Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

A Multiple-Index Model and Dimension Reduction

Yingcun Xia

A Multiple-Index Model and Dimension Reduction

Yingcun XIA

Dimension reduction can be used as an initial step in statistical modeling. Further specification of model structure is imminent and important when the reduced dimension is still greater than 1. In this article we investigate one method of specification that involves separating the linear component from the nonlinear components, leading to further dimension reduction in the unknown link function and, thus, better estimation and easier interpretation of the model. The specified model includes the popular econometric multiple-index model and the partially linear single-index model as its special cases. A criterion is developed to validate the model specification. An algorithm is proposed to estimate the model directly. Asymptotic distributions for the estimators of the parameters and the nonparametric link function are derived. Air pollution data in Chicago are used to illustrate the modeling procedure and to demonstrate its advantages over the existing dimension reduction approaches.

KEY WORDS: Asymptotic distribution; Convergence of algorithm; Dimension reduction; Local linear smoother; Semiparametric model.

1. INTRODUCTION

Suppose that Y is a response and $\mathbf{X}$ is a p-dimensional covariate vector. An essential association between Y and $\mathbf{X}$ is the conditional mean function $M(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, leading to a general model $Y = M(\mathbf{X}) + \varepsilon$, where $E(\varepsilon|\mathbf{X}) = 0$ almost surely. Due to the "curse of dimensionality," the general model with $p > 1$ is difficult to estimate well and, thus, hard to use in practice. To lessen the effect of dimensionality, a popular approach is dimension reduction of the conditional mean (Samarov 1993; Hristache, Juditski, Polzehl, and Spokoiny 2001; Cook and Li 2002; Xia, Tong, Li, and Zhu 2002; Yin and Cook 2002; among others), which searches for d_A linear combinations of $\mathbf{X}$: $\bar\beta_1^\top\mathbf{X}, \dots, \bar\beta_{d_A}^\top\mathbf{X}$ with $d_A < p$, such that $M(\mathbf{X})$ can be well approximated by a lower dimensional function $m(\bar\beta_1^\top\mathbf{X}, \dots, \bar\beta_{d_A}^\top\mathbf{X})$ or, more ideally, $M(\mathbf{X}) = m(\bar\beta_1^\top\mathbf{X}, \dots, \bar\beta_{d_A}^\top\mathbf{X})$, leading to the *dimension reduction model in the regression mean function*:

$$\text{Model (A):}\quad Y = m\big(\bar\beta_1^\top\mathbf{X}, \dots, \bar\beta_{d_A}^\top\mathbf{X}\big) + \varepsilon.$$

This model has received much attention and was investigated intensively. For example, all the references cited previously concern this model. Dimension reduction is an initial step for model construction.

If $d_A > 1$, model (A) needs further specification in order to be estimated better; see, for example, Samarov (1993), Carroll, Fan, Gijbels, and Wand (1997), and Samarov, Spokoiny, and Vial (2005). It is very likely that one of the dimension reduction components $\bar\beta_1^\top\mathbf{X}, \dots, \bar\beta_{d_A}^\top\mathbf{X}$, or its linear combination $c_1\bar\beta_1^\top\mathbf{X} + \cdots + c_{d_A}\bar\beta_{d_A}^\top\mathbf{X} \equiv \gamma_0^\top\mathbf{X}$, affects the response linearly, and the other orthogonal $d_A - 1$ combinations affect the response nonlinearly, leading to the model:

$$\text{Model (B):}\quad Y = G\big(\tilde\beta_1^\top\mathbf{X}, \dots, \tilde\beta_{d_A-1}^\top\mathbf{X}\big) + \gamma_0^\top\mathbf{X} + \varepsilon,$$

where $E(\varepsilon|\mathbf{X}) = 0$ almost surely and G is an unknown link function. A well-known model in econometrics proposed by Ichimura and Lee (1991) and Horowitz (1998) has a similar form. Following them, we call model (B) the multiple-index model. Compared with model (A), the dimension of the nonparametric function in model (B) is reduced by 1. Therefore, $G(\cdot)$ can be estimated much more accurately than $m(\cdot)$ in model (A). For ease of exposition, let

$$\mathbf{B}_A = \big(\bar\beta_1, \dots, \bar\beta_{d_A}\big), \qquad \mathbf{B}_B = \big(\tilde\beta_1, \dots, \tilde\beta_{d_A-1}\big), \qquad d_B = d_A - 1.$$

Model (B) is very general, including the partially linear model (Speckman 1988; Wang 2003) and the single-index model (Ichimura 1993) as its special cases. An important special case of model (B) is when $d_B = 1$, that is, $Y = \gamma_0^\top\mathbf{X} + g(\beta_1^\top\mathbf{X}) + \varepsilon$, where $g(\cdot)$ is a univariate unknown function, which again includes the partially linear single-index model (Carroll et al. 1997; Yu and Ruppert 2002) as its special case. Note that the partially linear single-index model (Carroll et al. 1997) takes a slightly different form: $Y = \mathbf{a}^\top\mathbf{X}_1 + g(\mathbf{b}^\top\mathbf{X}_2) + \varepsilon$, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, in which the linear variables $\mathbf{X}_1$ and nonlinear variables $\mathbf{X}_2$ need to be specified before modeling. In contrast, model (B) allows the data to choose the linear component and the nonlinear components and is, thus, more data driven.

It is easy to see that model (B) is not unique. For example, we can rewrite $G(\mathbf{B}_0^\top\mathbf{X}) := G(\mathbf{B}_0^\top\mathbf{X}) - \mathbf{c}^\top\mathbf{B}_0^\top\mathbf{X}$ and $\gamma_0 := \gamma_0 + \mathbf{B}_0\mathbf{c}$ for any $q \times 1$ vector $\mathbf{c}$. There are several methods to make the model identifiable that *do not impose any restriction to the model flexibility*. The next proposition presents one of those methods.

Proposition 1.1 (Identification). Suppose $\mathbf{X}$ has a continuous density function with nonsingular covariance matrix $\boldsymbol{\Sigma}_0$. The gradient $\nabla M(\mathbf{x}) = \partial M(\mathbf{x})/\partial\mathbf{x}$ and $\boldsymbol{\Omega} = E[\{\nabla M(\mathbf{X}) - E\nabla M(\mathbf{X})\}\{\nabla M(\mathbf{X}) - E\nabla M(\mathbf{X})\}^\top]$ exist. If $\boldsymbol{\Omega}$ has d_B nonzero eigenvalues, then we have

(I) Model (B) can always be rewritten in such a way that (i1) $\mathbf{B}_0^\top\boldsymbol{\Sigma}_0\mathbf{B}_0 = \mathbf{I}_{d_B}$ and the first nonzero element in each column of $\mathbf{B}_0$ is positive; (i2) $\gamma_0^\top\boldsymbol{\Sigma}_0\mathbf{B}_0 = 0$; and (i3) matrix $\boldsymbol{\Omega}_0 \overset{\text{def}}{=} E[\{\nabla G(\mathbf{B}_0^\top\mathbf{X}) - E\nabla G(\mathbf{B}_0^\top\mathbf{X})\}\{\nabla G(\mathbf{B}_0^\top\mathbf{X}) - E\nabla G(\mathbf{B}_0^\top\mathbf{X})\}^\top]$ is a full ranked diagonal matrix, where $\nabla G(\mathbf{u}) = \partial G(\mathbf{u})/\partial\mathbf{u}$.

(II) If model (B) satisfying (i1)–(i3) also holds with γ_0, $\mathbf{B}_0$, and G replaced by $\bar\gamma_0$, $\bar{\mathbf{B}}_0$, and $\bar G$, respectively, then there exists a rotation matrix $\mathbf{Q}: d_B \times d_B$ such that

$$\bar\gamma_0 = \gamma_0, \qquad \bar{\mathbf{B}}_0 = \mathbf{B}_0\mathbf{Q}^\top, \qquad \bar G(\mathbf{u}) = G(\mathbf{Qu}).$$

Furthermore, if the nonzero eigenvalues of $\boldsymbol{\Omega}$ differ from one another, then $\mathbf{Q} = \mathbf{I}_{d_B}$.

Yingcun Xia is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546 (E-mail: *staxyc@nus.edu.sg*). The author thanks an associate editor and two referees for their valuable comments. This work was supported by NUS FRG R-155-000-063-112, the National University of Singapore, and the National Natural Science Foundation of China (grant 10471061).

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Theory and Methods
DOI 10.1198/016214508000000805

Throughout this article we assume that model (B) is rewritten according to (i1)–(i3), which, as pointed out in Proposition 1.1, *do not represent any restriction.*

Model (B) identifies the linear component and the nonlinear components in a data-driven way. This identification is itself important in statistics and in many other disciplines of science in order to understand different mechanisms between the linear and nonlinear factors. See, for example, Grenfell, Finkenstädt, Wilson, Coulson, and Crawley (2000), Gao and Tong (2004), Samarov et al. (2005), and Maestri et al. (2005). The well-known partially linear model $Y = \beta_1^\top \mathbf{X}_1 + G(\mathbf{X}_2) + \varepsilon$ might be the first semiparametric model that tried to separate the linear and nonlinear components; see Speckman (1988) and Wang (2003). The partially linear single-index model (Carroll et al. 1997) is one of the recent models with a similar motivation. For these models, however, the selection of linear and nonlinear components is rather arbitrary. Tong (1990) used plots of the response against each covariate to identify linear components and nonlinear components in a time series setting. More recently, Samarov et al. (2005) tried to identify the linear and nonlinear variables using nonparametric kernel smoothing.

Another advantage of separating the linear components and nonlinear components as in model (B) is for model interpretation. It is known that model (A), or the general dimension reduction model (Li 1991), is sometimes not so informative and is hard to visualize when $d_A > 1$. In contrast, for model (B) the plot of the response against the linear component has clear meaning. On the other hand, because the dimension in the nonlinear part is further reduced by 1, its interpretation is relatively easier. One real dataset will be employed to demonstrate this point later.

In terms of estimation, because the structure of the dimension reduction directions are specified in model (B), it is expected that the estimation efficiency can be significantly improved. Note that a naive estimation, which applies a linear regression to the response first and then dimension reduction to the fitted residuals of the linear regression, cannot guarantee the consistency. As an example, consider $\mathbf{X} = (X_1, X_2, X_3)^\top$, where $X_k = \eta_k + \eta_0$ with $\eta_0, \eta_1, \eta_2, \eta_3 \overset{\text{iid}}{\sim} \text{Uniform}(-.5, .5)$, and model $Y = (X_1 + X_2)^3 + \varepsilon$. The linear regressor is $L(\mathbf{X}) = \beta_0 + \beta^\top \mathbf{X}$, where

$$\begin{aligned}(\beta_0, \beta) &= \operatorname*{arg\,min}_{\beta_0, \beta} E\{Y - \beta_0 - \beta^\top \mathbf{X}\}^2 \\ &= (0, 1.25, 1.25, -.15)^\top.\end{aligned}$$

The residual $E(Y|\mathbf{X}) - L(\mathbf{X}) = (X_1 + X_2)^3 - 1.25(X_1 + X_2) + .15X_3$ has a two-dimensional dimension reduction space. Instead, the true nonlinear part in the model is one dimensional. Thus, the method cannot estimate the model consistently. Other relevant estimation methods such as those of Härdle and Stoker (1989), Li (1991), Samarov (1993), Newey and Stoker (1993), Hristache et al. (2001), Yin and Cook (2002), Xia et al. (2002), Donkers and Schafgans (2003), and Banerjee (2007) cannot be applied directly to model (B). In this article we propose a method to estimate model (B) directly. We also develop a criterion to determine the dimension d_B and to choose between model (A) and model (B). The asymptotic theories for the estimation and model selection are obtained. An algorithm for the model estimation is also proved to be convergent.

2. INITIAL ESTIMATORS

Because the dimension d_B is unknown, we can try a working dimension q: $1 \le q \le p$; the choice of q will be discussed later. Recall that $M(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. By model (B), we have $M(\mathbf{x}) = \gamma_0^\top \mathbf{x} + G(\mathbf{B}_0^\top \mathbf{x})$ and, thus,

$$\nabla M(\mathbf{x}) \overset{\text{def}}{=} \frac{\partial M(\mathbf{x})}{\partial \mathbf{x}} = \gamma_0 + \mathbf{B}_0 \nabla G(\mathbf{B}_0^\top \mathbf{x}). \tag{1}$$

It follows that $\nabla M(\mathbf{X}) - E\nabla M(\mathbf{X}) = \mathbf{B}_0\{\nabla G(\mathbf{B}_0^\top \mathbf{X}) - E\nabla G(\mathbf{B}_0^\top \mathbf{X})\}$. The average of its outer product is

$$\begin{aligned}\mathbf{\Omega} &\overset{\text{def}}{=} E\big[\{\nabla M(\mathbf{X}) - E\nabla M(\mathbf{X})\}\{\nabla M(\mathbf{X}) - E\nabla M(\mathbf{X})\}^\top\big] \\ &= \mathbf{B}_0 \mathbf{\Omega}_0 \mathbf{B}_0^\top,\end{aligned} \tag{2}$$

where $\mathbf{\Omega}_0$ is defined in Proposition 1.1. Therefore, $\mathbf{B}_0$ can be estimated by the eigenvectors of $\mathbf{\Omega}$. By (1), we have $E\{\nabla M(\mathbf{X})\} = \gamma_0 + \mathbf{B}_0 E\{\nabla G(\mathbf{B}_0^\top \mathbf{X})\}$. Because the model is written according to the identification format, $(\mathbf{I} - \mathbf{\Sigma}_0^{1/2} \mathbf{B}_0 \mathbf{B}_0^\top \times \mathbf{\Sigma}_0^{1/2})$ is a projection matrix, and $(\mathbf{I} - \mathbf{\Sigma}_0^{1/2} \mathbf{B}_0 \mathbf{B}_0^\top \mathbf{\Sigma}_0^{1/2}) \mathbf{\Sigma}_0^{1/2} \times E\{\nabla M(\mathbf{X})\} = \mathbf{\Sigma}_0^{1/2} \gamma_0$. Thus, γ_0 can be obtained easily by

$$\gamma_0 = \mathbf{\Sigma}_0^{-1/2} \big(\mathbf{I} - \mathbf{\Sigma}_0^{1/2} \mathbf{B}_0 \mathbf{B}_0^\top \mathbf{\Sigma}_0^{1/2}\big) \mathbf{\Sigma}_0^{1/2} E\{\nabla M(\mathbf{X})\}. \tag{3}$$

To implement the previous idea, one can use multivariate local polynomial smoothing to estimate the gradients (Fan and Gijbels 1996). Here we consider local linear smoothing only, while a higher order polynomial may not be preferred due to its instability. The details can be stated as follows. Consider a multiple-density kernel function $K(u_1, \ldots, u_p)$ and bandwidth $(h_1, \ldots, h_p)$. For simplicity of bandwidth selection and ease of exposition, after the following standardization, a common bandwidth is used for all variables, that is, $h_1 = \cdots = h_p = h$. Let $K_h(\mathbf{u}) = h^{-p} K(\mathbf{u}/h)$, where $\mathbf{u} = (v_1, \ldots, v_p)^\top$. Suppose that $\{(\mathbf{X}_i, Y_i), i = 1, 2, \ldots, n\}$ is a random sample from $(\mathbf{X}, Y)$. Let $\mathbf{S}_\mathbf{X} = n^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$, with $\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{X}_i$. Standardize $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ by setting

$$\tilde{\mathbf{X}}_i := \mathbf{S}_\mathbf{X}^{-1/2} (\mathbf{X}_i - \bar{\mathbf{X}}). \tag{4}$$

For any $\mathbf{x}$, the principle of the local linear smoother suggests minimizing

$$n^{-1} \sum_{i=1}^n \{Y_i - a - \mathbf{b}^\top \tilde{\mathbf{X}}_{i\mathbf{x}}\}^2 K_h(\tilde{\mathbf{X}}_{i\mathbf{x}}), \tag{5}$$

with respect to a and $\mathbf{b}$ to estimate $M(\mathbf{x})$ and $\partial M(\mathbf{x})/\partial \mathbf{x}$, respectively, where $\tilde{\mathbf{X}}_{i\mathbf{x}} = \tilde{\mathbf{X}}_i - \mathbf{x}$. Denote the solution of $(a, \mathbf{b})$ in (5) at $\mathbf{x} = \tilde{\mathbf{X}}_j$ by $(\hat{a}_j, \hat{\mathbf{b}}_j)$.

Let $\bar{\mathbf{b}} = \sum_{j=1}^n \rho(\hat{f}(\tilde{\mathbf{X}}_j)) \hat{\mathbf{b}}_j / \sum_{j=1}^n \rho(\hat{f}(\tilde{\mathbf{X}}_j))$, where $\hat{f}(\mathbf{x}) = n^{-1} \sum_{i=1}^n K_h(\tilde{\mathbf{X}}_{i\mathbf{x}})$ and $\rho(v)$ is a trimming function introduced for technical purposes to handle the notorious boundary points. The trimming function $\rho(\cdot)$ is any bounded function with bounded second-order derivatives on $\mathcal{R}$ such that $\rho(v) > 0$ if $v > \omega_0 > 0$; $\rho(v) = 0$ if $v \le \omega_0$. One example of the trimming function is $\rho(v) = 1$ if $v \ge \omega_0'$; $\exp(-(v - \omega_0)^{-1} + (\omega_0' - v)^{-1}) / \{1 + \exp(-(v - \omega_0)^{-1} + (\omega_0' - v)^{-1})\}$ if $\omega_0' > v > \omega_0$;

0 if $v \le \omega_0$. In practice, ω_0' can be very small, and, thus, no observations are trimmed off. Analogous to $\mathbf{\Omega}$ in (2), we consider an average of outer products of $\hat{\mathbf{b}}_j$'s,

$$\hat{\mathbf{\Omega}} = n^{-1}\sum_{j=1}^{n}\rho(\hat{f}(\tilde{\mathbf{X}}_j))\{\hat{\mathbf{b}}_j - \bar{\mathbf{b}}\}\{\hat{\mathbf{b}}_j - \bar{\mathbf{b}}\}^\top.$$

Calculate the first q eigenvectors of $\hat{\mathbf{\Omega}}$. In accordance with the requirements in Proposition 1.1, each eigenvector is written such that its first nonzero element is positive. Denote them by $\tilde{\mathbf{B}}_{(1)}$. The estimators of $\mathbf{B}_0$ and γ_0 are then, respectively,

$$\hat{\mathbf{B}}_{(1)} = \mathbf{S}_{\mathbf{X}}^{-1/2}\tilde{\mathbf{B}}_{(1)}, \qquad \hat{\gamma}_{(1)} = \mathbf{S}_{\mathbf{X}}^{-1/2}\big(\mathbf{I} - \tilde{\mathbf{B}}_{(1)}\tilde{\mathbf{B}}_{(1)}^\top\big)\bar{\mathbf{b}}.$$

Note that multiple kernel smoothing is used above, which is known to be inefficient. To improve the efficiency, one can employ the idea of an adaptive kernel as in Hristache et al. (2001) and Xia et al. (2002). However, the estimation combining the two ideas is not as efficient as the method introduced in Xia et al. (2002); see the discussion in Xia (2006).

3. REFINED ESTIMATORS

If model (B) holds, then the gradients $\partial G(\mathbf{B}_0^\top\mathbf{x})/\partial\mathbf{x}$ for all $\mathbf{x}$ are in a common d_B-dimensional subspace, $\{\mathbf{B}_0\mathbf{u} : \mathbf{u} \in \mathcal{R}^{d_B}\}$, as shown in (1). To use this information, we can replace $\mathbf{b}$ in (5), which is an estimate of the gradient, by $\mathbf{Bd}(\mathbf{x})$ in order to be in line with the local linear approximation $M(\mathbf{X}_i) \approx \gamma_0^\top\mathbf{X}_i + G(\mathbf{B}_0^\top\mathbf{x}) + \{\mathbf{B}_0\nabla G(\mathbf{B}_0^\top\mathbf{x})\}^\top\mathbf{X}_{i\mathbf{x}}$ for $\mathbf{X}_i$ close to $\mathbf{x}$, where $\mathbf{X}_{i\mathbf{x}} = \mathbf{X}_i - \mathbf{x}$. Because $G(\mathbf{B}_0^\top\mathbf{x})$ and $\nabla^\top G(\mathbf{B}_0^\top\mathbf{x})$ are unknown, they can be estimated by minimizing the local linear approximation error

$$n^{-1}\sum_{i=1}^{n}\{Y_i - \gamma^\top\mathbf{X}_i - a - \mathbf{d}^\top\mathbf{B}^\top\mathbf{X}_{i\mathbf{x}}\}^2 K_h(\mathbf{X}_{i\mathbf{x}}),$$

with respect to a and $\mathbf{d}$. Because $(\gamma, \mathbf{B})$ is common for all $\mathbf{x}$, it should be estimated by minimizing the approximation errors for all $\mathbf{x} = \mathbf{X}_j$, $j = 1, \dots, n$. As a consequence, we propose to estimate $(\gamma_0, \mathbf{B}_0)$ by minimizing

$$n^{-2}\sum_{j=1}^{n}\mathcal{I}_{nj}\rho_j\sum_{i=1}^{n}\{Y_i - \gamma^\top\mathbf{X}_i - a_j - \mathbf{d}_j^\top\mathbf{B}^\top\mathbf{X}_{ij}\}^2 w_{ij}, \qquad (6)$$

with respect to $\gamma, a_j, \mathbf{d}_j = (d_{j1}, \dots, d_{jq})^\top$, $j = 1, \dots, n$, and $\mathbf{B} : \mathbf{B}^\top\mathbf{\Sigma}_0\mathbf{B} = \mathbf{I}_q$, where $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j$. The weight function w_{ij} should be adaptive to the structure, that is, $w_{ij} = K_h(\mathbf{B}^\top\mathbf{X}_{ij})$. Here we abuse the notation and say $K(\cdot)$ is the same function defined previously but with p replaced by q. The trimming function is $\rho_j = \rho(\hat{f}_{\mathbf{B}}(\mathbf{X}_j))/\hat{f}_{\mathbf{B}}(\mathbf{X}_j)$, with $\hat{f}_{\mathbf{B}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_h(\mathbf{B}^\top\mathbf{X}_{i\mathbf{x}})$. Note that we introduce another function $\mathcal{I}_{nj} = 1(|\mathbf{X}_j| \le n)$. Hereafter, for any matrix $\mathbf{A}$, $|\mathbf{A}|$ denotes its largest singular value, which is the Euclidean norm if $\mathbf{A}$ is a vector. It is easy to see that, under some mild conditions, the observations trimmed off by $\mathcal{I}_{nj}$ are negligible. These two trimming functions are used here for technical purposes. The preceding estimation procedure is similar to the minimum average (conditional) variance estimation method (Xia et al. 2002).

The minimization problem in (6) can be solved by fixing $(a_j, \mathbf{d}_j)$, $j = 1, \dots, n$, and fixing $(\gamma, \mathbf{B})$ alternatively. As a consequence, the minimization can be decomposed into two quadratic programming problems both of which have simple analytic solutions. For any matrix $\mathbf{B} = (\beta_1, \dots, \beta_q)$, define operators $\ell(\cdot)$ and $\mathcal{M}(\cdot)$, respectively, as

$$\ell(\mathbf{B}) = (\beta_1^\top, \dots, \beta_q^\top)^\top \qquad \text{and} \qquad \mathcal{M}(\ell(\mathbf{B})) = \mathbf{B}.$$

The following algorithm implements the estimation.

Step 0 (Initializing). Calculate $\mathbf{S}_{\mathbf{X}}$ and standardize $\mathbf{X}_i$ to $\tilde{\mathbf{X}}_i$ as in (4). Let $\tilde{\mathbf{B}}_{(1)}$ be the matrix calculated in Section 2 and let $\tilde{\gamma}_{(1)} = (\mathbf{I} - \tilde{\mathbf{B}}_{(1)}\tilde{\mathbf{B}}_{(1)}^\top)\bar{\mathbf{b}}$. Set $t = 1$, $\mathbf{B}_{(1,0)} = \tilde{\mathbf{B}}_{(1)}$, and $\gamma_{(1,0)} = \tilde{\gamma}_{(1)}$. Let h_0 be the bandwidth used in the initial estimation.

Step I (Outer loop). Set $\tau = 0$ and $h_t = \max\{c_n h_{t-1}, \hbar_t\}$, where $1 > c_n \gg n^{-1/(p+4)}$ and $\hbar_t$ is another bandwidth discussed later.

Step I.1 (Inner loop). Fix $\mathbf{B} = \mathbf{B}_{(t,\tau)}$ and $\gamma = \gamma_{(t,\tau)}$ and calculate the solutions of $(a_j, \mathbf{d}_j)$, $j = 1, \dots, n$, to the minimization problem in (6):

$$\begin{pmatrix} a_j^{(t,\tau)} \\ \mathbf{d}_j^{(t,\tau)} h_t \end{pmatrix} = \left\{\sum_{i=1}^{n} K_{h_t}\big(\mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}\big)\begin{pmatrix} 1 \\ \mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}/h_t \end{pmatrix} \times \begin{pmatrix} 1 \\ \mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}/h_t \end{pmatrix}^\top\right\}^{-1} \sum_{i=1}^{n} K_{h_t}\big(\mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}\big) \times \begin{pmatrix} 1 \\ \mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}/h_t \end{pmatrix}\{Y_i - \tilde{\mathbf{X}}_i^\top\gamma_{(t,\tau)}\}.$$

Step I.2 (Inner loop). Let

$$\hat{f}_{\mathbf{B}_{(t,\tau)}}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_{h_t}\big(\mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{i\mathbf{x}}\big) \qquad \text{and}$$

$$\rho_j^{(t,\tau)} = \rho\big(\hat{f}_{\mathbf{B}_{(t,\tau)}}(\tilde{\mathbf{X}}_j)\big)/\hat{f}_{\mathbf{B}_{(t,\tau)}}(\tilde{\mathbf{X}}_j).$$

Fix $a_j = a_j^{(t,\tau)}$ and $\mathbf{d}_j = \mathbf{d}_j^{(t,\tau)}$ and calculate the solution of $(\gamma, \mathbf{B})$ or $(\gamma, \ell(\mathbf{B}))$ in (6):

$$\begin{pmatrix} \gamma^{(t,\tau+1)} \\ \mathbf{b}^{(t,\tau+1)} \end{pmatrix} = \left\{\sum_{j,i=1}^{n}\mathcal{I}_{nj}\rho_j^{(t,\tau)} K_{h_t}\big(\mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}\big) \times \begin{pmatrix} \tilde{\mathbf{X}}_i \\ \tilde{\mathbf{X}}_{ij}^{(t,\tau)} \end{pmatrix}\begin{pmatrix} \tilde{\mathbf{X}}_i \\ \tilde{\mathbf{X}}_{ij}^{(t,\tau)} \end{pmatrix}^\top\right\}^{-1} \times \sum_{j,i=1}^{n}\mathcal{I}_{nj}\rho_j^{(t,\tau)} K_{h_t}\big(\mathbf{B}_{(t,\tau)}^\top\tilde{\mathbf{X}}_{ij}\big) \times \begin{pmatrix} \tilde{\mathbf{X}}_i \\ \tilde{\mathbf{X}}_{ij}^{(t,\tau)} \end{pmatrix}\{Y_i - a_j^{(t,\tau)}\},$$

where $\tilde{\mathbf{X}}_{ij}^{(t,\tau)} = \mathbf{d}_j^{(t,\tau)} \otimes \tilde{\mathbf{X}}_{ij}$.

Step I.3 (Inner loop). Calculate

$$\Lambda_{(t,\tau+1)} = \big\{\mathcal{M}\big(\mathbf{b}^{(t,\tau+1)}\big)\big\}^\top\mathcal{M}\big(\mathbf{b}^{(t,\tau+1)}\big) \quad \text{and}$$

$$\mathbf{B}_{(t,\tau+1)} = \mathcal{M}\big(\mathbf{b}^{(t,\tau+1)}\big)\Lambda_{(t,\tau+1)}^{-1/2};$$

$$\gamma_{(t,\tau+1)} = \big(\mathbf{I} - \mathbf{B}_{(t,\tau+1)}\mathbf{B}_{(t,\tau+1)}^\top\big)\gamma^{(t,\tau+1)}.$$

If a convergence criterion is satisfied, stop; otherwise, set $\tau := \tau + 1$ and go to step I.1.

Step II (Outer loop). Repeat steps I.1–I.3 until convergence. Let $(\gamma_{(t+1,0)}, \mathbf{B}_{(t+1,0)})$ be the final value of $(\gamma_{(t,\tau)}, \mathbf{B}_{(t,\tau)})$. If a convergence criterion is met, stop; otherwise, set $t := t + 1$ and go to step I.

Let $(\tilde{\gamma}, \tilde{\mathbf{B}})$ be the final value of $(\gamma_{(t,0)}, \mathbf{B}_{(t,0)})$. Then the final estimators are $\hat{\gamma} = \mathbf{S}_{\mathbf{X}}^{-1/2}\tilde{\gamma}$ and $\hat{\mathbf{B}} = \mathbf{S}_{\mathbf{X}}^{-1/2}\tilde{\mathbf{B}}$, respectively. In the calculation, as usual the convergence is regarded as being achieved whenever the changes of $(\gamma_{(t,\tau)}, \mathbf{B}_{(t,\tau)})$ are very small, for example, $|\gamma_{(t,\tau+1)} - \gamma_{(t,\tau)}| + |\mathbf{B}_{(t,\tau+1)}\mathbf{B}_{(t,\tau+1)}^\top - \mathbf{B}_{(t,\tau)}\mathbf{B}_{(t,\tau)}^\top| < 10^{-6}$, in the last few iterations. The estimated link function $G(\mathbf{u})$ and the gradient are

$$\begin{pmatrix} \hat{G}(\mathbf{u}) \\ \widehat{\nabla G(\mathbf{u})}h \end{pmatrix} = \left\{ \sum_{i=1}^{n} K_h(\hat{\mathbf{B}}^\top \mathbf{X}_i - \mathbf{u}) \times \begin{pmatrix} 1 \\ (\hat{\mathbf{B}}^\top \mathbf{X}_i - \mathbf{u})/h \end{pmatrix} \begin{pmatrix} 1 \\ (\hat{\mathbf{B}}^\top \mathbf{X}_i - \mathbf{u})/h \end{pmatrix}^\top \right\}^{-1} \times \sum_{i=1}^{n} K_h(\hat{\mathbf{B}}^\top \mathbf{X}_i - \mathbf{u}) \begin{pmatrix} 1 \\ (\hat{\mathbf{B}}^\top \mathbf{X}_i - \mathbf{u})/h \end{pmatrix} (Y_i - \hat{\gamma}^\top \mathbf{X}_i),$$

where h is the bandwidth used in the last iteration of the algorithm.

Bandwidths $h_0, \hbar_t, t = 1, \ldots,$ need to be selected in the algorithm. After fixing $\gamma_{(t,0)}$ and $\mathbf{B}_{(t,0)}$, the bandwidths are actually selected for the estimation of $E(Y|\mathbf{X} = \mathbf{x})$ and $E(Y - \gamma_{(t,0)}^\top \mathbf{X}|\mathbf{B}_{(t,0)}^\top \mathbf{X} = \mathbf{u}), t = 1, 2, \ldots,$ respectively. Most existing bandwidth selection methods can be used. Details can be found in Fan and Gijbels (1996) and Yang and Tschernig (1999). In the following calculation $c_n = 1/1.5$, and the simple rule of thumb of Silverman (1986) is used to select the bandwidth; that is, after standardizing $\mathbf{X}_i$, we take $h_0 = n^{-1/(p+4)}$ and $\hbar_t = n^{-1/(q+4)}$ for all t.

4. MODEL SELECTION

One purpose of this article is to further specify a dimension reduction model such that it can be estimated better. Therefore, it is essential to validate the specification and select one model between (A) and (B). First, we need to find the dimension in model (A). There are a number of methods for this purpose; see, for example, Yin and Cook (2002, 2004) and Xia et al. (2002). Here the method in Xia et al. (2002) is used. The method is based on a semiparametric cross-validation (CV) criterion. Suppose for a working dimension q, the estimated directions for model (A) are $\hat{\mathbf{B}}_A = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$. The estimation details can be found in Xia et al. (2002). For simplicity, we use the delete-one-observation N–W estimator

$$\hat{m}_j^A(\hat{\mathbf{B}}_A^\top \mathbf{x}) = \sum_{i=1, i\neq j}^{n} K_{h_A}(\hat{\mathbf{B}}_A^\top \mathbf{X}_{i\mathbf{x}}) Y_i \Big/ \sum_{i=1, i\neq j}^{n} K_{h_A}(\hat{\mathbf{B}}_A^\top \mathbf{X}_{i\mathbf{x}}).$$

The CV value for model (A) is defined as

$$CVA(q) = n^{-1} \sum_{j=1}^{n} \rho(\hat{f}(\mathbf{X}_j))\{Y_j - \hat{m}_j^A(\hat{\mathbf{B}}_A^\top \mathbf{X}_j)\}^2.$$

Then the dimension for model (A) is selected as

$$\hat{d}_A = \arg\min_{q \geq 0} CVA(q).$$

Xia et al. (2002) proved that this selection is consistent; that is, $\hat{d}_A \to d_A$ in probability.

Suppose the estimator of $\mathbf{B}_0$ in model (B) is $\hat{\mathbf{B}}_B$ with working dimension q. Following the idea of Speckman (1988), calculate the leave-one-out estimator of γ_0 and $G(\cdot)$, respectively, by

$$\hat{\gamma}_j = \left\{ \sum_{i=1, i\neq j}^{n} (\mathbf{X}_i - \hat{m}_j^B(\mathbf{X}_i))(\mathbf{X}_i - \hat{m}_j^B(\mathbf{X}_i)) \right\}^{-1} \times \sum_{i=1, i\neq j}^{n} (\mathbf{X}_i - \hat{m}_j^B(\mathbf{X}_i))\{Y_i - \tilde{m}_j^B(\mathbf{X}_i)\}$$

and $\hat{G}_j(\mathbf{x}) = \tilde{m}_j^B(\mathbf{x}) - \hat{\gamma}_j^\top \hat{m}_j^B(\mathbf{x})$, where $\hat{m}_j^B(\mathbf{x}) = \{n \hat{f}_{\hat{\mathbf{B}}_B, j}(\mathbf{x})\}^{-1} \sum_{i=1, i\neq j}^{n} K_{h_B}(\hat{\mathbf{B}}_B^\top \mathbf{X}_{i\mathbf{x}})\mathbf{X}_i$, $\tilde{m}_j^B(\mathbf{x}) = \{n \times \hat{f}_{\hat{\mathbf{B}}_B, j}(\mathbf{x})\}^{-1} \sum_{i=1, i\neq j}^{n} K_{h_B}(\hat{\mathbf{B}}_B^\top \mathbf{X}_{i\mathbf{x}})Y_i$, and $\hat{f}_{\hat{\mathbf{B}}_B, j}(\mathbf{x}) = n^{-1} \times \sum_{i=1, i\neq j}^{n} K_{h_B}(\hat{\mathbf{B}}_B^\top \mathbf{X}_{i\mathbf{x}})$. The CV value for model (B) is defined as

$$CVB(q) = n^{-1} \sum_{j=1}^{n} \rho(\hat{f}(\mathbf{X}_j))\{Y_j - \hat{\gamma}_j^\top \mathbf{X}_j - \hat{G}_j(\hat{\mathbf{B}}^\top \mathbf{X}_j)\}^2.$$

Note that the same trimming function $\rho(\hat{f}(\mathbf{X}_j))$ is used for the two models in order to compare their CV values.

Finally, our selection criterion is as follows.

If $CVB(\hat{d}_A - 1) < CVA(\hat{d}_A)$, then we select model (B)*; otherwise model* (A).

The dimension of model (B) is estimated by

$$\hat{d}_B = \begin{cases} \hat{d}_A - 1 & \text{if } CVB(\hat{d}_A - 1) \leq CVA(\hat{d}_A) \\ \hat{d}_B = \hat{d}_A & \text{otherwise.} \end{cases}$$

Equivalently, the linear component in model (B) exists if $CVB(\hat{d}_A - 1) \leq CVA(\hat{d}_A)$, and the nonlinear components exist if $\hat{d}_B \geq 1$.

As noticed by an anonymous referee, there are other alternatives for the selection of dimension and models. For example, one can select $\hat{d}_B = \arg\min_{q\geq 0} CVB(q)$ and select model (B) if $CVB(\hat{d}_B) < CVA(\hat{d}_B + 1)$. One can also select $\hat{d}_A = \arg\min_{q\geq 0} CVA(q)$ and $\hat{d}_B = \arg\min_{q\geq 0} CVB(q)$ and select model (B) if $CVB(\hat{d}_B) < CVA(\hat{d}_A)$. Simulations suggest that these methods also work with about the same efficiency as the preceding method.

5. ASYMPTOTICS

Some asymptotic properties are presented in this section. Their proofs can be obtained from the author on request or downloaded from the supplemental materials website at *http://www.amstat.org/publications/jasa/supplemental_materials*. For any $\mathbf{B} : p \times q$ such that $\mathbf{B}^\top \mathbf{\Sigma}_0 \mathbf{B} = \mathbf{I}_q$, let $f_{\mathbf{B}}(\mathbf{u})$ be the density function of $\mathbf{B}^\top \mathbf{X}$, $\mu_{\mathbf{B}}(\mathbf{u}) = E(\mathbf{X}|\mathbf{B}^\top \mathbf{X} = \mathbf{u})$, and $w_{\mathbf{B}}(\mathbf{u}) = E(\mathbf{X}\mathbf{X}^\top|\mathbf{B}^\top \mathbf{X} = \mathbf{u})$. For ease of exposition, denote $\mu_{\mathbf{B}}(\mathbf{B}^\top \mathbf{x})$ and $f_{\mathbf{B}}(\mathbf{B}^\top \mathbf{x})$ by $\mu_{\mathbf{B}}(\mathbf{x})$ and $f_{\mathbf{B}}(\mathbf{x})$, respectively. Let $\nu_{\mathbf{B}}(\mathbf{x}) = \mathbf{x} - \mu_{\mathbf{B}}(\mathbf{x})$. For ease of exposition, we assume that $\mu_0 = \int K(v_1, \ldots, v_q)\, dv_1 \cdots dv_q = 1$ and $\mu_{2,k} =$

$\int K(v_1, \dots, v_q) v_k^2 \, dv_1 \cdots dv_q = 1$ for $k = 1, \dots, q$; otherwise, redefine $K(v_1, \dots, v_q) := \mu_0^{-1} K(v_1/\sqrt{\mu_{2,1}}, \dots, v_q/\sqrt{\mu_{2,q}})/\sqrt{\mu_{2,1} \cdots \mu_{2,q}}$. For any square matrix $\mathbf{A}$, denote its Moore–Penrose inverse matrix by $\mathbf{A}^+$.

Proposition 5.1 (Convergence of the algorithm). Suppose assumptions (C1)–(C6) in the Appendix hold and let $q = d_A$. Denote the estimators at outer loop t and inner loop τ by $(\mathbf{B}_{(t,\tau)}, \gamma_{(t,\tau)})$. Then, for any $t > 1$ and $\tau \geq 1$, there is a rotation matrix $\mathbf{Q}$ such that

$$\begin{aligned}
&\begin{pmatrix} \gamma_{(t,\tau+1)} - \gamma_0 \\ \ell(\mathbf{B}_{(t,\tau+1)} - \mathbf{B}_0\mathbf{Q}) \end{pmatrix} \\
&\quad = \{\tilde{\mathbf{D}}_0^+ \mathbf{D}_0 + o(1)\} \begin{pmatrix} \gamma_{(t,\tau)} - \gamma_0 \\ \ell(\mathbf{B}_{(t,\tau)} - \mathbf{B}_0\mathbf{Q}) \end{pmatrix} + \tilde{\mathbf{D}}_0^+ \Phi_n \\
&\qquad + O\left(h_t^4 + \frac{\log n}{n h_t^q}\right) + o(n^{-1/2})
\end{aligned}$$

almost surely, where $\tilde{\mathbf{D}}_0$ and $\mathbf{D}_0$ are two semipositive definite matrices (details can be found in the proof). Moreover, the eigenvalues of $\tilde{\mathbf{D}}_0^+ \mathbf{D}_0$ are all smaller than 1; that is, the algorithm is contractive.

Remark 5.1. It is known that the basis of a linear space is not unique. The estimator $\hat{\mathbf{B}}$ only converges to one of the bases. If the nonzero eigenvalues of $\mathbf{M}_0$ in the Appendix differ from one another and the model is written according to (i1)–(i3), then $\mathbf{Q} = \mathbf{I}_{d_B}$. Proposition 5.1 only indicates the convergence of the inner loop. For the outer loop, because $h_t \to h$ as $t \to \infty$, its convergence is also implied by the proposition.

Theorem 5.1 (Consistency of the estimators). Suppose assumptions (C1)–(C6) in the Appendix hold. If $q = d_B$ and the final bandwidth is h_B, then there is a rotation matrix $\mathbf{Q}$ such that

$$|\hat{\gamma} - \gamma_0| + |\hat{\mathbf{B}} - \mathbf{B}_0\mathbf{Q}| = O\left(h_B^4 + \frac{\log n}{n h_B^{d_B}} + n^{1/2}\right)$$

in probability.

Remark 5.2. Note that, for model (A), the consistency rate for the directions is $O_P\{h_A^4 + \log n (n h_A^{d_A})^{-1} + n^{1/2}\}$. Because $d_A = d_B + 1$, the consistency rate for the estimation of model (B) can be faster than that for model (A). In other words, identifying model (B) correctly can indeed improve the estimation efficiency.

Corollary 5.1 (Asymptotic distributions). Suppose assumptions (C1)–(C6) in the Appendix hold, $q = d_B$, and the nonzero eigenvalues of $\mathbf{M}_0$ differ from one another. Suppose the model is written according to (i1)–(i3). If the density function of $\mathbf{B}_0^\top \mathbf{X}$ is positive at $\mathbf{u} = (v_1, \dots, v_{d_B})^\top$, then we have

$$\begin{aligned}
&(n h_B^{d_B})^{1/2} \left\{ \hat{G}(\mathbf{u}) - G(\mathbf{u}) - \frac{1}{2} \sum_{\iota=1}^{d_B} \nabla_{\iota,\iota}^2 G(\mathbf{u}) h_B^2 \right\} \\
&\qquad \xrightarrow{D} \mathrm{N}\left\{ 0, \frac{\sigma^2(\mathbf{u})}{f_{\mathbf{B}_0}(\mathbf{u})} \int K^2(\mathbf{u}) \, dv_1 \cdots dv_{d_B} \right\},
\end{aligned}$$

where $\sigma^2(\mathbf{u}) = E(\varepsilon^2 | \mathbf{B}_0^\top \mathbf{X} = \mathbf{u})$. Furthermore, if $d_B \leq 3$ then we have

$$\begin{aligned}
&\sqrt{n} \begin{pmatrix} \hat{\gamma} - \gamma_0 \\ \ell(\hat{\mathbf{B}} - \mathbf{B}_0) \end{pmatrix} \\
&\quad \xrightarrow{D} \mathrm{N}\Bigg[0, \begin{pmatrix} \mathbf{I}_p & -\mathbf{B}_0(\mathbf{I}_{d_B} \otimes \gamma_0^\top) \\ 0 & \mathbf{I}_{pd_B} \end{pmatrix} \\
&\qquad \times \mathbf{W}_0^+ \mathbf{W}_2 \mathbf{W}_0^+ \begin{pmatrix} \mathbf{I}_p & -\mathbf{B}_0(\mathbf{I}_{d_B} \otimes \gamma_0^\top) \\ 0 & \mathbf{I}_{pd_B} \end{pmatrix}^\top \Bigg],
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{W}_0 = E\Bigg\{ &\rho(f_{\mathbf{B}_0}(\mathbf{X})) \begin{pmatrix} \nu_{\mathbf{B}_0}(\mathbf{X}) \\ \nabla G(\mathbf{B}_0^\top \mathbf{X}) \otimes \nu_{\mathbf{B}_0}(\mathbf{X}) \end{pmatrix} \\
&\times \begin{pmatrix} \nu_{\mathbf{B}_0}(\mathbf{X}) \\ \nabla G(\mathbf{B}_0^\top \mathbf{X}) \otimes \nu_{\mathbf{B}_0}(\mathbf{X}) \end{pmatrix}^\top \Bigg\}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{W}_2 = E\Bigg\{ &\rho^2(f_{\mathbf{B}_0}(\mathbf{X})) \begin{pmatrix} \nu_{\mathbf{B}_0}(\mathbf{X}) \\ \nabla G(\mathbf{B}_0^\top \mathbf{X}) \otimes \nu_{\mathbf{B}_0}(\mathbf{X}) \end{pmatrix} \\
&\times \begin{pmatrix} \nu_{\mathbf{B}_0}(\mathbf{X}) \\ \nabla G(\mathbf{B}_0^\top \mathbf{X}) \otimes \nu_{\mathbf{B}_0}(\mathbf{X}) \end{pmatrix}^\top \varepsilon^2 \Bigg\}.
\end{aligned}$$

Remark 5.3. Corollary 5.1 indicates that, if $d_B \leq 3$, then the root-n consistency for the estimators of the parameters can be achieved. If higher order local polynomial smoothing is used, the root-n consistency can also be achieved for $d_B > 3$. However, the model with $d_B > 3$ is not attractive in practice because of the "curse of dimensionality." Instead, $d_B = 1$ is more appealing. In this case, the asymptotic distribution is similar to that in Carroll et al. (1997), where the variance and covariance matrix is $\mathbf{W}_0^+ \mathbf{W}_2 \mathbf{W}_0^+$. The difference in the variance–covariance matrices results from the identification requirement in our model.

Remark 5.4. To utilize the asymptotic distribution for statistical inference of the parameters, we need to estimate the variance–covariance matrix. Replace the values $G(\mathbf{B}_0^\top \mathbf{X}_j)$, $\nabla G(\mathbf{B}_0^\top \mathbf{X}_j)$, and $\nu_{\mathbf{B}_0}(\mathbf{X}_j)$, respectively, by G_j, ∇_j, and μ_j with

$$\begin{aligned}
&\begin{pmatrix} G_j \\ \nabla_j h_B \end{pmatrix} \\
&\quad = \left\{ \sum_{i=1}^n K_{h_B}(\hat{\mathbf{B}}^\top \mathbf{X}_{ij}) \begin{pmatrix} 1 \\ \hat{\mathbf{B}}^\top \mathbf{X}_{ij}/h_B \end{pmatrix} \begin{pmatrix} 1 \\ \hat{\mathbf{B}}^\top \mathbf{X}_{ij}/h \end{pmatrix}^\top \right\}^{-1} \\
&\qquad \times \sum_{i=1}^n K_{h_B}(\hat{\mathbf{B}}^\top \mathbf{X}_{ij}) \begin{pmatrix} 1 \\ \hat{\mathbf{B}}^\top \mathbf{X}_{ij}/h_B \end{pmatrix} \{Y_i - \mathbf{X}_i^\top \hat{\gamma}\}
\end{aligned}$$

and $\mu_j = \{n f_j\}^{-1} \sum_{i=1}^n K_{h_B}(\hat{\mathbf{B}}^\top \mathbf{X}_{ij}) \mathbf{X}_i$. Let $f_j = n^{-1} \times \sum_{i=1}^n K_{h_B}(\hat{\mathbf{B}}^\top \mathbf{X}_{ij})$ and $\nu_j = \mu_j - \mathbf{X}_j$. Then $\mathbf{W}_0$ and $\mathbf{W}_2$ can be estimated, respectively, by

$$\hat{\mathbf{W}}_0 = n^{-1} \sum_{j=1}^n \mathcal{I}_{nj} \rho(f_j) \begin{pmatrix} \nu_j \\ \nabla_j \otimes \nu_j \end{pmatrix} \begin{pmatrix} \nu_j \\ \nabla_j \otimes \nu_j \end{pmatrix}^\top$$

and

$$\hat{\mathbf{W}}_2 = n^{-1}\sum_{j=1}^{n}\mathcal{I}_{nj}\rho(f_j)\begin{pmatrix}\nu_j\\ \nabla_j\otimes\nu_j\end{pmatrix}\begin{pmatrix}\nu_j\\ \nabla_j\otimes\nu_j\end{pmatrix}^{\top} \times (Y_j - G_j)^2.$$

Finally, the variance–covariance matrix can be estimated by

$$\hat{\Xi} = \begin{pmatrix}\mathbf{I}_p & -\hat{\mathbf{B}}(\mathbf{I}_{d_B}\otimes\hat{\gamma}^{\top})\\ 0 & \mathbf{I}_{pd_B}\end{pmatrix}\hat{\mathbf{W}}_0^{+}\hat{\mathbf{W}}_2\hat{\mathbf{W}}_0^{+} \times \begin{pmatrix}\mathbf{I}_p & -\hat{\mathbf{B}}(\mathbf{I}_{d_B}\otimes\hat{\gamma}^{\top})\\ 0 & \mathbf{I}_{pd_B}\end{pmatrix}^{\top}.$$

It is easy to see that $\hat{\Xi}$ is a consistent estimator of the variance-covariance matrix in Corollary 5.1. The standard errors (S.E.) for elements in $(\hat{\gamma}, \hat{\mathbf{B}})$ are asymptotically the main diagonals of $\hat{\Xi}/\sqrt{n}$, respectively. These standard errors can be used for the statistical inference about the parameters separately; see Examples 6.2 and 6.3.

Theorem 5.2 (Consistency of model selection). Suppose assumptions (C1)–(C6) in the appendix hold. For every working dimension q, we use bandwidth $h \propto n^{-1/(q+4)}$. If model (B) is true, then we have $P\{CVB(\hat{d}_A - 1) < CVA(\hat{d}_A)\} \to 1$ as $n \to \infty$; otherwise, we have $P\{CVB(\hat{d}_A - 1) > CVA(\hat{d}_A)\} \to 1$. Moreover, the selected dimension is consistent, that is,

$$P(\hat{d}_B = d_B) \to 1 \quad \text{as } n \to \infty.$$

Theorem 5.2 indicates that our selection criterion for models (A) and (B) is consistent and that the selected dimension for model (B) is also consistent. Theorem 5.2 also implies that the estimation method searches for the dimension reduction space in the conditional mean exhaustively (Xia 2007).

6. NUMERICAL STUDIES

In the following calculations, we use the quadratic kernel $K(\mathbf{u}) = \mu_0^{-1}(1 - |\mathbf{u}|^2/\mu_2)^2 I(|\mathbf{u}|/\mu_2 < 1)/\mu_2$, where $\mu_0 = \int(1 - |\mathbf{u}|^2)I(|\mathbf{u}| < 1)\,d\mathbf{u}$ and $\mu_2 = \mu_0^{-1}\int(1 - |\mathbf{u}|^2)I(|\mathbf{u}| < 1)v_1^2\,d\mathbf{u}$. We use the trimming function in Section 2 with $\omega_0' = .01$. Thus, all observations have equal weight. The bandwidths by this rule of thumb are used for the selection of h_0 and $\hbar_t$, while $c_n = 1/1.5$; see Sliverman (1986) and Scott (1992). A computer code is available at *www.stat.nus.edu/˜ycxia/mim.m*. Besides comparing model (A) and the estimation method in Xia et al. (2002), we also consider the general dimension reduction model:

$$\text{Model (A}^*\text{):}\quad Y = m^*\big(\beta_1^{\top}\mathbf{X}, \ldots, \beta_{d_A^*}^{\top}\mathbf{X}, \varepsilon\big),$$

where ε is independent of $\mathbf{X}$; see Li (1991). There are a number of methods to estimate model (A*) and to select the dimension d_A^*. In the following, we mainly consider the sliced inverse regression (SIR) and the principal Hessian direction (pHd) methods; see Li (1991, 1992). For pHd applied to the residuals of a linear regression, the preceding model has a roughly similar form to that of model (B). Therefore, pHd must have better performance than the other inverse regression methods. A code in R (*http://www.r-project.org/*) for the inverse regression methods is used in the following calculations.

Example 6.1 (Simulated data). Let us first check the estimation consistency by model

$$Y = c\gamma_0^{\top}\mathbf{X} + \frac{\beta_1^{\top}\mathbf{X}}{.5 + (1.5 + \beta_2^{\top}\mathbf{X})^2} + \sigma\varepsilon,$$

where $\gamma_0 = (1, -1, 0, 0, 0, 0, 0, 0, -1, 1)^{\top}$, $\beta_1 = (0, 0, .5, .5, .5, .5, 0, 0, 0, 0)^{\top}$, and $\beta_2 = (0, 0, .5, -.5, .5, -.5, 0, 0, 0, 0)^{\top}$. The predictor $\mathbf{X} = (X_1, \ldots, X_{10})^{\top} = \boldsymbol{\Sigma}_0^{1/2}(\xi_1, \ldots, \xi_{10})^{\top}$, with $\xi_i, i = 1, \ldots, 10 \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$. The covariance matrix is $\boldsymbol{\Sigma}_0 = (.5^{|i-j|})_{1\le i,j\le 10}$. The constant c in the model is 1 or 0 depending on whether the linear part exists or not. If $c = 0$, the model was used by Li (1991). Note that the model is also a special case of models (A) and (A*) with $d_A = d_A^* = 3$ when $c = 1$ and $d_A = d_A^* = 2$ when $c = 0$.

With $c = 1$ and different sample size n, we estimate the parameters γ_0 and $\mathbf{B}_0 = (\beta_1, \beta_2)$. The estimation errors are defined as $|\gamma_0 - \hat{\gamma}|$ and $|(\hat{\gamma}, \hat{\mathbf{B}})(\hat{\gamma}, \hat{\mathbf{B}})^{\top} - (\gamma_0, \mathbf{B}_0)(\gamma_0, \mathbf{B}_0)^{\top}|$ for model (B) and $|\hat{\mathbf{B}}_A\hat{\mathbf{B}}_A^{\top} - (\gamma_0, \mathbf{B}_0)(\gamma_0, \mathbf{B}_0)^{\top}|$ for models (A) and (A*). With 200 replications for each size n, the calculation results are shown in Figure 1, (a) and (c). As the sample size increases, the estimation errors of model (B) drop rapidly. Multiplying the errors by a factor of root n, $\sqrt{n}/10$, the values still tend to decrease when n is small and remain roughly constant as n increases. This dynamical pattern supports that the estimation error has a consistency rate of root n. Figure 1, (a) and (c), also indicates clearly that model (B) can indeed give much more efficient estimators than model (A), while both of them are much better than the pHd method (Li 1992).

With $c = 0$ or $c = 1$, the proposed selection method can select the model and dimension quite accurately as shown in Figure 1, (b) and (d). The frequencies of correct selection tend to 100% as the sample size increases, lending support to the consistency of the selection method. When the model has dimension 2 (i.e., $c = 0$), the χ^2 test (Li 1991) has the same efficiency in selecting the dimension as our method. However, when the dimension is 3 (i.e., $c = 1$), the χ^2 test is much worse than our method.

Example 6.2 (Simulated data). In this example we check the asymptotic distribution by the model

$$Y = X_1 - X_2 - 2\exp\{-(X_3 + X_4 + X_5)^2\} + \sigma\varepsilon,$$

where $X_k = \eta_k + \eta_0, k = 1, 2, 3, 4, 5$, with $\eta_0, \ldots, \eta_5$ being independent and distributed uniformly on $[-.5, .5]$. Thus, the correlation coefficient between any two covariates is .5. In this example $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{05})^{\top} = (1, -1, 0, 0, 0)^{\top}$ and $\mathbf{B}_0 = \beta_1 = (\beta_{01}, \ldots, \beta_{05})^{\top} = (0, 0, 1, 1, 1)^{\top}/\sqrt{3}$.

With different sample size n and magnitude of noise σ, 500 random samples are generated and used to estimate the model. The calculation results are listed in Table 1. The estimates of γ_0 and $\mathbf{B}_0$ are quite accurate and stable by checking the mean and standard deviation of the estimates. When $\sigma = 1$ and $n = 100$, the frequency that the model and its dimension are correctly selected is .995; for the other three cases the frequencies are all 100%. Next, we check the asymptotic distribution by calculating the frequencies of accepting $\gamma_{0k} = 0$ and $\beta_{0k} = 0$, that is, frequencies of $|\hat{\gamma}_{0k}| < 1.96\{\hat{\Xi}_{k,k}/n\}^{1/2}$ and $|\hat{\beta}_{0k}| < 1.96\{\hat{\Xi}_{p+k,p+k}/n\}^{1/2}$, respectively. Based on the

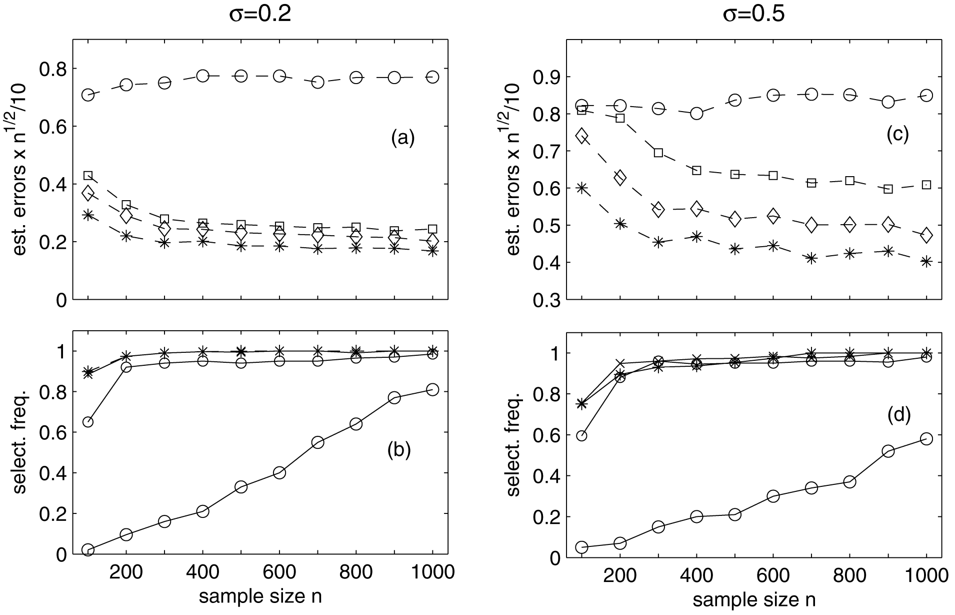

Figure 1. Calculation results for Example 6.1. The left panels are the results for $\sigma = .2$; the right panels for $\sigma = .5$. In (a) and (c), the dashed lines with circles, squares, diamonds, and stars represent the estimation errors multiplied by $\sqrt{n}/10$ of $(\gamma_0, \mathbf{B}_0)$ using model (A*) with pHd, model (A) with MAVE, and model (B) with the new method, and of γ_0 using model (B) with the new method, respectively. In (b) and (d), the solid lines with stars and crosses are the frequencies of correct selections of model (B) and its dimension for $c = 1$ and $c = 0$, respectively. The solid lines with circles are the frequencies of correct selections of the dimension using the χ^2 test (with a 5% significance level) of Li (1991) with $c = 0$ (upper panel) and $c = 1$ (lower panel), respectively.

asymptotic distribution in Corollary 5.1 and Remark 5.4, the frequencies should be around 95% as n is large enough. The frequencies reported in the table lend support to the asymptotic distributions.

Example 6.3 (Real data). A variety of atmospheric pollutants are known or suspected to have serious effects on public health and the environment; see the report of World Health Organization (WHO) (2003). The main pollutants include nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), particulate matter (PM_x; particles with diameter smaller than x micrometers), ozone (O_3), among others. Pollutants can be classified as either primary or secondary pollutants. Primary pollutants are substances directly produced by a process, such as ash from a volcanic eruption or carbon monoxide gas from motor vehicle exhaust. The primary pollutants can be controlled by reducing the emission of harmful gas. Secondary pollutants are generated in the air when primary pollutants *react* or *interact* with weather conditions; see also the WHO report (2003). An important example of secondary pollutants is ozone, one of the many secondary pollutants that make up photochemical smog. Therefore, it is of great interest to model the dependence of the ozone concentration on the primary pollutants and weather conditions, whereby we can control the ozone concentration by controlling the primary pollutants. For this purpose, we use the data collected in Chicago from 1987 to 1997 (available at *http://www.ihapss.jhsph.edu/data/data.htm*). The weather con-

Table 1. Mean, standard deviation (in parentheses) of estimates, and frequencies of accepting $\gamma_{0k} = 0$ and $\beta_{0k} = 0$ at 5% level (in square brackets) for Example 6.2

σ	n	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}
.5	100	.993 (.186) [.004]	−.99 (.183) [0]	−.008 (.175) [.946]	.001 (.169) [.946]	−.003 (.167) [.952]	.002 (.092) [.984]	.004 (.093) [.976]	.572 (.081) [0]	.565 (.082) [0]	.561 (.087) [0]
	400	1.002 (.080) [0]	−1.002 (.079) [0]	.002 (.076) [.954]	.001 (.083) [.932]	.001 (.082) [.930]	−.001 (.039) [.976]	.001 (.039) [.974]	.574 (.036) [0]	.576 (.036) [0]	.576 (.034) [0]
1.0	100	.973 (.352) [.202]	−1.007 (.347) [.188]	.029 (.333) [.942]	.000 (.342) [.918]	−.015 (.336) [.944]	.017 (.180) [.952]	.001 (.184) [.944]	.517 (.174) [.182]	.540 (.168) [.134]	.540 (.151) [.142]
	400	.995 (.164) [0]	−1.003 (.159) [0]	.003 (.157) [.938]	−.001 (.153) [.952]	−.001 (.158) [.936]	.002 (.083) [.968]	.003 (.084) [.966]	.572 (.069) [0]	.563 (.073) [0]	.570 (.071) [0]

Table 2. Selection procedure for dimension and models

Working dimension q	Model and methods			
	(A): $CVA(q)$	(B): $CVB(q-1)$	(A*): SIR p value	(A*): pHd p value
1	.5943	.5744*	.0000	.0000
2	.4484	.4450	.0000	.0000
3	.4431	.4392	.3299	.0000
4	.4565	.4410	.7472	.0000
5	.4699	.4474	.9207	.0000
6	.4819	.4583	.9515	.0002
7	.4981	.4611	.9862	.1583
8	.5065	.4725	—	—

*CV value of linear regression model.

ditions include daily temperature (T), humidity (H), and their difference between the highest and lowest values, denoted by VT and VH, respectively.

Let Y be the daily average ozone levels and $\mathbf{X} = (\mathrm{PM}_{10}, \mathrm{SO}_2, \mathrm{NO}_2, \mathrm{CO}, \mathrm{T}, \mathrm{H}, \mathrm{VT}, \mathrm{VH})^\top$. All the variables (including Y) are standardized separately such that each has mean 0 and variance 1 in order to make it easy to compare the estimated coefficients in the dimension reduction directions. First, we apply the model selection method to models (A) and (B). Table 2 lists the calculation results. Based on the calculation, we choose $\hat{d}_A = 3$ for model (A). However, model (B) with $d_B = 2$ is preferred because $CVB(2) = .4392 < CVA(3) = .4431 \leq \min\{CVA(1), \dots, CVA(8)\}$. If we use SIR and the χ^2 test, the selected dimension is also 3 for model (A*); see Table 2. If we use pHd, the selected number of dimensions is 7, which does not seem so reasonable.

Next, we fit the data to models (A) and (A*) with $d_A = d_A^* = 3$ and model (B) with $d_B = 2$. The estimation results are shown in Table 3 and Figure 2.

Based on the previous estimation and plots in Figure 2, we have the following observations. First, the plots of model (B) show a much clearer relationship between the level of O_3 and the covariates than the plots of models (A) and (A*), indicating that specifying the model structure as in model (B) is very helpful in recovering the relationships between the response and its covariates. Second, primary pollutants combined with weather conditions have a linear effect on the level of O_3 as shown in panel 7 of Figure 2. Third, weather conditions play important roles (and seem to predominate) in the nonlinear part. There are thresholds for weather conditions (at about 0 in panels 8 and 9 of Fig. 2), suggesting certain weather conditions are necessary for the chemical reaction between the primary pollutants to create ozone. These findings support the chemistry-based claim that ozone is generated in the air when primary pollutants react or interact under certain weather conditions.

7. CONCLUSION

Dimension reduction is a useful tool for statistical inference. However, it is still far from the goal of statistical modeling. By separating the linear component from the nonlinear components, this article considered one method of specification, leading to model (B). Based on the investigation (theory, numerical performance, and real data analysis) in this article, the specification can achieve three goals. First is better estimation of the model. See Remark 5.2 for the detailed discussion and Example 6.1 for numerical comparison. Second is easier interpretation of the estimated model. See Example 6.3, where the specified model can give a much clearer hint to the possible relationship between the response and the dimension reduction directions than dimension reduction model (A) or model (A*). Third is identification of the linear component and nonlinear components, which is the interest of the other disciplines.

Table 3. Estimated coefficients (and corresponding S.E.) of models (A), (B), and (A*)

Model		Variables							
		PM_{10}	SO_2	CO	NO_2	T	H	VT	VH
Model (A)	$\bar{\beta}_1$	.009	.089	.581	−.410	−.260	−.583	−.229	−.162
	$\bar{\beta}_2$	−.379	−.467	−.266	−.239	.313	−.633	.029	.101
	$\bar{\beta}_3$	.194	.188	−.020	−.026	.869	.159	.118	.362
Model (B)	γ_0	.113	.097	.039	.147	−.079	.226	.073	.029
		(.019)	(.015)	(.021)	(.018)	(.019)	(.023)	(.041)	(.047)
	$\tilde{\beta}_1$	−.247	−.112	−.222	.167	−.129	.622	.603	.286
		(.121)	(.069)	(.104)	(.096)	(.084)	(.168)	(.251)	(.341)
	$\tilde{\beta}_2$	−.010	.043	−.137	.111	.545	.488	.489	.438
		(.089)	(.087)	(.111)	(.109)	(.075)	(.203)	(.274)	(.293)
Model (A*)	β_1	.208	−.096	.072	−.424	.757	−.322	.091	.277
	β_2	.445	.264	−.242	.299	.048	.500	.461	.345
	β_3	−.329	−.228	−.143	.458	.434	.409	.429	.262

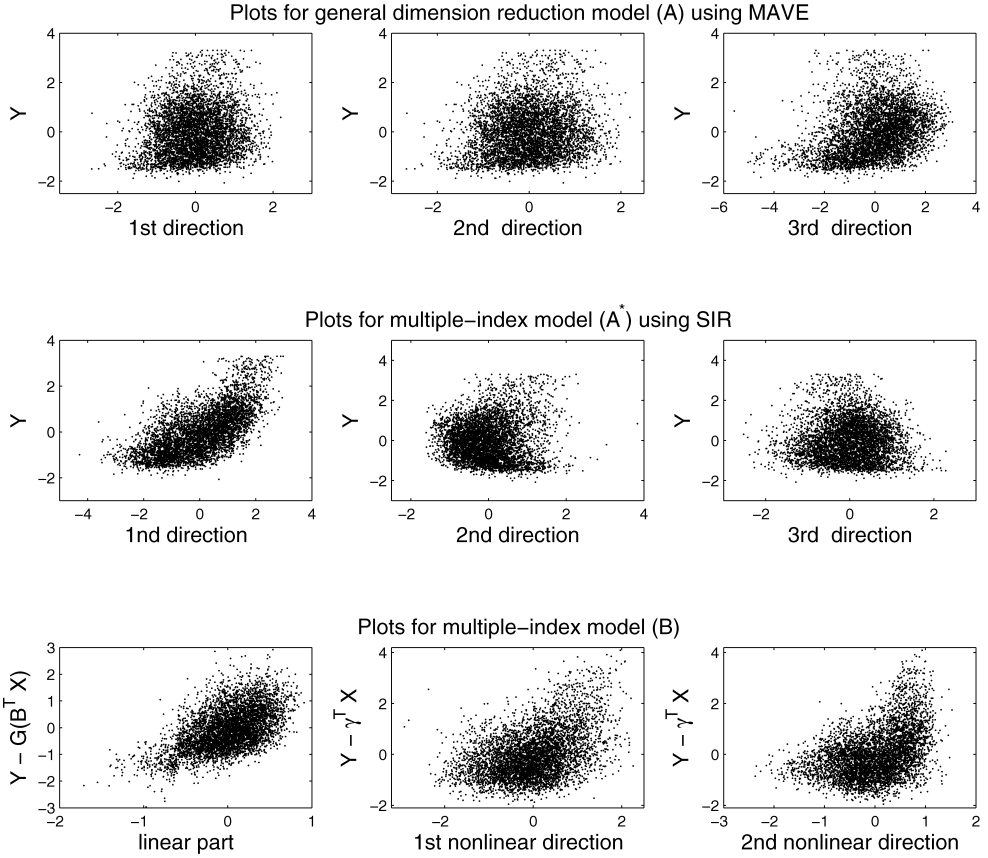

Figure 2. Calculation results for Example 6.3. The three upper panels are plots of Y against the three directions of the dimension reduction space in model (A). The three panels in the middle are plots of Y against the three directions of the dimension reduction space in model (A*). The three lower panels are plots of the linear part and two nonlinear parts in model (B).

This article only considered the specification of dimension reduction in the conditional mean function (Samarov 1993; Hristache et al. 2001; Cook and Li 2002; Xia et al. 2002; Yin and Cook 2002; among others). A parallel specification for general dimension reduction (Li 1991; Cook 1998; among others) would be more relevant and needs to be investigated. Other methods of specification are also possible. These specifications will push dimension reduction one step closer to statistical modeling.

APPENDIX: ASSUMPTIONS AND PROOF OF PROPOSITION 1.1

We need the following conditions to prove our theoretical results.

(C1) (Design of $\mathbf{X}$) The density function $f(\mathbf{x})$ of $\mathbf{X}$ has bounded second-order derivatives on $\mathcal{R}^p$ and $E|\mathbf{X}|^4 < \infty$; functions $\mu_{\mathbf{B}}(\mathbf{u})$ and $w_{\mathbf{B}}(\mathbf{u})$ have bounded derivatives with respect to $\mathbf{u}$ and $\mathbf{B}$ in a small neighborhood of $\mathbf{B}_0$.

(C2) (Moments of Y) Response Y has moment $E|Y|^5 < \infty$.

(C3) (Conditional function) Function $E(Y - \gamma^\top\mathbf{X}|\mathbf{B}^\top\mathbf{X} = \mathbf{u})$ has bounded fourth-order derivatives with respect to $\mathbf{u}$, γ, and $\mathbf{B}$ for $(\gamma, \mathbf{B})$ in a small neighborhood of $(\gamma_0, \mathbf{B}_0)$.

(C4) (Nonlinear dimension) Matrix $\mathbf{M}_0 = E[\rho(f(\mathbf{X}))\{\nabla G(\mathbf{B}_0^\top \times \mathbf{X}) - \bar{\nabla}\}\{\nabla G(\mathbf{B}_0^\top\mathbf{X}) - \bar{\nabla}\}^\top]$ has full rank of d_B, where $\bar{\nabla} = E\{\rho(f(\mathbf{X}))\nabla G(\mathbf{B}_0^\top\mathbf{X})\}/E\rho(f(\mathbf{X}))$.

(C5) (Kernel function) $K(\mathbf{u})$ is a multivariate density function with bounded second-order derivatives and a compact support.

(C6) (Bandwidths) Bandwidths $h_0 \propto n^{-\varrho_0}$ and $\hbar_t \propto n^{-\varrho_1}$, where $\varrho_0 = 1/(p+4)$ and $\varrho_1 = 1/(d_B+4)$.

In (C1), the requirement of $E|\mathbf{X}|^4 < \infty$ is used to handle the boundary points in the proof with trimming function $\mathcal{I}_{nj}$. If we adopt the fixed trimming scheme of Härdle, Hall, and Ichimura (1993), this requirement can be removed. Requiring (C2) is to meet the conditions for the uniform consistency results in a uniform consistency result. Härdle et al. (1993) even required the existence of all moments of Y. Lower order of smoothness than (C3) is sufficient if we are only interested in the estimation consistency. Condition (C4) indicates that the dimension d_B cannot be further reduced. The popular kernel functions such as the Epanechnikov kernel and the quadratic kernel are included in (C5). The Gaussian kernel can be used with some modifications to the proofs. Bandwidths satisfying (C6) can be found easily; see, for example, Yang and Tschernig (1999). Actually, a wider range of bandwidths is allowed; see the proofs.

Proof of Proposition 1.1

(I) Without loss of generality, assume that $\mathbf{\Sigma}_0 = \mathbf{I}_p$; otherwise, we can standardize the covariates. Let $\bar{\gamma}_0 = \gamma_0 - \mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top\gamma_0$ and $\bar{\mathbf{B}}_0 = \mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1/2}\mathbf{Q}\mathbf{Q}_0$, where $\mathbf{Q}$ is the eigenvector of $E[\{\nabla G(\mathbf{B}_0^\top\mathbf{X}) + (\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top\gamma_0\}\{\nabla G(\mathbf{B}_0^\top\mathbf{X}) + (\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top\gamma_0\}^\top]$ and $\mathbf{Q}_0 = \mathrm{diag}(e_1, \dots, e_q)$ with $e_k = \pm 1$ having the same sign as the first nonzero element in the kth column of $\mathbf{B}_0(\mathbf{B}_0^\top\mathbf{B}_0)^{-1/2}\mathbf{Q}$. Let $\bar{G}(\mathbf{u}) = G(\mathbf{B}_0^\top\mathbf{B}_0\mathbf{Q}\mathbf{Q}_0\mathbf{u}) + \gamma_0^\top\mathbf{B}_0\mathbf{Q}\mathbf{Q}_0\mathbf{u}$. We have

$$Y = \bar{\gamma}_0^\top\mathbf{X} + \bar{G}(\bar{\mathbf{B}}_0^\top\mathbf{X}) + \varepsilon,$$

and $\bar{\gamma}_0$ and $\bar{\mathbf{B}}_0$ satisfy (i1)–(i3).

(II) Note that $\mathbf{\Omega} = \mathbf{B}_0\mathbf{Q}\mathbf{Q}_0\Lambda(\mathbf{B}_0\mathbf{Q}\mathbf{Q}_0)^\top$, where $\mathbf{Q}\Lambda\mathbf{Q}^\top$ is the eigenvalue–eigenvector decomposition of $\mathbf{\Omega}_0$. Therefore, $\mathbf{B}_0$ must be a basis in the space spanned by the first d_B eigenvectors of $\mathbf{\Omega}$. Any two bases differ by a rotating matrix. Thus, the second equation of (II) holds. By (1), we have $\gamma_0 = E\nabla M(\mathbf{X}) - \mathbf{B}_0 E\nabla G(\mathbf{B}_0^\top\mathbf{X})$. Thus, by (i2),

$$\gamma_0 = (\mathbf{I} - \mathbf{B}_0\mathbf{B}_0^\top)\gamma_0 = (\mathbf{I} - \mathbf{B}_0\mathbf{B}_0^\top)E\{\nabla M(\mathbf{X})\}$$

is unique because $\mathbf{B}_0\mathbf{B}_0^\top$ is unique by the second equation of (II). Therefore, the first equation of (II) holds. The last equation of (II) follows from the first two.

[Received November 2007. Revised July 2008.]

REFERENCES

Banerjee, A. N. (2007), "A Method of Estimating the Average Derivative," *Journal of Econometrics*, 136, 65–88.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Cook, R. D. (1998), *Regression Graphics*, New York: Wiley.

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474.

Delecroix, M., Hristache, M., and Patilea, V. (2005), "On Semiparametric M-Estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136, 730–769.

Donkers, B., and Schafgans, M. M. A. (2003), "A Derivative Based Estimator for Semiparametric Index Models," CentER Discussion Paper 22, Tilburg University.

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

Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057.

Gao, J., and Tong, H. (2004), "Semiparametric Nonlinear Time Series Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 66, 321–336.

Grenfell, B. T., Finkenstädt, B. F., Wilson, K., Coulson, T. N., and Crawley, M. J. (2000), "Ecology—Nonlinearity and the Moran Effect," *Nature*, 406, 847.

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178.

Horowitz, J. (1998), *Semiparametric Methods in Econometrics*, New York: Springer.

Hristache, M., Juditski, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics*, 29, 1537–1566.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H., and Lee, L. F. (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, eds. W. A. Barnett, J. Powell, and G. Tauchen, Cambridge, U.K.: Cambridge University Press.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

——— (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

Maestri, R., et al. (2005), "Linear and Non-Linear Indices of Heart Rate Variability in Chronic Heart Failure: Mutual Interrelationships and Prognostic Value," *Computers in Cardiology*, 3, 981–984.

Newey, W. K., and Stoker, T. M. (1993), "Efficiency of Weighted Average Derivative Estimation and Index Models," *Econometrica*, 61, 1199–1223.

Rao, C. R., and Mitra, S. K. (1971), *Generalized Inverse of Matrices and Its Applications*, New York: Wiley.

Samarov, A. M. (1993), "Exploring Regression Structure Using Nonparametric Functional Estimation," *Journal of the American Statistical Association*, 88, 836–847.

Samarov, A., Spokoiny, V., and Vial, C. (2005), "Component Identification and Estimation in Nonlinear High-Dimensional Regression Models by Structure Adaptation," *Journal of the American Statistical Association*, 100, 429–445.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

Sliverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Speckman, P. (1988), "Kernel Smoothing in Partially Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

Tong, H. (1990), *Non-Linear Time Series: A Dynamical System Approach*, New York: Oxford University Press.

Wang, Q. (2003), "Dimension Reduction in Partly Linear Error-in-Response Models With Validation Data," *Journal of Multivariate Analysis*, 85, 234–252.

World Health Organization (2003), report of a WHO/HEI working group, Bonn, Germany. Available at *www.euro.who.int/document/e78992.pdf*.

Xia, Y. (2006), "Asymptotic Distributions of Two Estimators of the Single-Index Model," *Econometric Theory*, 22, 1112–1137.

——— (2007), "A Constructive Approach to the Estimation of Dimension Reduction Directions," *The Annals of Statistics*, 35, 2654–2690.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

Yang, L., and Tschernig, R. (1999), "Multivariate Bandwidth Selection for Local Linear Regression," *Journal of the Royal Statistical Society*, Ser. B, 61, 793–815.

Yin, X., and Cook, R. D. (2002), "Dimension Reduction for the Conditional k-th Moment in Regression", *Journal of the Royal Statistical Society*, Ser. B, 64, 159–175.

——— (2004), "Asymptotic Distribution of Test Statistic for the Covariance Dimension Reduction Methods in Regression," *Statistics and Probability Letters*, 68, 421–427.

Yu, Y., and Ruppert, D. (2002), "Penalized Spline Estimation for Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 97, 1042–1054.