Optimal Smoothing for a Computationally and Statistically Efficient Single Index Estimator

Yingcun Xia
(National University of Singapore)

Wolfgang Härdle
(Humboldt-Universität zu Berlin)

Oliver Linton
(London School of Economics)

DP No: EM 2009 537

2009

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel: 020 7955 6674

Abstract

In semiparametric models it is a common approach to under-smooth the nonparametric functions in order that estimators of the finite dimensional parameters can achieve root-*n* consistency. The requirement of under-smoothing may result as we show from inefficient estimation methods or technical difficulties. Based on local linear kernel smoother, we propose an estimation method to estimate the single-index model without under-smoothing. Under some conditions, our estimator of the single-index is asymptotically normal and most efficient in the semi-parametric sense. Moreover, we derive higher expansions for our estimator and use them to define an optimal bandwidth for the purposes of index estimation. As a result we obtain a practically more relevant method and we show its superior performance in a variety of applications.

Key words and phrases: ADE; Asymptotics; Bandwidth; MAVE method; Semi-parametric efficiency.

1 Introduction

Single index models (SIMs) are widely used in the applied quantitative sciences. Although the context of applications for SIMs almost never prescribes the functional or distributional form of the involved statistical error, the SIM is commonly fitted with (low dimensional) likelihood principles. Both from a theoretical and practical point of view such fitting approach has been criticized and has led to semiparametric modelling. This approach involves high dimensional parameters (nonparametric functions) and a finite dimensional index parameter. Consider the following single-index model,

$$Y = g(\theta_0^\top X) + \varepsilon, \tag{1}$$

where $E(\varepsilon|X) = 0$ almost surely, g is an unknown link function, and θ_0 is a single-index parameter with length one and first element positive for identification. In this model there is a single linear combination of covariates X that can capture most information about the relation between response variable Y and covariates X, thereby avoiding the "curse of dimensionality". Estimation of the single-index model is very attractive both in theory and in practice. In the last decade a series of papers has considered estimation of the parametric index and the nonparametric part with focus on root-n estimability and efficiency issues, see Carroll, Fan, Gijbels and Wand (1997) for an overview. There are numerous methods proposed or can be used for the estimation of the model. Amongst them, the most popular ones are the average derivative estimation (ADE) method investigated by Härdle and Stoker (1989), the sliced inverse regression (SIR) method proposed by Li (1989); the semiparametric least squares (SLS) method of Ichimura (1993) and the simultaneous minimization method of Härdle, Hall and Ichimura (1993).

The existing estimation methods are all subject to some or other of the following four critiques: (1) *Heavy computational burden:* see, for example, Härdle, Hall and Ichimura (1993), Delecroix, Härdle and Hristache (2003), Xia and Li (1999) and Xia *et al.* (1999). These methods include complicated optimization techniques (iteration between bandwidth choice and parameter estimation) for which no simple and effective algorithm is available up to now. (2) *Strong restrictions on link functions or design of covariates X:* Li (1991) required the covariate to have a symmetric distribution; Härdle and Stoker (1989) and Hristache *et al.* (2001) needed a non-symmetric structure for the link function, i.e., $|Eg'(\theta_0^\top X)|$ is bounded away from 0. If these conditions are violated, the corresponding methods are inconsistent. (3) *Inefficiency:* The ADE method of Härdle and Stoker (1989) or the improved ADE method of Hristache *et al.* (2001) is not asymptotically efficient in the semi-parametric sense, Bickel *et al.* (1993). Nishiyama and Robinson (2000,

2005) considered the Edgeworth correction to the ADE methods. Härdle and Tsybakov (1993) discussed the sensitivity of the ADE. Since this method involves high dimensional smoothing and derivative estimation, its higher order properties are poor. (4) *Under-smoothing:* Let h_g^{opt} be the optimal bandwidth in the sense of MISE for the estimation of link function g and let h_θ be the bandwidth used for the estimation of θ_0. Most of the methods mentioned above require the bandwidth h_θ to be much smaller than the bandwidth h_g^{opt}, i.e. $h_\theta/h_g^{opt} \to 0$ as $n \to \infty$, in order that estimators of θ_0 can achieve root-n consistency, see, Härdle, and Stoker (1989) and Hristache *et al.* (2002), Robinson (1988), Hall (1989) and Carroll *et al.* (1997) among others. Due to technical complexities, there are few investigations about how to select the bandwidth h_θ for the estimation of the single-index. Thus it could be the case that even if $h_\theta = h_g^{opt}$ allows for root-n consistent estimation of θ, that $h_\theta^{opt}/h_g^{opt} \to 0$ or $h_g^{opt}/h_\theta^{opt} \to 0$, where h_θ^{opt} is the optimal bandwidth for estimation of θ. This would mean that using a single bandwidth h_g^{opt} would result in suboptimal performance for the estimator of θ. Higher order properties of other semiparametric procedures have been studied in Linton (1995) *inter alia*.

Because the estimation of θ_0 is based on the estimation of the link function g, we might expect that a good bandwidth for the link function should be a good bandwidth for the single-index, i.e., under-smoothing should be unnecessary. Unfortunately, most of the existing estimation methods involve for technical reason "under-smoothing" the link function in order to obtain a root-n consistent estimator of θ_0. See, for example, Härdle and Stoker (1989), Hristache *et al.* (2001, 2002), Carroll *et al.* (1997) and Xia and Li (1999). Härdle, Hall and Ichimura (1993) investigated this problem for the first time and proved that the optimal bandwidth for the estimation of the link function in the sense of MISE can be used for the estimation of the single-index to achieve root-n consistency. As mentioned above, for its computational complexity the method of Härdle, Hall and Ichimura (1993) is hard to implement in practice.

This paper presents a method of joint estimation of the parametric and nonparametric parts. It avoids undersmoothing and the computational complexity of former procedures and achieves the semiparametric efficiency bound. It is based on the MAVE method of Xia et al (2002), which we outline in the next section. Using local linear approximation and global minimization, we give a very simple iterative algorithm. The proposed method has the following advantages: (i) the algorithm only involves one-dimensional smoothing and is proved to converge at a geometric rate; (ii) with normal errors in the model, the estimator of θ_0 is asymptotically normal and efficient in the semiparametric sense; (iii) the optimal bandwidth for the estimation of the link function in the sense of MISE can be used to estimate θ_0 with root-n consistency;

(iv) by a second order expansion, we further show that the optimal bandwidth for the estimation of the single-index θ_0, h_θ^{opt}, is of the same magnitude as h_g^{opt}.

Therefore, the commonly used "under-smoothing" approach is inefficient in the sense of second order approximation. Powell and Stoker (1996) investigated bandwidth selection for the ADE methods. We also propose an automatic bandwidth selection method for our estimator of θ. Xia (2006) has recently shown the first order asymptotic properties of this method. Our theoretical results are proven under weak moment conditions.

In section 3 we present our main results. We show the speed of convergence, give the asymptotic estimation and derive a smoothing parameter selection procedure. In the following section we investigate the proposed estimator in simulation and application. Technical details are deferred to the appendix.

2 The MAVE method

Suppose that $\{X_i, Y_i : i = 1, 2, \ldots, n\}$ is a random sample from model (1). The basic idea of our estimation method is to linearly approximate the smooth link function g and to estimate θ_0 by minimizing the overall approximation errors. Xia et al (2002) proposed a procedure via the so called minimum average conditional variance estimation (MAVE). The single index model (1) is a special case of what they considered, and we can estimate it as follows. Assuming function g and parameter θ_0 are known, then the Taylor expansion of $g(\theta_0^\top X_i)$ at $g(\theta_0^\top x)$ is

$$g(\theta_0^\top X_i) \approx a + d\theta_0^\top (X_i - x),$$

where $a = g(\theta_0^\top x)$ and $d = g'(\theta_0^\top x)$. With fixed θ, the local estimator of the conditional variance is then

$$\sigma_n^2(x|\theta) = \min_{a,d} \{n\hat{f}_\theta(x)\}^{-1} \sum_{i=1}^{n} [Y_i - \{a + d\theta^\top (X_i - x)\}]^2 K_h\{\theta^\top (X_i - x)\},$$

where $\hat{f}_\theta(x) = n^{-1} \sum_{i=1}^n K_h\{\theta^\top (X_i - x)\}$, where K is a univariate density function, h is the bandwidth and $K_h(u) = K(u/h)/h$; see Fan et al (1996). The value $\sigma_n^2(x|\theta)$ can also be understood as the local departure of Y_i with X_i close to x from a local linear model with given θ. Obviously, the best approximation of θ should minimize the overall departure at all $x = X_j, j = 1, \cdots, n$. Thus, our estimator of θ_0 is to minimize

$$Q_n(\theta) = \sum_{j=1}^{n} \sigma_n^2(X_j|\theta) \tag{2}$$

with respect to $\theta : |\theta| = 1$. This is the so-called minimum average conditional variance estimation (MAVE) in Xia et al (2002). In practice it is necessary to include some trimming in covariate regions where density is low, so we weight $\sigma_n^2(X_j|\theta)$ by a sequence $\hat{\rho}_j^\theta$, where $\hat{\rho}_j^\theta = \rho_n\{\hat{f}_\theta(X_j)\}$, that is discussed further below.

The corresponding algorithm can be stated as follows. Suppose θ_1 is an initial estimate of θ_0. Set the number iteration $\tau = 1$ and bandwidth h_1. We also set a final bandwidth h. Let $X_{ij} = X_i - X_j$.

Step 1: With bandwidth h_τ, calculate $\hat{f}_\theta(X_j) = n^{-1}\sum_{i=1}^n K_{h_\tau}(\theta^\top X_{ij})$ and the solutions of a_j and d_j to the inner problem in (2)

$$\begin{pmatrix} a_j^\theta \\ d_j^\theta h_\tau \end{pmatrix} = \Big\{ \sum_{i=1}^n K_{h_\tau}(\theta^\top X_{ij}) \begin{pmatrix} 1 \\ \theta^\top X_{ij}/h_\tau \end{pmatrix} \begin{pmatrix} 1 \\ \theta^\top X_{ij}/h_\tau \end{pmatrix}^\top \Big\}^{-1} \sum_{i=1}^n K_{h_\tau}(\theta^\top X_{ij}) \begin{pmatrix} 1 \\ \theta^\top X_{ij}/h_\tau \end{pmatrix} Y_i.$$

Step 2: Fix the weight $K_{h_\tau}(\theta^\top X_{ij}), f_\theta(X_j), a_j^\theta$ and d_j^θ. Calculate the solution of θ to (2)

$$\theta = \{ \sum_{i,j=1}^n K_{h_\tau}(\theta^\top X_{ij})\hat{\rho}_j^\theta \{d_\theta(X_j)\}^2 X_{ij}X_{ij}^\top \hat{f}_\theta(\theta^\top X_j)\}^{-1} \sum_{i,j=1}^n K_{h_\tau}(\theta^\top X_{ij})\hat{\rho}_j^\theta d_\theta(X_j) X_{ij}(y_i - a_j^\theta)/\hat{f}_\theta(\theta^\top X_j),$$

where $\hat{\rho}_j^\theta = \rho_n\{\hat{f}_\theta(X_j)\}$.

Step 3: Set $\tau = \tau + 1$, $\theta := \theta/|\theta|$ and $h_\tau := \max\{h, h_\tau/\sqrt{2}\}$, go to Step 1.

Repeat steps 1 and 2 until convergence.

The iteration can be stopped by the common rule. For example, if the calculated θ's are stable at a certain direction, we can stop the iteration. The final vector $\theta := \theta/|\theta|$ is the MAVE estimator of θ_0, denoted by $\hat{\theta}$. Note that these steps are an explicit algorithm of the Xia et al (2002) method for the single-index model with some version of what the called 'refined kernel weighting' and boundary trimming. Similar to the other direct estimation methods, the calculation above is easy to implement. See Horowitz and Härdle (1996) for more discussions. After θ is estimated, the link function can be then estimated by the local linear smoother as $g^{\hat{\theta}}(v)$, where

$$\hat{g}^\theta(v) = [n\{s_2^\theta(v)s_0^\theta(v) - (s_1^\theta(v))^2\}]^{-1} \sum_{i=1}^n \{s_2^\theta(v) - s_1^\theta(v)(\theta^\top X_i - v)/h_\tau\} K_{h_\tau}(\theta^\top X_i - v) Y_i, \tag{3}$$

and $s_k^\theta(v) = n^{-1}\sum_{i=1}^n K_{h_\tau}(\theta^\top X_i - v)\{(\theta^\top X_i - v)/h_\tau\}^k$ for $k = 0, 1, 2$. Actually, $\hat{g}^{\hat{\theta}}(v)$ is the final value of a_j^θ in Step 1 with $\theta^\top X_j$ replaced by v.

In the algorithm, $\rho_n(.)$ is a trimming function employed to handle the boundary points. There are many choices for the estimator to achieve the root-n consistency; see e.g. Härdle and Stocker (1989) and HHI

(1993). However, to achieve the efficiency bound, $\rho_n(v)$ must tend to 1 for all v. In this paper, we take $\rho_n(v)$ as a bounded function with third order derivatives on $\mathbb{R}$ such that $\rho_n(v) = 1$ if $v > 2c_0 n^{-\varsigma}$; $\rho_n(v) = 0$ if $v \leq c_0 n^{-\varsigma}$ for some constants $\varsigma > 0$ and $c_0 > 0$. As an example, we can take

$$\rho_n(v) = \begin{cases} 1, & \text{if } v \geq 2c_0 n^{-\varsigma}, \\ \frac{\exp\{(2c_0 n^{-\varsigma} - v)^{-1}\}}{\exp\{(2c_0 n^{-\varsigma} - v)^{-1}\} + \exp\{(v - c_0 n^{-\varsigma})^{-1}\}}, & \text{if } 2c_0 n^{-\varsigma} > v > c_0 n^{-\varsigma}, \\ 0, & \text{if } v \leq c_0 n^{-\varsigma}. \end{cases} \tag{4}$$

The choice of ς will be given below.

3 Main Results

We impose the following conditions to obtain the asymptotics of the estimators.

[(C1)] [Initial estimator] The initial estimator is in $\Theta_n = \{\theta : |\theta - \theta_0| \leq n^{-\alpha}\}$ for some $0 < \alpha \leq 1/2$.

[(C2)] [Design] The density function $f_\theta(v)$ of $\theta^\top X$ and its derivatives up to 6th order are bounded on $\mathbb{R}$ for all $\theta \in \Theta_n$, $E|X|^6 < \infty$ and $E|Y|^3 < \infty$. Furthermore, $\sup_{v \in R, \theta \in \Theta_n} |f_\theta(v) - f_{\theta_0}(v)| \leq c|\theta - \theta_0|$ for some constant $c > 0$.

[(C3)] [Link function] The conditional mean $g_\theta(v) = E(Y|\theta^\top X = v), E(X|\theta^\top X = v)$, $E(XX^\top|\theta^\top X = v)$ and their derivatives up to 6th order are bounded for all $\theta : |\theta - \theta_0| < \delta$ where $\delta > 0$.

[(C4)] [Kernel function] $K(v)$ is a symmetric density function with finite moments of all orders.

[(C5)] [Bandwidth and trimming parameter] Trimming parameter $\varsigma \leq 1/20$ and bandwidth $h \propto n^{-\rho}$ for some ρ with $1/5 - \epsilon \leq \rho \leq 1/5 + \epsilon$ for some $\epsilon > 0$.

Assumption (C1) is feasible because such an initial estimate is obtainable using existing methods, such as Härdle and Stoker (1989), Powell et al. (1989) and Horowitz and Härdle (1996). Actually, Härdle, Hall and Ichimura (1993) even assumed that the initial value is in a root-n neighborhood of θ_0, $\{\theta : |\theta - \theta_0| \leq C_0 n^{-1/2}\}$. Assumption (C2) means that X may have discrete components providing that $\theta^\top X$ is continuous for θ in a small neighborhood of θ_0; see also Ichimura (1993). The moment requirement on X is not strong. Härdle, Hall and Ichimura (1993) obtained their estimator in a bounded area of $\mathbb{R}^p$, which is equivalent to assume that X is bounded; see also Härdle and Stoker (1989). We impose slightly higher order moment requirement than second moment for Y to ensure the optimal bandwidth in (C5) can be used in applying Lemma 6.1 in

section 6. The smoothness requirements on the link function in (C3) can be relaxed to the existence of a bounded second order derivative at the cost of more complicated proofs and smaller bandwidth. Assumption (C4) includes the Gaussian kernel and the quadratic kernel. Assumption (C5) includes the commonly used optimal bandwidth in both the estimation of the link function and the estimation of the index θ_0. Actually, imposing these constraints on the bandwidth is for ease of exposition in the proofs.

Let $\mu_\theta(x) = E(X|\theta^\top X = \theta^\top x)$, $\nu_\theta(x) = \mu_\theta(x) - x$, $w_\theta(x) = E(XX^\top|\theta^\top X = \theta^\top x)$, $W_0(x) = \nu_{\theta_0}(x)\nu_{\theta_0}(x)$. Let A^+ denote the Moore-Penrose inverse of a symmetric matrix A. Recall that K is a symmetric density function. Thus, $\int K(v)dv = 1$ and $\int vK(v)dv = 0$. For ease of exposition, we further assume that $\mu_2 = \int v^2K(v)dv = 1$. Otherwise, we can redefine $K(v) := \mu_2^{1/2}K(\mu_2^{1/2}v)$.

We have the following asymptotic results for the estimators.

Theorem 3.1 (Speed of algorithm) *Let θ_τ be the value calculated in Step 3 after τ iterations. Suppose assumptions (C1)-(C5) hold. If $h_\tau \to 0$ and $|\theta_\tau - \theta_0|/h_\tau^2 \to 0$, we have*

$$\theta_{\tau+1} - \theta_0 = \frac{1}{2}\{(I - \theta_0\theta_0^\top) + o(1)\}(\theta_\tau - \theta_0) + \frac{1}{2\sqrt{n}}N_n + O(n^{2\varsigma}h_\tau^4)$$

almost surely, where $N_n = [E\{g'(\theta_0^\top X)^2W_0(X)\}]^+ n^{-1/2}\sum_{i=1}^n g'(\theta_0^\top X_i)\nu_{\theta_0}(X_i)\varepsilon_i = O_p(n^{-1/2})$.

Theorem 3.1 indicates that the algorithm converges at a geometric rate, i.e. after each iteration, the estimation error reduces by half approximately. By Theorem 3.1 and the bandwidth requirement in the algorithm, we have

$$|\theta_{\tau+1} - \theta_0| \quad = \quad \{\frac{1}{2} + o(1)\}|\theta_{\tau+1} - \theta_0| + O(n^{-1/2} + n^{2\varsigma}h_\tau^4).$$

Starting with $|\theta_1 - \theta_0| = Cn^{-\alpha}$, in order to achieve root-n consistency, say $|\theta_k - \theta_0| \le cn^{-1/2}$ i.e. $2^{-k}Cn^{-\alpha} \le cn^{-1/2}$, the number of iterations k can be calculated roughly by

$$k = \{(\frac{1}{2} - \alpha)\log n + \log(C/c)\}/\log 2. \tag{5}$$

Based on Theorem 3.1, we immediately have the following limiting distribution.

Theorem 3.2 (Efficiency of estimator) *Under the conditions (C1)-(C5), we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Sigma_0),$$

where $\Sigma_0 = [E\{g'(\theta_0^\top X)^2W_0(X)\}]^+ E\{g'(\theta_0^\top X)^2W_0(X)\varepsilon^2\}[E\{g'(\theta_0^\top X)^2W_0(X)\}]^+$.

By choosing a similar trimming function, the estimators in Härdle, Hall and Ichimura (1993) and Ichimura (1993) have the same asymptotic covariance matrix as Theorem 3.2. If we further assume that the conditional distribution of Y given X belongs to a canonical exponential family

$$f_{Y|X}(y|x) = \exp\{y\eta(x) - \mathcal{B}(\eta(x)) + \mathcal{C}(y)\}$$

for some known functions $\mathcal{B}$, $\mathcal{C}$ and η, then Σ_0 is the lower information bound in the semiparametric sense (Bickel, Klaassen, Ritov and Wellner, 1993). See also the proofs in Carroll, Fan, Gijbels and Wand (1997) and Härdle, Hall and Ichimura (1993). In other words, our estimator is the most efficient in the semiparametric sense.

For the estimation of the single-index model, it was generally believed that undersmoothing the link function must be employed in order to allow the estimator of the parameters to achieve root-n consistency. However, Härdle, Hall and Ichimura (1993) established that undersmoothing the link function is not necessary. They derived an asymptotic expansion of the sum of squared residuals. We also derive an asymptotic expansion but of the estimator $\widehat{\theta}$ itself. This allows us to measure the higher order cost of estimating the link function. We use the expansion to propose an automatic bandwidth selection procedure for the index. Let $f_{\theta_0}(.)$ be the density function of $\theta_0^\top X$.

Theorem 3.3 (Higher Order Expansion) *Under conditions (C1)-(C5) and ε_i is independent of X_i, we have almost surely*

$$\hat{\theta} - \theta_0 = \mathcal{E}_n + \frac{c_{1,n}}{nh} + c_{2,n}h^4 + \mathcal{H}_n + O\{n^{2\varsigma}\gamma_n^3\},$$

where $\gamma_n = h^2 + (nh/\log n)^{-1/2}$,

$$\mathcal{E}_n = (W_n)^+ \sum_{i=1}^n \rho_n\{f_{\theta_0}(X_j)\} g'(\theta_0^\top X_i)\nu_{\theta_0}(\theta_0^\top X_i)\varepsilon_i,$$

with $W_n = n^{-1}\sum_{j=1}^n \rho_n\{f_{\theta_0}(X_j)\}(g'(\theta_0^\top X_i))^2\nu_{\theta_0}(X_j)\nu_{\theta_0}^\top(X_j)$, $\mathcal{H}_n = O\{n^{-1/2}\gamma_n + n^{-1}h^{-1/2}\}$ *with* $E\{\mathcal{H}_n\mathcal{E}_n\}$ $= o\{(nh)^{-2} + h^8\}$ *and*

$$c_{1,n} = \int K^2(v)v^2 dv\sigma^2(nW_n)^{-1}\sum_{j=1}^n \rho_n\{f_\theta(X_j)\}\{\nu'_{\theta_0}(X_j) + f'_0(X_j)\nu_{\theta_0}(X_j)/f_{\theta_0}(X_j)\},$$

$$c_{2,n} = \frac{1}{4}(\int K(v)v^4 dv - 1)(nW_n)^{-1}\sum_{j=1}^n \rho_n\{f_\theta(X_j)\}g'(\theta_0^\top X_j)g''(\theta_0^\top X_j)\nu''_{\theta_0}(X_j).$$

Because $K(v)$ is a density function and we constrain that $\int v^2 K(v) = 1$, it follows that $\mu_4 = \int K(v)v^4 dv > 1$. In the expansion of $\hat{\theta} - \theta_0$, the first term $\mathcal{E}_n$ does not depend on h. The second and third terms are the leading term among the remainders. The higher order properties of this estimator are better than those of the AD method, see Nishiyama and Robinson (2000), and indeed do not reflect a curse of dimensionality.

To minimize the stochastic expansion, it is easy to see that the bandwidth should be proportional to $n^{-1/5}$. Moreover, by Theorem 3.2 we consider the Mahalanobis distance

$$(\hat{\theta} - \theta_0)^\top \Sigma_0^+ (\hat{\theta} - \theta_0) = T_n + o\{h^8 + (nh)^{-2}\},$$

where

$$T_n = (\mathcal{E}_n + \frac{c_{1,n}}{nh} + c_{2,n}h^4 + \mathcal{H}_n)^\top \Sigma_0^+ (\mathcal{E}_n + \frac{c_{1,n}}{nh} + c_{2,n}h^4 + \mathcal{H}_n)$$

is the leading term. We have by Theorem 3.3 that

$$ET_n = E(\mathcal{E}_n^\top \Sigma_0^+ \mathcal{E}_n) + (\frac{c_1}{nh} + c_2 h^4)^\top \Sigma_0^+ (\frac{c_1}{nh} + c_2 h^4) + o\{h^8 + (nh)^{-2}\},$$

where $c_1 = \int K^2(v)v^2 dv \sigma^2 W_0^+ E\{\nu_0'(X) + f^{-1}(X)f'(X)\nu_0(X)\}$, $W_0 = E\{(g'(\theta_0^\top X))^2 \nu_{\theta_0}(X)\nu_{\theta_0}^\top(X)\}$ and

$$c_2 = \frac{1}{4}(\int K(v)v^4 dv - 1) W_0^+ E[g'(\theta_0^\top X) g''(\theta_0^\top X) \nu_{\theta_0}''(X)].$$

Note that $E(\mathcal{E}_n^\top \Sigma_0^+ \mathcal{E}_n)$ does not depend on h. By minimizing ET_n with respective to h, the optimal bandwidth should be

$$h_\theta = \left\{ \frac{(9r_2^2 + 16r_1)^{1/2} - 3r_2}{8} \right\}^{1/5} n^{-1/5},$$

where $r_1 = c_1^\top \Sigma_0^+ c_1 / (c_2^\top \Sigma_0^+ c_2)$ and $r_2 = c_1^\top \Sigma_0^+ c_2 / c_2^\top \Sigma_0^+ c_2$. As a comparison, we consider the optimal bandwidth for the estimation of the link function g. By Lemma 5.1 and Theorem 3.2, if $f_{\theta_0}(v) > 0$ we have

$$\hat{g}(v) = g(v) + \frac{1}{2} g''(v)^2 h^2 + \frac{1}{n f_{\theta_0}(v)} \sum_{i=1}^n K_h(\theta_0^\top X_i - v)\varepsilon_i + O_P(n^{-1/2} + h^2 \gamma_n). \tag{6}$$

In other words, the link function can be estimated with the efficiency as if the index parameter vector is known. A brief proof for (6) is given in section 5. It follows that

$$|\hat{g}(v) - g(v)|^2 = S_n(v) + O_P\{(n^{-1/2} + h^2\gamma_n)\gamma_n\}.$$

where the leading term is $S_n(v) = [\frac{1}{2} g''(v)^2 + \{n f_{\theta_0}(v)\}^{-1} \sum_{i=1}^n K_h(\theta_0^\top X_i - v)\varepsilon_i]^2$. Suppose we are interested in constant bandwidth in region $[a, b]$ with weight $w(v)$. Minimizing $\int_{[a,b]} ES_n(v)w(v)dv$ with respect to h,

we have the optimal bandwidth for the estimation of the link function is

$$h_g = \left[\frac{\int K^2(v)dv \int_{[a,b]} f_{\theta_0}^{-1}(v)\sigma_{\theta_0}^2(v)w(v)dv}{\int_{[a,b]} g''(v)^2 w(v)dv} \right]^{1/5} n^{-1/5}.$$

It is noticeable that the optimal bandwidth for the estimation of the parameter vector θ_0 is of the same order as that for the estimation of the link function. In other words, under-smoothing may lose efficiency for the estimation of θ_0 in the higher order sense. These optimal bandwidth h_θ^{opt} and h_g^{opt} can be consistently estimated by plug-in methods; see Ruppert et al (1995).

Although the optimal bandwidth for the estimation of θ is different from that for the link function, its estimation such as the plug-in method may be very unstable because of the estimation of second order derivatives. Moreover, its estimation needs another pilot parameter which is again hard to choose. In practice it is convenient to apply h_g^{opt} for h_θ^{opt} directly, and since h_g^{opt} and h_θ^{opt} have the same order, the loss of efficiency in doing so should be small. For the former, there are a number of estimation methods such as CV and GCV methods. If CV methods is used, in each iteration with the latest estimator θ, the bandwidth is selected by minimizing

$$\hat{h}_g = \underset{h}{\text{argmin}} n^{-1} \sum_{j=1}^{n} \{Y_j - \hat{g}_j^\theta(\theta^\top X_j)\}^2$$

where $\hat{g}_j^\theta(v)$ is the delete-one-observation estimator of the link function, i.e. the estimator of $\hat{g}^\theta(v)$ in (3) using data $\{(X_i, Y_i), i \neq j\}$. Another advantage for this approach is that we can also obtain the estimator for the link function.

4 Numerical Results

In the following calculation, the Gaussian kernel function and the trimming function (4) with $\varsigma = 1/20$ and $c_0 = 0.01$ are used. A MATLAB code rMAVE.m for the calculations below is available at

http://www.stat.nus.edu.sg/%7Estaxyc

In the first example, we check the behavior of bandwidths h_g and h_θ. We consider two sets of simulations to investigate the finite performance of our estimation method, and to compare the bandwidths for the estimation of the link function g and the single-index θ_0. Our models are

$$\text{model A: } y = (\theta_0^\top X)^2 + 0.2\varepsilon, \qquad \text{model B: } y = \cos(\theta_0^\top X) + 0.2\varepsilon,$$

where $\theta_0 = (3,2,2,1,0,0,-1,-2,-2,-3)^\top/6$, $X \sim N_{10}(0,I)$, and $\varepsilon \sim N(0,1)$ is independent of X. The ADE method was used to choose the initial value of θ. With different sample size n and bandwidth h, we estimate the model and calculate estimation errors

$$err_\theta = \{1 - |\theta_0^\top \hat{\theta}|\}^{1/2}, \qquad err_g = \frac{1}{n}\sum_{j=1}^{n} \rho_n\{\hat{f}_{\hat{\theta}}(\hat{\theta}^\top X_j)\}|\hat{g}^{\hat{\theta}}(\hat{\theta}^\top X_j) - g(\theta_0^\top X_j)|,$$

where $\hat{g}^{\hat{\theta}}(\hat{\theta}^\top X_j)$ is defined in (3). With 200 replications, we calculate the mean errors $mean(err_\theta)$ and $mean(err_g)$. The results are shown in Figure 1.

We have the following observations. (1) Notice that $n^{1/2} mean(err_\theta)$ tends to decrease as n increases, which means the estimation error err_θ enjoys a root-n consistency (and slightly faster for finite sample size). (2) Notice that the U-shape curves of err_θ has a wider bottom than those of err_g. Thus, the estimation of θ_0 is more robust to the bandwidth than the estimation of g. (3) Let $h_\theta^{opt} = \arg\min_h mean(err_\theta)$ and $h_g^{opt} = \arg\min_h mean(err_g)$. Then h_θ^{opt} and h_g^{opt} represent the best bandwidths respectively for the estimation of the link function g and the single-index θ_0. Notice that h_θ^{opt}/h_g^{opt} tends to increase as n increases, which means the optimal bandwidth for the estimation of θ_0 tends to zero not faster than that for the estimation of link function. Thus the under-smoothing bandwidth is not optimal.

Next, we compare our method with some of the existing estimation methods including ADE in Härdle and Stocker (1993), MAVE, the method in Hristache et al (2001), called HJS hereafter, the SIR and pHd methods in Li (1991, 1992) and SLS in Ichimura (1993). For SLS, we use the algorithm in Friedman (1984) in the calculation. The algorithm has best performance among those proposed for the minimization of SLS, such as Weisberg and Welsh (1994) and Fan and Yao (2003). We consider the following model used in Hristache et al (2001),

$$Y = (\theta_0^\top X)^2 \exp(a\theta_0^\top X) + \sigma\varepsilon, \tag{7}$$

where $X = (\mathbf{x}_1, \cdots, \mathbf{x}_{10})^\top$, $\theta_0 = (1,2,0,...,0)^\top/\sqrt{5}$, $\mathbf{x}_1, \cdots, \mathbf{x}_{10}, \varepsilon$ are independent and $\varepsilon \sim N(0,1)$. For the covariates X: $(\mathbf{x}_k + 1)/2 \sim Beta(\tau, 1)$ for $k = 1, \cdots, p$. Parameter a is introduced to control the shape of function. If $a = 0$, the structure is symmetric; the bigger it is, the more monotonic the function is.

Following Hristache et al (2001), we use the absolute deviation $\sum_{j=1}^{p} |\hat{\theta}_j - \theta_j|$ to measure the estimation errors. The calculation results for different σ and τ based on 250 replications are shown in Table 1. We have the following observations from Table 1. Our methods has much better performance than ADE and the method of Hristache et al (2001). For each simulation, the better one of SIR and pHd is reported in Table

Figure 1: *The wide solid lines are the values of* $\log\{n^{1/2}mean(err_\theta)\}$ *and the narrow lines are the values of* $\log\{n^{1/2}mean(err_g)\}$ *(re-scaled for easier visualisation). The dotted vertical lines correspond to the bandwidths* h_θ *and* h_g *respectively.*

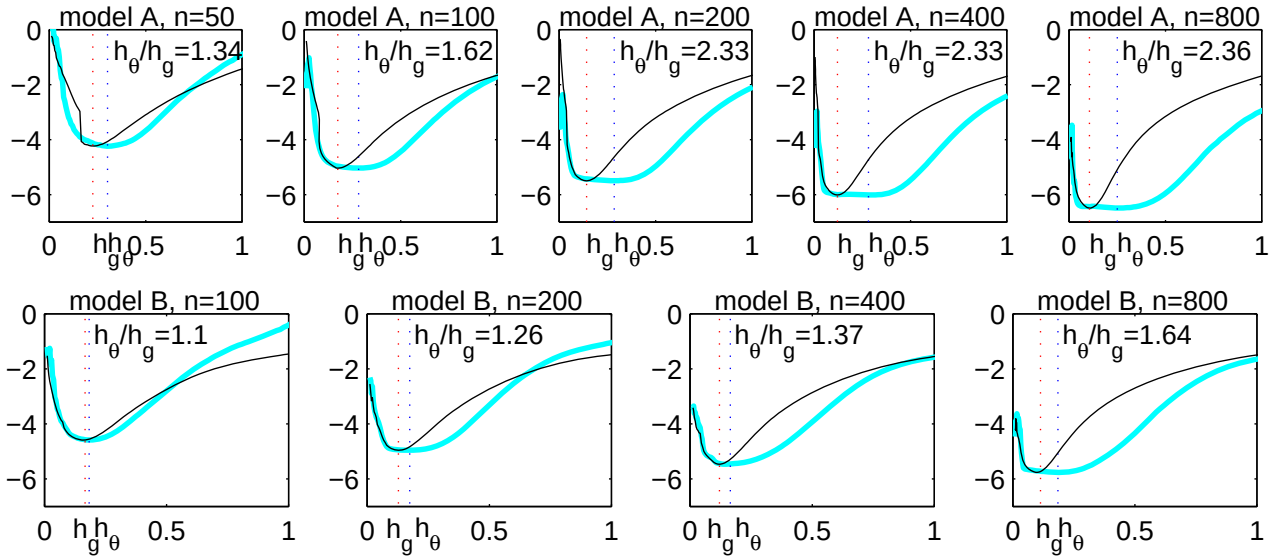

Table 1. Average estimation errors $\sum_{j=1}^{p}|\hat{\theta}_j - \theta_j|$
and their standard deviations (in square bracket) for model (7).

			$a=1$					$a=0$		
n	σ	τ	ADE*	HJS*	SIR/pHd	SLS	MAVE	SIR/pHd	SLS	MAVE
200	0.1	1	0.6094	0.1397	0.6521	0.0645	0.0514	0.7500	0.6910	0.0936
					[0.1569]	[0.0258]	[0.0152]	[0.1524]	[1.2491]	[0.0255]
200	0.2	1	0.6729	0.2773	0.6976	0.1070	0.0934	0.7833	0.8937	0.1809
					[0.1759]	[0.0375]	[0.0294]	[0.1666]	[1.3192]	[0.0483]
400	0.1	0.75	0.7670	0.1447	0.3778	0.1151	0.0701	0.6037	0.0742	0.0562
					[0.0835]	[0.0410]	[0.0197]	[0.1134]	[0.0193]	[0.0146]
400	0.1	1	0.4186	0.0822	0.4868	0.0384	0.0295	0.5820	0.5056	0.0613
					[0.1149]	[0.0125]	[0.0096]	[0.1084]	[1.0831]	[0.0167]
400	0.1	1.5	0.2482	0.0412	0.5670	0.0208	0.0197	0.5760	0.0923	0.0669
					[0.1524]	[0.0063]	[0.0056]	[0.1215]	[0.0257]	[0.0175]
400	0.2	1	0.4665	0.1659	0.5249	0.0654	0.0607	0.6084	0.7467	0.1229
					[0.1353]	[0.0207]	[0.0178]	[0.1064]	[1.2655]	[0.0357]
400	0.4	1	0.5016	0.3287	0.6328	0.1262	0.1120	0.6994	0.9977	0.2648
					[0.1386]	[0.0406]	[0.0339]	[0.1370]	[1.2991]	[0.1880]

* The values are adopted from Hristache et al (2001)

1, suggesting that these methods are not so competitive. Actually the main application of SIR and pHd is not in the estimation of single-index models. See Li (1991, 1992). For SLS, its performance depends much on the data and the model. If the model is easy to estimate (such as monotone and having big signal/noise ratio), it performance quite well. But overall SLS is still not so good as MAVE. The proposed method has the best performance in all the simulations we have done.

5 Proof of Theorems

Let $f_\theta(v)$ be the density function of $\theta^\top X$ and $\Lambda_n = \{x : |x| < n^c, f_\theta(x) > n^{-2\varsigma}, \theta \in \Theta_n\}$ where $c > 1/3$ and $\varsigma > 0$ is defined in (C5). Suppose A_n is a random matrix depending on x and θ. By $A_n = \mathcal{O}(a_n)$ (or $A_n = o(a_n)$) we mean that all elements in A_n are $O_{a.s.}(a_n)$ (or $o_{a.s.}(a_n)$) uniformly for $\theta \in \Theta_n$ and $x \in \Lambda_n$. Let $\delta_n = (nh/\log n)^{-1/2}$, $\gamma_n = h^2 + \delta_n$ and $\delta_\theta = |\theta - \theta_0|$. For any vector $V(v)$ of functions of v, we define $(V(v))' = dV(v)/dv$.

Suppose $(X_i, Z_i), i = 1, 2, \ldots, n$, are i.i.d. samples from (X, Z). Let $X_{ix} = X_i - x$,

$$s_k^\theta(x) = n^{-1}\sum_{i=1}^n K_h(\theta^\top X_{ix})\{\theta^\top X_{ix}/h\}^k, \quad t_k^\theta(x) = n^{-1}\sum_{i=1}^n K_h(\theta^\top X_{ix})\{\theta^\top X_{ix}/h\}^k X_i,$$

$$w_k^\theta(x) = n^{-1}\sum_{i=1}^n K_h(\theta^\top X_{ix})\{\theta^\top X_{ix}/h\}^k X_i X_i^\top, \quad e_k^\theta(x) = n^{-1}\sum_{i=1}^n K_h(\theta^\top X_{ix})\{\theta^\top X_{ix}/h\}^k \varepsilon_i,$$

$\epsilon_k^\theta = s_k^\theta(x) - Es_k^\theta(x)$, $\xi_k^\theta = t_k^\theta(x) - Et_k^\theta(x)$, $D_{n,k}^\theta(x) = s_2^\theta(x)s_k^\theta(x) - s_1^\theta(x)s_{k+1}^\theta(x)$, $E_{n,k}^\theta = s_0^\theta(x)s_{k+1}^\theta(x) - s_1^\theta(x)s_k^\theta(x)$ for $k = 1, 2, \ldots$. For any random variable Z and its random observations $Z_i, i = 1, ..., n$, let

$$T_{n,k}^\theta(Z|x) = s_2^\theta(x)n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})(\theta^\top X_{ix}/h)^k Z_i - s_1^\theta(x)n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})(\theta^\top X_{ix}/h)^{k+1} Z_i,$$

$$S_{n,k}^\theta(Z|x) = s_0^\theta(x)n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})(\theta^\top X_{ix}/h)^{k+1} Z_i - s_1^\theta(x)n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})(\theta^\top X_{ix}/h)^k Z_i.$$

By the Taylor expansion of $g(\theta_0^\top X_i)$ at $\theta_0^\top x$, we have

$$\begin{aligned} g(\theta_0^\top X_i) &= g(\theta_0^\top x) + \sum_{k=1}^5 \frac{1}{k!} g^{(k)}(\theta_0^\top x)\{\theta^\top X_{ix} + (\theta_0 - \theta)^\top X_{ix}\}^k + O(\{\theta^\top X_{ix} + (\theta_0 - \theta)^\top X_{ix}\}^6) \\ &= g(\theta_0^\top x) + A^\theta(x, X_i) + B^\theta(x, X_i)(\theta_0 - \theta) + O\{(\theta^\top X_{ix})^6 + \delta_\theta^3(|X_i|^6 + |x|^6)\}, \end{aligned} \tag{8}$$

where $A^\theta(x, X_i) = \sum_{\ell=1}^5 (k!)^{-1} g^{(k)}(\theta_0^\top x)(\theta^\top X_{ix})^k$ and

$$B^\theta(x, X_i) = \sum_{k=1}^5 \frac{1}{(k-1)!} g^{(k)}(\theta_0^\top x)(\theta^\top X_{ix})^{k-1} X_{ix}^\top + \frac{1}{2} g''(\theta_0^\top x)(\theta - \theta_0)^\top X_{ix} X_{ix}^\top.$$

For ease of exposition, we simplify the notation and abbreviate g for $g(\theta_0^\top x)$ and g', g'', g''' for $g'(\theta_0^\top x)$, $g''(\theta_0^\top x)$, $g'''(\theta_0^\top x)$ respectively. Without causing confusion, we write $f_\theta(\theta^\top x)$ as f_θ, $f_\theta(\theta^\top X_j)$ as $f_\theta(X_j)$ and $K_h(\theta^\top X_{ij})$ as $K_h^\theta(X_{ij})$. Similar notations are used for the other functions.

Lemma 5.1 (Link function) *Let*

$$\Sigma_n^\theta(x) = n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix}) \begin{pmatrix} 1 \\ \theta^\top X_{ix}/h \end{pmatrix} \begin{pmatrix} 1 \\ \theta^\top X_{ix}/h \end{pmatrix}^\top$$

and

$$\begin{pmatrix} a_\theta(x) \\ d_\theta(x)h \end{pmatrix} = \{n\Sigma_n^\theta(x)\}^{-1} \sum_{i=1}^n K_h^\theta(X_{ix}) \begin{pmatrix} 1 \\ \theta^\top X_{ix}/h \end{pmatrix} Y_i.$$

Under assumptions (C2)–(C5), we have

$$a_\theta(x) = g(\theta_0^\top x) + A_n^\theta(x)h^2 + B_n^\theta(x)(\theta_0-\theta) + V_n^\theta(x) + \mathcal{O}(h^2\gamma_n^2 + \delta_\theta^3)(1+|x|^6),$$

$$d_\theta(x)h = g'(\theta_0^\top x)h + \tilde{A}_n^\theta(x)h^2 + \tilde{B}_n^\theta(x)(\theta_0-\theta)h + \tilde{V}_n^\theta(x) + \mathcal{O}(h^2\gamma_n^2 + \delta_\theta^3)(1+|x|^6),$$

where

$$\begin{aligned}
A_n^\theta(x) &= \frac{1}{2}g'' + \frac{1}{4}\{(\mu_4-1)g''f_\theta^{-2}(f_\theta f_\theta'' - 2(f_\theta')^2) + \frac{1}{24}\mu_4 g^{(4)}\}h^2 + H_{1,n}^\theta(x), \\
\tilde{A}_n^\theta(x) &= \frac{1}{2}g''(\mu_4-1)f_\theta^{-1}f_\theta' h + \frac{1}{6}g^{(3)}\mu_4 h + \frac{1}{2}g''f_\theta^{-1}(\epsilon_3^\theta - \epsilon_1^\theta) + \mathcal{O}(h\gamma_n), \\
B_n^\theta(x) &= g'\nu_\theta + \mathcal{O}(\gamma_n+\delta_\theta), \quad \tilde{B}_n^\theta(x) = g'(\theta_0^\top x)f_\theta^{-1}\{f_\theta\nu_\theta(x)\}' + \mathcal{O}(\gamma_n),
\end{aligned}$$

where $H_{1,n}^\theta(x) = \frac{1}{2}g''(\theta_0^\top x)\{f_\theta^{-1}(\epsilon_2^\theta - \epsilon_0^\theta) + (2-\mu_4)f_\theta^{-2}f_\theta' h\epsilon_1^\theta - f_\theta^{-2}f_\theta' h\epsilon_3^\theta\} + \frac{1}{6}f_\theta^{-1}g'''h\epsilon_3^\theta$ *and* $V_n^\theta(x) = f_\theta^{-1}e_0^\theta - f_\theta^{-2}f_\theta' h e_1^\theta + \mu_4 f_\theta^{-2}f_\theta'' h^2 e_0^\theta/2 + f_\theta^{-2}(e_0^\theta\epsilon_2^\theta - e_1^\theta\epsilon_1^\theta) - \mu_4 f_\theta^{-2}f_\theta''' h^3 e_1^\theta + \{f_\theta^{-2}(f_\theta')^2 - (\mu_4+1)f_\theta^{-1}f_\theta''\}\{f_\theta^{-1}h^2 e_0^\theta - f_\theta^{-2}f_\theta' h^3 e_1^\theta\} - f_\theta^{-1}(\epsilon_0^\theta+\epsilon_1^\theta)\{f_\theta^{-1}e_0^\theta - f_\theta^{-2}f_\theta' e_1^\theta\} + 2f_\theta^{-2}f_\theta' h\epsilon_1^\theta f_\theta^{-1}e_0^\theta$ *and* $\tilde{V}_n^\theta(x) = f_\theta^{-1}e_1^\theta + f_\theta^{-2}f_\theta'' h^2 e_1^\theta/2 + f_\theta^{-2}(\epsilon_0^\theta e_1^\theta - \epsilon_1^\theta e_0^\theta) - f_\theta^{-2}f_\theta' h e_0^\theta + f_\theta^{-1}\epsilon_0^\theta[-(\mu_4+1)f_\theta^{-1}f_\theta'' h^2/2 - f_\theta^{-1}(\epsilon_0^\theta+\epsilon_1^\theta) + f_\theta^{-2}(f_\theta')^2 h^2]$.

Lemma 5.2 (Summations) *Let* $\eta_n^\theta(x) = n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})X_{ix}\varepsilon_i$. *Under conditions (C1)-(C5), we have*

$$\mathcal{A}_n^\theta \stackrel{def}{=} n^{-1}\sum_{j=1}^n \rho_n\{s_0^\theta(x_j)\}g'(\theta_0^\top X_j)\eta_n^\theta(X_j)/s_0^\theta(X_j) = \mathcal{E}_n^\theta + r_{n,0}^\theta(\theta-\theta_0) + Q_n^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

$$\mathcal{B}_n^\theta \stackrel{def}{=} (nh)^{-1}\sum_{j=1}^n \rho_n\{s_0^\theta(X_j)\}e_k^\theta(X_j)\eta_n^\theta(X_j)/s_0^\theta(X_j) = \frac{\tilde{c}_{k,n}}{nh} + R_n^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

$$\mathcal{C}_n^\theta \stackrel{def}{=} n^{-1}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))\epsilon_k^\theta(X_j)\eta_n^\theta(X_j)/s_0^\theta(X_j) = M_n^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

where $\mathcal{E}_n^\theta = \sum_{i=1}^n \rho_n\{f_\theta(X_j)\}g'(\theta^\top X_i)\nu_\theta(\theta^\top X_i)\varepsilon_i$, $r_{n,0}^\theta = o(1)$,

$$\mathcal{E}_n^\theta = \mathcal{O}\{(n/\log n)^{-1/2}\}, \quad Q_n^\theta = \mathcal{O}\{(n/\log n)^{-1/2}\gamma_n\}, \quad R_n^\theta = \mathcal{O}\{n^{-1/2}\delta_n\}, \quad M_n^\theta = \mathcal{O}\{n^{-1/2}\delta_n\},$$

with $E\{\mathcal{E}_n^\theta Q_n^\theta\} = o(h^8 + (nh)^{-2}), E\{\mathcal{E}_n^\theta R_n^\theta\} = o(h^8 + (nh)^{-2}), E\{\mathcal{E}_n^\theta M_n^\theta\} = o(h^8 + (nh)^{-2})$, *and* $\tilde{c}_{k,n} = \int v^{k+1}K^2(v)dvE[\rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(\nu_\theta(X_j)f_\theta(X_j))'(X_j)]$ *if* k *is odd, 0 otherwise.*

Lemma 5.3 (Denominator) *Let* $\mathcal{D}_n^\theta = n^{-2}\sum_{i,j=1}^n \rho_n(s_0^\theta(X_j))d_\theta^2(X_j)K_h^\theta(X_{ij})X_{ij}X_{ij}^\top/s_0^\theta(X_j)$ *in the algorithm. Suppose* $(\theta, B) : p \times p$ *is an orthogonal matrix. Then under (C1)-(C5), we have almost surely*

$$(\mathcal{D}_n^\theta)^{-1} = \theta\theta^\top d_{11}^\theta h^{-2} - \theta d_{12}^\theta B^\top h^{-1} - B(d_{12}^\theta)^\top\theta^\top h^{-1} + Bd_{22}^\theta B^\top,$$

where

$$d_{11}^\theta = (G_n^\theta)^{-1} + o(1), \quad d_{12}^\theta = H_n^\theta h + \mathcal{O}(\gamma_n), \quad d_{22}^\theta = \frac{1}{2}(B^\top W_n^\theta B)^{-1} + \mathcal{O}(\gamma_n),$$

with $G_n^\theta = n^{-1}\sum_{j=1}^n \rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(g'(\theta_0 X_j))^2$ *and* $H_n^\theta = \frac{1}{2}n^{-1}\sum_{j=1}^n \rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)\{(f_\theta\nu_\theta)'(X_j)\}^\top (G_n^\theta)^{-1}(g'(\theta_0^\top X_j))^2 B(B^\top W_n^\theta B)^{-1}$ *and* $W_n^\theta = n^{-1}\sum_{j=1}^n \rho_n\{f_\theta(X_j)\}(g'(\theta^\top X_i))^2\nu_\theta(X_j)\nu_\theta^\top(X_j)$.

Proof of Lemma 5.3 Let (θ, B) be an orthogonal matrix. It is easy to see that

$$n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})\theta^\top X_{ix}X_{ix}^\top\theta = s_2^\theta(x)h^2, \quad n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})\theta^\top X_{ix}X_{ix}^\top B = \{t_1^\theta(x) - s_1^\theta(x)x\}^\top Bh,$$

$$n^{-1}\sum_{i=1}^n K_h^\theta(X_{ix})B^\top X_{ix}X_{ix}^\top B = B^\top\{w_0^\theta(x) - t_0^\theta(x)x^\top - x(t_0^\theta(x))^\top + xx^\top s_0^\theta(x)\}B.$$

Thus

$$(\mathcal{D}_n^\theta)^{-1} = (\theta, B)\begin{pmatrix} D_{11}^\theta h^2 & (D_{12}^\theta)^\top Bh \\ B^\top D_{12}^\theta h & B^\top D_{22}^\theta B \end{pmatrix}^{-1}(\theta, B)^\top,$$

where

$$D_{11}^\theta = n^{-1}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))\{d^\theta(X_j)\}^2 s_2^\theta(X_j)/s_0^\theta(X_j),$$

$$D_{12}^\theta = n^{-1}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))\{d_\theta(X_j)\}^2\{t_1^\theta(X_j) - s_1^\theta(X_j)X_j\}^\top/s_0^\theta(X_j),$$

$$D_{22}^\theta = n^{-1}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))(d_\theta(X_j))^2\{w_0^\theta(X_j) - t_0^\theta(X_j)X_j^\top - X_j t_0^\theta(X_j) + X_j X_j^\top s_0^\theta(X_j)\}/s_0^\theta(X_j).$$

By the matrix inversion formula in blocks (Schott, 1997), we have the equation in Lemma 5.3 with $d_{11} = \{D_{11}^\theta - (D_{12}^\theta)^\top BB^\top (D_{22}^\theta)^{-1} BB^\top D_{12}^\theta\}^{-1}$, $d_{12} = d_{11}^\theta (D_{12}^\theta)^\top B(B^\top D_{22}^\theta B)^{-1}$, $d_{22}^\theta = \{B^\top D_{22}^\theta B\}^{-1} + d_{11}\{B^\top D_{22}^\theta B\}^{-1} B^\top D_{12}^\theta (D_{12}^\theta)^\top B\{B^\top D_{22}^\theta B\}^{-1}$. By Lemma 6.1, we have

$$D_{11}^\theta = G_n^{-1} + o(1), \quad D_{12}^\theta = H_n h + \mathcal{O}(\gamma_n), \quad D_{22}^\theta = 2W_n + \mathcal{O}(\gamma_n).$$

Thus, Lemma 5.3 follows. ■

Lemma 5.4 (Numerator) *Let* $\mathcal{N}_n^\theta = n^{-2} \sum_{i,j=1}^n \rho_n(s_0^\theta(X_j)) K_h^\theta(X_{ij}) X_{ij}\{Y_i - a_\theta(X_j) - d_\theta(X_j)\theta_0^\top X_{ij}\}/s_0^\theta(X_j)$. *Under assumptions (C1)–(C5), we have almost surely*

$$\mathcal{N}_n^\theta = \mathcal{E}_n^\theta + \frac{\tilde{c}_{1,n}}{nh} + \tilde{c}_{2,n} h^4 + \mathcal{R}_n^\theta + \mathcal{B}_n^\theta(\theta - \theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3 + \delta_\theta^3)\},$$

where $\mathcal{R}_n^\theta = \mathcal{O}\{n^{-1}(\log n/h)^{1/2} + (\log n/n)^{-1/2} h^2\}$, $\theta^\top \mathcal{R}_n^\theta = \mathcal{O}\{hn^{-1}(\log n/h)^{1/2} + (\log n/n)^{-1/2} h^3\}$ *and* $E\{\mathcal{R}_n^\theta \mathcal{E}_0^\theta\} = \mathcal{O}\{(nh)^{-2} + h^8\}$, $\mathcal{B}_n^\theta = W_n^\theta + o(1)$ *with* W_n^θ *defined in Lemma 5.3,* $\tilde{c}_{1,n}$ *and* $\mathcal{E}_0^\theta$ *are defined in Lemma 5.2 and*

$$\tilde{c}_{2,n} = \frac{1}{4}(\mu_4 - 1) \sum_{j=1}^n \rho_n\{f_\theta(X_j)\} g'(\theta_0^\top X_j) g''(\theta_0^\top X_j) \nu_\theta''(X_j).$$

Proof of Theorem 3.1 By assumption (C2), we have

$$\sum_{n=1}^\infty P(\bigcup_{i=1}^n \{X_i \notin \Lambda_n\}) \le \sum_{n=1}^\infty nP(X_i \notin \Lambda_n) \le \sum_{n=1}^\infty nP(|X_i| > n^c) < \sum_{n=1}^\infty nn^{-6c} E|X|^6 < \infty$$

for any $c > 1/3$. It follows from the Borel-Cantelli lemma that

$$P(\bigcap_{n=1}^\infty \bigcup_{i=1}^n \{X_i \notin \Lambda_n\}) = 0. \tag{9}$$

Let $\tilde{\Lambda}_n = \{x : f_\theta(\theta^\top x) > 2n^{-\epsilon}\}$. Similarly, we have

$$P(\bigcap_{n=1}^\infty \bigcup_{i=1}^n \{X_i \notin \tilde{\Lambda}_n\}) = 0. \tag{10}$$

Thus, we can exchange summations over $\{X_j : j = 1, \cdots, n\}$, $\{X_j : X_j \in \Lambda_n, j = 1, \cdots, n\}$ and $\{X_j : X_j \in \tilde{\Lambda}_n, j = 1, \cdots, n\}$ in the sense of almost surely consistency. On the other hand, we have by (C2)

$$n^{-1} \sum_{|X_j| < n^c} (1 + |X_j|^6) = O(1).$$

By the notation in Lemmas 5.3 and 5.4, after one iteration of Steps 1-3, the new θ is

$$\tilde{\theta} = \theta_0 + (\mathcal{D}_n^\theta)^{-1}\mathcal{N}_n^\theta. \tag{11}$$

Note that $\theta^\top \mathcal{E}_n^\theta = 0, \theta^\top c_{1,n}^\theta = 0$, $\theta^\top c_{2,n}^\theta = 0$, $\theta^\top W_n^\theta = 0$, $W_n^\theta (W_n^\theta)^+ = I - \theta\theta^\top$ and $\delta_\theta/h^2 \to 0$. We have

$$\begin{aligned}\tilde{\theta} =& \theta_0 + \theta[\theta^\top d_{11}^\theta h^{-2}\{\mathcal{R}_n^\theta + \mathcal{B}_n^\theta(\theta-\theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3+\delta_\theta^3)\}\} - d_{12}^\theta B^\top h^{-1}\mathcal{N}_n^\theta] \\ & - B(d_{12}^\theta)^\top \theta^\top h^{-1}[\mathcal{R}_n^\theta + \mathcal{B}_n^\theta(\theta-\theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3+\delta_\theta^3)\}] + B d_{22}^\theta B^\top \mathcal{N}_n^\theta \\ =& (1+a_n)\theta_0 + \{\frac{1}{2}(I-\theta_0\theta_0^\top) + b_n\}(\theta-\theta_0) + \frac{1}{2}\{W_n^\theta\}^+\mathcal{E}_n^\theta + \mathcal{O}(h^4),\end{aligned}$$

where $a_n = o(1)$ and $b_n = o(1)$.

By (25) below, we have $s_0^\theta(x) = f_\theta(\theta^\top x) + \mathcal{O}(\gamma_n)$. Thus by the smoothness of $\rho_n(.)$ and (10), we have

$$\rho_n(s_0^\theta(x)) = \rho_n(f_\theta(\theta^\top x)) + \mathcal{O}(n^\varsigma\gamma_n) = 1 + \mathcal{O}(n^\varsigma\gamma_n). \tag{12}$$

Since $\rho_n(.)$ is bounded, we have $E\{\rho_n(\hat{f}_\theta(\theta^\top x)) - 1\}^2 = o(1)$. By (C3) and Lemma 6.1, we have

$$\mathcal{E}_n^\theta \;=\; n^{-1}\sum_{i=1}^n g'(\theta_0^\top X_i)\nu_{\theta_0}(X_j)\varepsilon_i + o(n^{-1/2}).$$

Note that $W_n = W_0 + o(\delta_\theta)$. It is easy to check that $|\tilde{\theta}| = 1 + a_n + b_n + \mathcal{O}(h^4) = 1 + o(1)$. Thus

$$\tilde{\theta}/|\tilde{\theta}| = \theta_0 + \{\frac{1}{2}(I-\theta_0\theta_0^\top) + o(1)\}(\theta-\theta_0) + \frac{1}{2}n^{-1}W_0^+\sum_{i=1}^n g'(\theta_0^\top X_i)\nu_{\theta_0}(X_i)\varepsilon_i + o(h^3 + n^{-1/2}).$$

Let $\theta^{(k)}$ be the value of θ after k iteration. Because $h_{k+1} = \max\{h_k/c_h, h\}$. Therefore,

$$|\theta_{k+1} - \theta_0|/h_{k+1}^2 \to 0,$$

for all $k > 1$. We have

$$\theta^{(k+1)} = \theta_0 + \{\frac{1}{2}(I-\theta_0\theta_0^\top) + o(1)\}(\theta^{(k)}-\theta_0) + \frac{1}{2}n^{-1}W_0^+\sum_{i=1}^n g'(\theta_0^\top X_i)\nu_{\theta_0}(X_i)\varepsilon_i + o(h_k^3 + n^{-1/2}).$$

Recursing the above equation, we have

$$\begin{aligned}\theta^{(k+1)} = \theta_0 + \{\frac{1}{2^k}(I-\theta_0\theta_0^\top) + o(1)\sum_{\iota=1}^k\frac{1}{2^\iota}\}(\theta^{(1)}-\theta_0) + \{\sum_{\iota=1}^k\frac{1}{2^\iota}\}n^{-1}W_0^+\sum_{i=1}^n g'(\theta_0^\top X_i)\nu_{\theta_0}(X_i)\varepsilon_i \\ + o(\sum_{\iota=1}^k\frac{1}{2^\iota}h_{k-\iota}^3 + n^{-1/2}).\end{aligned}$$

Thus as the number of iterations $k \to \infty$, Theorem 3.1 follows immediately from the above equation and the central limit theorem. ■

Proof of Theorem 3.3 Based on Theorem 3.2, we can assume $\delta_\theta = (\log n/n)^{1/2}$. Note that $\theta^\top\{\mathcal{E}_n^\theta + c_{1,n}(nh)^{-1} + c_{2,n}h^4\} = 0$. We consider the product of each term in $(\mathcal{D}_n^\theta)^{-1}$ with $\mathcal{N}_n^\theta$. We have

$$\theta\theta^\top d_{11}^\theta h^{-2}\mathcal{N}_n^\theta = \theta\theta^\top d_{11}^\theta h^{-2}[\mathcal{R}_n^\theta + B_n^\theta(\theta-\theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3+\delta_\theta^3)\}] = a_n^\theta\theta_0 + a_n^\theta(\theta-\theta_0),$$
$$\theta d_{12}^\theta B^\top h^{-1}\mathcal{N}_n^\theta = b_n^\theta\theta_0 + b_n^\theta(\theta-\theta_0), \quad B(d_{12}^\theta)^\top\theta^\top h^{-1}$$

$$\begin{aligned}(\mathcal{D}_n^\theta)^{-1}\mathcal{N}_n &= \theta\{S_n^\top(\theta-\theta_0) + H_n\mathcal{E}_n^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^4)\}\\ &= \theta_0\{S_n^\top(\theta-\theta_0) + H_n\mathcal{E}_0^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^4)\} + c_n(\theta-\theta_0),\end{aligned}$$

where $S_n = \mathcal{O}(1)$ and $c_n = \mathcal{O}(\gamma_n/h)$. It is easy to see that $c_n =\circ(1)$ providing that $|\theta-\theta_0|/h^2 \to 0$. By Lemma 5.3 and 5.4, we have

$$\begin{aligned}\tilde\theta =&\theta_0\{1 + S_n^\top(\theta-\theta_0) + H_n\mathcal{E}_0^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^4)\} + \frac{1}{2}W_n^\theta\{\mathcal{E}_0^\theta + \frac{c'_{1,n}}{nh} + c'_{2,n}h^4 + \mathcal{R}_n^\theta + \mathcal{Q}_n^\theta\}\\ &+ \{\frac{1}{2}(I-\theta\theta^\top) + c_n\}(\theta-\theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3 + h\log n/n)\}.\end{aligned}$$

It is easy to see that $|\tilde\theta| = 1 + S_n^\top(\theta-\theta_0) + H_n\mathcal{E}_0^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^4)$. Thus

$$\tilde\theta/|\tilde\theta| = \theta_0 + \frac{1}{2}W_n^\theta\{\mathcal{E}_0^\theta + \frac{c'_{1,n}}{nh} + c'_{2,n}h^4 + \mathcal{R}_n^\theta + \mathcal{E}_0^\theta H_n^\top\mathcal{E}_0^\theta\} + \{\frac{1}{2}(I-\theta\theta^\top) + c'_n\}(\theta-\theta_0) + \mathcal{O}\{n^{2\varsigma}(\gamma_n^3 + h\log n/n)\},$$

where $c'_n =\circ(1)$. Similar to the proof of Theorem 3.1, we complete the proof with $c_{1,n} = W_n^{-1}c'_{1,n}$ and $c_{2,n} = W_n^{-1}c'_{2,n}$. ■

6 Proofs of the Lemmas

In this section, we first give some results about the uniform consistency. Based on these results, the Lemmas are proved.

Lemma 6.1 *Suppose $G_{n,i}(\chi)$ is a martingale with respect to $\mathcal{F}_i = \sigma\{G_{n,\ell}(\chi), \ell \le i\}$ with $\chi \in \mathcal{X}$ and $\mathcal{X}$ is a compact region in a multidimensional space such that (I) $|G_{n,i}(\chi)| < \xi_i$, where ξ_i are IID and $\sup E\xi_1^{2r} < \infty$ for some $r > 2$; (II) $EG_{n,k}^2(\chi) < a_n s(\chi)$ with $\inf s(\chi)$ positive, and (III) $|G_{n,i}(\chi) - G_{n,i}(\tilde\chi)| < n^{\alpha_1}|\chi-\tilde\chi|M_i$, where $M_i, i = 1, 2, ...$ are IID with $EM_1^2 < \infty$. If and $a_n = cn^{-\delta}$ with $0 \le \delta < 1 - 2/r$, then for any $\alpha_1' > 0$ we have*

$$\sup_{|\chi|\le n^{\alpha_1'}}\left|n^{-1}s^{-1/2}(\chi)\sum_{i=1}^n G_{n,i}(\chi)\right| = O\{(n^{-1}a_n\log n)^{1/2}\}$$

almost surely. Suppose for any fixed n and k, $G_{n,i,k}(\theta)$ is a martingale with respect to $\mathcal{F}_{i,k} = \sigma\{G_{n,\ell,k}(\theta), \ell \le i\}$ such that (I) $|G_{n,i,k}(\theta)| \le \xi_i$, (II) $EG^2_{n,i,k}(\theta) < a_n$ and (III) $|G_{n,i,k}(\theta) - G_{n,i,k}(\tilde{\theta})| < n^{\alpha_2}|\theta - \tilde{\theta}|M_i$, where ξ_i, a_n and M_i are defined above. If $E|\varepsilon_k|^{2r} < \infty$ and $E\{\varepsilon_k|G_{n,i,j}(\theta), i < j, j = 1, ..., k-1\} = 0$, then

$$\sup_{\theta\in\Theta}\Big|n^{-2}\sum_{k=2}^{n}\Big\{\sum_{i=1}^{k-1}G_{n,i,k}(\theta)\Big\}\varepsilon_k\Big| = O\{(a_n\log n)^{1/2}/n\}$$

almost surely.

Proof of Lemma 6.1 We give the details for the second part of the Lemma. The first part is easier and can be proved similarly. Let $\Delta_n(\theta)$ be the expression between the absolute symbols in the equation. By (III) and the strong low of large numbers, it is easy to see that there are $n_1 = n^{\alpha_3}$ balls centered at $\theta_\iota : B\iota = \{\theta : |\theta - \theta_\iota| < n^{-\alpha_4}\}$ with $\alpha_4 > \alpha_2 + 2$, such that $\bigcup_{\iota=1}^{n_1} B\iota \supset \Theta$. By the strong law of large numbers, we have

$$\max_{1\le\iota\le n_1}\sup_{\theta\in B\iota}|\Delta_n(\theta) - \Delta_n(\theta_\iota)| \le n^{\alpha_2}\max_{1\le\iota\le n_1}\sup_{\theta\in B\iota}|\theta-\theta_\iota|n^{-2}\sum_{k=1}^{n}|\varepsilon_k|\sum_{i=1}^{n}M_i = O\{(a_n\log n)^{1/2}/n\}$$

almost surely. Let $\Delta_{n,k}(\theta\iota) = \sum_{i=1}^{k-1}G_{n,i,k}(\theta\iota)$. Next, we show that there is a constant c_1 such that

$$p_n \overset{def}{=} P\Big(\bigcap_{\ell=1}^{\infty}\bigcup_{n=\ell}^{\infty}\{\max_{1<k\le n}\max_{1<\iota\le n_1}|\Delta_{n,k}(\theta\iota)| > c_1(na_n\log n)^{1/2}\}\Big) = 0. \tag{13}$$

Let $T_n = \{na_n\log(n)\}^{1/2}$, $G^I_{n,i,k}(\theta\iota) = G_{n,i,k}(\theta\iota)I(|G_{n,i,k}(\theta\iota)| \le T_n)$ and $G^O_{n,i,k}(\theta\iota) = G_{n,i,k}(\theta\iota) - G^I_{n,i,k}(\theta\iota)$. Write

$$\Delta_{n,k}(\theta\iota) = \sum_{i=1}^{k-1}\{G^I_{n,i,k}(\theta\iota) - EG^I_{n,i,k}(\theta\iota)\} + \sum_{i=1}^{k-1}\{G^O_{n,i,k}(\theta\iota) - EG^O_{n,i,k}(\theta\iota)\}. \tag{14}$$

Note that $E|G^O_{n,i,k}(\theta\iota)| \le T_n^{-r+1}E|\xi_1|^r = E|\xi_1|^r\{na_n\log(n)\}^{-(r-1)/2}$. If $a_n = cn^{-\delta}$ with $0 \le \delta < 1 - 2/r$ and $k \le n$, we have

$$|\sum_{i=1}^{k-1}EG^O_{n,i,k}(\theta\iota)| \le E|\xi_1|^r(k-1)\{na_n\log(n)\}^{-(r-1)/2} \le CE|\xi_1|^r\{na_n\log(n)\}^{1/2}. \tag{15}$$

Note that

$$\sum_{i=1}^{n}|G^O_{n,i,k}(\theta\iota)| \le \sum_{i=1}^{n}|\xi_i|I(|\xi_i| > T_n) \le T_n^{-r+1}\sum_{i=1}^{n}|\xi_i|^rI(|\xi_i| > T_n)$$

For fixed T, by the strong law of large numbers, we have

$$n^{-1}\sum_{i=1}^{n}|\xi_i|^rI(|\xi_i| > T) \to E\{|\xi_1|^rI(|\xi_1| > T)\}$$

almost surely. The right hand side above is dominated by $E\{|\xi_1|^r\}$ and $\to 0$ as $T \to \infty$. Note that T_n increase to ∞ with n. For large n such that $T_n > T$, we have

$$n^{-1}\sum_{i=1}^{n}|\xi_i|^r I(|\xi_i| > T_n) \le n^{-1}\sum_{i=1}^{n}|\xi_i|^r I(|\xi_i| > T) \to 0$$

almost surely as $T \to \infty$. It follows

$$\sum_{i=1}^{n}|G^O_{n,i,k}(\theta\iota)| = o(nT_n^{-r+1}) = o\{(na_n\log n)^{1/2}\} \tag{16}$$

almost surely. Thus by (15) and (16), if $c_1' > CE|\xi_1|^r$ we have

$$\begin{aligned} p_n' &\stackrel{def}{=} P\Big(\bigcap_{\ell=1}^{\infty}\bigcup_{n=\ell}^{\infty}\{\max_{1<k\le n}\max_{1<\iota\le n_1}|\sum_{i=1}^{k-1}\{G^O_{n,i,k}(\theta\iota) - EG^O_{n,i,k}(\theta\iota)\}| > c_1'(na_n\log n)^{1/2}\}\Big) \\ &\le P\Big(\bigcap_{\ell=1}^{\infty}\bigcup_{n=\ell}^{\infty}\{\sum_{i=1}^{n}|\xi_i|(|\xi_i| \ge T_n) > c_1'(na_n\log n)^{1/2}\}\Big) \\ &\qquad + P\Big(\bigcap_{\ell=1}^{\infty}\bigcup_{n=\ell}^{\infty}\{\max_{1<k\le n}\max_{1<\iota\le n_1}|\sum_{i=1}^{k-1}EG^O_{n,i,k}(\theta\iota)| > c_1'(na_n\log n)^{1/2}\}\Big) \\ &= 0. \end{aligned} \tag{17}$$

By condition (II), if $k \le n$ we have

$$\max_{1\le\iota\le n_1}\mathrm{Var}\sum_{i=1}^{k-1}\{G^I_{n,i,k}(\theta\iota) - EG^I_{n,i,k}(\theta\iota)\} \le c_2 na_n \stackrel{def}{=} N_1, \tag{18}$$

where c_2 is a constant. By the condition on a_n and the definition of $G^I_{n,i,k}(\theta\iota)$, we have constants c_3 and c_4 such that

$$\begin{aligned} \max_{1\le\iota\le n^\alpha}|\{G^I_{n,i,k}(\theta\iota) - EG^I_{n,i,k}(\theta\iota)\}| &\le c_3T_n \\ &= c_3\{na_n/\log n\}^{1/2}\{a_n^{-r}\log^{r+1}n/n^{r-2}\}^{1/(2(r-1))} \\ &\le c_4\{na_n/\log n\}^{1/2} \stackrel{def}{=} N_2. \end{aligned} \tag{19}$$

Let $N_3 = c_5\{na_n\log n\}^{1/2}$ with $c_5^2 > 2(\alpha_3+3)(c_2+c_4c_5)$. By the Bernstein's inequality (cf. DE LA Peña, 1999), we have from (18) and (19) that for any $k \le n$,

$$\begin{aligned} P(|\sum_{i=1}^{k-1}\{G^I_{n,i,k}(\theta\iota) - EG^I_{n,i,k}(\theta\iota)\}| > N_3) &\le 2\exp\left(\frac{-N_3^2}{2(N_1+N_2N_3)}\right) \\ &\le 2\exp\{-c_5^2\log n/(2c_2+2c_4c_5)\} \\ &\le c_6n^{-\alpha_3-3}. \end{aligned}$$

Let $c_1 > \max\{c_5, c_1'\}$. We have

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\Big\{ \max_{1<k\le n} \max_{1<\iota\le n_1} |\sum_{i=1}^{k-1}[G^I_{n,i,k}(\theta_\iota) - EG^I_{n,i,k}(\theta_\iota)]| > c_1(na_n \log n)^{1/2}\Big\} \\
&\le \sum_{n=1}^{\infty}\sum_{k=2}^{n}\sum_{\iota=1}^{n_1} P\Big\{ |\sum_{i=1}^{k-1}[G^I_{n,i,k}(\theta_\iota) - EG^I_{n,i,k}(\theta_\iota)]| > c_1(na_n \log n)^{1/2}\Big\} \\
&\le \sum_{n=1}^{\infty} c_6 n^{-\alpha_3-3} n^{1+\alpha_3} < \infty. \qquad (20)
\end{aligned}$$

By (14), (17) and (20) and the Borel-Cantelli lemma, we have

$$p_n \le P\Big\{ \bigcap_{\ell=1}^{\infty}\bigcup_{n=\ell}^{\infty} \max_{1<k\le n} \max_{1<\iota\le n_1} |\sum_{i=1}^{k-1}[G^I_{n,i,k}(\theta_\iota) - EG^I_{n,i,k}(\theta_\iota)]| > c_1(na_n \log n)^{1/2}\Big\} + p_n' = 0.$$

Therefore (13) follows.

Let $\Delta^I_{n,k}(\theta_\iota) = \Delta_{n,k}(\theta_\iota) I\{|\Delta_{n,k}(\theta_\iota)| \le c_1(na_n \log n)^{1/2}\}$ and $U_\ell(\theta_\iota) = \sum_{k=2}^{\ell} \Delta^I_{n,k}(\theta_\iota)\varepsilon_k$. Write

$$\Delta_n(\theta_\iota) = U_n(\theta_\iota) + \sum_{k=2}^{n} \Delta^O_{n,k}(\theta_\iota)\varepsilon_k,$$

where $\Delta^O_{n,k}(\theta_\iota) = \Delta_{n,k}(\theta_\iota) - \Delta^I_{n,k}(\theta_\iota)$. It is easy to see from (13) that for the second part on the right hand side above,

$$\max_{1<\iota\le n_1} |\sum_{k=2}^{n} \Delta^O_{n,k}(\theta_\iota)\varepsilon_k| = O\{n(a_n \log n)^{1/2}\} \qquad (21)$$

almost surely, since for any constant $c > 0$,

$$\begin{aligned}
\sum_{n=1}^{\infty} P\{\max_{1<\iota\le n_1} |\sum_{k=2}^{n} \Delta^O_{n,k}(\theta_\iota)\varepsilon_k| > cn(a_n \log n)^{1/2}\} &\le \sum_{n=1}^{\infty} P(\max_{1<\iota\le n_1}\max_{1<k\le n} |\Delta^O_{n,k}(\theta_\iota)| > 0) \\
&\le \sum_{n=1}^{\infty} P\{\max_{1<\iota\le n_1}\max_{1<k\le n} |\Delta_{n,k}(\theta_\iota)| > c_1(na_n \log n)^{1/2}\} \\
&< \infty.
\end{aligned}$$

Now consider the first term. Let $T_n'^{1/2}/\log n$,

$$U^I_\ell(\theta_\iota) = \sum_{k=2}^{\ell} \Delta^I_{n,k}(\theta_\iota)\{\varepsilon_k(|\varepsilon_k| \le T_n') - E[\varepsilon_k(|\varepsilon_k| \le T_n')]\}$$

and $U^O_\ell(\theta_\iota) = U_\ell(\theta_\iota) - U^I_\ell(\theta_\iota)$. Similar to the proof of (15) and (16), we have almost surely

$$|\sum_{k=2}^{\ell} \Delta^O_{n,k}(\theta_\iota) E\{\varepsilon_k(|\varepsilon_k| > T_n')\}| = O\{n(a_n \log n)^{1/2}\}, \qquad (22)$$

$$|\sum_{k=2}^{\ell} \Delta^O_{n,k}(\theta_\iota)\varepsilon_k(|\varepsilon_k| > T_n')| = O\{n(a_n \log n)^{1/2}\}. \qquad (23)$$

Note that

$$|\Delta^I_{n,k}(\theta_\iota)\{\varepsilon_k(|\varepsilon_k| \le T'_n) - E[\varepsilon_k(|\varepsilon_k| \le T'_n)]\}| < 2c_1(na_n \log n)^{1/2} T'_n = 2c_1 n(a_n/\log n)^{1/2} \stackrel{def}{=} N_4$$

and by (II), $\mathrm{Var}\{U^I_\ell(\theta_\iota)\} = c_2'^2 a_n \stackrel{def}{=} N_5$, where c'_2 is a constant. Let $N_6 = c'_3 n(a_n \log n)^{1/2}$ with ${c'_3}^2 > 2(\alpha_3+3)(2c_1c'_3 + c'_2)$. By the Berenstein's inequality, we have

$$P(|U^I_n(\theta_\iota)| \ge N_6) \le 2\exp\{-\frac{N_6^2}{2(N_6N_4 + N_5)}\} \le 2n^{-\alpha_3-3}.$$

Therefore

$$\sum_{n=1}^{n} P\{\max_{1\le\iota\le n_1} |U^I_n(\theta_\iota)| \ge N_6\} < \sum_{n=1}^{n} n_1 P\{|U^I_n(\theta_\iota)| \ge N_6\} < \infty.$$

By the Borel-Cantelli lemma, we have

$$\max_{1\le\iota\le n_1} |U^I_n(\theta_\iota)| = O(N_6) \tag{24}$$

almost surely. Lemma 6.1 follows from (21), (22), (23) and (24). ■

Proof of Lemma 5.1 Write $s^\theta_k(x) = \epsilon^\theta_k(x) + Es^\theta_k(x)$. By Taylor expansion, we have

$$s^\theta_k(x) = \sum_{\tau=0}^{3} \mu_{k+\tau} f^{(\tau)}_\theta(x) h^\tau + \epsilon^\theta_k(x) + \mathcal{O}(h^4). \tag{25}$$

Because $Var\{\epsilon^\theta_k(x)\} = O\{(nh)^{-1}\}$, it follows from Lemma 6.1 that $\epsilon^\theta_k(x) = \mathcal{O}(\delta_n)$. It is easy to check that

$$D^\theta_{n,0}(x) = f^2_\theta + \frac{1}{2}(\mu_4+1) f_\theta f''_\theta h^2 - (f'_\theta)^2 h^2 + f(\epsilon^\theta_0 + \epsilon^\theta_2) - 2f'_\theta h\epsilon^\theta_1 + \mathcal{O}(\gamma^2_n).$$

$$D^\theta_{n,2}(x) = f^2_\theta + \mu_4(f_\theta f''_\theta - (f'_\theta)^2)h^2 + 2f_\theta\epsilon^\theta_2 - f'_\theta h\epsilon^\theta_3 - \mu_4 f'_\theta h\epsilon^\theta_1 + \mathcal{O}(\gamma^2_n).$$

$$D^\theta_{n,3}(x) = f_\theta\epsilon^\theta_3 + \mathcal{O}(h\gamma_n), \quad D^\theta_{n,4}(x) = \mu_4 f^2_\theta + \mathcal{O}(\gamma_n), \quad D^\theta_{n,5}(x) = \mathcal{O}(h).$$

$$T^\theta_{n,0}(X|x) = f^2_\theta \nu_\theta(x) + \mathcal{O}(\gamma_n), \quad S^\theta_{n,0}(X|x) = \mathcal{O}(h), \quad T^\theta_{n,k}(X|x) = \mathcal{O}(1), \quad S^\theta_{n,k}(X|x) = \mathcal{O}(1), \text{ for } k \ge 1,$$

$$T^\theta_{n,0}(|\theta^\top X_{ix}|^6|x) = \mathcal{O}(h^6), \quad S^\theta_{n,0}(|\theta^\top X_{ix}|^6|x) = \mathcal{O}(h^6), \quad T^\theta_{n,0}(XX^\top|x) = \mathcal{O}(1), \quad S^\theta_{n,0}(XX^\top|x) = \mathcal{O}(h),$$

$$E_{n,2}(x) = (\mu_4 - 1) f_\theta f'_\theta h + f(\epsilon^\theta_3 - \epsilon^\theta_1) + \mathcal{O}(h\gamma_n), \quad E_{n,3}(x) = \mu_4 f^2_\theta + \mathcal{O}(\gamma_n), \quad E_{n,4}(x) = \mathcal{O}(h).$$

Note that

$$a^\theta(x) = T^\theta_{n,0}(Y|x)/D^\theta_{n,0}(x), \quad d^\theta(x)h = S^\theta_{n,0}(Y|x)/D^\theta_{n,0}(x).$$

and

$$A_n^\theta(x) = \sum_{k=2}^{5} \frac{1}{k!} g^{(k)}(\theta_0^\top x) \frac{D_{n,k}^\theta(x)}{D_{n,0}^\theta(x)} h^{k-2}, \quad B_n^\theta(x) = \sum_{k=0}^{4} \frac{1}{k!} g^{(k+1)}(\theta_0^\top x) \frac{T_{n,k}(X|x) - D_{n,k}(x)x}{D_{n,0}^\theta(x)} h^k,$$

$$C_n(x,\theta) = \frac{1}{2} g''(\theta_0^\top x) \{T_{n,0}(XX^\top|x) - T_{n,0}(X|x)x^\top - xT_{n,0}(X^\top|x) + xx^\top D_{n,0}^\theta(x)\} \{D_{n,0}^\theta(x)\}^{-1},$$

$$\tilde{A}_n^\theta(x) = \sum_{k=2}^{4} \frac{1}{k!} g^{(k)}(\theta_0^\top x) \frac{E_{n,k}^\theta(x)}{D_{n,0}^\theta(x)} h^{k-2}, \quad \tilde{B}_n^\theta(x) = \sum_{k=1}^{4} \frac{k}{k!} g^{(k)}(\theta_0^\top x) \frac{S_{n,k}(X|x) - E_{n,k}(x)x}{D_{n,0}^\theta(x)} h^k,$$

$$\tilde{C}_n(x,\theta) = \frac{1}{2} g''(\theta_0^\top x) \{S_{n,0}(XX^\top|x) - S_{n,0}(X|x)x^\top - xS_{n,0}(X^\top|x) + xx^\top E_{n,0}^\theta(x)\} \{D_{n,0}^\theta(x)\}^{-1}.$$

Lemma 5.1 follows from simple calculations based on the above equations. ■

Proof of Lemma 5.2 It follows from Lemma 6.1 that $\eta_n^\theta(x) = \mathcal{O}(\delta_n)(1+|x|)$ and $s_0^\theta = f_\theta + \tilde{\epsilon}_0^\theta$ where $\tilde{\epsilon}_0^\theta = \epsilon_k^\theta + (Es_k^\theta - f_\theta) = \mathcal{O}(\gamma_n)$. Because $|\rho_n''(.)| < n^{2\varsigma}$, we have

$$\rho_n(s_0^\theta(X_j)) = \rho_n(f_\theta(X_j)) + \rho_n'(f_\theta(X_j))\tilde{\epsilon}_0^\theta(X_j) + \mathcal{O}(n^{2\varsigma}\gamma_n^2). \tag{26}$$

Thus

$$\mathcal{A}_n^\theta = \tilde{\mathcal{E}}_n^\theta + Q_{n,1}^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

where $\tilde{\mathcal{E}}_n^\theta = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \rho_n(f_\theta(X_j)) f_\theta^{-1}(X_j) g'(\theta^\top X_j) K_h^\theta(X_{ij}) X_{ij}\varepsilon_i$, and $Q_{n,1}^\theta = n^{-1}\sum_{i=1}^n G_{n,i}^\theta$ with

$$G_{n,i}^\theta = n^{-1}\sum_{j=1}^{n} \Big[\frac{1}{2} f_\theta''(X_j)\{\rho_n'(f_\theta(X_j)) - \rho_n(f_\theta(X_j)) f_\theta^{-1}(X_j)\}h^2 + \{1 - \rho_n(f_\theta(X_j) f_\theta^{-1}(X_j)\}\tilde{\epsilon}_0^\theta(X_j)\Big] \times f_\theta^{-1}(X_j) g'(\theta^\top X_j) K_h^\theta(X_{ij}) X_{ij}\varepsilon_i.$$

Simple calculations lead to $E\tilde{\mathcal{E}}_n^\theta = 0, E(\tilde{\mathcal{E}}_n^\theta)^2 = O(n^{-1})$, $E(G_{n,i}^\theta) = 0$ and $E(G_{n,i}^\theta)^2 = O\{h^4 + (nh)^{-1}\}$. By the first part of Lemma 6.1, we have

$$\tilde{\mathcal{E}}_n^\theta = \mathcal{O}\{(\log n/n)^{1/2}\}, \quad Q_{n,1}^\theta = \mathcal{O}\{h^2(\log n/n)^{1/2} + n^{-1}(\log n/h)^{1/2}\}.$$

By Taylor expansion, $g'(\theta_0^\top x) = g'(\theta^\top x) + g''(v^*)(\theta_0 - \theta)^\top x$, where v^* is a value between $\theta^\top x$ and $\theta_0^\top x$. Write

$$\tilde{\mathcal{E}}_n^\theta = \mathcal{E}_n^\theta + Q_{n,2}^\theta + r_{n,0}(\theta - \theta_0),$$

where $Q_{n,2}^\theta = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \{\rho_n(f_\theta(X_j)) f_\theta^{-1}(X_j) g'(\theta^\top X_j) K_h^\theta(X_{ij}) X_{ij} - \rho_n(f_\theta(X_i)) g'(\theta^\top X_i)\nu_\theta(X_i)\}\varepsilon_i$ and $r_{n,0} = \mathcal{O}(\gamma_n/h)$. By Lemma 6.1 and that $Var(Q_{n,2}^\theta) = O\{h^4 + (nh)^{-1}\}$, we have

$$Q_{n,2}^\theta = \mathcal{O}\{(n/\log n)^{-1/2}\gamma_n\}.$$

Let $Q_n^\theta = Q_{n,1}^\theta + Q_{n,2}^\theta$. It is easy to check that $E\{Q_n^\theta \mathcal{E}_n^\theta\} = o(h^8 + (nh)^{-2})$. Therefore, the first part of Lemma 5.2 follows.

Similarly, we have from (26) that

$$\mathcal{B}_n^\theta = (nh)^{-1}\sum_{j=1}^n \{\rho_n(f_\theta(X_j)) + \rho_n'(f_\theta(X_j))\tilde{\epsilon}_0^\theta(X_j)\} e_k^\theta(X_j)\eta_n^\theta(X_j)/f_\theta(X_j) + \mathcal{O}(n^{2\varsigma}\gamma_n^4/h).$$

Let $\tilde{R}_n^\theta$ be the first term on the right hand side above. Then

$$\begin{aligned}\tilde{R}_n^\theta &= n^{-3}\sum_{j=1}^n \{\rho_n(f_\theta(X_j)) + \rho_n'(f_\theta(X_j))\tilde{\epsilon}_0^\theta(X_j)\}\sum_{i=1}^n K_h^2(\theta^\top X_{ij})(\theta^\top X_{ij}/h)^k X_{ij}\varepsilon_i^2/f_\theta(X_j)\\ &\quad + n^{-3}\sum_{j=1}^n \{\rho_n(f_\theta(X_j)) + \rho_n'(f_\theta(X_j))\tilde{\epsilon}_0^\theta(X_j)\}\sum_{i\neq\ell}^n K_h(\theta^\top X_{ij})(\theta^\top X_{ij}/h)^k K_h(\theta^\top X_{\ell j})X_{\ell j}\varepsilon_i\varepsilon_\ell/f_\theta(X_j)\\ &\stackrel{def}{=} \tilde{R}_{n,1}^\theta + \tilde{R}_{n,2}^\theta + \tilde{R}_{n,3}^\theta + \tilde{R}_{n,4}^\theta.\end{aligned}$$

If ε is independent of X, then

$$E_\theta(x) \stackrel{def}{=} E\{K_h^2(\theta^\top X_{ix})(\theta^\top X_{ij}/h)^k X_{ix}\varepsilon_i^2\} = h^{-1}\sum_{\ell=0}^2 \frac{1}{\ell!}\tilde{\mu}_{k+\ell}\{f_\theta(x)\nu_\theta(x)\}^{(\ell)}h^\ell\sigma^2 + O(h^2),$$

where $\tilde{\mu}_k = \int K^2(v)v^k dv$. By Lemma 6.1, we have

$$n^{-1}\sum_{i=1}^n K_h^2(\theta^\top X_{ix})(\theta^\top X_{ix}/h)^k X_{ix}\varepsilon_i^2 - E_\theta(x) = \mathcal{O}(h^{-1}\delta_n).$$

Thus

$$R_{n,0}^\theta \stackrel{def}{=} (n^2h)^{-1}\sum_{j=1}^n \rho_n(f_\theta(X_j))\Big[n^{-1}\sum_{i=1}^n K_h^2(\theta^\top X_{ij})(\theta^\top X_{ij}/h)^k X_{ij}\varepsilon_i^2 - E_\theta(X_j)\Big] = \mathcal{O}\{(nh^2)^{-1}\delta_n\}. \quad (27)$$

It is easy to check that $E\{\mathcal{E}_n^\theta R_{n,0}^\theta\} = 0$. Write

$$(n^2h)^{-1}\sum_{j=1}^n \rho_n(f_\theta(X_j))E_\theta(X_j) = (nh)^{-1}E\{\rho_n(f_\theta(X_j))E_\theta(X_j)\} + R_{n,1}^\theta,$$

where $E\{R_{n,1}^\theta \mathcal{E}_n^\theta\} = 0$ and

$$R_{n,1}^\theta = (n^2h)^{-1}\sum_{j=1}^n [\rho_n(f_\theta(X_j))E_\theta(X_j) - E\{\rho_n(f_\theta(X_j))E_\theta(X_j)\}] = \mathcal{O}\{(nh^2)^{-1}(n/\log n)^{-1/2}\}. \quad (28)$$

Note that $E\{\rho_n(f_\theta(X))\nu_\theta(X)\} = 0$. We have

$$(nh)^{-1}E\{\rho_n(f_\theta(X_j))E_\theta(X_j)\} = \frac{\tilde{c}_{k,n}}{nh} + R_{n,2}^\theta. \quad (29)$$

where $R^\theta_{n,2} = O(n^{-1})$ and $E\{R^\theta_{n,2}\mathcal{E}^\theta_0\} = 0$. By (27)-(29) and the fact that $(n/\log n)^{-1/2} = o(\gamma_n)$, we have

$$\tilde{R}^\theta_{n,1} = \frac{\tilde{c}_k}{nh} + R^\theta_{n,1} + R^\theta_{n,2}. \tag{30}$$

Similarly

$$\tilde{R}^\theta_{n,2} = \mathcal{O}\{(nh)^{-1}\gamma_n\}. \tag{31}$$

Let $G^\theta_{n,i,\ell} = n^{-1}\sum_{j=1}^n \rho_n(f_\theta(X_j))K_h(\theta^\top X_{ij})(\theta^\top X_{ij}/h)^k K_h(\theta^\top X_{\ell j})X_{\ell j}/f_\theta(X_j)$. Write $\tilde{R}^\theta_{n,3}$ as

$$\tilde{R}^\theta_{n,3} = n^{-2}\sum_{i\neq\ell}^n \frac{1}{2}(G^\theta_{n,i,\ell} + G^\theta_{n,\ell,i})\varepsilon_i\varepsilon_\ell = n^{-2}\sum_{\ell=1}^n \Big\{\sum_{i<\ell}\frac{1}{2}(G^\theta_{n,i,\ell} + G^\theta_{n,\ell,i})\varepsilon_i\Big\}\varepsilon_\ell.$$

By the second part of Lemma 6.1, we have

$$\tilde{R}^\theta_{n,3} = \mathcal{O}\{n^{-1/2}\delta_n\}. \tag{32}$$

Similarly, we have

$$\tilde{R}^\theta_{n,4} = \mathcal{O}\{n^{-1/2}\delta_n\}. \tag{33}$$

Thus the second part of Lemma 5.2 follows from (30) and (31).

The third part of Lemma 5.2 can be proved similarly as the proof of the second part. ■

Proof of Lemma 5.4 By (8), Lemma 5.1 and $\theta_0 = \theta + (\theta_0 - \theta)$, simple calculations lead to

$$\begin{aligned} Y_i - a_\theta(x) - d_\theta(x)\theta_0^\top X_{ix} = \varepsilon_i + \{\tilde{A}^\theta(x, X_i) - A^\theta_n(x)h^2\} + \{\tilde{B}^\theta(x, X_i) - B^\theta_n(x)\}^\top(\theta_0 - \theta) \\ - V^\theta_n(x) + \mathcal{O}\{h^2\gamma_n^2 + \delta_\theta^3\}, \end{aligned}$$

where $\tilde{A}^\theta(x, X_i) = A^\theta(x, X_i) - d_\theta(x)\theta^\top X_{ix}$ and $\tilde{B}^\theta(x, X_i) = B^\theta(x, X_i) - d_\theta(x)X_{ix}$. It follows from the Taylor expansion that

$$C^\theta_{n,k}(x) \stackrel{def}{=} n^{-1}\sum_{i=1}^n K^\theta_h(X_{ix})(\theta^\top X_{ix}/h)^k X_{ix} = \sum_{\ell=0}^5 \frac{1}{\ell!}\mu_{k+\ell}(f\mu_\theta)^{(\ell)}h^\ell + \tilde{\xi}^\theta_k + \mathcal{O}(h^6),$$

where $\tilde{\xi}^\theta_k = n^{-1}\sum_{i=1}^n\{K^\theta_h(X_{ix})(\theta^\top X_{ix}/h)^k X_{ix} - EK^\theta_h(X_{ix})(\theta^\top X_{ix}/h)^k X_{ix}\} = \xi^\theta_k - x\epsilon^\theta_k$. We have

$$\begin{aligned} n^{-1}\sum_{i=1}^n K^\theta_h(X_{ix})X_{ix}\tilde{A}^\theta(x, X_i) &= \{g'(\theta_0^\top x) - d_\theta(x)\}C_{n,1}(x)h + \sum_{k=2}^5 \frac{1}{k!}g^{(k)}(\theta_0^\top x)C^\theta_{n,k}(x)h^k \\ &= -\frac{1}{2}g''(\mu_4 - 1)f'_\theta f_\theta^{-1}(\nu_\theta f_\theta)'h^4 + \frac{1}{2}g''(\nu_\theta f_\theta)'(\epsilon^\theta_3 - \epsilon^\theta_1)h^2 + \tilde{V}^\theta_n\{(\nu_\theta f_\theta)'h + \frac{1}{6}\mu_4(\nu_\theta f_\theta)''h^3 + \tilde{\xi}^\theta_1\} \\ &\quad + \frac{1}{2}g''h^2\{f_\theta\nu_\theta + \frac{1}{2}\mu_4(f_\theta\nu_\theta)''h^2\} + \frac{1}{24}g^{(4)}\mu_4 f_\theta\nu_\theta h^4 + \frac{1}{2}g''h^2\tilde{\xi}^\theta_2 + \frac{1}{6}g'''h^3\tilde{\xi}^\theta_3 + \mathcal{O}(h^2\gamma_n^2). \end{aligned}$$

Thus

$$n^{-1}\sum_{i=1}^{n} K_h^\theta(X_{ix})X_{ix}\{\tilde{A}^\theta(x,X_i)-A_n^\theta(x)h^2\} =\frac{1}{4}(\mu_4-1)g''f_\theta\nu_\theta'' + B_{n,1}^\theta(\theta-\theta_0)+H_{2,n}^\theta+\mathcal{O}(h^2\gamma_n^2), \qquad (34)$$

where $B_{n,1}^\theta=\{(\nu_\theta f_\theta)'h+\frac{1}{6}\mu_4(\nu_\theta f_\theta)'''h^3+\tilde{\xi}_1^\theta\}B_n^\theta(x)^\top$ with $B_n^\theta(x)$ defined in Lemma 5.1, and

$$\begin{aligned}H_{2,n}^\theta =&\frac{1}{2}g''(\epsilon_3^\theta-\epsilon_1^\theta)h^2(\nu_\theta f_\theta)' + f_\theta^{-1}e_1^\theta(f_\theta\nu_\theta)'h+\frac{1}{2}f_\theta^{-2}f_\theta''(f\nu)'h^3e_1^\theta+f_\theta^{-2}(f\nu)'h(\epsilon_0^\theta e_1^\theta-\epsilon_1^\theta e_0^\theta)\\ &-f^{-2}f'(f\nu)'h^2e_0^\theta+f^{-1}(f\nu)'h\epsilon_0^\theta\{-\frac{1}{2}(\mu_4+1)f^{-1}f_\theta''h^2-f^{-1}(\epsilon_0^\theta+\epsilon_1^\theta)+f^{-2}(f')^2h^2\}\\ &+\frac{1}{6}\mu_4f^{-1}e_1^\theta(f\nu)''h^3+f^{-1}e_1^\theta\tilde{\xi}_1^\theta-f^{-2}f'he_0^\theta\tilde{\xi}_1^\theta+\frac{1}{2}g''h^2\tilde{\xi}_2^\theta+\frac{1}{6}g'''h^3\tilde{\xi}_3^\theta-\frac{1}{2}g''h^2\tilde{\xi}_0^\theta\\ &-\frac{1}{2}g''(\theta_0^\top x)\{f_\theta^{-1}(\epsilon_2^\theta-\epsilon_0^\theta)+(2-\mu_4)f_\theta^{-2}f_\theta'h\epsilon_1^\theta-f_\theta^{-2}f_\theta'h\epsilon_3^\theta\}\nu_\theta f_\theta h^2-\frac{1}{6}g'''\epsilon_3^\theta\nu_\theta h^3.\end{aligned}$$

By the expansions of $d_\theta(x)$ in Lemma 5.1, $\rho_n(s_0^\theta(x))$ in (26), and (34), we have

$$\begin{aligned}n^{-2}\sum_{j=1}^{n}\rho_n(s_0^\theta(X_j))d_\theta(X_j)\sum_{i=1}^{n}K_h^\theta(X_{ix})X_{ij}\{\tilde{A}^\theta(X_j,X_i)-A_n^\theta(X_j)h^2\}/s_0^\theta(X_j)&\\ =\tilde{c}_{2,n}h^4+(B_{n,2}^\theta)^\top(\theta-\theta_0)+\tilde{R}_{n,1}^\theta+\mathcal{O}\{n^{2\varsigma}(h^2\gamma_n^2+\delta_\theta^2h+\delta_\theta^3)\},&\end{aligned}$$

where $B_{n,2}^\theta=n^{-1}\sum_{j=1}^n\rho_n(s_0^\theta(X_j))d_\theta(X_j)B_{1,n}^\theta(X_j)/s_0^\theta(X_j)$ and $\tilde{R}_{n,1}^\theta=n^{-1}\sum_{j=1}^n\rho_n(s_n^\theta(X_j))d_\theta(X_j)H_{2,n}^\theta(X_j)$ $/s_0^\theta(X_j)$. Again by the expansion of d_θ and that $\tilde{\xi}_1^\theta=\mathcal{O}(\delta_n)$, we have $B_{n,2}^\theta=\mathcal{O}(h+\delta_n)$. It is easy to check that $H_{2,n}^\theta=\mathcal{O}(h\delta_n+\delta_n^2)$. We have

$$\begin{aligned}\tilde{R}_{n,1} =&n^{-1}\sum_{j=1}^{n}[\rho_n(f_\theta(X_j))+\rho_n'(f_\theta(X_j))\{f_\theta(X_j)+\frac{1}{2}f_\theta''(X_j)h^2+\epsilon_0^\theta(X_j)\}]\{g'(\theta_0^\top X_j)+\frac{1}{6}g'''(\theta_0^\top X_j)h^2\\ &+\tilde{V}_n^\theta(X_j)/h\}H_{2,n}^\theta(X_j)f_\theta^{-1}(X_j)\{1-\frac{1}{2}f_\theta^{-1}(X_j)f_\theta''(X_j)h^2-f_\theta^{-1}(X_j)\epsilon_0^\theta(X_j)\}+\mathcal{O}(n^{2\varsigma}\gamma_n^3)\\ \stackrel{def}{=}&R_{n,1}+\mathcal{O}(n^{2\varsigma}\gamma_n^3).\end{aligned}$$

Next, we need to consider the terms in $R_{n,1}$ one by one. Write

$$\begin{aligned}R_{n,1,1}^\theta\stackrel{def}{=}&n^{-1}\sum_{j=1}^{n}\rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(f_\theta(X_j)\nu_\theta(X_j))'e_1^\theta h\\ =&hn^{-2}\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n}K_h^\theta(X_{ij})\rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(f_\theta(X_j)\nu_\theta(X_j))'\Big\}\varepsilon_i.\end{aligned}$$

Note that $E\{\rho_n(f_\theta(X))f_\theta^{-1}(X)(f_\theta(X)\nu_\theta(X))'|\theta^\top X\}=0$. We have by Lemma 6.1

$$R_{n,1,1}^\theta=\mathcal{O}\{hn^{-1}(h^{-1}\log n)^{-1/2}\}$$

and

$$E\{\mathcal{E}_0^\theta R_{n,1,1}^\theta\} = hn^{-3}E\sum_{i=1}^n\Big\{\sum_{j=1}^n K_h^\theta(X_{ij})\rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(f_\theta(X_j)\nu_\theta(X_j))'\Big\}\rho_n(f_\theta(X_i))g'(\theta^\top X_i)\nu_\theta(X_i)\varepsilon_i^2$$
$$= hn^{-3}E\Big\{\sum_{j=1}^n K_h^\theta(0)\rho_n(f_\theta(X_j))f_\theta^{-1}(X_j)(f_\theta(X_j)\nu_\theta(X_j))'\rho_n(f_\theta(X_j))g'(\theta^\top X_j)\nu_\theta(X_j)\varepsilon_j^2\Big\}$$
$$= O(n^{-2}).$$

Applying similar approach to all the terms in $R_{n,1}^\theta$, we have

$$R_{n,1}^\theta = \mathcal{O}\{n^{-1}(\log n/h)^{1/2} + (\log n/n)^{1/2}h^2\} \quad \text{and} \quad E\{\mathcal{E}_n^\theta R_{n,1}^\theta\} = o\{(nh)^{-2} + h^8\}. \tag{35}$$

By Lemmas 5.1 and 6.1, we have

$$B_{n,3}^\theta \stackrel{def}{=} n^{-2}\sum_{j=1}^n \rho(s_0^\theta(X_j))d_\theta(X_j)\sum_{i=1}^n K_h^\theta(X_{ij})X_{ij}\{\tilde{B}^\theta(X_j,X_i) - B_n^\theta(X_j)\}^\top/s_0^\theta(X_j) = W_n^\theta + \mathcal{O}\{(\gamma_n+\delta_\theta)/h\}.$$

By Lemma 5.2, we have

$$n^{-2}\sum_{j=1}^n \rho(s_0^\theta(X_j))d_\theta(X_j)\sum_{i=1}^n K_h^\theta(X_{ij})X_{ij}\varepsilon_i/s_0^\theta(X_j) = \mathcal{E}_0^\theta + \frac{\tilde{c}_{1,n}}{nh} + B_{n,4}^\theta(\theta_0-\theta) + R_{n,2}^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

where $\tilde{c}_{1,n}$ is defined in the lemma, and

$$B_{4,n}^\theta = n^{-1}\sum_{j=1}^n\{\rho_n(f_\theta(X_j)) + \rho_n'(f_\theta(X_j))\epsilon_0^\theta(X_j)\}\eta_n^\theta(X_j)(\tilde{B}_n^\theta(X_j))^\top/h$$

and

$$R_{n,2}^\theta = n^{-1}\sum_{j=1}^n[\frac{1}{6}\rho_n(f_\theta(X_j))g'''(\theta_0^\top X_j)h^2 + \rho_n'(f_\theta(X_j))\epsilon_0^\theta(X_j)\{g'(\theta_0^\top X_j) + \tilde{V}_n^\theta(X_j)/h\}]\eta_n^\theta(X_j).$$

Noting that $\eta_n^\theta = \mathcal{O}(\delta_n)$, we have $B_{4,n}^\theta = \mathcal{O}(\delta_n/h)$. Similarly, we have

$$n^{-2}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))d_\theta(X_j)V_n^\theta(X_j)\sum_{i=1}^n K_h^\theta(X_{ij})X_{ij}/s_0^\theta(X_j) = R_{n,3}^\theta + \mathcal{O}(n^{2\varsigma}\gamma_n^3),$$

where

$$R_{n,3}^\theta = n^{-1}\sum_{j=1}^n \rho_n(s_0^\theta(X_j))d_\theta(X_j)V_n^\theta(X_j)[\nu_\theta(X_j) + \frac{1}{2}f_\theta^{-1}\{(f_\theta\nu_\theta)'' - f_\theta^{-1}f_\theta''\nu_\theta(X_j)\}h^2 + \xi_0^\theta(X_j) - \epsilon_0^\theta(X_j)].$$

By the same arguments leading to (35), we have

$$R_{n,2}^\theta = \mathcal{O}\{n^{-1}(\log n/h)^{1/2} + (\log n/n)^{1/2}h^2\} \quad \text{and} \quad E\{\mathcal{E}_n^\theta R_{n,2}^\theta\} = o\{(nh)^{-2} + h^8\}, \tag{36}$$

$$R_{n,3}^\theta = \mathcal{O}\{n^{-1}(\log n/h)^{1/2} + (\log n/n)^{1/2}h^2\} \quad \text{and} \quad E\{\mathcal{E}_n^\theta R_{n,3}^\theta\} = o\{(nh)^{-2} + h^8\}. \tag{37}$$

Lemma 5.4 follows from the above equations with $\mathcal{R}_n^\theta = R_{n,1}^\theta + R_{n,2}^\theta + R_{n,3}^\theta$ and $\mathcal{B}_n^\theta = B_{n,2}^\theta + B_{n,3}^\theta + B_{n,4}^\theta = W_n^\theta + \mathcal{O}\{n^{2\varsigma}(\gamma_n + \delta_\theta)/h\}$. ■

Acknowledgements The first author is most grateful to Professor V. Spokoiny for helpful discussions and NUS FRG R-155-000-048-112 and the Alexander von Humboldt Foundation for financial support. The second author thanks the Deutsche Forschungsgemeinschaft SFB 649 "Ökouomisches Risiko" for financial support. This paper was partly written while the third author was a Universidad Carlos III de Madrid-Banco Santander Chair of Excellence, and he thanks them for financial support.

References

[1] Bickel, P., Klaassen, A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Inference in Semiparametric Models,* Baltimore: Johns Hopkins University Press.

[2] Carroll, R.J., Fan. J. Gijbels, I. and Wand, M.P. (1997) Generalized partially linear single-index models. *J. Am. Statist. Ass.,* **92**, 477-489.

[3] Delecroix, M., Hristache, M. and Patilea, V. (2004) On semiparametric M-estimation in single-index regression. *J. Statist. Plann. and Infer.* (to appear).

[4] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications.* Chapman & Hall, London.

[5] Fan, J. and Yao, Q. (2003) *Nonlinear Time Series : nonparametric and parametric methods.* New York : Springer Verlag.

[6] Friedman, J. H. (1984) SMART User's Guide. Laboratory for Computational Statistics, Stanford University Technical Report No. 1.

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

[8] Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Amer. Stat. Ass.* **84** 986-995.

[9] Härdle, W. and A.B. Tsybakov (1993). How sensitive are average derivatives? *Journal of Econometrics* **58** 31-48.

[10] Horowitz, J.L. & Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. *J Amer. Stat. Assoc.*, **91** 1632-1640.

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

[12] Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71-120.

[13] Ichimura, H. and Lee, L. (1991) Semiparametric least squares estimation of multiple index models: Single equation estimation. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, edited by Barnett, W., Powell, J. and Tauchen, G.. Cambridge University Press.

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

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

[16] Nishiyama, Y., and P. M. Robinson (2000). Edgeworth expansions for semiparametric average derivatives. *Econometrica* 68, 931-980.

[17] Nishiyama, Y., and P. M. Robinson (2005). The Bootstrap and the Edgeworth Correction for semiparametric average derivatives. *Econometrica* 73, 903-948.

[18] Penrose, R. (1955) A generalized inverse for matrices, *Proc. Cambridge Philos. Soc.* **51**, 406-413.

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

[20] Powell, J.L. and T.M. Stoker (1996). Optimal bandwidth choice for density weighted averages. *Journal of Econometrics* 755, 291-316.

[21] Ruppert, D., Sheather, J., and Wand, P. M. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257-1270.

[22] Schott, J.R. (1997) *Matrix Analysis for Statistics*. John Wiley & Sons. New York.

[23] Weisberg, S. and Welsh, A. H. (1994) Estimating the missing link functions, *Ann. of Statist.* **22**, 1674-1700.

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

[25] Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* (to appear)

[26] Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275-1285.

[27] Yin, X. & Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.