Statistica Sinica **17**(2007), 265-287

THRESHOLD VARIABLE SELECTION USING NONPARAMETRIC METHODS

Yingcun Xia, Wai-Keung Li and Howell Tong

National University of Singapore, University of Hong Kong and London School of Economics

Abstract: Selecting the threshold variable is a key step in building a generalized threshold autoregressive (TAR) model. This paper proposes a semi-parametric method for this purpose that is based on a single-index functional coefficient model. The asymptotic distribution of the estimator is obtained. A simple algorithm is given and its convergence is proved. Some simulations are reported. Two data sets are analyzed, one of which gives strong statistical support for ratio-dependent predation in Ecology.

Key words and phrases: Local linear smoother, nonlinear time series, single-index coefficient models, threshold autoregressive (TAR) time series models.

1. Introduction

The threshold autoregressive (TAR) model is one of the popular models in nonlinear time series. As a generalized nonlinear TAR model, a semi-parametric single-index functional coefficient model has the form

$$y_t = g_0(\theta_0^T Z_t) + g_1(\theta_0^T Z_t)x_{t,1} + \cdots + g_p(\theta_0^T Z_t)x_{t,p} + \varepsilon_t, \quad t = 1, 2, \ldots, \quad (1.1)$$

where (X_t, Z_t, y_t) are $\mathbb{R}^p$, $\mathbb{R}^q$, and $\mathbb{R}$-valued random variables respectively, with $X_t = (x_{t,1}, \ldots, x_{t,p})^T$; $\theta_0 \in \Theta = \{\theta : |\theta| = 1\}$ is an unknown parameter vector, called a single-index direction; $g_k(\cdot)$, $k = 0, \ldots, p$, are unknown coefficient functions and $E(\varepsilon_t|X_t, Z_t) = 0$ almost surely. We further assume that the first element of θ_0 is positive for model identification. Model (1.1) is a generalized semi-parametric threshold autoregressive model if we take X_t and Z_t to be the lagged-variables of y_t. The model is also a single-indexing version of the varying functional coefficient model proposed by Hastie and Tibshirani (1993) under an IID setting, and the functional coefficient model proposed by Chen and Tsay (1993) under a time series setting. The model has been investigated by Xia and Li (1999) and Fan, Yao and Cai (2003). Model (1.1) can give sensible approximate relations between variables due to the single-indexing construction; see Xia and Li (1999) and Fan et al. (2003). Moreover, the model can be used to select the threshold variable $\theta_0^T Z_t$ in a generalized threshold model; see Tong (1990), Chen and Tsay (1993) and Xia and Li (1999). The estimation of the threshold variable is, generally speaking, non-trivial even under parametric setting; see, e.g., Chen (1995) and Chan and Tong (1986). The difficulty results from the flexible form of the varying coefficient functions. Fortunately, the semi-parametric approach can cope with such a flexibility.

Another motivation of this research is related to a recent debate in ecology about ratio-dependent predation; see, e.g., Bohannan and Lenski (1999), Abrams and Ginzburg (2000) and Jost and Ellner (2000). Ecologists try to use functional responses to describe prey-predator interactions and the complex dynamics. The term "prey-dependent" means that the consumption rate of each single predator is only a function of prey density, and a "predator-dependent" functional response is one in which both predator and prey densities affect the per-predator consumption rate. "Ratio dependence" means that consumption is a function of the ratio of prey to predator density. Theoretical studies have shown that the dynamics of models with predator-dependent functional response can differ considerably from the dynamics of correspondingly structured models with prey-dependent functional response; see Rogers and Hassell (1974) and Kuang and Beretta (1998). The protozoan predator-prey system of *P.aurelia* and *D.nastum* is a classic in population ecology. The three pairs of time series in Figure 1 are the longest time series reported in Rao (1973) (cf., Jost and Ellner (2000)) using a refined protozoan predator-prey system under three different conditions. The mechanism of the interactions between the prey and predator populations, denoted by Y_t and R_t respectively, can be described as

$$\frac{dR_t}{dt} = f_1(R_{t-\tau_1}, Y_{t-\tau_1})R_t; \quad \frac{dY_t}{dt} = f_2(R_{t-\tau_2}, Y_{t-\tau_1})Y_t + f_3(R_{t-\tau_3}, Y_{t-\tau_1})R_t, \tag{1.2}$$

where f_1, f_2 and f_3 are functional responses and $\tau_k, k = 1, 2, 3$, are time-delays. The classic functional responses are set to be some nonlinear functions up to some unknown parameters. For example $f(u, v) = a(1+bu)^{-1}u$ (Holling type II), $f(u, v) = a(v+bu)^{-1}u$ (ratio-dependent II) and $f(u, v) = a(v^m+bu)^{-1}u$ (Hasssell-Varley type II). Simply speaking, the above debate is about whether $f_k, k = 1, 2, 3$, are functions of u only as in Holling type II functional response or functions of u/v^m for some $m > 0$ as in the ratio-dependent II or Hasssell-Varley type II functional responses. Note that all the cases can be written as functions of linear combinations $\theta_1 \log(u) + \theta_2 \log(v)$. Correspondingly, the functional response can be written as $f(u, v) = \tilde{f}(\theta_{k1} \log(u) + \theta_{k2} \log(v))$ or $\tilde{f}(\theta_{k1} U + \theta_{k2} V)$, where $U = \log(u)$ and $V = \log(v)$. Using this approach and taking $Z_{t-\tau_1} = \log(Y_{t-\tau_1})$ and $S_{t-\tau_1} = \log(R_{t-\tau_1})$, the functions in (1.2) can be written as $f_k(R_{t-\tau_1}, Y_{t-\tau_1}) =$

$\tilde{f}_k(\theta_{k1}S_{t-\tau_1}+\theta_{k2}Z_{t-\tau_1}), k=1,2,3$. If we approximate the differential quotients by differences $R_{t+1}-R_t$ and $Y_{t+1}-Y_t$, respectively, we have the statistical model

$$R_{t+1} = (\tilde{f}_1(\theta_{11}S_{t-\tau_1}+\theta_{12}Z_{t-\tau_1})+1)R_t+\varepsilon_{t+1},$$
$$Y_{t+1} = (\tilde{f}_2(\theta_{21}S_{t-\tau_2}+\theta_{22}Z_{t-\tau_2})+1)Y_t+\tilde{f}_3(\theta_{31}S_{t-\tau_3}+\theta_{32}Z_{t-\tau_3})R_t+\epsilon_{t+1}.$$

These are special cases of (1.1). Statistically, the above debate is equivalent to a testing problem: $\theta_{k2}=0$ vs $\theta_{k2}\neq 0$, $k=1,2,3$.

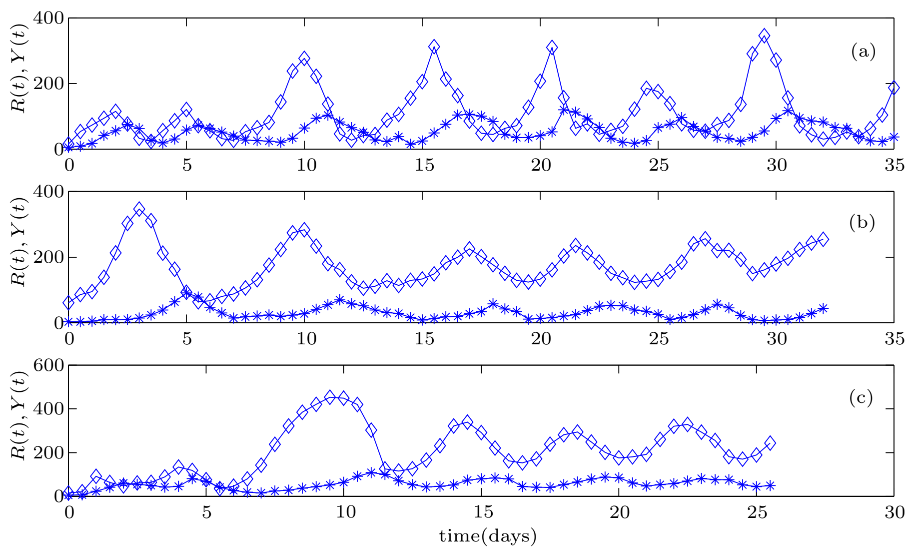

Figure 1. Original predator-prey data sets with different conditions under which they were run. Diamonds are the prey measurements and stars are the predator abundances.

The above discussion motivates us to investigate the estimation of the single-index in (1.1) and therefore the model. Xia and Li (1999) studied the estimation of model (1.1) following the method of Härdle, Hall and Ichimura (1993). The estimation method is very hard to implement. Fan et al. (2003) proposed another estimation method, but the asymptotic properties are unknown. Note that the estimation of model (1.1) is strongly related to the estimation of the single-index model $y=g(\theta_0^TX)+\varepsilon$; see Härdle et al. (1993). For the single-index model there are numerous estimation methods; see, for example, Härdle and Stoker (1989), Li (1991), Härdle et al. (1993), Carroll, Fan, Gijbels and Wand (1997), Hristache, Juditsky and Spokoiny (2001), Xia, Tong, Li and Zhu (2002) and the references therein. However, none of these methods can be used directly here and there are concerns with these methods, which we briefly summarize. (1)

Heavy computational burden: see, for example, Härdle et al. (1993), Carroll et al. (1997) and Xia and Li (1999); these methods entail complicated optimization techniques and no simple algorithm is available to-date. (2) *Strong restrictions on link functions or designs of covariates:* Li (1991) required strong restrictions on the distributions of the covariates; Härdle and Stoker (1989) and Hristache et al. (2001) needed a non-symmetric structure of the link function, i.e., $|Eg'(\theta_0^T X)|$ is away from 0; if these conditions are violated, their methods cannot obtain useful estimators. (3) *Under-smoothing:* Most of the methods mentioned above require under-smoothing the link function in order to achieve root-m consistency for the parameter estimators; see Härdle and Stoker (1989) and Hristache et al. (2001), Hall (1989) and Carroll et al. (1997) among others. More discussion on the selection of bandwidth for the partially linear model can be found in Linton (1995). In this paper we use the newly introduced minimum average variance estimation (MAVE) method (Xia et al. (2002)) to address the above concerns.

2. Estimation

For ease of exposition, rewrite $x_0 \equiv 1$ and, by an abuse of notation, $X = (x_0, \ldots, x_p)^T$. Let $G(\theta^T z) = (g_0(\theta^T z), g_1(\theta^T z), \ldots, g_p(\theta^T z))^T$. If $G(\cdot)$ is known, then the single-index direction θ_0 minimizes

$$E\Big[y - G(\theta^T Z)^T X\Big]^2. \tag{2.1}$$

The conditional variance given $\xi = \theta^T Z$ and θ is $\sigma_\theta^2(\theta^T Z) = E[\{y - G(\theta^T Z)^T X\}^2 \mid \theta^T Z = \xi]$. It follows that $E[y - G(\theta^T Z)^T X]^2 = E\sigma_\theta^2(\theta^T Z)$. Therefore, minimizing (2.1) is equivalent to minimizing, with respect to θ,

$$E\sigma_\theta^2(\theta^T Z) \quad \text{subject to} \quad \theta^T\theta = 1. \tag{2.2}$$

We call the estimation procedure the minimum average (conditional) variance estimation (MAVE) method; see Xia et al. (2002). Because g_k, $k = 0, \ldots, p$, are unknown, we may use a local linear function to approximate them. Let $\{(X_i, Z_i, y_i),\ i = 1, \ldots, n\}$ be a sample from (1.1). For any z, a local linear expansion of $g_k(\theta_0^T Z_i)$ at $\theta_0^T z$ is

$$g_k(\theta_0^T Z_i) = g_k(\theta_0^T z) + g_k'(\theta_0^T z)\theta_0^T Z_{i0} + O_P\{(\theta_0^T Z_{i0})^2\}, \quad k = 0, \ldots, p,$$

where $Z_{i0} = Z_i - z$. Let $G'(\theta_0^T z) = (g_0'(\theta_0^T z), \ldots, g_q'(\theta_0^T z))^T$. For Z_i close to z, we have

$$y_i - X_i^T G(\theta_0^T Z_i) \approx y_i - X_i^T G(\theta_0^T z) - X_i^T G'(\theta_0^T z) Z_{i0}^T \theta_0.$$

Following the idea of Nadaraya-Watson kernel estimation, we estimate $\sigma_\theta^2(\theta^T z)$ by

$$\hat{\sigma}_\theta^2(\theta^T z) = \min_{a,d} \sum_{i=1}^n \left\{ y_i - X_i^T a - X_i^T d Z_{i0}^T \theta \right\}^2 w_{i0}. \tag{2.3}$$

Here, $w_{i0} \geq 0$, $i = 1, \ldots, n$, are some weights, typically centered at z. Note that $\sum_{i=1}^n w_{i0} = 1$ is needed in (2.3). For simplicity, we remove this restriction in the following context. Write $a_j = (a_{j0}, \ldots, a_{jp})^T$ and $d_j = (d_{j0}, \ldots, d_{jp})^T$. By (2.2) and (2.3), our estimation procedure is to minimize

$$n^{-1} \sum_{j=1}^n \mathcal{I}(\bar{w}_j) \sum_{i=1}^n \left\{ y_i - X_i^T a_j - X_i^T d_j Z_{ij}^T \theta \right\}^2 w_{ij} \tag{2.4}$$

with respect to (a_j, d_j), $j = 1, \ldots, n$, and θ, where $Z_{ij} = Z_i - Z_j$, $\bar{w}_j = n^{-1} \sum_{i=1}^n w_{ij}$ and $\mathcal{I}(\cdot)$ is a bounded weight function employed to handle the boundary points of the observations. The trimming function $\mathcal{I}(\cdot)$ is adopted here for technical simplicity; see Härdle et al. (1993) and Powell, Stock and Stoker (1989). In our proofs, we take $\mathcal{I}(v) \geq 0$ to be any function with a bounded third order derivative and $\mathcal{I}(v) = 0$ if $v \leq c_0$, where c_0 is a small constant. Theoretically, c_0 can tend to 0 as $n \to \infty$ at a slow rate, but this will complicate the proof and benefit us with no more than the fixed c_0 in practice. The smoothness of $\mathcal{I}(v)$ is needed for ease of proofs. In practice, we can further take $\mathcal{I}(\cdot) \equiv 1$; or $\mathcal{I}(v) = 1$ if $v \geq c_0$, 0 otherwise. Note that we obtain the solution of θ and a_j simultaneously with just a single cost function, namely (2.4). This is different from existing estimation methods; see, e.g., Carroll et al. (1997) and Härdle et al. (1993).

Minimizing (2.4) is a quadratic problem that is easily solved. A simple algorithm to implement (2.4) is as follows. Let

$$\begin{pmatrix} a_j \\ d_j \end{pmatrix} = \left\{ \sum_{i=1}^n w_{ij} \begin{pmatrix} X_i \\ Z_{ij}^T \theta X_i \end{pmatrix} \begin{pmatrix} X_i \\ Z_{ij}^T \theta X_i \end{pmatrix}^T \right\}^{-1} \sum_{i=1}^n w_{ij} \begin{pmatrix} X_i \\ Z_{ij}^T \theta X_i \end{pmatrix} y_i, \tag{2.5}$$

$$\theta = \left\{ \sum_{j=1}^n \mathcal{I}(\bar{w}_j) \sum_{i=1}^n w_{ij} (X_i^T d_j)^2 Z_{ij} Z_{ij}^T \right\}^{-} \sum_{j=1}^n \mathcal{I}(\bar{w}_j) \sum_{i=1}^n w_{ij} X_i^T d_j Z_{ij} (y_i - X_i^T a_j), \tag{2.6}$$

where $\{\cdot\}^-$ denotes the Moore-Penrose inverse of a matrix. The minimization in (2.4) can be solved by iterating (2.5) and (2.6) until convergence; in each iteration θ is replaced by $\text{sign}_1(\theta)\theta/|\theta|$, where θ is the latest value given by (2.6) and $\text{sign}_1(\theta)$ is the sign of the first element of θ. The final value of $\text{sign}_1(\theta)\theta/|\theta|$ is our estimator of the single-index direction θ_0.

The choice of weight w_{ij} plays an important role for different estimation methods; see Hristache et al. (2001) and Xia et al. (2002). In this paper, we use two sets of weights. Suppose $H(\cdot)$ and $K(\cdot)$ are a q-variate and a univariate density function, respectively. We first use weight $w_{ij} = H_{b,i}(Z_j)$, where $H_{b,i}(z) = b^{-q}H(Z_{i0}/b)$ and b is a bandwidth, a multivariate dimensional kernel weight. Let $\tilde{\theta}$ be the final value of iterating (2.5) and (2.6). Because of the so-called "curse of dimensionality" in nonparametrics, the estimate $\tilde{\theta}$ based on this kind of weight is not efficient. However, $\tilde{\theta}$ is an appropriate initial estimate of θ_0. To refine the estimation, we further use a single-index kernel weight $w_{ij}^{\theta} = K_{h,i}^{\theta}(\theta^T Z_j)$, where $K_{h,i}^{\theta}(v) = h^{-1}K\{(\theta^T Z_i - v)/h\}$, h is the bandwidth and θ is the latest estimate of θ_0. Let $\hat{\theta}$ be the final value of θ in the iterations. We estimate θ_0 by $\hat{\theta}$.

Suppose $\{(X_i, Z_i, y_i),\ i = 1, \ldots, n\}$ is a set of observations. We make the following assumptions on the stochastic nature of the observations, the coefficient functions and the kernel functions. Let $X_{i(\ell)}$ and $Z_{i(\ell)}$ be the ℓth elements of X_i and Z_i, respectively, and take $\xi_i^{(\iota)} = X_{i(\ell_1)}^{k1} X_{i(\ell_2)}^{k2} Z_{i(\ell_3)}^{k3} Z_{i(\ell_4)}^{k4}$ with $\iota = k_1 + k_2 + k_3 + k_4$.

(C1) $\{(X_i, Z_i, y_i)\}$ is a strictly stationary (with the same marginal distribution as (X, Z, y)) and $\alpha-$mixing sequence with a geometrically decaying mixing rate $\alpha(k)$.

(C2) With probability 1, Z is distributed in a compact region $\mathcal{D}$; the density functions f of Z and f_θ of $\theta^T Z$ have bounded continuous derivatives and f_θ is Lipschitz continuous in $\theta \in \Theta$.

(C3) g_k, $k = 0, \ldots, p$, has a bounded, continuous third order derivative; for all $\iota \le 2r$ with some $r > 2$; the conditional expectations $E(\xi^{(\iota)}|Z = z)$ and $E(\xi^{(\iota)}|\theta^T Z = v)$ have bounded continuous derivatives and the latter is Lipschitz continuous in $\theta \in \Theta$; $E(|\xi_\ell^{(\iota)}||\xi_1^{(\iota)}| \mid Z_1 = z_1, Z_\ell = z_\ell)$ is bounded by a constant for all $\ell > 0$, z_1, z_ℓ and x_1.

(C4) $\sup_{x,z} E(\varepsilon^2|X = x, Z = z) < \infty$, $E\varepsilon^r < \infty$ and $E\{\varepsilon_i|(X_j, Z_j), j \le i\} = 0$ almost surely, where r is the same as in (C3).

(C5) $E(XX^T|Z)$ is positive definite; $P(G'^T(\theta_0^T Z)X = 0) = 0$.

(C6) H and K are symmetric density functions with compact supports $\{z : |z| \le a_0'\}$ and $\{v : |v| \le a_0\}$, respectively, for some $a_0, a_0' > 0$. The Fourier transform of K is absolutely integrable.

The mixing rate in (C1) can be relaxed to be algebraic, i.e., $\alpha(k) = O(k^{-\rho})$. Suppose the bandwidth $h \sim n^{-\delta}$. Then the mixing rate satisfying the following equation is sufficient.

$$\sum_{n=1}^{\infty} n^{-\{\frac{1}{2}-\frac{1}{r}-\delta(\frac{1}{2}+\frac{1}{r})\}\rho+2q+1+\frac{1}{r}+(\frac{1}{2}+\frac{1}{r})\delta}(\log n)^{\frac{\varrho}{2}} < \infty. \tag{2.7}$$

The first part of (C2) is a common assumption on density functions of kernel smoothers when uniform convergence rate is needed. See, e.g., Linton (1995). Our results can be extended to the case that Z is not bounded provided high order moments of Z exist. The Lipschitz condition on the density function can be fulfilled under some mild conditions on the density function f, see Hall (1989). The third order derivative in (C3) is needed for higher order expansion. Actually, existence of a second order derivative is sufficient for root-m consistency if we confine the bandwidth to a smaller range. The restriction on the expectation conditioned on cross-product terms over time is needed for the consistency of estimators when the observations are dependent. If $E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} \neq 0$ in (C4) then our asymptotic results still hold, but the distribution will have a more complicated variance matrix depending on the structure of the stochastic process of the observations. Assumption (C5) is imposed to ensure that the proposed algorithm has an attractor with a single direction. As discussed in Fan et al. (2003), there are identifiability problems if $X \equiv Z$. We can assume that the $g_k(\cdot)$ are not all linear when $X \equiv Z$ for the identification of the single-index, but we need some further constraints for the identification of the coefficient functions. For example, we can confine the conditional mean functions to $g_1(\theta_0^T X)x_1 + \cdots + g_p(\theta_0^T X)x_p$ or $g_0(\theta_0^T X) + g_1(\theta_0^T X)x_1 + \cdots + g_{p-1}(\theta_0^T X)x_{p-1}$ if $\theta_{0p} \neq 0$.

In this paper, we only employ kernel functions with compact support as in (C6). We further assume that $\kappa_2 \triangleq \int K(u)u^2 du = 1$ and $\mathcal{H}_2 \triangleq \int H(z)zz^T dz = I_{q\times q}$; otherwise we take $K(u) =: K(u/\sqrt{\kappa_2})/\sqrt{\kappa_2}$ and $H(z) =: H(\mathcal{H}_2^{-1/2}z)$ $(\det(\mathcal{H}_2))^{-1/2}$.

Lemma 1. *Suppose that* (C1)$-$(C6) *hold and* $\{z : f(z) \geq c_0\}$ *is non-empty,* $b \to 0$ *and* $nb^{q+2}/\log n \to \infty$*. Let* $\tilde{\theta}$ *be the estimator based on the multi-kernel weight. If we start the iteration with* θ *such that* $\theta^T\theta_0 \neq 0$*, then* $\tilde{\theta} - \theta_0 = o_P(1)$.

Let $\mu_\theta(z) = E(Z|\theta^T Z = \theta^T z)$, $\pi_\theta(z) = E(XX^T|\theta^T Z = \theta^T z)$, $V_\theta(z) = E[\{X^T\ G'(\theta_0^T z)\}X\ Z_{i0}^T|\theta^T Z = \theta^T z]$,

$$U_0 = E[\mathcal{I}_f^{\theta_0}(Z)\{G'(\theta_0^T Z)X\}^2 E\{(Z - \mu_{\theta_0}(Z))(Z - \mu_{\theta_0}(Z))^T|\theta_0^T Z\}],$$
$$W_k = E\left[\mathcal{I}_f^{\theta_0}(Z)\{G'(\theta_0^T Z)X\}^2\{Z - \mu_{\theta_0}(Z)\}\{Z - \mu_{\theta_0}(Z)\}^T\varepsilon^k\right], \quad k = 0, 2,$$

$\mathcal{I}_f^{\theta_0}(z) = \mathcal{I}(f_{\theta_0}({\theta_0}^T z))f_{\theta_0}({\theta_0}^T z)$ and $W_1 = W_0 + U_0 - E[\mathcal{I}_f^{\theta_0}(Z)V_{\theta_0}^T(Z)\{\pi_{\theta_0}(Z)\}^{-1}$ $V_{\theta_0}(Z)]$.

Theorem 1. *Suppose that* (C1)$-$(C6) *hold and* $\{z, f_\theta(\theta^T z) \geq c_0\}$ *is non-empty for all* $\theta \in \Theta$*,* $h \sim n^{-\delta}$ *with* $1/6 < \delta < 1/4$*. If we start the estimation procedure with single-index kernel weight and* $\theta = \tilde{\theta}$*, then* $n^{1/2}\{\hat{\theta} - \theta_0\} \xrightarrow{D} N(0, W_1^- W_2 W_1^-)$.

Theorem 1 still holds if we start with any consistent estimate θ. The proof of Theorem 1 is given in Section 4. The convergence of the algorithm is also implied in the proof. For statistical inference, we further give an estimator for the variance and covariance matrix in the asymptotic distribution, as follows. Take $\hat{f}_j = n^{-1}\sum_{i=1}^n K_h(\hat{\theta}^\top Z_{ij})$, $\hat{\mu}_j = (n\hat{f}_j)^{-1}\sum_{i=1}^n K_h(\hat{\theta}^\top Z_{ij})X_i$ and

$$\hat{\pi}(Z_j) = (n\hat{f}_j)^{-1}\sum_{i=1}^n K_h(\hat{\theta}^\top Z_{ij})X_iX_i^\top, \hat{V}_j = (n\hat{f}_j)^{-1}\sum_{i=1}^n K_h(\hat{\theta}^\top Z_{ij})X_i^T d_j X_i Z_{ij}^T,$$

$$\hat{W}_1 = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{I}(\hat{f}_j)\{d_j^\top X_i\}^2 K_h(\hat{\theta}^\top Z_{ij})Z_{ij}Z_{ij}^\top - n^{-1}\sum_{j=1}^n \mathcal{I}(\hat{f}_j)\hat{f}_j\hat{V}_j^T\hat{\pi}_j^{-1}\hat{V}_j,$$

$$\hat{W}_2 = \sum_{j=1}^n \mathcal{I}(\hat{f}_j)\hat{f}_j(d_j^T X_j)^2\{Z_j - \hat{\mu}_j\}\{Z_j - \hat{\mu}_j\}^T(y_j - a_j)^2.$$

Remark 1. In Xia and Li (1999), their estimator has the same distribution but with variance matrix $W_0^- W_2 W_0^-$. By Schwarz's Inequality, we have that $W_1 - W_0$ is a semi-positive definite matrix. Hence, $W_0^- W_2 W_0^- - W_1^- W_2 W_1^-$ is a semi-positive definite matrix and the proposed estimation method in this paper is more efficient than that in Xia and Li (1999) for (1.1).

Remark 2. Note that the bandwidth with rate $n^{-1/5}$ satisfies the requirement. This property confirms that many existing bandwidth selection methods can be employed here.

Remark 3. In Theorem 1, a consistent initial estimator $\tilde{\theta}$ based on the multi-dimension kernel is used. However, when the dimension of Z is high, we have the risk of suffering from a poor initial estimator $\tilde{\theta}$. To reduce this risk, we use the idea of elliptical kernels as proposed by Hristache et al. (2001) by taking $w_{ij} = K_h(|(\theta\theta^T + 2^{-k}I)Z_{ij}|)$ in step k of the iterations. Given a set of weights w_{ij} (or w_{ij}^θ), we need several iterations between (2.5) and (2.6) to obtain a better approximation of the solution of (2.4). Therefore, for the single-index kernel weights, we suggest fixing θ in weight w_{ij}^θ for several iterations before replacing it by the latest value of θ.

Remark 4. In the proof of the theorem, we further show that the algorithm has a very fast convergence rate. Let $\hat{\theta}_k$ be the value of θ after k'th iteration, see (2.5) and (2.6). Then we have $|\hat{\theta}_k - \theta_0| \le \Delta_k|\hat{\theta}_{k-1} - \theta_0|$, where $\max_k \Delta_k < 1$ as n is large enough. In other words, the algorithm has a geometric convergence rate.

After obtaining the estimate of θ_0, we can further estimate the coefficient functions with θ_0 replaced by $\hat{\theta}$. Because $\hat{\theta}$ is root-m consistent, we immediately have the following result; see Xia and Li (1999) and Cai, Fan and Yao (2000).

Corollary 1. *Suppose the assumptions of Theorem* 1 *hold and that the density function* f_{θ_0} *of* $\theta_0^T Z$ *is positive at* v *and the derivative of* $E(XX^T\varepsilon^2|\theta_0^T Z = v)$ *exists. Then*

$$(nh)^{\frac{1}{2}}\{\hat{G}(v)-G(v)-\frac{1}{2}G''(v)h^2\}\xrightarrow{D} N(0, f_{\theta_0}^{-1}(v)\Sigma_0^{-1}(v)\Sigma_2(v)\Sigma_0^{-1}(v)\int K^2(u)du),$$

where $\Sigma_k(v) = E(XX^T|\varepsilon|^k \mid \theta_0^T Z = v), k = 0, 2.$

3. Simulation Study

In this section, we use simulations to demonstrate the performance of our method for finite data sets. Some practical problems are addressed and some observations are made. Bandwidth selection is always an important practical issue for nonparametric kernel smoothing. Note that the optimal bandwidth for the estimation of the regression function, in the sense of minimizing the mean integrated squared error, can be used in our procedure. There are many methods available to estimate the optimal bandwidth. In our calculations, we use the cross-validation bandwidth selection method as follows. Corresponding to (2.5), calculate

$$\begin{pmatrix} a_{h,j} \\ d_{h,j} \end{pmatrix} = \Big\{\sum_{\substack{i=1\\ i\neq j}}^{n} K_{h,i}^{\theta}(\theta^T Z_j)\begin{pmatrix} X_i \\ Z_{ij}^T\theta X_i\end{pmatrix}\begin{pmatrix} X_i \\ Z_{ij}^T\theta X_i\end{pmatrix}^T\Big\}^{-1}\sum_{\substack{i=1\\ i\neq j}}^{n} K_{h,i}^{\theta}(\theta^T Z_j)\begin{pmatrix} X_i \\ Z_{ij}^T\theta X_i\end{pmatrix} y_i.$$

We take c_0 to be very small, such that all points are assigned to have weight $\mathcal{I}(\bar{w}_j) = 1$ in (2.4). When $\theta = \theta_0$, $a_{h,j}$ is actually a kernel estimate of $G(\theta_0^T Z_j)$ with the observation (X_j, Z_j, y_j) deleted. Our bandwidth for each iteration is chosen to be

$$h_\theta = \arg\inf_h \sum_{j=1}^{n} \mathcal{I}(\bar{w}_j)\{y_j - a_{h,j}^T X_j\}^2.$$

When $|\theta - \theta_0| = O_P(n^{-1/2})$, it can be shown that $h_\theta \sim n^{-1/5}$ under some mild conditions. In the calculations, the stopping rule is that $|\theta_k^T\theta_{k+1}|$ do not change for several consecutive iterations (3, in our calculations), where θ_k is the value of kth iteration.

Example 3.1. Consider the simulated model from Fan et al. (2003):

$$y_i = 3\exp\{-(\theta_0^T Z_i)^2\} + 0.8\{\theta_0^T Z_i\}x_{i1} + 1.5\sin(\pi\theta_0^T Z_i)x_{i3} + \sigma\varepsilon_i, \qquad (3.1)$$

where $X_i = Z_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^T$, $i = 1, \ldots, n$, are independent random vectors uniformly distributed on $[-1, 1]^{\otimes 4}$, $\{\varepsilon_i\}$ is a sequence of independent standard normal random variables, and $\theta_0 = (1/3, 2/3, 0, 2/3)^T$. Besides estimating the model, we also consider the hypotheses testing, at significant level $\alpha = 0.05$, of

$$H_{10} : \theta_{01} = 0 \quad \text{v.s.} \quad H_{11} : \theta_{01} \neq 0,$$
$$H_{30} : \theta_{03} = 0 \quad \text{v.s.} \quad H_{31} : \theta_{03} \neq 0,$$

based on the asymptotic distributions. We use $|\hat{\theta}^T\theta_0|$ to measure the estimation accuracy of $\hat{\theta}$. We take initial value $\theta = (1, 0, 0, 0)^T$ in all the calculations. With sample size 50, 100, 200 and 400 and noise magnitude $\sigma = 0.5, 1$ and 2, our simulation results of 200 replications for every combination of sample size and noise magnitude are shown in Figure 2. Some statistics are also listed in Table 1. With reasonable signal-noise ratio, the proposed method can estimate θ_0 quite well. Compared with Fan et al., (2003, Figure 3b), the distributions of the values in Figure 2 are much closer to 1 than theirs, suggesting better performance by our method for this model. We found that more extensive overlapping of Z_i and X_i worsen the estimation. If we take $X_i = (x_{i1}, x_{i3})^T$, the estimation results will improve substantially.

Table 1. Mean and standard deviation (in parentheses) of $|\hat{\theta}^\top\theta_0|$ and the rejection rates of H_{10} [in square brackets] and H_{30} {in braces}.

σ	$n = 50$	$n = 100$	$n = 200$	$n = 400$
0.5	0.8280 (0.1809)	0.9240 (0.1134)	0.9760 (0.0633)	0.9978 (0.0155)
	[0.775] {0.285}	[0.920] {0.115}	[0.995] {0.055}	[1.000] {0.035}
1.0	0.7380 (0.2191)	0.8706 (0.1464)	0.9297 (0.0996)	0.9850 (0.0474)
	[0.675] {0.440}	[0.670] {0.335}	[0.815] {0.070}	[0.980] {0.040}
2.0	0.5385 (0.2766)	0.7009 (0.2395)	0.8034 (0.1800)	0.8985 (0.1106)
	[0.500] {0.495}	[0.560] {0.385}	[0.635] {0.220}	[0.880] {0.115}

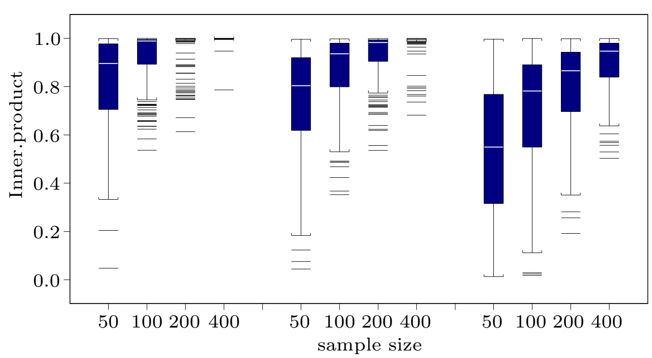

Figure 2. Simulation results for Example 3.1. The three sets of boxplots of the absolute inner products $\hat{\theta}^T\theta_0$ for models (3.1) for σ =0.5, 1, 2 with sample size n = 50, 100, 200 and 400 for each σ, respectively.

Example 3.2. We consider the SETAR time series model

$$y_t = (\Phi(-vz_t) - 0.5)y_{t-1} + (\Phi(2vz_t) - 0.6)y_{t-2} + \varepsilon_t, \tag{3.2}$$

where $z_t = y_{t-1} + y_{t-2} - y_{t-3} - y_{t-4}$, and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. (To ensure that the conditions in Theorem 1 are satisfied, we may further truncate to $\varepsilon_t =: \varepsilon_t I_{|\varepsilon_t| \le 4}$; this truncation actually does not affect the sampling for finite samples). The parameter v is employed here to control the difference between the TAR model and the SETAR model; see the first panel of Figure 3. Here, $X_t = (y_{t-1}, y_{t-2})^T$, $Z_t = (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})^T$, and $\theta_0 = (1, 1, -1, -1)^T/2$. We take initial value $\theta = (1, 2, 0, 0)^T/\sqrt{5}$ in the calculations. With sample size 50, 100, 200 and 400, our simulation results based on 200 replications for each combination of sample size and v are shown in Figure 3. Some statistics are listed in Table 2. Because θ_0 is a global parameter, it can be estimated well even when some of the coefficient functions are estimated poorly. Similar to the results under the parametric setting, the estimation accuracy tends to increase as the coefficient function becomes steeper; see Chen (1995) for more details under parametric settings.

Table 2. Mean and mean squared deviation (in parentheses) of the inner products of the estimates for model (3.2).

v	$n = 50$	$n = 100$	$n = 200$	$n = 400$
0.5	0.8058(0.2091)	0.9266(0.1120)	0.9770(0.0278)	0.9922(0.0079)
1.0	0.8984(0.1439)	0.9626(0.0719)	0.9869(0.0152)	0.9955(0.0044)
5.0	0.8864(0.1470)	0.9720(0.0279)	0.9894(0.0104)	0.9953(0.0051)

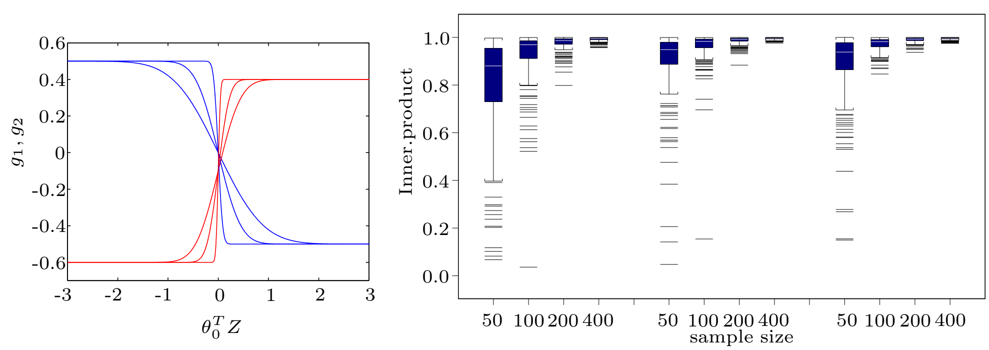

Figure 3. Simulation results for Example 3.2. The left panel are the coefficient functions in model (3.2); the decreasing lines are g_1 and the increasing lines are g_2. From flat to steep, the lines correspond to coefficient functions with $v = 0.5, 1$ and 5 respectively. In the right panel, there are three sets of boxplots of $|\hat{\theta}^T \theta_0|$ for models (3.2) for v = 0.5, 1 and 5 , respectively, and sample size n = 50, 100, 200, 400 for each v.

4. Data Analysis

In this section, we return to our motivating problems with two data sets. For the first one, we use our estimation method to search for a threshold variable and build a TAR model. For the second data set, we answer a question in ecology.

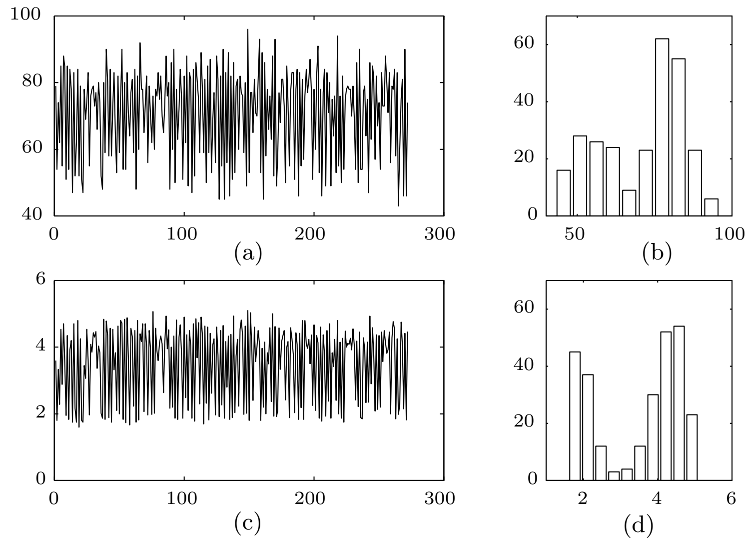

Figure 4. Data set of the Old Faithful Geyser. (a): the waiting time between the eruptions. (b): the histogram of the waiting time. (c): the duration of eruptions. (d): the histogram of the duration of eruptions.

Example 4.1.(The Old Faithful Geyser data set). There are two series in the data set: duration of eruption (x_t, in minutes) and waiting time (y_t, in minutes). They are shown in Figures 4(a) and (b), which also show the histograms. Here our primary focus is the series y_t. Note that the histogram shows two modes, suggesting the possibility of a mixture of distributions, perhaps due to a hidden threshold variable. Is it possible to find a reasonable proxy of the hidden variable? To this end, we use the following *single-index coefficient regression model* after standardization:

$$y_t = g_0(\theta^T Z_t) + \sum_{i=1}^{5} g_i(\theta^T Z_t) y_{t-i} + \varepsilon_t,$$

where $Z_t = (x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5})^T$. Using our estimation procedure, we estimate θ as

$$\hat{\theta} = \underset{(0.085)\ (0.082)\ (0.068)\ (0.052)\ (0.046)}{(0.6328,\ 0.6785,\ 0.3622,\ 0.0490,\ 0.0744)^T}.$$

Where the values in the parentheses are the corresponding standard errors of the estimates. The residual sum of squares is 0.5905. Note that the last two elements are quite small (and their t-values are less than 2). To simplify, we now take $Z_t = (x_{t-1}, x_{t-2}, x_{t-3})^T$ and consider

$$y_t = g_0(\theta_0^T Z_t) + g_1(\theta_0^T Z_t)y_{t-1} + g_2(\theta_0^T Z_t)y_{t-2} + g_3(\theta_0^T Z_t)y_{t-3} + g_4(\theta_0^T Z_t)y_{t-4} + g_5(\theta_0^T Z_t)y_{t-5} + \varepsilon_t. \quad (4.1)$$

We estimate θ_0 as $\hat{\theta} = (0.6355,\ 0.6758,\ 0.3732)^T$ (with corresponding standard errors of 0.0899, 0.0885 and 0.0782, respectively). The residual sum of squares is 0.6140. The coefficient functions are shown in Figure 5. It seems reasonable to approximate most of them by step functions with a common jump at about 0.0. This lends some support to the plausibility of a hidden threshold variable, a proxy for which might be $\hat{\theta}^T Z_t$, or $z_t = 0.6355x_{t-1} + 0.6758x_{t-2} + 0.3732x_{t-3}$. We can further build the following tentative threshold model for the waiting time y_t:

$$y_t = \begin{cases} \underset{(0.097)}{0.195} - \underset{(0.104)}{0.737}y_{t-1} - \underset{(0.127)}{0.174}y_{t-2} + \underset{(0.104)}{0.126}y_{t-3} - \underset{(0.082)}{0.203}y_{t-5} + \varepsilon_{1t}, & \text{if } z_t \geq -0.07; \\ \\ \underset{(0.007)}{-0.040} - \underset{(0.071)}{0.424}y_{t-1} - \underset{(0.084)}{0.245}y_{t-3} - \underset{(0.079)}{0.264}y_{t-4} + \varepsilon_{2t}, & \text{if } z_t < -0.07, \end{cases}$$

with Var $(\varepsilon_{1t}) = 0.6557$ and Var $(\varepsilon_{2t}) = 0.6354$, and pooled variance 0.6450. Note that the variance of ε_{1t} and ε_{2t} are about the same and we may pool them to

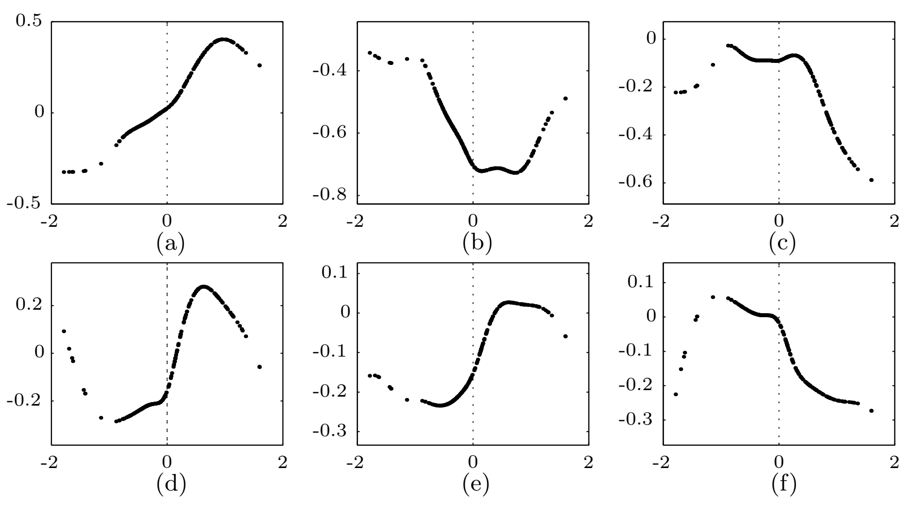

Figure 5. Calculation results for the Old Faithful geyser data in Example 4.1. (a)−(f) are the estimated coefficient functions in model (4.1).

form ε_t. We conduct a white noise test for the series using Bartlett's Kolmogorov-Smirnov statistic. See, e.g., Fuller (1976). The test statistics for $\{y_t\}$ and $\{\varepsilon_t\}$ are 0.3691 and 0.0415 respectively. At the significance level $\alpha = 0.05$, for which the critical value is 0.1174, $\{y_t\}$ is rejected as a white noise sequence but $\{\varepsilon_t\}$ is not rejected as such. The residual autocorrelations at lag $k = 1, \ldots, 6$ are $r_1 = 0.0213$, $r_2 = -0.0208$, $r_3 = 0.0039$, $r_4 = 0.0115$, $r_5 = 0.0288$ and $r_6 = 0.0324$. The corresponding standard errors for $r_1, \ldots, r_6$ are $0.0518, 0.0542, 0.0498, 0.0576, 0.0572$, and 0.0605, respectively. See Li (1992). These values also suggest that $\{\varepsilon_t\}$ may be a white noise process.

The previous analysis suggests that the threshold AR model is acceptable, as constructed, from a statistical point of view. Note that the estimated threshold variable is $z_t = 0.6328x_{t-1} + 0.6785x_{t-2} + 0.3622x_{t-3}$. The "upper regime" of the threshold AR model we have constructed corresponds to the longer waiting time, and the "lower regime" the shorter waiting time. Our threshold variable indicates that longer eruption durations will result in longer waiting time.

Example 4.2.(The protozoan predator-prey system). Now we join the debate in ecology using our proposed method. The lags are selected to be $t-1$, i.e., $\tau_1 = \tau_2 = 1$, according to some ecological background of the problem; see Jost and Ellner (2000). We further simplify the model to

$$R_{t+1} = g_1(\theta_1^T W_t) R_t + \varepsilon_t, \qquad Y_{t+1} = g_2(\theta_2^T W_t) Y_t + g_3(\theta_2^T W_t) R_t + \epsilon_t,$$

where $W_t = (\log(R_{t-1}),\ \log(Y_{t-1}))^T$. The estimated parameters are listed in Table 3. The estimates of the functional responses, i.e., g_1, g_2 and g_3 are shown in Figure 6.

Note that the signs of θ_{11} are positive and those of θ_{12} are negative for all the data sets in Table 3. Thus, the functions g_1 can be written as $\tilde{g}_1(R_{t-1}^b / Y_{t-1}^a)$ where $a, b > 0$ and the $\tilde{g}_1(\cdot)$'s are increasing functions for all the data sets; see Figures 6(a), 6(d) and 6(g). For example, $a = 0.7948$, $b = 0.6068$ and $\tilde{g}_1(v) = g_1(\log(v))$ for the first data set. This suggests that the prey (food for the predator) has a positive effect on the number of predators; the predators at the previous time point has negative effect on the current number of predator because of the limited food supply (i.e., the prey). Our results suggest that the dynamics of predator is typically ratio-dependent. Note that the signs of θ_{21} and θ_{22} are positive and that the functions g_2 and g_3 are decreasing functions (except for the estimate in Figure 6(f)) for all the data sets. This suggests that both the prey population and the predator population at the previous time point have a negative effect on the dynamics of the prey. A possible reason for this is that food competition among prey population and predation by predators affect the prey population. Thus, our statistical analysis suggests that the dynamics of prey is typically both prey and predator dependent.

Table 3. Estimates of the single-index (and the standard error) for different data sets in Example 4.2.

Data set	θ_{11}	θ_{12}	θ_{21}	θ_{22}
set 1	0.6068(0.1622)	-0.7948(0.2174)	0.9616(0.1563)	0.2745(0.0459)
set 2	0.1842(0.0645)	-0.9829(0.1746)	0.4230(0.0642)	0.9061(0.1393)
set 3	0.8411(0.1337)	-0.5409(0.0867)	0.4783(0.0679)	0.8782(0.0773)

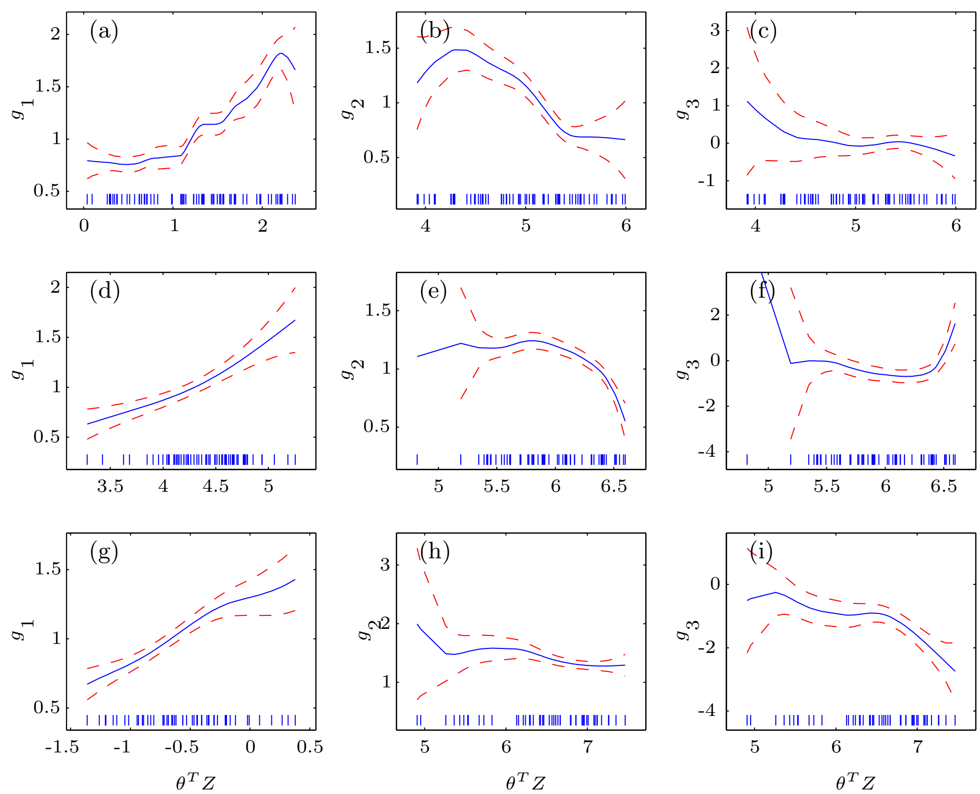

Figure 6. The estimation results for Example 4.2. (a)−(c) correspond to the first data set; (d)−(f) correspond to the second data set; (g)-(i) correspond to the third data set. The central lines in (a), (d) and (g) are the estimated g_1 for the three corresponding data sets. The central lines in (b), (e) and (h) are the estimated g_2 for the three corresponding data sets. The central lines in (c), (f) and (i) are the estimated g_3 for the three corresponding data sets. The upper and lower dashed lines are the corresponding 95% symmetric pointwise confidence intervals. The distribution of the single-indexes $\theta_1^T Z_t$ and $\theta_2^T Z_t$ are shown at the bottom of the panels.

5. Proofs

We give only an outline of the proof of Theorem 1. A complete proof, and those for the lemmas 1 and A.1-A.4 are available at http://stat.sinica.edu.tw/statistica/. A computer code sifc.m in Matlab is also available at http://www.stat.nus.edu.sg/˜staxyc. The idea of the proof can be stated as follows. Based on Lemmas A.1−A.3, we obtain uniform consistency rates for the local linear estimators of the coefficient functions; see (A.40). Based on the expansions and (2.6), we then build a recursive formula for the iteration in the algorithm, i.e., $\theta_{k+1} - \theta_0 = \Gamma_k(\theta_k - \theta_0) +$ Smaller terms, where θ_k is the estimator of θ_0 after the kth iteration and $\max_k |\Gamma_k| < 1$. See (A.43) for more details. This recursive formula indicates that the true direction θ_0 is the attractor of the algorithm. The formula is finally used to prove the convergence of the algorithm as well as the consistency and asymptotic normality of the estimator.

Let $\delta_\theta = |\theta - \theta_0|$. In Θ, δ_θ is bounded. Let $\delta_{qn} = \{\log n/(nb^q)\}^{1/2}$, $\tau_{qn} = b^2 + \delta_{qn}$, $\delta_n = \{\log n/(nh)\}^{1/2}$, $\tau_n = h^2 + \delta_n$ and $\delta_{0n} = (\log n/n)^{1/2}$. By the condition $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$, we have $\delta_{0n} \ll h^2 \ll h^{-1}\delta_n$ and $\delta_n \ll h$. We use these relations frequently in our calculations. Suppose A_n is a matrix. $A_n = O(a_n)$ means every element in A_n is $O(a_n)$ almost surely. We adopt consistency in the sense of "almost surely" because we need to prove the convergence of the algorithm, which theoretically needs infinite iteration. Let $c, c_1, c_2, \cdots$ be a set of constants. For ease of exposition, c may have different values at different places. We write $K_h(\theta^T Z_{i0}) = h^{-1}K(\theta^T(Z_i - z)/h)$ and $H_b(Z_{i0}) = h^{-q}H\{(Z_i - z)/h\}$ as $K^\theta_{h,i}(z)$ (or $K^\theta_{h,i}$) and $H_{b,i}(z)$ (or $H_{b,i}$) respectively in the following context, for simplicity.

Lemma A.1. *Suppose $\varphi(\theta)$ is a measurable function of (X, Z, y), such that $\sup_{\theta,\vartheta\in\Theta} |\varphi(\theta) - \varphi(\vartheta)| < M(X, Z, y)|\theta - \vartheta|$ a.s. with $EM^r(X, Z, y) < c$; $\sup_{\theta\in\Theta,v} E(|\varphi(\theta)|^r \mid \theta^T Z = v) < c$ for some $r \geq 3$. Let $\varphi_i(\theta)$ be the corresponding value of $\varphi(\theta)$ at (X_i, Z_i, y_i). Assume that $\sup_{\theta\in\Theta,u,v} E(|\varphi_i(\theta)\varphi_1(\theta)| \mid \theta^T Z_1 = u, \theta^T Z_i = v) < c$ for all $i > 1$. Let $g(v)$ be any function with continuous second order derivative, $m(u, v) = g(u) - g(v) - g'(v)(u - v) - g''(v)(u - v)^2/2$ and $\zeta_i^{k,\ell} = m(\theta_0^T Z_i, \theta_0^T z)\mathbf{x}_i^k(\theta^T Z_{i0})^\ell$, where $\mathbf{x}_i$ is any component of X_i, $k = 0, 1$ and $\ell = 0, 1$. If* (C1) *holds, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n \varphi_i(\theta) - E\varphi_i(\theta)\Big| = O(\delta_{0n}),$$

$$\sup_{|\theta-\theta_0|<a_n}\Big|\frac{1}{n}\sum_{i=1}^n \{\varphi_i(\theta) - \varphi_i(\theta_0)\} - E\{\varphi_i(\theta) - \varphi_i(\theta_0)\}\Big| = O(a_n\delta_{0n}),$$

where $a_n \to 0$ *as* $n \to \infty$. *If further* (C2) *and* (C6) *hold,* $h \sim n^{-\delta}$ *with* $0 < \delta < 1 - 2/r$, *then*

$$\sup_{\substack{\theta\in\Theta\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{H_{b,i}\varphi_i(\theta) - E(H_{b,i}\varphi_i(\theta))\}\Big| = O(\delta_{qn}),$$
$$\sup_{\substack{\theta\in\Theta\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{K^{\theta}_{h,i}\varphi_i(\theta) - E(K^{\theta}_{h,i}\varphi_i(\theta))\}\Big| = O(\delta_n),$$
$$\sup_{\substack{|\theta-\theta_0|<a_n\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{K^{\theta}_{h,i}\zeta_i^{k,\ell} - E(K^{\theta}_{h,i}\zeta_i^{k,\ell})\}\Big| = O\{\delta_n h^{\ell}(a_n^2 + h^2)\}.$$

For any measurable function $A(\xi,\eta)$, let $E_k A(\xi_i,\eta_k) = E\{A(v,\eta_k)\}|_{v=\xi_i}$.

Lemma A.2. *Let* $\xi(\theta)$ *be a measurable function of* (X,Z,y). *Suppose* $E\{\xi(\theta) \mid \theta^T Z\} = 0$ *for all* $\theta \in \Theta$ *and* $|\xi(\theta) - \xi(\vartheta)| \le |\theta - \vartheta|\tilde{\xi}$ *with* $E\tilde{\xi}^r < \infty$ *for some* $r > 2$. *Let* φ_i *be defined as in Lemma A.1. If (C1) and (C6) hold, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big\{K^{\theta}_{h,i}(Z_j)\varphi_j(\theta) - E_j(K^{\theta}_{h,i}(Z_j)\varphi_j(\theta))\Big\}\xi_i(\theta)\Big| = O(\delta_n^2).$$

Let $d(z,\mathcal{D}^c) = \min_{z'\in\mathbb{R}^q-\mathcal{D}}|z-z'|$, $J_0(z)$ and $J_\theta(v)$ be any bounded functions such that $J_0(z) = 0$ if $d(z,\mathbb{R}^q-\mathcal{D}) > b$ and $J_\theta(\theta^T z) = 0$ if $d(\theta^T z, \theta^T(\mathbb{R}^q-\mathcal{D})) > h$. By definition, we have

$$\frac{1}{n}\sum_{j=1}^{n} J_0(Z_j) = O(b), \qquad \frac{1}{n}\sum_{j=1}^{n} J_\theta(Z_j) = O(h). \tag{5.1}$$

Let $r(v_1,v_2,x) = G^T(v_1)x - G^T(v_2)x - \{G'^T(v_2)x\}(v_1 - v_2) - \{G''^T(v_2)x\}(v_1 - v_2)^2/2$. To cope with the boundary points, we give the following nonuniform rates of convergence.

Lemma A.3. *Suppose assumptions* (C2), (C3) *and* (C6) *hold. Then*

$$EH_{b,i}\{\frac{\theta^T Z_{i0}}{b}\}^k\{\vartheta^T Z_{i0}/b\}^{\ell} = v^{\theta,\vartheta}_{k,\ell} f(z) + J_0(z) + O(b),$$
$$EK^{\theta}_{h,i}\{\frac{\theta^T Z_{i0}}{h}\}^{\ell} = \tau_\ell f_\theta(\theta^T z) + J_\theta(z) + O(h),$$
$$EK^{\theta}_{h,i}\{\theta^T Z_{i0}\} r(\theta_0^T Z_i, \theta_0^T z, X_i) = O\{h(h + J_\theta(z))(\delta_\theta^2 + h^2)\},$$

uniformly for $\theta,\vartheta \in \Theta$ *with* $\theta \perp \vartheta$ *and* $z \in \mathcal{D}$, *where* $v^{\theta,\vartheta}_{k,\ell} = \int_{\mathbb{R}^q} H(U)(\theta^T U)^k(\vartheta^T U)^{\ell}\, dU$ *and* $\tau_\ell = \int K(u)u^{\ell}du$.

Lemma A.4. *Under assumptions* (C2) *and* (C5), *we have that* W_0 *is a semi-positive matrix with rank* $q-1$.

For ease of exposition, we abbreviate $\sup_{z\in\mathcal{D},\theta\in\Theta}|A_n(z,\theta)| = O(a_n)$ as $A_n(z,\theta) = O(a_n)$ in the following context.

Proof of Theorem 1. By Taylor expansion, write

$$y_i = \left(G^T(\theta_0^T z), G'^T(\theta_0^T z)\right)\begin{pmatrix} X_i \\ \theta^T Z_{i0} X_i \end{pmatrix} + R(Z_i, X_i, z, \theta) + \varepsilon_i,$$

where $R(Z_i, X_i, z, \theta) = G'^T(\theta_0^T z) X_i Z_{i0}^T(\theta_0 - \theta) + G''^T(\theta_0^T Z_i^*) X_i \{\theta_0^T Z_{i0}\}^2/2$. Note that this expansion is unique under the assumptions even $X \equiv Z$ with the assumption before Lemma 1. Let (a^T, d^T) be the value on the right hand side of (2.5), with Z_j replaced by z and

$$C_n(z) = n^{-1}\sum_{i=1}^n H_{b,i}\begin{pmatrix} X_i \\ Z_{i0}^T\theta X_i \end{pmatrix}\begin{pmatrix} X_i \\ Z_{i0}^T\theta X_i \end{pmatrix}^T. \tag{5.2}$$

We have

$$\begin{pmatrix} a \\ d \end{pmatrix} = \begin{pmatrix} G(\theta_0^T z) \\ G'(\theta_0^T z) \end{pmatrix} + C_n^{-1}(z) n^{-1}\sum_{i=1}^n H_{b,i}\begin{pmatrix} X_i \\ Z_{i0}^T\theta X_i \end{pmatrix}\{R(Z_i, X_i, z, \theta) + \varepsilon_i\}. \tag{5.3}$$

Let

$$\begin{aligned} R(X_i, Z_i, z, \theta) = {} & G'^T(\theta_0^T z) X_i Z_{i0}^T(\theta_0 - \theta) + \frac{1}{2} G''^T(\theta_0^T z) X_i \{\theta_0^T Z_{i0}\}^2 \\ & + r(\theta_0^T Z_i, \theta_0^T z, X_i). \end{aligned}$$

Write

$$y_i = \left(G^T(\theta_0^T z), G'^T(\theta_0^T z)\right)\begin{pmatrix} X_i \\ \theta^T Z_{i0} X_i \end{pmatrix} + R(X_i, Z_i, z, \theta) + \varepsilon_i.$$

Let $C_{\theta,n}(z)$ be the value of $C_n(z)$ in (A.29) with $H_{b,i}(Z_j)$ replaced by $K_{h,i}^{\theta}(Z_j)$ and

$$\begin{pmatrix} a_\theta \\ d_\theta \end{pmatrix} = \begin{pmatrix} G(\theta_0^T z) \\ G'(\theta_0^T z) \end{pmatrix} + C_{\theta,n}^{-1}(z)\sum_{i=1}^n K_{h,i}^{\theta}\begin{pmatrix} X_i \\ Z_{i0}^T\theta X_i \end{pmatrix}\{R(X_i, Z_i, z, \theta) + \varepsilon_i\}.$$

By Lemma A.1, we have $R_{3n}^{\theta}(z) = O(\delta_n)$ and $R_{4n}^{\theta}(z) = O(\delta_n)$. On D^θ,

$$\begin{aligned} a_\theta = {} & G(\theta_0^T z) + \frac{1}{2} G''(\theta_0^T z) h^2 + \pi_\theta^{-1}(z) V_\theta(z)(\theta_0 - \theta) + R_{3n}^{\theta}(z) \\ & + O\{(h + J_\theta(z))\delta_\theta + h^2(h + J_\theta(z) + \delta_n) + \delta_\theta^2\}, \\ d_\theta = {} & G'(\theta_0^T z) + h^{-1} R_{4n}^{\theta}(z) + O\{\tau_n + h^{-1}(\delta_n + J_\theta(z))\delta_\theta\}, \end{aligned} \tag{5.4}$$

where

$$R_{3n}^{\theta}(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(z) X_i \varepsilon_i, \quad R_{4n}^{\theta}(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(z)\{\frac{\theta^T Z_{i0}}{h}\} X_i \varepsilon_i.$$

Let $a_{\theta,j}$ and $d_{\theta,j}$ be the values above with z replaced by Z_j. Write

$$y_i - a_{\theta,j}^T X_i = (d_{\theta,j}^T X_i) Z_{ij}^T \theta_0 + \Delta_{i,j}^{(\theta,0)} + \Delta_{i,j}^{(\theta,1)} + \Delta_{i,j}^{(\theta,2)} + r_{ij} - X_i^T R_{3n}^{\theta}(Z_j) + \varepsilon_i,$$

where $\Delta_{i,j}^{(\theta,0)} = X_i^T \pi_{\theta}^{-1}(z) V_{\theta}(z)(\theta - \theta_0)$, $\Delta_{i,j}^{(\theta,1)} = \{G'(\theta_0^T Z_j) - d_{\theta,j}\}^T X_i \{\theta_0^T Z_{ij}\}$, $\Delta_{i,j}^{(\theta,2)} = \{G''(\theta_0^T Z_j)\}^T X_i \{(\theta_0^T (Z_i - Z_j))^2 - h^2\}/2$ and $|r_{ij}| \le c\{|\theta_0^T Z_{ij}|^3 + (h + J_{\theta}(Z_j))\delta_{\theta} + h^2(h + J_{\theta}(Z_j) + \delta_n) + \delta_{\theta}^2\}|X_i|$. Note that by Lemmas A.1 and A.3, $\sup_{z\in\mathcal{D}} |\hat{f}_{\theta}(z) - f_{\theta}(z) - J_{\theta}(z)| = O(h + \delta_n)$, where $\hat{f}_{\theta}(z) = n^{-1}\sum_{i=1}^{n} K_{h,i}^{\theta}(z)$. Therefore

$$\sup_{z\in\mathcal{D}} |\mathcal{I}(\hat{f}_{\theta}(z)) - \mathcal{I}(f_{\theta}(z)) - J_{\theta}(z)| = O(b + \delta_n). \tag{5.5}$$

Write $\mathcal{I}(\hat{f}_{\theta}(z))$ as $\mathcal{I}_{nj}^{\theta}$. We have

$$\begin{aligned}\theta = \theta_0 + D_{\theta,n}^{+} \sum_{j=1}^{n} \mathcal{I}_{nj}^{\theta} \sum_{i=1}^{n} (d_{\theta,j}^T X_i) K_{h,i}^{\theta}(Z_j) Z_{ij} \{\Delta_{i,j}^{(\theta,0)} + \Delta_{i,j}^{(\theta,1)} + \Delta_{i,j}^{(\theta,2)} + r_{ij} \\ - X_i^T R_{3n}^{\theta}(Z_j) + \varepsilon_i\},\end{aligned} \tag{5.6}$$

where $D_{\theta,n} = n^{-2}\sum_{j=1}^{n} \mathcal{I}_{nj}^{\theta} \sum_{i=1}^{n} (d_{\theta,j}^T X_i)^2 K_{h,i}^{\theta}(Z_j) Z_{ij} Z_{ij}^T$. By (A.40), we have $d_{\theta} = G'(\theta_0^T z) + O\{h^{-1}\delta_n + (1 + h^{-1}J_{\theta}(z))\delta_{\theta}\}$. Exchanging the order of summation we have, by Lemma A.1,

$$\begin{aligned} D_{\theta,n} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{I}_{nj}^{\theta} \{d_{\theta,j}^T X_i\}^2 K_{h,i}^{\theta}(Z_j) Z_{ij} Z_{ij}^T \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_f^{\theta}(Z_i) \{G'^T(\theta_0^T Z_i) X_i\}^2 \{Z_i - \mu_{\theta}(Z_i)\}\{Z_i - \mu_{\theta}(Z_i)\}^T \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_f^{\theta}(Z_i) \{G'^T(\theta_0^T Z_i) X_i\}^2 E\{(Z_i - \mu_{\theta}(Z_i))(Z_i - \mu_{\theta}(Z_i))^T\} \\ &\quad + O(h^{-1}\delta_n + h + \delta_{\theta}) \\ &= W_0 + U_0 + O(h^{-1}\delta_n + h + \delta_{\theta}), \end{aligned}$$

where $\mathcal{I}_f^\theta(z) = \mathcal{I}(f_\theta(z))f_\theta(z)$. By Lemmas A.1 and A.3, we have

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} \Delta_{ij}^{(\theta,0)} = E\{\mathcal{I}_f^\theta(Z)V_\theta(Z)\pi_\theta^{-1}(Z)V_\theta(Z)\}(\theta-\theta_0) + O(h^{-1}\tau_n\delta_\theta + \delta_\theta^2),$$

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} \Delta_{ij}^{(\theta,1)} = O(h^{-1}\tau_n\delta_\theta + h\tau_n + \delta_\theta^2).$$

For any d and d', by Lemmas A.1 and A.3,

$$\frac{1}{n}\sum_{i=1}^n d^T X_i X_i^T d' K_{h,i}^\theta Z_{i0}(\theta_0^T Z_{i0})^2 = \psi_\theta(z)h^2 + O\{h^2(J_\theta(z)+\tau_n) + h\delta_\theta + \delta_\theta^2\},$$

$$\frac{1}{n}\sum_{i=1}^n d^T X_i X_i^T d' K_{h,i}^\theta Z_{i0} = \psi_\theta(z) + O\{J_\theta(z)+\tau_n\},$$

where $\psi_\theta(z) = f_\theta(z)E(d^T X_i X_i^T d' Z_{i0}|\theta^T Z = \theta^T z)$. Therefore

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} \Delta_{ij}^{(\theta,2)} = O\{h^3 + h\delta_\theta + \delta_\theta^2\},$$

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} r_{ij} = O\{h^3 + \delta_\theta^2 + h\delta_\theta + h\delta_n\}.$$

Let $\tilde{V}_\theta(z) = \mathcal{I}^\theta(z)\{G'(\theta_0^T Z_i)\}^T X_i\{\mu_\theta(Z_i) - z\}$. Note that

$$\frac{1}{n}\sum_{j=1}^n \mathcal{I}_{nj}^\theta (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} = \tilde{V}_\theta(Z_i) + \frac{1}{n}\sum_{j=1}^n \{\mathcal{I}_{nj}^\theta (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} - \tilde{V}_\theta(Z_i)\}.$$

Exchanging the order of the summation, by Lemmas A.1 and A.2 we have

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} \varepsilon_i = \frac{1}{n}\sum_{i=1}^n \tilde{V}_\theta(Z_i)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n\delta_\theta)$$
$$= \frac{1}{n}\sum_{i=1}^n \tilde{V}_{\theta_0}(Z_i)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n\delta_\theta).$$

Similarly, we have

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj}^\theta \sum_{i=1}^n (d_{\theta,j}^T X_i) K_{h,i}^\theta(Z_j) Z_{ij} X_i^T R_{3n}^\theta(Z_j) = O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n\delta_\theta).$$

Therefore

$$\theta = \theta_0 + \{W_0 + U_0\}^{-} E\{\mathcal{I}_f^{\theta}(Z) V_{\theta_0}(Z) \pi_{\theta_0}^{-1}(Z) V_{\theta_0}(Z)\}(\theta - \theta_0)$$
$$+ n^{-1}\{W_0 + U_0\}^{-} \sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n \delta_\theta + \delta_\theta^2).$$

Let $D = (W_0 + U_0)^{-1/2} E\{\mathcal{I}_f^{\theta}(Z) V_{\theta_0}(Z) \pi_{\theta_0}^{-1}(Z) V_{\theta_0}(Z)\}(W_0 + U_0)^{-1/2}$. By the Schwarz inequality, we have that $W_0 + U_0 - E\{\mathcal{I}_f^{\theta}(Z) V_{\theta_0}(Z) \pi_{\theta_0}^{-1}(Z) V_{\theta_0}(Z)\}$ is a semi-positive matrix. We have, by Lemma A.4, the eigenvalues of D are less than 1, say $1 > \lambda_1 \geq \cdots \geq \lambda_{q-1} \geq 0$, so take an orthogonal matrix Γ such that $D = \Gamma \operatorname{diag}(\lambda_1, \ldots, \lambda_{q-1}, 0)\Gamma^T$. Let $\beta_k = (W_0 + U_0)^{-1/2}(\theta_k - \theta_0)$ so that

$$\beta_{k+1} = \Gamma \operatorname{diag}(\lambda_1, \ldots, \lambda_{p+q-1}, 0)\Gamma^T \beta_k + n^{-1}\{W_0 + U_0\}^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i$$
$$+ O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n \Delta_k + \Delta_k^2), \tag{5.7}$$

where $\Delta_k = |\beta_k|$. It follows that

$$\begin{aligned}\Delta_{k+1} &\leq \lambda_1 \Delta_k + \delta_{0n} + c(\Delta_k + h^{-1}\tau_n)\Delta_k + c(h^3 + h^{-1}\delta_n^2) \\ &= \delta_{0n} + \{\lambda_1 + c\Delta_k + c(h + h^{-1}\delta_n)\}\Delta_k + c(h\tau_n + h^{-1}\delta_n^2)\end{aligned} \tag{5.8}$$

almost surely, where c is a constant. We can further take $c > 1$. For sufficiently large n, we may assume that

$$c(h + h^{-1}\delta_n) \leq \frac{1 - \lambda_1}{3}, \quad \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2) \leq \frac{(1 - \lambda_1)^2}{9c}. \tag{5.9}$$

By Lemma 1, $\Delta_1 \to 0$ almost surely, and we may assume

$$\Delta_1 \leq \frac{1 - \lambda_1}{3c}. \tag{5.10}$$

Therefore, it follows from (A.44), (A.45) and (A.46) that

$$\Delta_2 \leq \{\lambda_1 + \frac{2}{3}(1 - \lambda_1)\}\frac{1 - \lambda_1}{3c} + \frac{(1 - \lambda_1)^2}{9c} = \frac{1 - \lambda_1}{3c}. \tag{5.11}$$

From (A.44), (A.45) and (A.47), we have $\Delta_3 \leq (1 - \lambda_1)/(3c)$. By induction, $\Delta_k \leq (1 - \lambda_1)/(3c)$ for all k. Therefore we have from (A.44) that $\Delta_{k+1} \leq \lambda_0 \Delta_k + \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)$ almost surely, where $0 \leq \lambda_0 < (2 + \lambda_1)/3 < 1$. It follows that

$$\Delta_k \leq \lambda_0^k \Delta_1 + \{\delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)\} \sum_{j=1}^{k} \lambda_0^j = O(\delta_{0n} + h\tau_n + h^{-1}\delta_n^2),$$

for sufficiently large k. By (A.43), we have

$$\{W_0 + U_0\}^{\frac{1}{2}}(\hat{\theta} - \theta_0) = D(\hat{\theta} - \theta_0) + n^{-1}\{W_0 + U_0\}^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2). \tag{5.12}$$

It follows from (A.48) that $W_1(\hat{\theta} - \theta_0) = n^{-1}\sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2)$, and we have completed the proof of the first part of Theorem 1.

Acknowledgements

We thank Professor J. Fan and his coauthors for making their manuscript available to us before its publication. YX thanks the Alexander von Humboldt foundation, NUS FRG R-155-000-048-112, and the National Natural Science Foundation of China (Grant No. 10471061) for partial support. WKL thanks the Hong Kong Research Grants Council (Grant No. 7036/06P) and an Outstanding Researcher Award (2002) bestowed by the University of Hong Kong for partial support. HT thanks the Hong Kong Research Grants Council and the Distinguished Research Achievement Award (2002) bestowed by the University of Hong Kong for partial support.

References

Abrams, P. A. and Ginzburg, L. R. (2000). The nature of predation: prey dependent, ratio dependent, or neither? *Trends Ecol. Evol.* **15**, 337-341.

Bohannan, B. J. M. and Lenski, R. E. (1999). Effect of prey heterogeneity on the response of a model food chain to resource enrichment. *Am. Nat.* **153**, 73-82.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95**, 941-956.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models. *J. Time Ser. Anal.* **7**, 178-190.

Chen, R. (1995). Threshold variable selection in open-loop threshold autoregressive models. *J. Time Ser. Anal.* **16**, 461-481.

Chen, R. and Tsay, S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298-308.

Fan, J. Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. Roy. Statist. Soc. Ser. B* **65**, 57-80.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* Wiley, New York.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17**, 573-588.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the single-index coefficients in single-index models. *Ann. Statist.* **29**, 1537-1566.

Jost, C. and Ellner, S. (2000). Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proc. R. Soc. Lond. Ser. B* **367**, 1611-1620.

Kuang, Y. and Beretta, E. (1998). Global qualitative analysis of ratio-dependnet predator-prey models. *J. Math. Biol.* **36**, 389-406.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Amer. Statist. Assoc.* **86**, 316-342.

Li, W. K. (1992). On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika* **79**, 435-437.

Linton, O. (1995). Second order approximation in the partially linear regression model. *Econometrica* **63**, 1079-1112.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

Rogers, D. and Hassell, M. P. (1974). General models for insect parasite and predator serching behaviour: interference. *J. Anim. Ecol.* **43**, 239-253.

Tong, H. (1990). *Nonlinear Time Series Analysis: a Dynamical System Approach.* Oxford University Press, Oxford.

Veilleux, B. G. (1976). The analysis of a predatory interaction between Didinium and Paramecium. Masters theis, University of Alberta.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94**, 1275-1285.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore.

E-mail: staxyc@nus.edu.sg

Department of Statistics and Actuarial Science, University of Hong Kong, China.

E-mail: hrntlwk@hkucc.hku.hk

Department of Statistics, London School of Economics, Houghton Street, London, WC2A2AE, U.K.

E-mail: h.tong@lse.ac.uk

(Received December 2004; accepted December 2005)

Statistica Sinica **17**(2007): Supplement, S39∼S57

THRESHOLD VARIABLE SELECTION USING NONPARAMETRIC METHODS

Yingcun Xia, Wai Keung Li and Howell Tong

National University of Singapore, University of Hong Kong and London School of Economics

Supplementary Material

Appendix. Proofs

The idea of proofs can be stated as follows. We first develop some Lemmas (Lemmas A.1−A.3) to obtain uniform consistency rates for the local linear estimators of the coefficient functions; see equation (5.40). Based on the expansions and equation (2.6), we then build a recursive formula for the iteration in the algorithm, i.e.,

$$\theta_{k+1} - \theta_0 = \Gamma_k(\theta_k - \theta_0) + \text{Smaller term},$$

where θ_k is the estimator of θ_0 after the kth iteration and $||\Gamma_k|| \le 1$. See (5.7) for more details. This recursive formula indicates that the true direction θ_0 is the attractor of the algorithm. The formula is finally used to prove the convergence of the algorithm as well as the consistency and asymptotic normality of the estimator.

Let $\delta_\theta = |\theta - \theta_0|$. In Θ, δ_θ is bounded. Let $\delta_{qn} = \{\log n/(nb^q)\}^{1/2}$, $\tau_{qn} = b^2 + \delta_{qn}$, $\delta_n = \{\log n/(nh)\}^{1/2}$, $\tau_n = h^2 + \delta_n$ and $\delta_{0n} = (\log n/n)^{1/2}$. By the condition $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$, we have $\delta_{0n} \ll h^2 \ll h^{-1}\delta_n$ and $\delta_n \ll h$. We shall use these relations frequently in our calculations. Suppose A_n is a matrix. $A_n = O(a_n)$ means every element in A_n is $O(a_n)$ almost surely. We adopt the consistency in the sense of "almost surely" because we need to prove the convergence of the algorithm, which theoretically needs infinite iterations. Let $c, c_1, c_2, \cdots$ be a set of constants. For ease of exposition, c may have different values at different places. We abbreviate $K_h(\theta^T Z_{i0}) = h^{-1}K(\theta^T(Z_i - z)/h)$ and $H_b(Z_{i0}) = h^{-q}H\{(Z_i - z)/h\}$ as $K^\theta_{h,i}(z)$ (or $K^\theta_{h,i}$) and $H_{b,i}(z)$ (or $H_{b,i}$) respectively in the following context for simplicity.

Lemma A.1. *Suppose $\varphi(\theta)$ is measurable function of (X, Z, y) such that $\sup_{\theta,\vartheta\in\Theta}$ $|\varphi(\theta) - \varphi(\vartheta)| < M(X, Z, y)|\theta - \vartheta|$ a.s. with $EM^r(X, Z, y) < c$; $\sup_{\theta\in\Theta,v} E(|\varphi(\theta)|^r \mid \theta^T Z = v) < c$ for some $r \ge 3$; Let $\varphi_i(\theta)$ be the corresponding value of $\varphi(\theta)$ at*

(X_i, Z_i, y_i). *Assume that* $\sup_{\theta\in\Theta,u,v} E(|\varphi_i(\theta)\varphi_1(\theta)| \mid \theta^T Z_1 = u, \theta^T Z_i = v) < c$ *for all* $i > 1$. *Let* $g(v)$ *be any function with continuous second order derivative,* $m(u,v) = g(u) - g(v) - g'(v)(u-v) - g''(v)(u-v)^2/2$ *and* $\zeta_i^{k,\ell} = m(\theta_0^T Z_i, \theta_0^T z)\mathbf{x}_i^k$ $(\theta^T Z_{i0})^\ell$ *where* $\mathbf{x}_i$ *is any component of* X_i, $k = 0, 1$ *and* $\ell = 0, 1$. *If* (C1) *holds, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n \varphi_i(\theta) - E\varphi_i(\theta)\Big| = O(\delta_{0n}),$$

$$\sup_{|\theta-\theta_0|<a_n}\Big|\frac{1}{n}\sum_{i=1}^n\{\varphi_i(\theta) - \varphi_i(\theta_0)\} - E\{\varphi_i(\theta) - \varphi_i(\theta_0)\}\Big| = O(a_n\delta_{0n}),$$

where $a_n \to 0$ *as* $n \to \infty$. *If further* (C2) *and* (C6) *hold,* $h \sim n^{-\delta}$ *with* $0 < \delta < 1 - 2/r$, *then*

$$\sup_{\substack{\theta\in\Theta\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^n\{H_{b,i}\varphi_i(\theta) - E(H_{b,i}\varphi_i(\theta))\}\Big| = O(\delta_{qn}),$$

$$\sup_{\substack{\theta\in\Theta\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta\varphi_i(\theta) - E(K_{h,i}^\theta\varphi_i(\theta))\}\Big| = O(\delta_n),$$

$$\sup_{\substack{|\theta-\theta_0|<a_n\\ z\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta\zeta_i^{k,\ell} - E(K_{h,i}^\theta\zeta_i^{k,\ell})\}\Big| = O\{\delta_n h^\ell(a_n^2 + h^2)\}.$$

Proof. The proofs of Lemma A.1 are quite standard; see, e.g., Härdle, Janssen and Serfling (1988) and Xia and Li (1999). We here give the details for the last two equations. Note that $\Theta\otimes\mathcal{D}\subset\mathbb{R}^{2q}$ is bounded. There are n^{2q} balls B_{n_k} centered at (θ_{n_k}, z_{n_k}), $1 \le k \le n^{2q}$, with diameter less then $cn^{-1/2}h^{3/2}(> c/n)$, such that $\Theta\otimes\mathcal{D}\subset\cup_{1\le k\le n^{2q}}B_{n_k}$. Then

$$\begin{aligned}
&\sup_{z\in\mathcal{D},\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta(z)\varphi_i(\theta) - E(K_{h,i}^\theta(z)\varphi_i(\theta))\}\Big|\\
&\le \max_{1\le k\le n^{2q}}\Big|\frac{1}{n}\sum_{i=1}^n\Big[K_{h,i}^{\theta_{n_k}}(z_{n_k})\varphi_i(\theta_{n_k}) - E\{K_{h,i}^{\theta_{n_k}}(z_{n_k})\varphi_i(\theta_{n_k})\}\Big]\Big|\\
&\quad + \max_{1\le k\le n^{2q}}\sup_{(\theta,z)\in B_{n_k}}\Big|\frac{1}{n}\sum_{i=1}^n\Big[\{K_{h,i}^\theta(z) - K_{h,i}^{\theta_{n_k}}(z_{n_k})\}\varphi_i(\theta) + \{\varphi_i(\theta) - \varphi_i(\theta_{n_k})\}K_{h,i}^{\theta_{n_k}}(z)\\
&\quad - E\{K_{h,i}^\theta(z) - K_{h,i}^{\theta_{n_k}}(z_{n_k})\}\varphi_i(\theta) - E\{\varphi_i(\theta) - \varphi_i(\theta_{n_k})\}K_{h,i}^{\theta_{n_k}}(z)\Big]\Big|\\
&\triangleq \max_{1\le k\le n^{2q}}|R_{n,k,1}| + \max_{1\le k\le n^{2q}}\sup_{(\theta,z)\in B_{n_k}}|R_{n,k,2}|. \qquad \text{(A.1)}
\end{aligned}$$

By assumption (C6), we have

$$\max_{\substack{1\le k\le n^{2q}\\ z\in\mathcal{D}}}\sup_{(\theta,z)\in B_{n_k}}|K_{h,i}^{\theta}(z)-K_{h,i}^{\theta_{n_k}}(z_{n_k})|\le \max_{\substack{1\le k\le n^{2q}\\ z\in\mathcal{D}}}\sup_{(\theta,z)\in B_{n_k}} ch^{-2}(|\theta-\theta_{n_k}|+|z-z_{n_k}|)$$
$$\le c(nh)^{-\frac{1}{2}},$$
$$\max_{1\le k\le n^{2q}}\sup_{\theta\in B_{n_k}}|\varphi_i(\theta)-\varphi_i(\theta_{n_k})|\le M(X_i,Z_i,y_i)n^{-\frac{1}{2}}h^{\frac{3}{2}}.$$

By the strong law of large numbers for dependent observations (see, e.g., Rio (1995)), we have

$$\max_{1\le k\le n^{2q}}\sup_{(\theta,z)\in B_{n_k}}|R_{n,k,2}|\le c(nh)^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^{n}\{|\varphi_i(\theta)|+M(X_i,Z_i,y_i)\}=O(\delta_n).\quad\text{(A.2)}$$

Write $\varphi(\theta_{n_k})$ as φ_i for simplicity. More clearly, we write h as h_n. Let $T_\ell=\{\ell/(h_\ell\log(\ell))\}^{\kappa}$, where $\kappa=1/(2r-2)$. Let $\varphi_{i,\ell}^{o}=\varphi_i I\{|\varphi_i|\ge T_\ell\}$ and $\varphi_{i,\ell}^{I}=\varphi_i-\varphi_{i,\ell}^{o}$. We have

$$R_{n,k,1}=\frac{1}{n}\sum_{i=1}^{n}\Big[K_{h,i}^{\theta}(z)\varphi_{i,i}^{o}-E\{K_{h,i}^{\theta}(z)\varphi_{i,i}^{o}\}\Big]+\frac{1}{n}\sum_{i=1}^{n}\xi_{n_k,i},\tag{A.3}$$

where $\xi_{n_k,i}=K_{h,i}^{\theta_{n_k}}(z_{n_k})\varphi_{i,i}^{I}-E\{K_{h,i}^{\theta_{n_k}}(z_{n_k})\varphi_{i,i}^{I}\}$.

It is easy to check that

$$\sum_{\ell=1}^{\infty}(\frac{\ell}{h_\ell})^{-\frac{1}{2}}E|\varphi_{\ell,\ell}^{o}|\le\sum_{\ell=1}^{\infty}(\frac{\ell}{h_\ell})^{-\frac{1}{2}}T_\ell^{-r+1}E|\varphi_\ell|^{r}<\infty.$$

Therefore (cf., Rao (1973, p.111))

$$\sum_{\ell=1}^{\infty}(\frac{\ell}{h_\ell})^{-\frac{1}{2}}|\varphi_{\ell,\ell}^{o}|<\infty$$

almost surely. By the Kronecker lemma, we have

$$\frac{1}{n}\sum_{\ell=1}^{n}E|\varphi_{\ell,\ell}^{o}|=O\{(\frac{n}{h})^{-\frac{1}{2}}\},\qquad \frac{1}{n}\sum_{\ell=1}^{n}|\varphi_{\ell,\ell}^{o}|=O\{(\frac{n}{h})^{-\frac{1}{2}}\}.$$

Note that $|\varphi_{\ell,n}^{o}|\le|\varphi_{\ell,\ell}^{o}|$ for all $\ell\le n$, and $|K_{h,i}^{\theta_{n_k}}(z)|<ch^{-1}$ by (C6). We have

$$\max_{1\le k\le n^{2q}}\frac{1}{n}\sum_{i=1}^{n}E|K_{h,i}^{\theta_{n_k}}(z)\varphi_{i,n}^{o}|=O\{(nh)^{-\frac{1}{2}}\},\tag{A.4}$$

$$\max_{1\le k\le n^{2q}}\frac{1}{n}\sum_{i=1}^{n}|K_{h,i}^{\theta_{n_k}}(z)\varphi_{i,n}^{o}|=O\{(nh)^{-\frac{1}{2}}\}.\tag{A.5}$$

Next, we shall show

$$\max_{1\le k\le n^{2q}} \mathrm{Var}(\sum_{i=1}^{n} \xi_{n_k,i}) \le \frac{c_1 n}{h}. \tag{A.6}$$

By stationarity in (C1), we have

$$\mathrm{Var}(\sum_{i=1}^{n} \xi_{n_k,i}) = n\mathrm{Var}(\xi_{n_k,i}) + 2\sum_{i=2}^{n}(n-i)\mathrm{Cov}(\xi_{n_k,1}, \xi_{n_k,i}). \tag{A.7}$$

Let $\tilde{\varphi}(u) = E(|\varphi(\theta_{n_k})|^\ell \mid \theta_{n_k}^T Z = u)$ and $\tilde{\varphi}(u,v|i) = E(|\varphi_1\varphi_i| \mid \theta_{n_k}^T Z_1 = u, \theta_{n_k}^T Z_i = v)$. By the conditions about φ in Lemma A.1 and assumption (C2), we have

$$\begin{aligned}
L(\ell) &\triangleq E\{(K_{h,i}^{\theta_{n_k}}(z_{n_k}))^\ell |\varphi_i|^\ell\} = E\{(K_{h,i}^{\theta_{n_k}}(z_{n_k}))^\ell E(|\varphi_i|^\ell \mid \theta_{n_k}^T Z_i)\} \\
&= h^{-\ell}\int (K_h(u-\theta_{n_k}^T z_{n_k}))^\ell \tilde{\varphi}_{\theta_{n_k}}(u) f_{\theta_{n_k}^T Z}(u)du \\
&= h^{-\ell+1}\int (K(u))^\ell \tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T z_{n_k}+hu) f_{\theta_{n_k}^T Z}(\theta_{n_k}^T z_{n_k}+hu)du \\
&\le ch^{-\ell+1}, \qquad 0\le \ell \le r, \\
M(i) &\triangleq E\left\{K_{h,1}^{\theta_{n_k}}(z_{n_k}) K_{h,i}^{\theta_{n_k}}(z_{n_k}) |\varphi_1\varphi_i|\right\} \\
&\le E\left\{K_{h,1}^{\theta_{n_k}}(z_{n_k}) K_{h,i}^{\theta_{n_k}}(z_{n_k}) E\left(|\varphi_1\varphi_i| \mid \theta_{n_k}^T Z_1, \theta_{n_k}^T Z_i\right)\right\} \\
&= h^{-2}\int K\{\frac{u-\theta_{n_k}^T z_{n_k}}{h}\} K\{\frac{v-\theta_{n_k}^T z_{n_k}}{h}\} \tilde{\varphi}_{\theta_{n_k}}(u,v|i) f_{\theta_{n_k}^T Z_1, \theta_{n_k}^T Z_i}(u,v)dudv \\
&= \int K(u)K(v)\tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T z_{n_k}+hu, \theta_{n_k}^T z_{n_k}+hv|i) \\
&\quad \times f_{\theta_{n_k}^T Z_1, \theta_{n_k}^T Z_i}(\theta_{n_k}^T z_{n_k}+hu, \theta_{n_k}^T z_{n_k}+hv)dudv \\
&\le c\int K(u)K(v)\tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T z_{n_k}+hu, \theta_{n_k}^T z_{n_k}+hv|i)dudv \le c, \quad i=2,3,\ldots,
\end{aligned}$$

where $f_{\theta_{n_k}^T Z}$ and $f_{\theta_{n_k}^T Z_1, \theta_{n_k}^T Z_i}$ are the density functions of $\theta_{n_k}^T Z$ and $(\theta_{n_k}^T Z_1, \theta_{n_k}^T Z_i)$ respectively. It follows that

$$\mathrm{Var}(\xi_{n_k,i}) \le L(2) \le \frac{c}{h}. \tag{A.8}$$

By the Davydov's lemma (Hall and Heyde (1980, Corollary 2)),

$$\begin{aligned}
|\mathrm{Cov}(\xi_{n_k,1}, \xi_{n_k,i})| &\le 8\{\alpha(i-1)\}^{1-\frac{2}{r}}(E|\xi_{n_k,1}|^r)^{\frac{2}{r}} \\
&\le 8\{\alpha(i-1)\}^{1-\frac{2}{r}}\{L(r)\}^{\frac{2}{r}} \\
&\le ch^{-2+2/r}\{\alpha(i-1)\}^{1-\frac{2}{r}}.
\end{aligned} \tag{A.9}$$

Let $N_1 = INT(h^{(-1+2/r)/(2q)})$, where $INT(v)$ denotes the integer part of v. From (A.7)−(A.9) and assumption (C1), we have

$$\begin{aligned}
\mathrm{Var}(\sum_{i=1}^{n}\xi_{n_k,i}) &= n\mathrm{Var}(\xi_{n_k,i}) + 2\Big(\sum_{i=2}^{N_1} + \sum_{i=N_1+1}^{n}\Big)(n-i)\mathrm{Cov}(\xi_{n_k,1},\xi_{n_k,i}) \\
&\le \frac{cn}{h} + 2cn\sum_{i=2}^{N_1} M(i) + 2cnh^{-2+\frac{2}{r}}\sum_{i=N_1+1}^{n}\{\alpha(i-1)\}^{1-\frac{2}{r}} \\
&\le \frac{cn}{h} + 2cnN_1 + 2cnh^{-2+\frac{2}{r}}N_1^{-2q}\sum_{i=N_1+1}^{n} i^{2q}\{\alpha(i-1)\}^{1-\frac{2}{r}} \\
&\le \frac{cn}{h}.
\end{aligned}$$

Note that c does not depend on k. Therefore (A.6) follows.

Let $N_2 = INT(n^{1/2-1/r}h^{1/2+1/r}(\log n)^{-1/2})$ and $N_3 = INT(n/(2N_2))$. Then $n = 2N_2N_3 + N_0$ and $0 \le N_0 < 2N_2$. We write

$$W_{n_k}(j) = \sum_{i=(j-1)N_2+1}^{j\cdot N_2} \xi_{n_k,i}, \quad j = 1,\ldots,2N_2.$$

Then

$$\sum_{i=1}^{n}\xi_{n_k,i} = \sum_{j=1}^{N_3} W_{n_k}(2j-1) + \sum_{j=1}^{N_3} W_{n_k}(2j) + S_{n,0}^T\ , \tag{A.10}$$

where $S_{n,0}^T$ is the residual and has less than $2N_2$ terms. Its contribution is negligible.

For every $\eta > 0$, we use the strong approximation theorem of Bradley (1983) to approximate the random variables $W_{n_k}(1), W_{n_k}(3),\ldots,W_{n_k}(2j-1)$ by independent random variables $W_{n_k}^*(1), W_{n_k}^*(3),\ldots,W_{n_k}^*(2j-1)$ defined as follows. By enlarging the probability space if necessary, introduce a sequence $(U_1, U_2,\cdots)$ of independent uniform $[0,1]$ random variables that are independent of $\{W_{n_k}(1),\ldots,W_{n_k}(2j-1)\}$. Define $W_{n_k}^*(0) = 0, W_{n_k}^*(1) = W_{n_k}(1)$. Then for each $j \ge 2$, there exists a random variable $W_{n_k}^*(2j-1)$ which is a measurable function of $W_{n_k}(1), W_{n_k}(3),\ldots,W_{n_k}(2j-1)$ and U_j such that $W_{n_k}^*(2j-1)$ is independent of $W_{n_k}^*(1),\ldots,W_{n_k}^*(2j-3)$, has the same distributions as $W_{n_k}(2j-1)$ and satisfies

$$P(|W_{n_k}^*(2j-1) - W_{n_k}(2j-1)| > \eta) \le 18(\frac{|W_{n_k}(2j-1)|_\infty}{\eta})^{\frac{1}{2}}\alpha(N_2), \tag{A.11}$$

where $|\cdot|_\infty$ is the sup-norm. It follows from the definition of $W^*_{n_k}(2j-1)$ and (A.6) that,

$$\begin{aligned} &EW^*_{n_k}(2j-1)=0,\\ &\max_{k,j}\mathrm{Var}(W^*_{n_k}(2j-1))\le c_2 n^{\frac{1}{2}-\frac{1}{r}}h^{-\frac{1}{2}+\frac{1}{r}}(\log n)^{-\frac{1}{2}} \overset{\Delta}{=} N_4. \end{aligned} \tag{A.12}$$

By the condition in Lemma A.1, we have $h^{-r}(n/\log n)^{-r+2}\to 0$. Hence

$$\begin{aligned} \max_{1\le k\le n^{2q}}|\xi_{n_k,i}| &\le ch^{-1}T_n = c\{\frac{n}{h\log n}\}^{\frac{1}{2}}\{h^{-r}(\frac{n}{\log n})^{-r+2}\}^{\kappa}\\ &\le c_3\{\frac{n}{h\log n}\}^{\frac{1}{2}} \triangleq N_5. \end{aligned} \tag{A.13}$$

Let $N_6 = c_4(nh^{-1}\log n)^{1/2}$. By the Bernstein's inequality, we have from (A.12) and (A.13)

$$\begin{aligned} P\Big(\Big|\sum_{j=1}^{N_3} W^*_{n_k}(2j-1)\Big| > N_6\Big) &\le \exp\left(\frac{-c_4^2 nh^{-1}\log n}{2(N_3N_4+N_5N_6)}\right)\\ &\le \exp\{-c_4^2\log\frac{n}{c_2+2c_3c_4}\}\\ &\le c_5 n^{-2q-2}. \end{aligned} \tag{A.14}$$

The last inequality holds if we choose c_4 sufficiently large. By (A.11), if (i) $N_6/N_3 \le |W^*_{n_k}(2j-1)|_\infty$, we have

$$\begin{aligned} \Pr(|W_{n_k}(2j-1)-W^*_{n_k}(2j-1)|>\frac{N_6}{N_3}) &\le 18\Big(\frac{N_2N_5}{\frac{N_6}{N_3}}\Big)^{\frac{1}{2}}\alpha(N_2)\\ &\le c_6(\frac{n}{\log n})^{\frac{1}{2}}\alpha(N_2); \end{aligned} \tag{A.15}$$

if (ii) $N_6/N_3 > |W^*_{n_k}(2j-1)|_\infty$, take $\eta = |W^*_{n_k}(2j-1)|_\infty$ in (A.11), we have

$$\Pr(|W_{n_k}(2j-1)-W^*_{n_k}(2j-1)|>\eta)\le 18\alpha(N_2)\ ,$$

which is smaller than the right hand side of (A.15) as $n\to\infty$. Therefore,

$$\begin{aligned} &\Pr(|\sum_{j=1}^{N_3}\{W_{n_k}(2j-1)-W^*_{n_k}(2j-1)\}|>N_6)\\ &\le \sum_{j=1}^{N_3}\Pr(|W_{n_k}(2j-1)-W^*_{n_k}(2j-1)|>\frac{N_6}{N_3})\\ &\le c_7N_3(\frac{n}{\log n})^{\frac{1}{2}}\alpha(N_2). \end{aligned} \tag{A.16}$$

From (A.14) and (A.16), we have

$$
\begin{aligned}
&\Pr(\max_{1\le k\le n^{2q}}|\sum_{j=1}^{N_3} W_{n_k}(2j-1)| \ge 2N_6)\\
&\le \sum_{k=1}^{n^{2q}} \Pr(|\sum_{j=1}^{N_3} W^*_{n_k}(2j-1)| \ge N_6) + \sum_{k=1}^{n^{2q}} \Pr(|\sum_{j=1}^{N_3} |W_{n_k}(2j-1) - W^*_{n_k}(2j-1)| \ge N_6)\\
&\le n^{2q}\{c_5 n^{-2q-2} + c_7 N_3 (\frac{n}{\log n})^{\frac{1}{2}} \alpha(N_2)\}.
\end{aligned}
$$

By (2.7), it follows that

$$
\sum_{n=1}^{\infty} \Pr(\max_{1\le k\le n^{2q}}|\sum_{j=1}^{N_3} W_{n_k}(2j-1)| \ge 2N_6) < \infty.
$$

By the Borel-Cantelli lemma, we have

$$
\max_{1\le k\le n^{2q}}|\sum_{j=1}^{N_3} W_{n_k}(2j-1)| = O(N_6). \tag{A.17}
$$

Similarly, we can show

$$
\max_{1\le k\le n^{2q}}|\sum_{j=1}^{N_3} W_{n_k}(2j)| = O(N_6). \tag{A.18}
$$

Combining (A.4), (A.5), (A.10), (A.17), (A.18) and (A.3), we have

$$
\max_{1\le k\le n^{2q}} |R_{n,k,1}| = O(\delta_n). \tag{A.19}
$$

Therefore, the fourth part of Lemma A.1 follows from (A.1), (A.2) and (A.19).

Note that the key steps in the proof above are the continuity of the related functions and bounded variance in (A.6). To prove the last part of Lemma A.1, it is sufficient to show

$$
\sup_{|\theta-\theta_0|\le a_n, z\in\mathcal{D}} E(K^{\theta}_{h,i}\zeta_i^{k,\ell})^{\tau} \le c h^{\tau\ell-\tau+1}(a_n^{2\tau} + h^{2\tau}), \quad 2\le\tau\le r. \tag{A.20}
$$

Write $\theta_0 = b_n\theta + e_n\vartheta$, where $\vartheta \perp \theta$ and $\theta, \vartheta \in \Theta$. It is easy to see that $|b_n| < c$ and $|e_n| \sim a_n$ when $|\theta - \theta_0| < a_n$. Let $(\theta, \vartheta, \Gamma)$ be an orthogonal matrix. Let $\tilde{f}(v, u_1, u_2, \ldots, u_p)$ and $\tilde{f}(v, u_1, u_2)$ be the density functions of $(\mathbf{x}, \theta^T Z, \vartheta^T Z, \Gamma^T Z)$

and $(\mathbf{x}, \theta^T Z, \vartheta^T Z)$ respectively. We have

$$\begin{aligned}
&E(K_{h,i}^{\theta}\zeta_i^{k,\ell})^{\tau} \\
&\quad= \int (K_h(u_1-\theta^T z))^{\tau}(u_1-\theta^T z)^{\tau\ell}v^{\tau k}m^{\tau}(b_n u_1+e_n u_2, b_n\theta^T z+e_n\vartheta^T z) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \times\tilde{f}(v,u_1,u_2\ldots,u_p)dvdu_1du_2\cdots du_p \\
&\quad= h^{\tau\ell-\tau+1}\int (K(v_1))^{\tau}v_1^{\tau\ell}v^{\tau k}m^{\tau}(b_n v_1 h+b_n\theta^T z+e_n u_2, e_n\theta^T z+b_n\vartheta^T z) \\
&\qquad\qquad\qquad\qquad\qquad \times\tilde{f}(v,\theta^T z+hv_1,u_2,\ldots,u_p)dvdv_1du_2\cdots du_p \\
&\quad= h^{\tau\ell-\tau+1}\int (K(v_1))^{\tau}v_1^{\tau\ell}v^{\tau k}m^{\tau}(b_n v_1 h+b_n\theta^T z+e_n u_2, b_n\theta^T z+e_n\vartheta^T z) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \times\tilde{f}(v,\theta^T z+hv_1,u_2)dvdv_1du_2.
\end{aligned}$$

Note that $|m(u,v)| \le c(u-v)^2$. Therefore by (C2)

$$\begin{aligned}
&E(K_{h,i}^{\theta}\zeta_i^{k,\ell})^{\tau} \\
&\quad\le ch^{\tau\ell-\tau+1}\int (K(v_1))^{\tau}v_1^{\tau\ell}v^{\tau k}(b_n^{2\tau}v_1^{2\tau}h^{2\tau}+e_n^{2\tau})\tilde{f}(v,\theta^T z+hv_1,u_2)dvdv_1du_2 \\
&\quad= O\{h^{\tau\ell-\tau+1}(a_n^{2\tau}+h^{2\tau})\}.
\end{aligned}$$

The equations in Lemma A.1 still hold if we replace $|\theta-\theta_0| < a_n$ with $|\theta+\theta_0| < a_n$. The latter is needed for the proof of Theorem 1 when $\tilde{\theta}^T\theta_0 < 0$. For any measurable function $A(\xi,\eta)$, let $E_k A(\xi_i,\eta_k) = E\{A(v,\eta_k)\}|_{v=\xi_i}$.

Lemma A.2. *Let $\xi(\theta)$ is a measurable function of (X,Z,y). Suppose $E\{\xi(\theta) \mid \theta^T Z\} = 0$ for all $\theta\in\Theta$ and $|\xi(\theta)-\xi(\vartheta)| \le |\theta-\vartheta|\tilde{\xi}$ with $E\tilde{\xi}^r < \infty$ for some $r > 2$. Let φ_i be defined in Lemma* A.1. *If* (C1) *and* (C6) *hold, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big\{K_{h,i}^{\theta}(Z_j)\varphi_j(\theta)-E_j(K_{h,i}^{\theta}(Z_j)\varphi_j(\theta))\Big\}\xi_i(\theta)\Big| = O(\delta_n^2).$$

Proof. Let $\Delta_n(\theta)$ be the value in the absolute symbols. By the continuity of $K_{h,i}^{\theta}$ in θ, there are $n_1 < cn^{2q}$ points $\theta_{n,1},\ldots,\theta_{n,n_1}$ in Θ such that $\cup_{k=1}^{n_1}\{\theta : |\theta-\theta_{n,k}| < h^2\delta_n^2\} \supset \Theta$ and

$$\max_{1\le k\le n_1}\sup_{|\theta-\theta_{n,k}|<h^2\delta_n^2}\Big|\Delta_n(\theta)-\Delta_n(\theta_{n,k})\Big| = O(\delta_n^2). \tag{A.21}$$

The Fourier transform $\phi(s) = \int \exp(\mathrm{i}sv)K(v)dv$ will be used in the following, where i is the imaginary unit. Thus $K(v) = \int \exp(-\mathrm{i}sv)\phi(s)ds$. We have

$$\Delta_n(\theta_{n,k}) = \frac{1}{n^2}h^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n}\int \Big[\exp\{-\mathrm{i}s\theta_{n,k}^T\frac{Z_{ij}}{h}\}\varphi_j(\theta_{n,k}) \\ -E_j\{\exp(-\mathrm{i}s\theta_{n,k}^T\frac{Z_{ij}}{h})\varphi_j(\theta_{n,k})\}\Big]\phi(s)ds\xi_i(\theta_{n,k}) \\ = h^{-1}\int \frac{1}{n}\sum_{i=1}^{n}\exp(-\mathrm{i}s\theta_{n,k}^T\frac{Z_j}{h})\xi_i(\theta_{n,k})\cdot\frac{1}{n}\sum_{j=1}^{n}\Big[\exp(\mathrm{i}s\theta_{n,k}^T\frac{Z_j}{h})\varphi_j(\theta_{n,k}) \\ -E\{\exp(\mathrm{i}s\theta_{n,k}^T\frac{Z_j}{h})\varphi_j(\theta_{n,k})\}\Big]\phi(s)ds.$$

Following the same steps leading to (A.19), we have

$$\max_{1\le k\le n_1}|\frac{1}{n}\sum_{i=1}^{n}\exp(-\mathrm{i}s\theta_{n,k}^T\frac{Z_i}{h})\xi_i(\theta_{n,k})| \le c_8\delta_{0n}, \\ \max_{1\le k\le n_1}\Big|\frac{1}{n}\sum_{j=1}^{n}\Big[\exp(\mathrm{i}s\theta_{n,k}^T\frac{Z_i}{h})\varphi_j(\theta_{n,k}) - E\{\exp(\mathrm{i}s\theta_{n,k}^T\frac{Z_i}{h})\varphi_j(\theta_{n,k})\}\Big]\Big| \le c_9\delta_{0n}$$

almost surely, where c_8 and c_9 are constants which do not depend on s. Hence

$$\max_{1\le k\le n_1}\Big|\Delta_n(\theta_{n,k})\Big| \le h^{-1}\int c_8\delta_{0n}c_9\delta_{0n}|\phi(s)|ds = O(h^{-1}\delta_{0n}^2) = O(\delta_n^2). \tag{A.22}$$

Note that

$$\sup_{\theta\in\Theta}|\Delta_n(\theta)| \le \max_{1\le k\le n_1}\Big|\Delta_n(\theta_{n,k})\Big| + \max_{1\le k\le n_1}\sup_{|\theta-\theta_{n,k}|<h^2\delta_n^2}\Big|\Delta_n(\theta)-\Delta_n(\theta_{n,k})\Big|. \tag{A.23}$$

Therefore, the second part of Lemma A.2 follows from (A.21), (A.22) and (A.23).

Let $d(z,\mathcal{D}^c) = \min_{z'\in\mathbb{R}^q-\mathcal{D}}|z-z'|$, and $J_0(z)$ and $J_\theta(v)$ be any bounded functions such that $J_0(z) = 0$ if $d(z,\mathbb{R}^q-\mathcal{D}) > b$ and $J_\theta(\theta^T z) = 0$ if $d(\theta^T z, \theta^T(\mathbb{R}^q-\mathcal{D})) > h$. By definition, we have

$$\frac{1}{n}\sum_{j=1}^{n}J_0(Z_j) = O(b), \qquad \frac{1}{n}\sum_{j=1}^{n}J_\theta(Z_j) = O(h). \tag{A.24}$$

Let $r(v_1,v_2,x) = G^T(v_1)x - G^T(v_2)x - \{G'^T(v_2)x\}(v_1-v_2) - \{G''^T(v_2)x\}(v_1-v_2)^2/2$. To cope with the boundary points, we give the following nonuniform rates of convergence.

Lemma A.3. *Suppose assumptions* (C2), (C3) *and* (C6) *hold. Then*

$$EH_{b,i}\{\theta^T \frac{Z_{i0}}{b}\}^k\{\vartheta^T \frac{Z_{i0}}{b}\}^\ell = v_{k,\ell}^{\theta,\vartheta} f(z) + J_0(z) + O(b),$$
$$EK_{h,i}^\theta\{\theta^T \frac{Z_{i0}}{b}\}^\ell = \tau_\ell f_\theta(\theta^T z) + J_\theta(z) + O(h),$$
$$EK_{h,i}^\theta\{\theta^T Z_{i0}\} r(\theta_0^T Z_i, \theta_0^T z, X_i) = O\{h(h + J_\theta(z))(\delta_\theta^2 + h^2)\}$$

uniformly for $\theta, \vartheta \in \Theta$ *with* $\theta \perp \vartheta$ *and* $z \in \mathcal{D}$, *where* $v_{k,\ell}^{\theta,\vartheta} = \int_{\mathbb{R}^q} H(U)(\theta^T U)^k (\vartheta^T U)^\ell dU$ *and* $\tau_\ell = \int K(u) u^\ell du$.

Proof. We here only give the details for the first and the third parts. If $d(z, \mathcal{D}^c) > a_0 b$, we define $J_0(z) = 0$. From (C6), we have

$$\int_{\mathcal{D}} H_b(U - z)\{\frac{\theta^T(U - z)}{b}\}^k\{\frac{\vartheta^T(U - z)}{b}\}^\ell f(U) dU$$
$$= \int_{\mathbb{R}^q} H(U)\{\theta^T U\}^k\{\theta^T U\}^\ell f(z + hU) dU = v_{k,\ell}^{\theta,\vartheta} f(z) + O(b).$$

If $d(z, \mathcal{D}^c) < a_0 b$, we have by (C3)

$$J_0(z) \triangleq \int_{\mathcal{D}} H_b(U - z)|\frac{\theta^T(U - x)}{b}|^k|\frac{\vartheta^T(U - x)}{b}|^\ell f(U) dU$$
$$\le \int_{\mathbb{R}^q} H(U)|\theta^T U|^k|\vartheta^T U|^\ell f(z + hU) dU = O(1).$$

Therefore, the first part of Lemma A.3 follows.

Let $\theta^T z = v_0$, $\theta_0^T z = v_0'$. Write $\theta_0 = b_n\theta + e_n\vartheta$, where $1 - b_n \sim \delta_\theta$ and $e_n \sim \delta_\theta$. Let $\mathcal{D}_\theta$ be the positive support of $f_\theta(v)$. Note that

$$|r(\theta_0^T Z_i, \theta_0^T z, X_i) \le c|X_i| \cdot |\theta_0^T Z_{i0}|^3 \le c|X_i|\{\delta_\theta^3 + |\theta^T Z_{i0}|^3\}. \qquad \text{(A.25)}$$

If $|\theta^T z - \mathcal{D}_\theta^c| < a_0' h$, then by (A.25)

$$E|K_{h,i}^\theta\{\theta^T Z_{i0}\} r(\theta_0^T Z_i, \theta_0^T z, X_i)| \le chE\{K_{h,i}^\theta|\theta^T \frac{Z_{i0}}{h}||X_i|(\delta_\theta^3 + |\theta^T Z_{i0}|^3)\}$$
$$= O\{hJ_\theta(z)(\delta_\theta^3 + h^3)\}. \qquad \text{(A.26)}$$

Let $\mathcal{X}(v_1, v_2) = E(X|\theta^T Z = v_1, \vartheta^T Z = v_2)$ and $r_0(v_1, v_2, v_0') = \{G(v_1) - G(v_0') - G'(v_0')(v_1 - v_0) - G''(v_0')(v_1 - v_0')^2/2\}^T \mathcal{X}(v_1, v_2)$. We have

$$\frac{\partial r_0}{\partial v_1} = \{G'(v_1) - G'(v_0') - G''(v_0')(v_1 - v_0')\}\mathcal{X}(v_1, v_2) + r_0(v_1, v_2, v_0')\frac{\partial}{\partial v_1}\mathcal{X}(v_1, v_2),$$
$$\frac{\partial r_0}{\partial v_2} = \{G(v_1) - G(v_0') - G'(v_0')(v_1 - v_0) - G''(v_0')\frac{(v_1 - v_0')^2}{2}\}\frac{\partial}{\partial v_2}\mathcal{X}(v_1, v_2).$$

By (C2) and (C3), it follows that

$$\tilde{f}(v_0 + hv_1, v_2) = \tilde{f}(v_0, v_2) + O(h),$$
$$|r_0| \le c|v_1 - v_0'|^3, \quad |\frac{\partial r_0}{\partial v_1}| \le c|v_1 - v_0'|^2, \quad |\frac{\partial r_0}{\partial v_2}| \le c|v_1 - v_0'|^3.$$

Note that Z is bounded. We have

$$\begin{aligned} &|r_0(b_n v_0 + e_n v_2 + b_n v_1 h, v_0 + hv_1, v_0')\tilde{f}(v_0 + hv_1, v_2) \\ &\quad - r_0(b_n v_0 + e_n v_2, v_0, v_0')\tilde{f}(v_0, v_2)| \\ &\le c\{(\delta_\theta + h)^2 h\}, \end{aligned} \tag{A.27}$$

where $\tilde{f}(v_1, v_2)$ is the density function of $(\theta^T Z, \vartheta^T Z)$. If $|\theta^T z - D_\theta^c| > a_0' h$, we have $\int K(v_1)v_1 r_0(b_n v_0 + e_n v_2, v_0, v_0')\tilde{f}(v_0, v_2)dv_1 dv_2 = 0$. Hence

$$\begin{aligned} &|EK_{h,i}^\theta\{\theta^T Z_{i0}\} r(\theta_0^T Z_i, \theta_0^T z, X_i)| \\ &= |h \int f(v_1)v_1 r_0(b_n v_0 + e_n v_2 + b_n v_1 h, v_0 + hv_1, v_0')\tilde{f}(v_0 + hv_1, v_2)dv_1 dv_2| \\ &\le h \int K(v_1)|v_1| r_0(b_n v_0 + e_n v_2, v_0, v_0')\tilde{f}(v_0, v_2)dv_1 dv_2 + O\{h^2(\delta_\theta + h)^2\} \\ &= O\{h^2(\delta_\theta + h)^2\}. \end{aligned}$$

Therefore the third part of Lemma A.3 follows from the above equation and (A.26).

Lemma A.4. *Under assumptions* (C2) *and* (C5), *we have that* W_0 *is a semi-positive matrix with rank* $q - 1$.

Proof. Note that $\theta_0^T[G'(\theta_0{}^T Z)X\{Z - \mu_{\theta_0}(Z)\}] = 0$ almost surely. It follows that the rank of W_0 is not greater than $q-1$. To complete the proof, we need to show that for any vector $\vartheta \in \Theta$ such that $\vartheta^T\theta_0 = 0$,

$$\vartheta^T W_0 \vartheta > 0. \tag{A.28}$$

If $\vartheta^T W_0 \vartheta = 0$, i.e., $E[\{G'(\theta_0^T Z)X\}^2\{\vartheta^T Z - \vartheta^T\mu_{\theta_0}(Z)\}^2] = 0$, we have $\{G'(\theta_0^T Z)^T X\}\{\vartheta^T Z - \vartheta^T\mu_{\theta_0}(Z)\} \equiv 0$ almost surely. Because $P(G'(\theta_0^T Z)X = 0) = 0$ as assumed in (C5), we have $\vartheta^T Z - \vartheta^T\mu_{\theta_0}(Z) \equiv 0$ almost surely, which contradicts with the existence of the density function of Z in assumption (C2). Therefore (A.28) follows.

For ease of exposition, we abbreviate $\sup_{z\in\mathcal{D},\theta\in\Theta}|A_n(z,\theta)| = O(a_n)$ as $A_n(z, \theta) = O(a_n)$ in the following context.

Proof of Lemma 1. By Taylor expansion, write

$$y_i = \left(G^T(\theta_0^T z), G'^T(\theta_0^T z)\right)\begin{pmatrix} X_i \\ \theta^T Z_{i0} X_i \end{pmatrix} + R(Z_i, X_i, z, \theta) + \varepsilon_i,$$

where $R(Z_i, X_i, z, \theta) = G'^T(\theta_0^T z) X_i Z_{i0}^T(\theta_0 - \theta) + G''^T(\theta_0^T Z_i^*) X_i \{\theta_0^T Z_{i0}\}^2/2$. Note that this expansion is unique under the assumptions even $X \equiv Z$ with the assumption before Lemma 1. Let (a^T, d^T) be the value on the right hand side of (2.5) with Z_j replaced by z, and

$$C_n(z) = n^{-1} \sum_{i=1}^n H_{b,i} \begin{pmatrix} X_i \\ Z_{i0}^T \theta X_i \end{pmatrix} \begin{pmatrix} X_i \\ Z_{i0}^T \theta X_i \end{pmatrix}^T. \tag{A.29}$$

We have

$$\begin{pmatrix} a \\ d \end{pmatrix} = \begin{pmatrix} G(\theta_0^T z) \\ G'(\theta_0^T z) \end{pmatrix} + C_n^{-1}(z) n^{-1} \sum_{i=1}^n H_{b,i} \begin{pmatrix} X_i \\ Z_{i0}^T \theta X_i \end{pmatrix} \{R(Z_i, X_i, z, \theta) + \varepsilon_i\}. \tag{A.30}$$

Let $\pi(z) = E(XX^T | Z = z) f(z)$. For any ϑ, it follows from Lemmas A.1, A.3 and assumption (C1)-(C3) that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n H_{b,i} X_i X_i^T \{\theta^T Z_{i0}\}^k \{\vartheta^T Z_{i0}\}^\ell \\ &= \begin{cases} \pi(z)(\theta^T \vartheta)^k b^{k+\ell} + O\{b^{k+\ell}(\tau_{qn} + J_0(z))\}, & k = \ell = 0, 1, \\ \pi(z) b^{k+\ell} + O\{b^{k+\ell}(\tau_{qn} + J_0(z))\}, & k + \ell = 2, k \neq 1 \\ O(b^{k+\ell+1} + b^{k+\ell}(\tau_{qn} + J_0(z))\}, & k + \ell = 1, 3. \end{cases} \end{aligned}$$

It follows that on $\{f(z) \geq c_0\}$

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n H_{b,i} \begin{pmatrix} X_i \\ Z_{i0}^T \theta X_i \end{pmatrix} R(Z_i, X_i, z, \theta_0) \\ &= \begin{pmatrix} O\{b(b + J_0(z))\} \\ b^2 \theta^T(\theta_0 - \theta) \pi(z) G'(\theta_0^T z) + O\{b^2(b^2 + J_0(z))\} \end{pmatrix}, \end{aligned} \tag{A.31}$$

and

$$\{C_n(z)\}^{-1} = \begin{pmatrix} \pi^{-1}(z) + O\{\tau_{qn} + J_0(z)\} & O\{b^2 + b^{-1}(\delta_{qn} + J_0(z))\} \\ O\{b^2 + b^{-1}(\delta_{qn} + J_0(z))\} & b^{-2}\{\pi^{-1}(z) + J_0(z) + O(\delta_{qn})\} \end{pmatrix}. \tag{A.32}$$

By Lemma A.1 and assumption (C4), we have

$$\frac{1}{n} \sum_{i=1}^n H_{b,i} \begin{pmatrix} X_i \\ Z_{i0}^T \theta X_i \end{pmatrix} \varepsilon_i = \begin{pmatrix} O(\delta_{qn}) \\ O(b\tau_{qn}) \end{pmatrix}. \tag{A.33}$$

It follows from (A.30)$-$(A.33) that on $\{z : f(z) > c_0\}$,

$$\begin{pmatrix} a \\ d \end{pmatrix} = \begin{pmatrix} G(\theta_0^T z) \\ G'(\theta_0^T z) \end{pmatrix} + \begin{pmatrix} O\{\tau_{qn} + bJ_0(z)\} \\ \{\theta^T(\theta_0 - \theta)\} G'(\theta_0^T z) + O\{b^{-1}\tau_{qn} + J_0(z)\} \end{pmatrix}. \tag{A.34}$$

Write $r_{i0} = \{G(\theta_0^T z) - a\}^T X_i + \{G'(\theta_0^T z) - d\}^T X_i \{\theta_0^T Z_{i0}\}$ and r_{ij} the value of r_{i0} with z replaced by Z_j. By (A.34), we have

$$\begin{aligned} r_{i0} &= O(\tau_{qn} + bJ_0(z))|X_i| - \{\theta^T(\theta_0 - \theta)\}G'^T(\theta_0^T z)X_i Z_{i0}^T \theta_0 \\ &\quad + O(b^{-1}\tau_{qn} + J_0(z))|X_i| \cdot |Z_{i0}|. \end{aligned}$$

By Lemma A.1, for any d and d', we have

$$\frac{1}{n}\sum_{i=1}^n (d^T X_i X_i^T d') H_{b,i} Z_{i0} Z_{i0}^T = b^2 d^T \pi(z) d' f(x) I + O(b^2\tau_{qn} + b^2 J_0(z)), \quad \text{(A.35)}$$

$$\frac{1}{n}\sum_{i=1}^n (d^T X_i) H_{b,i} Z_{i0} |X_i| = O(b), \quad \frac{1}{n}\sum_{i=1}^n (d^T X_i) H_{b,i} Z_{i0} |X_i| \cdot |Z_{i0}| = O(b^2),$$

$$\frac{1}{n}\sum_{i=1}^n (d^T X_i) H_{b,i} Z_{i0} |Z_{i0}|^2 = O(b^3), \quad \frac{1}{n}\sum_{i=1}^n (d^T X_i) H_{b,i} Z_{i0} \varepsilon_i = O(b\delta_{qn}), \quad \text{(A.36)}$$

where I is the identity matrix. Thus

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n (d^T X_i) H_{b,i} Z_{i0} r_{i0} \\ &\quad = -b^2 d^T \pi(z) G'({\theta_0}^T z)\theta^T(\theta_0 - \theta)\theta_0 + O(b\tau_{qn} + b^2 J_0(z)). \end{aligned} \quad \text{(A.37)}$$

Note that by Lemmas A.1 and A.3,

$$\sup_{z\in\mathcal{D}} |n^{-1}\sum_{i=1}^n H_{b,i}(z) - f(z) - J_0(z)| = O(b + \delta_{qn}).$$

Therefore

$$\sup_{z\in\mathcal{D}} |\mathcal{I}(\bar{w}_j) - \mathcal{I}(f(z)) - \tilde{J}_0(z)| = O(b + \delta_{qn}), \quad \text{(A.38)}$$

where $\tilde{J}_0(z) = \mathcal{I}(f(z) + J_0(z)) - \mathcal{I}(f(z))$ satisfies (A.24). Write $\mathcal{I}(\bar{w}_j)$ as $\mathcal{I}_{nj}$. By (A.34), (A.35), (A.36), (A.37) and (A.38), we have

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj} \sum_{i=1}^n (d_j^T X_i)^2 H_{b,i}(Z_j) Z_{ij} Z_{ij}^T = b^2(\theta^T\theta_0)^2 C_0 I + O(b^2\tau_{qn} + b^3),$$

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj} \sum_{i=1}^n (d_j^T X_i) H_{b,i}(Z_j) Z_{ij} r_{ij} = b^2\theta^T\theta_0\theta^T(\theta_0 - \theta) C_0 I + O(b\tau_{qn}),$$

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj} \sum_{i=1}^n (d_j^T X_i) H_{b,i}(Z_j) Z_{ij} |Z_{ij}|^2 = O(b^3),$$

$$\frac{1}{n^2}\sum_{j=1}^n \mathcal{I}_{nj} \sum_{i=1}^n (d_j^T X_i) H_{b,i}(Z_j) Z_{ij} \varepsilon_i = O(b\delta_n),$$

where $C_0 = E\{\mathcal{I}(f(Z))f(Z)G'^T(\theta_0^T Z)X\}^2$. By (C3), write $y_i - a_j^T X_i = (d_j^T X_i) Z_{ij}^T\theta_0 + r_{ij} + O(|Z_{ij}|^2|X_i|) + \varepsilon_i$. By (2.6) and the foregoing four equations, if $\theta^T\theta_0 \neq 0$, we have

$$\begin{aligned}\tilde{\theta} &= \theta_0 + \Big\{\sum_{j=1}^n \mathcal{I}_{nj}\sum_{i=1}^n H_{b,i}(Z_j)(X_i^T d_j)^2 Z_{ij}Z_{ij}^T\Big\}^+ \\ &\quad\times\sum_{j=1}^n \mathcal{I}_{nj}\sum_{i=1}^n H_{b,i}(Z_j)(X_i^T d_j)Z_{ij}\{r_{ij}+\varepsilon_i\} \\ &= \theta_0 - \{\frac{\theta^T(\theta_0-\theta)}{(\theta^T\theta_0)}\}\theta_0 + O(b^{-1}\tau_{qn}) = (\theta^T\theta_0)^{-1}\theta_0 + O(b^{-1}\tau_{qn}).\end{aligned}$$

It follows that

$$\tilde{\theta} =: \frac{\text{sign}_1(\theta)\theta}{|\theta|} = \theta_0 + O(b^{-1}\tau_{qn}). \tag{A.39}$$

The proof of Lemma 1 is now completed.

Proof of Theorem 1. Let

$$\begin{aligned}R(X_i, Z_i, z, \theta) &= G'^T(\theta_0^T z)X_i Z_{i0}^T(\theta_0-\theta) + \frac{1}{2}G''^T(\theta_0^T z)X_i\{\theta_0^T Z_{i0}\}^2 \\ &\quad + r(\theta_0^T Z_i, \theta_0^T z, X_i).\end{aligned}$$

Write

$$y_i = \Big(G^T(\theta_0^T z), G'^T(\theta_0^T z)\Big)\begin{pmatrix}X_i\\ \theta^T Z_{i0}X_i\end{pmatrix} + R(X_i, Z_i, z, \theta) + \varepsilon_i.$$

Let $C_{\theta,n}(z)$ be the value of $C_n(z)$ in (A.29) with $H_{b,i}(Z_j)$ replaced by $K_{h,i}^\theta(Z_j)$ and

$$\begin{pmatrix}a_\theta\\ d_\theta\end{pmatrix} = \begin{pmatrix}G(\theta_0^T z)\\ G'(\theta_0^T z)\end{pmatrix} + C_{\theta,n}^{-1}(z)\sum_{i=1}^n K_{h,i}^\theta\begin{pmatrix}X_i\\ Z_{i0}^T\theta X_i\end{pmatrix}\{R(X_i, Z_i, z, \theta)+\varepsilon_i\}.$$

Let $\pi_{\theta 1}(z) = f_\theta(z)\pi_\theta'(z) - f_\theta'(\theta^T z)\pi_\theta(z)$. By Lemmas A.1, A.3 and assumptions (C1)−(C3), we have uniformly on $\mathcal{D}^\theta = \{z : f_\theta(z) > c_0\}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta X_i X_i^T = \pi_\theta(z)f_\theta(z) + O(\tau_n + J_\theta(z)),\\ &\frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta\{\theta^T Z_{i0}\}X_i X_i^T = \pi_{\theta 1}(z)h^2 + O(h\tau_n + hJ_\theta(z)),\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}\{\theta^T Z_{i0}\}^2 X_i X_i^T = \pi_\theta(z) f_\theta(z) h^2 + O(h^2\tau_n + h^2 J_\theta(z)),$$

$$C_{\theta,n}^{-1}(z) =$$

$$\begin{pmatrix} \{\pi_\theta(z) f_\theta(z)\}^{-1} + O(\tau_n + J_\theta(z)) & \pi_{\theta 2}(z) + O(h^{-1}\tau_n + h^{-1}J_\theta(z)) \\ \pi_{\theta 2}(z) + O(h^{-1}\tau_n + h^{-1}J_\theta(z)) & h^{-2}\{(\pi_\theta(z) f_\theta(z))^{-1} + O(\tau_n + J_\theta(z))\} \end{pmatrix},$$

where $\pi_{\theta 2}(z) = \{\pi_\theta(z) f_\theta(z)\}^{-1}\pi_{\theta 1}(z)\{\pi_\theta(z) f_\theta(z)\}^{-1}$. Let $V_\theta(z)$ is defined before Theorem 1 and $V_{\theta 1}(z) = f_\theta(\theta^T z)V_\theta'(z) - f_\theta'(\theta^T z)V_\theta(z)$. By Lemmas A.1 and A.3, we have

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} X_i\{G'(\theta_0^T z)X_i\}Z_{i0}^T(\theta_0 - \theta)$$
$$= f_\theta(\theta^T z)V_\theta(z)(\theta_0 - \theta) + O\{(\tau_n + J_\theta(z))\delta_\theta\},$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} X_i G''^T(\theta_0^T z)X_i\{\theta_0^T Z_{i0}\}^2$$
$$= f_\theta(\theta^T z)\pi_\theta(z)G''(\theta_0^T z)h^2 + O\{h^2(J_\theta(z) + \tau_n) + \delta_\theta^2\},$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} r(\theta_0^T Z_i, \theta_0^T z, X_i) = O\{\delta_\theta^2 + h^3\},$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}\{\theta^T Z_{i0}\}^k\{(\theta_0 - \theta)^T Z_{i0}\}^\ell X_i X_i^T$$
$$= \begin{cases} h^2\delta_\theta^\ell, & k = 2, \\ h^k(h + J_\theta(z) + \delta_n)\delta_\theta^\ell, & k = 1, 3, \end{cases}$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}\{\theta^T Z_{i0}\} r(\theta_0^T Z_i, \theta_0^T z, X_i)$$
$$= O\{h(h + J_\theta(z))(\delta_\theta^2 + h^2) + h\delta_n(\delta_\theta^2 + h^2)\}.$$

By Lemma A.1 and (C4), we have

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}\begin{pmatrix} X_i \\ Z_{i0}^T\theta X_i \end{pmatrix}\varepsilon_i = \begin{pmatrix} R_{3n}^\theta(z) + O(\tau_n\delta_n) \\ hR_{4n}^\theta(z) + O(h\tau_n\delta_n) \end{pmatrix},$$

where

$$R_{3n}^\theta(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(z)X_i\varepsilon_i, \quad R_{4n}^\theta(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(z)\{\frac{\theta^T Z_{i0}}{h}\}X_i\varepsilon_i.$$

By Lemma A.1, we have $R_{3n}^{\theta}(z) = O(\delta_n)$ and $R_{4n}^{\theta}(z) = O(\delta_n)$. We have on D^{θ},

$$\begin{aligned} a_\theta &= G(\theta_0^T z) + \frac{1}{2}G''(\theta_0^T z)h^2 + \pi_\theta^{-1}(z)V_\theta(z)(\theta_0 - \theta) + R_{3n}^{\theta}(z) \\ &\quad + O\{(h + J_\theta(z))\delta_\theta + h^2(h + J_\theta(z) + \delta_n) + \delta_\theta^2\}, \\ d_\theta &= G'(\theta_0^T z) + h^{-1}R_{4n}^{\theta}(z) + O\{\tau_n + h^{-1}(\delta_n + J_\theta(z))\delta_\theta\}. \end{aligned} \tag{A.40}$$

Let $a_{\theta,j}$ and $d_{\theta,j}$ be the values above with z replaced by Z_j. Write

$$y_i - a_{\theta,j}^T X_i = (d_{\theta,j}^T X_i)Z_{ij}^T\theta_0 + \Delta_{i,j}^{(\theta,0)} + \Delta_{i,j}^{(\theta,1)} + \Delta_{i,j}^{(\theta,2)} + r_{ij} - X_i^T R_{3n}^{\theta}(Z_j) + \varepsilon_i,$$

where $\Delta_{i,j}^{(\theta,0)} = X_i^T\pi_\theta^{-1}(z)V_\theta(z)(\theta - \theta_0)$, $\Delta_{i,j}^{(\theta,1)} = \{G'(\theta_0^T Z_j) - d_{\theta,j}\}^T X_i\{\theta_0^T Z_{ij}\}$, $\Delta_{i,j}^{(\theta,2)} = \{G''(\theta_0^T Z_j)\}^T X_i\{(\theta_0^T(Z_i - Z_j))^2 - h^2\}/2$ and $|r_{ij}| \le c\{|\theta_0^T Z_{ij}|^3 + (h + J_\theta(Z_j))\delta_\theta + h^2(h + J_\theta(Z_j) + \delta_n) + \delta_\theta^2\}|X_i|$. Note that by Lemmas A.1 and A.3,

$$\sup_{z\in\mathcal{D}} |\hat{f}_\theta(z) - f_\theta(z) - J_\theta(z)| = O(h + \delta_n),$$

where $\hat{f}_\theta(z) = n^{-1}\sum_{i=1}^n K_{h,i}^{\theta}(z)$. Therefore

$$\sup_{z\in\mathcal{D}} |\mathcal{I}(\hat{f}_\theta(z)) - \mathcal{I}(f_\theta(z)) - J_\theta(z)| = O(b + \delta_n). \tag{A.41}$$

Write $\mathcal{I}(\hat{f}_\theta(z))$ as $\mathcal{I}_{nj}^{\theta}$. We have,

$$\begin{aligned} \theta &= \theta_0 + D_{\theta,n}^{+}\sum_{j=1}^n \mathcal{I}_{nj}^{\theta}\sum_{i=1}^n (d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}\{\Delta_{i,j}^{(\theta,0)} + \Delta_{i,j}^{(\theta,1)} \\ &\quad + \Delta_{i,j}^{(\theta,2)} + r_{ij} - X_i^T R_{3n}^{\theta}(Z_j) + \varepsilon_i\}, \end{aligned} \tag{A.42}$$

where $D_{\theta,n} = n^{-2}\sum_{j=1}^n \mathcal{I}_{nj}^{\theta}\sum_{i=1}^n (d_{\theta,j}^T X_i)^2 K_{h,i}^{\theta}(Z_j)Z_{ij}Z_{ij}^T$. By (A.40), we have $d_\theta = G'(\theta_0^T z) + O\{h^{-1}\delta_n + (1 + h^{-1}J_\theta(z))\delta_\theta\}$. Exchanging the order of summation, we have by Lemma A.1

$$\begin{aligned} D_{\theta,n} &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{I}_{nj}^{\theta}\{d_{\theta,j}^T X_i\}^2 K_{h,i}^{\theta}(Z_j)Z_{ij}Z_{ij}^T \\ &= \frac{1}{n}\sum_{i=1}^n \mathcal{I}_f^{\theta}(Z_i)\{G'^T(\theta_0^T Z_i)X_i\}^2\{Z_i - \mu_\theta(Z_i)\}\{Z_i - \mu_\theta(Z_i)\}^T \\ &\quad + \frac{1}{n}\sum_{i=1}^n \mathcal{I}_f^{\theta}(Z_i)\{G'^T(\theta_0^T Z_i)X_i\}^2 E\{(Z_i - \mu_\theta(Z_i))(Z_i - \mu_\theta(Z_i))^T\} \\ &\quad + O(h^{-1}\delta_n + h + \delta_\theta) \\ &= W_0 + U_0 + O(h^{-1}\delta_n + h + \delta_\theta), \end{aligned}$$

where $\mathcal{I}_f^{\theta}(z) = \mathcal{I}(f_\theta(z))f_\theta(z)$. By Lemmas A.1 and A.3, we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}\Delta_{ij}^{(\theta,0)}$$
$$= E\{\mathcal{I}_f^{\theta}(Z)V_\theta(Z)\pi_\theta^{-1}(Z)V_\theta(Z)\}(\theta-\theta_0) + O(h^{-1}\tau_n\delta_\theta + \delta_\theta^2),$$
$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}\Delta_{ij}^{(\theta,1)} = O(h^{-1}\tau_n\delta_\theta + h\tau_n + \delta_\theta^2).$$

For any d and d' we have by Lemmas A.1 and A.3

$$\frac{1}{n}\sum_{i=1}^{n} d^T X_i X_i^T d' K_{h,i}^{\theta} Z_{i0}(\theta_0^T Z_{i0})^2 = \psi_\theta(z)h^2 + O\{h^2(J_\theta(z)+\tau_n) + h\delta_\theta + \delta_\theta^2\},$$
$$\frac{1}{n}\sum_{i=1}^{n} d^T X_i X_i^T d' K_{h,i}^{\theta} Z_{i0} = \psi_\theta(z) + O\{J_\theta(z)+\tau_n\}.$$

where $\psi_\theta(z) = f_\theta(z)E(d^T X_i X_i^T d' Z_{i0}|\theta^T Z = \theta^T z)$. Therefore

$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}\Delta_{ij}^{(\theta,2)} = O\{h^3 + h\delta_\theta + \delta_\theta^2\},$$
$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}r_{ij} = O\{h^3 + \delta_\theta^2 + h\delta_\theta + h\delta_n\}.$$

Let $\tilde{V}_\theta(z) = \mathcal{I}^{\theta}(z)\{G'(\theta_0^T Z_i)\}^T X_i\{\mu_\theta(Z_i) - z\}$. Note that

$$\frac{1}{n}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}$$
$$= \tilde{V}_\theta(Z_i) + \frac{1}{n}\sum_{j=1}^{n}\{\mathcal{I}_{nj}^{\theta}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij} - \tilde{V}_\theta(Z_i)\}.$$

Exchanging the order of the summation, by Lemmas A.1 and A.2 we have,

$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^T X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}\varepsilon_i$$
$$= \frac{1}{n}\sum_{i=1}^{n}\tilde{V}_\theta(Z_i)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n\delta_\theta)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\tilde{V}_{\theta_0}(Z_i)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2 + h^{-1}\tau_n\delta_\theta).$$

Similarly, we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\mathcal{I}_{nj}^{\theta}\sum_{i=1}^{n}(d_{\theta,j}^{T}X_i)K_{h,i}^{\theta}(Z_j)Z_{ij}X_i^{T}R_{3n}^{\theta}(Z_j)=O(h^3+h^{-1}\delta_n^2+h^{-1}\tau_n\delta_\theta).$$

Therefore

$$\theta=\theta_0+\{W_0+U_0\}^{-}E\{\mathcal{I}_f^{\theta}(Z)V_{\theta_0}(Z)\pi_{\theta_0}^{-1}(Z)V_{\theta_0}(Z)\}(\theta-\theta_0)$$
$$+n^{-1}\{W_0+U_0\}^{-}\sum_{i=1}^{n}\tilde{V}_{\theta_0}(z)\varepsilon_i+O(h^3+h^{-1}\delta_n^2+h^{-1}\tau_n\delta_\theta+\delta_\theta^2).$$

Let $D=(W_0+U_0)^{-1/2}E\{\mathcal{I}_f^{\theta}(Z)V_{\theta_0}(Z)\pi_{\theta_0}^{-1}(Z)V_{\theta_0}(Z)\}(W_0+U_0)^{-1/2}$. By the Schwarz's inequality, we have $W_0+U_0-E\{\mathcal{I}_f^{\theta}(Z)V_{\theta_0}(Z)\pi_{\theta_0}^{-1}(Z)V_{\theta_0}(Z)\}$ is a semi-positive matrix. We have, by Lemma A.4, the eigenvalues of D are less than 1. There are $1>\lambda_1\geq\lambda_2\geq\cdots,\lambda_{q-1}\geq 0$ and an orthogonal matrix Γ such that

$$D=\Gamma diag(\lambda_1,\ldots,\lambda_{q-1},0)\Gamma^T.$$

Let $\beta_k=(W_0+U_0)^{-1/2}(\theta_k-\theta_0)$. We have

$$\beta_{k+1}=\Gamma diag(\lambda_1,\ldots,\lambda_{p+q-1},0)\Gamma^T\beta_k+n^{-1}\{W_0+U_0\}^{-\frac{1}{2}}\sum_{i=1}^{n}\tilde{V}_{\theta_0}(z)\varepsilon_i$$
$$+O(h^3+h^{-1}\delta_n^2+h^{-1}\tau_n\Delta_k+\Delta_k^2), \tag{A.43}$$

where $\Delta_k=|\beta_k|$. It follows that

$$\Delta_{k+1}\leq\lambda_1\Delta_k+\delta_{0n}+c(\Delta_k+h^{-1}\tau_n)\Delta_k+c(h^3+h^{-1}\delta_n^2)$$
$$=\delta_{0n}+\{\lambda_1+c\Delta_k+c(h+h^{-1}\delta_n)\}\Delta_k+c(h\tau_n+h^{-1}\delta_n^2) \tag{A.44}$$

almost surely, where c is a constant. We can further take $c>1$. For sufficiently large n, we may assume that

$$c(h+h^{-1}\delta_n)\leq\frac{1-\lambda_1}{3},\quad \delta_{0n}+c(h\tau_n+h^{-1}\delta_n^2)\leq\frac{(1-\lambda_1)^2}{9c}. \tag{A.45}$$

Since by (A.39) $\Delta_1\to 0$ almost surely, we may assume

$$\Delta_1\leq\frac{1-\lambda_1}{3c}. \tag{A.46}$$

Therefore, it follows that from (A.44), (A.45) and (A.46)

$$\Delta_2\leq\{\lambda_1+\frac{2}{3}(1-\lambda_1)\}\frac{1-\lambda_1}{3c}+\frac{(1-\lambda_1)^2}{9c}=\frac{1-\lambda_1}{3c}. \tag{A.47}$$

From (A.44), (A.45) and (A.47), we have that

$$\Delta_3 \leq \frac{1-\lambda_1}{3c}.$$

Consequently, $\Delta_k \leq (1-\lambda_1)/(3c)$ for all k. Therefore we have from (A.44) that

$$\Delta_{k+1} \leq \lambda_0 \Delta_k + \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)$$

almost surely, where $0 \leq \lambda_0 < (2+\lambda_1)/3 < 1$. It follows that

$$\Delta_k \leq \lambda_0^k \Delta_1 + \{\delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)\} \sum_{j=1}^{k} \lambda_0^j = O(\delta_{0n} + h\tau_n + h^{-1}\delta_n^2),$$

for sufficiently large k. By (A.43), we have

$$\begin{aligned} &\{W_0 + U_0\}^{\frac{1}{2}}(\hat{\theta} - \theta_0) \\ &\quad = D(\hat{\theta} - \theta_0) + n^{-1}\{W_0 + U_0\}^{-\frac{1}{2}} \sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2). \end{aligned} \tag{A.48}$$

It follows from (A.48) that

$$W_1(\hat{\theta} - \theta_0) = n^{-1} \sum_{i=1}^{n} \tilde{V}_{\theta_0}(z)\varepsilon_i + O(h^3 + h^{-1}\delta_n^2).$$

We have completed the proof of the first part of Theorem 1.

References

Bradley, R. C. (1983). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics: A Survey of Resent Results* (Edited by E. Eberlein and M. S. Taqqu), 165-192. Birhauser, Boston.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications.* Academic Press, New York.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.* **16**, 1428-1449.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Rio, E. (1995). The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Probab.* **23**, 1188-1203.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94**, 1275-1285.