that by (4.21),

$$\widehat{\mu}_{p,1}(x)-p\left[\mu_p(x)+B_\mu(x)\right]=\widehat{\mu}_{p,2}(x)+o_p\left(h^2+h_0^2\right),$$

and

$$\widehat{\mu}_p(x)-\mu_p(x)-B_\mu(x)=p^{-1}\widehat{\mu}_{p,2}(x)+o_p\left(h^2+h_0^2\right).$$

Finally, by Lemma 4.5 in Section 4.7, we have

$$\sqrt{nh}\left[\widehat{\mu}_p(x)-\mu_p(x)-B_\mu(x)+o_p\left(h^2+h_0^2\right)\right]=\frac{1}{pg(x)}I_{10}\left\{1+o_p(1)\right\} \quad\rightarrow\quad N\left\{0,\sigma_\mu^2(x)\right\},$$

where $I_{10}=\sqrt{h/n}\sum_{t=1}^{n}\varepsilon_{t,3}c_t(x)$. Thus, we prove the theorem. □

4.6.2 Proofs of Lemmas

In this section, we present the proofs of Lemmas 4.2 - 4.5. Note that we use the same notation as in Sections 4.2 - 4.6. Also, throughout this section, we denote a generic constant by C, which may take different values at different appearances.

Proof of Lemmas 4.2. Let $\xi_t = c_t(x)\left(X_t - x\right)^j / h^j$. It is easy to verify by the Taylor expansion that

$$E\left(J_j\right) = E\left(\xi_t\right) = \int \frac{v^j W(v) g(x-hv)}{1+h\lambda_0 v W(v)} dv = g(x)\mu_j(W) + O\left(h^2\right), \tag{4.23}$$

and

$$E\left(\xi_t^2\right) = h^{-1} \int \frac{v^{2j} W^2(v) g(x-hv)}{\left[1+h\lambda_0 v W(v)\right]^2} dv = O\left(h^{-1}\right).$$

Also, by the stationarity, a straightforward manipulation yields

$$n\mathrm{Var}\left(J_j\right) = \mathrm{Var}\left(\xi_1\right) + \sum_{t=2}^{n} l_{n,t} \mathrm{Cov}\left(\xi_1, \xi_t\right), \tag{4.24}$$

where $l_{n,t} = 2(n-t+1)/n$. Now decompose the second term on the right hand side of (4.24) into two terms as follows

$$\sum_{t=2}^{n} \left|\mathrm{Cov}\left(\xi_1, \xi_t\right)\right| = \sum_{t=2}^{d_n} (\cdots) + \sum_{t=d_n+1}^{n} (\cdots) \equiv J_{j1} + J_{j2}, \tag{4.25}$$

where $d_n = O\left(h^{-1/(1+\delta/2)}\right)$. For J_{j1}, it follows by Assumption A4 that $\left|\mathrm{Cov}\left(\xi_1, \xi_t\right)\right| \leq C$, so that $J_{j1} = O\left(d_n\right) = o\left(h^{-1}\right)$. For J_{j2}, Assumption A2 implies that $\left|\left(X_t - x\right)^j W_h\left(x - X_t\right)\right| \leq Ch^{j-1}$, so that $\left|\xi_t\right| \leq Ch^{-1}$. Then, it follows from the Davydov's inequality (see, e.g., Lemma 1.1) that $\left|\mathrm{Cov}\left(\xi_1, \xi_{t+1}\right)\right| \leq Ch^{-2}\alpha(t)$, which, together with Assumption A5, implies that

$$J_{j2} \leq Ch^{-2} \sum_{t \geq d_n} \alpha(t) \leq Ch^{-2} d_n^{-(1+\delta)} = o\left(h^{-1}\right).$$

This, together with (4.24) and (4.25), therefore implies that $\mathrm{Var}\left(J_j\right) = O\left((nh)^{-1}\right) = o(1)$. This completes the proof of the lemma. □

Lemma 4.3: Under Assumptions A1 - A6 with h in A3 and A6 replaced by $h\, h_0$, we have

$$I_4 = \sqrt{\frac{h_0 h}{n}} \sum_{t=1}^{n} \varepsilon_{t,1} c_t(x) \quad \rightarrow \quad N\left\{0, \sigma_f^2(y \mid x) g^2(x)\right\}.$$

Proof of Lemmas 4.3. It follows by using the same lines as those used in the proof of Lemma 4.2 and Theorem 1 in Cai (2002), omitted. The outline is described as follows. First, similar to the proof of Lemma 4.2, it is easy to see that

$$\operatorname{Var}\left(I_4\right)=h_0 h \operatorname{Var}\left(\varepsilon_{t, 1} c_t(x)\right)+h_0 h \sum_{t=2}^n l_{n, t} \operatorname{Cov}\left(\varepsilon_{1,1} c_1(x), \varepsilon_{t, 1} c_t(x)\right) . \tag{4.26}$$

Next, we compute $\operatorname{Var}\left(\varepsilon_{t, 1} \mid X_t=x\right)$. Note that

$$h_0 E\left[Y_t^*(y)^2 \mid X_t=x\right]=\int_{-\infty}^{\infty} K^2(u) f\left(y-h_0 u \mid x\right) d u=\mu_0\left(K^2\right) f(y \mid x)+O\left(h_0^2\right),$$

which, together with the fact that

$$\operatorname{Var}\left(\varepsilon_{t, 1} c_t(x)\right)=E\left[c_t^2(x) E\left\{\varepsilon_{t, 1}^2 \mid X_t\right\}\right]=E\left[c_t^2(x) \operatorname{Var}\left(\varepsilon_{t, 1} \mid X_t\right)\right]$$

and (4.2), implies that

$$h h_0 \operatorname{Var}\left(\varepsilon_{t, 1} c_t(x)\right)=\mu_0\left(K^2\right) \mu_0\left(W^2\right) f(y \mid x) g(x)+O\left(h_0^2\right)=\sigma_f^2(y \mid x) g^2(x)+O\left(h_0^2\right) .$$

As for the second term on the right hand side of (4.26), similar to (4.25), it is decomposed into two summons. By using Assumption A4 for the first summon and using the Davydov's inequality and Assumption A5 to the second summon, we can show that the second term on the right hand side of (4.26) goes to zero as n goes to infinity. Thus, $\operatorname{Var}\left(I_4\right) \rightarrow \sigma_f^2(y \mid x) g^2(x)$ by (4.26). To show the normality, we employ Doob's small-block and large-block technique (see, e.g., Ibragimov and Linnik, 1971, p. 316). Namely, partition $\{1, \ldots, n\}$ into $2 q_n+1$ subsets with large-block of size $r_n=\left\lfloor\left(n h h_0\right)^{1 / 2}\right\rfloor$ and small-block of size $s_n=\left\lfloor\left(n h h_0\right)^{1 / 2} / \log n\right\rfloor$, where $q_n=\left\lfloor n /\left(r_n+s_n\right)\right\rfloor$ with $\lfloor x\rfloor$ denoting the integer part of x. By following the same steps as in the proof of Theorem 1 in Cai (2002), we can accomplish the rest of proofs: the summands for the large-blocks are asymptotically independent, two summands for the small-blocks are asymptotically negligible in probability, and the standard Lindeberg-Feller conditions hold for the summands for the large-blocks. See Cai (2002) for details. So, the proof of the lemma is complete. □

Lemma 4.4: Under Assumptions A1 - A6, we have

$$I_9=h_0 \sum_{t=1}^n W_{c, t}(x, h) G_{1, h_0}\left(\nu_p(x)-Y_t\right)=h_0^2 \mu_2(K) f\left(\nu_p(x) \mid x\right)+o_p\left(h_0^2\right) .$$

Proof of Lemmas 4.4. Define $\xi_{t,1} = c_t(x)G_{1,h_0}(\nu_p(x) - Y_t)$. Then, by Lemma 4.1, $I_9 = I_{10}\{1 + o_p(1)\}$, where $I_{10} = g^{-1}(x)h_0\sum_{t=1}^{n}\xi_{t,1}/n$. Similar to (4.23),

$$\begin{aligned}E(\xi_{t,1}) &= E\left[c_t(x)E\left\{G_{1,h_0}(\nu_p(x) - Y_t) \mid X_t\right\}\right]\\&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{K(u)W(v)uS(\nu_p(x) - h_0u) \mid x)\, g(x - hv)}{1 + h\lambda_0 vW(v)}dudv\\&= h_0\mu_2(K)f(\nu_p(x) \mid x)\, g(x) + O\left(h_0h^2\right),\end{aligned}$$

and

$$E\left(\xi_{t,1}^2\right) = E\left[b_t^2(x)W_h^2(x - X_t)\, E\left\{G_{1,h_0}^2(\nu_p(x) - Y_t) \mid X_t\right\}\right] = O(h_0/h),$$

so that $\mathrm{Var}(\xi_{t,1}) = O(h_0/h)$. By following the same arguments in the derivation of $\mathrm{Var}(J_j)$ in Lemma 4.2, one can show that $\mathrm{Var}(I_{10}) = O\left((nh)^{-1}\right) = o(1)$. This proves the lemma. □

Lemma 4.5: Under Assumptions A1 - A4 and B2 - B5, we have Under Assumptions A1 - A6, we have

$$I_{10} = \sqrt{\frac{h}{n}}\sum_{t=1}^{n}\varepsilon_{t,3}c_t(x) \quad\to\quad N\left\{0, p^2g^2(x)\sigma_\mu^2(x)\right\}.$$

Proof of Lemmas 4.5. It follows by using the same lines as those used in the proof of Lemma 4.1 and Theorem 1 in Cai (2001), omitted. The main idea is as follows. First, similar to the proof of Lemmas 4.2 and 4.3, we will show by Assumptions B1 - B3 that

$$\mathrm{Var}(I_{10}) \quad\to\quad p^2\sigma_\mu^2(x)g^2(x). \tag{4.27}$$

Finally, we need to compute $\mathrm{Var}(\varepsilon_{t,3}c_t(x))$. Since

$$\mathrm{Var}(\varepsilon_{t,3}c_t(x)) = E\left[c_t^2(x)E\left\{\varepsilon_{t,3}^2 \mid X_t\right\}\right] = E\left[c_t^2(x)\mathrm{Var}(\zeta_t(x) \mid X_t)\right],$$

then, we first need to calculate $\mathrm{Var}(\zeta_t(x) \mid X_t)$. To this effect, by (4.22),

$$\begin{aligned}&\mathrm{Var}(\zeta_t(x) \mid X_t = v) = \mathrm{Var}\left[(Y_t - \nu_p(x))\,\bar{G}_{h_0}(\nu_p(x) - Y_t) \mid X_t = v\right]\\=&E\left[(Y_t - \nu_p(x))^2\,\bar{G}_{h_0}^2(\nu_p(x) - Y_t) \mid X_t = v\right] - \left[l_1(\nu_p(x) \mid v) - \nu_p(x)S(\nu_p(x) \mid v)\right]^2 + O\left(h_0^2\right).\end{aligned}$$

Similar to (4.19),

$$\begin{aligned}&E\left[(Y_t - \nu_p(x))^2\,\bar{G}_{h_0}^2(\nu_p(x) - Y_t) \mid X_t = v\right] = \int_{-\infty}^{\infty}G_{h_0}^2(\nu_p(x) - y)\,(y - \nu_p(x))^2 f(y \mid v)dy\\=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(u_1)\,K(u_2)\,\tau\left(\max(\nu_p(x) - h_0u_1, \nu_p(x) - h_0u_2) \mid v\right)du_1du_2\\=&\tau(\nu_p(x) \mid v) - 2h_0\tau^{1,0}(\nu_p(x) \mid v)\,\alpha(K) + O\left(h_0^2\right) = \tau(\nu_p(x) \mid v) + O\left(h_0^2\right),\end{aligned}$$

since $\tau^{1,0}\left(\nu_p(x) \mid v\right) = 0$, where $\tau(u \mid v) = l_2(u \mid v) - 2\nu_p(x) l_1(u \mid v) + \nu_p^2(x) S(u \mid v)$. Therefore,

$$\operatorname{Var}\left(\zeta_t(x) \mid X_t = v\right) = \operatorname{Var}\left[\left(Y_t - \nu_p(x)\right) I\left(Y_t \geq \nu_p(x)\right) \mid X_t = v\right] + O\left(h_0^2\right),$$

and

$$h \operatorname{Var}\left(\varepsilon_{t,3} c_t(x)\right) = \mu_0\left(W^2\right) \operatorname{Var}\left[\left(Y_t - \nu_p(x)\right) I\left(Y_t \geq \nu_p(x)\right) \mid X_t = x\right] g(x) + o(1).$$

Similar to Lemmas 4.2 and 4.3, clearly, we have,

$$\operatorname{Var}\left(I_{10}\right) = h \operatorname{Var}\left(\varepsilon_{t,3} c_t(x)\right) + h \sum_{t=2}^{n} l_{n,t} \operatorname{Cov}\left(\varepsilon_{1,3} c_1(x), \varepsilon_{t,3} c_t(x)\right),$$

and the first term on right hand side of the above equation converges to $p^2 \sigma_\mu^2(x) g^2(x)$. As for the second term on the right hand side of the above equation, similar to (4.25), it is decomposed into two summons. By using Assumptions A4 and B2 for the first summon and using the Davydov's inequality and Assumptions A5 and B3 to the second summon, we can show that the second term on the right hand side of the above equation goes to zero as n goes to infinity. Thus, (4.27) holds. To show the normality, we employ Doob's small-block and large-block technique (see, e.g., Ibragimov and Linnik, 1971, p. 316). Namely, partition $\{1, \ldots, n\}$ into $2q_n + 1$ subsets with large-block of size r_n and small-block of size s_n, where s_n is given in Assumption B4, $q_n = \lfloor n/\left(r_n + s_n\right) \rfloor$, and $r_n = \left\lfloor (nh)^{1/2}/\gamma_n \right\rfloor$ with γ_n satisfying followings: γ_n is a sequence of positive numbers $\gamma_n \to \infty$ such that $\gamma_n s_n/\sqrt{nh} \to 0$ and $\gamma_n (n/h)^{1/2} \alpha\left(s_n\right) \to 0$ by Assumption B4. By following the same steps as in the proof of Theorem 1 in Cai (2001) and using Assumption B5, we can accomplish the rest of proofs: the summands for the large-blocks are asymptotically independent, two summands for the small-blocks are asymptotically negligible in probability, and the standard Lindeberg-Feller conditions hold for the summands for the large-blocks. See Cai (2001) for details. Therefore, the lemma is proved. □

4.7 Computer Codes

Please see the files chapter4-1.r, chapter4-2.r, chapter4-3.r, and chapter4-4.r for making figures. If you want to learn the codes for computation, they are available upon request.

4.8 References

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, **26**(7), 1487-1503.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, **9**(3), 203-228.

Boente, G. and Fraiman, R. (1995). Asymptotic distribution of smoothers based on local means and local medians under dependence. *Journal of Multivariate Analysis*, **54**(1), 77-90.

Cai, Z. and Wang, X. (2006). Nonparametric methods for estimating conditional value-at-risk and expected shortfall. *Journal of Econometrics*, **147**(1), 120-230.

Cai, Z. (2001). Weighted Nadaraya-Watson regression estimation. *Statistics & probability letters*, **51**(3), 307-318.

Cai, Z. (2002). Regression quantiles for time series. *Econometric theory*, **18**(1), 169-192.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, **136**(1), 163-188.

Cai, Z. and L. Qian (2000). Local estimation of a biometric function with covariate effects. In *Asymptotics in Statistics and Probability (M. Puri, ed)*, 47-70.

Cai, Z. and Tiwari, R. C. (2000). Application of a local linear autoregressive model to bod time series. *Environmetrics*, **11**(3), 341-350.

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, **103**(484), 1595-1608.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of statistics*, **19**(2), 760-777.

Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, **6**(1), 87-107.

Chen, S. X. and Tang, C. Y. (2005). Nonparametric inference of value-at-risk for dependent financial returns. *Journal of Financial Econometrics*, **3**(2), 227-255.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, **26**(1), 271-292.

Cosma, A., Scaillet, O. and Von Sachs, R. (2007). Multivariate wavelet-based shape-preserving estimation for dependent observations. *Bernoulli*, **13**(2), 301-329.

Duffie, D. and Pan, J. (1997). An overview of value at risk. *The Journal of Derivatives*, **4**(3), 7-49.

Duffie, D. and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement, and Management.* Princeton University Press, Princeton, NJ.

Embrechts, P., C. Klüppelberg and T. Mikosch (1997). *Modeling Extremal Events For Finance and Insurance.* Springer-Verlag, New York.

Engle, R. F. and Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantile. *Journal of Business and Economics Statistics*, **22**(4), 367-381.

Fan, J. and Gijbels, I. (2018). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J. and Gu, J. (2003). Semiparametric estimation of value at risk. *Econometrics Journal*, **6**(2), 261-290.

Fan, J., Hu, T.-C. and Truong, Y. K. (1994). Robust non-parametric function estimation. *Scandinavian journal of statistics*, **21**(4), 433-446.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**(1), 189-206.

Frey, R. and McNeil, A. J. (2002). Var and expected shortfall in portfolios of dependent credit risks: Conceptual and practical insights. *Journal of Banking & Finance*, **26**(7), 1317-1334.

Hall, P., Wolff, R. C. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**(445), 154-163.

Hürlimann, W. (2003). A Gaussian exponential approximation to some compound Poisson distributions. *ASTIN Bulletin: The Journal of the IAA*, **33**(1), 41-55.

Ibragimov, I.A. and Yu. V. Linnik (1971). *Independent and Stationary Sequences of Random Variables. Groningen*, Walters-Noordhoff, Netherlands.

Jorion, P. (2001). *Value at Risk*, 2nd Edition. McGraw-Hill, New York.

Jorion, P. (2003). *Financial Risk Manager Handbook*, 2nd Edition. John Wiley & Sons, New York.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, **46**(1), 33-50.

Lejeune, M. G. and Sarda, P. (1988). Quantile regression: A nonparametric approach. *Computational Statistics & Data Analysis*, **6**(3), 229-239.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, **24**(2), 165-179.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA*, **27**(1), 117-137.

Morgan, J.P. (1996). *Risk Metrics - Technical Documents*, 4th Edition. New York.

Oakes, D. and Dasu, T. (1990). A note on residual life. *Biometrika*, **77**(2), 409-410.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, **2**(1), 21-42.

Roussas, G. G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Annals of Mathematical Statistics*, **40**(4), 1386-1400.

Roussas, G. G. (1991). Estimation of transition distribution function and its quantiles in Markov processes: Strong consistency and asymptotic normality. In *Nonparametric Functional Estimation and Related Topics*(pp. 443-462). Springer-Verlag, Berlin.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters*, **7**(5), 407-412.

Scaillet, O. (2004). Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance*, **14**(1), 115-129.

Scaillet, O. (2005). Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal*, **74**(1), 639-660.

Truong, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Annals of Statistics*, **17**(2), 606-617.

Truong, Y. K. and Stone, C. J. (1992). Nonparametric function estimation involving time series. *Annals of Statistics*, **20**(1), 77-97.

Yu, K. and Jones, M. (1997). A comparison of local constant and local linear regression quantile estimators. *Computational statistics & data analysis*, **25**(2), 159-166.

Yu, K. and Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**(441), 228-237.

Chapter 5

Nonparametric Regression Models with Integrated Covariates

5.1 Introduction

Nonparametric estimation techniques have become cornerstone research topics in statistics for the last three decades since they offer numerous advantages relative to parametric techniques and have more flexibility and robustness to functional form misspecification, and have been embraced by applied researchers in many fields; see the books by Fan and Gijbels (1996) and Fan and Yao (2003). Asymptotic theory underlying various nonparametric estimators and test statistics for many commonly used models have been well established for independent and identically distributed (iid) data and some weak and strong dependent time series. The only nonparametric asymptotic analysis when covariates are integrated or unit root (denoted by I(1)) time series that we are aware of includes the papers, to name just a few, by Phillips and Park (1998), Park and Hahn (1999), Chang and Martinez-Chombo (2003), Chang and Park (2003), Juhl (2005), Cai, Li and Park (2009), Xiao (2009), Phillips (2009a, 2009b) and Phillips (2009). Particularly, Phillips and Park (1998), Juhl (2005), Phillips (1999), and Wang and Phillips (2009a, 2009b) considered the case when the true data generating process is a linear unit root process, while Park and Hahn (1999), Chang and Martinez-Chombo (2003), Chang and Park (2003), Cai, Li and Park (2009), Xiao (2009), Cai and Wang (2014), Cai, Wang and Wang (2015), and Sun, Cai and Li (2013) studied the models linearized in the nonstationary variables.

In this chapter, for the observed data $\{(Y_t, Z_t)\}$ for $t = 1, \ldots, n$, I study a nonparametric regression function with integrated covariate as follows,

$$Y_t = \beta(Z_t) + \varepsilon_t, \tag{5.1}$$

where $E(\varepsilon_t|Z_t) = 0$, $\{\varepsilon_t\}$ is stationary and $\beta(\cdot)$ is an unknown regression function. Here, Z_t is an integrated process satisfying

$$Z_t = \rho\, Z_{t-1} + u_t, \tag{5.2}$$

where $\rho = 1$ and $\{u_t\}$ is a stationary sequence. Clearly, Z_t is persistent and nonstationary. Indeed, model (5.1) is not new in literature but its asymptotics developed in the present chapter is novel when Z_t is persistent and nonstationary. For example, if Z_t is stationary, model (5.1) has been studied extensively in the literature; see Fan and Gijbels (1996) and Fan and Yao (2003) for details, while it was investigated by Karlsen and Tjøstheim (2001) for Z_t being null recurrent time series and Karlsen, Myklebust and Tjøstheim (2007) for the ϕ-irreducible Markov chain time series and by Bandi (2002) and Cai, Jing, Kong and Liu (2017) for nearly integrated time series (ρ in (5.2) is assumed to be $\rho = 1 + c/n$ with $c < 0$). A functional coefficient type model with I(1) covariates and nonlinear cointegration is investigated by Cai, Li and Park (2009), Xiao (2009), Cai, Wang and Wang (2015), and Sun, Cai and Li (2013), respectively. Finally, note that Wang and Phillips (2009a, 2009b) considered the case to allow $E(\varepsilon_t|Z_t) \neq 0$ and Cai, Jing, Kong and Liu (2017) extended the situation to allow $\{u_t\}$ in (5.2) to be a long memory process and ρ to be nearly one. For simplicity of notation, I consider only one-dimensional case since extension to multivariate Z_t involves fundamentally no new ideas but complicated notations.

Model (5.1) might have a great potential in many applications. For example, in macroeconomics, a particular parametric form of (5.1) can be used for forecasting inflation rate based on some persistent and nonstationary covariates such as velocity of monetary supply; see Bachmeier, Leelahanon and Li (2006), which showed that the velocity is an I(1) process. Also, using a semiparametric regression model with integrated covariates, Sun, Cai and Li (2013) considered the purchasing power parity hypothesis using Canadian and U.S. price and exchange rate data. Indeed, they showed that the difference between the two countries' 10-year Treasury bond rates is an I(1) process. Finally, it can be employed for the predictability of stock returns using various lagged financial variables, such as the dividend yield, term and default premia, the dividend-price ratio, the earning-price ratio, the book-to-market ratio, and interest rates; see Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Bandi (2002), Torous, Valkanov, and Yan (2004), Campbell and Yogo (2006), Polk, Thompson, and Vuolteenho (2006), Rossi (2007), Cai and Wang (2014), Cai, Wang and Wang(2015), and Cai, Jing, Kong and Liu (2017), and among others. In fact, Campbell and Yogo (2006) showed that the 95% confidence intervals for ρ in (5.2) are $[0.957, 1.007]$ and $[0.939, 1.000]$ for the log dividend-price ratio and the log earnings-price ratio, respectively; see Panel A in Table 4 of Campbell and Yogo (2006). As advocated by Campbell and Yogo (2006), Bachmeier, Leelahanon and Li (2006), and Cai, Li and Park (2009), the predictive power of using integrated or nearly integrated (highly persistent) covariates in a regression model can be improved significantly due to less noise.

The main purpose of the current chapter is to estimate the nonparametric regression $\beta(\cdot)$ by using the local linear (polynomial) and local constant (Nadaraya-Watson) fitting schemes and the main contribution of present chapter to the literature is to derive the asymptotic theory for both estimators. For simplicity, the main results can be summarized as follows. First, the optimal rate of convergence is $n^{1/5}$ slower than the usual $n^{2/5}$ rate for stationary case. Consequently, the order of the asymptotic mean-squared error (AMSE) is $n^{-2/5}$ rather than the standard rate $n^{-4/5}$. The intuitive explanation to this phenomenon is that an I(1) time series takes longer to revisit levels in its range. Second, the asymptotic bias term, similar to the stationary case, is independent of the stationary density of the regressor and is due to the linear approximation, which is typical for a local linear fitting scheme; see, for example, Fan and Gijels (1996) for details. Third, the limiting distribution is a mixed-normal (conditional normal) with the asymptotic variance depending inversely on the local time of a Brownian motion in which the unit root series can be embedded. Furthermore, the integrated covariate requires the larger bandwidths. Indeed, the optimal (in the AMSE sense) bandwidth is $O_p(n^{-1/10})$ implying a larger optimal bandwidth than in conventional kernel regressions with stationary regressors where the optimal bandwidth is known to be $O(n^{-1/5})$. Clearly, the use of conventional bandwidth has the theoretical potential of undersmoothing in the presence of I(1) covariates. Finally, it is very interesting that both local linear and local constant estimators share exactly same asymptotic properties at both interior and boundary points.

5.2 Statistical Properties

5.2.1 Local Linear Estimation

$\beta(\cdot)$ is estimated using local linear fitting from observations $\{(Y_t,\, Z_t)\}_{t=1}^n$. Our motivation of using local linear fitting is its high statistical efficiency in an asymptotic minimax sense, design adaptation and automatic correction for edge effects, as discussed in Fan and Gijbels (1996). Although a general local polynomial technique is applicable as well, it is well known that the local linear fitting will suffice for many applications; see Fan and Gijbels (1996) for a very comprehensive discussion, and that the theory developed for the local linear estimator continues to hold for the local polynomial estimator with only slight modification. Another virtue of using local polynomials is that both the unknown functions as well as their derivatives can be estimated simultaneously. For simplicity, the focus is only on local linear estimation and leave the generalization for additional research.

It is assumed throughout this chapter that $\beta(\cdot)$ is twice continuously differentiable, so that at any given z, a local approximation is used as $\beta(Z_t) \simeq \beta(z) + \beta'(z)\,(Z_t - z)$, when Z_t is a neighborhood of z, where $\simeq$ denotes the first order Taylor approximation and $\beta'(z)$ is the first derivative of $\beta(z)$. Hence, (5.1) is approximated by

$$Y_t \simeq \theta_0 + (Z_t - z)\,\theta_1 + \varepsilon_t,$$

and it becomes a local linear model. Therefore, the locally weighted sum of squares is

$$\sum_{t=1}^n \left[Y_t - \theta_0 - (Z_t - z)\,\theta_1\right]^2\, K_h(Z_t - z), \tag{5.3}$$

where $K_h(z) = K(z/h)/h$, $K(\cdot)$ is the kernel function, and $h = h_n > 0$ is the bandwidth satisfying $h \to 0$ and $n\,h \to \infty$ as $n \to \infty$, which controls the amount of smoothing used in the estimation. By minimizing (5.3) with respect to θ_0 and θ_1, the local linear estimate of $\beta(z)$ is obtained and is denoted by $\widehat{\beta}(z)$, and the local linear estimator of the derivative of $\beta(z)$ is denoted by $\widehat{\beta}'(z)$. It is easy to show that the minimizer of (5.3) is given by

$$\begin{pmatrix} \widehat{\beta}(z) \\ \widehat{\beta}'(z) \end{pmatrix} = \left[\sum_{t=1}^n \begin{pmatrix} 1 \\ Z_t - z \end{pmatrix}^{\otimes 2} K_h(Z_t - z)\right]^{-1} \sum_{t=1}^n \begin{pmatrix} 1 \\ Z_t - z \end{pmatrix} Y_t\, K_h(Z_t - z), \tag{5.4}$$

where $A^{\otimes 2} = A\,A^T$ ($A^{\otimes 1} = A$) for a vector or matrix A.

5.2.2 Notations and Assumptions

Since Z_t is an I(1) process, it can be re-expressed as $Z_t = Z_0 + \sum_{s=1}^{t} u_s$, where $\{u_s\}$ is a stationary process with mean zero and variance σ_u^2. In what follows, it is assumed that the process $\{u_t\}$ is a stationary linear process as $u_s = \sum_{j=0}^{\infty} c_j\, \omega_{s-j}$, where ω_j is a white noise with mean zero and $\sigma_\omega^2 = \mathrm{Var}(\omega_j) < \infty$, and $\{c_j\}$ satisfies, for some $0 < \tau \le 1$,

$$\sum_{j=0}^{\infty} |c_j|^{\tau} < \infty, \qquad \text{and} \qquad \sum_{j=0}^{\infty} c_j = 1. \tag{5.5}$$

Then, $\sigma_u^2 = \mathrm{Var}(u_s) = \sigma_\omega^2 \sum_{j=0}^{\infty} c_j^2$ and $\mathrm{Cov}(u_s, u_{s+t}) = \sigma_\omega^2 \sum_{j=0}^{\infty} c_j\, c_{j+t}$ for any s and t. Note that the assumption on $\{u_t\}$ being a linear process is due to an application of some results from Jeganathan (2004). Of course, it can be relaxed at the cost of involving lengthier mathematical proofs. Clearly, one has

$$Z_t/\sqrt{n} = Z_0/\sqrt{n} + \frac{1}{\sqrt{n}} \sum_{s=1}^{[nr]} u_s$$

for $r = t/n$. An application of Donsker's theorem (see, for example, Theorem 14.1 in Billinsley (1999) for iid $\{u_t\}$ with the existence of the second moment of u_t) leads to

$$Z_t/\sqrt{n} \implies W_u(r), \tag{5.6}$$

where " $\implies$ " represents weak convergence, $W_u(\cdot) = \sigma_0\, W(\cdot)$ with $W(\cdot)$ being a standard Brownian motion on [0, 1] and $\sigma_0^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2} \sum_{t=1}^{n} u_t)$, which is assumed to exist and be finite. In particular, it follows from Merlevéde, Peligrad and Utev (2006) that (5.6) holds if $\{u_t\}$ is stationary strong mixing sequence and satisfies, for some $\delta_0 > 0$,

$$E|u_t|^{2+\delta_0} < \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} k^{(2+\delta_0)/\delta_0}\, \alpha(k) < \infty, \tag{5.7}$$

where $\alpha(\cdot)$ is the mixing coefficient; see, e.g., Hall and Heyde (1980) for the definition.

Define $\eta_{t,z} = (Z_t - z)/\sqrt{t}$ for any z and let $f_{t,z}(\cdot)$ denote the density of $\eta_{t,z}$. Also, let $f_{t,s,z}(\cdot,\cdot)$ represent the joint density function of $(\eta_{t,z}, \eta_{s,z})$. Furthermore, let $\mathcal{F}_t$ be the smallest sigma field generated by $\{(Y_s,\, Z_s)\}_{s=-\infty}^{t}$. The following assumptions are listed.

Assumptions:

(C1) $E(\varepsilon_t|Z_t, \mathcal{F}_{t-1}) = 0$, $E(\varepsilon_t^2|Z_t, \mathcal{F}_{t-1}) = \sigma_\epsilon^2$, $E(\varepsilon_t^4|Z_t, \mathcal{F}_{t-1}) < C$ a.s., and $\{u_t\}$ is a stationary and mixing process satisfying constraints as imposed by (5.5) and (5.7), where σ_ϵ^2, C and σ_0^2 are finite positive constants.

(C2) Both $f_{t,z}(\cdot)$ and $f_{t,s,z}(\cdot, \cdot)$ have bounded continuous derivative functions (for all t, s and fixed z).

(C3) $K(\cdot)$ is a kernel function with a finite support, say $[-1,\ 1]$ and it is symmetric.

(C4) $n\,h \to \infty$ and $n\,h^{10} = O(1)$.

Next, we discuss the above conditions. Condition C1 requires that $\{\varepsilon_t\}$ is a martingale difference process with conditional homogenous variance and a finite fourth moment. The martingale difference assumption can be relaxed to a mixing process, and the assumption on the conditional homogenous error can be relaxed to the case that $E(\varepsilon_t^2|Z_t)$ is non-constant, with a lengthier proof. C2 is a very mild assumption. Indeed, it is satisfied if $\{u_t\}$ is commonly assumed to be iid normal. Finally, Assumptions C3 and C4 are commonly imposed in the kernel estimation literature and Assumption C4 is satisfied for the optimal bandwidth $h = O(n^{-1/10})$ (see later).

Finally, the local time $L(t, x)$ for a standard Brownian motion is defined in (1.4), which is

$$L(t, x) = \lim_{\Delta \to 0} \frac{1}{2\Delta} \int_0^t I_{\{|W(s)-x| \le \Delta\}} ds, \quad 0 \le t \le 1, \quad \text{and} \quad x \in \mathbb{R},$$

where I_A is the indicator function of an event A and $W(\cdot)$ is a standard Brownian motion; see Karatzas and Shreve (1991), Phillips and Park (1998), and Park and Phillips (1999) for details. Finally, define

$$\mu_j(K) = \int u^j\, K(u)du \quad \text{and} \quad \nu_j(K) = \int u^j\, K^2(u)du.$$

Note that $L(t, z)$ can be consistently estimated by $S_{n,0}(z)$; see Lemma 5.1 in Section 5.5.

5.2.3 Asymptotic Results

Now, main result is stated below and the proof is relegated to Section 5.5.

Theorem 5.1:. *Under Assumptions (C1) – (C4), one has*

$$\sqrt{n^{1/2}\,h}\,\left[\widehat{\beta}(z)-\beta(z)-h^2B(z)\right]\xrightarrow{d} MN(\sigma_\beta^2),$$

where $B(z)=\mu_2(K)\beta''(z)/2$ and $MN(\sigma_\beta^2)$ is a mixed normal distribution with mean zero and variance $\sigma_\beta^2=\sigma_\varepsilon^2\,\sigma_0\,\nu_0(K)\,/L(1,0)$ with σ_ε^2 defined in Assumption C1.

Remark 5.1: *First, ξ_t is called to be a mixed normal with mean μ_t and covariance Σ_t if the conditional distribution of ξ_t given μ_t and Σ_t is $N(\mu_t,\,\Sigma_t)$; see Phillips and Park (1998) for details. Note that the asymptotic properties for $\widehat{\beta}'(z)$ can be obtained as the same fashion as those in Theorem 5.1 and omitted. By comparing the results in Theorem 5.1 and conventional findings in Fan and Gijbels (1996) and Fan and Yao (2003) for the stationary covariates, the new results can be summarized as follows. Clearly, $h^2\mu_2(K)\beta''(z)/2$ serves as the asymptotic bias, which is the same as that for stationary case when one uses a local linear estimation method; see Fan and Yao (2003). However, the convergence rate is the order of $n^{1/4}h^{1/2}$ much lower with a factor $n^{1/4}$ by comparing with that for stationary covariates. Also, the stochastic asymptotic variance is independent of the grid point z. Indeed, one can show that the results in Theorem 5.1 hold true as long as any $z=z_n$ satisfies $z_n/\sqrt{n}\to 0$ and $n^{1/4}\,h^{5/2}\,\beta''(z_n)=O(1)$; see Theorem 5.2 later. Furthermore, from the asymptotic bias and variance presented in Theorem 5.1, the stochastic AMSE is given by*

$$AMSE = Var + bias^2=\sigma_\beta^2\,n^{-1/2}h^{-1}+\frac{h^4}{4}\,\mu_2^2(K)\,[\beta''(z)]^2\,.$$

The minimization of the AMSE with respect to h yields the optimal bandwidth

$$h_{opt}=\left(\frac{\sigma_\beta}{\mu_2(K)|\beta''(z)|}\right)^{2/5}n^{-1/10}=O_p(n^{-1/10}),$$

which is stochastic and much larger than the conventional optimal bandwidth $h_{opt,s}=O(n^{-1/5})$ for the stationary case; see Fan and Yao (2003). Therefore, if $h_{opt,s}$ were be used in estimating $\beta(\cdot)$ in (5.1), the nonparametric estimator given in (5.6) would be under-smoothing. Hence, it would be a very interesting future research topic on how to select the data-driven (optimal) bandwidth theoretically and empirically.

Now, the focus is on investigating the asymptotic behaviors at boundaries. When Z_t is $I(1)$, it follows from (5.6) that when $z=a\sqrt{n}$ $(a\neq 0)$ and $r=t/n$,

$$P(Z_t\geq z)=P(Z_t\geq a\sqrt{n})\;\to\;P(W_u(r)\geq a/\sigma_u)=1-\Phi(a/\sqrt{r}\sigma_0)>0,$$

where $\Phi(\cdot)$ is the distribution of the standard normal random variable. This means that there is a great chance that $|Z_t|$ can take large values. In other words, an I(1) time series takes longer to revisit levels in its range. Now, the question is how the asymptotic behaviors of the estimator look like when z is large like $z = a\sqrt{n}$ for any fixed a. To this end, the following asymptotic results is obtained at boundary $z = a\sqrt{n}$ for any fixed a. However, the detailed proofs are not provided since they follow closely the same arguments as those used in the proof of Theorem 5.1.

Theorem 5.2: . *If Assumptions (C1) – (C4) hold and $n^{1/4}\, h^{5/2}\, \beta''(a\sqrt{n}) = O(1)$ for any a, then, one has*

$$\sqrt{n^{1/2}\, h}\, \left[\widehat{\beta}(a\sqrt{n}) - \beta(a\sqrt{n}) - h^2\, B(a\sqrt{n})\right] \xrightarrow{d} MN(\sigma_a^2),$$

where $MN(\sigma_a^2)$ is a mixed normal distribution with mean zero and variance $\sigma_a^2 = \sigma_\varepsilon^2\, \sigma_0\, \nu_0(K)/L(1, a/\sigma_0)$.

Remark 5.2: *Comparing Theorem 5.2 with Theorem 5.1, one can observe that the magnitude of the asymptotic variance of $\widehat{\beta}(\cdot)$ at the boundary points ($z = O(n^{1/2})$) differs from that for the interior points ($z = o(n^{1/2})$). This finding is different from its stationary counterpart; see Fan and Gijbels (1996) for the stationary case.*

5.2.4 Nadaraya-Watson Estimation

Now, the turn is to discussing the asymptotic properties for the local constant estimator of $\beta(\cdot)$. It is well documented that the Nadaraya-Watson estimator is given by

$$\widetilde{\beta}(z) = \sum_{t=1}^{n} Y_t\, K_h(Z_t - z) / \sum_{t=1}^{n} K_h(Z_t - z). \tag{5.8}$$

For $\widetilde{\beta}(z)$, the following theorem can be established.

Theorem 5.3: . *Under the assumptions of Theorem 5.1, both $\widetilde{\beta}(z)$ and $\widehat{\beta}(z)$ share the exact same asymptotic properties. That is,*

$$\sqrt{n^{1/2}\, h}\, \left[\widetilde{\beta}(z) - \beta(z) - h^2 B(z)\right] \xrightarrow{d} MN(\sigma_\beta^2),$$

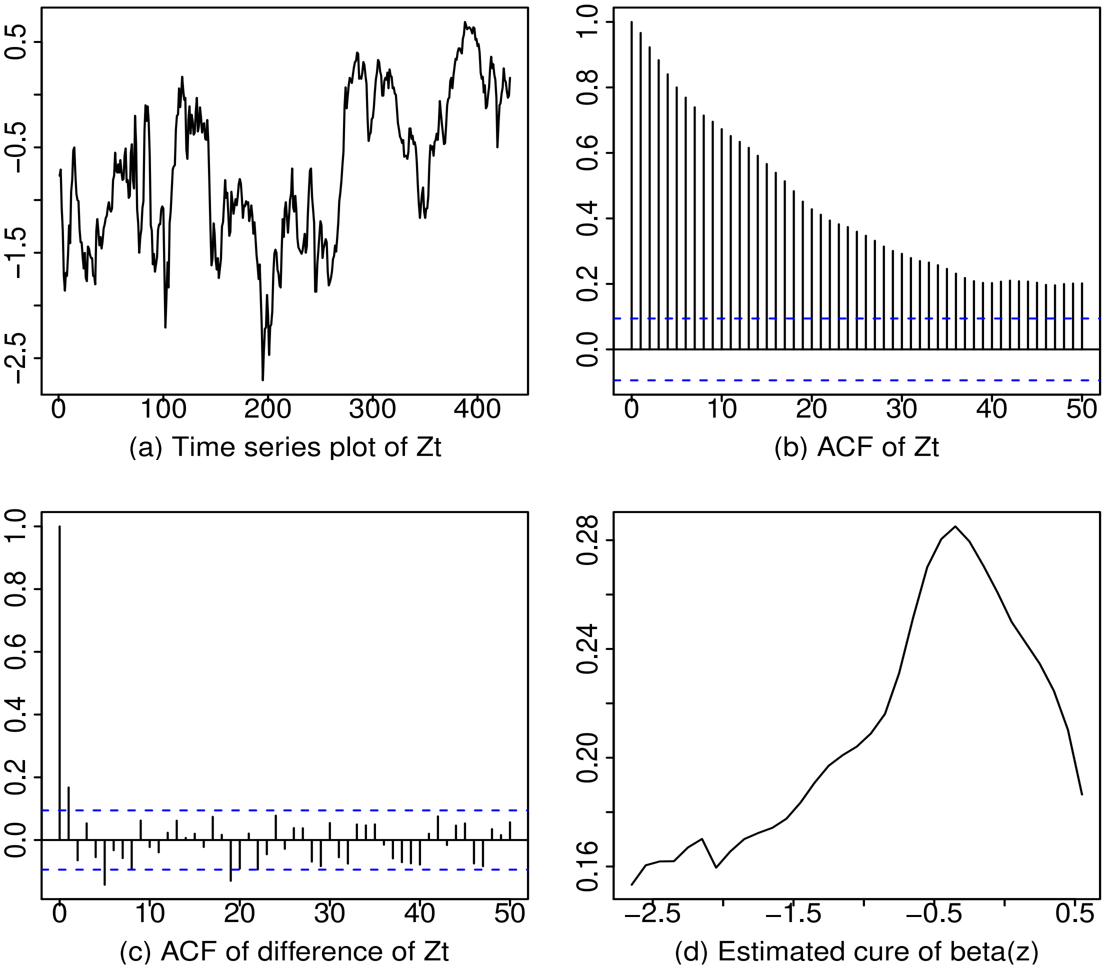
where $B(z) = \mu_2(K)\beta''(z)/2$ and $MN(\sigma_\beta^2)$ is a mixed normal distribution with mean zero and variance $\sigma_\beta^2 = \sigma_\varepsilon^2\, \sigma_0\, \nu_0(K)\, / L(1, 0)$. Further, Theorem 5.2 holds for $\widetilde{\beta}(z)$.

Remark 5.3: *It is clear that $h^2\mu_2(K)\beta''(z)/2$ serves as the asymptotic bias, which is the same as that case when one uses a local linear estimation method (see Theorem 5.1). However, for the stationary Z_t case with a local constant estimation method, there is an additional leading bias term which has the form of $h^2\mu_2(K)f_z'(z)\beta'(z)/2f_z(z)$, where $f_z(\cdot)$ is the stationary density of Z_t when Z_t is stationary; see Fan and Gijbels (1996). Theorem 5.3 shows that for non-stationary Z_t, the local constant estimator has the same leading bias as that of a local linear method. This is an interesting new finding that is not shared by a local constant estimator if Z_t is stationary. It can be shown that with nonstationary Z_t, the bias term associated with $f_{t,z}'(z)\beta'(z)$ has an order of $h^2n^{-1/2}\ln(n)$, which is smaller than h^2; see Lemma 5.4 in 5.5. Therefore, the leading bias contains only one term associated with $\beta''(z)$ with the order h^2. Interestingly, as in the case of standard local polynomial methods, the Nadaraya-Watson estimator is design-adaptive too in the sense of Fan and Gijbels (1996). Clearly, this property should be interpreted as follows. The clustered designs are not expected to occur in the presence of integrated (highly persistent) processes. Therefore, the theoretical relevance of the design-adaptation property and the theoretical appeal of local polynomial methods over the standard Nadaraya-Watson kernel estimates seem to vanish.*

5.3 An Illustrative Empirical Application

Sun, Cai and Li (2013) investigated the purchasing power parity (PPP) hypothesis using Canadian and U.S. price index and exchange rate data. The PPP theory says that the following setup holds $s_t = \beta_1 + \beta_1 p_t + \beta_2 p_t^* + u_t$, where s_t, p_t, and p_t^* are the logarithm of the nominal exchange rate expressed as Canadian dollars per unit of U.S. dollar, the Canadian and U.S. aggregate price levels, respectively. The aggregate price index is measured by the producer price index (PPI) base-weighted to the year 2000. Sun, Cai and Li (2013) used monthly data for the period from January 1974 to December 2009 so that there are 432 observations.

Sun, Cai and Li (2013) argued that based on the sticky-price theory of exchange rate determination, exchange rate movements also respond to monetary shocks. Due to sticky prices, the goods markets adjust to the monetary shocks slower than asset markets. Hence, in addition to the aggregate price levels, some other economic variables, such as interest rate differentials between two nations, also affect exchange rate formation and adjust more quickly to monetary shocks than the aggregate price indexes do. Therefore, to verify this economic theory, I will examine whether exchange rate depends on the interest rate differential between U.S. and Canada. Specifically, $Z_t = T_{US,t} - T_{CN,t}$ denotes the difference between the two countries' 10-year Treasury bond rates. The time series plot of Z_t is given in Figure 5.1(a) and its autocorrelation function (ACF) plot is displayed in Figure 5.1(b). Applying the augmented Dickey-Fuller (ADF) test statistic to the interest rate differential, I can not reject the null hypothesis $H_0 : \rho = 1$, so that Z_t follows a unit root process at the 5% significance level. Therefore, Z_t is treated as an integrated series. Indeed, one might evidence visually



(a) Time series plot of Zt

(b) ACF of Zt

(c) ACF of difference of Zt

(d) Estimated cure of beta(z)

Figure 5.1: (a) Time series plot of Z_t; (b) ACF plot of Z_t; (c) ACF plot of $Z_t - Z_{t-1}$; (d) Estimated curve of $\beta(z)$.

from Figures 5.1(a) and 5.1(b) that Z_t is a unit root process. Also, one can see from Figure 5.1(c) that $u_t = Z_t - Z_{t-1}$ is autocorrelated. Indeed, the Ljung-Box test rejects the null hypothesis of unautocorrelation of u_t.

Thus, for simplicity, the following nonparametric model is considered, by ignoring the price indices (p_t and p_t^*) for two countries

$$s_t = \beta(Z_t) + \varepsilon_t.$$

The Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ is used, and the smoothing parameter h is selected by the least squares cross-validation method so that $h = 0.425$. Figure 5.1(d) depicts the nonparametric estimate of $\beta(z)$. From Figure 5.1(d), it is very interesting to learn that $\widehat{\beta}(z)$ is nonlinear and reaches its max when $z = -0.35$. Further, it is increasing if $z \leq -0.35$ and then it is decreasing when $z > -0.35$. Also, it is asymmetric and the left side has a longer tail. The reaction of exchange rate to the the 10-year Treasury bond rate differential between the two nations is different based on the differential value. This means that when the U.S. 10-year Treasury bond rate becomes much lower or higher than that for the Canadian bond rate, the exchange rate between two nations becomes lower. In other words, the Canadian dollar is appreciated. Therefore, my analysis confirms that exchange rate between U.S. and Canada depends on the interest rate differential between two nations.

5.4 Discussion

This chapter studies a nonparametric regression model for integrated time series data by considering using the local polynomial local constant fitting schemes to estimate the nonparametric function and derives the asymptotic properties of the proposed estimators. The theoretical results show that the asymptotic bias has the same as that for stationary covariates. But, the convergence rate for the nonstationary covariates is slower than that for the stationary covariates by a factor of $n^{-1/4}$. Further, the asymptotic distribution is not normal any more but just a mixed normal associated with the local time of a standard Brownian motion. Moreover, it shows that the asymptotic properties for both the local linear and local constant estimators are exactly same.

It would like to mention some interesting future research topics related to this chapter. First, it would be very useful and important to discuss how to select the data-driven (optimal) bandwidth theoretically and empirically. Second, it should allow the errors $\{\varepsilon_t\}$ to be serially correlated time series, say α-mixing, to be heteroscedastic, and to be correlated with covariates as in Wang and Phillips (2009a, 2009b). Third, the model should include both stationary and nonstationary covariates. Finally, it is warranted to consider some extensions to other types of models like additive models, index models and varying coefficient models, and other types of nonstationarity such as nearly integrated processes; see, e.g., Bandi (2002), Torous, Valkanov and Yan (2005), Campbell and Yogo (2006), Polk, Thompson and Vuolteenaho (2006), Rossi (2007), Cai and Wang (2014), Cai, Wang and Wang (2015), and Cai, Jing, Kong and Liu (2017), which have a potential application in applied fields like economics and finance.

5.5 Proofs

Before proving the main results of this paper, we first give a few lemmas that will be used frequently in the proofs below. Throughout this section, C denotes a generic positive constant and it may take different values at different appearances.

To prove Theorem 5.1, define $G_j(u) = u^j\, K(u)$ for any $j \geq 0$. Then, it is easy to verify that $G_j(\cdot)$ is continuous and has a compact support. Also, both $G_j(\cdot)$ and $G_j^2(\cdot)$ are integrable. Also, define $S_n(z)$ as follows

$$S_n(z) = n^{-1/2} \sum_{t=1}^{n} K_h(Z_t - z) \begin{pmatrix} 1 \\ Z_{t,z,h} \end{pmatrix}^{\otimes 2} = \begin{pmatrix} S_{n,0}(z) & S_{n,1}(z) \\ S_{n,1}(z) & S_{n,2}(z \end{pmatrix}$$

where $Z_{t,z,h} = (Z_t - z)/h$ and for $0 \leq j \leq 2$,

$$S_{n,j}(z) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} K_{j,h}(Z_t - z)$$

with $K_{j,h}(u) = G_j(u/h)/h$. Then, re-express $S_{n,j}(z)$ as

$$S_{n,j}(z) = \frac{\beta_n}{n} \sum_{t=1}^{n} G_j(\beta_n(\gamma_n^{-1}\, Z_t + x_n)),$$

where $\beta_n = \sqrt{n}/h$, $\gamma_n = \sqrt{n}$, and $x_n = -z/\sqrt{n}$. Clearly, $x_n \to 0$ for any fixed z and $x_n = -a$ if $z = a\,\sqrt{n}$. Finally, let $\phi_\delta(x) = \exp\left(-x^2/2\,\delta^2\right)/\sqrt{2\,\pi\,\delta^2}$ for any $\delta > 0$ and $o_{L_2}(1)$ denote the convergence in L_2. Before proving the theorem, I first present some preliminary results. In what follows, it is assumed that Z_t satisfies (5.5).

Lemma 5.1: *Under assumption that the density of $\eta_{t,z}$ is bounded for all t,*

$$(i) \quad S_{n,j}(z) \xrightarrow{p} \begin{cases} \mu_j(K)\, L(1,0)/\sigma_0, & \text{if } z \text{ is fixed,} \\ \mu_j(K)\, L(1, a/\sigma_0)/\sigma_0, & \text{if } z = a\sqrt{n}, \end{cases}$$

and for any $p > 0$ and z,

$$(ii) \quad E\left[S_{n,j}(z)\right] = O(1), \qquad \textit{and} \qquad (iii) \quad E\left[|K_{j,h}(Z_t - z)|^p\right] = O(t^{-1/2}\, h^{1-p}).$$

Note that the above results still hold if fixed z is changed to be any z_n satisfying $z_n/\sqrt{n} \to 0$.

Proof: To establish the first assertion, I use some results from Jeganathan (2004). Indeed, by Proposition 6 and Lemma 7 of Jeganathan (2004), for each $\delta > 0$,

$$S_{n,j}(z) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(\gamma_n^{-1}\, Z_t + x_n) + o_{L_2}(1).$$

Since $\phi_\delta(z)$ satisfies the Lipschitz condition and $x_n \to 0$,

$$S_{n,j}(z) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(\gamma_n^{-1}\, Z_t) + o_{L_2}(1) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(W_u(t/n)) + o_{L_2}(1)$$

in view of (5.6) and (5.7). By Lemma 9 of Jeganathan (2004), one has

$$S_{n,j}(z) = \mu_j(K) \int_0^1 \phi_\delta(W_u(s))ds + o_{L_2}(1).$$

An application of Proposition 11 of Jeganathan (2004) gives

$$S_{n,j}(z) = \mu_j(K)\, L(1,0)/\sigma_0 + o_{L_2}(1)$$

as $\delta \downarrow 0$. By the same token, it is easy to show the case of $x_n = -a$ $(z = a\sqrt{n})$. For assertion (ii), one has

$$\begin{aligned} E\left[S_{n,j}(z)\right] &= n^{-1/2} \sum_{t=1}^{n} E\left[K_{j,h}(Z_t - z)\right] \\ &= n^{-1/2}\, h^{-1} \sum_{t=1}^{n} \int G_j(t^{1/2}u/h)\, f_{t,z}(u)du \\ &= n^{-1/2} \sum_{t=1}^{n} t^{-1/2} \int G_j(v)\, f_{t,z}(ht^{-1/2}v)dv \\ &\le C\, n^{-1/2} \sum_{t=1}^{n} t^{-1/2} = O(1). \end{aligned}$$

Finally, recall that $K_{j,h}(u) = h^{-1}G_j(u/h)$ and $G_j(u) = u^j\, K(u)$. It can be shown easily by the boundedness of $f_{t,z}(\cdot)$ that

$$\begin{aligned} E\left[|K_{j,h}(Z_t - z)|^p\right] = h^{-p}\int |K_{j,h}(t^{1/2}u/h)|^p f_t(u)du \\ = \quad t^{-1/2}h^{1-p}\int |G_j(v)|^p f_{t,z}(t^{-1/2}hv)dv \le Ct^{-1/2}h^{1-p}. \end{aligned}$$

This proves the lemma. □

By Lemma 5.1, one has

$$S_n(z) = \begin{pmatrix} S_{n,0}(z) & S_{n,1}(z) \\ S_{n,1}(z) & S_{n,2}(z) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & h^2\,\mu_2(K) \end{pmatrix}\; L(1,0)/\sigma_0\,\{1 + o_p(1)\},$$

which, by replacing Y_t in (5.4) by $Y_t = \beta(Z_t) + \varepsilon_t$, implies that

$$\widehat{\beta}(z) - \beta(z) \equiv \left[L(1,0)/\sigma_0\right]^{-1}\{B_n + C_n\}\,\{1 + o_p(1)\}, \qquad (5.9)$$

where $B_n = n^{-1/2}\sum_{t=1}^n\left[\beta(Z_t) - \beta(z) - \beta'(z)(Z_t - z)\right]K_h(Z_t - z)$ and $C_n = n^{-1/2}\sum_{t=1}^n \varepsilon_t$ $K_h(Z_t - z)$. I analyze B_n and C_n in Lemma B.2 and Lemma B.3 below.

Lemma 5.2:. *Under Assumptions given in Theorem 5.1, then,*

$$B_n = h^2 B(z)[S_{n,2}(z)] + o_p(h^2) = h^2 B(z)L(1,0)/\sigma_0 + o_p(h^2).$$

Proof: Note that the proof is similar to that for Lemma 5.1. Similar to Lemma 5.1, one can show that

$$\begin{aligned} B_n \quad &= \quad n^{-1/2}\sum_{t=1}^n\left[\beta(Z_t) - \beta(z) - \beta'(z)(Z_t - z)\right]K_h(Z_t - z) \\ &= \quad \frac{h^2}{2}\,\beta''(z)\,S_{n,2}(z)\{1 + o_p(1)\} \\ &= \quad \frac{h^2}{2}\;L(1,0)\,\beta''(z)\,\mu_2(K)/\sigma_0\{1 + o_p(1)\}. \end{aligned}$$

This completes the proof of Lemma 5.2. □

Lemma 5.3: *Under Assumptions given in Theorem 5.1, then,*

$$n^{3/4}h^{1/2}C_n \;\xrightarrow{d}\; MN(\sigma_1^2),$$

where $MN(\sigma_1^2)$ is a mixed normal with mean zero and covariance matrix $\sigma_1^2 = \sigma_\varepsilon^2\nu_0(K)$ $L(1,0)/\sigma_0$.

Proof: Clearly, $E[C_n] = 0$ since $E(\varepsilon_t | Z_t) = 0$. Also, by the assumptions that $\{\varepsilon_t\}$ is a martingale difference and $E(\varepsilon_t^2 | Z_t) = \sigma_\varepsilon^2$, one can conclude that the conditional variance of $n^{1/4} h^{1/2}\, C_n$, given $\{Z_t\}$, is

$$D_n = \frac{\sigma_\varepsilon^2\, h}{\sqrt{n}} \sum_{t=1}^{n} K_h^2(Z_t - z).$$

Similar to the proof of Lemma 5.1, one can show that

$$D_n = \sigma_\varepsilon^2\, \nu_0(K)\, L(1,0)/\sigma_0 + o_p(1).$$

Finally, by the central limit theorem for a martingale difference (see, e.g., Hall and Heyde (1980, p.58)), one obtains the conditional limiting distribution of C_n given $\{Z_t\}$,

$$n^{1/4} h^{1/2} C_n \xrightarrow{d} MN(\sigma_1^2).$$

This proves the lemma. □

Proof of Theorem 5.1: It is easy to check from Lemmas 5.1 and 5.2 that

$$C_n = h^2 B(z)\, L(1,0)/\sigma_0 + o_p(h^2).$$

Therefore, by (5.9) and Lemma 5.3, one has

$$\begin{aligned} & n^{1/4} h^{1/2} \left[\widehat{\beta}(z) - \beta(z) - h^2 B(z) + o_p(h^2) \right] \\ = \;& \sigma_0\ [L(1,0)]^{-1}\ n^{1/4} h^{1/2}\, C_n \{1 + o_p(1)\} \xrightarrow{d} MN(\sigma_\beta^2), \end{aligned}$$

which concludes the proof of the theorem. □

Proof of Theorem 5.3: It is easy to see from Lemma 5.1 that

$$\widetilde{\beta}(z) - \beta(z) \equiv \{E_n + C_n\}/S_{n,0}(z) = [L(1,0)/\sigma_0]^{-1} \{E_n + C_n\}\{1 + o_p(1)\},$$

where $E_n = n^{-1/2} \sum_{t=1}^{n} [\beta(Z_t) - \beta(z)]\, K_h(Z_t - z)$. Similar to Lemma 5.2, one has

$$E_n = \left[h\, \beta'(z)\, S_{n,1}(z) + \frac{h^2}{2}\, \beta''(z)\, S_{n,2}(z) \right] \{1 + o_p(1)\} = \frac{h^2}{2}\, L(1,0)\, \beta''(z)\, \mu_2(K)/\sigma_0 \{1 + o_p(1)\}$$

by Lemma 5.4 below. By Lemma 5.3, similar to the proof of Theorem 5.1, Theorem 5.3 is proved. □

Lemma 5.4:. *Under Assumptions given in Theorem 5.1, then,*

$$E[E_n] = O(h^2 n^{-1/2}\ln(n)) + O(h^2).$$

Proof: I first compute the following intermediate quantity. A simple calculation leads to

$$\begin{aligned} & E[(\beta(Z_t)-\beta(z))K_h(Z_t-z)] \\ = & \; t^{-1/2}\int [\beta(z+h\,v)-\beta(z)]\,K(v)\,f_{t,z}(t^{-1/2}hv)dv. \\ \approx & \; t^{-1/2}\int [\beta'(z)hv + h^2\beta''(z)v^2/2][f_{t,z}(0)+f'_{t,z}(0)t^{-1/2}hv]K(v)dv \\ = & \; h^2t^{-1}\beta'(z)f'_{t,z}(0)\mu_2(K) + \frac{1}{2}\,h^2t^{-1/2}\beta''(z)f_{t,z}(0)\mu_2(K), \end{aligned}$$

which implies that the order of the second term dominates the order of the first term. Therefore,

$$\begin{aligned} E[E_n] & \approx h^2\beta'(z)\mu_2(K)n^{-1/2}\sum_{t=1}^{n} t^{-1}f'_{t,z}(0) + \frac{1}{2}\,h^2\mu_2(K)\beta''(z)n^{-1/2}\sum_{t=1}^{n} t^{-1/2}f_{t,z}(0) \\ & = h^2\beta'(z)\mu_2(K)^{-1/2}n^{-1/2}O(\ln(n)) + \frac{1}{2}\,h^2\mu_2(K)\beta''(z)n^{-1/2}O(n^{1/2}) \\ & = O(h^2n^{-1/2}\ln(n)) + O(h^2). \end{aligned}$$

This concludes the proof of the lemma. □

5.6 References

Bachmeier, L., S. Leelahanon and Q. Li (2006). Money growth and inflation in the United States. *Macroeconomic Dynamics*, **11**, 113-127.

Bandi, F.M. (2002). On persistence and nonparametric estimation (with an application to stock return predictability). *Working paper*, Graduate School of Business, University of Chicago.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd Edition. Wiley, New York.

Cai, Z., B.-Y. Jing, X.-B. Kong and Z. Liu (2017). Nonparametric regression with nearly integrated regressors under long run dependence. *Econometrics Journal*, **20**, 118-138.

Cai, Z., Q. Li and J.Y. Park (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, **148**, 101-113.

Cai, Z. and Y. Wang (2014). Testing predictive regression models with nonstationary regressors. *Journal of Econometrics*, **178**, 4-14.

Cai, Z., Y. Wang and Y. Wang (2015). Testing instability in predictive regression model with nonstationary regressors. *Econometric Theory*, **31**, 953-980.

Campbell, J.Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, **81**, 27-60.

Cavanagh, C.L., G. Elliott and J.H. Stock (1995). Inference in models with nearly integrated regressors. *Econometric Theory*, **11**, 1131-1147.

Chang, Y. and E. Martinez-Chombo (2003). Electricity demand analysis using cointegration and error-correction models with time varying parameters: The Mexican case. *Working paper*, Department of Economics, Indiana University.

Chang, Y. and J. Park (2003). Index models with integrated time series. *Journal of Econometrics*, **114**, 73-106.

Elliott, G. and J.H. Stock (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory*, **10**, 672-700.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Hall, P. and C.C. Heyde (1980). *Martingale Limit Theory and its Applications.* Academic Press, New York.

Jeganathan, P. (2004). Convergence of functionals of sums of random variables to local times of fractional stable motions. *Annals of Probability*, **32**, 1771-1795.

Juhl, T. (2005). Functional coefficient models under unit root behavior. *Econometrics Journal*, **8**, 197-213.

Karatzas, I. and S.E. Shreve (1991). *Brownian Motion and Stochastic Calculus*, Second Edition. Springer-Verlag, New York.

Karlsen, H.A., T. Myklebust and D. Tjøstheim (2007). Nonparametric estimation in a nonlinear cointegration type model. *Annals of Statistics*, **35**, 252-299.

Karlsen, H. and D. Tjøstheim (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics*, **29**, 372-416.

Merlevéde, F., M. Peligrad and S. Utev (2006). Recent advances in invariance principles for stationary sequences. *Probability Surveys*, **3**, 1-36.

Park, J.Y. and S.B. Hahn (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory* **15**, 664-703.

Park, J.Y. and P.C.B. Phillips (1999). Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory*, **15**, 269-298.

Phillips, P.C.B. (2009). Local limit theory and spurious nonparametric regression. *Econometric Theory*, **25**, 1466-1497.

Phillips, P.C.B. and J. Park (1998). Nonstationary density and kernel autoregression. *Cowles Foundation discuss paper No. 1181.*

Polk, C, S. Thompson and T. Vuolteenaho (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics*, **81**, 101-141.

Rossi, B. (2007). Expectation hypothesis tests and predictive regressions at long horizons. *Econometrics Journal*, **10**, 1-26.

Sun, Y., Z. Cai and Q. Li (2010). Semiparametric functional coefficient models with integrated covariates. *Econometric Theory*, **29**, 659-672.

Torous, W., R. Valkanov and S. Yan (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business*, **77**, 937-966.

Wang, Q. and P.C.B. Phillips (2009a). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory*, **25**, 710-738.

Wang, Q. and P.C.B. Phillips (2009b). Structural nonparametric cointegrating regression. *Econometrica*, **77**, 1901-1948.

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics*, **152**, 81-92.