Statistica Sinica **11**(2001), 705-722

ROBUST LOCAL POLYNOMIAL REGRESSION FOR DEPENDENT DATA

Jiancheng Jiang and Y. P. Mack

Peking University and University of California

Abstract: Let $(X_j, Y_j)_{j=1}^n$ be a realization of a bivariate jointly strictly stationary process. We consider a robust estimator of the regression function $m(x) = E(Y|X = x)$ by using local polynomial regression techniques. The estimator is a local M-estimator weighted by a kernel function. Under mixing conditions satisfied by many time series models, together with other appropriate conditions, consistency and asymptotic normality results are established. One-step local M-estimators are introduced to reduce computational burden. In addition, we give a data-driven choice for minimizing the scale factor involving the ψ-function in the asymptotic covariance expression, by drawing a parallel with the class of Huber's ψ-functions. The method is illustrated via two examples.

Key words and phrases: Data-driven, local M-estimator, local polynomial regression, mixing condition, one-step, robustness.

1. Introduction

Consider a bivariate sequence of jointly strictly stationary random vectors $\{(X_j, Y_j), j = 1, \ldots, n\}$. Let $m(x) = E(Y|X = x)$ denote the regression function of Y on X, assumed to exist. When the sequence is i.i.d., the nonparametric estimation of $m(x)$ was first introduced independently by Nadaraya (1964) and Watson (1964): for a sample of size n,

$$\hat{m}(x) = \sum_{j=1}^{n} w_n(x, X_j) Y_j, \tag{1.1}$$

where

$$w_n(x, X_j) = K(\frac{X_j - x}{h_n}) / \sum_{i=1}^{n} K(\frac{X_i - x}{h_n}). \tag{1.2}$$

The *kernel function* $K(\cdot)$ is typically a symmetric density function, and the sequence of positive real numbers h_n, *bandwidths*, control the amount of smoothing. For the present discussion, we label (1.1) with weights (1.2) the *Nadaraya-Watson*, or *N-W* estimator. More generally, the *N-W* estimator can be thought

of as the solution of the following optimization problem:

$$\hat{m}(x) = \arg\min_{\alpha} \sum_{j=1}^{n} \zeta(Y_j - \alpha) K(\frac{X_j - x}{h_n}), \tag{1.3}$$

where, for the *N-W* estimator, $\zeta(v) = v^2$, suggesting the least squares criterion. Other choices of ζ are possible, for example in the context of robustness (see Hampel, Ronchetti, Rousseeaus and Stahl (1986) and Huber (1981)), leading to the local M-type regression estimators (see Cleveland (1979), Härdle (1990), Boente and Fraiman (1989), for example). For the purpose of this discussion, we label the solution of (1.3) under general ζ the *local constant kernel regression estimator* (*LOCKRE*).

When the data exhibit dependence a structural assumption, widely adopted in the study of nonparametric regression estimation, is the notion of *strong-mixing* (or *α-mixing*). Introduced by Rosenblatt (1956), strong mixing is a property shared by many time series models, including "generally" the autoregressive linear processes (see Athreya and Pantala (1986), Boente and Fraiman (1990).) Among a variety of mixing conditions, strong mixing is a mild restriction toward achieving asymptotic normality (see Bradley (1986), Doukhan (1994)). Under appropriate assumptions on the mixing rates, it was demonstrated that the performance of *LOCKREs* under mixing is essentially the same as under the i.i.d. assumption in terms of convergence rates, asymptotic normality and choice of data-dependent bandwidths. In particular, *LOCKREs* under strong mixing exhibit the same bias and boundary problems as the i.i.d. case (see Baek and We -hyly (1993) and Boente and Fraiman (1995)).

In the i.i.d. setting, Fan (1993) showed that local polynomial regression estimators have advantages over *LOCKREs* in terms of design adaptation and high asymptotic efficiency. The local polynomial regression estimators are constructed according to the following criterion: find a_k's so as to minimize

$$\sum_{j=1}^{n} [Y_j - \sum_{k=0}^{p} a_k (X_j - x)^k]^2 K(\frac{X_j - x}{h_n}). \tag{1.4}$$

Then $m(x)$ is estimated by the solution, $\hat{a}_0$, of a_0 from (1.4). Similarly, $m^{(j)}(x)$, the j-th derivative of m at x, is estimated by $j!\hat{a}_j$. We label the regression estimators so constructed as *local polynomial kernel regression estimators (LOPKREs)*. Obviously, the class of *LOPKREs* includes *LOCKREs*, but for the case $p = 0$ the advantages of reducing bias and boundary adjustment are lost. So for *LOPKREs* we will assume $p > 0$. For a detailed account of *LOPKREs* in the i.i.d. case, see Fan and Gijbels (1996). For mixing processes, Masry and Fan (1997) showed that *LOPKREs* share essentially the same asymptotic behavior as in i.i.d. cases.

As in *LOCKREs*, *LOPKREs* can be robustified by using a general ζ function instead of quadratic loss in (1.4) to form local M-type regression estimators. We will continue to call such estimators *LOPKREs*. Fan, Hu and Truong (1994), Fan and Jiang (2000), and Welsh (1996) pointed out that, for i.i.d. cases, such *LOPKREs* cope well with edge effects and are effective methods for derivative estimation. It is interesting to point out that the concept of robust local polynomial regression was previously introduced by Cleveland (1979), leading to the so-called LOWESS or Locally Weighted Smoothing Scatterplots. Theoretical endorsement of Cleveland's LOWESS was given in Fan and Jiang (2000).

For the present investigation, we present asymptotic results on *LOPKREs* under certain mixing conditions. This not only extends the results of Fan and Jiang (2000) to non-i.i.d. cases, but overcomes the lack of robustness for the estimators in Masry and Fan (1997). Pointwise asymptotic normality of our estimators enables one to find the asymptotically optimal variable bandwidth choice, and thereafter allows one to develop data-driven optimal variable bandwidth by using the idea of Fan and Gijbels (1995). To reduce computational burden, we study one-step local M-estimators which share the same asymptotic behavior as fully iterative M-estimators. A data-driven method for choosing the ψ-function (derivative function of ζ) is proposed. Our assumptions on the ψ-function are considerably weaker than in earlier works, and we do not require the symmetry of the conditional distribution of the error given X (see details in Section 2). The dependence structures assumed in this study are similar to the ρ-mixing and α-mixing conditions in Masry and Fan (1997).

The outline of this paper is as follows. In Section 2, we introduce the notation and assumptions used throughout the paper. Section 3 concentrates on the asymptotic properties of the proposed estimators, including pointwise consistency and joint asymptotic normality. In Section 4, one-step local M-estimators are studied and shown to have the same asymptotic behavior as their corresponding M-estimators. Section 5 includes the data-driven method for choosing the ψ-function within the class of Huber's ψ-functions to minimize the asymptotic variance, hence the asymptotic mean squared error, since asymptotic bias does not depend on the ψ-function. The idea is illustrated by examples in Section 6. Technical proofs are given in the Appendix.

2. Notations and Assumptions

As mentioned earlier, regression estimators constructed according to (1.4) are generally not robust. To overcome this shortcoming, we employ an outlier-resistant function ζ and propose to find a_j's to minimize

$$\sum_{i=1}^{n} \zeta(Y_i - \sum_{j=0}^{p} a_j (X_i - x)^j) K(\frac{X_i - x}{h_n}). \tag{2.1}$$

Equivalently, when ζ is differentiable with derivative ψ, we find a_j's that satisfy the local estimation equations:

$$\Psi_{nk}(\mathbf{a}(x)) := \sum_{i=1}^{n} \psi(Y_i - \sum_{j=0}^{p} a_j (X_i - x)^j) \frac{1}{h_n} K(\frac{X_i - x}{h_n})(X_j - x)^k = 0, \quad (2.2)$$

for $k = 0, \ldots, p$. Define $\Psi_n(\mathbf{a}(x)) := (\Psi_{n0}(\mathbf{a}(x)), \ldots, \Psi_{np}(\mathbf{a}(x)))^T$.

The local M-type estimator of $\mathbf{a}(x) \equiv (m(x), \ldots, m^{(p)}(x)/p!)^T$ is the solution to (2.2). We denote it by $\hat{\mathbf{a}}(x) = (\hat{a}_0(x), \ldots, \hat{a}_p(x))^T$.

For a given point x_0 in the interior of the support of the marginal density $f_X(x)$, the following notation and assumptions are needed.

(A1) The kernel function K is a continuous probability density function with bounded support $[-1, 1]$, say. Let $s_\ell = \int_{-1}^{1} K(u) u^\ell du$, $v_\ell = \int_{-1}^{1} u^\ell K^2(u) du$, for $\ell \geq 0$.

(A2) The regression function $m(\cdot)$ has a continuous $(p+1)th$ derivative at the point x_0.

(A3) The sequence of bandwidths h_n tends to zero and $nh_n \to +\infty$ as $n \to +\infty$.

(A4) $E[\psi(\varepsilon)|X = x_0] = 0$ with $\varepsilon = Y - m(X)$.

(A5) The marginal density $f_X(\cdot)$ of X_i is continuous at the point x_0 and $f_X(x_0) > 0$.

(A6) The function $\psi(\cdot)$ is continuous and has a derivative $\psi'(\cdot)$ almost everywhere. Further, assume that $\Gamma_1(x) = E[\psi'(\varepsilon)|X = x]$ and $\Gamma_2(x) = E[\psi^2(\varepsilon)|X = x]$ are positive and continuous at x_0, and there exists $\gamma > 0$ such that $E[|\psi^{2+\gamma}(\varepsilon)| \, |X = x]$ is bounded in a neighborhood of x_0.

(A7) The function $\psi'(\cdot)$ satisfies that $E[\sup_{|z| \leq \delta} |\psi'(\varepsilon + z) - \psi'(\varepsilon)| \, |X = x] = o(1)$ and $E[\sup_{|z| \leq \delta} |\psi(\varepsilon + z) - \psi(\varepsilon) - \psi'(\varepsilon) z| \, |X = x] = o(\delta)$, as $\delta \to 0$, uniformly in x in a neighborhood of x_0.

(B1) Either the process $\{(X_j, Y_j)\}$ is ρ-mixing with $\sum_\ell \rho(\ell) < +\infty$, or is strongly mixing with $\sum_\ell \ell^a [\alpha(\ell)]^b < +\infty$, for some $0 < b < 1$ and $a > b$, where $\rho(\ell)$, $\alpha(\ell)$, and the definitions of ρ-mixing and strongly mixing are the same as in Masry and Fan (1997).

(B2) $f(u, v; \ell) \leq M_1 < +\infty$, $E\{\psi^2(\varepsilon_1) + \psi^2(\varepsilon_\ell) \mid X_1 = u, X_\ell = v\} \leq M_2 < +\infty$, $\forall \ell \geq 1$, for u and v in a neighbourhood of x_0, where $f(u, v; \ell)$ is the joint density of X_1 and $X_{\ell+1}$.

(B3) For ρ-mixing and strongly mixing processes, we assume there exists a sequence of positive integers satisfying $s_n \to +\infty$ and $s_n = o(\sqrt{nh_n})$ such that $\sqrt{n/h_n}\rho(s_n) \to 0$ and $\sqrt{n/h_n}\alpha(s_n) \to 0$, as $n \to +\infty$.

(B4) The conditional distribution of ε given $X = x$ is continuous at the point $x = x_0$.

The above conditions are satisfied in many applications. Conditions (A1)-A(7) were proposed by Fan and Jiang (2000), where monotonicity and boundedness of $\psi(x)$ are not required. Condition (A7) is weaker than the Lipschitz continuity of the function $\psi'(x)$. It appears to be a minimal smoothness assumption on $\psi(x)$. In particular, Huber's $\psi(x)$ function satisfies this requirement. The bounded support restriction on $K(\cdot)$ is not essential, it is imposed to avoid technicalities of proofs and can be removed if we put restriction on the tail of $K(\cdot)$. We do not need the convexity of $\zeta(\cdot)$ required in Fan, Hu and Truong (1994). Also, we do not need the symmetry of the conditional distribution of ε given X. That is required by Härdle and Tsybakov (1988). The conditions (B1), (B3) and (B4) are the same as those in Masry and Fan (1997). Condition (B2) is a natural modification of the condition 2(ii) in Masry and Fan (1997). It is worth pointing out that the conditions we employ on the ψ-function are considerably weaker than those of Bianco and Boente (1998).

3. Asymptotic Properties

In this section, we establish the consistency and joint asymptotic normality of *LOPKREs*. Let $\mathbf{H} = \mathrm{diag}(1, h_n, \ldots, h_n^p)$, $\mathbf{c}_p = (s_{p+1}, \ldots, s_{2p+1})^T$, $\mathbf{S} = (s_{i+j-2})$ and $\mathbf{S}^* = (v_{i+j-2})$, $(1 \le i \le p+1;\ 1 \le j \le p+1)$ be $(p+1) \times (p+1)$ matrices.

Theorem 3.1. *Under* (A1)-(A7) *and* (B1)-(B2)*, there exist solutions, denoted by* $\hat{\mathbf{a}}(x_0)$*, to* (2.2) *such that* $\mathbf{H}(\hat{\mathbf{a}}(x_0) - \mathbf{a}(x_0)) \xrightarrow{P} 0$, as $n \to \infty$. *If in addition* (B3) *and* (B4) *hold, then*

$$\sqrt{nh_n}\,\{\mathbf{H}(\hat{\mathbf{a}}(x_0) - \mathbf{a}(x_0) - \frac{m^{(p+1)}(x_0)h_n^{p+1}}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_p(1+o_p(1))\} \xrightarrow{L} \mathcal{N}(0, \sigma^2(x_0)\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}/f_X(x_0)), \tag{3.1}$$

where $\sigma^2(x_0) = \Gamma_2(x_0)/\Gamma_1^2(x_0)$.

Remark 3.1. Assume the design density has the bounded support $[0,1]$. Consider the local polynomial fitting at the point $x_0 = dh_n$ in the left boundary region for some positive constant $1 \le d \le 0$. Then (3.1) continues to hold with slight modifications on the definition of moments:

$$s_i = \int_{-d}^{1} u^i K(u)du, \quad \text{and} \quad v_i = \int_{-d}^{1} u^i K^2(u)du. \tag{3.2}$$

A similar result holds for right boundary points. This property implies that the local polynomial M-estimation shares a similar boundary adaptation with least-squares local polynomial fitting (see Ruppert and Wand (1994)). (3.1) implies

that the optimal bandwidth for estimating $m^{(k)}(x_0)$, in the sense of minimizing the mean squared error of the asymptotic distribution, is

$$h_{k,opt} = n^{-\frac{1}{2p+3}} \{ \frac{[(p+1)!]^2 V_k \sigma^2(x_0)/f_X(x_0)}{2(p+1-k)[m^{(p+1)}(x_0)]^2 B_k^2} \}^{1/(2p+3)}, \tag{3.3}$$

where B_k and V_k are, respectively, the kth element of $\mathbf{S}^{-1}\mathbf{c}_p$ and the kth diagonal element of $\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}$.

4. One-Step Local M-estimators

The previous section establishes asymptotic properties of *LOPKREs* under certain conditions. It is clear that these properties reflect those mentioned in Fan and Jiang (2000) for i.i.d. data.

In practice, the computation of the estimator $\hat{\mathbf{a}}(x_0)$ is a data issue. We use Newton's method as in Fan and Jiang (2000), with initial value ${}_0\mathbf{a}(x_0)$ given by the local least squares estimator as in Masry and Fan (1997). Then the first iteration has the form

$${}_1\mathbf{a}(x_0) = {}_0\mathbf{a}(x_0) - \mathbf{W}_n^{-1}\Psi_n({}_0\mathbf{a}(x_0)), \tag{4.1}$$

where $\mathbf{W}_n = (w_{\ell m})$ is a $(p+1)\times(p+1)$ matrix with $w_{\ell m} = \frac{\partial}{\partial a_m}\Psi_{n\ell}({}_0\mathbf{a}(x_0))$, for $\ell = 0,\ldots,p$ and $m = 0,\ldots,p$. We label ${}_1\mathbf{a}(x_0)$ in (4.1) the *one-step local M-estimator*.

One-step local estimators have the same computational expediency as local least squares estimators. We now show that one-step local M-estimators have the same asymptotic performance as local M-estimators $\hat{\mathbf{a}}(x_0)$, as long as the initial estimators are good enough (i.e., ${}_0\mathbf{a}(x_0)$ satisfies the assumption in Theorem 4.1 below.) In other words, one-step local M-estimators reduce computational cost without downgrading performance.

Theorem 4.1. *Assume* ${}_0\mathbf{a}(x_0)$ *satisfies* $\mathbf{H}[{}_0\mathbf{a}(x_0) - \mathbf{a}(x_0)] = O_p(h_n^{p+1} + \frac{1}{\sqrt{nh_n}})$. *Then, under conditions* (A1)-(A7) *and* (B1)-(B4)*, the normalized one-step local M-estimators satisfy*

$$\begin{aligned}\sqrt{nh_n}\,\{\mathbf{H}({}_1\mathbf{a}(x_0) - \mathbf{a}(x_0) - \frac{m^{(p+1)}(x_0)h_n^{p+1}}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_p(1+o_p(1))\} \\ \xrightarrow{L} \mathcal{N}(0, \sigma^2(x_0)\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}/f_X(x_0)),\end{aligned} \tag{4.2}$$

where $\sigma^2(x_0)$ *is the same as that in Theorem* 3.1.

Remark 4.1. The condition on the initial estimators in Theorem 4.1 is mild. Most commonly used nonparametric regression estimators satisfy the condition

(see, e.g., Boente and Fraiman (1995), Masry and Fan (1997)). Especially, the local median estimator is a sensible and robust choice for the initial estimator.

5. Choice of ψ-function

In this section, we consider a minimax choice of the ψ-function with respect to Huber's ψ_k-class (see Huber (1964)). We indicate a data-driven selection of the parameter k.

Let F be the distribution function of ε, $\mathcal{C}$ be the set of all symmetric contaminated normal distributions $F = (1-t)\Phi + tH$, where $0 \le t \le 1$ is fixed and H varies over all symmetric distributions. By Theorem 3.2, we define an estimator of ψ to be the ψ^* which minimaxes the asymptotic variance parameter $\sigma^2(\psi, x_0, F) = \Gamma_2(x_0)/\Gamma_1^2(x_0) = E(\psi^2(\varepsilon)|X = x_0)/[E(\psi'(\varepsilon)|X = x_0)]^2$, that is

$$\sup_{F\in\mathcal{C}} \sigma^2(\psi^*, x_0, F) = \inf_{\psi\in\text{“}nice\text{”}} \sup_{F\in\mathcal{C}} \sigma^2(\psi, x_0, F), \tag{5.1}$$

where $\psi \in$ “*nice*” means that the ψ-function satisfies all conditions related to $\psi(x)$ in Section 2.

For fixed x_0, since the asymptotic variance parameter is similar to Huber's $\sigma^2(T, F)$ for the location model (see Huber (1964)), it can be shown that ψ^* corresponds to the Huber's ψ-functions: $\psi_k(u) = \max\{-k, \min(k, u)\}$, where k is a parameter. Here we consider the following data-driven choice for the parameter k in Huber's ψ_k-function. Theoretically, one should choose k_{opt} to minimize $\sigma^2(\psi_k, x_0, F)$. Unfortunately, $\sigma^2(\psi_k, x_0, F)$ includes the unknown distribution of the error. However, $\sigma^2(\psi_k, x_0, F)/s_0 f_X(x_0)$ can be consistently estimated by

$$\hat{\sigma}^2(\psi_k, x_0, F) = \frac{n^{-1}h_n^{-1}\sum_{j=1}^n \psi_k^2(\hat{\varepsilon}_j)K((X_j - x_0)/h_n)}{[n^{-1}h_n^{-1}\sum_{j=1}^n \psi_k'(\hat{\varepsilon}_j)K((X_j - x_0)/h_n)]^2}, \tag{5.2}$$

where $\hat{\varepsilon}_j = Y_j - \hat{m}(x_0)$, and $\hat{m}(x_0)$ is any consistent estimator of $m(x_0)$, such as the initial estimator in (4.1). Therefore, one viable choice for k is to find $\hat{k}$ to minimize $\hat{\sigma}^2(\psi_k, x_0, F)$. Certainly, one may study the optimal choice of ψ, in the minimax sense, for the robust local polynomial regression under other contamination distributions classes, such as the asymmetric contamination considered in Jaeckel (1971), by using the same idea. On the other hand, the regression case considered in this article is much more complicated, and we will not pursue it here.

6. Numerical Illustrations

In this section, we present two examples for the case $p = 1$. The objective is to illustrate our method rather than giving a completely data-driven recipe.

Example 1. The first example is based on a Monte Carlo experiment. We generated 400 samples, each of size $n = 200$ from the model:

$$Y_t = 0.3\exp\{-4(X_t+1)^2\} + 0.7\exp\{-16(X_t-1)^2\} + \varepsilon_t, \tag{6.1}$$

where $X_t = -0.4X_{t-1} + u_t$, $u_t \overset{iid}{\sim} \mathcal{N}(0, 0.8^2)$ for all $t = 1, \ldots, 200$, and the error process ε_t is independent of X_t, generated according to the contaminated Gaussian model $\varepsilon_t \overset{iid}{\sim} 0.1\mathcal{N}(0, 25\tau^2) + 0.9\mathcal{N}(0, \tau^2)$. We chose $\tau = 0.075$ so that $Var(\varepsilon_t) = 0.248^2$. The model was used in Fan and Gijbels (1995) except that $X_t \sim Uniform(-2, 2)$ and $\varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, 0.1^2)$.

For each sample, we use the optimal bandwidth h_{opt} as described in (3.3) and the standard Gaussian kernel. For robust implementation, Huber's ψ_k- function was employed with $\hat{k}$ selected according to (5.2) for each sample. We obtained the one-step local M-estimator by using the estimator of Masry and Fan (1997) as the initial value. Then the one-step local M-estimator was used as initial value to get the two-step local M-estimator, which was shown to be nearly as efficient as the fully iterative M-estimator in Fan and Jiang (2000). Our experience shows that, when the data is contaminated with thick-tailed noise and a commonly used nonparametric estimator, for instance that in Masry and Fan (1997), is employed as the initial estimator of $\mathbf{a}(x_0)$, the one-step local M-estimator is not good, but the two-step local M-estimator performs well. For comparison, we ran a local linear least squares regression with the same kernel. The entire procedure was repeated 400 times. For each sample, we measured the performance of the estimators according to the mean absolute deviation error (MADE) criterion:

$$MADE(\hat{m}) = N^{-1}\sum_{j=1}^{N} |\hat{m}(x_j) - m(x_j)|,$$

where $x_j, j = 1, \ldots, N$ ($N = 60$) are grid points. The typical sample chosen is the one with which the local linear least squares estimator of Masry and Fan (1997) has its median performance, in terms of MADE, among 400 simulations.

We display in Figure 1(a) the "typical" sample out of the 400 samples generated as above, together with the true function (6.1), and the two regression fits. The estimated $\hat{k}$ according to (5.2) is 0.219 for this sample.

Figure 1(b) includes plots of the true function, the *median curves* (i.e., the median of the estimators among 400 simulations) for the local linear least squares fit as well as the two-step local M-estimator fit. Figure 2 shows the *median curves* with envelopes formed via pointwise 2.5% and 97.5% sample percentiles for the two fits. It is evident that our robust estimator is the better one.

Example 2. We analyze an economic data set from the United Kingdom. The Y-variable of this bivariate data set is the rate of change of money wages, and the X-variable is the corresponding unemployment rate, for the period 1861-1913. This data set has been studied in Phillips (1958) and gave rise to the famous, perhaps controversial, *Phillips curve*. The sources of our data are Phillips (1958), Lipsey (1960), and the British Ministry of Labour Gazette.

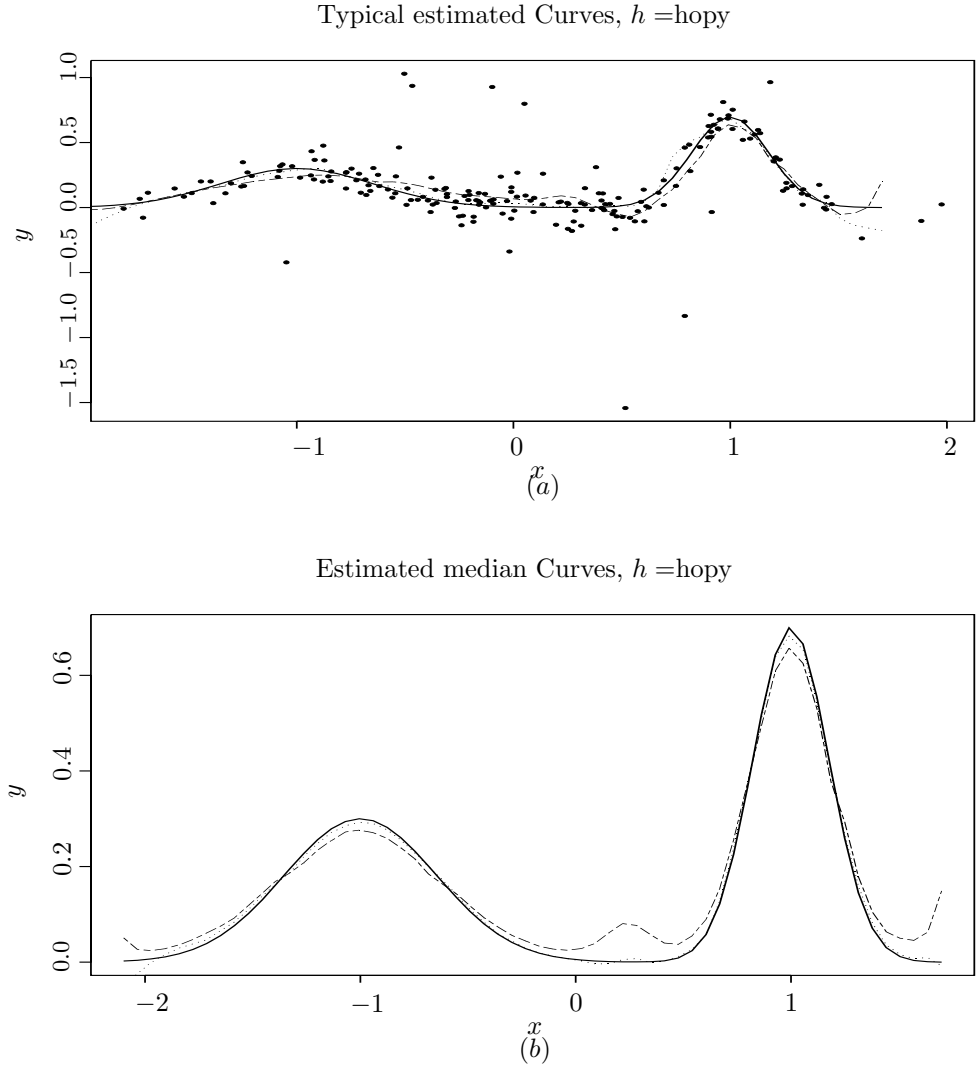
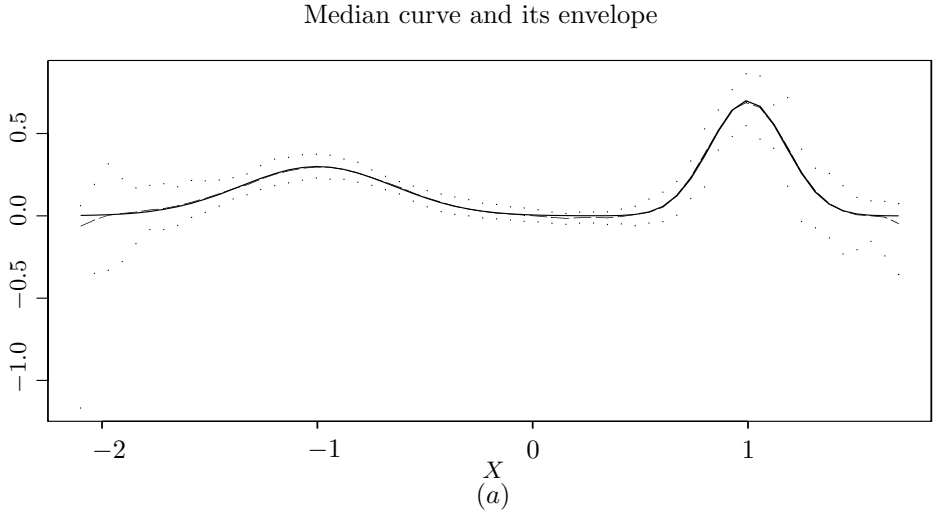


Figure 1. Simulation results for Example 1. Typical estimated curves among 400 simulations are presented in (a); the median curves (out of 400 simulations) and the true curve are shown in (b). Solid curve: true regression function, dash: local linear least squares estimator, dotted: two-step local M-estimator.



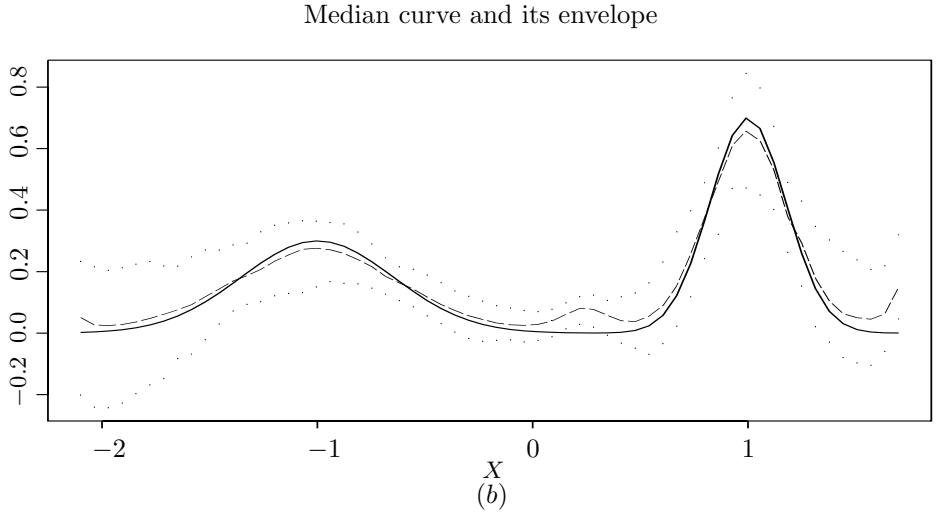


Figure 2. The median curves (out of 400 simulations) with envelopes formed via pointwise 2.5% and 97.5% sample percentiles for Example 1. (a)– two-step local M-estimator fit; (b) – local linear least squares fit. Solid curve: ture curves; dash: median curves ; dotted: envelopes.

A nonlinear model for the data from 1861 – 1913 was obtained by some unconventional fitting techniques in Phillips (1958). A simplified version of this model has the form

$$y = a + bx^{-c}, \tag{6.2}$$

with well-accepted parameter values $a = -0.9223$, $b = 8.9679$, $c = 1.3506$. We

call (6.2), with the given parameter values, our *reference model* for the discussion of the second example.

For nonparametric regression estimation, we again employed the standard Gaussian kernel, the bandwidth was based on the optimal one in (3.3), $\sigma^2(x)$ was assumed to be constant and estimated by the trimmed mean (with 5% trimming) of the squared residuals from the reference model (6.2), $m''(x)$ was also calculated using the reference model (6.2) and $f_X(x)$ was estimated using a consistent kernel estimator $\hat{f}_X(x)$ from the X-data.

A local linear least squares regression, as in Masry and Fan (1997), as well as the robust procedure suggested in the present study, were implemented. For the robust approach, the value of k in Huber's ψ_k function was determined from the data ($k = 3.4995$) as described in Section 5. The two nonparametric fits plus the Phillips curve were plotted on the same scales in Figure 3 together with the raw data points. It can be seen that the local linear least squares fit is influenced by the large negative Y-data values toward the right tail.

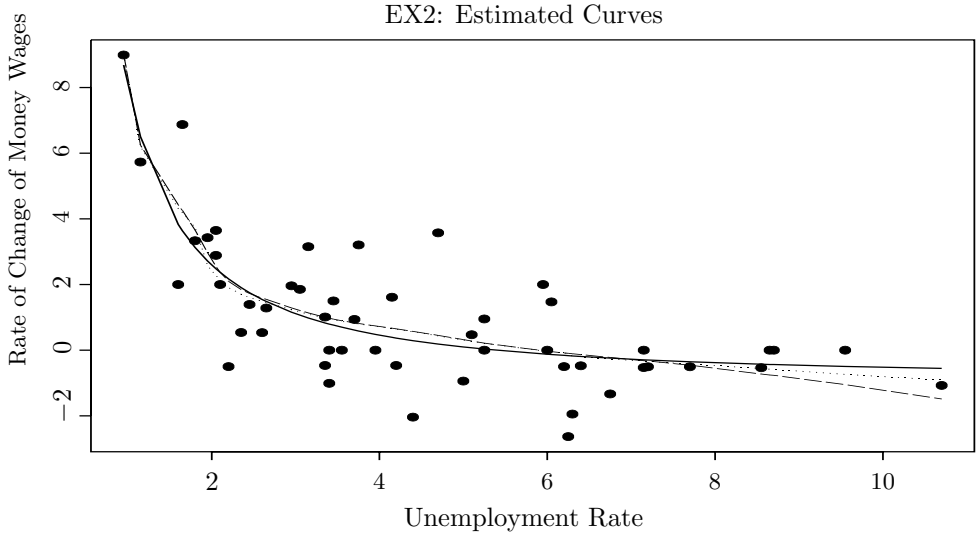


Figure 3. Numerical results for Example 2. Solid curve: Phillips curve; dash: local least squares; dotted: local two-step.

Acknowledgements

The authors are grateful to the Associate Editor and referees for their comments. Part of the research was carried out while both authors were visiting the Statistics Department of the University of California at Berkeley. The authors would like to express their gratitude for the support they received during the visit. Supported in part by Chinese NSF Grants 39930160 and 10001004.

7. Appendix

In this section, we give proofs of Theorems 3.1 and 4.1. Even though our technical devices are analogous to those in Fan and Jiang (2000), for general polynomial fitting (solving (2.2) for arbitrary p) under mixing conditions, the derivation of the asymptotic distributions of the resulting estimators is considerably more involved. The following notations and lemmas will be used for our technical proofs. Let $K_h(X_j) = \frac{1}{h_n}K(\frac{X_j-x_0}{h_n})$, and let $R(X_j) = m(X_j) - \sum_{\ell=0}^{p} \frac{1}{\ell!}m^{(\ell)}(x_0)(X_j-x_0)^\ell$, $\tilde{\mathbf{a}}(x) = \mathbf{H}\mathbf{a}(x)$, $\mathbf{x}_j = (1, X_j - x_0, \ldots, (X_j - x_0)^p)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\tilde{\mathbf{X}} \equiv (\tilde{\mathbf{x}}_\mathbf{1}, \ldots, \tilde{\mathbf{x}}_\mathbf{n})^T = \mathbf{H}^{-1}\mathbf{X}$.

Lemma 7.1. *Assume* (A1)-(A7) *and* (B1)-(B2) *hold. Let* $\Gamma_3(x) = E[(\psi'(\varepsilon))^2|X = x]$, $Z_{i\ell} = \psi'(\varepsilon_i)K_h(X_i)\left(\frac{X_i-x_0}{h_n}\right)^\ell$, *and* $Q_{n\ell} = n^{-1}\sum_{j=1}^{n} Z_{i\ell}$. *Then*

(1) $h_n Var(Z_{1\ell}) = \Gamma_3(x_0)f_X(x_0)v_{2\ell}(1 + o(1))$;
(2) $h_n \sum_{j=1}^{n} |cov(Z_{1\ell}, Z_{(j+1)\ell})| = o(1)$;
(3) $nh_n Var(Q_{n\ell}) = \Gamma_3(x_0)f_X(x)v_{2\ell}(1 + o(1))$.

Proof. The results hold by using the argument of Theorem 2 in Masry and Fan (1997).

Lemma 7.2. *Assume* (A1)-(A7) *and* (B1)-(B2) *hold. For any random sequence* $\{\eta_j\}_{j=1}^n$, *if* $\max_{1\le j\le n}|\eta_j| = o_p(1)$ *we have*

$$n^{-1}\sum_{j=1}^{n}\psi'(\varepsilon_j+\eta_j)K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell = \Gamma_1(x_0)f_X(x_0)s_\ell(1+o_p(1)),$$

$$n^{-1}\sum_{j=1}^{n}\psi'(\varepsilon_j+\eta_j)R(X_j)K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell$$

$$= \frac{1}{(p+1)!}h_n^{p+1}\Gamma_1(x_0)m^{(p+1)}(x_0)f_X(x_0)s_{\ell+p+1}(1+o_p(1)).$$

Proof. We give the proof of the first conclusion, the second can be shown by the same arguments. It is obvious that

$$n^{-1}\sum_{j=1}^{n}\psi'(\varepsilon_j+\eta_j)K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell$$

$$= n^{-1}\sum_{j=1}^{n}\psi'(\varepsilon_j)K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell + \sum_{j=1}^{n}[\psi'(\varepsilon_j+\eta_j)-\psi'(\varepsilon_j)]K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell$$

$$\equiv T_{n,1} + T_{n,2}.$$

By taking iterative expectation, we get

$$ET_{n,1} = n^{-1}E[\sum_{j=1}^{n}K_h(X_j)(\frac{X_j-x_0}{h_n})^\ell E(\psi'(\varepsilon_j)|X_j)] = \Gamma_1(x_0)f_X(x_0)s_\ell(1+o(1)).$$

By Lemma 7.1, we know $Var(T_{n,1}) = Var(Q_{n\ell}) = O(\frac{1}{nh_n})$. It follows that $T_{n,1} = \Gamma_1(x_0)f_X(x_0)s_\ell(1 + o_p(1))$, so it suffices to show that $T_{n,2} = o_p(1)$. For any given $\eta > 0$, let $\Delta_n = (\delta_1, \ldots, \delta_n)^T$, $D_\eta = \{\Delta_n : |\delta_j| \le \eta, \forall j \le n\}$, and

$$V(\Delta_n) = n^{-1}\sum_{j=1}^{n}[\psi'(\varepsilon_j + \delta_j) - \psi'(\varepsilon_j)]K_h(X_j)(\frac{X_j - x_0}{h_n})^\ell.$$

Then

$$\sup_{D_\eta}|V(\Delta_n)| \le n^{-1}\sum_{j=1}^{n}\sup_{D_\eta}|\psi'(\varepsilon_j) - \psi'(\varepsilon_j + \delta_j)|K_h(X_j)|\frac{X_j - x_0}{h_n}|^\ell.$$

By (A7), noticing that $|X_j - x_0| \le h_n$ in the above expression, we have

$$E[\sup_{D_\eta}|V(\Delta_n)|] \le a_\eta n^{-1}E[\sum_{j=1}^{n}K_h(X_j)|\frac{X_j - x_0}{h_n}|^\ell]$$

where a_η and b_η are two sequences of positive numbers, tending to zero as $\eta \to 0$. Since $\sup_{1\le j\le n}|\eta_j| = o_p(1)$, it follows that $V(\hat{\Delta}_n) = o_p(1)$ with $\hat{\Delta}_n = (\eta_1, \ldots, \eta_n)^T$. The conclusion follows from the fact $T_{n,2} = V(\hat{\Delta}_n) = o_p(1)$.

Lemma 7.3. *Assume* (A1)-(A7) *and* (B1)-(B4) *hold. Let*

$$\mathbf{J}_n \equiv \begin{pmatrix} J(0) \\ \vdots \\ J(p) \end{pmatrix} \equiv \begin{pmatrix} n^{-1}\sum_{j=1}^{n}\psi(\varepsilon_j)K_h(X_j) \\ \vdots \\ n^{-1}\sum_{j=1}^{n}\psi(\varepsilon_j)K_h(X_j)(X_j - x_0)^p/h_n^p \end{pmatrix}.$$

Then $\sqrt{nh_n}\mathbf{J}_n$ *is asymptotically normal with mean zero and covariance matrix* $D = \Gamma_2(x_0)f_X(x_0)\mathbf{S}^*$.

Proof. For any linear combination of $J(0), \ldots, J(p)$: $Q_n = \sum_{\ell=0}^{p}c_\ell J(\ell) = n^{-1}\sum_{i=1}^{n}\xi_i$, where $\xi_i = \psi(\varepsilon_i)\sum_{\ell=0}^{p}c_\ell K_h(X_i)(\frac{X_i - x_0}{h_n})^\ell$. By the argument of Theorem 3 in Masry and Fan (1997), we get $\sqrt{nh_n}Q_n \xrightarrow{L} N(0, \theta^2(x_0))$, where $\theta^2(x_0) = \Gamma_2(x_0)f_X(x_0)\int_{-1}^{1}(\sum_{\ell=0}^{p}c_\ell u^\ell K(u))^2du$. That is, $\sqrt{nh_n}\mathbf{J}_n$ is asymptotically normal. By computing the variance-covariance matrix of $\sqrt{nh_n}\mathbf{J}_n$, we get the result of the lemma.

Proof of Theorem 3.1. Note that (2.1) can be written as $\ell_n(\tilde{\mathbf{a}}) = \sum\zeta(Y_j - \tilde{\mathbf{x}}_\mathbf{j}^T\tilde{\mathbf{a}})K(\frac{X_j - x_0}{h_n})$. Let $\mathbf{S}_\delta = \{\tilde{\mathbf{a}} : \|\tilde{\mathbf{a}} - \tilde{\mathbf{a}}(x_0)\| \le \delta\}$. Denote by $r_j = (\tilde{\mathbf{a}} - \tilde{\mathbf{a}}(x_0))^T\tilde{\mathbf{x}}_\mathbf{j}$. Then $Y_j - \tilde{\mathbf{x}}_\mathbf{j}^T\tilde{\mathbf{a}} = \varepsilon_j + R(X_j) - r_j$.

We show that, for any sufficiently small δ,

$$\lim_{n\to\infty}P\{\inf_{\tilde{\mathbf{a}}\in\mathbf{S}_\delta}\ell_n(\tilde{\mathbf{a}}) \ge \ell_n(\tilde{\mathbf{a}}(x_0))\} = 1. \tag{7.1}$$

In fact, by integration, we have

$$
\begin{aligned}
& n^{-1}[\ell_n(\tilde{\mathbf{a}}) - \ell_n(\tilde{\mathbf{a}}(x_0))] \qquad (7.2)\\
&= n^{-1}\sum_{j=1}^{n}[K_h(X_j)\int_{\varepsilon_j+R(X_j)}^{\varepsilon_j+R(X_j)-r_j}\psi(t)\,dt]\\
&= n^{-1}\sum_{j=1}^{n}K_h(X_j)\int_{\varepsilon_j+R(X_j)}^{\varepsilon_j+R(X_j)-r_j}[\psi(\varepsilon_j)+\psi'(\varepsilon_j)(t-\varepsilon_j)+(\psi(t)-\psi(\varepsilon_j)\\
&\quad -\psi'(\varepsilon_j)(t-\varepsilon_j))]\,dt\\
&\equiv K_{n1}+K_{n2}+K_{n3}. \qquad (7.3)
\end{aligned}
$$

By the argument of Theorem 2(c) in Masry and Fan (1997), we have

$$
\begin{aligned}
K_{n1} &= -(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T n^{-1}\sum_{j=1}^{n}K_h(X_j)\psi(\varepsilon_j)\tilde{\mathbf{x}}_{\mathbf{j}}\\
&= o_p(1)\delta. \qquad (7.4)
\end{aligned}
$$

By the Mean Value Theorem for integration, we have

$$
K_{n3} = -(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T n^{-1}\sum_{j=1}^{n}K_h(X_j)[\psi(\varepsilon_j+z_j)-\psi(\varepsilon_j)-\psi'(\varepsilon_j)z_j]\tilde{\mathbf{x}}_{\mathbf{j}},
$$

where z_j lies between $R(X_j)$ and $R(X_j)-r_j$, for $j=1,\ldots,n$. Note that for $|X_j-x_0|\le h_n$, we have $\max_j|z_j|\le\max_j|R(X_j)|+2\delta$. Then by condition (A7), we obtain

$$
K_{n3} = o_p(1)\delta^2. \qquad (7.5)
$$

Note that by simple integration $K_{n2} = \frac{1}{2n}\sum_{j=1}^{n}K_h(X_j)\psi'(\varepsilon_j)\ [(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T\ \tilde{\mathbf{x}}_{\mathbf{j}}\ \tilde{\mathbf{x}}_{\mathbf{j}}^T\ (\tilde{\mathbf{a}}-\ \tilde{\mathbf{a}}(x_0))-2R(X_j)r_j] \equiv M_{n1}+M_{n2}$. It is obvious from Lemma 7.2 that $M_{n1} = \frac{1}{2}\Gamma_1(x_0)f_X(x_0)(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T\mathbf{S}(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))(1+o_p(1))$ and $M_{n2} = -(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T n^{-1}\sum_{j=1}^{n}K_h(X_j)\psi'(\varepsilon_j)R(X_j)\tilde{\mathbf{x}}_{\mathbf{j}} = O_p(h_n^{p+1})\delta$. Therefore $K_{n2} = \frac{1}{2}\Gamma_1(x_0)f_X(x_0)(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))^T\mathbf{S}(\tilde{\mathbf{a}}-\tilde{\mathbf{a}}(x_0))(1+o_p(1))+O_p(h_n^{p+1})\delta$. Let a be the largest eigenvalue of the positive definite matrix S. Then, for any $\tilde{\mathbf{a}}\in\mathbf{S}_\delta$, we have for sufficiently small δ that $\lim_{n\to\infty}P\{\inf_{\tilde{\mathbf{a}}\in\mathbf{S}_\delta}K_{n2} > -\frac{1}{2}af_X(x_0)\Gamma_1(x_0)\delta^2\}=1$. This together with (7.3), (7.4) and (7.5) establish (7.1).

By (7.1), $\ell_n(\tilde{\mathbf{a}})$ has a local minimum in the interior of $\mathbf{S}_\delta$. Since at a local minimum, (2.2) must be satisfied. Let $\tilde{\mathbf{a}}^*(x_0)$ be the closest root to $\tilde{\mathbf{a}}(x_0)$, and $\hat{\mathbf{a}}(x_0) = \mathbf{H}^{-1}\tilde{\mathbf{a}}^*(x_0)$. Then $\lim_{n\to\infty}P\{||\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))||\le\delta^2\}=1$. This implies the weak consistency part of Theorem 3.1.

For the asymptotic distribution part, let $\hat{\eta}_j = R(X_j)-\mathbf{x}_j{}^T(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))$. Then $Y_j-\mathbf{x}_j{}^T\hat{\mathbf{a}}(x_0)=\varepsilon_j+\hat{\eta}_j$. It follows from (2.2) that

$$
\sum_{j=1}^{n}\{\psi(\varepsilon_j)+\psi'(\varepsilon_j)\hat{\eta}_j+[\psi(\varepsilon_j+\hat{\eta}_j)-\psi(\varepsilon_j)-\psi'(\varepsilon_j)\hat{\eta}_j]\}K_h(X_j)\tilde{\mathbf{x}}_j=0. \qquad (7.6)
$$

Note that the second term in the left hand side of (7.6) is

$$\sum_{j=1}^{n}\psi'(\varepsilon_j)R(X_j)K_h(X_j)\tilde{\mathbf{x}}_j-\sum_{j=1}^{n}\psi'(\varepsilon_j)K_h(X_j)\tilde{\mathbf{x}}_\mathbf{j}\tilde{\mathbf{x}}_\mathbf{j}^T\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))\equiv L_{n1}+L_{n2}.$$

Applying Lemma 7.2, we obtain $L_{n1}=nh_n^{p+1}\frac{\Gamma_1(x_0)}{(p+1)!}f_X(x_0)m^{(p+1)}(x_0)\mathbf{c}_p(1+o_p(1))$ and $L_{n2}=-\Gamma_1(x_0)f_X(x_0)n\mathbf{SH}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))(1+o_p(1))$. Note that by the consistency of $\mathbf{H}\hat{\mathbf{a}}(x_0)$

$$\begin{aligned}\sup_{j:|X_j-x_0|\le h_n}|\hat{\eta}_j|&\le\sup_{j:|X_j-x_0|\le h_n}(|R(X_j)|+||\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))||)\\&=O_p(h_n^{p+1}+||\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))||).\end{aligned}$$

Then by (A7) and the argument in Lemma 5.2, the third term on the left hand side of (7.6) is given by $o_p(n)[h_n^{p+1}+||\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))||]=o_p(L_{n1}+L_{n2})$. Let $\mathbf{B}_n=\frac{m^{(p+1)}(x_0)h_n^{p+1}}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_p(1+o_p(1))$. Then, it follows from (7.6) that

$$\mathbf{H}(\hat{\mathbf{a}}(x_0)-\mathbf{a}(x_0))=\mathbf{B}_n+\Gamma_1^{-1}(x_0)f_X^{-1}(x_0)\mathbf{S}^{-1}\mathbf{J}_n(1+o_p(1)),\tag{7.7}$$

where $\mathbf{J}_n$ is given in Lemma 7.3. The conclusion follows from (7.7), Lemma 7.3 and Slutsky's Theorem.

Proof of Theorem 4.1. Let $\hat{\delta}_j=R(X_j)-({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))^T\mathbf{x}_j$. Then

$$\begin{aligned}\max_{j:|X_j-x_0|\le h_n}|\hat{\delta}_j|&\le\max_{j:|X_j-x_0|\le h_n}(|R(X_j)|+||\mathbf{H}({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0)||)\\&=O_p(h_n^{p+1}+||\mathbf{H}({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))||)=O_p(h_n^{p+1}+\frac{1}{\sqrt{nh_n}}).\end{aligned}\tag{7.8}$$

By the definitions of $\Psi_n(\mathbf{a}(x_0))$ and Lemma 7.2, we have

$$\begin{aligned}w_{\ell m}&=-\sum_{j=1}^{n}\psi'(Y_j-{}_0\mathbf{a}(x_0)^T\mathbf{x}_j)K_h(X_j)(X_j-x_0)^{\ell+m}\\&=-\sum_{j=1}^{n}\psi'(\varepsilon_j+\hat{\delta}_j)K_h(X_j)(X_j-x_0)^{\ell+m}\\&=-nh_n^{\ell+m}\Gamma_1(x_0)f_X(x_0)s_{\ell+m}(1+o_p(1)).\end{aligned}$$

Therefore $\mathbf{W}_n=-\mathbf{H}SHn\Gamma_1(x_0)f_X(x_0)(1+o_p(1))$ and $\mathbf{W}_n{}^{-1}=-\mathbf{H}^{-1}\mathbf{S}^{-1}\mathbf{H}^{-1}(n\Gamma_1(x_0)f_X(x_0))^{-1}(1+o_p(1))$. In addition, by the definitions of $\hat{\delta}_j$ and $\Psi_{n\ell}(\mathbf{a}(x_0))$, we have

$$\Psi_n({}_0\mathbf{a}(x_0))=\sum_{j=1}^{n}\psi(\varepsilon_j+\hat{\delta}_j)K_h(X_j)(X_j-x_0)^{\ell}$$

$$=\sum_{j=1}^{n}\psi(\varepsilon_j)K_h(X_j)(X_j-x_0)^\ell+\sum_{j=1}^{n}\psi'(\varepsilon_j)\hat{\delta}_jK_h(X_j)(X_j-x_0)^\ell$$
$$+\sum_{j=1}^{n}[\psi(\varepsilon_j+\hat{\delta}_j)-\psi(\varepsilon_j)-\psi'(\varepsilon_j)\hat{\delta}_j]K_h(X_j)(X_j-x_0)^\ell\equiv I_{n1}+I_{n2}+I_{n3}.$$

Further, Lemma 7.2 with $\eta_j=0$ yields

$$I_{n2}=-n\Gamma_1(x_0)f_X(x_0)({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))^T\mathbf{H}(s_\ell,\ldots,s_{\ell+p})^Th_n^\ell$$
$$+\frac{\Gamma_1(x_0)}{(p+1)!}nh_n^{\ell+p+1}s_{\ell+p+1}m^{(p+1)}(x_0)f_X(x_0)(1+o_p(1)).$$

By (7.8), (A7) and the argument in Lemma 7.2, we obtain $I_{n3}=o_p(nh_n^{\ell+p+1})+o_p(n)[\mathbf{H}({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))]$. Substituting the expressions of I_{n1}, I_{n2} and I_{n3}, we get

$$\Psi_{n\ell}({}_0\mathbf{a}(x_0))$$
$$=\sum_{j=1}^{n}\psi(\varepsilon_j)K_h(X_j)(X_j-x_0)^\ell+\frac{\Gamma_1(x_0)}{(p+1)!}nh_n^{\ell+p+1}s_{\ell+p+1}m^{(p+1)}(x_0)f_X(x_0)(1+o_p(1))$$
$$-n\Gamma_1(x_0)f_X(x_0)[\mathbf{H}({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))]^T\begin{pmatrix}s_\ell\\ \vdots\\ s_{\ell+p}\end{pmatrix}h_n^\ell(1+o_p(1)). \tag{7.9}$$

Therefore

$$\Psi_n({}_0\mathbf{a}(x_0))=\frac{\Gamma_1(x_0)}{(p+1)!}nh_n^{p+1}f_X(x_0)\mathbf{H}c_p(1+o_p(1))+\sum_{j=1}^{n}\psi(\varepsilon_j)K_h(X_j)\mathbf{x}_j$$
$$-n\Gamma_1(x_0)f_X(x_0)\mathbf{H}SH({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))(1+o_p(1)). \tag{7.10}$$

It follows from Lemma 7.3 that

$$\mathbf{H}{W_n}^{-1}\Psi_n({}_0\mathbf{a}(x_0))=-\frac{h_n^{p+1}}{(p+1)!}m^{(p+1)}(x_0)\mathbf{S}^{-1}\mathbf{c}_p+\mathbf{H}({}_0\mathbf{a}(x_0)-\mathbf{a}(x_0))$$
$$+(\Gamma_1(x_0)f_X(x_0))^{-1}\mathbf{S}^{-1}\mathbf{J}_n+o_p(h_n^{p+1}+\frac{1}{\sqrt{nh_n}}). \tag{7.11}$$

Hence, by (4.1) and (7.11), we get

$$\mathbf{H}({}_1\mathbf{a}(x_0)-\mathbf{a}(x_0))=(\Gamma_1(x_0)f_X(x_0))^{-1}\mathbf{S}^{-1}\mathbf{J}_n+\frac{h_n^{p+1}}{(p+1)!}m^{(p+1)}(x_0)\mathbf{S}^{-1}\mathbf{c}_p$$
$$+o_p(h_n^{p+1}+\frac{1}{\sqrt{nh_n}}). \tag{7.12}$$

The conclusion follows from (7.12), Lemma 7.3 and Slutsky's Theorem.

References

Athreya, K. B. and Pantala, S. G. (1986). A note on strong mixing of ARMA processes. *Statist. Probab. Lett.* **4**, 187.

Baek, J. and Wehrly, T. E. (1993). Kernel estimation for additive models under dependence. *Stochastic Process Appl.* **47**, 95-112.

Bianco A. M. and Boente, G. (1998). Robust kernel estimators for additive models with dependent observations. *Canad. J. Statist.* **26**, 239-255.

Boente, G. and Fraiman, R. (1989). Robust nonparametric regression. *J. Multivariate Anal.* **29**, 180-198.

Boente, G. and Fraiman, R. (1990). Asymptotic distribution of robust estimators for nonparametric models from mixing processes. *Ann. Statist.* **18**, 891-906.

Boente, G. and Fraiman, R. (1995). Asymptotic distribution of smoothers based on local means and local medians under dependence. *J. Multivariate Anal.* **54**, 77-90.

Bradley, R. A. (1986). Basic properties of strong mixing condition. *Proc. 7th Conference on Probability Theory (ed. M. Lasifeson)*, 65-72.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.

Doukhan, P. (1994). *Mixing: Properties and Examples.* Lecture Notes in Statistics **85**. Springer-Verlag, New York.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J., Hu, T. and Truong, Y. (1994). Robust Non-parametric function estimation. *Scand. J. Statist.* **21**, 433-446.

Fan, J. and Jiang, J. (2000). Variable bandwidth and one-step local M-estimator. *Science in China, Ser. A* **43**, 65-81.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation.* Lecture Notes in Mathematics **757**, 23-68. Springer-Verlag, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeaus, P. J. and Stahl, W. A. (1986). *Robust Statistics: The Approximation Based on Influence Functions.* Wiley, New York.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, New York.

Härdle, W. and Tsybakov, A. B. (1988). Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.* **16**, 120-135.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

Huber, P. J. (1981). *Robust Estimation.* Wiley, New York.

Jaeckel, L. A. (1971). Robust estimates of location: symmetry and asymmetry contamination. *Ann. Math. Statist.* **42**, 1020-1034.

Jaeckel, L. A. (1971). Some flexible estimates of location. *Ann. Math. Statist.* **42**, 1540-1552.

Lipsey, R. G. (1960). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1862-1957: a further analysis. *Economica* **27**, 1-31.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scand. J. Statist.* **24**, 165-179.

Nadaraya, E. A. (1964). On estimating regression. *J. Probab. Appl.* **9**, 141-142.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. *Economica* **25**, 283-299.

Robinson, P. M. (1984). Robust nonparametric regression. In *Robust and Nonlinear Time Series Analysis* (Edited by J. Franke, W. Härdle and D. Martin). Springer-Verlag, New York.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.*, U.S.A. **42**, 43-47.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, Ser. A **26**, 359-372.

Welsh, A. H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statist. Sinica* **6**, 347-366.

Department of Probability and Statistics, Peking University, Beijing 100871, China.

E-mail: jiang@math.pku.edu.cn

Division of Statistics, University of California, Davis, CA 95616, U.S.A.

E-mail: ypmack@ucdavis.edu

(Received April 2000; accepted March 2001)