Statistica Sinica **14**(2004), 1-39

A GOODNESS-OF-FIT TEST FOR SINGLE-INDEX MODELS

Yingcun Xia[1], W. K. Li[2], Howell Tong[2,3] and Dixin Zhang[4]

[1]*National University of Singapore,* [2]*University of Hong Kong,*
[3]*London School of Economics and* [4]*Nanjing University*

Abstract: The single-index model with an unknown link function is a generalized linear model that has been intensively investigated. This article considers a goodness-of-fit test for this model. Cramér-von Mises tests are constructed and the bootstrap method is used to provide $p-$values. The problem of bias in nonparametric estimation is tackled by the bootstrap method. Therefore, we do not need to undersmooth or oversmooth the link function. Some simulations are reported and some data are used for illustration.

Key words and phrases: Bias correction, bootstrap, Cramér-von Mises test, goodness of fit, local linear smoother, single-index model.

1. Introduction

Nonparametric methods have eased the problem of model mis-specification. Because of the curse of dimensionality in nonparametrics, many semi-parametric models have been introduced. Amongst them the single-index model and the projection pursuit method have proved to be effective. A single-index model can be written as

$$Y = g(X^T\theta) + \varepsilon, \tag{1.1}$$

where X is a $p \times 1$ covariate, θ is a vector with $\|\theta\| = 1$ and $E(\varepsilon|X) = 0$ almost surely. Both the link function $g(\cdot)$ and the parameter vector θ are unknown. If the model is correct, then it is known that root-n consistent estimator of θ can be obtained, where n is the sample size. Furthermore, the estimator of $g(\cdot)$ can achieve a consistent rate of $O_P(n^{-2/5})$ if the local linear or constant kernel smoother is used. See, for example, Ichimura and Lee (1991).

Developing appropriate goodness-of-fit tests for these models is an important and relevant problem. When the link function $g(\cdot) = g_0(\cdot)$ is known, Su and Wei (1991) investigated a supremum type test for

$\tilde{H}_0$: There exists a constant θ such that $E[Y - g_0(X^T\theta)|X] = 0$ almost surely.

Their alternative hypothesis, $\tilde{H}_1$, is that there does not exist a constant vector θ such that $E(Y|X) = g_0(X^T\theta)$. $\tilde{H}_1$ is quite a narrow alternative hypothesis.

Härdle, Spokoiny and Sperlich (1997) extended the alternative hypothesis to one which states that there is another function $\varphi(\cdot)$ ($\neq g_0(\cdot)$) and constant vector ϑ such that $Y = \varphi(X^T\vartheta) + \eta$ with $E(\eta|X) = 0$ almost surely. More recently, Stute, Manteiga and Quindimil (1998) considered a similar test using a different approach.

In this article, we consider the hypothesis

H_0 : There exist a constant vector θ and a link function $g(\cdot)$ such that $E[Y - g(X^T\theta)|X] = 0$ almost surely.

Note that H_0 does not specify the function $g(\cdot)$. We believe that this set-up is much more relevant in practical applications. Unlike Su and Wei (1991) and Stute, Manteiga and Quindimil (1998) who address the test problems under a parametric set-up, we investigate the test under semi-parametric assumptions in which the function $g(\cdot)$ is unknown and estimated nonparametrically.

The bias problem in estimation is an important issue in nonparametric inference but is notoriously difficult. This is in sharp contrast to the case of a parametric family of g's. There are two main techniques to deal with the bias problem: undersmoothing and bias correction. The former method is quite arbitrary, because the bandwidth is chosen to be smaller than the optimal one in the sense of minimizing the mean of integrated squared errors (MISE). As far as we know, there is as yet no generally accepted guidance on the bandwidth selection for undersmoothing. Some discussion can be found in Neumann and Kreiss (1998). For the latter method, we have to estimate the bias term, thus making the calculation more difficult. Moreover, the bias term is not so easy to estimate. See, for example, Xia (1998). For bootstraps, the bias problem is sometimes handled by the use of an oversmoothed estimator. See, for example, Härdle (1990, p.108). The method does not use the data driven bandwidth and is not completely free from being arbitrary. In this article, we propose a new method. Since we are going to use the bootstrap method to mimic the null distribution of the test statistic, we further use it to estimate the bias terms by a simple average of the bootstrap values. By doing so, we need neither undersmooth nor oversmooth the link function. Therefore, this method is easier to implement and is totally data driven.

The rest of this paper is organized as follows. In Section 2, we construct a Cramér-von Mises test statistic and give its asymptotic distribution. In Section 3, we propose some bootstrap test statistics and show why we may use these to assist us in obtaining the $p-$values of the previous statistic. The proofs are relegated to the Appendix. Some simulations and data analysis are given in Section 4. The programs are available at http://www.hku.hk/statistics/paper/test_SIM.

2. The Cramér-von Mises Test

Throughout this paper, we use v to denote a scalar variable and x a vector. Let $g_\theta(v) = E(Y|X^T\theta = v)$ and $\theta_0 = \arg\min_{\theta:\|\theta\|=1} E\Big[Y - g_\theta(X^T\theta)\Big]^2$. Then H_0 holds if and only if

$$E\Big[Y - g_{\theta_0}(X^T\theta_0)\Big|X\Big] = 0 \quad a.s. \tag{2.1}$$

To construct a Cramér-von Mises statistic, we further note that (2.1) is equivalent to

$$E\Big\{[Y - g_{\theta_0}(X^T\theta_0)]I(X < x)\Big\} \equiv 0, \tag{2.2}$$

where "$X < x$" means that every component of X is less than the corresponding component of x. Suppose that $\{(X_i,\ Y_i) : i = 1, \ldots, n\}$ is a random sample. Let $\hat{Y}_i$'s be the fitted values from the model (1.1) using some nonparametric method. Corresponding to (2.2), we construct the following residual marked empirical process

$$S_n(x) = n^{-1/2}\sum_{j=1}^{n}(Y_j - \hat{Y}_j)I(X_j < x).$$

To calculate the fitted value $\hat{Y}_j$, we need to estimate $g(\cdot)$ and θ in (1.1). For fixed θ, we estimate $g_\theta(v)$ using local linear kernel smoothing (see, e.g., Fan and Gijbels (1996)) by

$$\hat{g}_\theta(v) = \frac{\sum_{i=1}^n W_{n,h}(X_i^T\theta - v)Y_i}{\sum_{i=1}^n W_{n,h}(X_i^T\theta - v)}, \tag{2.3}$$

where $W_{n,h}(X_i^T\theta - v) = s_{n,\theta,2}(v)n^{-1}K_h(X_i^T\theta - v) - s_{n,\theta,1}(v)n^{-1}K_h(X_i^T\theta - v)\{(X_i^T\theta - v)/h\}$ with $s_{n,\theta,k}(v) = n^{-1}\sum_{j=1}^n K_h(X_j^T\theta - v)\{(X_j^T\theta - v)/h\}^k$, $k = 0, 1, 2$. Here and later, $K(\cdot)$ is a kernel function, $K_h(\cdot) = h^{-1}K(\cdot/h)$ and h is a bandwidth. For ease of exposition, we further assume that $\mu_2 = \int v^2K(v) = 1$. Otherwise, we may use kernel $\mu_2^{1/2}K(\mu_2^{1/2}v)$.

There are many methods to estimate the parameter θ. See for example Härdle and Stoker (1989), Ichimura and Lee (1991), Härdle, Hall and Ichimura (1993) and Weisberg and Welsh (1994). For simplicity, we here only consider estimators admitting the expression in the Appendix. Most estimation methods mentioned above admit such an expression. Actually, ℓ_n in (C6) can be a more general appropriate function. See Carroll, Fan, Gijbels and Wand (1997). Having obtained an estimate of θ we estimate $g(v)$ by $\hat{g}_{\hat{\theta}}(v)$ as in (2.3), and obtain the fitted value of Y_j as $\hat{Y}_j = \hat{g}_{\hat{\theta}}(X_j^T\hat{\theta})$ and hence the process $S_n(x)$.

To avoid the troublesome problem arising from the denominator of $\hat{g}_\theta(X_i^T\theta)$ being near 0, i.e., $\sum_{j=1}^n W_h(X_j^T\theta - X_i^T\theta) \approx 0$, a commonly used approach is to delete the boundary points, e.g., those satisfying $\|X_i\| > C$ for some constant C. See, for example, Härdle, Hall and Ichimura (1993) and Weisberg and Welsh

(1994). Here we only delete the observations for which $X_i^T\hat{\theta} \notin \mathcal{D}$ and $\mathcal{D}$ is a compact region on which $X^T\theta_0$ has positive density. We further define

$$S_{\mathcal{D}}(x) = n^{-1/2} \sum_{X_i^T\hat{\theta}\in\mathcal{D}} (Y_i - \hat{Y}_i) I(X_i < x).$$

Accordingly, we define $I_{\mathcal{D}}(X < x) = I(X^T\theta_0 \in \mathcal{D})I(X < x)$. Following Stute, Manteiga and Quindimil (1998), let

$$H(x) = \Big\{ I_D(X < x) - E\Big([g'(X^T\theta_0) I_{\mathcal{D}}(X < x)\{X - E(X|X^T\theta_0)\}]^T \Big) \ell(X, g, \theta_0) - E[I_{\mathcal{D}}(X < x) \Big| X^T\theta_0] \Big\} \varepsilon,$$

where ℓ is defined in the appendix. Let $\bar{\mathbb{R}}^{\otimes p} = [-\infty,\ \infty]^{\otimes p}$ and $D(\bar{\mathbb{R}}^{\otimes p})$ be the Skorokhod space. Let '$\Rightarrow$' denote the weak convergence (see, e.g., Billingsley (1968)).

Theorem 1. *Assume that conditions* (C1)−(C6) *hold. Then under* H_0*, we have* $S_{\mathcal{D}}(x) + B_{\mathcal{D}}(x) \Rightarrow Q(x)$ *in* $D(\bar{\mathbb{R}}^{\otimes p})$*, where* $B_{\mathcal{D}}(x) = n^{1/10} E[g''(X^T\theta_0) I_{\mathcal{D}}(X < x)]/2$ *and* $Q(x)$ *is a mean-zero Gaussian process with covariance function* $E[Q(x_1)Q(x_2)] = E[H(x_1)H(x_2)]$.

There is a bias term for the residual marked empirical process $S_{\mathcal{D}}(x)$, namely $B_{\mathcal{D}}(x)$. We have to remove it if we want to use the process $S_{\mathcal{D}}(x)$ for the purpose of testing. Therefore, we define the bias-corrected statistic as $CCV_{\mathcal{D}} = \int_{-\infty}^{\infty} [S_{\mathcal{D}}(x) + B_{\mathcal{D}}(x)]^2 F_n(dx)$, where $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i < x)$. By Theorem 1, we have $CCV_{\mathcal{D}} \to \int_{-\infty}^{\infty} [Q(x)]^2 F(dx)$ in distribution, where $F(x)$ is the cumulative distribution function of X.

As we have commented previously, the bias term will cause trouble in practice. In principle, it can be estimated using the usual method which, however, necessitates the selection of another bandwidth. Moreover, the limiting distribution still depends on the derivative of the unknown function $g(\cdot)$, which is difficult to estimate. It is hard to give a closed form for the distribution. Instead, we adopt the bootstrap approach to obtain an estimate of the bias and mimic the unknown distribution.

3. The Bootstrap Method

In this section, we adopt the wild bootstrap approach which is relevant for inference about regression models. See, e.g., Wu (1986) and Härdle and Mammen (1993). Suppose that $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ is a random sample from (1.1) under H_0. We first estimate θ_0 and $g(\cdot)$ as given in Section 2. We then generate independent bootstrap observations from the model

$$Y_i^* = \hat{g}_{\hat{\theta}}(X_i^T\hat{\theta}) + \varepsilon_i^*, \quad i = 1, \ldots, n, \tag{3.1}$$

with $\varepsilon_i^* = (Y_i - \hat{Y}_i)\epsilon_i^*$, where ϵ_i^*'s are i.i.d. random variables, each with zero mean, unit variance, finite moments of all orders and independent of $\{(X_i, Y_i), i = 1, \ldots, n\}$. We can re-estimate $\hat{\theta}$ and $\hat{g}(\cdot)$ as given in Section 2, and denote the estimators by $\hat{\theta}^*$ and $\hat{g}^*_{\hat{\theta}^*}(v)$, respectively. The bootstrap counterpart of $S_{\mathcal{D}}(x)$ is

$$S^*_{\mathcal{D}}(x) = n^{-1/2} \sum_{X_i^T\hat{\theta}\in\mathcal{D}} (Y_i^* - \hat{Y}_i^*)I(X_i < x).$$

3.1. The idea of bias-correction

Next, for ease of explanation of our basic idea, we temporarily assume that θ_0 is known. Let $z_i = X_i^T\theta_0$. Then $\hat{g}_{\hat{\theta}}(X_i^T\hat{\theta})$ in (3.1) changes to $\hat{g}_{\theta_0}(z_i)$. Under some assumptions (see the appendix for detail), we have

$$\hat{g}_{\theta_0}(v) = g_{\theta_0}(v) + \frac{1}{2}g''_{\theta_0}(v)h^2 + \frac{1}{nf_{\theta_0}(v)} \sum_{i=1}^n K_h(z_i - v)\varepsilon_i + o_P(h^2),$$

where $f_{\theta_0}(v)$ is the density function of $X^T\theta_0$. For each bootstrap sample, we have

$$\begin{aligned}\hat{g}^*_{\theta_0}(v) = g_{\theta_0}(v) + g''_{\theta_0}(v)h^2 + \frac{1}{nf_{\theta_0}(v)} \sum_{i=1}^n K * K_h(z_i - v)\varepsilon_i \\ + \frac{1}{nf_{\theta_0}(v)} \sum_{i=1}^n K_h(z_i - v)\varepsilon_i^* + o_P(h^2),\end{aligned} \tag{3.2}$$

where $K * K$ denotes the convolution of K. The bias for $\hat{g}^*_{\theta_0}(x)$ is

$$\begin{aligned}&E\left[\hat{g}^*_{\theta_0}(v) - \hat{g}_{\theta_0}(v)\Big|(z_i, Y_i),\ i = 1, \ldots, n\right] \\ &= \frac{1}{2}g''(v)h^2 + \frac{1}{nf_{\theta_0}(v)} \sum_{i=1}^n \{K * K_h(z_i - v) - K_h(z_i - v)\}\varepsilon_i + o_P(h^2).\end{aligned} \tag{3.3}$$

See Lemma A.1 in the appendix. Note that the second term on the right hand side above is $O_P(h^2)$ (as h is proportional to $n^{-1/5}$). Equation (3.3) implies that $\hat{g}_{\theta_0}(\cdot)$ and $\hat{g}^*_{\theta_0}(\cdot)$ have different bias terms. Therefore if we try to make a pointwise inference about the regression function g, we have to use another bandwidth and oversmooth the regression function such that the second term in (3.3) is $o_P(h^2)$. See Härdle (1990, p.107). However, the difference in (3.3) can be reduced by the summation of the residual marked empirical process in our problem, namely

$$n^{-1/2} \sum_{z_i\in\mathcal{D}} \{nf_{\theta_0}(z_i)\}^{-1} \sum_{j=1}^n \{K * K_h(z_i - v) - K_h(z_i - v)\}\varepsilon_i = o_P(1) \tag{3.4}$$

because $\int\{K*K(v)-K(v)\}dv = 0$. By (3.4), we can show that $S^*_{\mathcal{D}}(x)$ and $S_{\mathcal{D}}(x)$ have the same bias term asymptotically. Note that by (3.2), the bias $E[\hat{g}^*_{\theta_0}(v) - \hat{g}_{\theta_0}(v)|(z_i, Y_i),\ i = 1, \ldots, n]$ can be obtained by the average of the resample. Therefore the bias terms in the $S^*_{\mathcal{D}}(x)$ and $S_{\mathcal{D}}(x)$ can be easily calculated and removed.

3.2. The asymptotic distributions

Let $B_i = E[\hat{g}^*_{\hat{\theta}}(X_i^T\hat{\theta}) - \hat{g}_{\hat{\theta}}(X_i^T\hat{\theta})|(X_j, Y_j),\ j = 1, \ldots, n]$, $\tilde{Y}_i = \hat{Y}_i - B_i$ and $\tilde{Y}^*_i = \hat{Y}^*_i - B_i$ where $\hat{Y}^*_i = \hat{g}^*_{\hat{\theta}^*}(X_i^T\hat{\theta}^*)$. Then $\tilde{Y}_i$ and $\tilde{Y}^*_i$ can be seen as the bias corrected fitted values. Let

$$\tilde{S}_{\mathcal{D}}(x) = n^{-1/2}\sum_{X_i^T\hat{\theta}\in\mathcal{D}}(Y_i - \tilde{Y}_i)I(X_i < x).$$

Its bootstrap counterpart is

$$\tilde{S}^*_{\mathcal{D}}(x) = n^{-1/2}\sum_{X_i^T\hat{\theta}\in\mathcal{D}}(Y^*_i - \tilde{Y}^*_i)I(X_i < x).$$

Note that the summations in bootstrap statistics $S^*_{\mathcal{D}}(x)$ and $\tilde{S}^*_{\mathcal{D}}(x)$ should be taken over $\{X_i^T\hat{\theta}^* \in \mathcal{D}\}$ accordingly. However, by the results of Lemma A.3, the summations over $[\{X_i^T\hat{\theta} \in \mathcal{D}\} - \{X_i^T\hat{\theta}^* \in \mathcal{D}\}] \cup [\{X_i^T\hat{\theta}^* \in \mathcal{D}\} - \{X_i^T\hat{\theta} \in \mathcal{D}\}]$ are negligible. Finally, we have the following results.

Theorem 2. *Suppose* (C1)−(C6) *hold. Then under* H_0*, we have* $\tilde{S}_{\mathcal{D}}(x) \Rightarrow Q(x)$ *and* $\tilde{S}^*_{\mathcal{D}}(x) \Rightarrow Q(x)$ *in* $D(\bar{\mathbb{R}}^{\otimes p})$*, where* $Q(x)$ *is as defined in Theorem* 1*.*

Because $\tilde{S}^*_{\mathcal{D}}(x)$ has the same limiting distribution as $\tilde{S}_{\mathcal{D}}(x)$, we can use them to test our hypothesis. Consider the following Cramér-von Mises statistics

$$CVS_{\mathcal{D}} = \int_{-\infty}^{\infty}\left[\tilde{S}_{\mathcal{D}}(x)\right]^2 F_n(dx), \quad CVS^*_{\mathcal{D}} = \int_{-\infty}^{\infty}\left[\tilde{S}^*_{\mathcal{D}}(x)\right]^2 F_n(dx).$$

By Theorem 2, $CVS_{\mathcal{D}}$ and $CVS^*_{\mathcal{D}}$ have the same limiting distribution. We can mimic the distribution of $CVS_{\mathcal{D}}$ by its counterpart $CVS^*_{\mathcal{D}}$.

3.3. Test statistics with bias terms

Note that $S^*_{\mathcal{D}}(x) + B_{\mathcal{D}}(x) \Rightarrow Q(x)$ in $D(\bar{\mathbb{R}}^{\otimes p})$. Both $S^*_{\mathcal{D}}(x)$ and $S_{\mathcal{D}}(x)$ have the same bias term and the same limiting distribution when the bias is removed. For simplicity of calculation, we can construct the test statistics (with bias terms)

$$CVT_{\mathcal{D}} = \int_{-\infty}^{\infty}[S_{\mathcal{D}}(x)]^2 F_n(dx), \quad CVT^*_{\mathcal{D}} = \int_{-\infty}^{\infty}[S^*_{\mathcal{D}}(x)]^2 F_n(dx).$$

Our simulations (not reported here) show that tests based on $CVT_{\mathcal{D}}$ and $CVT^*_{\mathcal{D}}$ also work well.

3.4. Single-indexing test statistics

Let $r(x) = E[Y - g_{\theta_0}(X^T\theta_0)|X = x]$. That $r(x) \neq 0$ can usually be detected by $E[r(X)|X^T\theta]$. See Huber (1985, Section III). Let $r_\theta(v) = E[Y - g_{\theta_0}(X^T\theta_0)|X^T\theta = v]$ and $\theta_1 = \arg\min_{\theta:\|\theta\|=1} E[Y - g_{\theta_0}(X^T\theta_0) - r_\theta(X^T\theta)]^2$. Then $r_{\theta_1}(v)$ is just the second component of the projection pursuit regression. We can repeat the estimation method for unknown θ_0 to estimate θ_1, obtaining $\hat\theta_1$ say. Similarly, we can estimate its bootstrap counterpart say $\hat\theta_1^*$. Let

$$S'_{\mathcal{D}}(v) = n^{-1/2} \sum_{X_i^T\hat\theta\in\mathcal{D}} (Y_i - \tilde Y_i) I(X_i^T\hat\theta_1 < v),$$
$$S'^*_{\mathcal{D}}(v) = n^{-1/2} \sum_{X_i^T\hat\theta\in\mathcal{D}} (Y_i^* - \tilde Y_i^*) I(X_i^T\hat\theta_1^* < v).$$

Similarly, let $CVS'_{\mathcal{D}} = \int_{-\infty}^{\infty} [S'_{\mathcal{D}}(v)]^2 F_{n,\hat\theta_1}(dv)$ and $CVS'^*_{\mathcal{D}} = \int_{-\infty}^{\infty} [S'^*_{\mathcal{D}}(v)]^2 F_{n,\hat\theta_1}(dv)$, where $F_{n,\theta}(v) = n^{-1}\sum_{i=1}^n I(X_i^T\theta < v)$. Our simulations show that the test of H_0 based on $CVS'_{\mathcal{D}}$ and $CVS'^*_{\mathcal{D}}$ is more stable and powerful than that based on $CVS_{\mathcal{D}}$ and $CVS^*_{\mathcal{D}}$.

4. Simulations and Data Analyses

To check the performance of our bootstrap method for finite samples, we carry out the following simulations and data analyses. In the following examples we use the normal density kernel $K(v) = (2\pi)^{-1/2}\exp(-v^2/2)$, so $\int K''(v)dv = 0$. We use the method of Härdle, Hall and Ichimura (1993) to obtain $\hat\theta$ and $\hat\theta^*$.

Example 1. Consider the model

$$Y = x_1 + x_2 + 4\exp\{-(x_1 + x_2)^2\} + a(x_1^2 + x_2^2)^{1/2} + \sigma\varepsilon, \tag{4.1}$$

where x_1, x_2 and $\varepsilon \overset{i.i.d.}{\sim} N(0,\ 1)$. A typical data set with $n = 100$, $a = 0$ and $\sigma = 0.3$ is shown in Figure 1. Note that when $a = 0$, (4.1) is a single-index model with link function $g(v) = \sqrt{2}v + 4\exp(-2v^2)$ and $\theta_{01} = \theta_{02} = \sqrt{2}/2$.

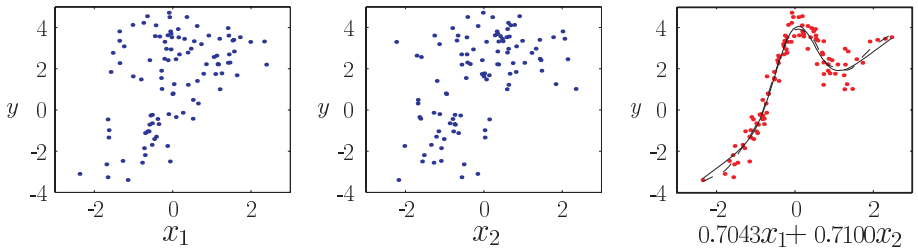

Figure 1. The first two panels are the plots of observations. In the third panel, dot denotes $(X_i^T\hat\theta, Y_i)$; solid line denotes the true function $g(\cdot)$; broken line denotes the estimated function $\hat g(\cdot)$. The estimated θ is $\hat\theta = (0.7043\ \ 0.7100)^T$.

Now, we carry out our test of H_0 : There exist a $(\theta_{01}, \theta_{02})^T$ and a link function $g(\cdot)$ such that $E[Y - g(\theta_{01}x_1 + \theta_{02}x_2)\big|(x_1,\ x_2)] = 0$ almost surely. We set the region $\mathcal{D} = [-2.5, 2.5]$. For each sample size $n = 50$, 100 and 300, 1,000 independent samples are drawn. Table 1 presents the percentages of times H_0 is rejected for $\alpha = 0.05$ and 0.10 and selected values of σ and a. (The percentage points of the null distribution are by reference to the bootstrap distribution based on 1,000 samples.) Table 1 shows that our bootstrap method works quite well if we use a bandwidth selected by cross-validation. It also has reasonable performance for bandwidths around the optimal ones in the MISE sense. Therefore any bandwidth comparable with the one obtained by MISE can be used. Some simulations on the power of the test with respect to different values of a are shown in Figure 2.

Table 1. Percentage of times H_0 is rejected out of 1,000 Monte Carlo replications.

n	σ	a	Nominal level α							
			0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05
50			$h = 0.11$		$h = 0.16^*$		$h = 0.21$		CV bandwidth	
	0.3	0.00	0.138	0.061	0.118	0.063	0.084	0.037	0.124	0.076
		0.25	0.309	0.174	0.195	0.099	0.095	0.044	0.162	0.084
		0.50	0.714	0.541	0.526	0.376	0.328	0.186	0.468	0.334
			$h = 0.15$		$h = 0.20^*$		$h = 0.25$		CV bandwidth	
	0.5	0.00	0.126	0.053	0.102	0.043	0.069	0.031	0.121	0.053
		0.25	0.204	0.107	0.106	0.043	0.070	0.031	0.120	0.056
		0.50	0.412	0.259	0.289	0.163	0.184	0.097	0.296	0.163
100			$h = 0.10$		$h = 0.14^*$		$h = 0.16$		CV bandwidth	
	0.3	0.00	0.131	0.065	0.094	0.045	0.106	0.037	0.113	0.052
		0.25	0.561	0.371	0.314	0.208	0.146	0.089	0.333	0.198
		0.50	0.961	0.921	0.905	0.806	0.707	0.554	0.856	0.753
			$h = 0.12$		$h = 0.16^*$		$h = 0.20$		CV bandwidth	
	0.5	0.00	0.126	0.055	0.101	0.057	0.095	0.041	0.120	0.047
		0.25	0.250	0.151	0.175	0.082	0.115	0.060	0.183	0.100
		0.50	0.788	0.632	0.613	0.445	0.410	0.275	0.535	0.418
300			$h = 0.07$		$h = 0.10^*$		$h = 0.13$		CV bandwidth	
	0.3	0.00	0.109	0.059	0.097	0.043	0.091	0.042	0.115	0.052
		0.25	0.968	0.903	0.912	0.856	0.698	0.553	0.885	0.800
		0.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			$h = 0.10$		$h = 0.12^*$		$h = 0.14$		CV bandwidth	
	0.5	0.00	0.124	0.055	0.103	0.048	0.102	0.045	0.111	0.047
		0.25	0.584	0.461	0.472	0.318	0.328	0.204	0.464	0.260
		0.50	0.998	0.991	0.990	0.977	0.985	0.950	0.985	0.964

* The optimal bandwidth in the sense of MISE.

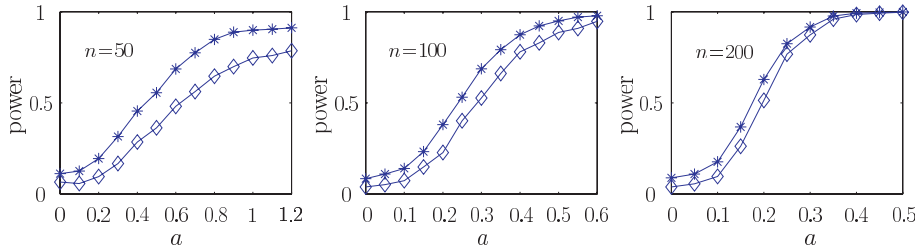

Figure 2. Power for Example 1 with $\sigma = 0.3$ and different values of a based on 1,000 replications. Asterisks are for $\alpha = 0.05$ and diamonds for $\alpha = 0.10$.

Example 2. We consider the data set from a study in which the respondents were adult residents in Los Angeles county. The major objective of the study was to provide estimates of the prevalence and incidence of depression and to identify causal factors and outcomes associated with the conditions. There were 294 respondents. See Afifi and Virinia (1984). Here we consider the regression of the state "depress" (Y) a person feels on "his/her age" (x_1) and "income" (x_2).

The original model for this data set is logistic, which is a single-index model with a known link function. Based on Figure 3(a), we take $\mathcal{D}$ sufficiently large to include all of the observations. We have $CVS_{\mathcal{D}} = 17.14$ and the $p-$value is $\Pr(CVS^*_{\mathcal{D}} > 17.14) = 0.613$. These values together with the histogram in Figure 3(b), suggest that the single-index model (see Figure 3(a)) fits the data set adequately.

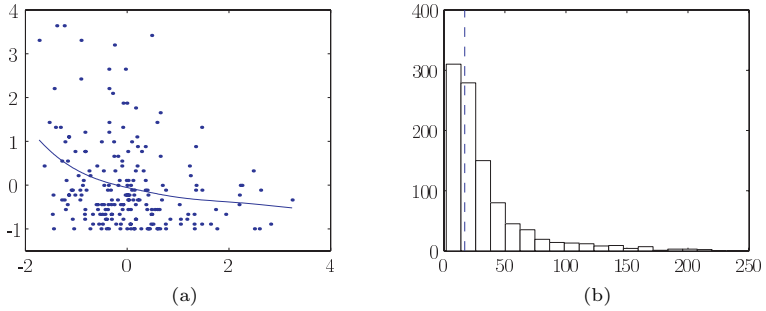

Figure 3. Calculation results for Example 2. (a) The estimated single-index function with estimate $\hat{\theta} = (0.4625,\ 0.8866)^T$ and $h = 0.73$. (b) The histogram of the bootstrap distribution, the dashed line is the value of the $CVS_{\mathcal{D}}$ statistic.

Example 3. Bell et al. (1989) (see Hastie and Tibshirani (1990, p.282)) studied multiple level thoracic and lumbar laminectomy, a corrective spinal surgery. The purpose of the study is to delineate the true incidence and nature of spinal deformities following this surgery and to assess the importance of age at the time

of surgery, as well as the effect of the number and location of vertebrae levels decompressed. The data in the study consist of 83 patients. The specific outcome of interest here is the relation of the presence of kyphosis (Y) with the age (x_1) in months at the time of the operation and the starting (x_2) and ending (x_3) range of vertebrae levels involved in the operation. The relation was previously fitted by a generalized additive model (Hastie and Tibshirani (1990, p.282)), which does not have a single-index form.

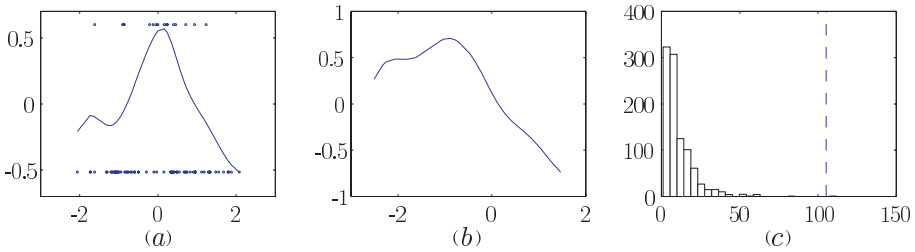

Figure 4. Calculation results for Example 3. (a) dot denotes the observation (a shift to Y is added to make it easier to be visualizable); solid line denotes the estimated single-index link function with estimate $\hat{\theta} = (0.9409,\ 0.3386)^T$ and $h = 0.4$. (b) The fitted single-index model with projection pursuit direction $(-0.2173,\ 0.9761)^T$ against the errors in (a). (c) The histogram of the bootstrap distribution, the dashed line being the value of the $CVS_{\mathcal{D}}$ statistic.

Based on Figure 4(a), we take $\mathcal{D}$ sufficiently large to include all of the observations. For simplicity, we only consider the relation of Y with $(x_1,\ x_3)$, for which $CVS_{\mathcal{D}} = 105.15$. The $p-$value is then $\Pr(CVS^*_{\mathcal{D}} > 105.15) = 0.003$. This, together with the histogram in Figure 4(c), suggests that we should reject the single-index model for the data set. Actually, the residual of the single-index model shows some relation with the covariates $(x_1,\ x_3)$ as in Figure 4(b). Therefore our results are consistent with the additive model fitted by Hastie and Tibshirani (1990) to this data set.

5. Conclusions

In this article, we have constructed a Cramér-von Mises test to check the goodness-of-fit of the popular single-index model. Choosing the kernel function appropriately, we can avoid the troublesome problem of having to oversmooth or undersmooth the underlying link function. Moreover, we can remove the bias term easily. To improve the power of the test for multivariate explanatory variables, we recommend the use of the single-indexing idea. From our proofs, the supremum type test statistics, such as $\sup_x |\tilde{S}_{\mathcal{D}}(x)|$, can also be employed. Their

distributions can be obtained by the bootstrap methods. Our limited simulations suggest that our method might also be used for other semiparametric models.

The proposed methods in this paper is computationally intensive because we need bootstrap to estimate the critical values for the test statistics. Developing more convenient testing methods is of interest. Another area for further investigation is the test for a multi-index model, as mentioned by one of the referees. It would also be interesting to verify the feasibility of the Neyman smooth test investigated by Ledwina (1994) and Fan (1996) under the present setting, and to compare their powers.

Appendix. Assumptions and Outline of the Proofs

(C1) $E|\varepsilon|^k < \infty$, $E\|X\|^k < \infty$ for all $k > 0$ and $\mathrm{var}(\varepsilon|X = x) = \sigma^2(x)$ is a continuous bounded function.

(C2) The density function f_θ of $\theta^T X$ has bounded fourth derivatives for all $\|\theta\| = 1$, and f_{θ_0} is bounded away from 0 on $\mathcal{D}$.

(C3) $E(Y|X = x)$ has bounded, continuous fourth derivatives.

(C4) The bandwidth h is proportional to $n^{-1/5}$.

(C5) $K(\cdot)$ is a symmetric probability density function with a compact support. Furthermore, the Fourier transformation of $K(\cdot)$ is absolutely integrable.

(C6) Under H_0, (C1)−(C2) and that g has bounded continuous derivative, $\hat{\theta}$ admits the expression

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} \sum_{i=1}^{n} \ell_n(X_i, g, \theta_0)\varepsilon_i + o_P(1),$$

where $\ell_n(X_i, g, \theta_0) = W_n^- w(X_i)\{X_i - C(X_i, \theta_0)\}g'(X_i^T\theta_0)$, $W_n = n^{-1}\sum_{i=1}^n w(X_i)\{X_i - C(X_i, \theta_0)\}\{X_i - C(X_i, \theta_0)\}^T g'(\theta_0^T X_i)^2$, W_n^- denotes a generalized inverse of W_n, and $C(x, \theta_0) = \sum_{j=1}^n K_h(\theta_0^T(X_j - x))X_j/(\sum_{j=1}^n K_h(\theta_0^T(X_j - x)))$ and $w(x) \geq 0$ is a bounded weight function.

The first part of assumption (C1) is made for simplicity of proof. See, for example, Härdle, Hall and Ichimura (1993). The existence of finite moments is sufficient. The second part is made to handle the boundary points of X. Since we only consider a finite region of g in many applications, this assumption is not unduly restrictive. If we standardize X, then X having a positive density function with a bounded fourth derivative near the origin can guarantee (C2). A drawback of (C2) is that it rules out dummy variables. Assumption (C3) is made to meet the requirement of continuity for kernel smoothing. As for the bandwidth assumption (C4), our results still hold for a larger range of bandwidths. However, the restriction will make the exposition easier. Since we use the data driven bandwidth, which is proportional to $n^{-1/5}$ asymptotically, the assumption is

quite natural. The kernel assumption (C5) is satisfied by the Gaussian kernel and the triweight kernel. For ease of exposition, we standardize the kernel such that the variance is 1. The expression in (C6) was given in Härdle, Hall and Ichimura (1993). The weight function $w(x)$ is used to handle points with small density. In Härdle, Hall and Ichimura (1993), $w(x) = 1$ when $x \in D_p$, 0 otherwise, where D_p is a compact region on which the density function $f(x)$ of X is continuous and positive. In Weisberg and Welsh (1994), $w(x)$ is a smooth version of the indicator function. We do not use their final expression in order to avoid some notational problem in the bootstrap counterpart. It is easy to see that $\ell_n(x, g, \theta_0)$ in (C6) can be replaced by

$$\ell(x,g,\theta_0)=\Big[\int w(z)\{z-\mu(z,\theta_0)\}\{z-\mu(z,\theta_0)\}^T g'(\theta_0^T z)^2 f(z)dz\Big]^{-}\{x-\mu(x,\theta_0)\},$$

where $\mu(x, \theta_0) = E(X|\theta_0^T X = \theta_0^T x)$. See the proof of Theorem 2 below. For some other estimation methods, such as Carroll et al. (1997), the estimators of θ admit a similar expression.

Let $\mathcal{B} = \{\theta \in \mathbb{R}^p : \|\theta - \theta_0\| \le Cn^{-1/2+\tau}\}$, where $C > 0$ is a constant and $0 < \tau < 1/10$. By (C6), we can restrict $\hat{\theta}$ to $\mathcal{B}$ almost surely. See also Weisberg and Welsh (1994). Similarly, we can restrict $\hat{\theta}^*$ to $\mathcal{B}$. See (A.39) below. For ease of exposition, we write $I_{\mathcal{D}}(X_i < x) = I(\theta_0^T X_i \in \mathcal{D})I(X_i < x)$.

Lemma A.1. *Under assumptions* (C1)−(C5), *we have*

$$\begin{aligned}
\hat{g}_\theta(v) &= g_\theta(v) + g'_\theta(v)(\theta_0-\theta)^T\mu_\theta(v) + \frac{1}{2}g''_\theta(v)h^2 \\
&\quad + \frac{1}{nf_\theta(v)}\sum_{i=1}^n K_h(X_i^T\theta - v)\varepsilon_i + O_P(h^3(\log n)^{1/2}), \\
\hat{g}^*_\theta(v) &= \hat{g}_{\hat{\theta}}(v) + g'_\theta(v)(\hat{\theta}-\theta)^T\mu_\theta(v) + \frac{1}{2}g''_\theta(v)h^2 + \frac{1}{nf_\theta(v)}\sum_{i=1}^n H_h(X_i^T\theta - v)\varepsilon_i \\
&\quad + \frac{1}{nf_\theta(v)}\sum_{i=1}^n K_h(X_i^T\theta - v)\varepsilon_i^* + O_P(h^3(\log n)^{1/2}), \\
\hat{g}'_\theta(v) &= g'_\theta(v) + \frac{1}{nhf_\theta(v)}\sum_{i=1}^n K'_h(X_i^T\theta - v)\varepsilon_i + O_P(h^2(\log n)^{1/2}), \\
\hat{g}''_\theta(v)h^2 &= g''_\theta(v)h^2 + \frac{1}{nf_\theta(v)}\sum_{i=1}^n K''_h(X_i^T\theta - v)\varepsilon_i + O_P(h^3(\log n)^{1/2})
\end{aligned}$$

uniformly for $\theta \in \mathcal{B}$ *and* $v \in \mathcal{D}$, *where* $\mu_\theta(v) = E(X|\theta^T X = v)$ *and* $H = K * K - K$.

Proof. Let $\mu_k = \int v^k K(v)dv$, $k = 0, 1, 2$, and $\delta_n = h^2(\log n)^{1/2}$, the same order as $\{\log n/(nh)\}^{1/2}$ under (C4). Following the steps of Lemma A.2 in Xia and Li

(1999) or Masry (1996), we have

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n} K_h(\theta^T X_i - v)\{(\theta^T X_i - v)/h\}^{\ell}\varepsilon_i = O_P(\delta_n),\\
&n^{-1}\sum_{i=1}^{n} K_h(\theta^T X_i - v)(\theta^T X_i - v)^{\ell} X_i = \mu_{\ell} h^{\ell} f_{\theta}(v)\mu_{\theta}(v) + O_P(h^{\ell}(h+\delta_n)),\\
&n^{-1}\sum_{i=1}^{n} K_h(\theta^T X_i - v)|\{(\theta^T X_i - v)/h\}^{\ell}| = O_P(1),\\
&s_{n,\theta,\ell}(v) = \mu_{\ell} f_{\theta}(v) + \mu_{\ell+1} f'_{\theta}(v)h + O_P(h^2+\delta_n),\ \ell = 0,1,2, \qquad \text{(A.1)}
\end{aligned}
$$

uniformly for $v \in \mathcal{D}$ and $\theta \in \mathcal{B}$ where $\mu_{\ell} = \int K(v)v^{\ell}dv$. By writing

$$
\begin{aligned}
y_i &= g_{\theta}(\theta^T X_i) + g'_{\theta}(\theta^T X_i)(\theta_0 - \theta)^T X_i + \varepsilon_i + O_P(n^{-1+2\tau})\\
&= g_{\theta}(v) + g'_{\theta}(v)(\theta^T X_i - v) + \frac{1}{2} g''_{\theta}(v)(\theta^T X_i - v)^2 + g'_{\theta}(v)(\theta_0 - \theta)^T X_i\\
&\quad + \varepsilon_i + O_P\{|\theta^T X_i - v|^3 + |\theta^T X_i - v| n^{-1/2+\tau} + n^{-1+2\tau}\},
\end{aligned}
$$

the first part of Lemma 1 follows from (2.3), the set of equations in (A.1) and the fact that $h^3 = o(n^{-1/2})$ and $h\delta_n = o(n^{-1/2})$. Similarly, write

$$
\begin{aligned}
y_i^* &= g_{\theta}(\theta^T X_i) + g'_{\theta}(\theta^T X_i)(\theta_0 - \hat{\theta})^T \mu_{\theta}(\theta^T X_i) + g'_{\theta}(\theta^T X_i)(\hat{\theta} - \theta)^T X_i\\
&\quad + \frac{1}{2} g''_{\theta}(\theta^T X_i)h^2 + \frac{1}{n f_{\theta}(\hat{\theta}^T X_i)} \sum_{j=1}^{n} K_h(\hat{\theta}^T(X_j - X_i))\varepsilon_j + \varepsilon_i^* + O_P\{|\theta^T X_i - v|^3\\
&\quad + |\theta^T X_i - v| n^{-1/2+\tau} + h^3 (\log n)^{1/2} + n^{-1+2\tau}\}.
\end{aligned}
$$

The second part of Lemma 1 follows from (2.3), (A.1) and the fact that

$$
\begin{aligned}
&\frac{1}{n f_{\theta}(\hat{\theta}^T X_i)} \sum_{j=1}^{n} K_h(\hat{\theta}^T(X_j - X_i))\varepsilon_j\\
&= \frac{1}{n f_{\theta}(\theta^T X_i)} \sum_{j=1}^{n} K_h(\theta^T(X_j - X_i))\varepsilon_j + O_P(h^3(\log n)^{1/2}),\\
&\frac{1}{n}\sum_{i=1}^{n} K_h(\theta^T X_i - v) \frac{1}{n f_{\theta}(\theta^T X_i)} \sum_{j=1}^{n} K_h(\theta^T(X_j - X_i))\varepsilon_j\\
&= n^{-1}\sum_{i=1}^{n} H_h(\theta^T X_i - v)\varepsilon_i + O_P(h\delta_n).
\end{aligned}
$$

See Xia and Li (1999) and Linton and Nielsen (1994). Next, we prove the last part. Let $\delta_n = h^2(\log n)^{1/2}$, the same order as $\{\log n/(nh)\}^{1/2}$ under (C4). For ease of exposition, all the derivatives in the following context of the proof are

taken with respect to v. Write $K_{(k)}(v) = K(v)v^k$ and $K_{(k),h}(v) = h^{-1}K(v/h)\{v/h\}^k$. Let $s_{n,\theta,k}(v) = n^{-1}\sum_{i=1}^n K_{(k),h}(\theta^T X_i - v)$, where $k = 0, 1, 2, 3$ and

$$\begin{aligned}
\tilde{g}_\theta(X_i) &= g(X_i^T\theta_0) - g_\theta(X_i^T\theta),\\
\Delta_{\theta,i}(v) &= g_\theta(X_i^T\theta) - \{g_\theta(v) + g'_\theta(v)(X_i^T\theta - v) + \frac{1}{2}g''_\theta(v)(X_i^T\theta - v)^2\},\\
W_{n,h}(X_i^T\theta - v) &= s_{n,\theta,2}(v)n^{-1}K_h(X_i^T\theta - v) - s_{n,\theta,1}(v)n^{-1}K_{(1),h}(X_i^T\theta - v),\\
D_{n,\theta,0}(v) &= s_{n,\theta,0}(v)s_{n,\theta,2}(v) - s^2_{n,\theta,1}(v),\\
D_{n,\theta,2}(v) &= s^2_{n,\theta,2}(v) - s_{n,\theta,1}(v)s_{n,\theta,3}(v).
\end{aligned}$$

Since $\sum_{i=1}^n W_{n,h}(X_i^T\theta - v)(X_i^T\theta - v) = 0$, the estimate of $g_\theta(v)$ can be written as (see, e.g., Fan and Gijbels (1996))

$$\begin{aligned}
\hat{g}_\theta(v) &= \frac{\sum_{i=1}^n W_{n,h}(X_i^T\theta - v)Y_i}{D_{n,\theta,0}(v)}\\
&= g_\theta(v) + \frac{1}{2}g''_\theta(v)h^2\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)} + \sum_{i=1}^n \frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\Delta_{\theta,i}(v)\\
&\quad + \sum_{i=1}^n \frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\tilde{g}_\theta(X_i) + \sum_{i=1}^n \frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\varepsilon_i.
\end{aligned}$$

It is easy to see that $\Delta'_{\theta,i}(v) = -(1/2)g^{(3)}_\theta(v)(X_i^T\theta - v)^2$, $\Delta''_{\theta,i}(v) = g^{(3)}_\theta(v)(X_i^T\theta - v) - (1/2)g^{(4)}_\theta(v)(X_i^T\theta - v)^2$. We have

$$\begin{aligned}
\hat{g}''_\theta(v) &= g''_\theta(v) + \frac{1}{2}g^{(4)}_\theta(v)h^2\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)} + g^{(3)}_\theta(v)h^2\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)'\\
&\quad + \frac{1}{2}g''_\theta(v)h^2\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)'' + \sum_{i=1}^n \frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\Delta''_{\theta,i}(v)\\
&\quad + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)'\Delta'_{\theta,i}(v) + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)''\Delta_{\theta,i}(v)\\
&\quad + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)''\tilde{g}_\theta(X_i) + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)''\varepsilon_i\\
&= g''_\theta(v) + g^{(3)}_\theta(v)h^2\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)' + \frac{1}{2}g''_\theta(v)h^2\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)''\\
&\quad - g^{(3)}_\theta(v)h^2\sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)'\{(X_i^T\theta - v)/h\}^2\\
&\quad + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)''\Delta_{\theta,i}(v) + \sum_{i=1}^n \left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)''\tilde{g}_\theta(X_i)
\end{aligned}$$

$$+\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\varepsilon_i. \tag{A.2}$$

Note that

$$\begin{aligned}W'_{n,h}(X_i^T\theta-v) &= s'_{n,\theta,2}(v)n^{-1}K_h(X_i^T\theta-v)-s'_{n,\theta,1}(v)n^{-1}K_{(1),h}(X_i^T\theta-v)\\ &\quad -s_{n,\theta,2}(v)(nh)^{-1}K'_h(X_i^T\theta-v)+s_{n,\theta,1}(v)(nh)^{-1}K'_{(1),h}(X_i^T\theta-v), \quad (A.3)\\ W''_{n,h}(X_i^T\theta-v) &= s''_{n,\theta,2}(v)n^{-1}K_h(X_i^T\theta-v)-s''_{n,\theta,1}(v)n^{-1}K_{(1),h}(X_i^T\theta-v)\\ &\quad -2(nh)^{-1}\{s'_{n,\theta,2}(v)K'_h(X_i^T\theta-v)-s'_{n,\theta,1}(v)K'_{(1),h}(X_i^T\theta-v)\}\\ &\quad +(nh^2)^{-1}\{s_{n,\theta,2}(v)K''_h(X_i^T\theta-v)-s_{n,\theta,1}(v)K''_{(1),h}(X_i^T\theta-v)\}. \quad (A.4)\end{aligned}$$

Then by Lemma A.2 of Xia and Li (1999), for any integrable function $\tilde{K}$ with $\int\tilde{K}(v)v^2dv\le\infty$ and any continuous function $m_\theta(v,u)$ with $E(m_\theta(X_i,\varepsilon_i)|X_i)=0$ almost surely and $E|m_\theta(X_i,\varepsilon_i)|^r<\infty$ for all $r>0$, we have

$$n^{-1}\sum_{i=1}^{n}\tilde{K}_h(\theta^TX_i-v)=f_\theta(v)\int\tilde{K}(v)dv+hf'_\theta(v)\int\tilde{K}(v)vdv+O_P(\delta_n), \tag{A.5}$$

$$n^{-1}\sum_{i=1}^{n}\tilde{K}_h(\theta^TX_i-v)m_\theta(X_i,\varepsilon_i)=O_P(\delta_n), \tag{A.6}$$

uniformly for $\theta\in\mathcal{B}$ and $v\in\mathcal{D}$. By assumption (C5), simple calculation leads to

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n}K_h(X_i^T\theta-v)=f_\theta(v)+O_P(\delta_n), \quad (A.7)\\ &n^{-1}\sum_{i=1}^{n}K_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}=f'_\theta(v)h+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}^2=f_\theta(v)+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}^3=k_4f'_\theta(v)h+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K'_h(X_i^T\theta-v)=-f'_\theta(v)h+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K'_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}=-f_\theta(v)+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K'_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}^2=-3f'_\theta(v)h+O_P(\delta_n),\\ &n^{-1}\sum_{i=1}^{n}K'_h(X_i^T\theta-v)\{(X_i^T\theta-v)/h\}^3=-3f_\theta(v)+O_P(\delta_n),\end{aligned}$$

$$n^{-1}\sum_{i=1}^{n} K_h''(X_i^T\theta - v) = O_P(h),$$
$$n^{-1}\sum_{i=1}^{n} K_h''(X_i^T\theta - v)\{(X_i^T\theta - v)/h\} = O_P(h),$$
$$n^{-1}\sum_{i=1}^{n} K_h''(X_i^T\theta - v)\{(X_i^T\theta - v)/h\}^2 = O_P(1),$$
$$n^{-1}\sum_{i=1}^{n} K_h''(X_i^T\theta - v)\{(X_i^T\theta - v)/h\}^3 = O_P(h),$$

where $k_4 = \int v^4 K(v)dv$. It follows that

$$\begin{aligned}
D_{n,\theta,0}(v) &= f_\theta^2(v) + O_P(\delta_n),\\
D'_{n,\theta,0}(v) &= s'_{n,\theta,2}(v)s_{n,\theta,0}(v) + s_{n,\theta,2}(v)s'_{n,\theta,0}(v) - 2s_{n,\theta,1}(v)s'_{n,\theta,1}(v)\\
&= h^{-1}\{2f_\theta(v)f'_\theta(v)h + O_P(\delta_n)\},\\
D''_{n,\theta,0}(v) &= s''_{n,\theta,2}(v)s_{n,\theta,0}(v) + s_{n,\theta,2}(v)s''_{n,\theta,0}(v) + 2s'_{n,\theta,2}(v)s'_{n,\theta,0}(v)\\
&\quad -2s_{n,\theta,1}(v)s''_{n,\theta,1}(v) - 2\{s'_{n,\theta,1}(v)\}^2\\
&= h^{-2}O_P(\delta_n),\\
D_{n,\theta,2}(v) &= f_\theta^2(v) + O_P(\delta_n),\\
D'_{n,\theta,2}(v) &= 2s_{n,\theta,2}(v)s'_{n,\theta,2}(v) - s'_{n,\theta,1}(v)s_{n,\theta,3}(v) - s'_{n,\theta,1}(v)s'_{n,\theta,3}(v)\\
&= h^{-1}\{2f_\theta(v)f'_\theta(v)h + O_P(\delta_n)\},\\
D''_{n,\theta,2}(v) &= 2\{s'_{n,\theta,2}(v)\}^2 + 2s_{n,\theta,2}(v)s''_{n,\theta,2}(v) - 2s'_{n,\theta,}(v)s'_{n,\theta,3}(v)\\
&\quad -s_{n,\theta,1}(v)s''_{n,\theta,3}(v) - s''_{n,\theta,1}(v)s_{n,\theta,3}(v)\\
&= h^{-2}O_P(\delta_n) = O_P((\log n)^{1/2}).
\end{aligned}$$

Therefore

$$\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)' = D_{n,\theta,0}^{-2}(v)\{D'_{n,\theta,2}(v)D_{n,\theta,0}(v) - D_{n,\theta,2}(v)D'_{n,\theta,0}(v)\} = O_P(1), \quad \text{(A.8)}$$

$$\begin{aligned}
\left(\frac{D_{n,\theta,2}(v)}{D_{n,\theta,0}(v)}\right)'' &= \{D_{n,\theta,0}^{-1}(v)\}''D_{n,\theta,2}(v) + 2\{D_{n,\theta,0}^{-1}(v)\}'D'_{n,\theta,2}(v)\\
&\quad + D_{n,\theta,0}^{-1}(v)D''_{n,\theta,2}(v) = O_P((\log n)^{1/2}), \quad \text{(A.9)}
\end{aligned}$$

$$\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta - v)}{D_{n,\theta,0}(v)}\right)'\{(X_i^T\theta - v)/h\}^2 = O_P(1). \quad \text{(A.10)}$$

Similarly, by (A.5), (A.3) and (A.4), we have

$$\sum_{i=1}^{n}|W_{n,h}(X_i^T\theta - v)\{X_i^T\theta - v)\}^3| = O_P(h^3),$$

$$\sum_{i=1}^{n}|W'_{n,h}(X_i^T\theta-v)\{X_i^T\theta-v)\}^3|=O_P(h^2),$$
$$\sum_{i=1}^{n}|W''_{n,h}(X_i^T\theta-v)\{X_i^T\theta-v)\}^3|=O_P(h).$$

Note that $|\Delta_{\theta,i}(v)|\le c(X_i^T\theta-v)^3$, where c is a constant. Therefore

$$\begin{aligned}
&|\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\Delta_{\theta,i}(v)|\\
&\le c\{D_{n,\theta,0}^{-1}(v)\}''\sum_{i=1}^{n}|W_{n,h}(X_i^T\theta-v)\{X_i^T\theta-v\}^3|\\
&\quad+2c\{D_{n,\theta,0}^{-1}(v)\}'\sum_{i=1}^{n}|W'_{n,h}(X_i^T\theta-v)\{X_i^T\theta-v\}^3|\\
&\quad+cD_{n,\theta,0}^{-1}(v)\sum_{i=1}^{n}|W''_{n,h}(X_i^T\theta-v)\{X_i^T\theta-v\}^3|\\
&=O_P(h). \qquad\text{(A.11)}
\end{aligned}$$

Similarly, by (A.6), (A.3) and (A.4), we have

$$\sum_{i=1}^{n}W_{n,h}(X_i^T\theta-v)\varepsilon_i=O_P(\delta_n),\quad \sum_{i=1}^{n}W'_{n,h}(X_i^T\theta-v)\varepsilon_i=O_P(h(\log n)^{1/2}),$$
$$\sum_{i=1}^{n}W''_{n,h}(X_i^T\theta-v)\varepsilon_i=f_\theta(v)n^{-1}h^{-2}\sum_{i=1}^{n}K''_h(X_i^T\theta-v)\varepsilon_i+O_P(h(\log n)^{1/2}).$$

Therefore

$$\begin{aligned}
\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\varepsilon_i&=\{D_{n,\theta,0}^{-1}(v)\}''\sum_{i=1}^{n}W_{n,h}(X_i^T\theta-v)\varepsilon_i\\
&\quad-2\{D_{n,\theta,0}^{-1}(v)\}'\sum_{i=1}^{n}W'_{n,h}(X_i^T\theta-v)\varepsilon_i+D_{n,\theta,0}^{-1}(v)\sum_{i=1}^{n}W_{n,h}(X_i^T\theta-v)\varepsilon_i\\
&=h^{-2}\frac{1}{nf_\theta(v)}\sum_{i=1}^{n}K''_h(X_i^T\theta-v)\varepsilon_i+O_P(h(\log n)^{1/2}). \qquad\text{(A.12)}
\end{aligned}$$

By (4.9) and (4.10) of Härdle, Hall and Ichimura (1993), we have

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\tilde g(X_i)\\
&=(\theta_0-\theta)^T\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\{X_i-E(X_i|X_i^T\theta)\}g'_\theta(X_i^T\theta)
\end{aligned}$$

$$+\sum_{i=1}^{n}\Big|\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\Big|O_P(n^{-1+2\tau}).$$

Note that $E(\{X_i-E(X_i|X_i^T\theta)\}g_\theta(X_i^T\theta)|X_i^T\theta)=0$ almost surely. By the same approach which leads to (A.11) and (A.12), we have

$$\sum_{i=1}^{n}\left(\frac{W_{n,h}(X_i^T\theta-v)}{D_{n,\theta,0}(v)}\right)''\tilde{g}(X_i)=O_P(h). \tag{A.13}$$

Finally, combining (A.2), (A.8) (A.9), (A.10), (A.11), (A.12) and (A.13), we have $\hat{g}''_\theta(v)-g''_\theta(v)=\{nf_\theta(v)h^2\}^{-1}\sum_{i=1}^n K''_h(X_i^T\theta-v)\varepsilon_i+O_P(h(\log n)^{1/2})$. This completes the proof of the last part of Lemma A.1.

Lemma A.2. *Let* $B_i=E[\hat{g}^*_{\hat\theta}(X_i^T\hat\theta)-\hat{g}_{\hat\theta}(X_i^T\hat\theta)\Big|(X_j,Y_j),\ j=1,\ldots,n]$. *Then under assumptions* (C1)−(C6), *we have*

$$B_i=\frac{1}{2}g''_{\hat\theta}(\hat\theta^TX_i)h^2+\frac{1}{nf_{\hat\theta}(\hat\theta^TX_i)}\sum_{j=1}^{n}H_h(\hat\theta^T(X_j-X_i))\varepsilon_i+O_P(h^3(\log n)^{1/2}),$$

$$n^{-1/2}\sum_{X_i^T\hat\theta\in\mathcal{D}}\left[\frac{1}{2}g''_{\hat\theta}(X_i^T\hat\theta)h^2-B_i\right]I(X_i<x)=o_P(1),$$

uniformly for $x\in\bar{\mathbb{R}}^{\otimes p}$.

Proof. By Lemma A.1, we have

$$\hat{g}^*_{\hat\theta}(X_i^T\hat\theta)=\hat{g}_{\hat\theta}(X_i^T\hat\theta)+\frac{1}{2}\hat{g}''_{\hat\theta}(X_i^T\hat\theta)h^2+\frac{1}{nf_{\hat\theta}(v)}\sum_{i=1}^{n}H_h(X_i^T\hat\theta-v)\varepsilon_i$$
$$+\frac{1}{nf_{\hat\theta}(X_i^T\hat\theta)}\sum_{j=1}^{n}K_h(X_j^T\hat\theta-X_i^T\hat\theta)\varepsilon^*_j+O_P(h^3(\log n)^{1/2}) \tag{A.14}$$

uniformly for $X_i^T\hat\theta\in\mathcal{D}$. By the definition of B_i, the first part of Lemma A.2 follows automatically from $n^{1/2}h^3(\log n)^{1/2}=o_P(1)$. By Lemma A.1, we have

$$\frac{1}{2}g''_{\hat\theta}(X_i^T\hat\theta)h^2-B_i=\frac{1}{nf_{\hat\theta}(X_i^T\hat\theta)}\sum_{j=1}^{n}H_h(X_j^T\hat\theta-X_i^T\hat\theta)\varepsilon_j+O_P(h^3(\log n)^{1/2}) \tag{A.15}$$

uniformly for all $X_i^T\hat\theta\in\mathcal{D}$. Therefore

$$n^{-1/2}\sum_{X_i^T\hat\theta\in\mathcal{D}}\left[\frac{1}{2}g''_{\hat\theta}(X_i^T\hat\theta)h^2-B_i\right]I(X_i<x)$$
$$=n^{-3/2}\sum_{j=1}^{n}\varepsilon_j\sum_{X_i^T\hat\theta\in\mathcal{D}}H_h(X_j^T\hat\theta-X_i^T\hat\theta)f^{-1}_{\hat\theta}(X_i^T\hat\theta)I(X_i<x)+o_P(1). \tag{A.16}$$

Let $G_\theta(v) = E_{\mathcal{D}}\{H_h(v - X_i^T\theta)f_\theta^{-1}(X_i^T\theta)I(X_i < x)\}$. Then

$$\begin{aligned} & n^{-3/2}\sum_{j=1}^{n}\varepsilon_j \sum_{X_i^T\theta\in\mathcal{D}} H_h(X_j^T\theta - X_i^T\theta)f_\theta^{-1}(X_i^T\theta)I(X_i < x) \\ =\ & n^{-3/2}\sum_{j=1}^{n}\varepsilon_j \sum_{X_i^T\theta\in\mathcal{D}} \Big[H_h(X_j^T\theta - X_i^T\theta)f_\theta^{-1}(X_i^T\theta)I(X_i < x) - G(X_j^T\theta)\Big] \\ & + n^{-1/2}\sum_{j=1}^{n}\varepsilon_j G_\theta(X_j^T\theta). \end{aligned} \tag{A.17}$$

Because $\int H(v)dv = 0$, we have $E_{\mathcal{D}}\left\{H_h((x - X_i)^T\theta)f_\theta^{-1}(X_i^T\theta)I(X_i < x)\right\} = O(h)$, uniformly for all $x \in \bar{\mathbb{R}}^{\otimes p}$ and $\theta \in \mathcal{B}$. By Lemma A.2 of Xia and Li (1999), the second term on the right hand side of (A.17) is $o_P(1)$ uniformly for $\theta \in \mathcal{B}$ and $x \in \bar{\mathbb{R}}^{\otimes p}$. The first term on the right hand side of (A.17) is also $o_P(1)$ uniformly for $\theta \in \mathcal{B}$ and $x \in \bar{\mathbb{R}}^{\otimes p}$. Therefore

$$n^{-3/2}\sum_{j=1}^{n}\varepsilon_j \sum_{X_i^T\theta\in\mathcal{D}} H_h(X_j^T\theta - X_i^T\theta)f_\theta^{-1}(X_i^T\theta)I(X_i < x) = o_P(1) \tag{A.18}$$

uniformly for $\theta \in \mathcal{B}$ and $x \in \bar{\mathbb{R}}^{\otimes p}$. The second part of Lemma A.2 follows from (A.18) and (A.16).

Lemma A.3. *Let $\mathcal{I}_i(\theta) = I(X_i^T\theta \in \mathcal{D}) - I(X_i^T\theta_0 \in \mathcal{D})$. Under assumptions* (C1)−(C5), *we have $n^{-1/2}\sup_{\theta\in\mathcal{B},x\in\bar{\mathbb{R}}^{\otimes p}}\sum_{i=1}^{n}\mathcal{I}_i(\theta)\varepsilon_i I(X_i < x) = O_P(n^{-1/4+\tau'/2})$ for any $\tau' > \tau$. Furthermore, if ξ_i is any measurable function of X_i and has finite second moment, then $n^{-1/2}\sup_{\theta\in\mathcal{B},x\in\bar{\mathbb{R}}^{\otimes p}}\sum_{i=1}^{n}\mathcal{I}_i(\theta)\xi_i I(X_i < x) = O_P(n^{\tau'})$.*

Proof. We follow the method of "continuous argument" as in Härdle, Hall and Ichimura (1993). See also Xia and Li (1999) and Masry (1996). Let $c_1, c_2, \ldots$ be positive constants in the following. Let $F_1, \ldots, F_p$ be the distributions of the components of $X = (x_1, \ldots, x_p)^T$, respectively. Let $Z_i = (z_{i1}, \ldots, z_{ip})^T = (F_1(x_{i1}), \ldots, F_p(x_{ip}))^T$. Then z_{ik}, $k = 1, \ldots, p$, are uniformly distributed on [0, 1]. Note that

$$\begin{aligned} \sum_{n=1}^{\infty} P(|\varepsilon_n| > n^{\delta_1}) &\le \sum_{n=1}^{\infty} E|\varepsilon_n|^t n^{-(t-1)\delta_1} < \infty, \\ \sum_{n=1}^{\infty} P(\|X_n\| > n^{\delta_1}) &\le \sum_{n=1}^{\infty} E\|X_n\|^t n^{-(t-1)\delta_1} < \infty \end{aligned}$$

for any $\delta_1 > 0$ by taking t sufficiently large. By the Borel-Cantelli Lemma and that n^{δ_1} is increasing in n, we need only consider the summation in Lemma A.3

on $\{\|X_i\| \le n^{\delta_1}\} \cap \{|\varepsilon_i| \le n^{\delta_1}\}$, i.e., $\sup_{\theta\in\mathcal{B},z\in[0,1]^{\otimes p}} n^{-1/2}\sum_{i=1}^n \mathcal{I}_i(\theta)\tilde{\varepsilon}_i I(Z_i < z)I(\|X_i\| < n^{\delta_1})$, where $\tilde{\varepsilon}_i = \varepsilon_i I(|\varepsilon_i| \le n^{\delta_1})$ (cf. Masry (1996)). Let $S_n(\theta, z) = n^{-1/2}\sum_{i=1}^n \mathcal{I}_i(\theta)\tilde{\varepsilon}_i I(Z_i < z)I(\|X_i\| < n^{\delta_1})$. Now, consider the bounded set $\mathcal{B}\otimes[0,1]^{\otimes p}$. It is easy to see that there are n^{2p} balls B_{n_k} with diameters less than $c_1 n^{-1}$ and center (θ_{n_k}, z_{n_k}) (in $\mathcal{B}\otimes[0,1]^{\otimes p}$) such that $\mathcal{B}\otimes[0,1]^{\otimes p} \subset \cup_{1\le k\le n^{2p}} B_{n_k}$. Then

$$\sup_{\substack{\theta\in\mathcal{B},\\ z\in[0,1]^{\otimes p}}} |S_n(\theta,z)| \le \max_{1\le k\le n^{2p}} |S_n(\theta_{n_k},z_{n_k})| + \max_{1\le k\le n^{2p}} \sup_{(\theta,z)\in B_{n_k}} |S_n(\theta,z) - S_n(\theta_{n_k},z_{n_k})|. \tag{A.19}$$

Note that for $(\theta, z) \in B_{n_k}$ and $\|X_i\| \le n^{\delta_1}$, we have $|(\theta-\theta_{n_k})^T X_i| \le c_1 n^{-1+\delta_1}$ and $\theta_{n_k}^T X_i - c_1 n^{-1+\delta_1} \le \theta^T X_i \le \theta_{n_k}^T X_i + c_1 n^{-1+\delta_1}$. Hence

$$\begin{aligned}
&|S_n(\theta,z) - S_n(\theta_{n_k},z_{n,k})| \\
&\le n^{-1/2}\sum_{i=1}^n |I(\theta^T X_i \in \mathcal{D}) - I(\theta_{n_k}^T X_i \in \mathcal{D})| I(\|X_i\| \le n^{\delta_1})|\tilde{\varepsilon}_i| \\
&\le n^{-1/2}\sum_{i=1}^n I(a - c_1 n^{-1+\delta_1} \le \theta_{n_k}^T X_i < a + c_1 n^{-1+\delta_1})|\tilde{\varepsilon}_i| \\
&\quad + n^{-1/2}\sum_{i=1}^n I(b - c_1 n^{-1+\delta_1} \le \theta_{n_k}^T X_i < b + c_1 n^{-1+\delta_1})|\tilde{\varepsilon}_i| \\
&\triangleq Q_{1,n,k} + Q_{2,n,k}.
\end{aligned} \tag{A.20}$$

Note that by assumptions (C1) and (C2),

$$EQ_{1,n,k} \le E|\tilde{\varepsilon}_i| n^{1/2} \int_{a-c_1 n^{-1+\delta_1}}^{a+c_1 n^{-1+\delta_1}} f_{\theta_{n_k}}(v)dv \le c_2 n^{-1/2+\delta_1}, \tag{A.21}$$

$$EQ_{2,n,k} \le E|\tilde{\varepsilon}_i| n^{1/2} \int_{b-c_1 n^{-1+\delta_1}}^{b+c_1 n^{-1+\delta_1}} f_{\theta_{n_k}}(v)dv \le c_3 n^{-1/2+\delta_1}. \tag{A.22}$$

Since $E\varepsilon_i = 0$, we have $|E\tilde{\varepsilon}_i| = |E\{\varepsilon_i I(|\varepsilon_i| > n^{\delta_1})\}| \le E|\varepsilon_i|^t n^{-(t-1)\delta_1}$ for any t and $|ES_n(\theta_{n_k}, z_{n_k})| \le E|\varepsilon_i|^t n^{-(t-1)\delta_1 - 1/2+1}$. Choosing t sufficiently large, $ES_n(\theta_{n_k}, z_{n_k})$ is negligible as $n \to \infty$. By (A.19) and (A.20), to finish the proof, it is sufficient to show that

$$\max_{1\le k\le n^{2p}} |S_n(\theta_{n_k},z_{n_k}) - ES_n(\theta_{n_k},z_{n_k})| = O_P(n^{-1/4+\tau'/2}), \tag{A.23}$$

$$\max_{1\le k\le n^{2p}} (Q_{1,n,k} - EQ_{1,n,k}) = O_P(n^{-1/2+\delta'}), \tag{A.24}$$

$$\max_{1\le k\le n^{2p}} (Q_{2,n,k} - EQ_{2,n,k}) = O_P(n^{-1/2+\delta'}), \tag{A.25}$$

where δ' is any small positive number. To save space, we only give the details for (A.25), which are more complicated than those for (A.23). Equation (A.24) can be proved similarly. Let $w_{k,i} = I(b - c_1 n^{-1+\delta_1} \le \theta_{n_k}^T X_i < b + c_1 n^{-1+\delta_1})|\tilde{\varepsilon}_i|$. Note that $Var\{n^{-1/2}\sum_{i=1}^n w_{k,i}\} = c_4 n^{-1+\delta_1}$. By Bernstein's inequality (cf. Chow and Teicher (1988, p.111)), we have

$$\Pr(|\sum_{j=1}^n \{w_{k,i} - E(w_{k,i})\}| > n^{\delta_2}) \le 2\exp\Big(\frac{-n^{2\delta_2}}{2(c_4 n n^{-1+\delta_1} + c_5 n^{\delta_1+\delta_2})}\Big).$$

Let $\delta_2 > \delta_1$, we have $\Pr(|\sum_{j=1}^n \{w_{k,i} - E(w_{k,i})\}| > n^{\delta_2}) \le c_6 \exp(-c_7 n^{\delta_3})$, where $\delta_3 > 0$. Thus

$$\begin{aligned}
&\Pr(\max_{1\le k\le n^{2p}} |\sum_{j=1}^n \{w_{k,i} - E(w_{k,i})\}| > n^{\delta_2}) \\
\le &\sum_{1\le k\le n^{2p}} \Pr(|\sum_{j=1}^n \{w_{k,i} - E(w_{k,i})\}| > n^{\delta_2}) \\
\le &\, c_6 n^{2p} \exp(-c_7 n^{\delta_3}) \to 0. \qquad \text{(A.26)}
\end{aligned}$$

Note that $\delta_2 > 0$ can be chosen to be sufficiently small. Equation (A.25) follows from (A.26).

For ease of exposition, we use $\sum_{\mathcal{D}}$ to denote $\sum_{X_i^T\theta_0\in\mathcal{D}}$ throughout the rest of this section.

Lemma A.4. *Under assumptions* (C1)-(C6), *we have* $S_{\mathcal{D}}(x) = n^{-1/2}\sum_{\mathcal{D}}(Y_i - \hat{Y}_i)I(X_i < x) + o_P(1)$, $\tilde{S}_{\mathcal{D}}(x) = n^{-1/2}\sum_{\mathcal{D}}(Y_i - \tilde{Y}_i)I(X_i < x) + o_P(1)$ *and* $\tilde{S}^*_{\mathcal{D}}(x) = n^{-1/2}\sum_{\mathcal{D}}(Y_i^* - \tilde{Y}_i^*)I(X_i < x) + o_P(1)$ *uniformly for* $x \in \bar{\mathbb{R}}^{\otimes p}$.

Proof. We only prove the first two equations here. Write

$$\begin{aligned}
S_{\mathcal{D}}(x) = &\, n^{-1/2} \sum_{X_i^T\hat{\theta}\in\mathcal{D}} \varepsilon_i I(X_i < x) \\
&+ n^{-1/2} \sum_{X_i^T\hat{\theta}\in\mathcal{D}} \Big[g(X_i^T\theta_0) - \hat{g}_{\theta_0}(X_i^T\theta_0)\Big] I(X_i < x) \\
&+ n^{-1/2} \sum_{X_i^T\hat{\theta}\in\mathcal{D}} \Big[\hat{g}_{\theta_0}(X_i^T\theta_0) - \hat{g}_{\hat{\theta}}(X_i^T\hat{\theta})\Big] I(X_i < x) \\
\triangleq &\, R_A + R_B + R_C,
\end{aligned}$$

$$\begin{aligned}
&n^{-1/2}\sum_{\mathcal{D}}(Y_i - \hat{Y}_i)I(X_i < x) \\
= &\, n^{-1/2}\sum_{\mathcal{D}} \varepsilon_i I(X_i < x) + n^{-1/2}\sum_{\mathcal{D}} \Big[g(X_i^T\theta_0) - \hat{g}_{\theta_0}(X_i^T\theta_0)\Big] I(X_i < x)
\end{aligned}$$

$$+n^{-1/2}\sum_{\mathcal{D}}\left[\hat{g}_{\theta_0}(X_i^T\theta_0)-\hat{g}_{\hat{\theta}}(X_i^T\hat{\theta})\right]I(X_i<x)$$
$$\stackrel{\Delta}{=} R_1+R_2+R_3. \tag{A.27}$$

We have by Lemma A.3,

$$R_A-R_1=n^{-1/2}\sum_{i=1}^{n}\mathcal{I}_i(\hat{\theta})\varepsilon_i I(X_i<x) \;=\; o_P(1) \tag{A.28}$$

uniformly for $x\in\bar{\mathbb{R}}^{\otimes p}$. By Lemma A.1, we have

$$\begin{aligned}R_B-R_2&=n^{-1/2}\sum_{i=1}^{n}\mathcal{I}_i(\hat{\theta})\left[g(X_i^T\theta_0)-\hat{g}_{\theta_0}(X_i^T\theta_0)\right]I(X_i<x)\\&=n^{-1/2}\sum_{i=1}^{n}\mathcal{I}_i(\hat{\theta})g(\theta_0^TX_i)\mu_{\theta_0}(\theta_0^TX_i)(\hat{\theta}-\theta_0)I(X_i<x)\\&\quad-\frac{1}{2}n^{-1/2}h^2\sum_{i=1}^{n}\mathcal{I}_i(\hat{\theta})g''(X_i^T\theta_0)I(X_i<x)\\&\quad-n^{-3/2}\sum_{j=1}^{n}\varepsilon_j\sum_{i=1}^{n}K_h(X_j^T\theta_0-X_i^T\theta_0)\mathcal{I}_i(\hat{\theta})f_{\theta_0}^{-1}(X_i^T\theta_0)I(X_i<x)+o_P(1).\end{aligned}$$

By Lemma A.3, the first term on the right hand side above is $o_P(1)$. By Lemma A.3 of Xia and Li (1999), the second term on the right hand side is $o_P(1)$. Hence

$$R_B-R_2=o_P(1) \tag{A.29}$$

uniformly for $x\in\bar{\mathbb{R}}^{\otimes p}$. By Lemma A.1 we have

$$\begin{aligned}\hat{g}_{\hat{\theta}}(v)&=g_{\hat{\theta}}(v)+g'_{\hat{\theta}}(v)(\theta_0-\hat{\theta})^T\mu_{\hat{\theta}}(v)+\frac{1}{2}g''_{\hat{\theta}}(v)h^2\\&\quad+\frac{1}{nf_{\hat{\theta}}(v)}\sum_{i=1}^{n}K_h(X_i^T\hat{\theta}-v)\varepsilon_i+O_P(h^3(\log n)^{1/2})\end{aligned}$$

uniformly for $v\in\mathcal{D}$. Hence,

$$\begin{aligned}&\hat{g}_{\theta_0}(v)-\hat{g}_{\hat{\theta}}(v)\\&=\left[g(v)-g_{\hat{\theta}}(v)\right]+\frac{1}{2}\left[g''(v)-g''_{\hat{\theta}}(v)\right]h^2\\&\quad+n^{-1}\left[f_{\theta_0}^{-1}(v)-f_{\hat{\theta}}^{-1}(v)\right]\sum_{i=1}^{n}K_h(X_i^T\theta_0-v)\varepsilon_i+g'_{\hat{\theta}}(v)(\hat{\theta}-\theta_0)^T\mu_{\hat{\theta}}(v)\\&\quad+\frac{1}{nf_{\hat{\theta}}(v)}\sum_{i=1}^{n}\left[K_h(X_i^T\theta_0-v)-K_h(X_i^T\hat{\theta}-v)\right]\varepsilon_i+O_P(h^3(\log n)^{1/2}).\end{aligned}$$

By (C2) and (C6), we have $n^{-1}[f_{\theta_0}^{-1}(v)-f_{\hat\theta}^{-1}(v)]\sum_{i=1}^n K_h(X_i^T\theta_0-v)\varepsilon_i = O_P(n^{-1+\tau}$ $h^{-1/2}(\log n)^{1/2})$. Following the proof of Lemma A.2 of Xia and Li (1999), we have $\sum_{i=1}^n[K_h(X_i^T\theta_0 - v) - K_h(X_i^T\theta - v)]\varepsilon_i = O_P((nhn^{-1/2+\tau}\log n)^{1/2})$ uniformly for $\theta \in \mathcal{B}$ and $v \in \mathcal{D}$. By (4.9) and (4.10) of Härdle, Hall and Ichimura (1993), we have $g(X_i^T\theta_0) - g_{\hat\theta}(X_i^T\hat\theta) = (\theta_0 - \hat\theta)^T[X_i - E(X_i|X_i^T\theta_0)]g'(X_i^T\theta_0) + O_P(n^{-1+2\tau})$ and

$$g''(X_i^T\theta_0) - g''_{\hat\theta}(X_i^T\hat\theta) = O_P(n^{-1/2+\tau}) \tag{A.30}$$

uniformly for $X_i^T\hat\theta \in \mathcal{D}$. Combining the last five equations above, we have

$$\hat g_{\theta_0}(X_i^T\theta_0) - \hat g_{\hat\theta}(X_i^T\hat\theta) = g'(X_i^T\theta_0)\left[X_i - E(X_i|X_i^T\theta_0)\right]^T(\theta_0 - \hat\theta) + O_P(n^{-1+2\tau}) \tag{A.31}$$

uniformly for $X_i^T\hat\theta \in \mathcal{D}$. By Lemma A.3, (A.31) and assumption (C6), we have

$$\begin{aligned} R_C - R_3 &= n^{-1/2}\sum_{i=1}^n \mathcal{I}_i(\hat\theta)\left\{\hat g_{\theta_0}(X_i^T\theta_0) - \hat g_{\hat\theta}(X_i^T\hat\theta)\right\} I(X_i < x) \\ &= n^{-1/2}(\theta_0 - \hat\theta)^T\sum_{i=1}^n \left\{\mathcal{I}_i(\hat\theta)g'(X_i^T\theta_0)\left[X_i - E(X_i|X_i^T\theta_0)\right] I(X_i < x)\right\} \\ &\quad + O_P(n^{-\frac{1}{2}+2\tau'}) \\ &= o_P(1) \end{aligned} \tag{A.32}$$

uniformly for $x \in \bar{\mathbb{R}}^{\otimes p}$. The first equation in Lemma A.4 follows from (A.28), (A.29) and (A.32). By (A.15), we have

$$\begin{aligned} & n^{-1/2}\Big[\sum_{X_i^T\hat\theta\in\mathcal{D}} - \sum_{\mathcal{D}}\Big]B_i I(X_i < x) \\ &= n^{-1/2}h^2\sum_{i=1}^n \mathcal{I}_i(\hat\theta)g''_{\hat\theta}(X_i^T\hat\theta)I(X_i < x) \\ &\quad + n^{-3/2}\sum_{i=1}^n \mathcal{I}_i(\hat\theta)f_{\hat\theta}^{-1}(X_i^T\hat\theta)I(X_i < x)\sum_{j=1}^n K''_h(X_j^T\hat\theta - X_i^T\hat\theta)\varepsilon_j \\ &\quad + O_P(n^{1/2}h^3(\log n)^{1/2}). \end{aligned} \tag{A.33}$$

The first term on the right hand side above is $o_P(1)$ by Lemma A.3. The second term on the right hand side above is also $o_P(1)$ by (A.18). Therefore the second equation of Lemma A.4 follows immediately from the first equation and (A.33).

Proof of Theorem 1. By Lemma A.4 and (A.27), we have

$$S_{\mathcal{D}}(x) = R_1 + R_2 + R_3 + o_P(1) \tag{A.34}$$

uniformly for $x \in \bar{\mathbb{R}}^{\otimes p}$. By Lemma A.1, we have

$$\begin{aligned} R_2 = &-\frac{1}{2}h^2 n^{-1/2}\sum_{\mathcal{D}} g''(X_i^T\theta_0)I(X_i < x) - n^{-1/2}\sum_{\mathcal{D}}\Big[\{nf_{\theta_0}(X_i^T\theta_0)\}^{-1} \\ &\times \sum_{j=1}^{n} K_h(X_j^T\theta_0 - X_i^T\theta_0)\varepsilon_j I(X_i < x)\Big] + o_P(1) \\ \triangleq\ & R_{21} + R_{22} + o_P(1). \end{aligned} \tag{A.35}$$

By the Law of Large Numbers, we have

$$R_{21} = -\frac{1}{2}h^2 n^{1/2} E\left\{g''(X_i^T\theta_0)I_{\mathcal{D}}(X_i < x)\right\} + o_P(1). \tag{A.36}$$

By Lemma A.2 of Xia and Li (1999), we have

$$\begin{aligned} R_{22} &= -n^{-1/2}\sum_{j=1}^{n}\left\{n^{-1}\sum_{\mathcal{D}} K_h(X_j^T\theta_0 - X_i^T\theta_0)f_{\theta_0}^{-1}(X_i^T\theta_0)I(X_i < x)\right\}\varepsilon_j \\ &= -n^{-1/2}\sum_{j=1}^{n} E\left\{I_{\mathcal{D}}(X_j < x)|X_j^T\theta_0\right\}\varepsilon_j + o_P(1). \end{aligned} \tag{A.37}$$

It follows from (A.35), (A.36) and (A.37) that

$$R_2 = -B_{\mathcal{D}}(x) - n^{-1/2}\sum_{j=1}^{n} E\left\{I_{\mathcal{D}}(X_j < x)|X_j^T\theta_0\right\}\varepsilon_j + o_P(1). \tag{A.38}$$

By (A.31) and condition (C6) and its remarks, we have

$$\begin{aligned} R_3 &= n^{-1/2}(\theta_0 - \hat{\theta})^T\sum_{\mathcal{D}} g'(X_i^T\theta_0)\{X_i - E(X_i|X_i^T\theta_0)\}I(X_i < x) + o_P(1) \\ &= -n^{-1}\sum_{\mathcal{D}} g'(X_i^T\theta_0)\{X_i - E(X_i|X_i^T\theta_0)\}I(X_i < x)n^{-1/2}\sum_{j=1}^{n}\ell_n(X_j, g, \theta_0)\varepsilon_j \\ &\quad + o_P(1) \\ &= -E[g'(X^T\theta_0)\{X - E(X|X^T\theta_0)\}I_{\mathcal{D}}(X < x)]n^{-1/2}\sum_{i=1}^{n}\ell(X_i, g, \theta_0)\varepsilon_i + o_P(1). \end{aligned}$$

From (A.38), (A.39), the definition of R_1 and (A.34), it follows that

$$\begin{aligned} &S_{\mathcal{D}}(x) + B_{\mathcal{D}}(x) \\ &= n^{-1/2}\sum_{i=1}^{n}\Big[I_{\mathcal{D}}(X_i < x) - E\{g'(X^T\theta_0)(X - E(X|X^T\theta_0))I_{\mathcal{D}}(X < x)\}^T\ell(X_i, g, \theta_0) \\ &\qquad - E\{I_{\mathcal{D}}(X_i < x)|X_i^T\theta_0\}\Big]\varepsilon_i + o_P(1). \end{aligned}$$

It is easy to see that any finite dimensional distribution of $S_{\mathcal{D}}(x) + B_{\mathcal{D}}(x)$ tends to that of $Q(x)$. The proof of tightness in $D(\bar{\mathbb{R}}^{\otimes p})$ of the summation on the right hand side above is very standard. Now, we give the details for the first part when the marginal distribution of X is uniform on $[0,1]$ as discussed in the proof Lemma A.3. We use the definition in Bickel and Wichura (1971). Suppose that B and C are two disjoint blocks and they are pairs of neighbors in $[0,1]^{\otimes p}$. Let $L(A) = \int_{x\in A} \sigma^2(x)f(x)dx$ for any Borel set $A \in [0,1]^{\otimes p}$. We have

$$\begin{aligned} &E\Big[\{n^{-1/2}\sum_{i=1}^{n} I_{\mathcal{D}}(X_i \in B)\varepsilon_i\}^2\{n^{-1/2}\sum_{i=1}^{n} I_{\mathcal{D}}(X_i \in C)\varepsilon_i\}^2\Big] \\ &\le L(B)L(C) \le \{L(B\cup C)\}^2. \end{aligned}$$

The tightness follows from the remark below Theorem 3 of Bickel and Wichura (1971). Therefore we have finished the proof of Theorem 1.

Proof of Theorem 2. By (2.3), (C5) and Lemma A.1, $\hat{g}_{\hat{\theta}}(v)$ has a bounded derivative in probability. The conditions in (C6) hold for the bootstrap model (3.1). Thus $\hat{\theta}^* - \hat{\theta} = (W_n^*)^- \sum_{i=1}^n w(X_i)\{X_i - C(X_i,\hat{\theta})\}\hat{g}'_{\hat{\theta}}(\hat{\theta}^T X_i)\varepsilon_i^* + o_P(n^{-1/2})$, where $W_n^* = n^{-1}\sum_{i=1}^n w(X_i)\{X_i - C(X_i,\hat{\theta})\}\{X_i - C(X_i,\hat{\theta})\}^T \hat{g}'_{\hat{\theta}}(\hat{\theta}^T X_i)^2$. By Lemma 8 of Weisberg and Welsh (1994), we have $n^{-1}\sum_{i=1}^n w(X_i)[\hat{g}'_{\hat{\theta}}(\hat{\theta}^T X_i) - g'(\theta_0^T X_i)]^2 = o_P(1)$. Using Lemma 11 of Weisberg and Welsh (1994), we have $n^{-1}\sum_{i=1}^n w(X_i)[C(X_i,\hat{\theta}) - C(X_i,\theta_0)]^2 = o_P(1)$. Thus $n^{-1}\sum_{i=1}^n w(X_i)[\{X_i - C(X_i,\hat{\theta})\}\hat{g}'_{\hat{\theta}}(\hat{\theta}^T X_i) - \{X_i - C(X_i,\theta_0)\}g'(\theta_0^T X_i)]^2 = o_P(1)$. It follows that

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n} w(X_i)\Big[\{X_i - C(X_i,\hat{\theta})\}\hat{g}'_{\hat{\theta}}(\hat{\theta}^T X_i) - \{X_i - C(X_i,\theta_0)\}g'(\theta_0^T X_i)\Big]\varepsilon_i^* \\ &= o_P(n^{-1/2}). \end{aligned}$$

Similarly, we have $W_n^* - W_n = o_P(1)$. Therefore

$$\hat{\theta}^* - \hat{\theta} = (W_n)^- \sum_{i=1}^{n} w(X_i)\{X_i - C(X_i,\theta_0)\}g'(\theta_0^T X_i)\varepsilon_i^* + o_P(n^{-1/2}). \tag{A.39}$$

By Lemma A.2 and Theorem 1, the first part of Theorem 2 follows. According to (A.34), by the second part of Lemma A.4, we write $S_{\mathcal{D}}^*(x) = R_1^* + R_2^* + R_3^* + o_P(1)$, where $R_1^* = n^{-1/2}\sum_{\mathcal{D}} \varepsilon_i^* I(X_i < x)$, $R_2^* = n^{-1/2}\sum_{\mathcal{D}}\{\hat{g}_{\hat{\theta}}(X_i^T\hat{\theta}) - \hat{g}^*_{\hat{\theta}}(X_i^T\hat{\theta})\}I(X_i < x)$, $R_3^* = n^{-1/2}\sum_{\mathcal{D}}\{\hat{g}^*_{\hat{\theta}}(X_i^T\hat{\theta}) - \hat{g}^*_{\hat{\theta}^*}(X_i^T\hat{\theta}^*)\}I(X_i < x)$. By (A.14), we have

$$R_2^* = -\frac{1}{2}h^2 n^{-1/2}\sum_{\mathcal{D}} \hat{g}''_{\hat{\theta}}(X_i^T\hat{\theta})I(X_i < x)$$

$$-n^{-1/2}\sum_{\mathcal{D}}\{nf_{\hat\theta}(X_i^T\hat\theta)\}^{-1}\sum_{j=1}^{n}K_h(X_j^T\hat\theta-X_i^T\hat\theta)\varepsilon_j^*I(X_i<x)$$
$$+O_P(n^{1/2}h^3(\log n)^{1/2}).$$

By Lemma A.2 and (A.30), we have $n^{-1/2}\sum_{\mathcal{D}}\hat g''_{\hat\theta}(X_i^T\hat\theta)I(X_i<x)=n^{-1/2}\sum_{\mathcal{D}}g''_{\hat\theta}(X_i^T\hat\theta)I(X_i<x)+o_P(1)=n^{1/2}E[g''(X^T\theta_0)I_{\mathcal{D}}(X<x)]+O_P(n^{\tau})$. Therefore

$$R_2^*=-B_{\mathcal{D}}(x)-n^{-1/2}\sum_{\mathcal{D}}\{nhf_{\hat\theta}(X_i^T\hat\theta)\}^{-1}\sum_{j=1}^{n}K_h(X_j^T\hat\theta-X_i^T\hat\theta)\varepsilon_j^*I(X_i<x)+o_P(1). \tag{A.40}$$

Similar to R_3, we have

$$R_3^*=n^{-1/2}(\hat\theta-\hat\theta^*)^T\sum_{\mathcal{D}}\hat g'_{\hat\theta}(X_i^T\hat\theta)\{X_i-E(X_i|X_i^T\theta_0)\}I(X_i<x)+o_P(1)$$
$$=-n^{-1}\sum_{\mathcal{D}}\hat g'_{\hat\theta}(X_i^T\hat\theta)\{X_i-E(X_i|X_i^T\theta_0)\}I(X_i<x)n^{-1/2}\sum_{i=1}^{n}\ell_n(X_i,\hat g,\hat\theta)\varepsilon_i^*+o_P(1).$$

Note that by (A.39) and Lemma A.1, we have

$$R_3^*=E[g'(X^T\theta_0)\{X-E(X|X^T\theta_0)\}I_{\mathcal{D}}(X<x)]n^{-1/2}\sum_{i=1}^{n}\ell_n(X_i,g,\theta_0)\varepsilon_i^*+o_P(1)$$
$$=E[g'(X^T\theta_0)\{X-E(X|X^T\theta_0)\}I_{\mathcal{D}}(X<x)]n^{-1/2}\sum_{i=1}^{n}\ell(X_i,g,\theta_0)\varepsilon_i^*+o_P(1).$$

From the expressions of R_1^*, R_2^* and R_3^*, we have

$$S_{\mathcal{D}}^*(x)+B_{\mathcal{D}}(x)$$
$$=n^{-1/2}\sum_{i=1}^{n}\Big[I_{\mathcal{D}}(X_i<x)-E\{g'(X^T\theta_0)(X-E(X|X^T\theta_0))I_{\mathcal{D}}(X<x)\}^T\ell(X_i,g,\theta_0)$$
$$-E\{I_{\mathcal{D}}(X_i<x)|X_i^T\theta_0\}\Big]\varepsilon_i^*+o_P(1).$$

It is easy to see that any finite dimensional distribution of $S_{\mathcal{D}}^*(x)+B_{\mathcal{D}}(x)$ tends to that of $Q(x)$. For the tightness, we prove the case as in the proof of Theorem 1. Define $L_n(x)=n^{-1}\sum_{i=1}^{n}I_{\mathcal{D}}(X_i<x)\hat\varepsilon_i^2$ and write $L_n(B)$ as the Borel measure generated from $L_n(x)$. Let B and C be defined as in the proof of Theorem 1. We have

$$E\Big[\{n^{-1}\sum_{i=1}^{n}I_{\mathcal{D}}(X_i\in B)\varepsilon_i^*\}^2\{n^{-1}\sum_{i=1}^{n}I_{\mathcal{D}}(X_i\in C)\varepsilon_i^*\}^2|(X_j,Y_j),j=1,\ldots,n\Big]$$
$$\le L_n(B)^2L_n(C)^2\le\{L_n(B\cup C)\}^2.$$

It is easy to check that $L_n(x) \to L(x)$ uniformly in x using the same method as in Lemma A.2. We proved the tightness by the remarks below Theorem 3 of Bickel and Wichura (1971). Therefore $S^*_{\mathcal{D}}(x) + B_{\mathcal{D}}(x) \Rightarrow Q(x)$ in $D(\bar{\mathbb{R}}^p)$. In combination with Lemma A.2, the second part of Theorem 2 follows.

Acknowledgements

We thank the BBSRC/EPSRC of UK, the Research Grants Council of Hong Kong (HKU7149/98H), the CRCG of the University of Hong Kong, the Friends of London School of Economics (Hong Kong) and NUS research grant R-155-000-032-112 for partial support. We are most grateful to one referee and the Editor for thorough comments and constructive suggestions.

References

Afifi, A. A. and Virginia, C. (1984). *Computer-Aided Multivariate Analysis*. Lifetime Learning Publications. Belmont, California.

Bickel, P. L. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42**, 1656-1670.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons, New York.

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.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neymans Truncation. *J. Amer. Stat. Assoc.* **91**, 674-688.

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

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

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

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

Härdle, W. and Mammen, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21**, 1926-1947.

Härdle, W., Spokoiny, V. and Sperlich, S. (1997). Semiparametric single index versus fixed function modelling. *Ann. Statist.* **25**, 212-243

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

Hastie, J. T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13**, 435- 525.

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

Ledwina, T. (1994). Data-driven version of Neymans smooth test of fit. *J. Amer. Stat. Assoc.* **89**, 1000-1005.

Linton, O. and Nielsen, J. P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statist. Probab. Lett.* **19**, 181-187.

Masry, E.(1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

Neumann, M. H. and Kreiss, J. P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.* **4**, 1570-1613.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**, 613-641.

Stute, W., Manteiga, G. and Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Stat. Assoc.* **93**, 141-149.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *J. Amer. Stat. Assoc.* **86**, 420-426.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the Missing link. *Ann. Statist.* **22**, 1674-1700.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261-1343.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Statist. Soc. B* **60**, 797-811.

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

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

E-mail: staxyc@nus.edu.sg

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

E-mail: hrntlwk@hkucc.hku.hk

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

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

Department of Finance, Nanjing University, China.

E-mail: dxzhang@gdcc.edu.cn

(Received May 2002; accepted July 2003)

COMMENTS

Lexin Li and Christopher J. Nachtsheim

University of Minnesota

In this paper, the authors develop a goodness-of-fit test for the single-index model

$$y = g(\theta^T X) + \varepsilon, \tag{1}$$

where y is a univariate response, X is a p-dimensional vector of predictors, $g(\cdot)$ is the unknown link function, θ is the unknown $p \times 1$ vector with $||\theta|| = 1$, and ε is the error term assumed to satisfy $\mathrm{E}(\varepsilon \mid X) = 0$. They propose a Cramér-von Mises test statistic, CVS_D, based on nonparametric kernel estimation of θ. Bootstrapping is then employed both to mimic the null distribution of the test statistic and to provide a bias correction for the kernel estimate. We applaud the authors for their interesting work in this important area.

Our purpose in this discussion is to explore the robustness of CVS_D and to compare its performance with a class of sufficient-dimension-reduction-based tests previously developed by Li (1991) and Cook and Weisberg (1991), with enhancements by Cook (1998) and others. We begin with a brief introduction to these techniques.

1. Dimension Reduction Alternatives

The alternative tests that we now describe are based on the theory of *sufficient dimension reduction*, where the objective is to identify a subspace $\mathcal{S}$ of the predictor space such that

$$y \perp\!\!\!\perp X \mid P_{\mathcal{S}} X, \tag{2}$$

where $P_{(\cdot)}$ stands for a projection operator with respect to the standard inner product. A subspace satisfying (2) is called a dimension reduction subspace. When the intersection of all subspaces satisfying (2) also satisfies (2), it is called the *central subspace* (CS) and denoted by $\mathcal{S}_{y|X}$. Under some minor conditions (Cook (1998)), $\mathcal{S}_{y|X}$ exists and its dimension $d = \dim(\mathcal{S}_{y|X})$ is called the *structural dimension* of the regression. For the single-index model in (1) it is easy to see that $\mathcal{S}_{y|X} = \mathrm{Span}(\theta)$, with $d = 1$ if we assume no predictor effects in ε.

Cook and Li (2002) also considered situations where only the conditional mean $\mathrm{E}(y \mid X)$ is of interest. Analogously, the mean dimension reduction subspace is defined as subspace $\mathcal{S}$ that satisfies

$$y \perp\!\!\!\perp \mathrm{E}(y \mid X) \mid P_{\mathcal{S}} X. \tag{3}$$

The intersection of all the mean dimension reduction subspaces is called the *central mean subspace* (CMS), denoted as $\mathcal{S}_{E(y|X)}$, if the intersection satisfies (3) also. For single-index model (1), if we impose the additional restriction that $\varepsilon \perp\!\!\!\perp X$, we then have $\mathcal{S}_{E(y|X)} = \mathrm{Span}(\theta)$.

There are a variety of approaches for estimating $\mathcal{S}_{y|X}$, for instance sliced inverse regression (SIR, Li (1991)), and sliced average variance estimation (SAVE, Cook and Weisberg (1991)). Methods for estimating $\mathcal{S}_{E(y|X)}$ include ordinary least squares (OLS, Li and Duan (1989)), principal Hessian directions (PHD, Li (1992)), and iterative Hessian transformation (IHT, Cook and Li (2002)). There

are also inferential methods for testing hypotheses about the structural dimension d. The testing procedure is generally sequential. That is, we first test hypotheses $d = 0$ versus $d > 0$. If the null hypothesis is rejected, we then test $d = 1$ versus $d > 1$, and so on. The estimate of d is the first integer m such that we fail to reject the null hypothesis $d = m$. Obviously, this test procedure includes the test of single-index model as a special case.

The dimension reduction approaches just described generally make two key assumptions: (a) the linearity condition, and (b) the coverage condition. Consider SIR as an example. The linearity condition states that the marginal distribution of predictors satisfies

$$\mathrm{E}\,(X \mid P_{\mathcal{S}_{y|X}} X) = P_{\mathcal{S}_{y|X}} X. \tag{4}$$

This condition implies that the inverse mean subspace $\mathcal{S}_{E(X|y)} = \mathrm{Span}(\Sigma_x^{-1}$ $\mathrm{E}\,((X - \mathrm{E}\,(X)) \mid y))$ is a subspace of the central subspace $\mathcal{S}_{y|X}$, where Σ_x is the covariance matrix of X. Therefore $\mathcal{S}_{E(X|y)}$ provides partial estimation of $\mathcal{S}_{y|X}$. The coverage condition then requires

$$\mathcal{S}_{E(X|y)} = \mathcal{S}_{y|X}. \tag{5}$$

That is, it assumes the inverse mean subspace coincides with the central subspace. Note that the coverage condition fails for models of the form $y = (\theta^T X)^2 + \varepsilon$, where X follows a standard normal distribution and ε is an independent error.

A comprehensive account of the sufficient dimension reduction theory is given by Cook (1998).

2. SAVE-based Test

Xia et al. carry out extensive simulations to test their method using the following model.

Example 1. Assume $x_1, x_2 \sim \mathrm{Normal}(0, 1)$, $\varepsilon \sim \mathrm{Normal}(0, 1)$ and $y = x_1 + x_2 + 4e^{-(x_1+x_2)^2} + a(x_1^2 + x_2^2)^{1/2} + \sigma\varepsilon$.

Note that the coverage condition is not met for this model. SIR is therefore not applicable, and we employ the SAVE-based permutation test (Cook and Weisberg (1991), Cook and Yin (2001)) as an alternative to CVS_D. When $a = 0$, the structural dimension $d = 1$, and when $a \neq 0$, $d = 2$. Consider the hypotheses $d = 1$ versus $d > 1$. Table 1 gives rejection rates (per $1,000$ Monte Carlo replications) for the SAVE-based permutation test with 49 permutations. The numbers in parentheses are the corresponding values reported by Xia et al. in their Table 1.

Table 1. Comparison of Rejection Rates (per 1,000 Monte Carlo replications) for SAVE-based Test and CVS_D (in parentheses) for Subset of Cases from Example 1 of Xia et al. ($\alpha = 0.05$ and optimal bandwidth used for CVS_D)

	$\sigma = 0.3$			$\sigma = 0.5$		
n	$a = 0$	$a = 0.25$	$a = 0.5$	$a = 0$	$a = 0.25$	$a = 0.5$
50	0.077 (.063)	0.071 (.099)	0.072 (.376)	0.055 (.043)	0.055 (.043)	0.062 (.163)
100	0.067 (.045)	0.136 (.208)	0.272 (.806)	0.055 (.057)	0.120 (.082)	0.225 (.445)
300	0.055 (.043)	0.298 (.856)	0.755 (1.000)	0.062 (.048)	0.207 (.318)	0.615 (.977)

Clearly, when $a = 0$ and the model is indeed a single-index model, two classes of tests have comparable Type I error rates. However, when $a \neq 0$ and the model has dimension greater than 1, the test proposed by Xia et al. has considerably more power. For example, when $n = 300, \sigma = 0.3, a = 0.5$, Xia et al. report the power of their test is 1.0. In this case the power for the SAVE-based permutation test is just 0.755, not surprising to us since CVS_D has the home field advantage in this example.

3. SIR-based Test

We next employ a series of six new examples (numbered 2 through 7) in which the coverage condition (5) is met. This permits the use of SIR-based test rather than the SAVE-based test in the previous section. Details concerning the SIR-based test statistic and its asymptotic distribution can be found in Li (1991). Throughout, we use the authors' Matlab computer code as published on the web to carry out our simulations of the CVS_D test procedure. Bandwidth $h = 0.15$ and sample size $n = 100$ are employed in all test cases. The results are shown in Table 2.

3.1. Effect of noise

In Examples 2 and 3, the response model takes the following form.

Examples 2 and 3 Assume $x_1, \ldots, x_4 \sim \text{Normal}(0, 1)$, $\varepsilon \sim \text{Normal}(0, 1)$ and $y = e^{0.3\,(x_1 - x_2) + 1} + \sigma\varepsilon$.

In experimenting with CVS_D, it became apparent that its performance can be affected by the magnitude of σ. Evidence of this is provided in Table 2. In Example 2, with $\sigma = 1.5$ both the SIR-based test and CVS_D reject the null hypothesis that $d = 1$ at the nominal levels. However, in Example 3 with $\sigma = 3$, CVS_D is rejecting the null too frequently, while the SIR-based test continues

to reject at the nominal levels. This may suggest that CVS_D will fail if θ is ineffectively estimated.

3.2. Predictor effects in variance

Consider the error term ε in single-index model (1). The authors assume that $\mathrm{E}(\varepsilon \,|\, X) = 0$ almost surely. This allows for predictor effects not only in the conditional mean $\mathrm{E}(y \,|\, X)$ but also in the conditional variance $\mathrm{Var}(y \,|\, X)$. This can be easily seen by considering a univariate response y and p-dimensional predictors X with $\mathrm{E}(y \,|\, X)$ finite. Then

$$y = \mathrm{E}(y \,|\, X) + (y - \mathrm{E}(y \,|\, X)) = g(\theta^T X) + \varepsilon^*, \tag{6}$$

where we assume $\mathrm{E}(y \,|\, X) = g(\theta^T X)$ for some $g(\cdot)$ and θ, and $\varepsilon^* = y - \mathrm{E}(y \,|\, X)$. Obviously $\mathrm{E}(\varepsilon^* \,|\, X) = 0$ almost surely. We next examine two examples which are single-index models, while both variance terms depend on $\theta^T X$.

Example 4. Assume $x_1, x_2 \sim \mathrm{Normal}(0,1)$, $\varepsilon \sim \mathrm{Normal}(0,1)$ and $y = x_1 + x_2 + e^{(x_1+x_2)/2} \times \varepsilon$.

Example 5. Assume $x_1, \ldots, x_4 \sim \mathrm{Normal}(0,1)$, $g(X) = e^{(x_1+x_2+x_3)/2} - 1.5$, $p(X) = (1 + e^{-g(X)})^{-1}$, and $y \sim \mathrm{Binomial}(2, p(X))$.

Table 2 displays the rejection rates (per 200 Monte Carlo replications) of two tests for the null hypothesis $d = 1$. In both examples, CVS_D rejects the true single-index model too often, while the SIR-based test performs as expected.

Table 2. Comparison of Rejection Rates (per 200 Monte Carlo replications for CVS_D (results in parentheses) and SIR-based test for Examples 2 through 7.

Discussion Example	Linearity Condition Met?	Predictor Effects in Variance?	True d	H_0: $d = 1$ H_a: $d > 1$ $\alpha = 0.05$	 $\alpha = 0.1$
2	Yes	No (small σ)	1	.050 (.060)	.110(.130)
3	Yes	No (large σ)	1	.060 (.270)	.090(.360)
4	Yes	Yes	1	.060 (.135)	.115(.190)
5	Yes	Yes	1	.050 (.195)	.080(.270)
6	Yes	Yes	2	.720 (.280)	.810(.385)
7	No	No	1	.215 (.155)	.340(.215)

We further consider predictor effects in the variance in Example 6, this time with structural dimension $d = 2$. This example is used to assess the power of the CVS_D test relative to that of the SIR-based test.

Example 6. Assume $x_1, \ldots, x_4 \sim \text{Normal}(0,1)$, $\varepsilon \sim \text{Normal}(0,1)$ and $y = e^{0.5\,(x_1+x_2)+1.5} + 0.25(x_3+x_4) + e^{0.85\,(x_3+x_4)+1} \times \varepsilon$.

Note two linear combinations of predictors, $\theta_1^T X = x_1 + x_2$ and $\theta_2^T X = x_3 + x_4$, are required to characterize the conditional distribution of $y \,|\, X$. Both are present in the conditional mean $\mathrm{E}\,(y \,|\, X)$ but $\theta_1^T X$ has a stronger effect than $\theta_2^T X$. Meanwhile the conditional variance $\mathrm{Var}\,(y \,|\, X)$ depends on $\theta_2^T X$. The results in Table 2 indicate that the power of the CVS_D test (e.g., 0.280 for $\alpha = 0.05$) is less than half of that for the corresponding SIR-based test (0.720 for $\alpha = 0.05$).

We note that another advantage of the sufficient-dimension-reduction-based testing approach, in Example 6, is that a further test of hypotheses $d = 2$ versus $d > 2$ can be conducted when the null $d = 1$ is rejected in the first stage. Our simulation shows that the SIR-based test produces rejection rates 0.040 and 0.075 for nominal levels 0.05 and 0.1 respectively. Thus it is clear that SIR-based testing approach not only correctly tests if a model is single-index, it can simultaneously test for multi-index model also.

Examples 4, 5 and 6 have shown that heteroscedasticity adversely affects the performance of the test method proposed by Xia et al. Our conjecture is that the test procedure works best for model (1) with the restriction that ε is independent of X. Thus the predictor effects are in the conditional mean $\mathrm{E}\,(y \,|\, X)$ only. This corresponds to the study of the central mean subspace $\mathcal{S}_{E(y|X)}$. If this conjecture is true, the applicability of the proposed test will be limited considerably. On the other hand, the approaches associated with the estimation and inference of the central subspace $\mathcal{S}_{y|X}$ such as SIR and SAVE place no restriction on the conditional distribution of $y \,|\, X$.

3.3. Nonlinearity among predictors

With model (1) and the additional assumption $\varepsilon \perp\!\!\!\perp X$, we now consider a situation in which the linearity condition is not met. We therefore expect that the SIR-based test will perform poorly relative to CVS_D. We consider the following example with structural dimension $d = 1$.

Example 7. Assume $x_1 \sim \text{Uniform}(0,1)$, $e \sim \text{Uniform}(-0.3, 0.3)$, $x_2 = \log(x_1) + e$, $x_3, x_4, x_5 \sim \text{Normal}(0,1)$, $\varepsilon \sim \text{Normal}(0,1)$ and $y = e^{0.5\,(x_1-x_2-x_3)+1} + \varepsilon$.

The results in Table 2 confirm our expectation that the SIR-based test will not perform well here. For example, the rejection rates for the null $d = 1$ are 0.215 when $\alpha = 0.05$ and 0.340 when $\alpha = 0.10$. We were surprised to see that the CVS_D also rejects the null too often in this example. The rejection rates for $\alpha = 0.05$ and $\alpha = 0.10$ (0.155 and 0.215) are closer than those for the SIR-based test, but are still significantly different from nominal. Extensions of the sufficient

dimension reduction approaches when there is nonlinearity among predictors can be found in Cook and Nachtsheim (1994) and Li (2003).

4. Conclusions

In this discussion, we have examined the performance of the goodness-of-fit test proposed by Xia et al. for single-index models. Our simulation results suggest that the performance of the proposed test is adversely affected when the error variance is large, when there are predictor effects in the variance, and/or when predictors are nonlinearly related. Alternative testing approaches, based on the theory of sufficient dimension reduction, avoid many of these difficulties. Moreover they are capable of identifying the structural dimension of the regression for both single and multi-index models.

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN55455, U.S.A.

E-mail: lexinl@stat.umn.edu

Operations and Mgmt Saevces, Carlson of Management, University of Minnesota, 3-245 Carl SMgmt, 321 19th Ave. S, Minneapolis, MN55455, U.S.A.

E-mail: CNachtsheim@csom.umn.edu

REJOINDER

Yingcun Xia, W. K. Li, Howell Tong and Dixin Zhang

We welcome the opportunity of an open discussion of our paper. The purpose of the original paper was quite modest as it was intended only to solve a simple testing problem using a new approach. The discussion of Professors L. Li and C. J. Nachtsheim (called LN, hereafter) has, however, widened our perspective and allowed us to delve deeper into the problem. We are therefore most grateful to the Editor and the above discussants.

From the point of view of kernel smoothing, our proposed method provides a data-driven semiparametric test. The method needs no under-smoothing or over-smoothing of the link functions. For finite samples, our method enjoys a higher testing power than all the methods we know of.

The concern of LN is that our test is too sensitive in that it may reject H_0 too frequently when the conditional variance, $\sigma(x)^2$, of the model is large. Their simulations seem to add substance to this concern.

Bearing in mind the fact that when using kernel smoothing, a reasonable bandwidth should be proportional to $\sigma(x)^{2/(p+4)}$ when the N-W estimator or the local linear smoother is used, where p is the dimension of a link function, we would not ourselves share their concern. (See, for example, Fan and Gijbles (1996).) LN's concern is a product of their choice of the bandwidth. The bandwidth used by LN for the CVS_D test, $h = 0.15$, is too small for all the examples. Note also that their examples are with monotone link functions. This choice of the link functions favours the SIR-type approach because the monotonicity tends to reduce the curvature of the functions and thus our test would require a larger bandwidth. Figure 1 shows the optimal bandwidth h_{opt} for the local linear smoothing of y on $\theta^T X$ in the sense of MISE. Apart from our own example (Example 1 in LN), the bandwidth $h = 0.15$ is obviously too small for Examples 2−7.

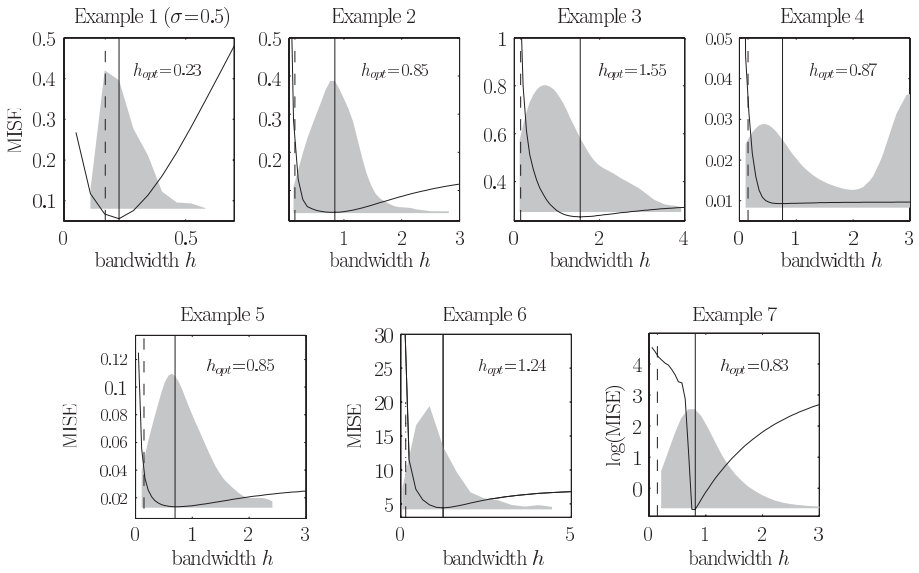

Figure 1. Visualisation of bandwidth selection for all the Examples based on 500 replications each has sample size n=100. The curves in the panels are the average of $ASE(h) = n^{-1}\sum_{i=1}^{n}\{\hat{m}(\theta_0^T x_i) - m(\theta_0^T x_i)\}^2$. The shaded area refers to the density function (rescaled) of the selected bandwidths by the CV method. The solid vertical lines mark the corresponding h_{opt} for each example. The dash vertical lines mark the bandwidth used by us (Example 1) and LN (Examples 2−7).

One way to obtain a reasonable bandwidth in practice is to apply the cross-validation bandwidth selection method to the regression of y on $\hat{\theta}^T X$, where $\hat{\theta}$ is the estimate of θ. The shades in Figure 1 show the distribution of the selected bandwidths using CV methods. Except for Example 4, for which an appropriate

bandwidth is infinite theoretically, the CV bandwidths for all the other examples are not far away from the respective h_{opt}. The other bandwidth that has to be selected is for the estimation of the single-index θ. In the code we put on the website, we used the average derivative estimation method (Härdle and Stoker (1989)). Again, if no information about the bandwidth is available then a data driven bandwidth is suggested. Under this circumstance we can resort to using the CV bandwidth selection method although it might not be the best approach. (The CV bandwidth selection method for both steps is now available at the same website.)

Armed with the above observations, we can now check the examples provided by LN. Table 1 shows that both CV bandwidths and fixed bandwidths have reasonable rejection rate for Examples 2−5. Moreover, for a wide range of bandwidths, e.g., $h_{opt}/2$ to h_{opt}, the rate changes little. This further indicates that our test method is robust to the choice of bandwidth provided that it is not too small or too large. We shall discuss Examples 6−7 below.

Table 1. Rejection rate for Examples 2−7 based on 500 replications with different bandwidths.

Example	CV bandwidth		Bandwidth $h = h_{opt}$		Bandwidth $h = h_{opt}/2$	
	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
2	0.024	0.090	0.024	0.064	0.032	0.080
3	0.030	0.088	0.028	0.072	0.028	0.080
4	0.042	0.092	0.040	0.088	0.050	0.108
5	0.034	0.072	0.024	0.082	0.036	0.096
6	0.108	0.180	0.074	0.156	0.122	0.194
7	0.160	0.228	0.160	0.226	0.160	0.232

For Example 6, we can write it as $y = m_1(\theta_1^T X) + m_2(\theta_2^T X) + \sigma(\theta_2^T X)\varepsilon$. The signal-noise ratio (S/N) for the first and the second dimensions are, respectively,

$$\Big(\frac{\mathrm{Var}\,\{m_1(\theta_1^T X)\}}{\mathrm{Var}\,\{\sigma(\theta_2^T X)\varepsilon\}}\Big)^{1/2} = 0.40 \quad \text{and} \quad \Big(\frac{\mathrm{Var}\,\{m_2(\theta_2^T X)\}}{\mathrm{Var}\,\{\sigma(\theta_2^T X)\varepsilon\}}\Big)^{1/2} = 0.03.$$

The second S/N is very small for sample size n = 100. See Figure 2. The second dimension is hard to detect through the conditional mean (signal). However, the second dimension has a strong effect on the conditional variance. It can be detected via the conditional variance specification (Xia, Tong and Li (2002)). Thus, it is not surprising that our method cannot detect the second dimension with high frequency because our method only concentrates on the conditional mean. It is also not surprising that the χ^2 test based on SIR estimates can detect the second dimension via the conditional variance. To confirm the fact

that the χ^2 test detects the second dimension through the conditional variance, we set $\sigma(\theta_2^T X) \equiv 1$. Our simulation results in Table 2 (in the block corresponding to the row Example 6* and the column $a = 0.25$) suggest that the χ^2 test also has difficulties in detecting the second dimension merely through the conditional mean.

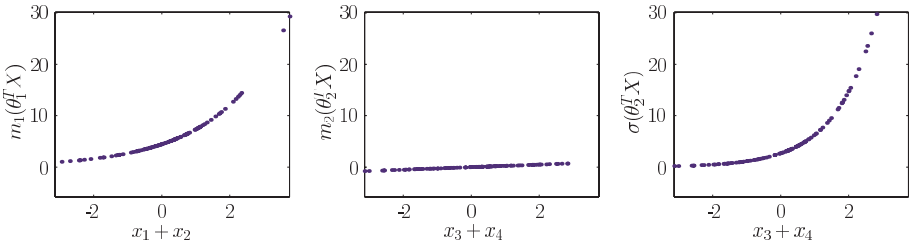

Figure 2. A typical sample with $n = 100$ from Example 6.

The nonlinear confounding effect in predictors is a real problem for the estimation of the efficient dimension reduction space. See, Li (1997). We believe that the problem becomes more difficult for testing problems. LN mentioned the work of Cook and Nachtsheim (1994) about nonlinear confounding in the predictors. However, as far as we know, their main interest is about the estimation of the directions, and little is discussed about testing. Although our method cannot solve the testing problem completely, it provides a way to improve the existing methods.

It is interesting to compare the power using the models provided by LN when H_0 is not true. We are not aware of reports on the power of the tests used by LN and our examples below are designed only to give us some initial ideas about the power of these tests. Note that in LN, only Example 6 gives rise to an alternative H_a. To see the effect of heteroscedasticity, we further extend Example 4 and consider the following models.

$$\text{Example } 4^* : \; y = x_1 + x_2 + ae^{(x_1 - x_2)/2} + e^{(x_1+x_2)/2}\varepsilon,$$
$$\text{Example } 6^* : \; y = e^{0.5(x_1+x_2)+1.5} + a(x_3 + x_4) + \varepsilon,$$

where x_1, x_2, x_3, x_4 and ε are defined in LN. The parameter a is employed to control the departure of the model from H_0. In all the simulations below, we use CV bandwidth for our method. For the SIR estimation, we take the number of slices to be 10. For the permutation test (Cook (1998)), the number of permutation is set to be 50. We also tried some other settings for the number of slices and the number of permutations as well as some other testing methods as implemented by the programs provided in Weisberg (2002). There are no big differences in the results. The programs by Weisberg (2002) are used for the χ^2

test and the permutation test. Table 2 shows that our method has quite reasonable rejection rates when H_0 is true even with heteroscedasticity, while the power is much higher than the χ^2 test and the permutation test. The low power of the permutation and χ^2 tests are somewhat puzzling and we would be happy to amend our results if we have misunderstood the procedures.

Table 2. Comparisons of rejection rate for Examples 4* and 6* based on 500 replications.

Md.	Metohd	$a=0$		$a=0.25$		$a=1$		$a=2$	
		α=0.05	α=0.10	α=0.05	α=0.10	α=0.05	α=0.10	α=0.05	α=0.10
4*	CVS_D	0.042	0.090	0.134	0.254	0.548	0.718	0.562	0.758
	Perm.	0.070	0.124	0.054	0.112	0.062	0.126	0.066	0.126
	χ^2	0.048	0.094	0.054	0.112	0.052	0.098	0.031	0.082
6*	CVS_D	0.042	0.100	0.056	0.126	0.424	0.590	0.800	0.920
	Perm.	0.054	0.106	0.042	0.100	0.182	0.296	0.226	0.360
	χ^2	0.040	0.090	0.060	0.116	0.136	0.236	0.214	0.336

Conclusions

The method developed in this paper is not intended to compete with the SIR method in the domain of dimension reduction, although they have something in common. However, in the common area, all the simulations suggest that our method is indeed more powerful than existing test methods. The concern of Professor L. Li and Professor C. J. Nachtsheim does not arise at all as long as the pilot parameter, h, is not too unreasonable and we have given some guidance to what constitutes a reasonable choice. This paper considers only single-index models. Dimension reduction with multi-indices is considered in Xia, Tong, Li and Zhu (2002).

References

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Amer. Statist. Assoc.* **89**, 592-600.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Austral. New Zealand J. Statist.* **43**, 147-200.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Ann. Statist.* **87**, 1025-1039.

Li, K. C. (1997). Nonlinear confounding in high dimensional regression. *Ann. Statist.* **17**, 1009-1052.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.

Li, L. (2003). Sufficient dimension reduction in high-dimensional data. Ph.D. Dissertation, School of Statistics, University of Minnesota.

Weisberg, S. (2002). Dimension reduction regression in R. *J. Statist. Soft.* **7** (website: http://www.jstatsoft.org/index.php?vol=7).

Xia, Y., Tong, H. and Li, W. K. (2002). Single index diffusion models and thier estimation. *Statist. Sinica* **12**, 785-799.

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