ESTIMATION FOR A PARTIAL-LINEAR SINGLE-INDEX MODEL

Jane-Ling Wang[1], Liugen Xue[2], Lixing Zhu[3], and Yun Sam Chong[4]

[1] *University of California at Davis*

[2] *Beijing University of Technology, Beijing, China*

[3] *Hong Kong Baptist University, Hong Kong, China*

[4] *Wecker Associate*

In this paper, we study the estimation for a partial-linear single-index model. A two-stage estimation procedure is proposed to estimate the link function for the single index and the parameters in the single index, as well as the parameters in the linear component of the model. Asymptotic normality is established for both parametric components. For the index, a constrained estimating equation leads to an asymptotically more efficient estimator than existing estimators in the sense that it is of a smaller limiting variance. The estimator of the nonparametric link function achieves optimal convergence rates; and the structural error variance is obtained. In addition, the results facilitate the construction of confidence regions and hypothesis testing for the unknown parameters. A simulation study is performed and an application to a real dataset is illustrated. The extension to multiple indices is briefly sketched.

[0] Lixing Zhu is the corresponding author. Email: lzhu@hkbu.edu.hk. Jane-Ling Wang's research was partially supported by NSF grant DMS-0406430. Liugen Xue's research was supported by the National Natural Science Foundation of China (10571008, 10871013), the Natural Science Foundation of Beijing (1072004) and Ph. D. Program Foundation of Ministry of Education of China (20070005003). Lixing Zhu's research was supported by a grant of The Research Grant Council of Hong Kong, Hong Kong, China (HKBU7060/04P and HKBU 2030/07P). The first three authors have equal contribution to this research. The authors thank the Editor, the Associate Editor, and the two referees for their insightful comments and suggestions which have led to substantial improvements in the presentation of the manuscript.

[0] *AMS 2000 subject classifications.* Primary 62G05; secondary 62G20.

[0] *Key words and phrases.* Dimension reduction, local linear smoothing, bandwidth, two-stage estimation, kernel smoother.

1 Introduction

Partial linear models have attracted lots of attention due to their flexibility to combine traditional linear models with nonparametric regression models. See, e.g. Heckman (1986), Rice (1986), Chen (1988), Bhattacharya and Zhao (1997), Xia and Härdle (2006), and the recent comprehensive books by Härdle, Gao, and Liang (2000) and Ruppert, Wand and Carroll (2003) for additional references. However, the nonparametric components are subject to the curse of dimensionality and can only accommodate low dimensional covariates X. To remedy this, a dimension reduction model which assumes that the influence of the covariate X can be collapsed to a single index, $X^{\mathrm{T}}\beta$, through a nonparametric link function g is a viable option and termed the partial-linear single-index model. Specifically, it takes the form:

$$Y = Z^{\mathrm{T}}\theta_0 + g(X^{\mathrm{T}}\beta_0) + e, \tag{1.1}$$

where $(X, Z) \in R^p \times R^q$ are covariates of the response variable Y, g is an unknown link function for the single index, and e is the error term with $E(e) = 0$ and $0 < \mathrm{Var}(e) = \sigma^2 < \infty$. For the sake of identifiability, it is often assumed that $\|\beta_0\| = 1$ and the rth component of β_0 is positive, where $\|\cdot\|$ denotes the Euclidean metric.

This model is quite general, it includes the aforementioned partial-linear model when the dimension of X is one and also the popular single-index model in the absence of the linear covariate Z. There is an extensive literature for the single-index model with three main approaches: projection pursuit regression (PPR) [Friedman and Stuetzle (1981), Hall (1989), Härdle, Hall and Ichimura (1993)]; the average derivative approach [Stoker (1986), Doksum and Samarov (1995), and Hristache, Juditsky and Spokoiny (2001)]; and sliced inverse regression (SIR) and related methods [Li (1991), Cook and Li (2002), Xia, Tong, Li and Zhu (2002), and Yin and Cook (2002)]. All these approaches rely on the assumption that the predictors in X are continuous variables, while model (1.1) compensates for this by allowing discrete or other continuous variables to be linearly associated with the response variable. To our knowledge, Carroll, Fan, Gijbels and Wand (1997) were the first to explore model (1.1) and they actually considered a generalized version, where a known link function is employed in the regression function while model (1.1) assumes an identity link function.

However, their approaches may become computationally unstable as observed by Yu and Ruppert (2002) and confirmed by our simulations in Section 3. The theory of Carroll, Fan, Gijbels and Wand (1997) also relies on the strong assumption that their estimator for θ_0 is already $\sqrt{n}$-consistent. Yu and Ruppert (2002) alleviated both difficulties by employing a link function g which falls in a finite-dimensional spline space, yielding essentially a flexible parametric model. Xia and Härdle (2006) used a method that is based on a local polynomial smoother and a modified version of least squares in Härdle, Hall and Ichimura (1993).

In this paper, we propose a new estimation procedure. Our approach requires no iteration and works well under the mild condition that a few indices based on X suffice to explain Z. Namely,

$$Z = \phi(X^{\mathrm{T}}\beta_Z) + \eta, \tag{1.2}$$

where $\phi(\cdot)$ is an unknown function from R^d to R^q, β_Z is a $p \times d$ matrix with orthonormal columns, η has mean zero and is independent of X. The dimension d is often much smaller than the dimension p of X. Such an assumption is not stringent and common in most dimension reduction approaches in the literature. A theoretical justification is provided in Li, Wen and Zhu (2008). Model (1.2) implies that a few indices of X suffice to summarize all the information carried in X to predict Z, which is often the case in reality, such as for the Boston Housing data in section 4, where a single index was selected for model (1.2) and Z is a discrete variable. In this data, first analyzed in Harrison and Rubinfeld (1978), the response variable is the median value of houses in 506 census tracts in the Boston area. The covariates include: average number of rooms, the proportion of houses built before 1940, eight variables describing the neighborhood, two variables describing the accessibility to highways and employment centers, and two variables describing air pollution. A key covariate of interest is a binary variable that specifies whether a house borders the river or not. Our analysis presented in Section 4 based on the dimension reduction assumptions of (1.1) and with Z equal to this binary variable in (1.2) demonstrates the advantages of our model assumption, only one index ($d = 1$) was needed in model (1.2) for this data.

To avoid the computational complications that we experienced with the procedure in Carroll et al. (1997), who aim at estimating β_0 and θ_0 simultaneously, we choose to estimate

β_0 and θ_0 sequentially. The idea is simple: θ_0 can be estimated optimally through approaches developed for partial linear models once we have a $\sqrt{n}$ estimate of β_0 and plug it in (1.1). However, β_0 and θ_0 may be correlated, leading to difficulties in identifying β_0. This is where model (1.2) comes in handy, as it allows us to remove the part of Z that is related to X so that the residual η in (1.2) is independent of X. Again, we need to impose the identifiability condition that β_Z has norm one and a positive first component. The procedure is as follows: First estimate β_Z via any dimension reduction approach, such as SIR or PPR for $q = 1$, and the projective resampling method in Li, Wen and Zhu (2008) for $q > 1$. Once β_Z has been estimated we proceed to estimate ϕ via a d-dimensional smoother and then obtain the residual for η. Since $\eta = Z - \phi(X^{\mathrm{T}}\beta_Z)$, plugging this into (1.1) we get

$$Y = \eta^{\mathrm{T}}\theta_0 + h(X^{\mathrm{T}}\beta_0, X^{\mathrm{T}}\beta_Z) + e,$$

where h is an unknown function, but now η and X are independent of each other. It is thus possible to employ a least squares approach to estimate θ_0 and the resulting estimate will be $\sqrt{n}$-consistent. We then employ a dimension reduction procedure to $Y - Z^{\mathrm{T}}\hat{\theta}_0$ and X to obtain an estimate for β_0 and g. This concludes the first stage , where the resulting estimates for θ_0 and β_0 are already $\sqrt{n}$ consistent but will serve the role as initial estimates for the next stage, where we update all the estimates but use a more sophisticated approach. Specifically for θ_0 we apply the profile method, also called partial regression in Speckman (1988), to estimate θ_0. Theoretical results in Section 2.2 indicate that the two-stage procedure is fully efficient, so there is no need for iteration. More importantly, to estimate the index β_0, we use an estimating equation to obtain asymptotic normality, which takes the constraint $\|\beta_0\| = 1$ into account. The estimator based on this new estimating equation performs better in several ways, summarized as follows.

1. Our estimation procedure directly targets the model parameters θ_0, β_0, β_Z, $\phi(\cdot)$ and $g(\cdot)$ and no iteration is needed.

2. We obtain the asymptotic normality of the estimator of β_0 and the optimal convergence rate of the estimator of $g(\cdot)$, as well as the asymptotic normality of the estimator of θ_0. The most attractive feature of this new method is that the estimator of β_0 has

smaller limiting variance when compared to three existing approaches in : Härdle et al. (1993) when the model is reduced to the single-index model, Carroll et al.(1997) if their link function is the identity function, and Xia and Härdle (2006) when their model is homoscedastic. This is the first result providing such a small limiting variance in this area.

3. We also provide the asymptotic normality of the estimator of σ^2. It allows us to consider the construction of confidence regions and hypothesis testing for θ_0 and β_0.

The rest of the paper is organized as follows. In Section 2, we elaborate on the new methodology and then present the asymptotic properties for the estimators. Section 3 reports the results of a simulation study and Section 4 an application to a real data example for illustration. Section 5 gives the proofs of the main theorems. Some lemmas and their proofs are relegated to the Appendix.

2 Methodology and Main Results

2.1 Estimating Procedures

The observations are $\{(X_i, Y_i, Z_i); 1 \le i \le n\}$, a sequence of independent and identically distributed (i.i.d.) samples from (1.1), i.e.

$$Y_i = Z_i^{\mathrm{T}}\theta_0 + g(X_i^{\mathrm{T}}\beta_0) + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \cdots, e_n$ are i.i.d. random errors with $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2 > 0$, $\{\varepsilon_i; 1 \le i \le n\}$ are independent of $\{(X_i, Z_i); 1 \le i \le n\}$, $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, $Z_i = (Z_{i1}, \ldots, Z_{iq})^{\mathrm{T}}$, $\beta_0 \in R^p$ and $\theta_0 \in R^q$. For simplicity of presentation, we initially assume that Z can be recovered from a single-index of X. That is, $d = 1$ in (1.2). The general case will be explored at the end of this section in Remarks 2. Below we first outline the steps for each stage and then elaborate on each of these steps.

Algorithm for Stage One:

1. Apply a dimension reduction method for the regression of Z_i versus X_i to find an estimator $\hat{\beta}_Z$ of β_Z;

2. Smooth the Z_i over $\hat{X}_i^{\mathrm{T}}\beta_Z$ to get an estimator $\hat{\phi}(\cdot)$ of $\phi(\cdot)$, then compute the residuals $\hat{\eta}_i = Z_i - \hat{\phi}(X_i^{\mathrm{T}}\hat{\beta}_Z)$;

3. Perform a linear regression of Y_i versus $\hat{\eta}_i$'s to find an initial estimator $\hat{\theta}_0$ of θ_0;

4. Apply a dimension reduction method to the regression of $Y_i - Z_i^{\mathrm{T}}\hat{\theta}_0$ versus X_i to find an initial estimator $\hat{\beta}_0$ of β_0

5. Smooth the $Y_i - Z_i^{\mathrm{T}}\hat{\theta}_0$ versus the $X_i^{\mathrm{T}}\hat{\beta}_0$ to obtain an estimator for g and for its derivative g'.

Algorithm for Stage Two:

6. Use the initial estimate $\hat{\beta}_0$ from Step 4 to update the estimate of θ_0 through a profile approach for the partial linear model by minimizing (2.5).

7. Use the updated estimate $\hat{\theta}$ of θ_0 from Step 6 to form the new residual $Y - Z^{\mathrm{T}}\hat{\theta}$, then update the estimate of β_0 by solving the estimating equation (2.10).

8. Use the updated estimates of θ_0 and β_0 in Steps 6 and 7 to update the estimate of g, following the procedure as described in Step 5.

This completes the algorithm and, as we show in Section 2.2, the resulting estimators are already theoretically efficient. However, the practical performance can be improved by iterating Steps 6 and 7 one or more times. Our experience, through simulation studies not reported in this paper, reveals limited benefits when iterating more than once.

Next, we elaborate on each of the steps in the above algorithms for the simple case of a single index ($d = 1$). For the dimension reduction method in Step 4, one can use any of several existing methods, such as SIR or one of its variants, PPR, or the minimum average variance estimator (MAVE) of Xia, Tong, Li and Zhu (2002). These methods are for univariate responses and hence can also be applied in Step 1 when $q = 1$. However,

when $q > 1$, a different method is needed in Step 1 for the case of a multivariate response, and we recommend the dimension reduction method in Li, Wen and Zhu (2008). This and other results in the literature already demonstrate the $\sqrt{n}$-consistency of these dimension reduction methods.

For the smoothing involved in Step 5, one can choose any one-dimensional smoother. We employ the local polynomial smoother (Fan and Gijbels, 1996) to obtain estimators of the link function g and its derivative g', which will be used in the second stage of the estimation procedure. Specifically, for a kernel function $K(\cdot)$ on R^1 and a bandwidth sequence $b = b_n$, define $K_b(\cdot) = b^{-1}K(\cdot/b)$. For a fixed β and θ, the local linear smoother aims at minimizing the weighted sum of squares

$$\sum_{i=1}^{n} [Y_i - Z_i^{\mathrm{T}}\theta - d_0 - d_1(X_i^{\mathrm{T}}\beta - t)]^2 K_b(X_i^{\mathrm{T}}\beta - t)$$

with respect to the parameters d_ν, $\nu = 0, 1$. Let $h = h_n$ and $h_1 = h_{1n}$ denote the bandwidths for estimating $g(\cdot)$ and $g'(\cdot)$, respectively. A simple calculation shows that the local linear smoother with these specifications can be represented as

$$\hat{g}(t; \beta, \theta) = \sum_{i=1}^{n} W_{ni}(t, \beta)(Y_i - Z_i^{\mathrm{T}}\theta), \tag{2.1}$$

and

$$\hat{g}'(t; \beta, \theta) = \sum_{i=1}^{n} \widetilde{W}_{ni}(t, \beta)(Y_i - Z_i^{\mathrm{T}}\theta), \tag{2.2}$$

where

$$W_{ni}(t; \beta) = \frac{K_h(X_i^{\mathrm{T}}\beta - t)[S_{n,2}(t; \beta, h) - (X_i^{\mathrm{T}}\beta - t)S_{n,1}(t; \beta, h)]}{S_{n,0}(t; \beta, h)S_{n,2}(t; \beta, h) - S_{n,1}^2(t; \beta, h)]}, \tag{2.3}$$

$$\widetilde{W}_{ni}(t; \beta) = \frac{K_{h_1}(X_i^{\mathrm{T}}\beta - t)[(X_i^{\mathrm{T}}\beta - t)S_{n,0}(t; \beta, h_1) - S_{n,1}(t; \beta, h_1)]}{S_{n,0}(t; \beta, h_1)S_{n,2}(t; \beta, h_1) - S_{n,1}^2(t; \beta, h_1)]}, \tag{2.4}$$

and

$$S_{n,l}(t; \beta, h) = \frac{1}{n}\sum_{i=1}^{n} (X_i^{\mathrm{T}}\beta - t)^l K_h(X_i^{\mathrm{T}}\beta - t), \quad l = 0, 1, 2.$$

The above estimators are for generic fixed values of β and θ. To obtain the estimates needed in Step 5, one replaces them with the initial values $\hat{\beta}_0$ obtained in Step 1 and $\hat{\theta}_0$ obtained in Step 3, respectively. We will show in Theorem 2 that this results in standard convergence rates for the estimate of g.

Likewise, a local linear smoother can be employed in Step 2 for estimating the unknown function ϕ in model (1.2). The resulting estimator is defined as

$$\hat{\phi}(t;\hat{\beta}_Z) = \sum_{i=1}^{n} W_{ni}(t;\hat{\beta}_Z) Z_i.$$

Several possibilities are available for the estimator of θ_0 in Step 6, such as the profile approach (termed "partial regression" in Speckman, 1988) or the partial spline approach (Heckman, 1986). Here the the partial spline approach is not suitable for correlated X and Z, so we adopt a profile approach and a local linear smoother. In short, this amounts to minimizing, over all θ, the sum of squared errors,

$$\sum_{i=1}^{n} [Y_i - Z_i^{\mathrm{T}}\theta - \hat{g}(X_i^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0, \theta)]^2, \tag{2.5}$$

where $\hat{g}$ is the estimator in (2.1) of g, obtained by smoothing $Y_i - Z_i^{\mathrm{T}}\theta$ versus $X_i^{\mathrm{T}}\hat{\beta}_0$, and $\hat{\beta}_0$ is an initial estimator of β_0, which could be the initial estimator $\hat{\beta}_0$ in Step 4 or the refined estimator from Step 7 when an iterated estimator for θ_0 is desirable. Because this smoother is expressed as a function of θ, the estimate derived from (2.5) is a profile estimate. More details about the derivation and advantages of the profile approach can be found in Speckman (1988). Specifically, let $\hat{\beta}_0$ be the current estimator, $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^{\mathrm{T}}$, $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)^{\mathrm{T}}$, where

$$\begin{aligned} \tilde{Y}_i &= Y_i - \hat{g}_1(X_i^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0), & \tilde{Z}_i &= Z_i - \hat{g}_2(X_i^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0), \\ \hat{g}_1(t;\hat{\beta}_0) &= \textstyle\sum_{i=1}^n W_{ni}(t;\hat{\beta}_0) Y_i, & \hat{g}_2(t;\hat{\beta}_0) &= \textstyle\sum_{i=1}^n W_{ni}(t;\hat{\beta}_0) Z_i, \end{aligned}$$

with $\hat{g}_1$ and $\hat{g}_2$ the respective estimators of $g_1(t) = E(Y|X^{\mathrm{T}}\beta_0 = t)$ and $g_2(t) = E(Z|X^{\mathrm{T}}\beta_0 = t)$. The resulting partial regression estimator is thus

$$\hat{\theta} = (\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Y}}. \tag{2.6}$$

For the estimator of β_0 in Step 7, we propose a novel method that takes advantage of the constraint $\|\beta_0\| = 1$ and hence is more efficient than existing approaches, including the PPR approach in Härdle et al (1993), the MAVE method in Xia et al. (2002), and the least squares approaches of Carroll et al (1997) and Xia and Härdle (2006) for the single-index partial linear model in (1.1). It is worth mentioning that Xia and Härdle (2006) allow possible heteroscadestic structure in (1.1), and least squares approaches have been

standard dimension methods and lead to the same asymptotic variances for estimators of β_0. For instance, in the homoscadestic case, the estimator in Xia and Härdle (2006) has an asymptotic variance that is identical to that of Härdle et al (1993). Our approach, based on an estimating equation under the constraint $\|\beta_0\| = 1$, is computationally stable and asymptotically more efficient, i.e., its asymptotic variance is smaller. The efficiency gain can be attributed to a re-parametrization, making use of the constraint $\|\beta_0\| = 1$ by transferring restricted least squares to un-restricted least squares, which makes it possible to search for the solution of the estimating equation over a restricted region in the Euclidean space R^{p-1}.

Without loss of generality, we may assume that the true parameter β_0 has a positive component (otherwise, consider $-\beta_0$), say $\beta_{0r} > 0$ for $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$ and $1 \le r \le p$. For $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, let $\beta^{(r)} = (\beta_1, \ldots, \beta_{r-1}, \beta_{r+1}, \ldots, \beta_p)^{\mathrm{T}}$ be a $p-1$ dimensional parameter vector after removing the rth component β_r in β. Then we may write

$$\beta = \beta(\beta^{(r)}) = (\beta_1, \ldots, \beta_{r-1}, (1 - \|\beta^{(r)}\|^2)^{1/2}, \beta_{r+1}, \ldots, \beta_p)^{\mathrm{T}}. \tag{2.7}$$

The true parameter $\beta_0^{(r)}$ must satisfy the constraint $\|\beta_0^{(r)}\| < 1$, and β is infinitely differentiable in a neighborhood of $\beta_0^{(r)}$. This "remove-one-component" method for β has also been applied in Yu and Ruppert (2002).

To obtain the estimator, consider a Jacobian matrix of β with respect to $\beta^{(r)}$,

$$\mathbf{J}_{\beta^{(r)}} = \frac{\partial \beta}{\partial \beta^{(r)}} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}, \tag{2.8}$$

where γ_s $(1 \le s \le p, s \ne r)$ is a $p-1$ dimensional unit vector with sth component 1, and $\gamma_r = -(1 - \|\beta^{(r)}\|^2)^{-1/2}\beta^{(r)}$. To motivate the estimating equation, we start with the least squares criterion:

$$D(\beta) := \sum_{i=1}^{n} [Y_i - Z_i^{\mathrm{T}}\hat{\theta} - \hat{g}(X_i^{\mathrm{T}}\beta; \beta, \hat{\theta})]^2. \tag{2.9}$$

From (2.7) and (2.9) we find $D(\beta) = D(\beta(\beta^{(r)})) = \tilde{D}(\beta^{(r)})$. Therefore, we may obtain an estimator of $\beta_0^{(r)}$, say $\hat{\beta}^{(r)}$, by minimizing $\tilde{D}(\beta^{(r)})$, and then obtain an estimator of β_0, $\hat{\beta}$, via a transformation. This means that we transform a restricted least squares problem to an unrestricted least squares problem by solving the estimation equation:

$$\sum_{i=1}^{n} [Y_i - Z_i^{\mathrm{T}}\hat{\theta} - \hat{g}(X_i^{\mathrm{T}}\beta; \beta, \hat{\theta})]\hat{g}'(X_i^{\mathrm{T}}\beta; \beta, \hat{\theta})\mathbf{J}_{\beta^{(r)}}^T X_i = 0. \tag{2.10}$$

We define the resulting estimator $\hat{\beta}$ of β_0 as the final target estimator. Theorem 3 implies that our estimator for β_0 has a smaller limiting variance than the estimators in Xia and Härdle (2006) and Carroll et al. (1997).

With $\hat{\theta}$ and $\hat{\beta}$, the final estimator $\hat{g}^*$ of g in Step 8 can be defined by

$$\hat{g}^*(t) := \hat{g}(t; \hat{\beta}, \hat{\theta}) = \sum_{i=1}^{n} W_{ni}(t; \hat{\beta})(Y_i - Z_i^{\mathrm{T}}\hat{\theta}),$$

and the estimator $\hat{\sigma}^2$ of σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[Y_i - Z_i^{\mathrm{T}}\hat{\theta} - \hat{g}^*(X_i^{\mathrm{T}}\hat{\beta})]^2$. Asymptotic results for the final parameter estimates of θ and β are established in Theorem 1 and Theorem 2, and results for the link estimate of g follow from Theorem 4.

REMARK 1 We consider a homoscedastic model of (1.1) with $d = 1$ in model (1.2). While the estimation procedure can be extended easily to heteroscedastic errors, an additional dimension reduction assumption on the variance function of of η, given X, is needed to avoid the curse of high dimensional smoother needed in Step 2 to estimate ϕ. This assumption requires that this variance function is also a function of a few indices based on X. Moreover, the extension of asymptotic theory is not straightforward. For instance, the asymptotic efficiency of the estimator β_0 is technically challenging in the heteroscedastic case and its study is beyond the scope of this paper.

REMARK 2 So far, we have assumed that $d = 1$. This assumption can be extended without difficulty to the general case where d might be greater than 1. In this case, a multivariate smoother will be employed for estimating $\phi(\cdot)$. The asymptotic results for the parameter estimates of β and θ remain unchanged, except that the rate of convergence for the link estimate of $\phi(\cdot)$ changes with the dimension of d.

REMARK 3 Other dimension reduction approaches, such as MAVE (Xia et al., 2002) and other variants of SIR, such as SIR2 (Li, 1991) and SAVE (Cook and Wiseberg, 1991), could be employed in Steps 1 and 4 for the case of $q = d = 1$ in (1.2), especially when SIR fails for the case of symmetric design of X. While MAVE is perhaps the most efficient method of all, the benefits over SIR are limited, as all estimates are updated in Stage 2, and it is in this step where the major efficiency gains occur. In addition, MAVE is computationally

more intensive than SIR and encounters difficulties in estimating β_Z, unless the covariate Z is one-dimensional and the dimension d of β_Z is also small. In fact, the $\sqrt{n}$-consistency may not hold when $d > 3$ in (1.2) as shown in Xia, Tong, Li and Zhu (2002).

Also, SIR2/SAVE was shown in Li and Zhu (2007) to be not $\sqrt{n}$-consistent, unless a bias correction is performed. In contrast, either SIR or pHd (Li, 1992) can be employed to identify the directions when $d > 1$ and $q = 1$, and both lead to $\sqrt{n}$-consistency.

REMARK 4 When the dimension q of Z is greater than 1, a multivariate extension of SIR (Li et al., 2003) can be employed conceptually in Step 1 of the algorithm. However, the number of observations per slice may become sparse, so we recommend an alternative multivariate approach as in Li, Wen and Zhu (2008) or Zhu, Zhu, Ferré and Wang (2008) in Step 1.

REMARK 5 The single-index assumption in (1.1) can be easily extended to multiple indices through SIR or its variants, but the estimation of the multivariate link function g would encounter the curse of high dimensionality. Since no more than three indices will be needed in many applications, the approach in this paper can indeed be extended in practice to multiple indices.

2.2 Main results

In this section, the $\sqrt{n}$ asymptotics for initial estimates of β_0 and θ_0 in Stage 1 are taken for granted as they follow from existing results, so we do not formally list the needed assumptions for this to hold but have provided sources after Theorem 1 below. However, the asymptotics for the initial estimate of g and each of the parametric and nonaprametric estimates in Stage 2 are fully developed in Section 2.2 with detailed assumptions listed for each estimator.

In order to study the asymptotic behavior of the estimators, we list the following conditions:

C1. (i) The distribution of X has a compact support set A.

(ii) The density function of $X^{\mathrm{T}}\beta$ is positive and satisfies a Lipschitz condition of order 1 for β in a neighborhood of β_0. Further, $X^{\mathrm{T}}\beta_0$ has a positive and bounded density function $f(t)$ on $\mathcal{T}$, where $\mathcal{T} = \{t = x^{\mathrm{T}}\beta_0 : x \in A\}$.

C2. (i) The functions g and g_{2i} have two bounded and continuous derivatives, where g_{2i} is the ith component of $g_2(t)$, $1 \le i \le q$;

(ii) g_{3j} satisfies a Lipschitz condition of order 1, where g_{3j} is the jth component of $g_3(t)$, and $g_3(t) = E(X|X^{\mathrm{T}}\beta_0 = t)$, $1 \le j \le p$.

C3. (i) The kernel K is a bounded, continuous and symmetric probability density function, satisfying

$$\int_{-\infty}^{\infty} u^2 K(u)du \neq 0, \quad \int_{-\infty}^{\infty} |u|^2 K(u)du < \infty;$$

(ii) K satisfies a Lipschitz condition on R^1.

C4. (i) $\sup_t E(\|Z\|^2|X_1^{\mathrm{T}}\beta_0 = t) < \infty$;

(ii) $E(e) = 0$, $\mathrm{Var}(e) = \sigma^2 < \infty$, $E(e^4) < \infty$.

C5. (i) $nh^2/\log^2 n \to \infty$, $\limsup_{n\to\infty} nh^5 < \infty$;

(ii) $nhh_1^3/\log^2 n \to \infty$, $nh^4 \to 0$, $\limsup_{n\to\infty} nh_1^5 < \infty$.

C6. (i) $\mathbf{\Sigma} = \mathrm{Cov}(Z - E(Z|X^{\mathrm{T}}\beta_0))$ is a positive definite matrix;

(ii) $\mathbf{V} = E[g'(X^{\mathrm{T}}\beta_0)^2 \mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}} X X^{\mathrm{T}} \mathbf{J}_{\beta_0^{(r)}}]$ is a positive definite matrix, where $\mathbf{J}_{\beta_0^{(r)}}$ is defined by (2.8).

REMARK 6 The Lipschitz condition and the two derivatives in C1 and C2 are standard smoothness conditions. C3 is the usual assumption for second-order kernels. C1 is used to bound the density function of $X^{\mathrm{T}}\beta$ away from zero. This ensures that the denominators of $\hat{g}(t;\beta,\theta_0)$ and $\hat{g}'(t;\beta,\theta_0)$ are, with high probability, bounded away from 0 for $t = x^{\mathrm{T}}\beta$, $x \in A$ and β near β_0. C4 is a necessary condition for the asymptotic normality of an estimator. In C5(i), the range of h for the estimators $\hat{\theta}$ and $\hat{g}$ is fairly large and contains the rate $n^{-1/5}$ of "optimal" bandwidths. However, when analyzing the asymptotic properties of the estimator $\hat{\beta}$ of β_0, we have to estimate the derivative g' of g. As is well known, the convergence rate of the estimator of g' is slower than that of the estimator of g if the same

bandwidth is used. This leads to a slower convergence rate for $\hat{\beta}$ than $\sqrt{n}$, unless we use a kernel of order 3 or *undersmoothing* to deal with the bias of the estimator. This motivates the introduction of another bandwidth h_1 in C5(ii) to control the variability of the estimator of g', and condition C5(ii) for bandwidths h and h_1. Chiou and Müller (1998) also consider the use of two bandwidths to construct the estimator of β in a relevant model. C6 ensures that the limiting variances for the estimators $\hat{\theta}$ and $\hat{\beta}$ exist.

The following theorems state the asymptotic behavior of the estimators proposed in Section 2.1. We first establish the asymptotic efficiency of $\hat{\theta}$.

THEOREM 1 *Suppose that conditions* C1, C2(i), C3(i), C4(i), C5(i) *and* C6(i) *hold. When* $\|\hat{\beta}_Z - \beta_Z\| = O_P(n^{-1/2})$ *and* $\|\hat{\beta}_0 - \beta_0\| = O_P(n^{-1/2})$, *we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \sigma^2 \mathbf{\Sigma}^{-1}).$$

REMARK 7 Carroll et al.(1997) give similar results with $\beta = 1$ and $p = 1$ (The case of a partially linear model). Theorem 1 generalizes their Theorems 2 and 3.

In Theorem 1, when we start with $\sqrt{n}$-consistent estimators for β_Z and β_0, $\hat{\theta}$ is consistent for θ_0 with the same asymptotic efficiency as an estimator that we would have obtained had we known β_0 and g, and thus the oracle property. Numerous examples of $\sqrt{n}$- consistent estimators already exist in the literature. For instance, Hall (1989) showed that one can obtain a $\sqrt{n}$-consistent estimator for β_0 using projection pursuit regression. Under the linearity condition that is slightly weaker than elliptical symmetry of X, Li (1991), Hsing and Carroll (1992) and Zhu and Ng (1995) proved that SIR, proposed by Li (1991), leads to a $\sqrt{n}$-consistent estimator of β_Z and of β_0, the latter when Z is not present in (1.1). Li and Zhu (2007) further show that, when including a bias-correction and under a condition almost equivalent to normality of X, sliced average variance estimation (SAVE, Cook and Weisberg 1991) performs similarly. We expect the results for β_0 to hold when Z is dependent of X, provided a good estimator of β_Z is available. Under very general regularity conditions and for $q = 1$, Xia, Tong, Li, and Zhu (2002) proposed the minimum average variance estimation

(MAVE) and Xia (2006) a refined version of MAVE, and both methods can provide $\sqrt{n}$-consistent estimators for the single-index β_0. However, there is no result in the literature regarding MAVE when the dimension of Z is larger than 1, and the $\sqrt{n}$-consistency needs further study when d is larger than or equal to 3, even for univariate Z. Therefore, for general theory, SIR may be a good choice for the initial estimators of β_Z and β_0.

THEOREM 2 *Suppose that conditions* C1–C6 *hold. If the rth component of β_0 is positive, we have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \sigma^2 \mathbf{J}_{\beta_0^{(r)}} \mathbf{V}^{-1} \mathbf{Q} \mathbf{V}^{-1} \mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}}),$$

where $\mathbf{Q} = E\{g'(X^{\mathrm{T}}\beta_0)^2 \mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}} [X - E(X|X^{\mathrm{T}}\beta_0)][X - E(X|X^{\mathrm{T}}\beta_0)]^{\mathrm{T}} \mathbf{J}_{\beta_0^{(r)}}\}$, $\mathbf{V}$ *and* $\mathbf{J}_{\beta_0^{(r)}}$ *are defined in condition C6.*

From Härdle et al (1993) and Carroll et al (1997), we can see that the estimator $\hat{\beta}$ of β has an asymptotic variance that corresponds to a generalized inverse $\sigma^2 \mathbf{Q}_1^-$ where

$$\mathbf{Q}_1 = E\left\{ g'(X^{\mathrm{T}}\beta_0)^2 \left[X - E(X|X^{\mathrm{T}}\beta_0)\right] \left[X - E(X|X^{\mathrm{T}}\beta_0)\right]^{\mathrm{T}} \right\}.$$

Note that there may be infinitely many inverse matrices of $\mathbf{Q}_1$, but there is a unique generalized inverse associated with the Jacobian $J_{\beta_0^{(r)}}$. The following theorem shows that the variance-cavariance matrix in Theorem 2 is smaller than $\sigma^2 \mathbf{Q}_1^-$, the variance associated with $\mathbf{J}_{\beta_0^{(r)}}$, in the sense that $\sigma^2 \mathbf{Q}_1^- - \sigma^2 \mathbf{J}_{\beta_0^{(r)}} \mathbf{V}^{-1} \mathbf{Q} \mathbf{V}^{-1} \mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}}$ is a non-negative definite matrix. We use the usual notation: for two non-negative matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ denotes that $\mathbf{A} - \mathbf{B}$ is a non-negative definite matrix.

THEOREM 3 *Under the conditions of Theorem 2, we have*

i) *there is a generalized inverse of $\mathbf{Q}_1$ that is of the form $\mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}} \mathbf{Q}^{-1} \mathbf{J}_{\beta_0^{(r)}}$;*

ii) $\mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}} \mathbf{Q}^{-1} \mathbf{J}_{\beta_0^{(r)}} \geq \mathbf{J}_{\beta_0^{(r)}} \mathbf{V}^{-1} \mathbf{Q} \mathbf{V}^{-1} \mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}}$.

REMARK 8 Theorem 3 shows that our estimator of β_0 is asymptotically more efficient than those of Härdle et al.(1993) and of Carroll et al. (1997). In addition, Carroll et al.(1997) use an iterated procedure to estimate β_0 and θ_0, while our estimation procedure does not require iteration.

From Theorem 2, we obtain an asymptotic result regarding the angle between $\hat{\beta}$ and β_0, which can be used to study issues of sufficient dimension reduction (SDR). We refer to Cook (1998, 2007) for more details.

COROLLARY 1 *Suppose that the conditions of Theorem 2 hold. Then*

$$\cos(\hat{\beta}, \beta_0) - 1 = O_P(n^{-1/2}),$$

where $\cos(\hat{\beta}, \beta_0)$ *is the cosine of the angle between* $\hat{\beta}$ *and* β_0.

The next two theorems provide the convergence rate of the estimator $\hat{g}^*(\cdot)$ of $g(\cdot)$ and the asymptotic normality of the estimator of σ^2.

THEOREM 4 *Suppose that the conditions of Theorem 1 hold. If* $\|\hat{\beta} - \beta_0\| = O_P(n^{-1/2})$. *Then*

$$\sup_{(x,\beta)\in\mathcal{A}_n} |\hat{g}^*(x^{\mathrm{T}}\beta) - g(x^{\mathrm{T}}\beta_0)| = O_P((nh/\log n)^{-1/2}),$$

where $\mathcal{A}_n = \{(x, \beta) : (x, \beta) \in A \times R^p, \|\beta - \beta_0\| \le cn^{-1/2}\}$ *for a constant* $c > 0$.

THEOREM 5 *Suppose that conditions C1–C6 hold and* $0 < \mathrm{Var}(e_1^2) < \infty$. *Then*

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2)/(\mathrm{Var}(e_1^2))^{1/2} \xrightarrow{D} N(0, 1).$$

Note that $n^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}} \xrightarrow{P} \mathbf{\Sigma}$ in Lemma A.5 of the Appendix. By Theorems 1 and 4 , we obtain

$$(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{1/2}(\hat{\theta} - \theta_0)/\hat{\sigma} \xrightarrow{D} N(0, \mathbf{I}_q).$$

We are now in the position to construct confidence regions for θ_0. From Theorem 10.2d in Arnold (1981) we obtain the following result.

THEOREM 6 *Under the conditions of Theorem 5, we have*

$$(\hat{\theta} - \theta_0)^{\mathrm{T}}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})(\hat{\theta} - \theta_0)/\hat{\sigma}^2 \xrightarrow{D} \chi_q^2,$$

where χ_q^2 *is chi-square distributed with* q *degrees of freedom. Let* $\chi_q^2(1-\alpha)$ *be the* $(1-\alpha)$*-quantile of* χ_q^2 *for* $0 < \alpha < 1$, *an asymptotic confidence region of* θ_0 *is*

$$R_\alpha = \{\theta : (\hat{\theta} - \theta)^{\mathrm{T}}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})(\hat{\theta} - \theta)/\hat{\sigma}^2 \le \chi_q^2(1-\alpha)\}.$$

To construct confidence regions for β_0, a plug-in estimator of the limiting variance of $\hat{\beta}$ is needed. We respectively define the following estimators $\hat{\mathbf{V}}$ and $\hat{\mathbf{Q}}$ of $\mathbf{V}$ and $\mathbf{Q}$ by

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}'(X_i^{\mathrm{T}}\hat{\beta}; \hat{\beta}, \hat{\theta})^2 \mathbf{J}_{\hat{\beta}^{(r)}}^{\mathrm{T}} X_i X_i^{\mathrm{T}} \mathbf{J}_{\hat{\beta}^{(r)}}$$

and

$$\hat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}'(X_i^{\mathrm{T}}\hat{\beta}; \hat{\beta}, \hat{\theta})^2 \mathbf{J}_{\hat{\beta}^{(r)}}^{\mathrm{T}} [X_i - \hat{g}_3(X_i^{\mathrm{T}}\hat{\beta}; \hat{\beta})][X_i - \hat{g}_3(X_i^{\mathrm{T}}\hat{\beta}; \hat{\beta})]^{\mathrm{T}} \mathbf{J}_{\hat{\beta}^{(r)}},$$

where $\hat{g}_3(t; \hat{\beta}) = \sum_{i=1}^{n} W_{ni}(t; \hat{\beta}) X_i$ is the estimator of $g_3(t) = E(X | X^{\mathrm{T}}\beta_0 = t)$ and $\mathbf{J}_{\hat{\beta}^{(r)}}$ is the estimator of $\mathbf{J}_{\beta_0^{(r)}}$. It is easy to prove that $\mathbf{J}_{\hat{\beta}^{(r)}} \xrightarrow{P} \mathbf{J}_{\beta_0^{(r)}}$, $\hat{\mathbf{V}} \xrightarrow{P} \mathbf{V}$ and $\hat{\mathbf{Q}} \xrightarrow{P} \mathbf{Q}$. Then for any $p \times l$ matrix $\mathbf{A}$ of full rank with $l < p$, Theorems 2 and 5 imply that

$$(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{J}_{\hat{\beta}^{(r)}}\hat{\mathbf{V}}^{-1}\hat{\mathbf{Q}}\hat{\mathbf{V}}^{-1}\mathbf{J}_{\hat{\beta}^{(r)}}^{\mathrm{T}}\mathbf{A})^{-1/2}\mathbf{A}^{\mathrm{T}}(\hat{\beta} - \beta_0)/\hat{\sigma} \xrightarrow{D} N(0, \mathbf{I}_l).$$

We again use Theorem 10.2d in Arnold (1981) to obtain the following limiting distribution.

THEOREM 7 *Suppose that the conditions of Theorem 5 hold. Then*

$$(\hat{\beta} - \beta_0)^{\mathrm{T}}\mathbf{A}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{J}_{\hat{\beta}^{(r)}}\hat{\mathbf{V}}^{-1}\hat{\mathbf{Q}}\hat{\mathbf{V}}^{-1}\mathbf{J}_{\hat{\beta}^{(r)}}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}(\hat{\beta} - \beta_0)/\hat{\sigma}^2 \xrightarrow{D} \chi_l^2.$$

The asymptotic confidence region of $\mathbf{A}^{\mathrm{T}}\beta_0$ is, letting $\chi_l^2(1-\alpha)$ be the $(1-\alpha)$-quantile of χ_l^2 for $0 < \alpha < 1$,

$$R_\alpha = \{\mathbf{A}^{\mathrm{T}}\beta : (\hat{\beta} - \beta)^{\mathrm{T}}\mathbf{A}(n^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{J}_{\hat{\beta}^{(r)}}\hat{\mathbf{V}}^{-1}\hat{\mathbf{Q}}\hat{\mathbf{V}}^{-1}\mathbf{J}_{\hat{\beta}^{(r)}}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}(\hat{\beta} - \beta)/\hat{\sigma}^2 \le \chi_l^2(1-\alpha)\}.$$

3 Simulation study

In this section, we examine the performance of the procedures in Section 2, for the estimation of both β_0 and θ_0. We report the accuracy of estimators using PPR and SIR as dimension-reduction methods. The sample size for the simulated data is $n = 100$ and the number of simulated samples is 2000 for the parametric components. When SIR is applied, using 5 or 10 elements per slice generally yields good results. In other words, each slice contains 10 to 20 points. A quadratic model of the form

$$Y = (X^{\mathrm{T}}\beta_0 - 0.5)^2 + Z\theta_0 + 0.2e,$$

was used, where $\theta_0 = 1$ is a scalar, $\beta_0 = (0.75, 0.5, -0.25, -0.25, 0.25)^{\mathrm{T}}$, X is a 5-dimensional vector with independent uniform [0,1] components, and e is a standard normal variable. The dependency between X and Z was prescribed by defining Z as a binary variable with probability $\exp(\beta_Z X)/(1 + \exp(X^{\mathrm{T}}\beta_Z))$ to be 1 and 0 otherwise. Two extreme cases of β_Z are reported in Table 1 and Table 2, one based n choosing the same value as β_0 with $\beta_Z = \beta_0$, and the other on $\beta_Z = (0.5, 0, 0.5, 0.5, -0.5)^{\mathrm{T}}$, so that β_Z is orthogonal to β_0. We also checked scenarios where β_Z and β_0 are neither orthogonal nor parallel to each other, and the results are in agreement with the two extreme cases reported here.

For the smoothing steps, we used a local linear smoother with a Gaussian kernel throughout. A product Gaussian kernel was used when bivariate smoothing was involved and equal bandwidths were selected for each kernel to save computing time. A pilot study revealed that the bandwidth chosen at the first stage to estimate the residual η has little effect on the accuracy of the final estimates of θ_0, so we choose an initial bandwidth of 0.5 to estimate ϕ in (1.2), as this value was frequently selected by generalized cross validation (GCV). The subsequent smoothing steps utilized the GCV method as proposed in Craven and Wahba (1979). For instance, when estimating g and θ_0 in the second stage, the GCV statistic is given by the formula

$$\mathrm{GCV}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Z_i^{\mathrm{T}}\hat{\theta} - \hat{g}_h(X_i^{\mathrm{T}}\hat{\beta}; \hat{\beta}, \hat{\theta})^2/(n^{-1}\mathrm{tr}(\mathbf{I} - \mathbf{S}_h))^2, \tag{3.1}$$

where $\hat{g}_h(\cdot)$ is the estimator of $g(\cdot)$ with a bandwidth h and $\mathbf{S}_h$ is the smoothing matrix corresponding to a bandwidth of h. The GCV bandwidth was selected to minimize (3.1). We use the optimal bandwidth, $\hat{h}_{\mathrm{opt}}$, for $\hat{g}$ and $\hat{\theta}$. When calculating the estimator $\hat{\beta}$, we chose the bandwidths,

$$h = \hat{h}_{\mathrm{opt}} n^{1/5} n^{-1/3} = \hat{h}_{\mathrm{opt}} n^{-2/15} \quad \text{and} \quad \hat{h}_1 = \hat{h}_{\mathrm{opt}}, \tag{3.2}$$

respectively, because this guarantees that the required bandwidth has the correct order of magnitude for optimal asymptotic performance [see Carroll al.(1997), Stute and Zhu (2005), and Zhu and Ng (2003)]. Note that choices (3.2) satisfy condition C5(ii). Relevant discussion on choosing two distinct bandwidths can be found in Chiou and Müller (1998).

In the simulation, PPR and SIR were used to obtain the initial estimators of β_0 and β_Z. The notation SIR_c means that when we used SIR to estimate β_Z, the number of data points per slice is c. The resulting estimates for θ_0 and the one-step iterated estimates are summarized in Tables 1 and 2, where we report bias, standard deviation (SD), and mean square error (MSE). The case with known β_0 is also reported in the last row and serves as a gold standard. The right columns under "One-step iterated estimate" in Tables 1 and 2 represent the results obtained when iterating the algorithms in Section 2.1 one more time after obtaining the estimates in the left columns.

Tables 1 and 2 are about here

From Tables 1 and 2 we find that the three methods have small mean square errors with projection pursuit regression outperforming both SIR procedures. This is expected, as the simulated model structure satisfies the additive assumption of PPR and the estimates of the β-directions were iteratively updated through estimates of the unknown link functions, ϕ and g. In other non-additive situations, SIR might be more reliable than PPR. Iterated estimates improved the results for all cases and markedly so for the orthogonal case. Compared to the case when β_0 is known, PPR typically attains 80% or more of the efficiency after one iteration.

For the estimation of β_0, we computed the angle (in radians) between $\hat{\beta}$ and β_0 as a measure of accuracy. The mean, standard deviation (SD), and mean squared error (MSE) of the angle between $\hat{\beta}$ and β_0 are reported in Table 3. Here, PPR leads to by far superior estimates compared to SIR.

Table 3 is about here

The performance of the nonparametric estimates for g is demonstrated in Figure 1. Again, GCV was used for bandwidth choice and compared to the estimates based on the optimal

fixed bandwidth. The true function g and the mean of each estimated g-function over the 2000 replicates are plotted. In general, GCV seems to work well for all parametric and nonparametric components. This is consistent with the results reported in Chen and Shiau (1994) for the analysis of partially linear models based on generalized cross validation (GCV). Theoretical properties of the current models in regard to GCV will be a topic for further investigation.

Figure 1 is about here

A final remark is that we tried to compare our procedure with that proposed in Carroll, *et al.* (1997), for the quadratic model used in the above simulations with β_Z and β_0 orthogonal. However, we were not able to obtain any results for the method in Carroll et al. (1997), as their procedure seems to be very sensitive to the choice of the initial estimates. We then used our estimates for β_0 and θ_0 as the initial values for their procedure. Nevertheless, we were still unable to obtain any meaningful comparison results as out of the seven attempted trials their procedure crashed six times on the first simulation and once on the second simulation. Since θ_0 is only a scalar, we postulate that their procedure has difficulties with high dimensional β_0, which is here a five-dimensional vector.

4 Data Example

We analyze the Boston Housing data mentioned in Section 1. The goal is to determine the effect of the various variables on housing price, including a binary variable, which describes whether the census tract borders the Charles River. According to Harrison and Rubinfeld (1978), bordering the river should have a positive effect on the median housing price of the census tract. They used a linear model that included a log transformation for the response variable and three of the covariates, and power transformations for three other covariates.

Their final model is

$$\begin{aligned}\log(MV) &= a_1 + a_2 RM^2 + a_3 AGE + a_4 \log(DIS) + a_5 \log(RAD) + a_6 TAX \\ &\quad + a_7 PTRATIO + a_8 (B - 0.63)^2 + a_9 \log(LSTAT) + a_{10} CRIM \\ &\quad + a_{11} ZN + a_{12} INDUS + a_{13} CHAS + a_{14} NOX^p + e.\end{aligned}$$

The coefficient a_{13} is estimated to be 0.088, which is significant with a p-value of less than 0.01 for the hypothesis $H_0 : a_{13} = 0$ versus $H_1 : a_{13} \neq 0$. The coefficient of determination R^2 attained by their analysis is 0.81, where R^2 is the squared correlation between the true dimension-reduction variable $X^{\mathrm{T}}\beta_0$ and the estimated dimension-reduction variable $X^{\mathrm{T}}\hat{\beta}_0$.

This data set was also analyzed by Chen and Li (1998), who used sliced inverse regression with all thirteen covariates. After examining the initial results, Chen and Li (1998) trimmed the data and then dropped some of the variables. We fit the data on the first SIR direction of the initial analysis reported in their article and obtained an R^2 of 0.705 using GCV bandwidth 0.43. Note that the assumptions of sliced inverse regression are probably not met because some of the covariates are discrete. We thus proposed to use a partial-linear single-index model. Several choices of Z were attempted, but they did not yield better results, in terms of R^2, than the one using only the Charles River variable as Z and the other covariates as X. We thus focus on this model, where a log transformation was applied on Y.

To select the number of observations per slice in the dimension reduction step of SIR, we borrow our experience in the simulation presented in Section 3, where 5 or 10 observations per slice worked well for a total sample size of 100, leading to about 20 to 10 slices. Since the sample size for the housing data is much larger, we use SIR with 20 data points per slice and this leads to a total of 26 slices. As Chen and Li (1998) pointed out, SIR is not sensitive to the choice of slice number, and they tried slicing with 10 or 30 points per slice leading to 17 or 50 slices, and obtain very similar results. The GCV bandwidth for estimating g and θ is 0.367, which is smaller than the bandwidth 0.43 chosen by the GCV method for the SIR approach of Chen and Li (1998). To estimate β by (2.10), the bandwidths selected by (3.2) for $h = 0.16$ and for h_1 is 0.367. The R^2 is 0.8047, which is essentially equal to that obtained by Harrison and Rubinfeld and higher than that using SIR on all thirteen variables.

The value of the test statistic for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ is 3.389 when the degrees of freedom are calculated according to Hastie and Tibshirani, and 3.419 when n degrees of freedom are used. Either way the result is significant with p-value < 0.01.

We also omitted the Charles River variable and used a dimension-reduction model on Y and X. After obtaining an estimate for β_0, we then estimate the relationship between Y and $X^{\mathrm{T}}\hat{\beta}$. GCV yields a bandwidth of 0.16, and we obtain $R^2 = 0.8021$. Even though the Charles River variable is significant, its inclusion leads to only a minor increase in R^2.

Figure 2 is about here

Figure 2 shows the estimated g along with the data. On the x-axis of the above graph, the estimated value $x^{\mathrm{T}}\hat{\beta}$ is given, and on the y-axis, the estimated value $\hat{g}^*(t)$. Figure 2 shows a downward trend in the effective dimension reduction (EDR) variate obtained. The upward curvature of the function at high values of the EDR variate may or may not be a real effect.

The advantage of our procedure over the one used by Harrison and Rubinfeld is that Harrison and Rubinfeld have to make choices regarding transformations for every variable in the model. We only need to choose the bandwidth or bandwidths used for smoothing.

5 Proofs of Theorems

Since the proofs of the theorems are rather long, the proofs of Theorems 1–4 are presented in this section, and more details of the proofs are divided into Lemmas A.2–A.7 in the Appendix.

In this section and the Appendix, we use $c > 0$ to represent any constant which may take different values for each appearance, and $a \wedge b = \min(a, b)$.

Proof of Theorem 1. Denote

$$\tilde{G} = (g(X_1^{\mathrm{T}}\beta_0) - \hat{g}(X_1^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0, \theta_0), \ldots, g(X_n^{\mathrm{T}}\beta_0) - \hat{g}(X_n^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0, \theta_0))^{\mathrm{T}}.$$

From (2.6) we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{G} + \sqrt{n}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}e.$$

Lemma A.5 in the Appendix implies

$$n(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1} \xrightarrow{P} \mathbf{\Sigma}^{-1}. \tag{5.1}$$

Therefore, Lemma A.6 in the Appendix leads to

$$\sqrt{n}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{G} \xrightarrow{P} 0.$$

It remains to show that

$$\sqrt{n}(\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}e \xrightarrow{D} N(0, \sigma^2\mathbf{\Sigma}^{-1}). \tag{5.2}$$

Since

$$\begin{aligned}\tilde{\mathbf{Z}}^{\mathrm{T}}e &= \sum_{i=1}^{n}[Z_i - g_2(X_i^{\mathrm{T}}\beta_0)]e_i + \sum_{i=1}^{n}[g_2(X_i^{\mathrm{T}}\beta_0) - \hat{g}_2(X_i^{\mathrm{T}}\hat{\beta}_0; \hat{\beta}_0)]e_i \\ &=: M_1 + M_2.\end{aligned}$$

The central limit theorem implies $n^{-1/2}M_1 \xrightarrow{D} N(0, \mathbf{\Sigma})$. Similarly to the proof of (A.17), it is easy to obtain that $n^{-1/2}M_2 \xrightarrow{P} 0$. This together with (5.1) and Slutsky's Theorem proves (5.2), and hence Theorem 1. □

Proof of Theorem 2. The proof is divided into two steps: From (2.9), step (I) provides the existence of the least squares estimator $\hat{\beta}$ of β_0, and from (3.1), step (II) proves the asymptotic normality of $\hat{\beta}$.

(I) **Proof of existence.** We prove the following fact: Under conditions C1–C5 and with probability one there exists an estimator of β_0 minimizing expression (2.9) in $\mathcal{B}_{1n}$, where $\mathcal{B}_{1n} = \{\beta : \|\beta - \beta_0\| = B_1 n^{-1/2}\}$ for some constant such that $0 < B_1 < \infty$.

In fact, let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$. We have

$$\begin{aligned} D(\beta) &= (\mathbf{Y} - \mathbf{Z}\hat{\theta})^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)(\mathbf{Y} - \mathbf{Z}\hat{\theta}) \\ &= (\mathbf{Y} - \mathbf{Z}\theta_0)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)(\mathbf{Y} - \mathbf{Z}\theta_0) \\ &\quad - 2(\mathbf{Y} - \mathbf{Z}\theta_0)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)\mathbf{Z}(\hat{\theta} - \theta_0) \\ &\quad + \{\mathbf{Z}(\hat{\theta} - \theta_0)\}^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)^{\mathrm{T}}(\mathbf{I} - \mathbf{S}_\beta)\mathbf{Z}(\hat{\theta} - \theta_0) \\ &=: D_1(\beta) - D_2(\beta) + D_3(\beta). \end{aligned}$$

The same arguments as in the proof of Theorem 1 can be used to obtain that $D_2(\beta) = R_0 + o_P(1)$ and $D_3(\beta) = o_P(1)$, where R_0 is a constant independent of β. This implies $D(\beta) = D_1(\beta) - R_0 + o_P(1)$. Thus, minimizing $D(\beta)$ simultaneously with respect to β is very much like separately minimizing $D_1(\beta)$ with respect to β. It follows from (2.7) that we only need to prove the existence of an estimator of $\beta_0^{(r)}$ in $\mathcal{B}_{2n}$, where $\mathcal{B}_{2n} = \{\beta^{(r)} : \|\beta^{(r)} - \beta_0^{(r)}\| = B_2 n^{-1/2}\}$ for some constant such that $0 < B_2 < \infty$. Since $R(\beta^{(r)}) = (-\frac{1}{2})\frac{\partial D_1(\beta)}{\partial \beta^{(r)}}$, where $R(\beta^{(r)})$ is defined in (A.19) of Lemma A.7. For an arbitrary $\beta^{(r)} \in \mathcal{B}_{2n}$ with the value of constant B_2 in $\mathcal{B}_{2n}$ to be determined, we have from Lemma A.7 below that

$$\begin{aligned} &(\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} R(\beta^{(r)}) \\ &\quad = (\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} U(\beta_0^{(r)}) - n(\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} \mathbf{V}(\beta^{(r)} - \beta_0^{(r)}) + o_P(1). \end{aligned} \tag{5.3}$$

The following arguments are similar to those used by Weisberg and Welsh (1994), which in turn use (6.3.4) of Ortega and Rheinboldt (1973). We note that term (5.3) is dominated by the term $\sim B_2^2$ because $\sqrt{n}\|\beta^{(r)} - \beta_0^{(r)}\| = B_2$, whereas $|(\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} U(\beta_0^{(r)})| = B_2 O_P(1)$ and $n(\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} \mathbf{V}(\beta^{(r)} - \beta_0^{(r)}) \sim B_2^2$. So, for any given $\eta > 0$, if B_2 is chosen large enough, then it will follows that $(\beta^{(r)} - \beta_0^{(r)})^{\mathrm{T}} R(\beta_0^{(r)}) < 0$ on an event with probability $1 - \eta$. From the arbitrariness of η, we can prove the existence of the least squares estimator of $\beta_0^{(r)}$ in $\mathcal{B}_{2n}$ as in the proof of Theorem 5.1 of Welsh (1989). The details are omitted.

(II) **Proof of asymptotic normality.** From step (I) we find that $\hat{\beta}^{(r)}$ is a solution in $\mathcal{B}_{2n}$ to the equation $R(\beta^{(r)}) = 0$. That is, $R(\hat{\beta}^{(r)}) = 0$. By Lemma A.7, we have

$$0 = U(\beta_0^{(r)}) - n\mathbf{V}(\hat{\beta}^{(r)} - \beta_0^{(r)}) + o_P(\sqrt{n}\ ),$$

and hence

$$\sqrt{n}(\hat{\beta}^{(r)} - \beta_0^{(r)}) = \mathbf{V}^{-1} n^{-1/2} U(\beta_0) + o_P(1).$$

We now consider the estimator $\hat{\beta}$. A simple calculation yields

$$\frac{2\sqrt{1 - \|\beta_0^{(r)}\|^2}}{\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} + \sqrt{1 - \|\beta_0^{(r)}\|^2}} - 1 = \frac{\sqrt{1 - \|\beta_0^{(r)}\|^2} - \sqrt{1 - \|\hat{\beta}^{(r)}\|^2}}{\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} + \sqrt{1 - \|\beta_0^{(r)}\|^2}} = O_P(n^{-1/2}),$$

and hence

$$\begin{aligned}
&\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} - \sqrt{1 - \|\beta_0^{(r)}\|^2} \\
&\quad = -\frac{(\hat{\beta}^{(r)} + \beta_0^{(r)})^{\mathrm{T}}(\hat{\beta}^{(r)} - \beta_0^{(r)})}{\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} + \sqrt{1 - \|\beta_0^{(r)}\|^2}} \\
&\quad = -\frac{2\beta_0^{(r)T}(\hat{\beta}^{(r)} - \beta_0^{(r)}) + \|\hat{\beta}^{(r)} - \beta_0^{(r)}\|^2}{\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} + \sqrt{1 - \|\beta_0^{(r)}\|^2}} \\
&\quad = -\frac{\beta_0^{(r)T}(\hat{\beta}^{(r)} - \beta_0^{(r)})}{\sqrt{1 - \|\beta_0^{(r)}\|^2}} + O_P(n^{-1}).
\end{aligned}$$

It follows from (2.7) and the above equation, that

$$\hat{\beta} - \beta_0 = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{r-1} \\ \sqrt{1 - \|\hat{\beta}^{(r)}\|^2} \\ \hat{\beta}_{r+1} \\ \vdots \\ \hat{\beta}_p \end{pmatrix} - \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0(r-1)} \\ \sqrt{1 - \|\beta_0^{(r)}\|^2} \\ \beta_{0(r+1)} \\ \vdots \\ \beta_{0p} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 - \beta_{01} \\ \vdots \\ \hat{\beta}_{r-1} - \beta_{0(r-1)} \\ -\frac{\beta_0^{(r)T}(\hat{\beta}^{(r)} - \beta_0^{(r)})}{\sqrt{1 - \|\beta_0^{(r)}\|^2}} \\ \hat{\beta}_{r+1} - \beta_{0(r+1)} \\ \vdots \\ \hat{\beta}_p - \beta_{0p} \end{pmatrix} + O_P(n^{-1}).$$

That is, from the definition of $\mathbf{J}_{\beta_0^{(r)}}$ of (2.8)

$$\hat{\beta} - \beta_0 = \mathbf{J}_{\beta_0^{(r)}}(\hat{\beta}^{(r)} - \beta_0^{(r)}) + O_P(n^{-1}).$$

Thus, we have

$$\sqrt{n}(\hat{\beta} - \beta_0) = \mathbf{J}_{\beta_0^{(r)}} \mathbf{V}^{-1} n^{-1/2} U(\beta_0^{(r)}) + o_P(1).$$

Theorem 2 follows from this, Central Limit Theorem and Slutsky's Theorem. □

Proof of Theorem 3. Recalling the definition of $\mathbf{Q}$, we can see that $\mathbf{Q} = \mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}}\mathbf{Q}_1\mathbf{J}_{\beta_0^{(r)}}$. Define

$$\mathbf{\Pi}_0 := \mathbf{J}_{\beta_0^{(r)}}\mathbf{Q}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}}, \quad \mathbf{\Pi}_1 := \mathbf{J}_{\beta_0^{(r)}}\mathbf{V}^{-1}\mathbf{Q}\mathbf{V}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}}.$$

We now prove that $\mathbf{\Pi}_0$ is a generalized inverse of $\mathbf{Q}_1$. To this end, we need to prove that $\mathbf{\Pi}_0\mathbf{Q}_1\mathbf{\Pi}_0 = \mathbf{\Pi}_0$ and $\mathbf{Q}_1\mathbf{\Pi}_0\mathbf{Q}_1 = \mathbf{Q}_1$. Note that

$$\mathbf{\Pi}_0\mathbf{Q}_1\mathbf{\Pi}_0 = \mathbf{J}_{\beta_0^{(r)}}\mathbf{Q}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}}\mathbf{Q}_1\mathbf{J}_{\beta_0^{(r)}}\mathbf{Q}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}} = \mathbf{J}_{\beta_0^{(r)}}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}} = \mathbf{\Pi}_0.$$

We now prove $\mathbf{Q}_1\mathbf{\Pi}_0\mathbf{Q}_1 = \mathbf{Q}_1$. First, by $\mathbf{QR}$ decomposition (see, e.g. Gentle 1998, Section 3.2.2, pages 95-97 for more details) for β_0, we can find its orthogonal complement such that $\mathbf{B} = (b_1, \beta_0)$ is an orthogonal matrix, and $\beta_0 = \mathbf{B}\begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix}$. Thus, $\mathbf{J}_{\beta_0^{(r)}} = \mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{J}_{\beta_0^{(r)}} =: \mathbf{B}\mathbf{R}$ where $\mathbf{R} = \begin{pmatrix}\mathbf{R}_1\\ \mathbf{R}_2\end{pmatrix}$ with $\mathbf{R}_1$ being a $(p-1)\times(p-1)$ nonsigular matrix. Further, note that

$$\begin{aligned}
\mathbf{Q} &= \mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}}\mathbf{Q}_1\mathbf{J}_{\beta_0^{(r)}} = \mathbf{R}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{Q}_1\mathbf{B}\mathbf{R}\\
&= \mathbf{R}^{\mathrm{T}}\begin{pmatrix} b_1^{\mathrm{T}}\mathbf{Q}_1 b_1 & 0\\ 0 & 0\end{pmatrix}\mathbf{R}\\
&= \mathbf{R}_1^{T} b_1^{\mathrm{T}}\mathbf{Q}_1 b_1 \mathbf{R}_1.
\end{aligned}$$

To prove the result, we rewrite $\mathbf{Q}_1$ in another form. Define $\mathbf{S} = \mathbf{B}\begin{pmatrix}\mathbf{R}_1 & 0\\ 0 & 1\end{pmatrix}$. $\mathbf{S}$ is a nonsingular matrix. Then

$$\begin{aligned}
\mathbf{Q}_1 &= (\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}_1\mathbf{S}\mathbf{S}^{-1} = (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix}\mathbf{R}_1^{\mathrm{T}} & 0\\ 0 & 1\end{pmatrix}\mathbf{B}^{\mathrm{T}}\mathbf{Q}_1\mathbf{B}\begin{pmatrix}\mathbf{R}_1 & 0\\ 0 & 1\end{pmatrix}\mathbf{S}^{-1}\\
&= (\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}_1\mathbf{S}\mathbf{S}^{-1} = (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix}\mathbf{R}_1^{\mathrm{T}} & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix} b_1^{\mathrm{T}}\mathbf{Q}_1 b_1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{R}_1 & 0\\ 0 & 1\end{pmatrix}\mathbf{S}^{-1}\\
&= (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix}\mathbf{Q} & 0\\ 0 & 0\end{pmatrix}\mathbf{S}^{-1}.
\end{aligned}$$

We now prove that $\mathbf{Q}_1\mathbf{\Pi}_0\mathbf{Q}_1 = \mathbf{Q}_1$ that is of the above form. From the above and noting that $\mathbf{S}^{-1} = \begin{pmatrix} \mathbf{R}_1^{-1} & 0 \\ 0 & 1 \end{pmatrix}\mathbf{B}^{\mathrm{T}}$ and $(\mathbf{S}^{\mathrm{T}})^{-1} = \mathbf{B}\begin{pmatrix} (\mathbf{R}_1^{\mathrm{T}})^{-1} & 0 \\ 0 & 1 \end{pmatrix}$, we have

$$
\begin{aligned}
& \mathbf{Q}_1\mathbf{\Pi}_0\mathbf{Q}_1 \\
= \; & (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{S}^{-1}\mathbf{B}\mathbf{R}Q^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{S}^{-1} \\
= \; & (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{R}_1^{-1} & 0 \\ 0 & 1 \end{pmatrix}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\begin{pmatrix} (\mathbf{R}_1^{\mathrm{T}})^{-1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{S}^{-1} \\
= \; & (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{1} \\ \mathbf{R}_2 \end{pmatrix}\mathbf{Q}^{-1}\begin{pmatrix} \mathbf{1} & \mathbf{R}_2 \end{pmatrix}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{S}^{-1} \\
= \; & (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} \\ 0 \end{pmatrix}\mathbf{Q}^{-1}\begin{pmatrix} \mathbf{Q} & 0 \end{pmatrix}\mathbf{S}^{-1} = (\mathbf{S}^{\mathrm{T}})^{-1}\begin{pmatrix} \mathbf{Q} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{S}^{-1} = \mathbf{Q}_1.
\end{aligned}
$$

Thus, $\mathbf{\Pi}_0$ is one of the solutions of $\mathbf{Q}_1^{-}$. To prove that the asymptotic variance-covariance matrix $\sigma^2\mathbf{\Pi}_1$ of our estimator is smaller than the corresponding matrix $\sigma^2\mathbf{\Pi}_0$ given in Härdle et al. (1993), we only need to show that $\mathbf{\Pi}_0 - \mathbf{\Pi}_1$ is a positive semi-definite matrix, that is, $\mathbf{\Pi}_0 > \mathbf{\Pi}_1$. Recall that $\mathbf{V} = \mathbf{J}^T_{\beta_0^{(r)}} E\left\{g'(X^{\mathrm{T}}\beta_0)^2 XX^{\mathrm{T}}\right\}\mathbf{J}_{\beta_0^{(r)}}$. Note that both $\mathbf{Q}$ and $\mathbf{V}$ are positive definite matrices and obviously $\mathbf{V} \geq \mathbf{Q}$. Thus, $\mathbf{Q}^{-1} \geq \mathbf{V}^{-1}$, and then $\mathbf{V}^{-1} \geq \mathbf{V}^{-1}\mathbf{Q}\mathbf{V}^{-1}$. From these two inequalities, it is easy to see that

$$
\mathbf{\Pi}_0 \geq \mathbf{J}_{\beta_0^{(r)}}\mathbf{V}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}} \geq \mathbf{J}_{\beta_0^{(r)}}\mathbf{V}^{-1}\mathbf{Q}\mathbf{V}^{-1}\mathbf{J}^{\mathrm{T}}_{\beta_0^{(r)}} = \mathbf{\Pi}_1.
$$

The proof is now complete. □

Proof of Corollary 1. Let $\bullet$ denote the inner product of two vectors. Theorem 2 implies $\|\hat{\beta} - \beta_0\| = O_P(n^{-1/2})$ and

$$
\begin{aligned}
|\cos(\hat{\beta}, \beta_0) - 1| &= |(\hat{\beta} - \beta_0) \bullet \beta_0/\|\hat{\beta}\| + (\|\beta_0\| - \|\hat{\beta}\|)/\|\hat{\beta}\|\,| \\
&\leq 3\|\hat{\beta} - \beta_0\|/\|\hat{\beta}\| = O_P(n^{-1/2}).
\end{aligned}
$$

This completes the proof of Corollary 1. □

Proof of Theorem 4. Denote $\theta_0 = (\theta_{01}, \ldots, \theta_{0q})^{\mathrm{T}}$, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)^{\mathrm{T}}$. Theorem 1 and

Lemma A.4 in the Appendix yield

$$\begin{aligned}
&\sup_{(x,\beta)\in\mathcal{A}_n} |\hat{g}^*(x^{\mathrm{T}}\beta) - g(x^{\mathrm{T}}\beta_0)| \\
&\quad \le \sum_{s=1}^{q} \sup_{(x,\beta)\in\mathcal{A}_n} |\hat{g}_{2s}(x^{\mathrm{T}}\beta;\beta) - g_{2s}(x^{\mathrm{T}}\beta_0)||\hat{\theta}_s - \theta_{0s}| \\
&\qquad + \sup_{(x,\beta)\in\mathcal{A}_n} |\hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0) - g(x^{\mathrm{T}}\beta_0)| \\
&\qquad + \sum_{s=1}^{q} \sup_{x\in A} |g_{2s}(x^{\mathrm{T}}\beta_0)||\hat{\theta}_s - \theta_{0s}| = O_P((nh/\log n)^{-1/2}),
\end{aligned}$$

and hence Theorem 4 follows. □

Proof of Theorem 5. Decomposing $\hat{\sigma}^2$ into several parts, we have

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 + \frac{1}{n}\sum_{i=1}^{n} [Z_i^{\mathrm{T}}(\theta_0 - \hat{\theta}) + g(X_i^{\mathrm{T}}\beta_0) - \hat{g}(X_i^{\mathrm{T}}\hat{\beta};\hat{\beta},\theta_0)]^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} e_i Z_i^{\mathrm{T}}(\theta_0 - \hat{\theta}) + \frac{2}{n}\sum_{i=1}^{n} e_i[g(X_i^{\mathrm{T}}\beta_0) - \hat{g}(X_i^{\mathrm{T}}\hat{\beta};\hat{\beta},\theta_0)] \\
&=: I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Note that $\sqrt{n}\|\hat{\theta} - \theta_0\| = O_P(1)$ and using (A.8) of Lemma A.4, we have

$$\begin{aligned}
\sqrt{n}|I_2| &\le \frac{1}{n}\sum_{i=1}^{n} \|Z_i\|^2 \sqrt{n}\|\hat{\theta} - \theta_0\|^2 \\
&\quad + \sqrt{n} \sup_{(x,\beta)\in\mathcal{A}_n} |g(x^{\mathrm{T}}\beta_0) - \hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0)|^2 \\
&= O_P(n^{-1/2}) + O_P((nh^2/\log^2 n)^{-1/2}) \xrightarrow{P} 0.
\end{aligned}$$

Since $Ee_i = 0$, we obtain $\sqrt{n}I_3 \xrightarrow{P} 0$. Similarly to the proof of (A.17) in the Appendix, we also have $\sqrt{n}I_4 \xrightarrow{P} 0$. This proves that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 + o_P(n^{-1/2})$. Therefore, we have

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(e_i^2 - \sigma^2) + o_P(1).$$

The proof can now be completed by employing the central limit theorem. □

APPENDIX

The following Lemmas A.1–A.7 are needed to prove Theorems 1, 2, 4, 5. Lemma A.1 gives an important probability inequality and Lemmas A.2 and A.3 provide bounds for the

moments of the relevant estimators. They are used to obtain the rates of convergence for the estimators of the nonparametric component, and are used in the proof of Lemmas A.4–A.7. Lemma A.4 presents the uniform rates of convergence in probability for the estimators $\hat{g}$, g_{2s} and $\hat{g}'$. These results are very useful for the nonparametric estimations. The proof of Lemma A.5–A.7, as well as Theorem 3 and 4, rely on Lemma A.4. To simplify the proof of Theorem 1, we divide the main steps of the proofs into Lemmas A.5 and A.6. Lemma A.5 is used to obtain the limiting variance of the estimator $\hat{\theta}$, and Lemma A.6 together with Lemma A.5 shows that the rate of convergence of the nonlinear section of e_i for $\hat{\theta} - \theta_0$ is $o_P(n^{-1/2})$. Lemma A.7 provides the main step for the proof of Theorem 2.

Lemma A.1 *Let $\xi_1(x,\beta),\ldots,\xi_n(x,\beta)$ be a sequence of random variables. Denote $f_{x,\beta}(V_i) = \xi_i(x,\beta)$ for $i = 1,\ldots,n$, where $V_1,\ldots,V_n$ be a sequence of random variables, and $f_{x,\beta}$ is a function on $\mathcal{A}_n$, where $\mathcal{A}_n = \{(x,\beta) : (x,\beta) \in A \times R^p, \|\beta - \beta_0\| \le cn^{-1/2}\}$ for a constant $c > 0$. Assume that $f_{x,\beta}$ satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} |f_{x,\beta}(V_i) - f_{x^*,\beta^*}(V_i)| \le cn^a[\|\beta - \beta^*\| + \|x - x^*\|] \tag{A.1}$$

for some constants x^*, β^*, $a > 0$ and $c > 0$. Let $\varepsilon_n > 0$ depend only on n. If

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right| > \frac{1}{2}\varepsilon_n\right\} \le \frac{1}{2}, \tag{A.2}$$

for $(x,\beta) \in \mathcal{A}_n$, then we have

$$\begin{aligned} &P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right| > \frac{1}{2}\varepsilon_n\right\} \\ &\quad\le c_1 n^{2pa}\varepsilon_n^{-2p} E\left\{\sup_{(x,\beta)\in\mathcal{A}_n} 2\exp\left(\frac{-n^2\varepsilon_n^2/128}{\sum_{i=1}^{n}\xi_i^2(x,\beta)}\right)\wedge 1\right\}, \end{aligned} \tag{A.3}$$

where $c_1 > 0$ is a constant.

Proof. Let $\{\xi_1'(x,\beta),\ldots,\xi_n'(x,\beta)\}$ be an independent version of $\{\xi_1(x,\beta),\ldots,\xi_n(x,\beta)\}$. Now generate independent sign random variables $\sigma_1,\ldots,\sigma_n$ for which $P\{\sigma_i = 1\} = P\{\sigma_i = -1\} = \frac{1}{2}$, and $\{\sigma_i, 1 \le i \le n\}$ independent of $\{\xi_i(x,\beta), \xi_i'(x,\beta), 1 \le i \le n\}$. By symmetry, $\sigma_i(\xi_i - \xi_i')$ has the same distribution as $(\xi_i - \xi_i')$. The symmetrization Lemma in Pollard

(1984) implies

$$
\begin{aligned}
&P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right|>\varepsilon_n\right\}\\
&\quad\le 2P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}[\xi_i(x,\beta)-\xi_i'(x,\beta)]\right|>\frac{1}{2}\varepsilon_n\right\}\\
&\quad= 2P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i[\xi_i(x,\beta)-\xi_i'(x,\beta)]\right|>\frac{1}{2}\varepsilon_n\right\}\\
&\quad\le 4P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\xi_i(x,\beta)\right|>\frac{1}{4}\varepsilon_n\right\}. \qquad \text{(A.4)}
\end{aligned}
$$

Let P_n be the empirical measure that puts equal mass $\frac{1}{n}$ at each of the n observations $V_1,\ldots,V_n$. Let $\mathcal{F}=\{f_{x,\beta}(\cdot):\|x\|\le C,\|\beta\|\le B\}$ be a class of functions indexed by x and β consisting of $f_{x,\beta}(V_i)=\xi_i(x,\beta)$. Denote $\mathcal{V}=(V_1,\ldots,V_n)$. Given $\mathcal{V}$, choose function $f_1^\circ,\ldots,f_m^\circ$, each in $\mathcal{F}$, such that

$$
\min_{j\in\{1,\ldots,m\}}\frac{1}{n}\sum_{i=1}^{n}|f_{x,\beta}(V_i)-f_j^\circ(V_i)|<\varepsilon_n \qquad \text{(A.5)}
$$

for each $f_{x,\beta}$ in $\mathcal{F}$. Let $N(\varepsilon_n,P_n,\mathcal{F})$ be the minimum m for all sets that satisfies (A.5). Denote $f_{x,\beta}^*$ for the f_j° at which the minimum is achieved, we then have

$$
\begin{aligned}
&P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\xi_i(x,\beta)\right|>\frac{1}{4}\varepsilon_n\Big|\mathcal{V}\right\}\\
&\quad= P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{x,\beta}(V_i)\right|>\frac{1}{4}\varepsilon_n|\mathcal{V}\right\}\\
&\quad\le P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{x,\beta}^*(V_i)\right|>\frac{1}{8}\varepsilon_n|\mathcal{V}\right\}\\
&\quad\le N(\varepsilon_n,P_n,\mathcal{F})\max_{j\in\{1,\ldots,N\}}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_j^\circ(V_i)\right|>\frac{1}{8}\varepsilon_n|\mathcal{V}\right\}. \qquad \text{(A.6)}
\end{aligned}
$$

Now we need to determine the order of $N(\varepsilon_n,P_n,\mathcal{F})$. For each set satisfying (A.5), each f_j° has a pair (x_j,β_j) such that $f_j^\circ(v)=f_{x_j,\beta_j}(v)$. Then for all $(x,\beta)\in\mathcal{A}_n$, we have from (A.1) that

$$
\frac{1}{n}\sum_{i=1}^{n}|f_{x,\beta}(V_i)-f_{x_j,\beta_j}(V_i)|\le cn^a(\|\beta-\beta_j\|+\|x-x_j\|).
$$

Next, we want to bound the right-hand side of the above formula by ε_n. Thus for each $(x,\beta)\in\mathcal{A}_n$, we need a pair (x_j,β_j) within radius $r_n=O(n^{-a}\varepsilon_n)$ of (x,β). Therefore, the

number N needed to satisfy (A.5) is bounded by $r_n^{-p} r_n^{-p} = cn^{2pa}\varepsilon_n^{-2p}$, i.e.

$$N(\varepsilon_n, P_n, \mathcal{F}) \le cn^{2pa}\varepsilon_n^{-2p}. \tag{A.7}$$

Now conditioning on $\mathcal{V}$, $\sigma_i f_j^\circ(V_i)$ is bounded. Hoeffding's inequality [(Hoeffding (1963)] yields

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f_j^\circ(V_i)\right| > \frac{1}{8}\varepsilon_n | \mathcal{V}\right\} \le 2\exp\left(\frac{-2n(\varepsilon_n/8)^2}{\sum_{i=1}^n 4f^2_{x_j,\beta_j}(V_i)}\right) \wedge 1.$$

This together with (A.4), (A.6) and (A.7) proves (A.3). □

Lemma A.2 *Suppose that conditions* C1, C2 and C3(i) *hold. If* $h = cn^{-a}$ *for any* $0 < a < 1/2$ *and some constants* $c > 0$*, then, for* $i = 1, \cdots, n$*, we have*

$$E\left[g(X_i^{\mathrm{T}}\beta_0) - \sum_{j=1}^n W_{nj}(X_i^{\mathrm{T}}\beta_0; \beta_0) g(X_j^{\mathrm{T}}\beta_0)\right]^2 = O(h^4),$$

$$E\left[g(x^{\mathrm{T}}\beta) - \sum_{j=1}^n W_{nj}(x^{\mathrm{T}}\beta; \beta) g(X_j^{\mathrm{T}}\beta)\right]^2 = O(h^4),$$

$$E\left[g'(X_i^{\mathrm{T}}\beta_0) - \sum_{j=1}^n \widetilde{W}_{nj}(X_i^{\mathrm{T}}\beta_0; \beta_0) g(X_j^{\mathrm{T}}\beta_0)\right]^2 = O(h_1^2)$$

and

$$E\left[\sum_{j=1}^n W_{ni}(X_j^{\mathrm{T}}\beta_0; \beta_0)\varphi(X_j^{\mathrm{T}}\beta_0) - \varphi(X_i^{\mathrm{T}}\beta_0)\right]^2 = O(\sqrt{h}\ ),$$

where $\varphi(t) = g'(t) g_{3s}(t)$ and g_{3s} is the sth component of $g_3(t) = E(X | X^{\mathrm{T}}\beta_0 = t)$.

Proof. See Lemma 1 of Zhu and Xue (2006). □

Lemma A.3 *Under the assumptions of Lemma A.2, we have*

$$\begin{cases} E\{W_{ni}^2(X_i^{\mathrm{T}}\beta_0; \beta_0)\} = O((nh)^{-2}), \\ E\left\{\sum_{j=1, j\neq i}^n W_{nj}^2(X_i^{\mathrm{T}}\beta_0; \beta_0)\right\} = O((nh)^{-1}), \end{cases}$$

$$E\left\{\sum_{j=1}^n W_{nj}^2(x^{\mathrm{T}}\beta; \beta)\right\} = O((nh)^{-1})$$

and

$$\begin{cases} E\{\widetilde{W}_{ni}^2(X_i^{\mathrm{T}}\beta_0; \beta_0)\} = O((nh_1)^{-2} + (n^3h_1^5)^{-1}), \\ E\left\{\sum_{j=1, j\neq i}^n \widetilde{W}_{nj}^2(X_i^{\mathrm{T}}\beta_0; \beta_0)\right\} = O((nh_1^3)^{-1}). \end{cases}$$

Proof. See Lemma 2 of Zhu and Xue (2006). □

Lemma A.4 *Suppose that conditions* C1–C4 *and* C5(i) *hold. We then have*

$$\sup_{(x,\beta)\in\mathcal{A}_n} |g(x^{\mathrm{T}}\beta_0) - \hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0)| = O_P((nh/\log n)^{-1/2}) \tag{A.8}$$

and

$$\sup_{(x,\beta)\in\mathcal{A}_n} |g_{2s}(x^{\mathrm{T}}\beta_0) - \hat{g}_{2s}(x^{\mathrm{T}}\beta;\beta)| = O_P((nh/\log n)^{-1/2}). \tag{A.9}$$

If in addition, C5(ii) also holds, then we have

$$\sup_{(x,\beta)\in\mathcal{A}_n} |g'(x^{\mathrm{T}}\beta_0) - \hat{g}'(x^{\mathrm{T}}\beta;\beta,\theta_0)| = O_P((nh_1^3/\log n)^{-1/2}), \tag{A.10}$$

where $\mathcal{A}_n = \{(x,\beta) : (x,\beta) \in A \times R^p, \|\beta - \beta_0\| \le cn^{-1/2}\}$ for a constant $c > 0$.

Proof. We only prove (A.8), the proofs for (A.9) and (A.10) are similar. Write $\tilde{g}(X_i, e_i) = g(x^{\mathrm{T}}\beta_0) - g(X_i^{\mathrm{T}}\beta_0) - e_i$, $i = 1, \ldots, n$. We have

$$g(x^{\mathrm{T}}\beta_0) - \hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0) = \sum_{i=1}^{n} W_{ni}(x^{\mathrm{T}}\beta;\beta)\tilde{g}(X_i, e_i). \tag{A.11}$$

Let $\xi_i(x,\beta) = n(nh/\log n)^{1/2} W_{ni}(x^{\mathrm{T}}\beta;\beta)\tilde{g}(X_i, e_i)$, $f_{x,\beta}(V_i) = \xi_i(x,\beta)$, $V_i = (X_i, e_i)$, $i = 1, \ldots, n$. Using lemma A.1, we have to verify (A.1) and (A.2). A simple calculation yields (A.1), so we now verify (A.2). By lemmas A.2 and A.3, and noting that $\sup_{(x,\beta)\in\mathcal{A}_n} |g(x^{\mathrm{T}}\beta) - g(x^{\mathrm{T}}\beta_0)| = O(n^{-1/2})$, we have

$$\begin{aligned} E[g(x^{\mathrm{T}}\beta_0) - \hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0)]^2 &= E\left[\sum_{i=1}^{n} W_{ni}(x^{\mathrm{T}}\beta;\beta)\tilde{g}(X_i, e_i)\right]^2 \\ &\le cE\left[g(x^{\mathrm{T}}\beta) - \sum_{i=1}^{n} W_{ni}(x^{\mathrm{T}}\beta;\beta) g(X_i^{\mathrm{T}}\beta)\right]^2 \\ &\quad + cE\left\{\sum_{i=1}^{n} W_{ni}^2(x^{\mathrm{T}}\beta;\beta)\right\} + O(n^{-1}) \\ &\le ch^4 + c(nh)^{-1}. \end{aligned} \tag{A.12}$$

Given a $M > 0$, by Chevbychev's inequality and (A.12), we have

$$\begin{aligned} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right| > \frac{1}{2}M\right\} &\le 4M^{-2}E\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right]^2 \\ &\le 4M^{-2}nh(\log n)^{-1}E\left[\sum_{i=1}^{n} W_{ni}(x^{\mathrm{T}}\beta;\beta)\tilde{g}(X_i, e_i)\right]^2 \\ &\le cM^{-2}(cnh^5 + c(\log n)^{-1}). \end{aligned} \tag{A.13}$$

Therefore, from C5(i), we can choose M large enough so that the right hand side of (A.13) is less than or equal to $\frac{1}{2}$. Hence, (A.2) is satisfied. We now can use (A.3) of Lemma A.1 to get (A.8). By Lemma A.3, we obtain

$$n^{-2}\sum_{i=1}^{n} E\xi_i^2(x,\beta) = nh(\log n)^{-1}\sum_{i=1}^{n} E\left[W_{ni}(x^{\mathrm{T}}\beta;\beta)\tilde{g}(X_i,e_i)\right]^2$$
$$\leq cnh(\log n)^{-1}\sum_{i=1}^{n} EW_{ni}^2(x^{\mathrm{T}}\beta;\beta) \leq c(\log n)^{-1}.$$

This implies that $n^{-2}\sum_{i=1}^{n}\xi_i^2(x,\beta) = O_P((\log n)^{-1})$. Hence, from Lemma A.1 we have

$$P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(x,\beta)\right| > \frac{1}{2}M\right\} \leq cn^{2pa}M^{-2p}\exp\left(-cM^2\log n\right).$$

The right-hand side of the above formula tends to zero when M is large enough. Therefore, (A.8) follows. □

Lemma A.5 *Under the assumptions of Theorem 1, we have*

$$n^{-1}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}} \xrightarrow{P} \mathbf{\Sigma}.$$

where $\mathbf{\Sigma}$ *is defined in condition* C6.

Proof. Noting that $\tilde{\mathbf{Z}} = (\mathbf{I}-\mathbf{S})\mathbf{Z}$, the (i,s) element of $\tilde{\mathbf{Z}}$ is

$$\tilde{Z}_{is} = [Z_{is} - g_{2s}(X_i^{\mathrm{T}}\beta_0)] + [g_{2s}(X_i^{\mathrm{T}}\beta_0) - \hat{g}_{2s}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)].$$

The (s,t) element of $\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}}$ is

$$\begin{aligned}\sum_{i=1}^{n}\tilde{Z}_{is}\tilde{Z}_{it} =& \sum_{i=1}^{n}[Z_{is} - g_{2s}(X_i^{\mathrm{T}}\beta_0)][Z_{it} - g_{2t}(X_i^{\mathrm{T}}\beta_0)]\\ &+\sum_{i=1}^{n}[Z_{is} - g_{2s}(X_i^{\mathrm{T}}\beta_0)][g_{2t}(X_i^{\mathrm{T}}\beta_0) - \hat{g}_{2t}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)]\\ &+\sum_{i=1}^{n}[Z_{it} - g_{2t}(X_i^{\mathrm{T}}\beta_0)][g_{2s}(X_i^{\mathrm{T}}\beta_0) - \hat{g}_{2s}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)]\\ &+\sum_{i=1}^{n}[g_{2s}(X_i^{\mathrm{T}}\beta_0) - \hat{g}_{2s}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)][g_{2t}(X_i^{\mathrm{T}}\beta_0) - \hat{g}_{2t}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)]\\ =:& I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{A.14}$$

By the law of large numbers, we have

$$n^{-1}I_1 \xrightarrow{P} E\{[Z_{1s} - E(Z_{1s}|X_1^{\mathrm{T}}\beta_0)][Z_{1t} - E(Z_{1t}|X_1^{\mathrm{T}}\beta_0)]\} =: \Sigma_{st}, \tag{A.15}$$

where Σ_{st} is the (s,t) element of $\mathbf{\Sigma}$. Noting that

$$\frac{1}{n}\sum_{i=1}^{n}|Z_{is}-g_{2s}(X_i^{\mathrm{T}}\beta_0)| \xrightarrow{P} E|Z_{1s}-g_{2s}(X_1^{\mathrm{T}}\beta_0)|<\infty,$$

this together with (A.9) of Lemma A.4 proves that

$$n^{-1}I_2 \le O_P(1)\sup_{(x,\beta)\in\mathcal{A}_n}|g_{2t}(x^{\mathrm{T}}\beta_0)-\hat{g}_{2t}(x^{\mathrm{T}}\beta;\beta)| \xrightarrow{P} 0.$$

Similarly, we can prove $n^{-1}I_3 \xrightarrow{P} 0$ and $n^{-1}I_4 \xrightarrow{P} 0$. This together with (A.14) and (A.15) proves Lemma A.5. □

Lemma A.6 *Under the assumptions of Theorem 1, we have*

$$n^{-1/2}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{G} := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{Z}_i[g(X_i^{\mathrm{T}}\beta_0)-\hat{g}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0,\theta_0)] \xrightarrow{P} 0.$$

Proof. The sth component of $\tilde{Z}^{\mathrm{T}}\tilde{G}$ is

$$\begin{aligned}
&\sum_{i=1}^{n}\tilde{Z}_{is}[g(X_i^{\mathrm{T}}\beta_0)-\hat{g}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0,\theta_0)]\\
&\quad=\sum_{i=1}^{n}[Z_{is}-g_{2s}(X_i^{\mathrm{T}}\beta_0)][g(X_i^{\mathrm{T}}\beta_0)-\hat{g}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0,\theta_0)]\\
&\qquad+\sum_{i=1}^{n}[g_{2s}(X_i^{\mathrm{T}}\beta_0)-\hat{g}_{2s}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0)][g(X_i^{\mathrm{T}}\beta_0)-\hat{g}(X_i^{\mathrm{T}}\hat{\beta}_0;\hat{\beta}_0,\theta_0)]\\
&\quad=: J_1+J_2. \qquad\qquad (A.16)
\end{aligned}$$

For J_2, from (A.8) and (A.9) of Lemma A.4 we have

$$\begin{aligned}
n^{-1/2}|J_2| \le&\sqrt{n}\sup_{(x,\beta)\in\mathcal{A}_n}|g_{2s}(x^{\mathrm{T}}\beta_0)-\hat{g}_{2s}(x^{\mathrm{T}}\beta;\beta)|\\
&\times\sup_{(x,\beta)\in\mathcal{A}_n}|g(x^{\mathrm{T}}\beta_0)-\hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0)| = O_P((nh^2/\log^2 n)^{-1}).
\end{aligned}$$

Noting that $nh^2/\log^2 n \to \infty$, we obtain $n^{-1/2}J_2 \xrightarrow{P} 0$. It remains to prove that

$$n^{-1/2}J_1 \xrightarrow{P} 0, \qquad (A.17)$$

as this together with (A.16) implies Lemma A.6. To prove (A.17) , we only need to show that

$$\sup_{\beta\in\mathcal{B}_n'}\left|\frac{1}{n}\sum_{i=1}^{n}\sqrt{n}[Z_{is}-g_{2s}(X_i^{\mathrm{T}}\beta_0)][g(X_i^{\mathrm{T}}\beta_0)-\hat{g}(X_i^{\mathrm{T}}\beta,\beta)]\right| \xrightarrow{P} 0, \qquad (A.18)$$

where $\mathcal{B}'_n = \{\beta : \|\beta - \beta_0\| \le cn^{-1/2}\}$ for a constant $c > 0$. Toward this goal, we note that Lemma A.1 can be used when the variable x is removed. Let

$$\xi_i(\beta) = \sqrt{n}[Z_{is} - g_{2s}(X_i^{\mathrm{T}}\beta_0)][g(X_i^{\mathrm{T}}\beta_0) - \hat{g}(X_i^{\mathrm{T}}\beta, \beta, \theta_0)],$$
$$f_\beta(V_i) = \xi_i(\beta), \qquad V_i = (X_i, Z_{is}, e_i), \quad i = 1, \ldots, n.$$

We now verify that (A.1) and (A.2) are satisfied. By the condition C3(ii) on the kernel function, we calculate that

$$\frac{1}{n}\sum_{i=1}^{n} |f_\beta(V_i) - f_{\beta^*}(V_i)| \le cn^{5/2}h^{-2}\|\beta - \beta^*\| = cn^a\|\beta - \beta^*|$$

where $a = \frac{5}{2} + 2\lambda(\frac{1}{5} \le \lambda < \frac{1}{2})$. Hence, (A.1) is satisfied.

We next verify that (A.2) is satisfied. Denote $\zeta_i = Z_{is} - g_{2s}(X_i^{\mathrm{T}}\beta_0)$. From condition C4, Lemmas A.2 and A3, we have

$$\begin{aligned}
&E\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)\right]^2 \\
&\quad\le 2n^{-1}\sum_{i=1}^{n} E\left\{\left[g(X_i^{\mathrm{T}}\beta_0) - \sum_{j=1}^{n} W_{nj}(X_i^{\mathrm{T}}\beta;\beta)g(X_j^{\mathrm{T}}\beta_0)\right]^2 E(\zeta_i^2|X_i^{\mathrm{T}}\beta_0)\right\} \\
&\qquad + 2n^{-1}\sum_i\sum_j\sum_k\sum_l E[W_{nj}(X_i^{\mathrm{T}}\beta;\beta)W_{nl}(X_k^{\mathrm{T}}\beta;\beta)\zeta_i\zeta_k e_j e_l] \\
&\quad\le ch^4 + cn^{-1} + cn^{-1}\left\{\sum_{i=1}^{n} EW_{ni}^2(X_i^{\mathrm{T}}\beta;\beta) + \sum_{i\ne j} EW_{nj}^2(X_i^{\mathrm{T}}\beta;\beta)\right\} \\
&\quad\le ch^4 + cn^{-1} + c(nh)^{-1} \longrightarrow 0.
\end{aligned}$$

Hence, we can obtain

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)\right| > \frac{1}{2}\varepsilon\right\} \le ch^4 + cn^{-1} + c(nh)^{-1} < \frac{1}{2}$$

when n large enough. Therefore, (A.2) is satisfied. By (A.8) of Lemma A.4, we have

$$\frac{1}{n^2}\sum_{i=1}^{n}\xi_i^2(\beta) = O_P(1)\sup_{(x,\beta)\in\mathcal{A}_n}[g(x^{\mathrm{T}}\beta_0) - \hat{g}(x^{\mathrm{T}}\beta;\beta,\theta_0)]^2 = O_P(nh/\log n)^{-1}).$$

By using Lemma A.1, we obtain

$$P\left\{\sup_{(x,\beta)\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)\right| > \frac{1}{2}\varepsilon\right\} \le cn^{2pa}\varepsilon^{-2p}\exp(-cnh/\log n) \longrightarrow 0.$$

by $nh/\log n \to \infty$. This proves (A.18) and thus completes the proof of Lemma A.6. □

Lemma A.7. Suppose that conditions C1–C6 are satisfied, then we have

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R(\beta^{(r)}) - U(\beta_0^{(r)}) + n\mathbf{V}(\beta^{(r)} - \beta_0^{(r)})\| = o_P(\sqrt{n}),$$

where $\mathcal{B}_n = \{\beta^{(r)} : \|\beta^{(r)} - \beta_0^{(r)}\| \le Cn^{-1/2}\}$ for a constant $C > 0$, $\mathbf{V}$ is defined in condition C6,

$$R(\beta^{(r)}) = \sum_{i=1}^{n} [Y_i - Z_i^{\mathrm{T}}\theta_0 - \hat{g}(X_i^{\mathrm{T}}\beta; \beta, \theta_0)]\hat{g}'(X_i^{\mathrm{T}}\beta; \beta, \theta_0)\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i, \tag{A.19}$$

and

$$U(\beta_0^{(r)}) = \sum_{i=1}^{n} e_i g'(X_i^{\mathrm{T}}\beta_0)\mathbf{J}_{\beta_0^{(r)}}^{\mathrm{T}}[X_i - E(X_i|X_i^{\mathrm{T}}\beta_0)].$$

Proof. Separating $R(\beta^{(r)})$, we have

$$\begin{aligned}
R(\beta^{(r)}) =& \sum_{i=1}^{n} e_i g'(X_i^{\mathrm{T}}\beta_0)\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}}[X_i - E(X_i|X_i^{\mathrm{T}}\beta_0)] \\
&+ \sum_{i=1}^{n} e_i[\hat{g}'(X_i^{\mathrm{T}}\beta; \beta, \theta_0) - g'(X_i^{\mathrm{T}}\beta_0)]\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i \\
&- \sum_{i=1}^{n} g'(X_i^{\mathrm{T}}\beta_0)\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i\{\hat{g}(X_i^{\mathrm{T}}\beta; \beta, \theta_0) - \hat{g}(X_i^{\mathrm{T}}\beta_0; \beta_0, \theta_0)\} \\
&- \sum_{i=1}^{n} g'(X_i^{\mathrm{T}}\beta_0)\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}}\{X_i[\hat{g}(X_i^{\mathrm{T}}\beta_0; \beta_0, \theta_0) - g(X_i^{\mathrm{T}}\beta_0)] - e_i g_3(X_i^{\mathrm{T}}\beta_0)\} \\
&- \sum_{i=1}^{n} [\hat{g}(X_i^{\mathrm{T}}\beta; \beta, \theta_0) - g(X_i^{\mathrm{T}}\beta_0)][\hat{g}'(X_i^{\mathrm{T}}\beta; \beta, \theta_0) - g'(X_i^{\mathrm{T}}\beta_0)]\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i \\
=:& R_1(\beta^{(r)}) + R_2(\beta^{(r)}) - R_3(\beta^{(r)}) - R_4(\beta^{(r)}) - R_5(\beta^{(r)}).
\end{aligned} \tag{A.20}$$

Noting that $\mathbf{J}_{\beta^{(r)}} - \mathbf{J}_{\beta_0^{(r)}} = O_P(n^{-1/2})$ for all $\beta^{(r)} \in \mathcal{B}_n$, we have

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R_1(\beta^{(r)}) - U(\beta_0^{(r)})\| = o_P(\sqrt{n}\,). \tag{A.21}$$

Since $\|\beta^{(r)} - \beta_0^{(r)}\| \le Cn^{-1/2}$ implies $\|\beta - \beta_0\| \le Cn^{-1/2}$ for all $\beta^{(r)} \in \mathcal{B}_n$, similar to the proof of (A.17) we can show that

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R_2(\beta^{(r)})\| = o_P(\sqrt{n}\,). \tag{A.22}$$

For $R_3(\beta^{(r)})$, by a Taylor expansion of $\beta^{(r)} - \beta_0^{(r)}$ with a suitable mean $\bar{\beta}^{(r)} \in \mathcal{B}_n$ and $\bar{\beta} = \bar{\beta}(\bar{\beta}^{(r)})$, we get

$$\begin{aligned}
R_3(\beta^{(r)}) =& \sum_{i=1}^{n} g'(X_i^{\mathrm{T}}\beta_0)\hat{g}'(X_i^{\mathrm{T}}\bar{\beta}; \bar{\beta}, \theta_0)\mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i X_i^{\mathrm{T}} \mathbf{J}_{\bar{\beta}^{(r)}}(\beta^{(r)} - \beta_0^{(r)}) \\
=& \sum_{i=1}^{n} g'(X_i^{\mathrm{T}}\beta_0)[\hat{g}'(X_i^{\mathrm{T}}\bar{\beta}; \bar{\beta}, \theta_0) - g'(X_i^{\mathrm{T}}\beta_0)] \\
& \times \mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i X_i^{\mathrm{T}} \mathbf{J}_{\bar{\beta}^{(r)}}(\beta^{(r)} - \beta_0^{(r)}) \\
& + \sum_{i=1}^{n} g'(X_i^{\mathrm{T}}\beta_0)^2 \mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} X_i X_i^{\mathrm{T}} \mathbf{J}_{\bar{\beta}^{(r)}}(\beta^{(r)} - \beta_0^{(r)}) \\
=:& R_{31}(\beta^{(r)}, \bar{\beta}^{(r)}) + R_{32}(\beta^{(r)}, \bar{\beta}^{(r)}).
\end{aligned}$$

By (A.10) of Lemma A.4 and the law of large numbers, we obtain that

$$\sup_{\beta^{(r)}, \bar{\beta}^{(r)} \in \mathcal{B}_n} \|R_{31}(\beta^{(r)}, \bar{\beta}^{(r)})\| = o_P(\sqrt{n}\,)$$

and

$$\sup_{\beta^{(r)}, \bar{\beta}^{(r)} \in \mathcal{B}_n} \|R_{32}(\beta^{(r)}, \bar{\beta}^{(r)}) - n\mathbf{V}(\beta^{(r)} - \beta_0^{(r)})\| = o_P(\sqrt{n}\,).$$

Therefore, we have

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R_3(\beta^{(r)}) - n\mathbf{V}(\beta^{(r)} - \beta_0^{(r)})\| = o_P(\sqrt{n}\,). \tag{A.23}$$

We now consider $R_4(\beta^{(r)})$. Write $R_4(\beta^{(r)}) = \mathbf{J}_{\beta^{(r)}}^{\mathrm{T}} R_4^*(\beta^{(r)})$. Let $R_{4,s}^*$ denote the sth component of $R_4^*(\beta^{(r)})$. First, from Lemma A.2 and A.3 we have

$$\begin{aligned}
n^{-1}E(R_{4,s}^{*2}) \le & cn^{-1}\sum_{i=1}^{n} E\left\{\sum_{j=1}^{n} W_{ni}(X_j^{\mathrm{T}}\beta_0; \beta_0) g'(X_j^{\mathrm{T}}\beta_0) X_{js} - g'(X_i^{\mathrm{T}}\beta_0) g_{3s}(X_i^{\mathrm{T}}\beta_0)\right\}^2 \\
& + c\sum_{i=1}^{n} E\left\{\sum_{j=1}^{n} W_{nj}(X_i^{\mathrm{T}}\beta_0; \beta_0) g(X_j^{\mathrm{T}}\beta_0) - g(X_i^{\mathrm{T}}\beta_0)\right\}^2 \\
\le & c(nh)^{-1} + c\sqrt{h} + cnh^4 \longrightarrow 0.
\end{aligned}$$

This implies

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R_4(\beta^{(r)})\| = o_P(\sqrt{n}), \tag{A.24}$$

and by Lemma A.4 we obtain

$$\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R_5(\beta^{(r)})\| = o_P(\sqrt{n}). \tag{A.25}$$

Substituting (A.21)–(A.25) into (A.20), we prove Lemma A.7. □

REFERENCES

Arnold, S. f. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley & Sons, New York.

Bhattacharya, P. K. and Zhao, P.-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* **25** 244-262.

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.

Chen, C.-H. and Li, K.-C.(1998). Can SIR be as popular as multiple linear regression. *Statistics Sinica* **8** 289–316.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–141.

Chen, H. and Shiau, J.-J. H. (1994). Data-driven efficient estimators for a Partially linear model. *Ann. Statist.* **22** 211–237.

Chiou, J. M. and Müller, H. G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* **93** 1376–1387.

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

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression1, 2 R. Dennis Cook. *Statist. Sci.* **22** 1-26.

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

Cook, R. D. and Wiseberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

Craven, P. and Wahba, G. (1979), Smoothing and noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* **31** 377-403.

Doksum, K. and Samarov, A. (1995), Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. of Statist.* **23** 1443-1473.

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

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

Gentle, J. E.(1998). *Numerical Linear Algebra for Applications in Statistics*. Berlin: Springer-Verlag.

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

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

Härdle, W., Gao, J. and Liang, H. (2000) *Partially Linear Models*. Springer, New York

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing pries and the demand for clean air *Environmental Economics and Management* **5** 81–102.

Heckman, N. (1986). Spline smoothing in a partly linear model. *J. Royal Statist. Soci.* Ser.A **48** 244–248.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061.

Li, B., Wen, S. Q. and Zhu L. X. (2008). On a projective resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **106** 1177-1186.

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

Li, K. C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98** 99–109.

Li, Y. X. and Zhu, L. X. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* **35** 41–69.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag New York Inc., New York.

Rice, J. (1986), Convergence rates for partially splined models. *Statist. Prob. Lett.* **4** 203-208.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press,

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soci.* Ser.B **50** 413–434.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481.

Stute,W. and Zhu, L. X. (2005). Nonparametric checks for single-index models. *Ann. Statist.* **33** 1048–1083.

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

Welsh, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17** 337–361. [Correction(1990) **18** 1500.]

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multi. Anal.* **97** 1162 - 1184

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

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137.

Yin, X. and Cool, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. *J. R. Statist. Soc.* B **64** 159–175.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics for Sliced Inverse Regression. *Statistica Sinica* **5** 727-736.

Zhu, L. X. and Ng, K. W. (2003) Checking the adequacy of a partial linear model. *Statist. Sinica* **13** 763-781.

Zhu L. X. and Xue L. G. (2006) Empirical likelihood confidence regions in a partially linear single-index model. *J. Roy. Statist. Soc.* ser. B, **68** 549–570.

Zhu, L., Zhu, L. X., Ferrë L., and Wang, T. (2008). Sufficient Dimension Reduction Through Discretization-Expectation Estimation. Unpublished manuscript, Hong Kong Baptist University.

TABLE 1

Simulation results for $\hat{\theta}$ with β_Z and β_0 parallel

	Resulting estimate			One-step iterated estimate		
	Bias	SD	MSE	Bias	SD	MSE
PPR	0.0058	0.0706	0.00502	0.0046	0.0701	0.00493
SIR_5	0.0095	0.0862	0.00753	0.0083	0.0869	0.00762
SIR_{10}	0.0113	0.0788	0.00634	0.0098	0.0808	0.00663
β_0 given	0.0031	0.0660	0.00436			

TABLE 2

Simulation results for $\hat{\theta}$ with β_Z and β_0 orthogonal

	Resulting estimate			One-step iterated estimate		
	Bias	SD	MSE	Bias	SD	MSE
PPR	-0.0087	0.0972	0.00952	-0.0047	0.0711	0.00508
SIR_5	-0.0115	0.1395	0.01960	-0.0072	0.0919	0.00850
SIR_{10}	-0.0102	0.1362	0.01865	-0.0083	0.0959	0.00926
β_0 given	-0.0024	0.0696	0.00485			

TABLE 3

Simulation results for the angles between $\hat{\beta}$ and β_0

	β_Z and β_0 parallel			β_Z and β_0 orthogonal		
	Mean	SD	MSE	Mean	SD	MSE
PPR	0.0148	0.0056	0.00025	0.0157	0.0066	0.00029
SIR_5	0.0467	0.0223	0.00268	0.0482	0.0232	0.00286
SIR_{10}	0.0496	0.0230	0.00299	0.0528	0.0229	0.00331

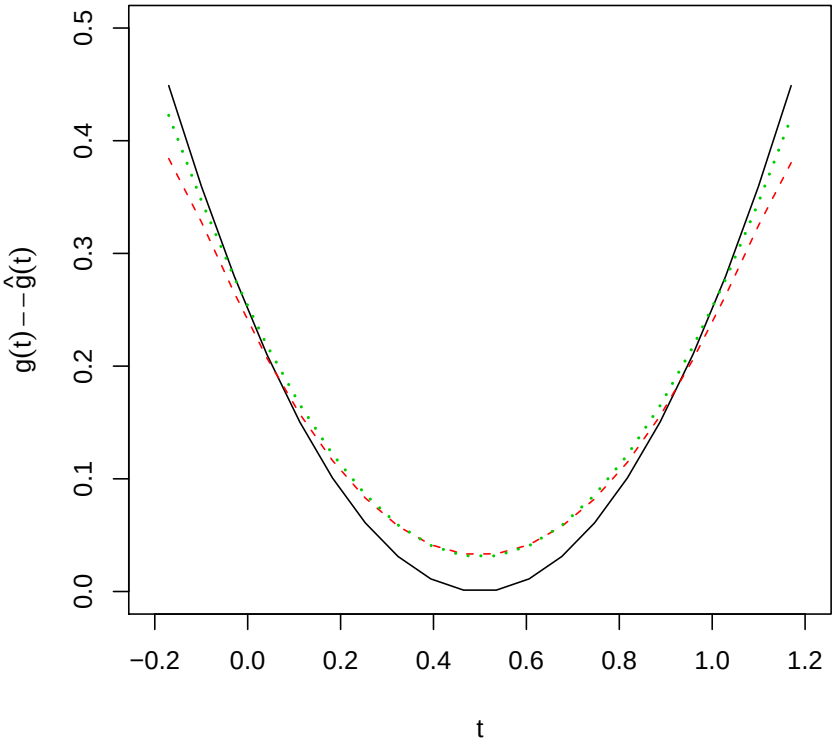

Figure 1: Curve estimate for a single replication of the quadratic model simulation study, with orthogonal β_Z and β_0. The true cure g(solid curve), the mean of $\hat{g}^*$ with GCV bandwidth (dashed curve) and a fixed optimal bandwidth $h_{opt} = 0.439$ (dotted curve) over 2000 simulations are shown.

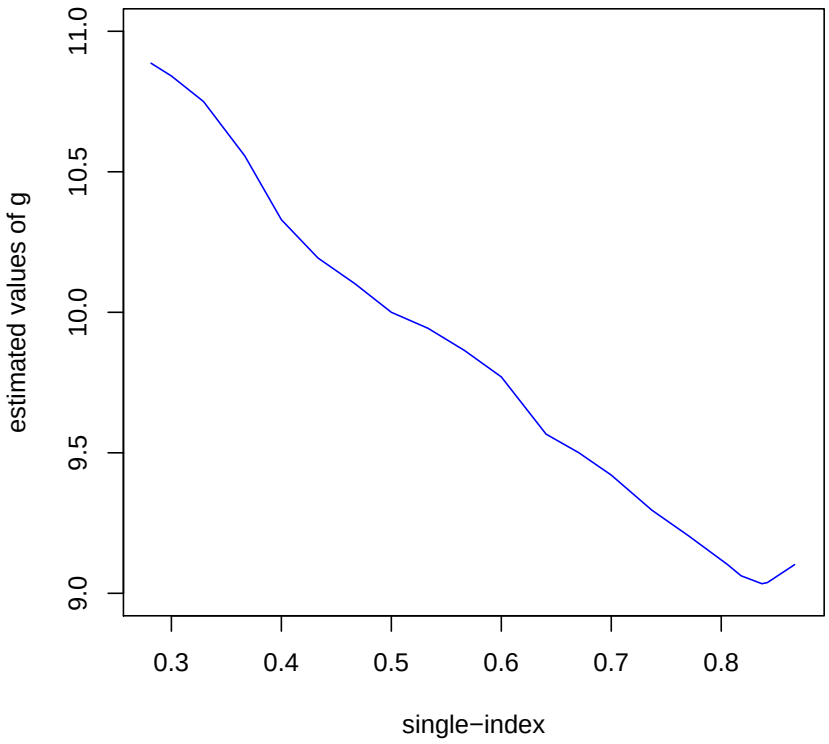

Figure 2: Curve estimate for the Boston Housing data, with $x^{\mathrm{T}}\hat{\beta}$ on the x-axis and $\hat{g}^*(t)$ on the y-axis.