ESTIMATION FOR A PARTIAL-LINEAR SINGLE-INDEX MODEL

BY JANE-LING WANG, LIUGEN XUE, LIXING ZHU AND YUN SAM CHONG

University of California at Davis, Beijing University of Technology, Hong Kong Baptist University and 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.

1. Introduction. Partial-linear models have attracted lots of attention due to their flexibility to combine traditional linear models with nonparametric regression models. See, for example, 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, $X^T \beta$, through a nonparametric link function $g$ is a viable option and termed the partial-linear single-index model. Specifically, it takes the form,

$$Y = Z^T \theta_0 + g(X^T \beta_0) + e,$$

where $Z$ is a $p \times 1$ vector of covariates, $\theta_0$ is a $p \times 1$ vector of regression coefficients, $\beta_0$ is a $d \times 1$ vector of single-index coefficients, and $e$ is the error term.

Received April 2008; revised March 2009.

1Supported in part by NSF Grant DMS-04-06430.

2Supported in part 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).

3Supported in part by a Grant of The Research Grant Council of Hong Kong, Hong Kong, China (HKBU7060/04P and HKBU 2030/07P).

AMS 2000 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. Dimension reduction, local linear smoothing, bandwidth, two-stage estimation, kernel smoother.
where \((X, Z) \in \mathbb{R}^p \times \mathbb{R}^q\) 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 \(0 < \text{Var}(e) = \sigma^2 < \infty\). For the sake of identifiability, it is often assumed that \(\|\beta_0\| = 1\) and the \(r\)th component of \(\beta_0\) 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 et al. (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 et al. (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 approach may become computationally unstable as observed by Yu and Ruppert (2002) and confirmed by our simulations in Section 3. The theory of Carroll et al. (1997) also relies on the strong assumption that their estimator for \(\theta_0\) 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,

\[
Z = \phi(X^T \beta_Z) + \eta,
\]

where \(\phi(\cdot)\) is an unknown function from \(\mathbb{R}^d\) to \(\mathbb{R}^q\), \(\beta_Z\) is a \(p \times d\) matrix with orthonormal columns and \(\eta\) 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 $\beta_0$ and $\theta_0$ simultaneously, we choose to estimate $\beta_0$ and $\theta_0$ sequentially. The idea is simple: $\theta_0$ can be estimated optimally through approaches developed for partial linear models once we have a $\sqrt{n}$ estimate of $\beta_0$ and plug it in (1.1). However, $\beta_0$ and $\theta_0$ may be correlated, leading to difficulties in identifying $\beta_0$. 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 $\eta$ in (1.2) is independent of $X$. Again, we need to impose the identifiability condition that $\beta Z$ has norm one and a positive first component. The procedure is as follows: First estimate $\beta Z$ 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 $\beta Z$ has been estimated, we proceed to estimate $\phi$ via a $d$-dimensional smoother and then obtain the residual for $\eta$. Since $\eta = Z - \phi(X^T \beta Z)$, plugging this into (1.1), we get

$$Y = \eta^T \theta_0 + h(X^T \beta_0, X^T \beta Z) + e,$$

where $h$ is an unknown function, but now $\eta$ and $X$ are independent of each other. It is thus possible to employ a least squares approach to estimate $\theta_0$, and the resulting estimate will be $\sqrt{n}$-consistent. We then employ a dimension-reduction procedure to $Y - Z^T \hat{\theta}_0$ and $X$ to obtain an estimate for $\beta_0$ and $g$. This concludes the first stage where the resulting estimates for $\theta_0$ and $\beta_0$ 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 $\theta_0$, we apply the profile method, also called partial regression in Speckman (1988), to estimate $\theta_0$. 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 $\beta_0$, 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 $\theta_0$, $\beta_0$, $\beta Z$, $\phi(\cdot)$ and $g(\cdot)$, and no iteration is needed.
2. We obtain the asymptotic normality of the estimator of $\beta_0$ and the optimal convergence rate of the estimator of $g(\cdot)$ as well as the asymptotic normality of
the estimator of $\theta_0$. The most attractive feature of this new method is that the estimator of $\beta_0$ has smaller limiting variance when compared to three existing approaches: Härdle, Hall and Ichimura (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 $\sigma^2$. It allows us to consider the construction of confidence regions and hypothesis testing for $\theta_0$ and $\beta_0$.

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 provides 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 $\{(X_i, Y_i, Z_i); 1 \leq i \leq n\}$, a sequence of independent and identically distributed (i.i.d.) samples from (1.1), that is,

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

where $e_1, \ldots, e_n$ are i.i.d. random errors with $E(e_i) = 0$ and $\text{Var}(e_i) = \sigma^2 > 0$, $\{e_i; 1 \leq i \leq n\}$ are independent of $\{(X_i, Z_i); 1 \leq i \leq n\}$, $X_i = (X_{i1}, \ldots, X_{ip})^T$, $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$, $\beta_0 \in \mathbb{R}^p$ and $\theta_0 \in \mathbb{R}^q$. 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 Remark 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 $\beta_Z$;
2. Smooth the $Z_i$ over $X_i^T \hat{\beta}_Z$ to get an estimator $\hat{\phi}(\cdot)$ of $\phi(\cdot)$, then compute the residuals $\hat{\eta}_i = Z_i - \hat{\phi}(X_i^T \hat{\beta}_Z)$;
3. Perform a linear regression of $Y_i$ versus $\hat{\eta}_i$’s to find an initial estimator $\hat{\theta}_0$ of $\theta_0$;
4. Apply a dimension-reduction method to the regression of $Y_i - Z_i^T \hat{\theta}_0$ versus $X_i$ to find an initial estimator $\hat{\beta}_0$ of $\beta_0$;
5. Smooth the $Y_i - Z_i^T \hat{\theta}_0$ versus the $X_i^T \hat{\beta}_0$ 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 $\theta_0$ through a profile approach for the partial linear model by minimizing (2.5);

7. Use the updated estimate $\hat{\theta}$ of $\theta_0$ from Step 6 to form the new residual $Y - Z^T\hat{\theta}$ then update the estimate of $\beta_0$ by solving the estimating equation (2.10);

8. Use the updated estimates of $\theta_0$ and $\beta_0$ 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 et al. (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 $\mathbb{R}$ and a bandwidth sequence $b = b_n$, define $K_b(\cdot) = b^{-1}K(\cdot/b)$. For a fixed $\beta$ and $\theta$, the local linear smoother aims at minimizing the weighted sum of squares

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

with respect to the parameters $d_v$, $v=0, 1$. Let $h = h_{n_1}$ 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

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

and

\begin{equation}
\hat{g}'(t; \beta, \theta) = \sum_{i=1}^{n} \tilde{W}_{ni}(t, \beta)(Y_i - Z_i^T \theta), \tag{2.2}
\end{equation}
where
\[ W_{ni}(t; \beta) = \frac{K_h(\beta^T X_i - t)[S_{n,2}(t; \beta, h) - (X_i^T \beta - t)S_{n,1}(t; \beta, h)]}{S_{n,0}(t; \beta, h)S_{n,2}(t; \beta, h) - S_{n,1}^2(t; \beta, h)}, \]
\[ \tilde{W}_{ni}(t; \beta) = \frac{K_h(\beta^T X_i - t)[(X_i^T \beta - t)S_{n,0}(t; \beta, h_1) - S_{n,1}(t; \beta, h_1)]}{S_{n,0}(t; \beta, h_1)S_{n,2}(t; \beta, h_1) - S_{n,1}^2(t; \beta, h_1)}, \]
and
\[ S_{n,l}(t; \beta, h) = \frac{1}{n} \sum_{i=1}^{n} (X_i^T \beta - t)^l K_h(\beta^T X_i - t), \quad l = 0, 1, 2. \]

The above estimators are for generic fixed values of \( \beta \) and \( \theta \). 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 \( \phi \) 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 \( \theta_0 \) 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 \( \theta \), the sum of squared errors,
\[ \sum_{i=1}^{n} \left[ Y_i - Z_i^T \theta - \hat{g}(X_i^T \hat{\beta}_0; \hat{\beta}_0, \theta) \right]^2, \]
where \( \hat{g} \) is the estimator in (2.1) of \( g \), obtained by smoothing \( Y_i - Z_i^T \theta \) versus \( X_i^T \hat{\beta}_0 \), and \( \hat{\beta}_0 \) is an initial estimator of \( \beta_0 \) which could be the initial estimator \( \hat{\beta}_0 \) in Step 4 or the refined estimator from Step 7 when an iterated estimator for \( \theta_0 \) is desirable. Because this smoother is expressed as a function of \( \theta \), 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, \( \tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T, \tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)^T \), where
\[ \tilde{Y}_i = Y_i - \hat{g}_1(X_i^T \hat{\beta}_0; \hat{\beta}_0), \quad \tilde{Z}_i = Z_i - \hat{g}_2(X_i^T \hat{\beta}_0; \hat{\beta}_0), \]
\[ \hat{g}_1(t; \hat{\beta}_0) = \sum_{i=1}^{n} W_{ni}(t; \hat{\beta}_0)Y_i, \quad \hat{g}_2(t; \hat{\beta}_0) = \sum_{i=1}^{n} W_{ni}(t; \hat{\beta}_0)Z_i, \]
with \( \hat{g}_1 \) and \( \hat{g}_2 \) the respective estimators of \( g_1(t) = E(Y|X^T\beta_0 = t) \) and \( g_2(t) = E(Z|X^T\beta_0 = t) \). The resulting partial regression estimator is thus

\[
\hat{\theta} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{Y}.
\]

For the estimator of \( \beta_0 \) in Step 7, we propose a novel method that takes advantage of the constraint \( \|eta_0\| = 1 \) and hence is more efficient than existing approaches, including the PPR approach in Härdle, Hall and Ichimura (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 heteroscedastic structure in (1.1), and least squares approaches have been standard dimension methods and lead to the same asymptotic variances for estimators of \( \beta_0 \). For instance, in the homoscedastic case, the estimator in Xia and Härdle (2006) has an asymptotic variance that is identical to that of Härdle, Hall and Ichimura (1993). Our approach, based on an estimating equation under the constraint \( \|eta_0\| = 1 \), is computationally stable and asymptotically more efficient, that is, its asymptotic variance is smaller. The efficiency gain can be attributed to a re-parametrization, making use of the constraint \( \|eta_0\| = 1 \) 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 \( \mathbb{R}^{p-1} \).

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

\[
\beta = \beta(\beta^{(r)}) = (\beta_1, \ldots, \beta_{r-1}, (1 - \|eta^{(r)}\|^2)^{1/2}, \beta_{r+1}, \ldots, \beta_p)^T.
\]

The true parameter \( \beta_0^{(r)} \) must satisfy the constraint \( \|eta_0^{(r)}\| < 1 \), and \( \beta \) is infinitely differentiable in a neighborhood of \( \beta_0^{(r)} \). This “remove-one-component” method for \( \beta \) has also been applied in Yu and Ruppert (2002).

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

\[
J_{\beta^{(r)}} = \frac{\partial \beta}{\partial \beta^{(r)}} = (\gamma_1, \ldots, \gamma_p)^T,
\]

where \( \gamma_s \) (\( 1 \leq s \leq p, s \neq r \)) is a \((p-1)\)-dimensional unit vector with \( s \)th component 1, and \( \gamma_r = -(1 - \|eta^{(r)}\|^2)^{-1/2} \beta^{(r)} \). To motivate the estimating equation, we start with the least squares criterion,

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

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 $\beta_0$, $\hat{\beta}$, via a transformation. This means that we transform a restricted least squares problem to an unrestricted least squares problem by solving the estimation equation,

\begin{equation}
(2.10) \quad \sum_{i=1}^{n} [Y_i - Z_i^T \hat{\theta} - \hat{g}(X_i^T \beta; \beta, \hat{\theta})] \hat{g}'(X_i^T \beta; \beta, \hat{\theta}) J_{\beta(\theta)}^T X_i = 0.
\end{equation}

We define the resulting estimator, $\hat{\beta}$ of $\beta_0$, as the final target estimator. Theorem 3 implies that our estimator for $\beta_0$ 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

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

and the estimator $\hat{\sigma}^2$ of $\sigma^2$ by $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [Y_i - Z_i^T \hat{\theta} - \hat{g}^*(X_i^T \hat{\beta})]^2$. Asymptotic results for the final parameter estimates of $\theta$ and $\beta$ are established in Theorems 1 and 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 to heteroscedastic errors, an additional dimension-reduction assumption on the variance function of $\eta$, given $X$, is needed to avoid the curse of high dimensional smoother needed in Step 2 to estimate $\phi$. 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 $\beta_0$ 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 $\beta$ and $\theta$ 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 $\beta_Z$, unless the covariate $Z$ is one-dimensional and the dimension $d$ of $\beta_Z$ is also small. In fact, the $\sqrt{n}$-consistency may not hold when $d > 3$ in (1.2) as shown in Xia et al. (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) 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 $\beta_0$ and $\theta_0$ 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 nonparametric 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 $X^T \beta$ is positive and satisfies a Lipschitz condition of order 1 for $\beta$ in a neighborhood of $\beta_0$. Further, $X^T \beta_0$ has a positive and bounded density function $f(t)$ on $T$ where $T = \{t = x^T \beta_0 : x \in A\}$.

C2. (i) The functions $g$ and $g_{2i}$ have two bounded and continuous derivatives where $g_{2i}$ is the $i$th component of $g_2(t)$, $1 \leq i \leq q$;
   (ii) $g_{3j}$ satisfies a Lipschitz condition of order 1 where $g_{3j}$ is the $j$th component of $g_3(t)$, and $g_3(t) = E(X|X^T \beta_0 = t)$, $1 \leq j \leq p$.

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 $R^1$. 

PARTIAL-LINEAR SINGLE-INDEX MODEL

C4. (i) sup \( E(\|Z\|^2|X^T\beta_0 = t) < \infty \);
(ii) \( E(e) = 0, \text{Var}(e) = \sigma^2 < \infty, E(e^4) < \infty \).

C5. (i) \( nh^2/\log^2 n \to \infty, \lim sup_{n \to \infty} nh^5 < \infty \);
(ii) \( nh_1^2/\log^2 n \to \infty, nh_1^4 \to 0, \lim sup_{n \to \infty} nh_1^5 < \infty \).

C6. (i) \( \Sigma_1 = \text{Cov}(Z - E(Z|X^T\beta_0)) \) is a positive definite matrix;
(ii) \( V = E[g'(X^T\beta_0)^2J_{\beta_0}^TXX^TJ_{\beta_0}] \) is a positive definite matrix, where \( J_{\beta_0}^{(r)} \) 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 \( X^T\beta \) 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 \( t = x^T\hat{\beta}, x \in A \) and \( \beta \) near \( \beta_0 \). 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 \( \beta_0 \), 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 \( \beta \) 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
\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \sigma^2 \Sigma^{-1}),
\]
where \( \Sigma \) is defined in condition (C6).

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 \( \beta_Z \) and \( \beta_0 \), \( \hat{\theta} \) is consistent for \( \theta_0 \) with the same asymptotic efficiency as an estimator that we would have obtained had we known \( \beta_0 \) 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 $\beta_0$ 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 $\beta_Z$ and of $\beta_0$, 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, see Cook and Weisberg (1991)] performs similarly. We expect the results for $\beta_0$ to hold when $Z$ is dependent of $X$, provided a good estimator of $\beta_Z$ is available. Under very general regularity conditions and for $q = 1$, Xia et al. (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 $\beta_0$. 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 $\beta_Z$ and $\beta_0$.

**Theorem 2.** Suppose that conditions C1–C6 hold. If the $r$th component of $\beta_0$ is positive, we have

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

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

From Härdle, Hall and Ichimura (1993) and Carroll et al. (1997), we can see that the estimator $\hat{\beta}$ of $\beta$ has an asymptotic variance that corresponds to a generalized inverse $\sigma^2 Q_1^{-1}$ where

$$Q_1 = E[g'(X^T \beta_0)^2 [X - E(X|X^T \beta_0)][X - E(X|X^T \beta_0)]^T].$$

Note that there may be infinitely many inverse matrices of $Q_1$, but there is a unique generalized inverse associated with the Jacobian $J_{\beta_0^{(r)}}$. The following theorem shows that the variance–covariance matrix in Theorem 2 is smaller than $\sigma^2 Q_1^{-1}$, the variance associated with $J_{\beta_0^{(r)}}$, in the sense that $\sigma^2 Q_1^{-1} - \sigma^2 J_{\beta_0^{(r)}} V^{-1} Q V^{-1} J_{\beta_0^{(r)}}^T$ is a nonnegative definite matrix. We use the usual notation: for two nonnegative matrices $A$ and $B$, $A \geq B$ denotes that $A - B$ is a nonnegative definite matrix.

**Theorem 3.** Under the conditions of Theorem 2, we have:

(i) there is a generalized inverse of $Q_1$ that is of the form $J_{\beta_0^{(r)}}^T Q^{-1} J_{\beta_0^{(r)}}$;

(ii) $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \sigma^2 J_{\beta_0^{(r)}} V^{-1} Q V^{-1} J_{\beta_0^{(r)}}^T)$;
\[(ii) \ J_{\beta_0(r)}^T Q^{-1} J_{\beta_0(r)} \geq \ J_{\beta_0(r)}^T V^{-1} Q V^{-1} J_{\beta_0(r)}^T.\]

**Remark 8.** Theorem 3 shows that our estimator of \(\beta_0\) is asymptotically more efficient than those of Härdle, Hall and Ichimura (1993) and of Carroll et al. (1997). In addition, Carroll et al. (1997) use an iterated procedure to estimate \(\beta_0\) and \(\theta_0\) while our estimation procedure does not require iteration.

From Theorem 2, we obtain an asymptotic result regarding the angle between \(\hat{\beta}\) and \(\beta_0\) 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 \(\beta_0\).

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 \(\sigma^2\).

**Theorem 4.** Suppose that the conditions of Theorem 1 hold. If \(\|\hat{\beta} - \beta_0\| = O_P(n^{-1/2})\), then
\[
\sup_{(x, \beta) \in A_n} |\hat{g}^*(x^T \beta) - g(x^T \beta_0)| = O_P ((nh/\log n)^{-1/2}),
\]
where \(A_n = \{(x, \beta) : (x, \beta) \in A \times \mathbb{R}^p, \|\beta - \beta_0\| \leq cn^{-1/2}\}\) for a constant \(c > 0\).

**Theorem 5.** Suppose that conditions C1–C6 hold and \(0 < \text{Var}(e_1^2) < \infty\). Then
\[
\sqrt{n}(\hat{\sigma}^2 - \sigma^2)/(\text{Var}(e_1^2))^{1/2} \xrightarrow{D} N(0, 1).
\]

Note that \(n^{-1}\tilde{Z}^T \tilde{Z} \xrightarrow{P} \Sigma\) in Lemma A.5 of the Appendix. By Theorems 1 and 4, we obtain
\[
(\tilde{Z}^T \tilde{Z})^{1/2}(\hat{\theta} - \theta_0)/\hat{\sigma} \xrightarrow{D} N(0, I_q).
\]

We are now in the position to construct confidence regions for \(\theta_0\). From Theorem 10.2d in Arnold (1981) we obtain the following result.

**Theorem 6.** Under the conditions of Theorem 5, we have
\[
(\hat{\theta} - \theta_0)^T (\hat{\Sigma}^T \hat{\Sigma})(\hat{\theta} - \theta_0)/\hat{\sigma}^2 \xrightarrow{D} \chi_q^2,
\]
where \(\chi_q^2\) is chi-square distributed with \(q\) degrees of freedom. Let \(\chi_q^2(1 - \alpha)\) be the \((1 - \alpha)\)-quantile of \(\chi_q^2\) for \(0 < \alpha < 1\), an asymptotic confidence region of \(\theta_0\) is
\[
R_\alpha = \{\theta : (\hat{\theta} - \theta)^T (\hat{\Sigma}^T \hat{\Sigma})(\hat{\theta} - \theta)/\hat{\sigma}^2 \leq \chi_q^2(1 - \alpha)\}.
\]
To construct confidence regions for $\beta_0$, a plug-in estimator of the limiting variance of $\hat{\beta}$ is needed. We respectively define the following estimators $\hat{V}$ and $\hat{Q}$ of $V$ and $Q$ by

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

and

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

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

$$(n^{-1} A^T J_{\hat{\beta}(r)} \hat{V}^{-1} \hat{Q}^{-1} J_{\hat{\beta}(r)}^T A)^{-1/2} A^T (\hat{\beta} - \beta_0) / \hat{\sigma} \xrightarrow{D} N(0, I_l).$$

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

$$(\hat{\beta} - \beta_0)^T A(n^{-1} A^T J_{\hat{\beta}(r)} \hat{V}^{-1} \hat{Q}^{-1} J_{\hat{\beta}(r)}^T A)^{-1} A^T (\hat{\beta} - \beta_0) / \hat{\sigma}^2 \xrightarrow{D} \chi^2_l.$$

The asymptotic confidence region of $A^T \beta_0$ is, letting $\chi^2_l(1 - \alpha)$ be the $(1 - \alpha)$-quantile of $\chi^2_l$ for $0 < \alpha < 1$,

$$R_{\alpha} = \{A^T \beta : (\hat{\beta} - \beta)^T A(n^{-1} A^T J_{\hat{\beta}(r)} \hat{V}^{-1} \hat{Q}^{-1} J_{\hat{\beta}(r)}^T A)^{-1} A^T (\hat{\beta} - \beta) / \hat{\sigma}^2 \leq \chi^2_l(1 - \alpha) \}.$$

### 3. Simulation study.

In this section, we examine the performance of the procedures in Section 2, for the estimation of both $\beta_0$ and $\theta_0$. 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 10–20 points. A quadratic model of the form

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

was used where $\theta_0 = 1$ is a scalar, $\beta_0 = (0.75, 0.5, -0.25, -0.25, 0.25)^T$, $X$ 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 \(\exp(\beta_Z X)/(1 + \exp(X^T \beta_Z))\) to be 1 and 0 otherwise. Two extreme cases of \(\beta_Z\) are reported in Tables 1 and 2; one based on choosing the same value as \(\beta_0\) with \(\beta_Z = \beta_0\), and the other on \(\beta_Z = (0.5, 0, 0.5, 0.5, -0.5)^T\), so that \(\beta_Z\) is orthogonal to \(\beta_0\). We also checked scenarios where \(\beta_Z\) and \(\beta_0\) 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 \(\eta\) has little effect on the accuracy of the final estimates of \(\theta_0\), so we choose an initial bandwidth of 0.5 to estimate \(\phi\) 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 \(\theta_0\) in the second stage, the GCV statistic is given by the formula

\[
GCV(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - Z_i^T \hat{\theta} - \hat{g}_h(X_i^T \hat{\beta}; \hat{\beta}, \hat{\theta})^2) / (n^{-1} \text{tr}(I - S_h))^2.
\]

### Table 1
Simulation results for \(\hat{\theta}\) with \(\beta_Z\) and \(\beta_0\) 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 |
| SIR₅             | 0.0095 | 0.0862 | 0.00753 | 0.0083 | 0.0869 | 0.00762 |
| SIR₁₀            | 0.0113 | 0.0788 | 0.00634 | 0.0098 | 0.0808 | 0.00663 |
| \(\beta_0\) given | 0.0031 | 0.0660 | 0.00436 |        |        |        |

### Table 2
Simulation results for \(\hat{\theta}\) with \(\beta_Z\) and \(\beta_0\) 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 |
| SIR₅             | −0.0115 | 0.1395 | 0.01960 | −0.0072 | 0.0919 | 0.00850 |
| SIR₁₀            | −0.0102 | 0.1362 | 0.01865 | −0.0083 | 0.0959 | 0.00926 |
| \(\beta_0\) given | −0.0024 | 0.0696 | 0.00485 |        |        |        |
where \( \hat{g}_h(\cdot) \) is the estimator of \( g(\cdot) \) with a bandwidth \( h \), and \( S_h \) is the smoothing matrix corresponding to a bandwidth of \( h \). The GCV bandwidth was selected to minimize (3.1). We use the optimal bandwidth, \( \hat{h}_{opt} \), for \( \hat{g} \) and \( \hat{\theta} \). When calculating the estimator \( \hat{\beta} \), we chose the bandwidths,

\[
(3.2) \quad h = \hat{h}_{opt} n^{1/5} n^{-1/3} = \hat{h}_{opt} n^{-2/15} \quad \text{and} \quad \hat{h}_1 = \hat{h}_{opt},
\]

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

In the simulation, PPR and SIR were used to obtain the initial estimators of \( \beta_0 \) and \( \beta_Z \). The notation \( SIR_c \) means that when we use SIR to estimate \( \beta_Z \), the number of data points per slice is \( c \). The resulting estimates for \( \theta_0 \) 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 \( \beta_0 \) 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.

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 \( \beta \)-directions were iteratively updated through estimates of the unknown link functions \( \phi \) and \( g \). In other nonadditive 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 \( \beta_0 \) is known, PPR typically attains 80% or more of the efficiency after one iteration.

For the estimation of \( \beta_0 \), we computed the angle (in radians) between \( \hat{\beta} \) and \( \beta_0 \) as a measure of accuracy. The mean, standard deviation (SD) and mean squared error (MSE) of the angle between \( \hat{\beta} \) and \( \beta_0 \) are reported in Table 3. Here, PPR leads to by far superior estimates compared to SIR.

**Table 3**

*Simulation results for the angles between \( \hat{\beta} \) and \( \beta_0 \)*

| \( \beta_Z \) and \( \beta_0 \) parallel | \( \beta_Z \) and \( \beta_0 \) orthogonal |
|-----------------------------------|-----------------------------------|
| Mean    | SD    | MSE    | Mean    | SD    | MSE    |
| PPR     | 0.0148 | 0.0056 | 0.00025 | 0.0157 | 0.0066 | 0.00029 |
| SIR5    | 0.0467 | 0.0223 | 0.00268 | 0.0482 | 0.0232 | 0.00286 |
| SIR10   | 0.0496 | 0.0230 | 0.00299 | 0.0528 | 0.0229 | 0.00331 |
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.

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 $\beta_Z$ and $\beta_0$ 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 $\beta_0$ and $\theta_0$ 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 $\theta_0$ is only a scalar, we postulate that their procedure has difficulties with high dimensional $\beta_0$, 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

$$\log(MV) = a_1 + a_2RM^2 + a_3AGE + a_4\log(DIS) + a_5\log(RAD) + a_6TAX$$
$$+ a_7PT \text{ RATIO} + a_8(B - 0.63)^2 + a_9\log(LSTAT) + a_{10}\text{CRIM}$$
$$+ a_{11}ZN + a_{12}\text{INDUS} + a_{13}\text{CHAS} + a_{14}\text{NOX}^P + e.$$  

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 $X^T\beta_0$ and the estimated dimension-reduction variable $X^T\hat{\beta}_0$.

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) point out, SIR is not sensitive to the choice of slice number, and they try 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 $\theta$ 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 $\beta$ 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 $\beta_0$, we then estimate the relationship between $Y$ and $X^T\hat{\beta}$. GCV yields a bandwidth of 0.16, and we obtain
Fig. 2. Curve estimate for the Boston Housing data, with $x^T \hat{\beta}$ on the x-axis and $\hat{g}^*(t)$ on the y-axis.

$R^2 = 0.8021$. Even though the Charles River variable is significant, its inclusion leads to only a minor increase in $R^2$.

Figure 2 shows the estimated $g$ along with the data. On the x-axis of the above graph, the estimated value $x^T \hat{\beta}$ 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

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

From (2.6) we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \tilde{G} + \sqrt{n}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T e.$$
Lemma A.5 in the Appendix implies
\[
(5.1) \quad n(\tilde{Z}^T \tilde{Z})^{-1} \xrightarrow{p} \Sigma^{-1}.
\]
Therefore, Lemma A.6 in the Appendix leads to
\[
\sqrt{n}(\tilde{Z}^T \tilde{Z})^{-1/2} \tilde{Z}^T \tilde{G} \xrightarrow{p} 0.
\]
It remains to show that
\[
(5.2) \quad \sqrt{n}(\tilde{Z}^T \tilde{Z})^{-1/2} \tilde{Z}^T e \xrightarrow{D} N\left(0, \sigma^2/\Sigma_1^{-1}\right).
\]
Since
\[
\tilde{Z}^T e = \sum_{i=1}^{n} [Z_i - g_2(X_i^T \beta_0)]e_i + \sum_{i=1}^{n} [g_2(X_i^T \beta_0) - \hat{g}_2(X_i^T \hat{\beta}_0; \hat{\beta}_0)]e_i =: M_1 + M_2.
\]
Central Limit Theorem implies \(n^{-1/2} M_1 \xrightarrow{D} N(0, \Sigma)\). Note that \(\hat{g}_2\) converges to \(g_2\) and \(e_i\) are independent identically distributed with mean zero. 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 \(\beta_0\), 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 \(\beta_0\) minimizing expression (2.9) in \(B_{1n}\) where \(B_{1n} = \{\beta : \|\beta - \beta_0\| = B_1 n^{-1/2}\}\) for some constant such that \(0 < B_1 < \infty\).

In fact, let \(Y = (Y_1, \ldots, Y_n)^T\) and \(Z = (Z_1, \ldots, Z_n)^T\). We have
\[
D(\beta) = (Y - Z\hat{\beta})^T (I - S_{\beta})^T (I - S_{\beta})(Y - Z\hat{\beta})
\]
\[
= (Y - Z\theta_0)^T (I - S_{\beta})^T (I - S_{\beta})(Y - Z\theta_0)
\]
\[
- 2(Y - Z\theta_0)^T (I - S_{\beta})^T (I - S_{\beta})Z(\hat{\theta} - \theta_0)
\]
\[
+ [Z(\hat{\theta} - \theta_0)]^T (I - S_{\beta})^T (I - S_{\beta})Z(\hat{\theta} - \theta_0)
\]
\[
=: D_1(\beta) - D_2(\beta) + D_3(\beta).
\]
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 \(\beta\). This implies \(D(\beta) = D_1(\beta) - R_0 + o_p(1)\). Thus minimizing \(D(\beta)\) simultaneously with respect to \(\beta\) is very much like separately minimizing \(D_1(\beta)\) with respect to \(\beta\).

It follows from (2.7) that we only need to prove the existence of an estimator of
$\beta_0^{(r)}$ in $B_{2n}$ where $B_{2n} = \{ \beta^{(r)} : \| \beta^{(r)} - \beta_0^{(r)} \| = B_{2n}^{-1/2} \}$ 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.7) of Lemma A.7. For an arbitrary $\beta^{(r)} \in B_{2n}$ with the value of constant $B_2$ in $B_{2n}$ to be determined, we have from Lemma A.7 in the Appendix that

$$
(\beta^{(r)} - \beta_0^{(r)})^T R(\beta^{(r)})
= (\beta^{(r)} - \beta_0^{(r)})^T U(\beta_0^{(r)}) - n(\beta^{(r)} - \beta_0^{(r)})^T V(\beta^{(r)} - \beta_0^{(r)}) + o_P(1).
$$

The following arguments are similar to those used by Weisberg and Welsh (1994). 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 $| (\beta^{(r)} - \beta_0^{(r)})^T U(\beta_0^{(r)}) | = B_2 O_P(1)$ and $n(\beta^{(r)} - \beta_0^{(r)})^T V(\beta^{(r)} - \beta_0^{(r)}) \sim B_2^2$. So, for any given $\eta > 0$, if $B_2$ is chosen large enough, then it will follows that $(\beta^{(r)} - \beta_0^{(r)})^T R(\beta_0^{(r)}) < 0$ on an event with probability $1 - \eta$. From the arbitrariness of $\eta$, we can prove the existence of the least squares estimator of $\beta_0^{(r)}$ in $B_{2n}$ 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 $B_{2n}$ to the equation $R(\beta^{(r)}) = 0$. That is, $R(\hat{\beta}^{(r)}) = 0$. By Lemma A.7, we have

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

and hence

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

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

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

and hence

$$
\sqrt{1 - \| \hat{\beta}^{(r)} \|^2} - \sqrt{1 - \| \beta_0^{(r)} \|^2}
= -\frac{2 \beta_0^{(r)T}(\hat{\beta}^{(r)} - \beta_0^{(r)}) + \| \hat{\beta}^{(r)} - \beta_0^{(r)} \|^2}{\sqrt{1 - \| \hat{\beta}^{(r)} \|^2 + \sqrt{1 - \| \beta_0^{(r)} \|^2}}}
= -\frac{\beta_0^{(r)T}(\hat{\beta}^{(r)} - \beta_0^{(r)})}{\sqrt{1 - \| \beta_0^{(r)} \|^2}} + O_P(n^{-1}).
$$
It follows from (2.7) and the above equation, that

\[
\hat{\beta} - \beta_0 = \begin{pmatrix}
\hat{\beta}_1 \\
\vdots \\
\hat{\beta}_{r-1} \\
\sqrt{1 - \|\hat{\beta}^{(r)}\|^2} \\
\hat{\beta}_{r+1} \\
\vdots \\
\hat{\beta}_p \\
\beta_1 - \beta_{01} \\
\vdots \\
\beta_0 - \beta_{0(r-1)} \\
\beta_0^{(r)}(\hat{\beta}^{(r)} - \beta_0^{(r)}) \\
\beta_0^{(r)}T(\hat{\beta}^{(r)} - \beta_0^{(r)}) \\
\sqrt{1 - \|\beta_0^{(r)}\|^2} \\
\hat{\beta}_{r+1} - \beta_{0(r+1)} \\
\vdots \\
\hat{\beta}_p - \beta_{0p}
\end{pmatrix}
- \begin{pmatrix}
\beta_{01} \\
\vdots \\
\beta_{0(r-1)} \\
\beta_{0(r+1)} \\
\vdots \\
\beta_{0p}
\end{pmatrix}
+ O_P(n^{-1}).
\]

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

\[
\hat{\beta} - \beta_0 = J^{(r)}_{\beta_0} \left( \hat{\beta}^{(r)} - \beta_0^{(r)} \right) + O_P(n^{-1}).
\]

Thus we have

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

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

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

\[
\Pi_0 := J^{(r)}_{\beta_0} Q^{-1} J^{T}_{\beta_0^{(r)}}, \quad \Pi_1 := J^{(r)}_{\beta_0} V^{-1} Q V^{-1} J^{T}_{\beta_0^{(r)}}.
\]

We now prove that \( \Pi_0 \) is a generalized inverse of \( Q_1 \). To this end, we need to prove that \( \Pi_0 Q_1 \Pi_0 = \Pi_0 \) and \( Q_1 \Pi_0 Q_1 = Q_1 \). Note that

\[
\Pi_0 Q_1 \Pi_0 = J^{(r)}_{\beta_0} Q^{-1} J^{T}_{\beta_0^{(r)}} Q_1 J^{(r)}_{\beta_0} Q^{-1} J^{T}_{\beta_0^{(r)}} = J^{(r)}_{\beta_0} Q^{-1} Q Q^{-1} J^{T}_{\beta_0^{(r)}} = \Pi_0.
\]

We now prove \( Q_1 \Pi_0 Q_1 = Q_1 \). First, by QR decomposition [see, e.g. Gentle (1998), Section 3.2.2, pages 95–97 for more details] for \( \beta_0 \), we can find its orthogonal complement such that \( B = (b_1, \beta_0) \) is an orthogonal matrix, and \( \beta_0 = B^{(6)} \).
Thus $J_{\beta_0^{(r)}} = BB^T J_{\beta_0^{(r)}} = BR$ where $R = (R_1 \ R_2)$ with $R_1$ being a $(p - 1) \times (p - 1)$ nonsingular matrix. Further, note that

$$Q = J_{\beta_0^{(r)}}^T Q_{\beta_0^{(r)}} = R^T B^T Q_1 B R$$

$$= R^T \left( b_1^T Q_1 b_1 \ 0 \right) R$$

$$= R_1^T b_1^T Q_1 b_1 R_1 .$$

To prove the result, we rewrite $Q_1$ in another form. Define $S = B \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix}$. $S$ is a nonsingular matrix. Then

$$Q_1 = (S^T)^{-1} S^T Q_1 S S^{-1} = (S^T)^{-1} \left( \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix} \right) B^T Q_1 B \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix} S^{-1}$$

$$= (S^T)^{-1} S^T Q_1 S S^{-1} = (S^T)^{-1} \left( \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix} \right) \left( b_1^T Q_1 b_1 \ 0 \right) \left( \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix} \right) S^{-1}$$

$$= (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1} .$$

We now prove that $Q_1 \Pi_0 Q_1 = Q_1$ is of the above form. From the above and noting that $S^{-1} = \left( \begin{pmatrix} R_1 & 0 \\ 0 & 1 \end{pmatrix} \right) B^T$ and $(S^T)^{-1} = B \begin{pmatrix} (R_1^{-1}) & 0 \\ 0 & 1 \end{pmatrix}$, we have

$$Q_1 \Pi_0 Q_1$$

$$= (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1} B R Q^{-1} R^T B^T (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1}$$

$$= (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) \left( \begin{pmatrix} R_1^{-1} & 0 \\ 0 & 1 \end{pmatrix} \right) B^T B R Q^{-1} R^T B^T$$

$$\times \left( \begin{pmatrix} R_1^{-1} & 0 \\ 0 & 1 \end{pmatrix} \right) \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1}$$

$$= (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) \left( \begin{pmatrix} I & R_2 \\ R_2 & 0 \end{pmatrix} \right) Q^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1}$$

$$= (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) Q^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1} = (S^T)^{-1} \left( \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \right) S^{-1} = Q_1 .$$

Thus $\Pi_0$ is one of the solutions of $Q_1^{-1}$. To prove that the asymptotic variance–covariance matrix $\sigma^2 \Pi_1$ of our estimator is smaller than the corresponding matrix $\sigma^2 \Pi_0$ given in Härdle, Hall and Ichimura (1993), we only need to show that $\Pi_0 - \Pi_1$ is a positive semi-definite matrix, that is, $\Pi_0 > \Pi_1$. Recall that $V = J_{\beta_0^{(r)}}^T E \left( g \left( X^T \beta_0 \right)^2 X X^T \right) J_{\beta_0^{(r)}}$. Note that both $Q$ and $V$ are positive definite matrices and obviously $V \geq Q$. Thus $Q^{-1} \geq V^{-1}$, and then $V^{-1} \geq V^{-1} Q V^{-1}$. 

From these two inequalities, it is easy to see that
\[ \Pi_0 \geq J_{\beta(r)}^T V^{-1} J_{\beta(r)} \geq J_{\beta(r)}^T V^{-1} Q V^{-1} J_{\beta(r)} = \Pi_1. \]

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
\[ |\cos(\hat{\beta}, \beta_0) - 1| = |(\hat{\beta} - \beta_0) \bullet \beta_0 / \| \hat{\beta} \| + (\| \beta_0 \| - \| \hat{\beta} \|) / \| \hat{\beta} \| | \leq 3 \| \hat{\beta} - \beta_0 \| / \| \hat{\beta} \| = O_P(n^{-1/2}). \]

This completes the proof of Corollary 1. □

**Proof of Theorem 4.** Denote \( \theta_0 = (\theta_{01}, \ldots, \theta_{0q})^T, \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)^T \). Theorem 1 and Lemma A.4 in the Appendix yield
\[
\sup_{(x, \beta) \in A_n} |\hat{g}^*(x^T \beta) - g(x^T \beta_0)| \\
\leq \sum_{s=1}^q \sup_{(x, \beta) \in A_n} |\hat{g}_{2s}(x^T \beta; \beta) - g_{2s}(x^T \beta_0)||\hat{\theta}_s - \theta_{0s}| \\
+ \sup_{(x, \beta) \in A_n} |\hat{g}(x^T \beta; \beta, \theta_0) - g(x^T \beta_0)| \\
+ \sum_{s=1}^q \sup_{x \in A} |g_{2s}(x^T \beta_0)||\hat{\theta}_s - \theta_{0s}| = O_P((nh/\log n)^{-1/2}),
\]
and hence Theorem 4 follows. □

**Proof of Theorem 5.** Decomposing \( \hat{\sigma}^2 \) into several parts, we have
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 + \frac{1}{n} \sum_{i=1}^n [Z_i^T (\theta_0 - \hat{\theta}) + g(X_i^T \beta_0) - \hat{g}(X_i^T \hat{\beta}; \hat{\beta}, \theta_0)]^2 \\
+ \frac{2}{n} \sum_{i=1}^n e_i Z_i^T (\theta_0 - \hat{\theta}) + \frac{2}{n} \sum_{i=1}^n e_i [g(X_i^T \beta_0) - \hat{g}(X_i^T \hat{\beta}; \hat{\beta}, \theta_0)]
\]
\[ =: I_1 + I_2 + I_3 + I_4. \]

Note that \( \sqrt{n}\| \hat{\theta} - \theta_0 \| = O_P(1) \) and using (A.4) of Lemma A.4, we have
\[
\sqrt{n} I_2 \leq \frac{1}{n} \sum_{i=1}^n \| Z_i \|^2 \sqrt{n}\| \hat{\theta} - \theta_0 \|^2 \\
+ \sqrt{n} \sup_{(x, \beta) \in A_n} |g(x^T \beta_0) - \hat{g}(x^T \beta; \beta, \theta_0)|^2 \\
= O_P(n^{-1/2}) + O_P((nh^2/\log^2 n)^{-1/2}) \xrightarrow{P} 0.
\]
Since $Ee_i = 0$, we obtain $\sqrt{n}I_3 \xrightarrow{P} 0$. Similarly to the proof of Lemma A.6 in the full version of this paper, 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 for the proofs of 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 $\hat{g}$, $g_2$, and $\hat{g}^\prime$. 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 $e_j$ for $\hat{\theta} - \theta_0$ is $o_P(n^{-1/2})$. Lemma A.7 provides the main step for the proof of Theorem 2. To save space, we do not provide the details of the proofs. The full version of this paper [Wang et al. (2009)] is available from arXiv.org at arXiv:0905.2042.

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

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

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

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

then for any $\beta_0 \in \mathbb{R}^p$, the sequence $\xi_i(x, \beta_0)$ converges weakly to a normal distribution with mean zero and variance $\frac{1}{n} \sum_{i=1}^n \xi_i(x, \beta_0)^2$ as $n \to \infty$.
for \((x, \beta) \in A_n\), then we have

\[
P \left\{ \sup_{(x, \beta) \in A_n} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i(x, \beta) \right| > \frac{1}{2} \varepsilon_n \right\} \leq c_1 n^{2p_a} \varepsilon_n^{-2p} E \left\{ \sup_{(x, \beta) \in A_n} 2 \exp \left( \frac{-n^2 \varepsilon_n^2 / 128}{\sum_{i=1}^{n} \xi_i^2(x, \beta)} \right) \wedge 1 \right\},
\]

(A.3)

where \(c_1 > 0\) is a constant.

The proof uses the arguments for the symmetrization lemma in Pollard (1984).

**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 \(i = 1, \ldots, n\), we have

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

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

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

and

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

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

The proof is similar to that of Lemma 1 of Zhu and Xue (2006).

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

\[
\begin{align*}
E \left[ W_{ni}^2(X_i^T \beta_0; \beta_0) \right] &= O((nh)^{-2}), \\
E \left\{ \sum_{j=1, j \neq i}^{n} W_{nj}^2(X_i^T \beta_0; \beta_0) \right\} &= O((nh)^{-1}), \\
E \left[ \sum_{j=1}^{n} W_{nj}^2(x^T \beta; \beta) \right] &= O((nh)^{-1})
\end{align*}
\]
and
\[
\begin{align*}
E\{\tilde{W}_{ni}^2(X_i^T \beta_0; \beta_0)\} &= O((nh_1)^{-2} + (n^3h_1^5)^{-1}), \\
E\left\{\sum_{j=1, j\neq i}^n \tilde{W}_{nj}^2(X_j^T \beta_0; \beta_0)\right\} &= O((nh_1)^{-1}).
\end{align*}
\]

The proof is similar to that of Lemma 2 of Zhu and Xue (2006).

The proofs of the following two lemmas are standard by applying the previous lemmas.

**Lemma A.4.** Suppose that conditions C1–C4 and C5(i) hold. We then have
\[
\sup_{(x, \beta) \in \mathcal{A}_n} |g(x^T \beta_0) - \hat{g}(x^T \beta; \beta, \theta_0)| = o_P((nh/\log n)^{-1/2})
\]
and
\[
\sup_{(x, \beta) \in \mathcal{A}_n} |g_{2s}(x^T \beta_0) - \hat{g}_{2s}(x^T \beta; \beta)| = o_P((nh/\log n)^{-1/2}).
\]

If, in addition, C5(ii) also holds, then we have
\[
\sup_{(x, \beta) \in \mathcal{A}_n} |g'(x^T \beta_0) - \hat{g}'(x^T \beta; \beta, \theta_0)| = o_P((nh^3/\log n)^{-1/2}),
\]
where $\mathcal{A}_n = \{(x, \beta) : (x, \beta) \in A \times R^p, \|\beta - \beta_0\| \leq cn^{-1/2}\}$ for a constant $c > 0$.

**Lemma A.5.** Under the assumptions of Theorem 1, we have
\[
n^{-1}\tilde{Z}^T \tilde{Z} \overset{P}{\longrightarrow} \Sigma,
\]
where $\Sigma$ is defined in condition C6.

**Lemma A.6.** Under the assumptions of Theorem 1, we have
\[
n^{-1/2}\tilde{Z}^T \tilde{G} := \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{Z}_i [g(X_i^T \beta_0) - \hat{g}(X_i^T \hat{\beta}_0; \hat{\beta}_0, \theta_0)] \overset{P}{\longrightarrow} 0.
\]

This can be proved by applying Lemmas A.1–A.4.

**Lemma A.7.** Suppose that conditions C1–C6 are satisfied, then we have
\[
\sup_{\beta^{(r)} \in \mathcal{B}_n} \|R(\beta^{(r)}) - U(\hat{\beta}_0^{(r)}) + nV(\beta^{(r)} - \hat{\beta}_0^{(r)})\| = o_P(\sqrt{n}),
\]
where $\mathcal{B}_n = \{\beta^{(r)} : \|\beta^{(r)} - \hat{\beta}_0^{(r)}\| \leq Cn^{-1/2}\}$ for a constant $C > 0$, $V$ is defined in condition C6,
\[
R(\beta^{(r)}) = \sum_{i=1}^n [Y_i - Z_i^T \theta_0 - \hat{g}(X_i^T \beta; \beta, \theta_0)]\hat{g}'(X_i^T \beta; \beta, \theta_0) \mathbf{J}_{\beta^{(r)}}^T X_i
\]
\begin{equation*}
U(\beta_0^{(r)}) = \sum_{i=1}^{n} e_i g'(X_i^T \beta_0) J_{\beta_0^{(r)}}^T [X_i - E(X_i | X_i^T \beta_0)].
\end{equation*}

The proof can be done by separating \( R(\beta^{(r)}) \) as
\begin{equation*}
R(\beta^{(r)}) = \sum_{i=1}^{n} e_i g'(X_i^T \beta_0) J_{\beta_0^{(r)}}^T [X_i - E(X_i | X_i^T \beta_0)]
\end{equation*}
\begin{equation*}
+ \sum_{i=1}^{n} e_i [\hat{g}'(X_i^T \beta; \beta_0) - g'(X_i^T \beta_0)] J_{\beta_0^{(r)}}^T X_i
\end{equation*}
\begin{equation*}
- \sum_{i=1}^{n} g'(X_i^T \beta_0) J_{\beta_0^{(r)}}^T [X_i (\hat{g}(X_i^T \beta; \beta, \theta_0) - \hat{g}(X_i^T \beta_0; \beta_0, \theta_0))]
\end{equation*}
\begin{equation*}
- \sum_{i=1}^{n} g'(X_i^T \beta_0) J_{\beta_0^{(r)}}^T [X_i [\hat{g}(X_i^T \beta_0; \beta_0, \theta_0) - g(X_i^T \beta_0)] - e_i g_3(X_i^T \beta_0)]
\end{equation*}
\begin{equation*}
- \sum_{i=1}^{n} [\hat{g}(X_i^T \beta; \beta, \theta_0) - g(X_i^T \beta_0)] [\hat{g}'(X_i^T \beta; \beta, \theta_0) - g'(X_i^T \beta_0)] J_{\beta_0^{(r)}}^T X_i
\end{equation*}
\begin{equation*}
= R_1(\beta^{(r)}) + R_2(\beta^{(r)}) - R_3(\beta^{(r)}) - R_4(\beta^{(r)}) - R_5(\beta^{(r)})
\end{equation*}
and applying Lemma A.4 to related estimators \( \hat{g} \) and \( \hat{g}' \). The details can be found in the full version.

**Acknowledgments.** 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.

**REFERENCES**

ARNOLD, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. Wiley, New York. [MR0606011]

BHATTACHARYA, P. K. and ZHAO, P.-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* 25 244–262. [MR1429924]

CARROLL, R. J., FAN, J., GIJBEELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* 92 477–489. [MR1467842]

CHEN, C.-H. and LI, K.-C. (1998). Can SIR be as popular as multiple linear regression. *Statist. Sinica* 8 289–316. [MR1624402]

CHEN, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* 16 136–141. [MR0924861]

CHEN, H. and SHHAU, J.-J. H. (1994). Data-driven efficient estimators for a partially linear model. *Ann. Statist.* 22 211–237. [MR1272081]
CHIOU, J. M. and MÜLLER, H. G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* 93 1376–1387. MR1666634

COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York. MR1645673

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression 1, 2. *Statist. Sci.* 22 1–26. MR2408655

COOK, R. D. and LI, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* 30 455–474. MR1902895

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. MR1137117

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. MR0516581

DOKSUM, K. and SAMAROV, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.* 23 1443–1473. MR1370291

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London. MR1383587

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* 76 817–823. MR0650892

GENTLE, J. E. (1998). *Numerical Linear Algebra for Applications in Statistics*. Springer, Berlin. MR1638791

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* 17 573–588. MR0994251

HÄRDLE, W., HALL, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* 21 157–178. MR1212171

HÄRDLE, W., GAO, J. and Ichimura, H. (2000). *Partially Linear Models*. Springer, New York.

HARRISON, D. and RUBINFELD, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5 81–102.

HECKMAN, N. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. A* 48 244–248. MR0868002

HRISTACHE, M., JUDITSKY, A. and SPEKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* 29 595–623. MR1865333

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* 20 1040–1061. MR1165605

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.* 103 1177–1186. MR2462891

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* 86 316–342. MR1137117

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. MR1209564

LI, K. C., CARROLL, R. J. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* 98 99–109. MR1965677

LI, Y. X. and ZHU, L. X. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* 35 41–69. MR2332268

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York. MR0762984

RICE, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* 4 203–208. MR0848718

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, New York. MR1998720

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* 50 413–434. MR0970977
STOKER, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481. MR0868152

STUTE, W. and ZHU, L. X. (2005). Nonparametric checks for single-index models. *Ann. Statist.* **33** 1048–1083. MR2195628

WANG, J. L., XUE, L. G., ZHU, L. X. and CHONG, Y. S. (2009). Estimation for a partial-linear single index model. Available at arXiv:0905.2042.

WEISBERG, S. and WELSH, A. H. (1994). Adapting for the missing linear link. *Ann. Statist.* **22** 1674–1700. MR1329165

WELSH, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17** 337–361. [Correction (1990) **18** 1500.] MR0981455

XIA, Y. and HÄRDLE, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97** 1162–1184. MR2276153

XIA, Y., TONG, H., LI, W. K. and ZHU, L. X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64** 363–410. MR1924297

XIA, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

YIN, X. and COOL, R. D. (2002). Dimension reduction for the conditional $k$-th moment in regression. *J. Roy. Statist. Soc. Ser. B* **64** 159–175. MR1904698

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

ZHU, L. X. and NG, K. W. (1995). Asymptotics for sliced inverse regression. *Statist. Sinica* **5** 727–736. MR1347616

ZHU, L. X. and NG, K. W. (2003). Checking the adequacy of a partial linear model. *Statist. Sinica* **13** 763–781. MR1997173

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. MR2278341

---

J.-L. WANG  
DEPARTMENT OF STATISTICS  
UNIVERSITY OF CALIFORNIA AT DAVIS  
DAVIS, CALIFORNIA 95616  
USA  
E-MAIL: wang@wald.ucdavis.edu

L. XUE  
SCHOOL OF APPLIED MATHEMATICS AND PHYSICS  
BEIJING UNIVERSITY OF TECHNOLOGY  
BEIJING  
CHINA  
E-MAIL: lgxue@bjut.edu.cn

L. ZHU  
DEPARTMENT OF MATHEMATICS  
HONG KONG BAPTIST UNIVERSITY  
HONG KONG  
CHINA  
E-MAIL: lzhu@hkbu.edu.hk

Y. S. CHONG  
WILLIAM E. WECCKER ASSOCIATES, INC.  
505 SAN MARIN DRIVE  
NOVATO, CALIFORNIA 94945  
USA