TI 2005-067/4
Tinbergen Institute Discussion Paper

Bahadur Representation for the Nonparametric *M*-Estimator Under Alpha-mixing Dependence

Yebin Cheng
Jan G. de Gooijer

Department of Quantitative Economics, Faculty of Economics and Econometrics, University of Amsterdam, and Tinbergen Institute.

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Bahadur Representation for the Nonparametric M-Estimator Under α-mixing Dependence

Yebin Cheng[1] and Jan G. De Gooijer[2*]

[1] Tinbergen Institute and Department of Quantitative Economics,
University of Amsterdam,
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
Telephone: +31-20-525 7110; Fax: +31-20-525 4349
e-mail: y.cheng@uva.nl

[2] Department of Quantitative Economics, University of Amsterdam,
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
Telephone: +31-20-525 4244; Fax: +31-20-525 4349
e-mail: j.g.degooijer@uva.nl

Abstract: Under the condition that the observations, which come from a high-dimensional population (X, Y), are strongly stationary and strongly-mixing, through using the local linear method, we investigate, in this paper, the strong Bahadur representation of the nonparametric M-estimator for the unknown function $m(x) = \arg\min_a \mathbb{E}(\rho(a, Y)|X = x)$, where the loss function $\rho(a, y)$ is measurable. Furthermore, some related simulations are illustrated by using the cross validation method for both bivariate linear and bivariate nonlinear time series contaminated by heavy-tailed errors. The M-estimator is applied to a series of S&P 500 index futures and spot prices to compare its performance in practice with the "usual" squared-loss regression estimator.

AMS 2000 subject classification: 62G20, 62G05.

Key words and phrases: Asymptotic representation; Kernel function; Robust estimator; Strongly-mixing.

[*]Corresponding author

1 Introduction

Assume that a sequence of random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, which comes from the population (X, Y) taking value in the space $I\!R^d \times I\!R^p$, is a strongly stationary and α-mixing process with the α-mixing coefficients $\alpha(n) \to 0$ as $n \to \infty$. Let $m(x) = I\!E\,(Y|X = x)$. Estimation of the regression function $m(x)$ plays an important role in statistics. Several methodologies have been proposed for this purpose including kernel smoothing methods, regression splines, local polynomial fitting, etc. During the past decades, much attention was also given to robust estimation of $m(x)$, especially when the sample is contaminated by heavy-tailed errors. When $p = 1$ and $\rho(a, y) = \| y - a \|^2$, the squared-loss nonparametric regression problem comes down to calculating the minimum value of the function $I\!E\,(\rho\,(a, Y)\,|X = x)$ on a, that is

$$m\,(x) = \arg\min_a I\!E\,(\rho\,(a, Y)\,|X = x)\,. \tag{1.1}$$

In a nonparametric setting this leads to the popular Nadaraya-Watson estimator. Besides the quadratic loss function $\rho(a, y)$, it is also possible to consider other functionals in (1.1), where the real function $\rho(a, y)$ is measurable with respect to a variable y for each fixed a. Such robust regression estimators have been investigated extensively in the literature, see, for instance, Fan, Hu & Truong (1994), Masry (1996a, 1996b), Masry & Fan (1997), Jiang & Mack (2001), Kozek & Pawlak (2002), and Cai (2003). In most of these papers the asymptotic normality for the proposed nonparametric M-estimator were considered when $\rho(a, y)$ is convex or non-convex on a. This applies to both i.i.d. and non-i.i.d., especially α-mixing, random variables. Also, some other topics related to the p-th conditional quantile, where $0 < p < 1$ and the corresponding $\rho(a, y)$ is equal to $|y - a| + (2p - 1)(y - a)$, can be found in Berlinet, Gannoun & Matzner-Løber (2001). Besides its asymptotic distribution, the Bahadur representation of the nonparametric M-estimator is another interesting topic since it not only provides a kind of asymptotic representation for the estimator but also gives the convergence rate for the remainder term. Using the local polynomial method and under the condition that both X and Y are in the univariate space, Hong (2003) obtained the strong and weak Bahadur representation of the nonparametric M-estimator for i.i.d. random samples. A similar problem, but for the p-th conditional quantile, was studied by Honda (2000).

In this paper, however, under the condition that the random sample is both α-mixing and is in a high-dimensional space, we will address the strong Bahadur representation for the nonparametric M-estimator of the regression function by using the local linear method. Suppose that the first order partial derivative $m'(x)$ of $m(x)$ exists. Then the corresponding local linear estimator $(\hat{m}_n(x), \hat{m}_n'(x))$ for the unknown function $(m(x), m'(x))$ can be defined by

$$\left(\hat{m}_n(x), \hat{m}_n'(x)\right) = \arg\min_{(a,b)} \sum_{i=1}^{n} w_{n,i} \rho\left(a + b\left(X_i - x\right), Y_i\right), \tag{1.2}$$

where $w_{n,i} = K\left(\frac{x-X_i}{h_n}\right) / \sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)$ is the weight function, $K(\cdot)$ is the kernel function and h_n is the bandwidth. Compared to the papers mentioned above, our result is interesting on its own since many stochastic processes are neither i.i.d. nor univariate. Also, our proof is different from that of Hong (2003). The basic idea of the proof comes from Jureckova & Sen (1996, Chapter 5), in which asymptotic representations for M-estimators were investigated in the case of univariate i.i.d. random variables. For the nonparametric problem related to α-mixing data, we refer to Bosq (1998) and Liebscher (1996, 2001).

The plan of the paper is as follows. In Section 2, we will establish the strong Bahadur representation of the nonparametric M-estimator and give its proof. From this, asymptotic normality will be obtained. In Section 3 some simulation results will be presented on the nonparametric M-estimator and the estimator obtained from (1.1) for both bivariate linear and bivariate nonlinear time series processes contaminated with heavy-tailed error distributions. We find that, in terms of the mean absolute deviation error (MADE), the nonparametric M-estimator outperforms the squared-loss nonparametric regression estimator. Both estimators are applied in Section 4 to a real example of heavy-tailed data in finance. Finally, in the appendix, some results on the nonparametric M-estimator for α-mixing processes are given.

Throughout this paper, denote by $\psi(a, y) = \partial\rho(a, y)/\partial a$, $G(a, b) = I\!E\left[\psi\left(m(x) + a, Y\right) \middle| X = b\right]$, $G_1(a, b) = \partial G(a, b)/\partial a^T$ and $G_2(a, b) = \partial G(a, b) / \left(\partial a \partial a^T\right)$. Also, assume that the density functions of (X, Y) and X are $f(x, y)$ and $f(x)$, respectively. $\|\cdot\|$ is the usual Euclidean norm for the corresponding matrix. Without otherwise specified, all limit relationships in this paper refer to $n \to \infty$.

2 Main Result and Its Proof

We first list some required conditions and then give our results and their corresponding proofs.

A1. *The density function $f(\cdot)$ is continuous at the point x and $f(x) > 0$.*

A2. *The kernel function $K(\cdot)$ is bounded from above. $\int K(z)dz = 1$. The support set of $K(\cdot)$ is contained in $[-1, 1]^d$.*

A3. *It holds that $\alpha_n = O\left(n^{-\theta}\right)$, $\sqrt{nh_n^d} \sim n^{\gamma}$, $\theta > 6$ and $\frac{1}{2} > \gamma > \frac{6}{6+\theta}$.*

A4. *$G_2(a,b)$ is bounded in the neighborhood of $(0,x)$.*

A5. $\int zK(z)\,dz = 0$, $\int zz^T K(z)\,dz > 0$ and $G_1(0,x) > 0$. Also, $G_1(a,b)$ is continuous in the neighborhood of $(0,x)$.

A6. *$\psi(\cdot)$ satisfies the Lipschitz condition of order one.*

A7. *Denote by $L(s,t) = I\!E\left(\|\psi(m(x)+s,Y)\|^r \,|\, X=t\right)$. Assume that there exists some $r > 4$ such that the function $L(s,t)$ is bounded in the neighborhood of $(0,x)$.*

A8. $\theta > \frac{4r+2}{r-4}$.

A9. $\frac{1}{2} > \gamma > \frac{4r-2+\theta}{4r-6+(r-2)\theta}$.

A10. *The function $L_j(s_1,s_2,t_1,t_2)$ is defined by*

$$I\!E\left(\left\|\psi(m(x)+s_1,Y_1)\psi^T(m(x)+s_2,Y_{j+1})\right\| \,\middle|\, X_1=t_1, X_{j+1}=t_2\right),$$

which is bounded from above in the neighborhood of $(0,0,x,x)$. The joint density of X_1 and X_{j+1} is bounded for all $j \geq 2$.

A11. *The second partial derivative of $m(x)$ exists and is bounded in the neighborhood of x.*

Our main result is as follows.

Theorem 2.1. *Suppose that Conditions A1–A9 hold. Then the following strong Bahadur representation*

$$\begin{aligned}
&\sqrt{nh_n^d}\,\mathrm{vec}\left(\hat{m}(x)-m(x), h_n\left(\hat{m}'(x)-m'(x)\right)\right)\\
= \;& -H^{-1}\sqrt{nh_n^d}\sum_{i=1}^n w_{n,i}\eta_{n,i} + O\left(\varepsilon_n + \frac{\log n}{\sqrt{nh_n^d}}\right) \qquad (2.1)
\end{aligned}$$

holds almost surely for all n sufficiently large, where

$$\eta_{n,i} = \begin{pmatrix} 1 \\ \frac{X_i-x}{h_n} \end{pmatrix} \otimes \psi(m(x)+m'(x)(X_i-x),Y_i),$$

$$\varepsilon_n = \frac{n^{\left(\frac{3}{2}-\frac{1}{2\theta}\right)/(pd+d)}(\log n)^{\frac{2}{(p+1)d}+\frac{1}{2}}}{(nh_n^d)^{\frac{\theta^2-1}{2\theta(p+1)d}}} \to 0 \qquad (2.2)$$

and

$$H = \int \begin{pmatrix} 1 & z^T \\ z & zz^T \end{pmatrix} \otimes G_1\left(h_n m'(x)z, x+h_n z\right) f(x+h_n z)\,K(z)\,dz.$$

Proof. For simplicity of notation, denote by $t = \mathrm{vec}\,(a, b)$,

$$\begin{aligned} \xi_{n,i}\,(t) &= w_{n,i} \begin{pmatrix} 1 \\ \frac{X_i - x}{h_n} \end{pmatrix} \otimes \Big\{ \psi\Big(m(x) + m'(x)(X_i - x) \\ &\quad + \frac{a + b\frac{X_i - x}{h_n}}{\sqrt{nh_n^d}}, Y_i \Big) - \psi(m(x) + m'(x)(X_i - x) \Big\}, \end{aligned}$$

$$I(t) = \sqrt{nh_n^d} \sum_{i=1}^{n} w_{n,i} \psi \left(m\,(x) + m'\,(x)\,(X_i - x) + \frac{a + b\frac{X_i - x}{h_n}}{\sqrt{nh_n^d}}, Y_i \right)$$

and $Z_{n,i}(t) = \xi_{n,i}\,(t) \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)$. From Condition $A5$ and the bound of the support set of $K\,(\cdot)$, we know that

$$H \to f\,(x) \begin{pmatrix} 1 & 0 \\ 0 & \int zz^T K\,(z)\,dz \end{pmatrix} \otimes G_1\,(0, x) = H_1 > 0.$$

Then, it can be inferred that $\inf_{\|t\| = M\sqrt{\log n}} t^T H t \geq \frac{1}{2} M^2 \lambda_0 \log n > 0$ for all sufficiently large n and some suitable constant $M > 0$ such that $M\lambda_0 > 2$, where λ_0 is the smallest eigenvalue of the matrix of H_1. From this, Lemma 2.4 and Lemma 3.2 below, it holds subsequently for all large n that

$$\begin{aligned} & \inf_{\|t\| = M\sqrt{\log n}} t^T I(t) \geq \inf_{\|t\| = M\sqrt{\log n}} t^T \left(\sqrt{nh_n^d} \sum_{i=1}^{n} \xi_{n,i}\,(t) - Ht \right) \\ & + \inf_{\|t\| = M\sqrt{\log n}} t^T H t + \inf_{\|t\| = M\sqrt{\log n}} \sqrt{nh_n^d} t^T \sum_{i=1}^{n} w_{n,i} \eta_{n,i} \\ \geq\ & -M\sqrt{\log n} \sup_{\|t\| = M\sqrt{\log n}} \left\| \sqrt{nh_n^d} \sum_{i=1}^{n} \xi_{n,i}\,(t) - Ht \right\| + \frac{\lambda_0}{2} M^2 \log n \\ & - M\sqrt{\log n}\sqrt{nh_n^d} \sup_{\|t\| = M\sqrt{\log n}} \left\| \sum_{i=1}^{n} w_{n,i} \eta_{n,i} \right\| \\ \geq\ & -M\sqrt{\log n}\,\varepsilon_n + \frac{\lambda_0}{2} M^2 \log n - M \log n > 0. \end{aligned}$$

According to Theorem 6.3.4 of Ortega and Rheinboldt (1970), it can be concluded that for all n sufficiently large, the system of equations $I(t) = 0$ has a root that lies in the sphere $\|t\| = M\sqrt{\log n}$ with probability one. Substituting this root into (2.7) below we see that (2.1) holds. □

Corollary 2.2. *Suppose that Conditions A1–A11 hold. Then the following asymptotic normality in distribution*

$$\begin{aligned} & \sqrt{nh_n^d} \left(\mathrm{vec}\left(\hat{m}(x) - m\,(x)\,, h_n \left(\hat{m}'\,(x) - m'\,(x) \right) \right) + h_n^2 H_1^{-1} \mu\,(x) + o\left(h_n^2\right) \right) \\ & \qquad \to N\left(0, H_1^{-1} H_2 H_1^{-1}\right) \end{aligned} \tag{2.3}$$

holds, where

$$\mu(x) = \int \begin{pmatrix} 1 \\ z \end{pmatrix} \otimes \left[G_2(0,x) \left(z^T \otimes I_d \otimes z^T \right) \mathrm{vec} \left(\frac{\partial^2 m(x)}{\partial x \partial x^T} \right) \right] K(z)\, dz$$

and

$$H_2 = f(x) \int K^2(z) \begin{pmatrix} 1 & z^T \\ z & zz^T \end{pmatrix} dz \otimes I\!E \left(\psi(m(X_i), Y_i) \psi^T(m(X_i), Y_i) \middle| X_i = x \right).$$

Proof. From the Taylor expansion and Condition $A11$, we know that

$$m(x) + h_n m'(x) z - m(x + h_n z) = -h_n^2 \left(z^T \otimes I_d \otimes z^T \right) \mathrm{vec} \left(\frac{\partial^2 m(x)}{\partial x \partial x^T} \right) + o\left(h_n^2\right).$$

From the definition on $m(x + h_n z)$, it can be seen that

$$\int \psi(m(x + h_n z), y) f_{Y|X}(y | X_i = x + h_n z)\, dy = 0.$$

Henceforth, from the two relationships above and the properties on $G_1(a, b)$, it holds that

$$\begin{aligned} & \int \psi(m(x) + h_n m'(x) z, y) f_{Y|X}(y | X_i = x + h_n z)\, dy \\ = \quad & -h_n^2 G_1(0,x) \left(z^T \otimes I_d \otimes z^T \right) \mathrm{vec} \left(\frac{\partial^2 m(x)}{\partial x \partial x^T} \right) + o\left(h_n^2\right). \end{aligned}$$

Therefore, it can be inferred that

$$\frac{I\!E K\left(\frac{x - X_i}{h_n}\right) \eta_{n,i}}{I\!E K\left(\frac{x - X_i}{h_n}\right)} = -h_n^2 \mu(x) + o\left(h_n^2\right).$$

From this, Theorem 2.1 and Lemma 3.3 below, we know that

$$\begin{aligned} & \sqrt{n h_n^d} \left(\mathrm{vec} \left(\hat{m}(x) - m(x), h_n \left(\hat{m}'(x) - m'(x) \right) \right) + h_n^2 H_1^{-1} \mu(x) + o\left(h_n^2\right) \right) \\ = \quad & -\frac{H^{-1}}{\sqrt{n h_n^d}} \sum_{i=1}^n \left(K\left(\frac{x - X_i}{h_n} \right) \eta_{n,i} - I\!E \left(K\left(\frac{x - X_i}{h_n} \right) \eta_{n,i} \right) \right). \end{aligned}$$

Let $\zeta_i = K\left(\frac{x - X_i}{h_n}\right) \eta_{n,i} / \sqrt{h_n^d}$. Similar to the proof of Lemma 1 of Cai (2003), under Conditions A7, A8 and A10 we have

$$\sum_{j=2}^n \left\| \mathrm{cov}\left(\zeta_1, \zeta_j\right) \right\| \to 0.$$

Thus, it can be shown that

$$\mathrm{var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\zeta_i - I\!E \zeta_i \right) \right) \to H_2.$$

From Condition A3, we know that there exists a sequence of positive integers s_n such that $s_n \to \infty$, $s_n = o(n^\gamma)$ and $n^{1-\gamma} s_n^{-\theta} \to 0$. Going along the same lines as the proof of Theorem 2 of Cai (2003), we can obtain (2.3). $\square$

For any nonzero vector $t \in I\!R^{(p+1)d}$, the asymptotic mean squared error of $t^T \mathrm{vec}\,(\hat{m}(x), h_n \hat{m}'(x))$ is equal to

$$h_n^4 t^T H_1^{-1} \mu(x) \mu^T(x) H_1^{-1} t + \frac{1}{n h_n^d} t^T H_1^{-1} H_2 H_1^{-1} t,$$

so that the corresponding optimal bandwidth h_{opt} can be chosen as

$$h_{opt} = \left(\frac{t^T H_1^{-1} H_2 H_1^{-1} t}{n t^T H_1^{-1} \mu(x) \mu^T(x) H_1^{-1} t} \right)^{\frac{1}{d+4}}.$$

The following Lemma A is a direct extension of Theorem 2.1 of Liebscher (1996) to the case of multi-dimensional space.

Lemma A. *Assume that the sequence of random variables $\{Z_i, i \geq 1\}$ is α-mixing with α-mixing coefficients α_n. $I\!EZ_i = 0$ and $\|Z_i\| \leq S(n) < +\infty$, a.s. Let $T_n = \sum_{i=1}^n Z_i$. Then, for all n, $N \in I\!N$, $1 \leq N \leq n$ and all $\varepsilon > 4NS(n)$,*

$$I\!P\left\{\|T_n\| > \varepsilon\right\} \leq 4d \exp\left\{ -\frac{\varepsilon^2}{64 d^2 \frac{n}{N} \max \mathrm{diag}(D(n,N)) + \frac{8}{3} \varepsilon d N S(n)} \right\} + 4d \frac{n}{N} \alpha_N$$

holds with $D(n,m) = \sup_{0 \leq j \leq n-1} I\!E(\sum_{i=j+1}^{(j+m)\wedge n} Z_i)^2$ $(m \leq n)$ where $a \wedge b$ means $\min\{a,b\}$ and $diag(A)$ denotes the diagonal matrix with the same elements in the diagonal as the matrix A.

In the $(p+1)\,d$-dimensional space, partition the sphere $\|t\| \leq M\sqrt{\log n}$ into a sequence of subrectangles such that the length of each side for each subrectangle is less than or equal to $C\varepsilon_n$, where and thereafter $C > 0$ denotes a constant which can take different values in different circumstances. And the set B_n is comprised by all the corresponding grid points of this division. It is easy to see that the number of the grid points is equal to $O\left(\left(\frac{\sqrt{\log n}}{\varepsilon_n}\right)^{(p+1)d}\right)$.

Lemma 2.3. *Under Conditions A1, A2, A6 and $\varepsilon_n \to 0$, it holds almost surely that*

$$\sup_{t \in B_n} \left\| \sum_{i=1}^n (Z_{n,i}(t) - I\!EZ_{n,i}(t)) \right\| = O\left(\varepsilon_n \sqrt{n h_n^d}\right). \tag{2.4}$$

Proof. To get the desired result, by the Borel-Cantelli lemma, it suffices to prove that there exists some constant $M_1 > 0$ such that

$$I = I\!P\left\{ \cup_{t \in B_n} \left\{ \left\| \sum_{i=1}^n (Z_{n,i}(t) - I\!EZ_{n,i}(t)) \right\| > M_1 \varepsilon_n \sqrt{n h_n^d} \right\} \right\} \leq \frac{1}{n (\log n)^2}. \tag{2.5}$$

We now apply Lemma A to prove this point. According to the properties of Kronecker products, it can be shown that $\|Z_{n,i}(t) - I\!EZ_{n,i}(t)\| = O\left(\sqrt{\frac{\log n}{n h_n^d}}\right)$ holds uniformly for $\|t\| \leq M\sqrt{\log n}$.

For $1 \leq N \leq n$, from Lemma 2.1 of Liebscher (1996), it holds that

$$\max \operatorname{diag}(D(n,N)) \leq \max \operatorname{diag}\left(I\!E\left[\sum_{i=1}^{N} Z_{n,i}(t) Z_{n,i}^{T}(t)\right]\right) \leq CNR_q^2(n)\log n, \qquad (2.6)$$

where

$$\begin{aligned} R_q^q(n) &= h_n^d \int\int \|K(z)\begin{pmatrix} 1 \\ z \end{pmatrix} \otimes \Big[\psi\Big(m(x) + h_n m'(x) z + \frac{a + h_n bz}{\sqrt{nh_n^d}}, y\Big) \\ &\quad - \psi(m(x) + h_n m'(x) z, y)|\Big]\|^q f(x + h_n z, y) dy dz. \end{aligned}$$

In view of A2 and A6, we know that $R_q^q(n) \leq C\left|\sqrt{\frac{\log n}{nh_n^d}}\right|^q h_n^d$. According to this and (2.6), it can be seen that

$$\max \operatorname{diag}(D(n,N)) \leq CNh_n^{2d/q}\left|\sqrt{\frac{\log n}{nh_n^d}}\right|^2 \log n.$$

Applying Lemma A, we get

$$I \quad \leq \quad \left\{\exp\left\{ - \frac{M^2\Big(\varepsilon_n\sqrt{nh_n^d}\Big)^2}{64d^2\frac{n}{N}\max \operatorname{diag}(D(n,N)) + \frac{8}{3}M\varepsilon_n\sqrt{nh_n^d}dNS(n)}\right\} + \frac{n}{N}\alpha_N\right\} \times 4d \times \#\{B_n\}.$$

Thus, we know that if N and ε_n satisfies

$$\frac{n}{N}CN\log n \times h_n^{2d/q}\left|\frac{b_n}{\sqrt{nh_n^d}}\right|^2 \leq \varepsilon_n\sqrt{nh_n^d}N\frac{C\sqrt{2}}{\sqrt{nh_n^d}}b_n,$$

$$\left(\frac{\sqrt{\log n}}{\varepsilon_n}\right)^{(p+1)d}\frac{n}{N}\alpha_N \leq \frac{1}{n(\log n)^2}$$

and

$$\left(\frac{\sqrt{\log n}}{\varepsilon_n}\right)^{(p+1)d}\exp\left\{-\frac{M\varepsilon_n\sqrt{nh_n^d}}{dNS(n)}\right\} \leq \frac{1}{n(\log n)^2},$$

then (2.5) will hold. Through solving these three inequalities, we can get the desired result. □

Lemma 2.4. *Under Conditions A1, A2, A3, A4 and A6, it holds that*

$$\sup_{\|t\|\leq M\sqrt{\log n}}\left\|\sqrt{nh_n^d}\sum_{i=1}^{n}\xi_{n,i}(t) - Ht\right\| = O\left(\varepsilon_n + \frac{\log n}{\sqrt{nh_n^d}}\right). \qquad (2.7)$$

Proof. From Condition A4 and by Taylor expansion, it can be shown that

$$\begin{aligned} nI\!EZ_{n,i}(t) &= nh_n^d \int \begin{pmatrix} 1 \\ z \end{pmatrix} \otimes \left[G\left(h_n m'(x) z + \frac{a + bz}{\sqrt{nh_n^d}}, x + h_n z\right)\right. \\ &\quad \left. - G\left(h_n m'(x) z, x + h_n z\right)\right] f_X(x + h_n z) K(z) dz \\ &= \sqrt{nh_n^d}Ht + O(\log n) \end{aligned}$$

holds uniformly for $\|t\| \le M\sqrt{\log n}$. Combining this and Lemma 2.3, we obtain

$$\sup_{t\in B_n}\left\|\frac{1}{\sqrt{nh_n^d}}\sum_{i=1}^{n} Z_{n,i}(t) - Ht\right\| = O\left(\varepsilon_n + \frac{\log n}{\sqrt{nh_n^d}}\right).$$

From this, Lemma 3.3 and the positive definition on matrix H, it can be inferred that

$$\begin{aligned}
&\sup_{t\in B_n}\left\|\sqrt{nh_n^d}\sum_{i=1}^{n}\xi_{n,i}(t) - Ht\right\| \\
\le\ & \left|\frac{nh_n^d}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} - 1\right| \sup_{t\in B_n}\left(\left\|\frac{1}{\sqrt{nh_n^d}}\sum_{i=1}^{n} Z_{n,i}(t) - Ht\right\| + \|Ht\|\right) \\
&+ \sup_{t\in B_n}\left\|\frac{1}{\sqrt{nh_n^d}}\sum_{i=1}^{n} Z_{n,i}(t) - Ht\right\| = O\left(\varepsilon_n + \frac{\log n}{\sqrt{nh_n^d}}\right).
\end{aligned} \tag{2.8}$$

It can be concluded that the function $\xi_{n,i}(t)$ relating to the covariable t satisfies the Lipschitz condition of order one from the condition that $\psi(\cdot)$ is a Lipschitz function. Because H is a positive definite matrix, $t^T Ht$ is a convex function and its partial derivative is equal to $2Ht$. In view of this, the partition on the sphere $\|t\| \le M\sqrt{\log n}$, (2.8) and Lemma 6 of Niemiro (1992), we know that (2.7) holds. $\square$

3 Some Numerical Illustrations

In this section, we present two examples for the case $p = 2$. The objective is to show that the nonparametric M-estimator is more accurate in estimating $m(x)$ than the commonly used squared-loss nonparametric regression (Nadaraya-Watson) estimator. Similar evidence for the univariate case ($p = 1$) can be found in Jiang and Mack (2001).

Crucial to the estimation procedures is the choice of the bandwidth h_n. Although it can be inferred from Corollary 2.2 that the convergence rate of the optimal bandwidth is of order $O(n^{-1/(d+4)})$, it also depends on some other unknown quantities, the estimation of which is beyond Theorem 2.1. As an alternative, we will adopt the cross-validation method. For a given weight function $w_{n,i}$ the method is similar in spirit to the cross-validation method suggested by Yao & Tong (1998). In the following two examples, the loss function $\rho(t)$ for the nonparametric M-estimator is taken as the well-known Huber function, which is equal to $\| t \|^2 /2$ for $\| t \| \le k$ and equal to $k(\| t \| - \frac{k}{2})$ for $\| t \| > k$, where $k > 0$ is a constant. Let $H = [a_1(\frac{n}{2})^{-1/(d+4)}, a_2(\frac{n}{2})^{-1/(d+4)}]$, where $a_1 < a_2$ are two positive constants. We will use the first set of $n/2$ observations for estimation and the second set of $n/2$ observations for validation. To be more specific, for each $h \in H$, the estimator given by (1.2) is applied to the first

$n/2$ observations and the resulting estimates are denoted by $(\hat{m}_{\frac{n}{2},h}, \hat{m}'_{\frac{n}{2},h})$. Next, the optimal bandwidth $h_{n/2}$ for the second set is chosen as

$$\arg\min_{h \in H} \sum_{i=\frac{n}{2}+1}^{n} w_{n,i} \rho(Y_i - \hat{m}_{\frac{n}{2},h} - \hat{m}'_{\frac{n}{2},h}(X_i - x)).$$

Finally, from Yao and Tong (1998) it can be observed that the optimal bandwidth, say h_{opt}, for the whole sample of size n is given by $h_{opt} = h_{n/2}/2^{1/(d+4)}$.

For the M-estimator, another crucial issue concerns the choice of the parameter k in the Huber function. According to Corollary 2.2, and following a similar approach as in Jiang and Mack (2001), one way to choose an optimal value of k is to minimize the positive matrix $H_1^{-1} H_2 H_1^{-1}$ in the sense of the determinant, i.e.,

$$k_{opt} = \arg\min_k \det \left(G_1^{-1}(0, x)\, \mathbb{E}\left[\psi_k \left(Y - m(X) \right) \psi_k^T \left(Y - m(X) \right) \middle| X = x \right] G_1^{-1}(0, x) \right), \quad (3.1)$$

where ψ_k is the first derivative of the Huber function. The quantity on the right of (3.1) can be estimated through plugging the related consistent estimators into it.

Example 3.1. We consider the following bivariate vector autoregressive (VAR) linear time series model of order 1:

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} 0.8 & -0.3 \\ 1.2 & 0.4 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} 0.2\varepsilon_{1,t} \\ 0.3\varepsilon_{2,t} \end{pmatrix},$$

where the errors $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are i.i.d. as $0.5N(0,1) + 0.5C$. Here, $N(0,1)$ is the standard normal distribution and C the Cauchy distribution. The sample size used is $n = 401$. The heavy-tailedness of the simulated data is obvious and does not need further support from basic statistics such as kurtosis, mean, median, 1st– and 3rd quartile. Denote the corresponding bivariate vector regression functions by $m_1(x_1, x_2)$ and $m_2(x_1, x_2)$, respectively, where x_1 is related to $X_{1,t-1}$ and x_2 is related to $X_{2,t-1}$. Figures 3.1.a)–3.1.d) show four estimated regression functions. In each graph, one of the covariates is fixed and the other one changes over the interval $[-2, 2]$. 33 equidistant grid points in $[-2, 2]$ are used to estimate the corresponding regression functions. For the squared-loss nonparametric regression estimator, we set $a_1 = 0.4$ and $a_2 = 20$. As for the nonparametric M-estimator, we set $a_1 = 2$ and $a_2 = 6$. Table 1, columns 2–5, contain the corresponding MADEs for the two estimators. We see from this Table and Figures 3.1. that the nonparametric M-estimator gives a better fit to the series than the squared-loss nonparametric regression (Nadaraya-Watson) estimator.

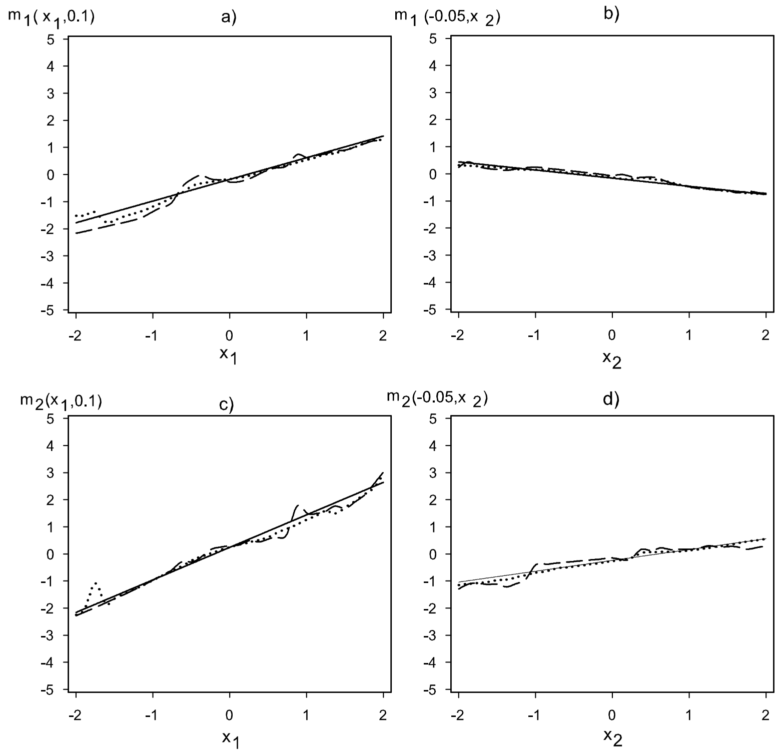

Figure 3.1. Simulation results for Example 3.1. Solid lines: true regression function, medium-dashed lines: estimated regression function corresponding to the squared-loss nonparametric regression (Nadaraya-Watson) estimator, dotted lines: estimated regression function corresponding to the M-estimator.

Table 1: Comparison of MADEs for the squared-loss nonparametric regression (Nadaraya-Watson) estimator and the nonparametric (robust) M-estimator.

	Figure 3.1				Figure 3.2			
Estimators	a	b	c	d	a	b	c	d
Squared-loss	0.2308	0.0866	0.1616	0.1679	0.3174	0.1504	0.1452	0.3220
Robust	0.1262	0.0466	0.1527	0.0581	0.2140	0.1061	0.0670	0.2696

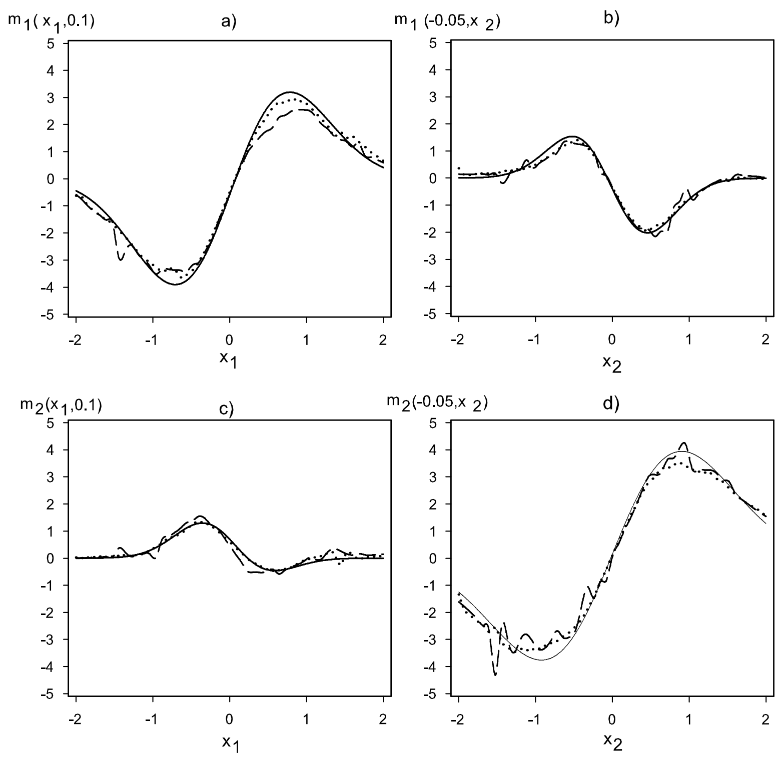

Figure 3.2. Simulation results for Example 3.2. Solid lines: true regression function, medium-dashed lines: estimated regression function corresponding to the squared-loss nonparametric regression (Nadaraya-Watson) estimator, dotted lines: estimated regression function corresponding to the M-estimator.

Example 3.2. To study the performance of the M-estimator in estimating nonlinear models with heavy-tailed errors, we consider the following stationary bivariate exponential AR model of order 1:

$$\begin{cases} X_{1,t} = (8X_{1,t-1} - 6X_{2,t-1}) \exp\left\{-3\left(0.3X_{1,t-1}^2 + 0.7X_{2,t-1}^2\right)\right\} + 0.18\varepsilon_{1,t}, \\ X_{2,t} = (7X_{1,t-1} - 3X_{2,t-1}) \exp\left\{-3\left(0.8X_{1,t-1}^2 + 0.2X_{2,t-1}^2\right)\right\} + 0.18\varepsilon_{2,t}, \end{cases}$$

where the errors $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are i.i.d. as $0.75N(0,1) + 0.25W$ and W is the symmetric Weibull distribution tailed as $P(|W| \geq x) = e^{-x^{1/3.3}}$ for any $x \geq 0$. The model was used by Harvill and Ray (2000) except that they generated the white noise disturbances from a bivariate (0,1) Gaussian distribution with cross-correlation 0. The sample size n was set at 401. Figures 3.2.a)-3.2.d) show the estimated regression functions. For each graph 50 equidistant grid points were chosen in the interval $[-2, 2]$. When dealing with the usual nonparametric squared-loss regression estimator, we set $a_1 = 0.2$ and $a_2 = 7$. For the nonparametric M-estimator, we set $a_1 = 1.4$ and $a_2 = 4$. Table 1, columns 6–10, contains the MADEs of the various estimated regression functions. It is evident that the nonparametric M-estimator provides a better approximation to the underlying structure than the commonly used regression estimator.

Remark: Note that both in theory and in practice the M-estimator is to be preferred over the squared-loss regression estimator when the observations are contaminated with heavy tailed errors. But, in practice, the squared-loss regression estimator still has its merits. The reason is that the squared-loss regression estimator can be expressed clearly. As a consequence it can be calculated much more quickly and accurately than that the M-estimator, which requires searching the minimum value of (1.2).

4 Real Data Example

As an illustration we analyze the transactions for the S&P500 stock index in May 1993 and its June futures contract traded at the Chicago Mercantile Exchange. The time interval is 1-minute (intraday). Several authors used this data to study index futures arbitrage. Tsay (1998) fitted a bivariate (nonlinear) threshold model to the first differences of the log(futures) and log(spot prices) using 7,060 observations. Here we shall focus on a randomly selected subset of 501 time series observations. Let $X_{1,t}$ and $X_{2,t}$ be the return series of the log futures and spot prices. Figure 4.1 shows the time plot of $X_{1,t}$ and $X_{2,t}$. The series fluctuate around a fixed mean and within a fixed range with some big outliers. Computing some basic statistics

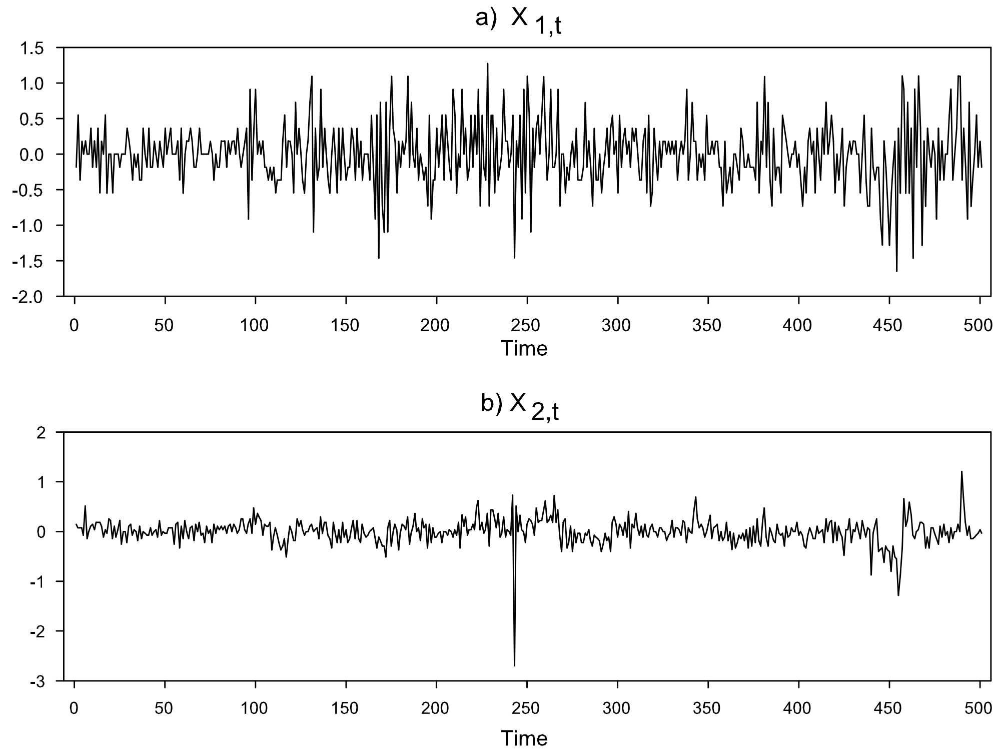

Figure 4.1. Time plots of minute-returns of S&P 500 index futures and prices.

revealed heavy-tailedness in $X_{2,t}$, with kurtosis 25.39. On the other hand, the kurtosis (0.93) of series $X_{1,t}$ indicates a distribution flatter than normal. Initial exploratory analysis of the sample auto– and cross-correlation function indicated that the series are highly correlated. A standard nonlinearity test suggested significant departures from linearity for both series, with a p-value of 0.0 for each series.

For simplicity of presentation, we assume that the bivariate time series come from the following bivariate AR model of order 1:

$$\begin{cases} X_{1,t} = m_1\left(X_{1,t-1}, X_{2,t-1}\right) + \varepsilon_{1,t}, \\ X_{2,t} = m_2\left(X_{1,t-1}, X_{2,t-1}\right) + \varepsilon_{2,t}. \end{cases}$$

Figure 4.2 shows plots of the four estimated regression functions. In each plot one of the covariates is set at a prefixed value and the values of the other covariate are left free. From graph 4.2.a) we see that there is a nonlinear relationship between $X_{1,t}$ and $X_{1,t-1}$, for values of $X_{1,t-1}$ in the interval $[-0.5, 0.5]$, when $X_{2,t-1}$ is fixed at 0.1. In graph 4.2.c), the regression function $m_2(\cdot)$ is almost equal to a constant. This means that the time series $X_{1,t-1}$ has a minimal effect on $X_{2,t}$, when $X_{2,t-1}$ is fixed at 0.1. Also, both estimated regression functions displayed in graphs 4.2.b) suggest a not-so-strong positive linear relationship between $X_{1,t}$ and

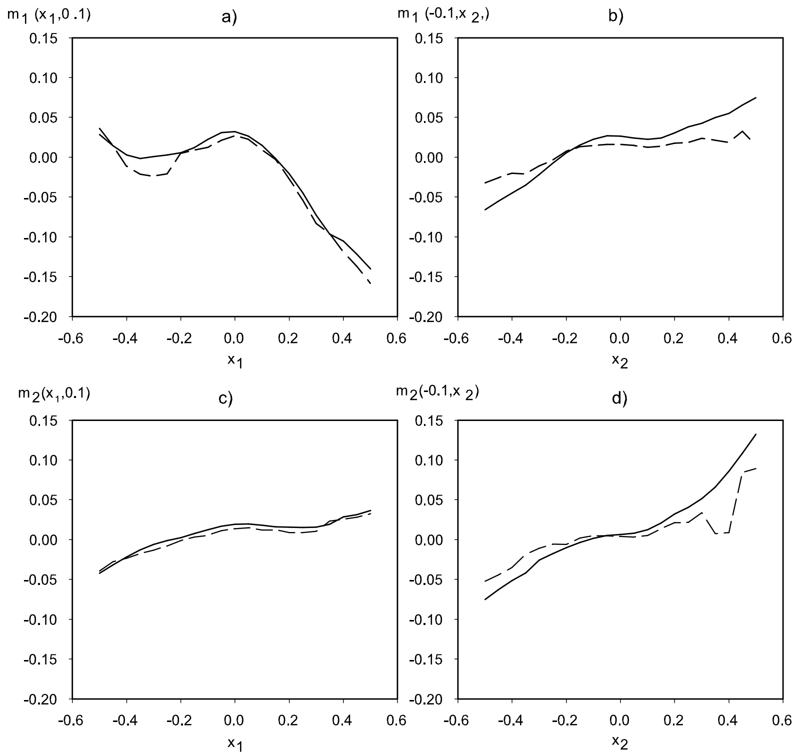

Figure 4.2. Estimated regression functions. Solid lines: estimated regression function corresponding to the squared-loss nonparametric regression (Nadaraya-Watson) estimator, dashed-dotted lines: estimated regression function corresponding to the M-estimator.

$X_{2,t-1}$, when $X_{1,t-1}$ is fixed at -0.1. Similarly the estimated regression functions in graph 4.2.d) indicate a weak positive linear relationship between $X_{2,t}$ and $X_{2,t-1}$, when $X_{1,t-1} = -0.1$. These results seem to be in line with Tsay's study, who entertained a bivariate threshold (nonlinear) model of order 8 for the full data set with separate higher-order linear AR processes in each regime. Note that the choice of the 501 observations is rather arbitrary. Nevertheless, results obtained for other subsets of data seem to support the above observations.

Finally, we would like to stress that the empirical application is intended as an example of the usefulness of the M-estimator for multivariate time series process with heavy-tailed errors and outliers. The application should not be considered as an in-depth data analysis. For instance, the functional form of the nonlinearity between the sequence of random variables $(X_{1,t}, X_{2,t}), (X_{1,t-1}, X_{2,t-1}), (X_{1,t-2}, X_{2,t-2}), \ldots$ is still open for further research. Also the application should not be linked to studies/models on futures arbitrage in finance.

Appendix

Let's introduce some necessary lemmas, which are used to get our main result. The following Lemma 3.1 is an extension of Theorem 4 of Rio (1995).

Lemma 3.1. *Let $\mathcal{A}$ be a σ-field of $(\Omega, \mathcal{F}, \mathbb{P})$ and let X be a real-valued random variable taking a.s. its values in $[a, b]$. Suppose furthermore that there exists a random variable ξ with uniform distribution over $[0, 1]$, independent of $\mathcal{A} \cup \sigma(X)$. Then, for any $\delta \geq 1$, there exists some random variable X^* independent of $\mathcal{A}$ and with the same distribution as X such that*

$$\mathbb{E}\left(|X - X^*|^{\delta}\right) \leq 2^{\frac{1}{\delta}} (b - a) \left[\alpha\left(\mathcal{A}, \sigma\left(X\right)\ \right)\right]^{\frac{1}{\delta}}. \tag{B.1}$$

Moreover, X^* is a $\mathcal{A} \cup \sigma(X\ ) \cup \sigma(\xi)$-measurable random variable.

Proof. Here, the same notations and methods as in the proof Theorem 4 of Rio (1995) are introduced to address the proof of this lemma. Analogous to the equality (2.4) in that paper, for any $\delta \geq 1$, it holds that

$$\int_0^1 \left|F_{\mathcal{A}}^{-1}(s) - F^{-1}(s)\right|^{\delta} ds = \int_a^b \left|F_{\mathcal{A}}(t) - F(t)\right|^{\frac{1}{\delta}} dt.$$

Also, according to Hölder's inequality and the definition of α-mixing, we can obtain that

$$\mathbb{E}\left(\left|F_{\mathcal{A}}(t) - F(t)\right|^{\frac{1}{\delta}}\right) \leq \left(\mathbb{E}\left|F_{\mathcal{A}}(t) - F(t)\right|\right)^{\frac{1}{\delta}} \leq \left(2\alpha\left(\mathcal{A}, \sigma\left(X\ \right)\right)\right)^{\frac{1}{\delta}}.$$

Going along the same lines as in the proof of Theorem 4 of Rio (1995), it can be shown that (B.1) holds. □

Lemma 3.2. *Under Conditions A1, A2, A3, A8, A9 and A10, it holds for large enough n with probability one that*

$$\sum_{i=1}^{n} w_{n,i}\eta_{n,i} = O\left(\sqrt{\frac{\log n}{nh_n^d}}\right). \tag{B.2}$$

Proof. In view of Lemma 3.3 below, it suffices to prove that

$$\frac{1}{\sqrt{nh_n^d}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\eta_{n,i} = O\left(\sqrt{\log n}\right).$$

Without loss of generality, we only prove that this relationship holds when $\eta_{n,i}$ is replaced by its first component η_i. Let $\psi_{1,i}$ be the first component of the random vector $\psi(m(x) + m'(x)(X_i - x), Y_i)$. For simplification, denote by $\varsigma_n = \sqrt{nh_n^d \log n}$ and

$$S_n = \sum_{i=1}^{n} \left(K\left(\frac{x - X_i}{h_n}\right)\eta_i - \mathbb{E}K\left(\frac{x - X_i}{h_n}\right)\eta_i\right),$$

respectively. From $\mathbb{E}K\left(\frac{x-X_i}{h_n}\right)\eta_{n,i}\sim Ch_n^d$ and Borel Cantelli's lemma, to get the desired result, it suffices to prove that there exists some constant $M>0$ such that $I_1=\sum_{n=1}^{\infty}\mathbb{P}(S_n>M\varsigma_n)<\infty$. In this proof, the constant C is independent of M. For $T_n>0$, we introduce the following notations $\bar{\psi}_{1,i}=\psi_{1,i}\mathbb{I}_{(|\psi_{1,i}|<T_n)}$,

$$V_{1,i}=\bar{\psi}_{1,i}K\left(\frac{x-X_i}{h_n}\right)-\mathbb{E}\bar{\psi}_{1,i}K\left(\frac{x-X_i}{h_n}\right)$$

and

$$V_{2,i}=\left(\psi_{1,i}-\bar{\psi}_{1,i}\right)K\left(\frac{x-X_i}{h_n}\right)-\mathbb{E}\left(\psi_{1,i}-\bar{\psi}_{1,i}\right)K\left(\frac{x-X_i}{h_n}\right),$$

respectively. Then, the equation $S_n=\sum_{i=1}^{n}\left(V_{1,i}+V_{2,i}\right)$ gives that

$$I_1\le\sum_{n=1}^{\infty}\mathbb{P}\left(\sum_{i=1}^{n}V_{1,i}>\frac{M\varsigma_n}{2}\right)+\sum_{n=1}^{\infty}\mathbb{P}\left(\sum_{i=1}^{n}V_{2,i}>\frac{M\varsigma_n}{2}\right)=I_{11}+I_{12}.$$

We first deal with I_{11}. Without loss of generality, assume that $n=2uv$, where $u=u_n$ and $v=v_n$ are positive integer numbers. Otherwise, there exists a remainder $r=r_n$ in the expression $n=2uv+r$, where $0\le r<2u$. In this case, it can be dealt with analogously according to the proof below. For $j=1,2,\ldots,2v$, denote by $W_n(j)=\sum_{i=(j-1)u+1}^{ju}V_{1,i}$ and $S_{n_i}=\sum_{j=1}^{v}W_n\left(i+2\left(j-1\right)\right)$ for $i=1,2$. Then, $\sum_{i=1}^{n}V_{1,i}=S_{n_1}+S_{n_2}$, which results in

$$I_{11}\le\sum_{i=1}^{2}\mathbb{P}\left(|S_{n_i}|>\frac{M\varsigma_n}{4}\right)=I_{13}+I_{14}.$$

We only dispose of I_{13} because I_{14} can be dealt with similarly. For simplification, let $W_j=W_n\left(1+2\left(j-1\right)\right),\ j=1,2,\ldots,v$. Analogous to Lemma 2.5 of Merlevède and Peligrad (2000), using Lemma 3.1, we present the following four assertions.

i). There exists a sequence of i.i.d. random variables $\{W_j^*,j=1,2,\ldots,v\}$ with the common distribution as W_1. Also, W_j^* is independent of $\sigma\left(W_1,\ldots,W_{j-1}\right)$.

ii). For any $\delta>0$, $\mathbb{E}\left|W_j-W_j^*\right|^{2+\delta}\le 2uT_n\left[2\alpha(u)\right]^{\frac{1}{2+\delta}}$.

Then, in view of the definition on α-mixing dependence, the following two conclusions can be inferred from the listed property i).

iii). The sequence $\left\{W_j-W_j^*,\ j=1,\ldots,v\right\}$ is still α-mixing with the α-mixing coefficients $\alpha\left(\left(j-1\right)u\right)$.

iv). For any $1\le i,j\le v$, it holds that

$$\mathbb{E}\left(W_i-W_i^*\right)\left(W_j-W_j^*\right)=\mathbb{E}\left(W_1-W_1^*\right)\left(W_{|j-i+1|}-W_{|j-i+1|}^*\right).$$

In the sequel, for simplification, u is chosen as

$$u^{r-1} = \frac{CM^{r-2}}{n} \frac{\left(\sqrt{nh_n^d}\right)^{r-2}}{(\log n)^{\frac{r+2}{2}}} \tag{B.3}$$

without restricting it to be an integral. Let

$$T_n = \frac{CM}{u} \sqrt{\frac{nh_n^d}{\log n}}. \tag{B.4}$$

From the known Condition A9, we can choose some suitable small constant $\delta > 0$ such that

$$\frac{2}{\delta} > \frac{(r-1)\,\theta}{r-(r-2)\,\gamma} - 1 \tag{B.5}$$

and

$$\gamma > \frac{\frac{2+\delta}{2} + \frac{\theta}{2+\delta} + (2+\delta)(r-1)}{(r-2)\left(\frac{2+\delta}{2} + \frac{\theta}{2+\delta}\right) + (1+\delta)(r-1)} \tag{B.6}$$

hold simultaneously. Substituting (B.3) into the inequality below gives

$$\begin{aligned} & \sum_{j=1}^{v-1} \left(\alpha\left((j-1)\,u\right)\right)^{\frac{\delta}{2+\delta}} \\ \leq\ & Cu^{\frac{-\theta\delta}{2+\delta}} v^{\frac{-\theta\delta}{2+\delta}+1} = \frac{Cn^{\left(1-\frac{\theta\delta}{2+\delta}\right)+\frac{1}{r-1}} (\log n)^{\frac{r+2}{2(r-1)}}}{M^{\frac{r-2}{r-1}} \left(\sqrt{nh_n^d}\right)^{\frac{r-2}{r-1}}} < \infty, \end{aligned} \tag{B.7}$$

where the last step is based on the inequality (B.5). Using $\alpha(i) = O\left(i^{-\theta}\right)$ and Davydov's (1968) inequality, we can get that

$$\begin{aligned} \mathrm{var}(W_1^*) &\leq I\!E\left(W_1^2\right) = u\mathrm{var}(V_{n,1}) + \sum_{j=1}^{u-1}(u-j)\mathrm{cov}(V_{n,1}, V_{n,1+j}) \\ &\leq Cuh_n^d \max\left\{1, u^{1-\theta\left(1-\frac{2}{r}\right)} h_n^{\frac{2d}{r}-d}\right\}. \end{aligned} \tag{B.8}$$

From Condition A9, we know that

$$\gamma > \frac{r^2 - 4r + 2 + \theta(r-2)}{(r-2)^2(\theta+1)},$$

so that the following relationship

$$\limsup \max\left\{1, u^{1-\theta\left(1-\frac{2}{r}\right)} h_n^{-\left(1-\frac{2}{r}\right)d}\right\} < \infty \tag{B.9}$$

holds because of (B.3). Thus, the following two facts

$$Cnh_n^d \max\left\{1, u^{1-\theta\left(1-\frac{2}{r}\right)} h_n^{\frac{2d}{r}-d}\right\} \leq uT_nM\varsigma_n$$

and $\frac{M\varsigma_n}{T_n u} \geq (1+\varepsilon)\log n$ hold, where $\varepsilon > 0$. Thus, according to Bernstein's inequality, we know that

$$I\!\!P\left(\left|\sum_{j=1}^{v} W_j^*\right| \geq \frac{M\varsigma_n}{8}\right) \leq 2\exp\left\{\frac{-\frac{1}{2}\left(\frac{M\varsigma_n}{8}\right)^2}{v\cdot \mathrm{var}(W_j^*) + 2uT_n\frac{M\varsigma_n}{8}/3}\right\}$$
$$\leq 2\exp\left\{-\frac{CM\varsigma_n}{T_n u}\right\} \leq 2n^{-(1+2\varepsilon)}.$$

Let

$$I_{15} = \frac{M^{2+\delta} h_n^{d\frac{2+\delta}{2}} u^{\frac{\delta}{2}}}{C\left[\alpha(u)\right]^{\frac{1}{2+\delta}}(n\log n)^{\frac{2+\delta}{2}}}.$$

From (B.3), we know that

$$\frac{I_{15}}{T_n} = \frac{CM^{(r-2)\left(\frac{2+\delta}{2}+\frac{\theta}{2+\delta}\right)+(1+\delta)(r-1)}\left(\sqrt{nh_n^d}\right)^{(r-2)\left(\frac{2+\delta}{2}+\frac{\theta}{2+\delta}\right)+(1+\delta)(r-1)}}{(\log n)^{\frac{r+2}{2}\left(\frac{2+\delta}{2}+\frac{\theta}{2+\delta}\right)+\frac{1+\delta}{2}(r-1)}\, n^{\frac{2+\delta}{2}+\frac{\theta}{2+\delta}+(2+\delta)(r-1)}}.$$

In view of (B.6), we know further that $T_n \leq I_{15}$. From the listed properties on W_j, Markov's inequality, Hölder's inequality, Davydov's (1968) inequality, it holds successively that

$$\begin{aligned}
& I\!\!P\left(\left|\sum_{j=1}^{v}\left(W_j - W_j^*\right)\right| \geq \frac{M\varsigma_n}{8}\right) \\
\leq\; & \frac{Cv}{M^2\varsigma_n^2}\left(I\!\!E\left|(W_1 - W_1^*)\right|^2 + \sum_{j=1}^{v-1} I\!\!E\left|(W_1 - W_1^*)\left(W_{j+1} - W_{j+1}^*\right)\right|\right) \\
\leq\; & \frac{Cv}{M^2\varsigma_n^2}\left(I\!\!E|W_j - W_j^*|^{2+\delta}\right)^{\frac{2}{2+\delta}}\left(1 + \sum_{j=1}^{v-1}\left(\alpha\left((j-1)\,u\right)\right)^{\frac{\delta}{2+\delta}}\right) \qquad \text{(B.10)} \\
\leq\; & \frac{Cv}{M^2\varsigma_n^2}\left(I\!\!E|W_j - W_j^*|^{2+\delta}\right)^{\frac{2}{2+\delta}} \leq \frac{Cv(uT_n)^{\frac{2}{2+\delta}}\left[\alpha(u)\right]^{\frac{2}{(2+\delta)^2}}}{M^2\varsigma_n^2} \leq \frac{1}{n(\log n)^2},
\end{aligned}$$

where in the third and the last inequalities above, (B.7) and $T_n \leq I_{15}$ are used, respectively. For I_{12}, by Markov's inequality, it holds for all sufficiently large n that

$$\begin{aligned}
I_{12} \;\leq\; & \frac{2nI\!\!E\left|\left(\psi_{1,i} - \bar{\psi}_{1,i}\right)K\left(\frac{x-X_j}{h_n}\right)\right|}{\frac{M}{2}\varsigma_n} \leq \frac{4}{M}\sqrt{\frac{nh_n^d}{\log n}}I\!\!E\left(\left|\left(\psi_{1,i} - \bar{\psi}_{1,i}\right)\right|\,\middle|\, X_j = x\right) \\
\leq\; & \frac{4}{MT_n^{r-1}}\sqrt{\frac{nh_n^d}{\log n}}I\!\!E\left(\left|\psi_1\right|^r \middle| X_j = x\right) \leq \frac{1}{n\left(\log n\right)^2},
\end{aligned}$$

where the last inequality is justified by (B.3), (B.4) and some large number $M > 0$. From the analysis above, we know that $I_1 < \infty$. This completes the proof. □

Lemma 3.3. *Under Conditions A1, A2 and A3, there exists some constant $M > 0$ such that*

$$\left|\sum_{i=1}^{n}\left(K\left(\frac{x-X_i}{h_n}\right) - I\!\!EK\left(\frac{x-X_i}{h_n}\right)\right)\right| \leq M\sqrt{nh_n^d\log n} \qquad \text{(B.11)}$$

holds almost surely for all sufficiently large n.

Proof. Here, a similar method to Lemma 3.2 is adopted. Assume that $n = 2uv$. Let $V_{n,i} = K\left(\frac{x-X_i}{h_n}\right) - \mathbb{E}K\left(\frac{x-X_i}{h_n}\right)$, $W_n(j) = \sum_{i=(j-1)u+1}^{ju} V_{n,i}$, $j = 1, 2, \ldots, 2v$, and $S_{n_i} = \sum_{j=1}^{v} W_n(i + 2(j-1))$, $i = 1, 2$. To prove (B.11), it suffices to get the bound from above on $\mathbb{P}\{S_{n_1} > \frac{M\varsigma_n}{2}\}$. Let $W_j = W_n(1 + 2(j-1))$, $j = 1, 2, \ldots, v$. And there exists a sequence of independent and identically distributed random variables W_j^*, $j = 1, 2, \ldots, v$, which have the same distribution as W_j, such that

$$\mathbb{E}\left|W_j - W_j^*\right|^{2+\delta} \leq Cu\left[\alpha(u)\right]^{\frac{1}{2+\delta}} \tag{B.12}$$

holds for some $\delta > 0$, which will be determined later. It holds trivially that $\mathbb{P}\left\{S_{n_1} > \frac{M\varsigma_n}{2}\right\} \leq I_{31} + I_{32}$, where $I_{31} = \mathbb{P}\left\{\sum_{j=1}^{v} W_j^* > \frac{M\varsigma_n}{4}\right\}$ and $I_{32} = \mathbb{P}\left\{\sum_{j=1}^{v}\left|W_j - W_j^*\right| > \frac{M\varsigma_n}{4}\right\}$. Under Conditions A2, A3 and Davydov's inequality, we can derive that

$$\left|\operatorname{cov}\left(V_{n,1}, V_{n,1+j}\right)\right| \leq C\left(\mathbb{E}V_{n,1}^{2+\delta_1}\right)^{\frac{2}{2+\delta_1}}\left[\alpha(j)\right]^{1-\frac{2}{2+\delta_1}} \sim C\left(h_n^d\right)^{\frac{2}{2+\delta_1}}\left[\alpha(j)\right]^{1-\frac{2}{2+\delta_1}}, \tag{B.13}$$

where the constant $\delta_1 > 0$. From this and analogous to (B.8), it can be shown that

$$\mathbb{E}\left(W_1^2\right) \leq C\left(uh_n^d + \left(h_n^d\right)^{\frac{2}{2+\delta_1}} u^{2-\theta\left(1-\frac{2}{2+\delta_1}\right)}\right). \tag{B.14}$$

Let $u = C\sqrt{\frac{nh_n^d}{\log n}}$. From $\theta > 3$, we can choose some suitable constant $\delta_1 > 0$ such that

$$\frac{1}{2} > \gamma > \frac{2}{1 + \theta - \frac{2}{\delta_1}},$$

so that

$$\limsup_{n\to\infty}\left(h_n^d\right)^{\frac{2}{2+\delta_1}-1} u^{1-\theta\left(1-\frac{2}{2+\delta_1}\right)} < \infty. \tag{B.15}$$

Combination of this, (B.14) and $n = 2uv$ gives that

$$2v \times \operatorname{var}(W_1^*) \leq 2u\frac{M\varsigma_n}{4}.$$

Therefore, from (B.14) and Bernstein's inequality (see, e.g., Serfling (1980), p.95), we know that

$$I_{31} \leq \exp\left\{-\frac{\left(\frac{M\varsigma_n}{4}\right)^2}{2v\cdot\operatorname{var}(W_1^*) + 2u\frac{M\varsigma_n}{4}}\right\} \leq \exp\left\{-\frac{CM\varsigma_n}{u}\right\} \leq \exp\left\{-CM\log n\right\}. \tag{B.16}$$

Similar to (B.10), for some $\delta > 0$, it holds that

$$\begin{aligned} I_{32} &\leq \frac{Cv}{M^2\varsigma_n^2}\left(\mathbb{E}\left|W_j - W_j^*\right|^{2+\delta}\right)^{\frac{2}{2+\delta}}\left(1 + \sum_{j=1}^{v-1}\left(\alpha\left((j-1)u\right)\right)^{\frac{\delta}{2+\delta}}\right) \\ &\leq \frac{Cu^{\frac{2}{2+\delta}}v}{M^2\varsigma_n^2}\left(1 + \sum_{j=1}^{v-1}\left(\alpha\left((j-1)u\right)\right)^{\frac{\delta}{2+\delta}}\right)\left[\alpha(u)\right]^{\frac{2}{(2+\delta)^2}}. \end{aligned}$$

We now begin to prove that the last expression above is less than or equal to $\frac{1}{n(\log n)^2}$. Because $n = 2uv$, $\varsigma_n = \sqrt{nh_n^d \log n}$, $u = C\sqrt{\frac{nh_n^d}{\log n}}$, $\alpha(u) = O(u^{-\theta})$ $\sum_{j=1}^{v-1} (j-1)^{\frac{-\delta}{2+\delta}} = O(v^{1-\frac{\delta}{2+\delta}})$, this is required equivalently to show that

$$\frac{M^2}{C} \geq n^2 u^{\frac{2}{2+\delta} - \frac{2\theta}{(2+\delta)^2} - 3} + n^{3-\frac{\theta\delta}{2+\delta}} u^{\frac{2}{2+\delta} - \frac{2\theta}{(2+\delta)^2} - 4} \tag{B.17}$$

In view of Condition A3, (B.17) holds for $\delta > 0$ sufficiently small. $\square$

References

Berlinet, A., Gannoun, A. and Matzner-Løber, E. (2001), "Asymptotic normality of convergent estimates of conditional quantiles", *Statistics*, Vol. 35, No. 2, 139–169.

Bosq, D. (1998), *"Nonparametric statistics for stochastic processes. Estimation and prediction"*, Second edition, Springer-Verlag, New York.

Cai, Z.W. and Ould-Said, E. (2003), "Local M-estimator for nonparametric time series", *Statistics & Probability Letters*, Vol. 65, 433-449.

Davydov, Y.A. (1968), "Convergence of distributions generalized by stationary stochastic processes", *Probab. Theory Appl. Statist.*, Vol. 13, 691-696.

Fan, J.Q., Hu, T.-C. and Truong, Y.K. (1994), "Robust nonparametric function estimation", *Scand. J. Statist.*, Vol. 21, 433-446.

Fan, J.Q., Heckman, N.E. and Wand, M.P. (1995), "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions", *J. Amer. Statist. Assoc.*, Vol. 90 , No. 429, 141-150.

Harvill, J.L. and Ray, B.K. (2000), "An investigation of lag identification tools for vector nonlinear time series", *Commun. Statist.: - Theory Meth.*, Vol. 29, No. 8, 1677-1701.

Honda, T. (2000), "Nonparametric estimation of a conditional quantile for α-mixing processes", *Ann. Inst. Statist. Math.*, Vol. 52, No. 3, 459-470.

Hong, S.Y. (2003), "Bahadur representation and its applications for local polynomial estimates in nonparametric M-regression", *J. of Nonparametric Statistics*, Vol. 15, No. 2, 237-251.

Jiang, J.C. & Mack Y.P. (2001), "Robust local polynomial regression for dependent data", *Statistica Sinica*, Vol. 11, 705-722.

Jureckova, J. & Sen, P.K. (1996), "*Robust Statistical Procedures: Asymptotics and Interrelations*", Wiley: New York.

Kozek, A.S. & Pawlak, M. (2002), "Distribution-free consistency of kernel non-parametric M-estimators", *Statistics & Probability Letters.*, No. 58, 343-353.

Liebscher, E. (1996), "Strong convergence of sums of α-mixing random variables with applications to density estimation", *Stochastic Process. Appl.*, No. 65, 69-80.

Liebscher, E. (2001), "Estimation of the density and the regression function under mixing conditions", *Statist. Decisions*, Vol. 16, No. 1, 9-26.

Masry, E. (1996a), "Multivariate local polynomial regression for time series: uniform strong consistency and rates", *J. Time Ser. Anal.*, Vol. 17, No. 6, 571–599.

Masry, E. (1996b), "Multivariate regression estimation—local polynomial fitting for time series", *Stochastic Process. Appl.*, Vol. 65, no. 1, 81–101.

Masry, E. & Fan, J. (1997), "Local polynomial estimation of regression functions for mixing processes", *Scand. J. Statist.*, Vol. 24, No. 2, 165–179.

Merlevède, F. & Peligrad, M. (2000), "The functional central limit theorem under the strong mixing condition.", *Ann. Probab.,* Vol. 28, No. 3, 1336–1352.

Niemiro, W. (1992), "Asymptotics for M-estimators defined by convex minimization", *Ann. Statist.*, Vol. 20, No. 3, 1514-1533.

Ortega, J.M. & Rheinboldt, W.C. (1970), "*Iterative solution of nonlinear equations in several variables*", Academic Press, New York-London.

Rio, M. (1995), "The functional law of the iterated logarithm for stationary strongly mixing sequences", *Ann. Probab.*, Vol. 23, No. 3, 1188-1203.

Serfling, R.J. (1980), *"Approximation Theorems of Mathematical Statistics"*, Wiley: New York.

Tsay, R.S. (1998), "Testing and modelling multivariate threshold models", *J. Amer. Statist. Ass.*, Vol. 93, No. 443, 1188-1202.

Yao, Q. & Tong, H. (1998), "Cross-validatory bandwidth selections for regression based on dependent data", *J. Statist. Planning Inf.*, Vol. 68, 387-415.