The Annals of Statistics
2011, Vol. 39, No. 3, 1658–1688
DOI: 10.1214/10-AOS871

THE EFM APPROACH FOR SINGLE-INDEX MODELS

BY XIA CUI[1], WOLFGANG KARL HÄRDLE[2] AND LIXING ZHU[3]

Sun Yat-sen University, Humboldt-Universität zu Berlin and National Central University, and Hong Kong Baptist University and Yunnan University of Finance and Economics

Single-index models are natural extensions of linear models and circumvent the so-called curse of dimensionality. They are becoming increasingly popular in many scientific fields including biostatistics, medicine, economics and financial econometrics. Estimating and testing the model index coefficients $\boldsymbol{\beta}$ is one of the most important objectives in the statistical analysis. However, the commonly used assumption on the index coefficients, $\|\boldsymbol{\beta}\| = 1$, represents a nonregular problem: the true index is on the boundary of the unit ball. In this paper we introduce the EFM approach, a method of estimating functions, to study the single-index model. The procedure is to first relax the equality constraint to one with $(d-1)$ components of $\boldsymbol{\beta}$ lying in an open unit ball, and then to construct the associated $(d-1)$ estimating functions by projecting the score function to the linear space spanned by the residuals with the unknown link being estimated by kernel estimating functions. The root-n consistency and asymptotic normality for the estimator obtained from solving the resulting estimating equations are achieved, and a Wilks type theorem for testing the index is demonstrated. A noticeable result we obtain is that our estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. [*J. Amer. Statist. Assoc.* **92** (1997) 447–489]. A fixed-point iterative scheme for computing this estimator is proposed. This algorithm only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Numerical studies based on simulation and on applications suggest that this new estimating system is quite powerful and easy to implement.

1. Introduction. Single-index models combine flexibility of modeling with interpretability of (linear) coefficients. They circumvent the curse of dimensionality and are becoming increasingly popular in many scientific fields. The reduction of dimension is achieved by assuming the link function to be a univariate function applied to the projection of explanatory covariate vector on to some direction.

Received April 2010; revised December 2010.

[1]Supported by NNSF project (11026194) of China, RFDP (20100171120042) of China and "the Fundamental Research Funds for the Central Universities" (11lgpy26) of China.

[2]Supported by Deutsche Forschungsgemeinschaft SFB 649 "Ökonomisches Risiko."

[3]Supported by a Grant (HKBU2030/07P) from Research Grants Council of Hong Kong, Hong Kong, China.

MSC2010 subject classifications. 62G08, 62G08, 62G20.

Key words and phrases. Single-index models, index coefficients, estimating equations, asymptotic properties, iteration.

In this paper we consider an extension of single-index models where, instead of a distributional assumption, assumptions of only the mean function and variance function of the response are made. Let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, denote the observed values with Y_i being the response variable and $\mathbf{X}_i$ as the vector of d explanatory variables. The relationship of the mean and variance of Y_i is specified as follows:

$$(1.1)\qquad E(Y_i|\mathbf{X}_i) = \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, \qquad \operatorname{Var}(Y_i|\mathbf{X}_i) = \sigma^2 V\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\},$$

where μ is a known monotonic function, V is a known covariance function, g is an unknown univariate link function and $\boldsymbol{\beta}$ is an unknown index vector which belongs to the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$. Here we assume the parameter space is Θ rather than the entire $\mathbb{R}^d$ in order to ensure that $\boldsymbol{\beta}$ in the representation (1.1) can be uniquely defined. This is a commonly used assumption on the index parameter [see Carroll et al. (1997), Zhu and Xue (2006), Lin and Kulasekera (2007)]. Another reparameterization is to let $\beta_1 = 1$ for the sign identifiability and to transform $\boldsymbol{\beta}$ to $(1, \beta_2, \dots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}$ for the scale identifiability. Clearly $(1, \beta_2, \dots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}$ can also span the parameter space Θ by simply checking that $\|(1, \beta_2, \dots, \beta_d)/(1 + \sum_{r=2}^d \beta_r^2)^{1/2}\| = 1$ and the first component $1/(1 + \sum_{r=2}^d \beta_r^2)^{1/2} > 0$. However, the fixed-point algorithm recommended in this paper for normalized vectors may not be suitable for such a reparameterization. Model (1.1) is flexible enough to cover a variety of situations. If μ is the identity function and V is equal to constant 1, (1.1) reduces to a single-index model Härdle, Hall and Ichimura (1993). Model (1.1) is an extension of the generalized linear model McCullagh and Nelder (1989) and the single-index model. When the conditional distribution of Y is logistic, then $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]^2$.

For single-index models: $\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = g(\boldsymbol{\beta}^\top \mathbf{X})$ and $V\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = 1$, various strategies for estimating $\boldsymbol{\beta}$ have been proposed in the last decades. Two most popular methods are the average derivative method (ADE) introduced in Powell, Stock and Stoker (1989) and Härdle and Stoker (1989), and the simultaneous minimization method of Härdle, Hall and Ichimura (1993). Next we will review these two methods in short. The ADE method is based on that $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x} = g'(\boldsymbol{\beta}^\top \mathbf{x})\boldsymbol{\beta}$ which implies that the gradient of the regression function is proportional to the index parameter $\boldsymbol{\beta}$. Then a natural estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)/\|n^{-1} \sum_{i=1}^n \widehat{\nabla G}(\mathbf{X}_i)\|$ with $\nabla G(\mathbf{x})$ denoting $\partial E(Y|\mathbf{X} = \mathbf{x})/\partial \mathbf{x}$ and $\|\cdot\|$ being the Euclidean norm. An advantage of the ADE approach is that it allows estimating $\boldsymbol{\beta}$ directly. However, the high-dimensional kernel smoothing used for computing $\widehat{\nabla G}(\mathbf{x})$ suffers from the "curse of dimensionality" if the model dimension d is large. Hristache, Juditski and Spokoiny (2001) improved the ADE approach by lowering the dimension of the kernel gradually. The method of Härdle, Hall and Ichimura (1993) is carried out by minimizing a least squares criterion based on nonparametric estimation of the link g with respect to $\boldsymbol{\beta}$ and bandwidth h. However, the minimization is difficult to implement since it depends on an optimization problem in a high-dimensional space. Xia et al. (2002) proposed to minimize average conditional variance (MAVE). Because the kernel used for computing $\boldsymbol{\beta}$ is a function of $\|\mathbf{X}_i - \mathbf{X}_j\|$, MAVE meets the problem of data sparseness. All the above estimators are consistent under some regular conditions. Asymptotic efficiency comparisons of the above methods have been discussed in Xia (2006) resulting in the MAVE estimator of $\boldsymbol{\beta}$ having the same limiting variance as the estimators of Härdle, Hall and Ichimura (1993), and claiming alternative versions of the ADE method having larger variance. In addition, Yu and Ruppert (2002) fitted the partially linear single-index models using a penalized spline method. Huh and Park (2002) used the local polynomial method to fit the unknown function in single-index models. Other dimension reduction methods that were recently developed in the literature are sliced inverse regression, partial least squares and canonical correlation method. These methods handle high-dimensional predictors; see Zhu and Zhu (2009a, 2009b) and Zhou and He (2008).

The main challenges of estimation in the semiparametric model (1.1) are that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional parameter $\boldsymbol{\beta}$, and the parameter $\boldsymbol{\beta}$ is on the boundary of a unit ball. For estimating $\boldsymbol{\beta}$ the former challenge forces us to deal with the infinite-dimensional nuisance parameter g. The latter one represents a nonregular problem. The classic assumptions about asymptotic properties of the estimates for $\boldsymbol{\beta}$ are not valid. In addition, as a model proposed for dimension reduction, the dimension d may be very high and one often meets the problem of computation. To attack the above problems, in this paper we will develop an estimating function method (EFM) and then introduce a computational algorithm to solve the equations based on a fixed-point iterative scheme. We first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form $\{((1-\sum_{r=2}^d \beta_r^2)^{1/2}, \beta_2, \dots, \beta_d)^\top : \sum_{r=2}^d \beta_r^2 < 1\}$. Then the derivatives of a function with respect to $(\beta_2, \dots, \beta_d)^\top$ are readily obtained by the chain rule and the classical assumptions on the asymptotic normality hold after transformation. The estimating functions (equations) for $\boldsymbol{\beta}$ can be constructed by replacing $g(\boldsymbol{\beta}^\top\mathbf{X})$ with $\hat{g}(\boldsymbol{\beta}^\top\mathbf{X})$. The estimate $\hat{g}$ for the nuisance parameter g is obtained using kernel estimating functions and the smoothing parameter h is selected using K-fold cross-validation. For the problem of testing the index, we establish a quasi-likelihood ratio based on the proposed estimating functions and show that the test statistics asymptotically follow a χ^2-distribution whose degree of freedom does not depend on nuisance parameters, under the null hypothesis. Then a Wilks type theorem for testing the index is demonstrated.

The proposed EFM technique is essentially a unified method of handling different types of data situations including categorical response variable and discrete explanatory covariate vector. The main results of this research are as follows:

(a) *Efficiency*. A surprising result we obtain is that our EFM estimator for $\boldsymbol{\beta}$ has smaller or equal limiting variance than the estimator of Carroll et al. (1997).

(b) *Computation*. The estimating function system only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Unlike the quasi-likelihood inference [Carroll et al. (1997)] where the maximization is difficult to implement when d is large, the reparameterization and the explicit formulation of the estimating functions facilitate an efficient computation algorithm. Here we use a fixed-point iterative scheme to compute the resultant estimator. The simulation results show that the algorithm adapts to higher model dimension and richer data situations than the MAVE method of Xia et al. (2002).

It is noteworthy that the EFM approach proposed in this paper cannot be obtained from the SLS method proposed in Ichimura (1993) and investigated in Härdle, Hall and Ichimura (1993). SLS minimizes the weighted least squares criterion $\sum_{j=1}^{n}[Y_j-\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}]^2V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}$, which leads to a biased estimating equation when we use its derivative if $V(\cdot)$ does not contain the parameter of interest. It will not in general provide a consistent estimator [see Heyde (1997), page 4]. Chang, Xue and Zhu (2010) and Wang et al. (2010) discussed the efficient estimation of single-index model for the case of additive noise. However, their methods are based on the estimating equations induced from the least squares rather than the quasi-likelihood. Thus, their estimation does not have optimal property. Also their comparison is with the one from Härdle, Hall and Ichimura (1993) and its later development. It cannot be applied to the setting under study. In this paper, we investigate the efficiency and computation of the estimates for the single-index models, and systematically develop and prove the asymptotic properties of EFM.

The paper is organized as follows. In Section 2, we state the single-index model, discuss estimation of g using kernel estimating functions and of $\boldsymbol{\beta}$ using profile estimating functions, and investigate the problem of testing the index using quasi-likelihood ratio. In Section 3 we provide a computation algorithm for solving the estimating functions and illustrate the method with simulation and practical studies. The proofs are deferred to the Appendix.

2. Estimating function method (EFM) and its large sample properties. In this section, which is concerned with inference based on the estimating function method, the model of interest is determined through specification of mean and variance functions, up to an unknown vector $\boldsymbol{\beta}$ and an unknown function g. Except for Gaussian data, model (1.1) need not be a full semiparametric likelihood specification. Note that the parameter space $\Theta=\{\boldsymbol{\beta}=(\beta_1,\dots,\beta_d)^\top:\|\boldsymbol{\beta}\|=1,\beta_1>0,$ $\boldsymbol{\beta}\in\mathbb{R}^d\}$ means that $\boldsymbol{\beta}$ is on the boundary of a unit ball and it represents therefore a nonregular problem. So we first choose an identifiable parameterization which transforms the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. By eliminating β_1, the parameter space Θ can be rearranged to a form

$\{((1-\sum_{r=2}^{d}\beta_r^2)^{1/2}, \beta_2, \ldots, \beta_d)^\top : \sum_{r=2}^{d}\beta_r^2 < 1\}$. Then the derivatives of a function with respect to $\boldsymbol{\beta}^{(1)} = (\beta_2, \ldots, \beta_d)^\top$ are readily obtained by chain rule and the classic assumptions on the asymptotic normality hold after transformation. This reparameterization is the key to analyzing the asymptotic properties of the estimates for $\boldsymbol{\beta}$ and to facilitating an efficient computation algorithm. We will investigate the estimation for g and $\boldsymbol{\beta}$ and propose a quasi-likelihood method to test the statistical significance of certain variables in the parametric component.

2.1. *The kernel estimating functions for the nonparametric part g.* If $\boldsymbol{\beta}$ is known, then we estimate $g(\cdot)$ and $g'(\cdot)$ using the local linear estimating functions. Let h denote the bandwidth parameter, and let $K(\cdot)$ denote the symmetric kernel density function satisfying $K_h(\cdot) = h^{-1}K(\cdot/h)$. The estimation method involves local linear approximation. Denote by α_0 and α_1 the values of g and g' evaluating at t, respectively. The local linear approximation for $g(\boldsymbol{\beta}^\top\mathbf{x})$ in a neighborhood of t is $\tilde{g}(\boldsymbol{\beta}^\top\mathbf{x}) = \alpha_0 + \alpha_1(\boldsymbol{\beta}^\top\mathbf{x} - t)$. The estimators $\hat{g}(t)$ and $\hat{g}'(t)$ are obtained by solving the kernel estimating functions with respect to α_0, α_1:

$$\begin{cases} \displaystyle\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top\mathbf{X}_j - t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\ \qquad \times [Y_j - \mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}] = 0, \\ \displaystyle\sum_{j=1}^{n} (\boldsymbol{\beta}^\top\mathbf{X}_j - t)K_h(\boldsymbol{\beta}^\top\mathbf{X}_j - t)\mu'\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}V^{-1}\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\} \\ \qquad \times [Y_j - \mu\{\tilde{g}(\boldsymbol{\beta}^\top\mathbf{X}_j)\}] = 0. \end{cases} \tag{2.1}$$

Having estimated α_0, α_1 at t as $\hat{\alpha}_0, \hat{\alpha}_1$, the local linear estimators of $g(t)$ and $g'(t)$ are $\hat{g}(t) = \hat{\alpha}_0$ and $\hat{g}'(t) = \hat{\alpha}_1$, respectively.

The key to obtain the asymptotic normality of the estimates for $\boldsymbol{\beta}$ lies in the asymptotic properties of the estimated nonparametric part. The following theorem will provide some useful results. The following notation will be used. Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, $\rho_l(z) = \{\mu'(z)\}^l V^{-1}(z)$ and $\mathbf{J} = \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\beta}^{(1)}}$ the Jacobian matrix of size $d \times (d-1)$ with

$$\mathbf{J} = \begin{pmatrix} -\boldsymbol{\beta}^{(1)\top}/\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} \\ \mathbf{I}_{d-1} \end{pmatrix}, \qquad \boldsymbol{\beta}^{(1)} = (\beta_2, \ldots, \beta_d)^\top.$$

The moments of K and K^2 are denoted, respectively, by, $j = 0, 1, \ldots,$

$$\gamma_j = \int t^j K(t)\,dt \quad \text{and} \quad \nu_j = \int t^j K^2(t)\,dt.$$

PROPOSITION 1. *Under regularity conditions* (a), (b), (d) *and* (e) *given in the Appendix, we have*:

(i) *With* $h \to 0$, $n \to \infty$ *such that* $h \to 0$ *and* $nh \to \infty$, $\forall \boldsymbol{\beta} \in \Theta$, *the asymptotic conditional bias and variance of* $\hat{g}$ *are given by*

$$\begin{aligned} E\{\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{x}) - g(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} &= \{\tfrac{1}{2}\gamma_2 h^2 g''(\boldsymbol{\beta}^\top \mathbf{x})\}^2 \\ &\quad + \nu_0 \sigma^2 / [nh f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\ &\quad + o_P(h^4 + n^{-1}h^{-1}). \end{aligned} \tag{2.2}$$

(ii) *With* $h \to 0$, $n \to \infty$ *such that* $h \to 0$ *and* $nh^3 \to \infty$, *for the estimates of the derivative* g', *it holds that*

$$\begin{aligned} &E\{\{\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x})\}^2 | \mathcal{X}\} \\ &\quad = \{\tfrac{1}{6}\gamma_4 \gamma_2^{-1} h^2 g'''(\boldsymbol{\beta}^\top \mathbf{x}) \\ &\qquad + \tfrac{1}{2}(\gamma_4 \gamma_2^{-1} - \gamma_2) h^2 g''(\boldsymbol{\beta}^\top \mathbf{x}) \\ &\qquad\quad \times [\rho_2'\{g(\boldsymbol{\beta}^\top \mathbf{x})\}/\rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\} + f'_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})/f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})]\}^2 \\ &\qquad + \nu_2 \gamma_2^{-2} \sigma^2 / [nh^3 f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x}) \rho_2\{g(\boldsymbol{\beta}^\top \mathbf{x})\}] \\ &\qquad + o_P(h^4 + n^{-1}h^{-3}). \end{aligned} \tag{2.3}$$

(iii) *With* $h \to 0$, $n \to \infty$ *such that* $h \to 0$ *and* $nh^3 \to \infty$, *we have that*

$$E\left\{ \left\| \frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} - g'(\boldsymbol{\beta}^\top \mathbf{x}) \mathbf{J}^\top \{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\} \right\|^2 \middle| \mathcal{X} \right\} = \mathcal{O}_P(h^4 + n^{-1}h^{-3}). \tag{2.4}$$

The proof of this proposition appears in the Appendix. Results (i) and (ii) in Proposition 1 are routine and similar to Carroll, Ruppert and Welsh (1998). In the situation where $\sigma^2 V = \sigma^2$ and the function μ is identity, results (i) and (ii) coincide with those given by Fan and Gijbels (1996). From result (iii), it is seen that $\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}$ converges in probability to $g'(\boldsymbol{\beta}^\top \mathbf{x})\mathbf{J}^\top\{\mathbf{x} - E(\mathbf{x}|\boldsymbol{\beta}^\top \mathbf{x})\}$, rather than $g'(\boldsymbol{\beta}^\top \mathbf{x})\mathbf{J}^\top \mathbf{x}$ as if g were known. That is, $\lim_{n\to\infty}\{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial \boldsymbol{\beta}^{(1)}\} \neq \partial\{\lim_{n\to\infty} \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})\}/\partial \boldsymbol{\beta}^{(1)}$, which means that the convergence in probability and the derivation of the sequence $\hat{g}_n(\boldsymbol{\beta}^\top \mathbf{x})$ (as a function of n) cannot commute. This is primarily caused by the fact that the support of the infinite-dimensional nuisance parameter $g(\cdot)$ depends on the finite-dimensional projection parameter $\boldsymbol{\beta}$. In contrast, a semiparametric model where the support of the nuisance parameter is independent of the finite-dimensional parameter is a partially linear regression model having form $Y = \mathbf{X}^\top \boldsymbol{\theta} + \eta(T) + \varepsilon$. It is easy to check that the limit of $\partial \hat{\eta}(T)/\partial \boldsymbol{\theta}$ is equal to $E(\mathbf{X}|T)$, which is the derivative of $\lim_{n\to\infty} \hat{\eta}(T) = E(Y|T) - E(\mathbf{X}^\top|T)\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$. Result (iii) ensures that the proposed estimator does not require undersmoothing of $g(\cdot)$ to obtain a root-n consistent estimator for $\boldsymbol{\beta}$ and it is also of its own interest in inference theory for semiparametric models.

2.2. *The asymptotic distribution for the estimates of the parametric part* $\boldsymbol{\beta}$. We will now proceed to the estimation of $\boldsymbol{\beta} \in \Theta$. We need to estimate the $(d-1)$-dimensional vector $\boldsymbol{\beta}^{(1)}$, the estimator of which will be defined via

$$\sum_{i=1}^{n} [\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}/\partial\boldsymbol{\beta}^{(1)}] V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] = 0. \tag{2.5}$$

This is the direct analogue of the "ideal" estimating equation for known g, in that it is calculated by replacing $g(t)$ with $\hat{g}(t)$. An asymptotically equivalent and easily computed version of this equation is

$$\begin{aligned} \hat{\mathbf{G}}(\boldsymbol{\beta}) &\stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbf{J}^\top \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}] \\ &= 0 \end{aligned} \tag{2.6}$$

with $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ the Jacobian mentioned above, $\hat{g}$ and $\hat{g}'$ are defined by (2.1), and $\hat{\mathbf{h}}(t)$ the local linear estimate for $\mathbf{h}(t) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = t) = (h_1(t), \dots, h_d(t))^\top$,

$$\hat{\mathbf{h}}(t) = \sum_{i=1}^{n} b_i(t)\mathbf{X}_i \Big/ \sum_{i=1}^{n} b_i(t),$$

where $b_i(t) = K_h(\boldsymbol{\beta}^\top \mathbf{X}_i - t)\{S_{n,2}(t) - (\boldsymbol{\beta}^\top \mathbf{X}_i - t)S_{n,1}(t)\}$, $S_{n,k} = \sum_{i=1}^{n} K_h(\boldsymbol{\beta}^\top \times \mathbf{X}_i - t)(\boldsymbol{\beta}^\top \mathbf{X}_i - t)^k, k = 1, 2$. We use (2.6) to estimate $\boldsymbol{\beta}^{(1)}$ in the single-index model, and then use the fact that $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ to obtain $\hat{\beta}_1$. The use of (2.6) constitutes in our view a new approach to estimating single-index models; since (2.6) involves smooth pilot estimation of g, g' and $\mathbf{h}$ we call it the Estimation Function Method (EFM) for $\boldsymbol{\beta}$.

REMARK 1. The estimating equations $\hat{\mathbf{G}}(\boldsymbol{\beta})$ can be represented as the gradient vector of the following objective function:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int_\mu^y \frac{s-y}{V\{\mu^{-1}(s)\}} \, ds$ and $\mu^{-1}(\cdot)$ the inverse function of $\mu(\cdot)$. The existence of such a potential function makes $\hat{\mathbf{G}}(\boldsymbol{\beta})$ to inherit properties of the ideal likelihood score function. Note that $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ is an open, connected subset of $\mathbb{R}^{d-1}$. By the regularity conditions assumed on $\mu(\cdot), g(\cdot), V(\cdot)$ (for details see the Appendix), we know that the quasi-likelihood function $\hat{Q}(\boldsymbol{\beta})$ is twice continuously differentiable on $\{\|\boldsymbol{\beta}^{(1)}\| < 1\}$ such that the global maximum of $\hat{Q}(\boldsymbol{\beta})$ can be achieved at some point. One may ask whether the solution is unique and also consistent. Some elementary calculations lead to the Hessian matrix $\partial^2 \hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}$, because the partial derivative $\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}} = \mu'\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}$, then

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}}\\
&\quad= \frac{1}{n}\frac{\partial\hat{\mathbf{G}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial[\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]}{\partial\boldsymbol{\beta}^{(1)}}[Y_i-\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]\\
&\qquad-\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)}}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\biggl[-\frac{\partial\{\boldsymbol{\beta}^{(1)}/\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}\}}{\partial\boldsymbol{\beta}^{(1)}}\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_{1i}-\hat{h}_1(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\\
&\qquad\qquad+\mathbf{J}^\top\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)\top}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\\
&\qquad\qquad+\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\frac{\partial\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}}{\partial\boldsymbol{\beta}^{(1)\top}}\\
&\qquad\qquad\qquad-\mathbf{J}^\top\hat{g}'(\boldsymbol{\beta}^\top\mathbf{X}_i)\frac{\partial\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)}{\partial\boldsymbol{\beta}^{(1)}}\rho_1\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\biggr]\\
&\qquad\qquad\times[Y_i-\mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]\\
&\qquad-\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}^\top\hat{g}'^2(\boldsymbol{\beta}^\top\mathbf{X}_i)\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\{\mathbf{X}_i-\hat{\mathbf{h}}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}^\top\rho_2\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}\mathbf{J}.
\end{aligned}
$$

By the regularity conditions in the Appendix, the multipliers of the residuals $[Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top\mathbf{X}_i)\}]$ in the first sum of (2.7) are bounded. Mimicking the proof of Proposition 1, the first sum can be shown to converge to 0 in probability as n goes to infinity. The second sum converges to a negative semidefinite matrix. If the Hessian matrix $\frac{1}{n}\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\top}}$ is negative definite for all values of $\boldsymbol{\beta}^{(1)}$, $\hat{\mathbf{G}}(\boldsymbol{\beta})$ has a unique root. At sample level, however, estimating functions may have more than one root. For the EFM method, the quasi-likelihood $\hat{Q}(\boldsymbol{\beta})$ exists, which can be used to distinguish local maxima from minima. Thus, we suppose (2.6) has a unique solution in the following context.

REMARK 2. It can be seen from the proof in the Appendix that the population version of $\hat{\mathbf{G}}(\boldsymbol{\beta})$ is

$$\mathbf{G}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{J}^\top g'(\boldsymbol{\beta}^\top \mathbf{X}_i)\{\mathbf{X}_i - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\rho_1\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}[Y_i - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}], \tag{2.7}$$

which is obtained by replacing $\hat{g}, \hat{g}', \hat{\mathbf{h}}$ with $g, g', \mathbf{h}$ in (2.6). One important property of (2.7) is that the second Bartlett identity holds, for any $\boldsymbol{\beta}$:

$$E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = -E\left\{\frac{\partial \mathbf{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)}}\right\}.$$

This property makes the semiparametric efficiency of the EFM (2.6) possible.

Let $\boldsymbol{\beta}^0 = (\beta_1^0, \boldsymbol{\beta}^{(1)0^\top})^\top$ denote the true parameter and $\mathbf{B}^+$ denote the Moore–Penrose inverse of any given matrix $\mathbf{B}$. We have the following asymptotic result for the estimator $\hat{\boldsymbol{\beta}}^{(1)}$.

THEOREM 2.1. *Assume the estimating function* (2.6) *has a unique solution and denote it by* $\hat{\boldsymbol{\beta}}^{(1)}$. *If the regularity conditions* (a)–(e) *in the Appendix are satisfied, the following results hold*:

(i) *With* $h \to 0$, $n \to \infty$ *such that* $(nh)^{-1}\log(1/h) \to 0$, $\hat{\boldsymbol{\beta}}^{(1)}$ *converges in probability to the true parameter* $\boldsymbol{\beta}^{(1)0}$.

(ii) *If* $nh^6 \to 0$ *and* $nh^4 \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}) \xrightarrow{\mathcal{L}} N_{d-1}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}), \tag{2.8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}} = \{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^+|_{\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}^{(1)0}}$, $\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(1)}}$ *and*

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top\mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top\mathbf{X})\}\rho_2\{g(\boldsymbol{\beta}^\top\mathbf{X})\}\{g'(\boldsymbol{\beta}^\top\mathbf{X})\}^2/\sigma^2].$$

REMARK 3. Note that $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} = 0$, so the nonnegative matrix $\boldsymbol{\Omega}$ degenerates in the direction of $\boldsymbol{\beta}$. If the mean function μ is the identity function and the variance function is equal to a scale constant, that is, $\mu\{g(\boldsymbol{\beta}^\top\mathbf{X})\} = g(\boldsymbol{\beta}^\top\mathbf{X})$, $\sigma^2 V\{g(\boldsymbol{\beta}^\top\mathbf{X})\} = \sigma^2$, the matrix $\boldsymbol{\Omega}$ in Theorem 2.1 reduces to be

$$\boldsymbol{\Omega} = E[\{\mathbf{X}\mathbf{X}^\top - E(\mathbf{X}|\boldsymbol{\beta}^\top\mathbf{X})E(\mathbf{X}^\top|\boldsymbol{\beta}^\top\mathbf{X})\}\{g'(\boldsymbol{\beta}^\top\mathbf{X})\}^2/\sigma^2].$$

Technically speaking, Theorem 2.1 shows that an undersmoothing approach is unnecessary and that root-n consistency can be achieved. The asymptotic covariance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(1)0}}$ in general can be estimated by replacing terms in its expression by estimates of those terms. The asymptotic normality of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\boldsymbol{\beta}}^{(1)^\top})^\top$ will follow from Theorem 2.1 with a simple application of the multivariate delta-method, since $\hat{\beta}_1 = \sqrt{1 - \|\hat{\boldsymbol{\beta}}^{(1)}\|^2}$. According to the results of Carroll et al. (1997), the asymptotic variance of their estimator is $\boldsymbol{\Omega}^+$. Define the block partition of matrix $\boldsymbol{\Omega}$ as follows:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}, \tag{2.9}$$

where $\boldsymbol{\Omega}_{11}$ is a positive constant, $\boldsymbol{\Omega}_{12}$ is a $(d-1)$-dimensional row vector, $\boldsymbol{\Omega}_{21}$ is a $(d-1)$-dimensional column vector and $\boldsymbol{\Omega}_{22}$ is a $(d-1) \times (d-1)$ nonnegative definite matrix.

COROLLARY 1. *Under the conditions of Theorem 2.1, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^0}) \tag{2.10}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} = \mathbf{J}\{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^+ \mathbf{J}^\top|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$. *Further,*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}^0} \le \boldsymbol{\Omega}^+|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$$

and a strict less-than sign holds when $\det(\boldsymbol{\Omega}_{22}) = 0$. *That is, in this case EFM is more efficient than that of Carroll et al.* (1997).

The possible smaller limiting variance derived from the EFM approach partly benefits from the reparameterization so that the quasi-likelihood can be adopted. As we know, the quasi-likelihood is often of optimal property. In contrast, most existing methods treat the estimation of $\boldsymbol{\beta}$ as if it were done in the framework of linear dimension reduction. The target of linear dimension reduction is to find the directions that can linearly transform the original variables vector into a vector of one less dimension. For example, ADE and SIR are two relevant methods. However, when the link function $\mu(\cdot)$ is identity, the limiting variance derived here may not be smaller or equal to the ones of Wang et al. (2010) and Chang, Xue and Zhu (2010) when the quasi-likelihood of (2.5) is applied.

2.3. *Profile quasi-likelihood ratio test.* In applications, it is important to test the statistical significance of added predictors in a regression model. Here we establish a quasi-likelihood ratio statistic to test the significance of certain variables in the linear index. The null hypothesis that the model is correct is tested against a full model alternative. Fan and Jiang (2007) gave a recent review about generalized likelihood ratio tests. Bootstrap tests for nonparametric regression, generalized partially linear models and single-index models have been systematically investigated [see Härdle and Mammen (1993), Härdle, Mammen and Müller (1998), Härdle, Mammen and Proenca (2001)]. Consider the testing problem:

$$H_0 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k\right) \tag{2.11}$$
$$\longleftrightarrow \quad H_1 : g(\cdot) = g\left(\sum_{k=1}^{r} \beta_k X_k + \sum_{k=r+1}^{d} \beta_k X_k\right).$$

We mainly focus on testing $\beta_k = 0, k = r+1, \ldots, d$, though the following test procedure can be easily extended to a general linear testing $\mathbf{B}\tilde{\boldsymbol{\beta}} = 0$ where $\mathbf{B}$ is a known matrix with full row rank and $\tilde{\boldsymbol{\beta}} = (\beta_{r+1}, \ldots, \beta_d)^\top$. The profile quasi-likelihood ratio test is defined by

$$T_n = 2\Big\{\sup_{\boldsymbol{\beta}\in\Theta} \hat{Q}(\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\Theta, \tilde{\boldsymbol{\beta}}=0} \hat{Q}(\boldsymbol{\beta})\Big\}, \tag{2.12}$$

where $\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^n Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$, $Q[\mu, y] = \int_\mu^y \frac{s-y}{V\{\mu^{-1}(s)\}} \, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. The following Wilks type theorem shows that the distribution of T_n is asymptotically chi-squared and independent of nuisance parameters.

THEOREM 2.2. *Under the assumptions of Theorem 2.1, if* $\beta_k = 0, k = r + 1, \ldots, d$, *then*

$$T_n \xrightarrow{\mathcal{L}} \chi^2(d-r). \tag{2.13}$$

3. Numerical studies.

3.1. *Computation of the estimates.* Solving the joint estimating equations (2.1) and (2.6) poses some interesting challenges, since the functions $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{X})$ depend on $\boldsymbol{\beta}$ implicitly. Treating $\boldsymbol{\beta}^\top X$ as a new predictor (with given $\boldsymbol{\beta}$), (2.1) gives us $\hat{g}, \hat{g}'$ as in Fan, Heckman and Wand (1995). We therefore focus on (2.6), as estimating equations. It cannot be solved explicitly, and hence one needs to find solutions using numerical methods. The Newton–Raphson algorithm is one of the popular and successful methods for finding roots. However, the computational speed of this algorithm crucially depends on the initial value. We propose therefore a fixed-point iterative algorithm that is not very sensitive to starting values and is adaptive to larger dimension. It is worth noting that this algorithm can be implemented in the case that d is slightly larger than n, because the resultant procedure only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high dimensionality.

Rewrite the estimating functions as $\hat{\mathbf{G}}(\boldsymbol{\beta}) = \mathbf{J}^\top \hat{\mathbf{F}}(\boldsymbol{\beta})$ with

$$\hat{\mathbf{F}}(\boldsymbol{\beta}) = (\hat{F}_1(\boldsymbol{\beta}), \ldots, \hat{F}_d(\boldsymbol{\beta}))^\top$$

and

$$\hat{F}_s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{X_{si} - \hat{h}_s(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\mu'\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}\hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_i)V^{-1}\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\} \times [Y_i - \mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}].$$

Setting $\hat{\mathbf{G}}(\boldsymbol{\beta}) = 0$, we have that

$$\begin{cases} -\beta_2 \hat{F}_1(\boldsymbol{\beta})/\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_2(\boldsymbol{\beta}) = 0, \\ -\beta_3 \hat{F}_1(\boldsymbol{\beta})/\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_3(\boldsymbol{\beta}) = 0, \\ \cdots \\ -\beta_d \hat{F}_1(\boldsymbol{\beta})/\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} + \hat{F}_d(\boldsymbol{\beta}) = 0. \end{cases} \tag{3.1}$$

Note that $\|\boldsymbol{\beta}^{(1)}\|^2 = \sum_{r=2}^{d} \beta_r^2$, $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}$ and after some simple calculations, we can get that

$$\begin{cases} \beta_1 = |\hat{F}_1(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|, & s = 1, \\ \beta_s^2 = \hat{F}_s^2(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2, & s \geq 2, \end{cases}$$

and $\operatorname{sign}\{\beta_s \hat{F}_1(\boldsymbol{\beta})\} = \operatorname{sign}\{\hat{F}_s(\boldsymbol{\beta})\}$, $s \geq 2$. The above equation can also be rewritten as

$$\boldsymbol{\beta} \frac{\hat{F}_1(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} = \frac{|\hat{F}_1(\boldsymbol{\beta})|}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|} \times \frac{\hat{\mathbf{F}}(\boldsymbol{\beta})}{\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|}. \tag{3.2}$$

Then solving the equation (2.6) is equivalent to finding a fixed point for (3.2). Though $\|\boldsymbol{\beta}^{(1)}\| < 1$ holds almost surely in (3.2) and always $\|\boldsymbol{\beta}\| = 1$, there will be some trouble if (3.2) is directly used as iterative equations. Note that the value of $\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$ is used as denominator that may sometimes be small, which potentially makes the algorithm unstable. On the other hand, the convergence rate of the fixed-point iterative algorithm derived from (3.2) depends on L, where $\|\frac{\partial\{\hat{\mathbf{F}}(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|\}}{\partial \boldsymbol{\beta}}\| \leq L$. For a fast convergence rate, it technically needs a shrinkage value L. An ad hoc fix introduces a constant M, adding $M\boldsymbol{\beta}$ on both sides of (3.2) and dividing by $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M$:

$$\boldsymbol{\beta} = \frac{M}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M}\boldsymbol{\beta} + \frac{|\hat{F}_1(\boldsymbol{\beta})|/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|^2}{\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M}\hat{\mathbf{F}}(\boldsymbol{\beta}),$$

where M is chosen such that $\hat{F}_1(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\| + M \neq 0$. In addition, to accelerate the rate of convergence, we reduce the derivative of the term on the right-hand side of the above equality, which can be achieved by choosing some appropriate M. This is the iteration formulation in Step 2. Here the norm of $\boldsymbol{\beta}_{new}$ is not equal to 1 and we have to normalize it again. Since the iteration in Step 2 makes $\boldsymbol{\beta}_{new}$

to violate the identifiability constraint with norm 1, we design (3.2) to include the whole $\boldsymbol{\beta}$ vector. The possibility of renormalization for $\boldsymbol{\beta}_{new}$ avoids the difficulty of controlling $\|\boldsymbol{\beta}_{new}^{(1)}\| < 1$ in each iteration in Step 2.

Based on these observations, the fixed-point iterative algorithm is summarized as:

Step 0. Choose initial values for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{old}$.

Step 1. Solve the estimating equation (2.1) with respect to $\boldsymbol{\alpha}$, which yields $\hat{g}(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$ and $\hat{g}'(\boldsymbol{\beta}_{old}^{\top}\mathbf{x}_i)$, $1 \le i \le n$.

Step 2. Update $\boldsymbol{\beta}_{old}$ with $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ by solving the equation (2.6) in the fixed-point iteration

$$\boldsymbol{\beta}_{new} = \frac{M}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\boldsymbol{\beta}_{old} + \frac{|\hat{F}_1(\boldsymbol{\beta}_{old})|/\|\hat{F}(\boldsymbol{\beta}_{old})\|^2}{\hat{F}_1(\boldsymbol{\beta}_{old})/\|\hat{F}(\boldsymbol{\beta}_{old})\| + M}\hat{\mathbf{F}}(\boldsymbol{\beta}_{old}),$$

where M is a constant satisfying $\hat{F}_1(\boldsymbol{\beta})/\|\hat{F}(\boldsymbol{\beta})\| + M \neq 0$ for any $\boldsymbol{\beta}$.

Step 3. Repeat Steps 1 and 2 until $\max_{1\le s\le d} |\beta_{new,s} - \beta_{old,s}| \le tol$ is met with tol being a prescribed tolerance.

The final vector $\boldsymbol{\beta}_{new}/\|\boldsymbol{\beta}_{new}\|$ is the estimator of $\boldsymbol{\beta}^0$. Similarly to other direct estimation methods [Horowitz and Härdle (1996)], the preceding calculation is easy to implement. Empirically the initial value for $\boldsymbol{\beta}$, $(1, 1, \dots, 1)^{\top}/\sqrt{d}$ can be used in the calculations. The Epanechnikov kernel function $K(t) = 3/4(1 - t^2)I(|t| \le 1)$ is used. The bandwidth involved in Step 1 can be chosen to be optimal for estimation of $\hat{g}(t)$ and $\hat{g}'(t)$ based on the observations $\{\boldsymbol{\beta}_{old}^{\top}\mathbf{X}_i, Y_i\}$. So the standard bandwidth selection methods, such as K-fold cross-validation, generalized cross-validation (GCV) and the rule of thumb, can be adopted. In this step, we recommend K-fold cross-validation to determine the optimal bandwidth using the quasi-likelihood as a criterion function. The K-fold cross-validation is not too computationally intensive while making K not take too large values (e.g., $K = 5$). Here we recommend trying a number of smoothing parameters that smooth the data and picking the one that seems most reasonable. As an adjustment factor, M will increase the stability of iteration. Ideally, in each iteration an optimum value for M should be chosen guaranteeing that the derivative on the right-hand side of the iteration formulation in Step 2 is close to zero. Following this idea, M will be depending the changes of $\boldsymbol{\beta}$ and $\hat{\mathbf{F}}(\boldsymbol{\beta})/\|\hat{\mathbf{F}}(\boldsymbol{\beta})\|$. This will be an expensive task due to the computation for the derivative on the right-hand side of the iteration formulation in Step 2. We therefore consider M as constant nonvarying in each iteration, and select M by the K-fold cross-validation method, according to minimizing the model prediction error. When the dimension d gets larger, M will get smaller. In our simulation runs, we empirically search M in the interval $[2/\sqrt{d}, d/2]$. This choice gives pretty good practical performance.

3.2. *Simulation results*.

EXAMPLE 1 (Continuous response). We report a simulation study to investigate the finite-sample performance of the proposed estimator and compare it with the rMAVE [refined MAVE; for details see Xia et al. (2002)] estimator and the EDR estimator [see Hristache et al. (2001), Polzehl and Sperlich (2009)]. We consider the following model similar to that used in Xia (2006):

$$\begin{aligned} &E(Y|\boldsymbol{\beta}^\top\mathbf{X}) = g(\boldsymbol{\beta}^\top\mathbf{X}), \qquad g(\boldsymbol{\beta}^\top\mathbf{X}) = (\boldsymbol{\beta}^\top\mathbf{X})^2\exp(\boldsymbol{\beta}^\top\mathbf{X}); \\ &\mathrm{Var}(Y|\boldsymbol{\beta}^\top\mathbf{X}) = \sigma^2, \qquad \sigma = 0.1. \end{aligned} \tag{3.3}$$

Let the true parameter $\boldsymbol{\beta} = (2, 1, 0, \dots, 0)^\top/\sqrt{5}$. Two sets of designs for $\mathbf{X}$ are considered: Design (A) and Design (B). In Design (A), $(X_s + 1)/2 \sim \mathrm{Beta}(\tau, 1)$, $1 \le s \le d$ and, in Design (B), $(X_1 + 1)/2 \sim \mathrm{Beta}(\tau, 1)$ and $P(X_s = \pm 0.5) = 0.5$, $s = 2, 3, 4, \dots, d$. The data generated in Design (A) are not elliptically symmetric. All the components of Design (B) are discrete except for the first component X_1. Y is generated from a normal distribution. This simulation data set consists of 400 observations with 250 replications. The results are shown in Table 1. All rMAVE, EDR and EFM estimates are close to the true parameter vector for $d = 10$. However, the average estimation errors from rMAVE and EDR estimates for $d = 50$ are about 2 and 1.5 times as large as those of the EFM estimates, respectively. This indicates that the fixed-point algorithm is more adaptive to high dimension.

EXAMPLE 2 (Binary response). This simulation design assumes an underlying single-index model for binary responses with

$$\begin{aligned} &P(Y = 1|\mathbf{X}) = \mu\{g(\boldsymbol{\beta}^\top\mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top\mathbf{X})\}], \\ &g(\boldsymbol{\beta}^\top\mathbf{X}) = \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 2)/\{1 + \exp(5\boldsymbol{\beta}^\top\mathbf{X} - 3)\} - 1.5. \end{aligned} \tag{3.4}$$

The underlying coefficients are assumed to be $\boldsymbol{\beta} = (2, 1, 0, \dots, 0)^\top/\sqrt{5}$. We consider two sets of designs: Design (C) and Design (D). In Design (C), X_1 and X_2

TABLE 1
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.3)

		Design (A)			Design (B)		
d	**τ**	**rMAVE**	**EDR**	**EFM**	**rMAVE**	**EDR**	**EFM**
10	0.75	0.0559*	0.0520	0.0792	0.0522*	0.0662	0.0690
10	1.5	0.0323*	0.0316	0.0298	0.0417*	0.0593	0.0457
50	0.75	0.9900	0.7271	0.5425	0.9780	0.7712	0.4515
50	1.5	0.3776	0.3062	0.1796	0.4693	0.4103	0.2211

*The values are adopted from Xia (2006).

TABLE 2
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.4)

	Design (C)			Design (D)		
d	**rMAVE**	**EDR**	**EFM**	**rMAVE**	**EDR**	**EFM**
10	0.5017	0.5281	0.4564	0.9614	0.9574	0.7415
50	2.0991	1.2695	1.1744	2.5040	2.4846	1.9908

follow the uniform distribution $U(-2, 2)$. In Design (D), X_1 is also assumed to be uniformly distributed in interval $(-2, 2)$ and $(X_2 + 1)/2 \sim \text{Beta}(1, 1)$. Similar designs for generalized partially linear single-index models are assumed in Kane, Holt and Allen (2004). Here a sample size of 700 is used for the case $d = 10$ and 3,000 is used for $d = 50$. Different sample sizes from Example 1 are used due to varying complexity of the two examples. For this example, 250 replications are simulated and the results are displayed in Table 2. In this set of simulations, the average estimation errors from rMAVE estimates and EDR estimates are about 1.5 and 1.2 times as large as EFM estimates, under both Design (C) and Design (D) for $d = 10$ or $d = 50$. The values in the row marked by $d = 50$ look a little bigger. However, it is reasonable because the number of summands in the average estimate error for $d = 50$ is five times as large as that for $d = 10$. Again it appears that the EFM procedure achieves more precise estimators.

EXAMPLE 3 (A simple model). To illustrate the adaptivity of our algorithm to high dimension, we consider the following simple single-index model:

$$Y = (\boldsymbol{\beta}^\top \mathbf{X})^2 + \varepsilon. \tag{3.5}$$

The true parameter is $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$; $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$. Both homogeneous errors and heterogeneous ones are considered. In the former case, $\varepsilon \sim N(0, 0.2^2)$ and in the latter case, $\varepsilon = \exp(\sqrt{5}\boldsymbol{\beta}^\top \mathbf{X}/14)\widetilde{\varepsilon}$ with $\widetilde{\varepsilon} \sim N(0, 1)$. The latter case is designed to show whether our method can handle heteroscedasticity. A similar modeling setup was also used in Wang and Xia (2008), Example 5. The simulated results given in Table 3 are based on 250 replicates with a sample of $n = 100$ observations. An important observation from this simulation is that the proposed EFM approach still works even when the dimension of the parameter is equal to or slightly larger than the number of observations. It can be seen from Table 3 that our approach also performs well under the heteroscedasticity setup.

EXAMPLE 4 (An oscillating function model). A single-index model is designed as

$$Y = \sin(a\boldsymbol{\beta}^\top \mathbf{X}) + \varepsilon, \tag{3.6}$$

TABLE 3
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.5)

ε		$d = 10$	$d = 50$	$d = 100$	$d = 120$
$\varepsilon \sim N(0, 0.2^2)$	rMAVE	0.0318	0.3484	—	—
	EDR	0.0363	0.5020	—	—
	EFM	0.0272	0.2302	2.9409	5.0010
$\varepsilon \sim N(0, \exp(\frac{2X_1+X_2}{7}))$	rMAVE	0.3427	4.6190	—	—
	EDR	0.2542	2.1112	—	—
	EFM	0.2201	1.7937	4.1435	6.4973

— means that the values cannot be calculated by rMAVE and EDR because of high dimension.

where $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)^\top/\sqrt{5}$, $\mathbf{X}$ is generated from $N_d(2, \mathbf{I})$ and $\varepsilon \sim N(0, 0.2^2)$. The number of replications is 250 and the sample size $n = 400$. The simulation results are shown in Table 4. In these chosen values for a, we see that EFM performs better than rMAVE and EDR. But as is understood, more oscillating functions are more difficult to handle than those less oscillating functions.

EXAMPLE 5 (Comparison of variance). To make our simulation results comparable with those of Carroll et al. (1997), we mimic their simulation setup. Data of size 200 are generated according to the following model:

$$Y_i = \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \alpha Z_i + \varepsilon_i, \tag{3.7}$$

where $\mathbf{X}_i$ are trivariate with independent $U(0, 1)$ components, Z_i are independent of $\mathbf{X}_i$ and $Z_i = 0$ are for i odd and $Z_i = 1$ for i even, and ε_i follow a normal distribution $N(0, 0.01)$ independent of both $\mathbf{X}_i$ and Z_i. The parameters are taken to be $\boldsymbol{\beta} = (1, 1, 1)^\top/\sqrt{3}$, $\alpha = 0.3$, $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. Note that the EFM approach can still be applicable for this model as the conditionally centered response Y given Z has the model as, because of the independence between $\mathbf{X}$ and Z,

$$Y_i - E(Y_i|Z_i) = a + \sin\{\pi(\boldsymbol{\beta}^\top \mathbf{X}_i - A)/(B - A)\} + \varepsilon_i.$$

TABLE 4
Average estimation errors $\sum_{s=1}^{d} |\hat{\beta}_s - \beta_s|$ *for model* (3.6)

	$a = \pi/2$			$a = 3\pi/4$		
d	rMAVE	EDR	EFM	rMAVE	EDR	EFM
10	0.0981	0.0918	0.0737	0.0970	0.0745	0.0725
50	0.5247	0.6934	0.4355	0.6350	1.8484	0.5407

TABLE 5
Estimation for $\boldsymbol{\beta}$ *of model* (3.7) *based on two randomly chosen samples*

	One group of sample			**Another group of sample**		
	X_1	X_2	X_3	X_1	X_2	X_3
GPLSIM est.	0.595*	0.568*	0.569*	0.563*	0.574*	0.595*
GPLSIM s.e.	0.013*	0.013*	0.013*	0.010*	0.010*	0.010*
EFM est.	0.579	0.575	0.577	0.573	0.577	0.580
EFM s.e.	0.011	0.011	0.011	0.010	0.010	0.010

*The values are adopted from Carroll et al. (1997). We abbreviate "estimator" to "est." and "standard error" to "s.e.," which are computed from the sample version of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ defined in (2.10).

As Z_i are dummy variables, estimating $E(Y_i|Z_i)$ is simple. Thus, when we regard $Y_i - E(Y_i|Z_i)$ as response, the model is still a single-index model. Here the number of replications is 100. The method derived from Carroll et al. (1997) is referred to be the GLPSIM approach. The numerical results are reported in Table 5. It shows that compared with the GPLSIM estimates, the EFM estimates have smaller bias and smaller (or equal) variance. Also in this example both EFM and GPLSIM can provide reasonably accurate estimates.

Performance of profile quasi-likelihood ratio test. To illustrate how the profile quasi-likelihood ratio performs for linear hypothesis problems, we simulate the same data as above, except that we allow some components of the index to follow the null hypothesis:

$$H_0 : \beta_4 = \beta_5 = \cdots = \beta_d = 0.$$

We examine the power of the test under a sequence of the alternative hypotheses indexed by parameter δ as follows:

$$H_1 : \beta_4 = \delta, \qquad \beta_s = 0 \qquad \text{for } s \geq 5.$$

When $\delta = 0$, the alternative hypothesis becomes the null hypothesis.

We examine the profile quasi-likelihood ratio test under a sequence of alternative models, progressively deviating from the null hypothesis, namely, as δ increases. The power functions are calculated at the significance level: 0.05, using the asymptotic distribution. We calculate test statistics from 250 simulations by employing the fixed-point algorithm and find the percentage of test statistics greater than or equal to the associated quantile of the asymptotic distribution. The pictures in Figures 1, 2 and 3 illustrate the power function curves for two models under the given significance levels. The power curves increase rapidly with δ, which shows the profile quasi-likelihood ratio test is powerful. When δ is close to 0, the test sizes are all approximately the significance levels.

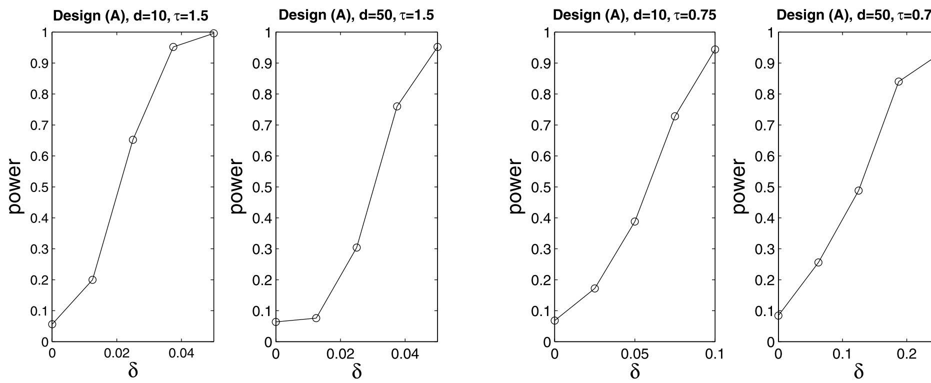

FIG. 1. *Simulation results for Design* (A) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* $\text{Beta}(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

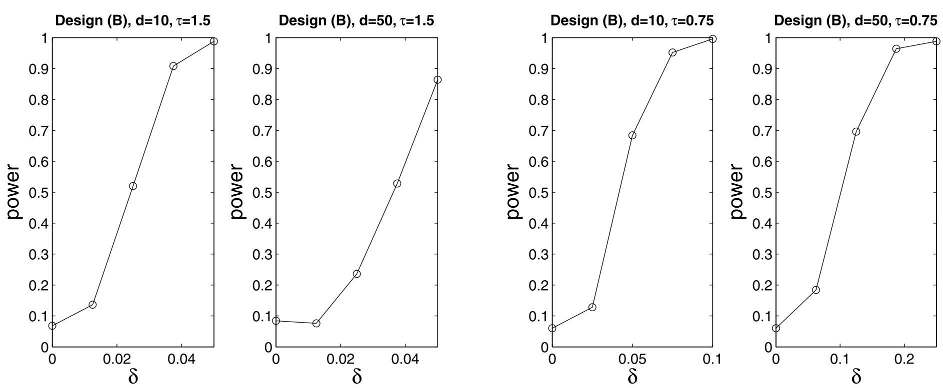

FIG. 2. *Simulation results for Design* (B) *in Example* 1. *The left graphs depict the case* $\tau = 1.5$ *with* τ *the first parameter in* $\text{Beta}(\tau, 1)$. *The right graphs are for* $\tau = 0.75$.

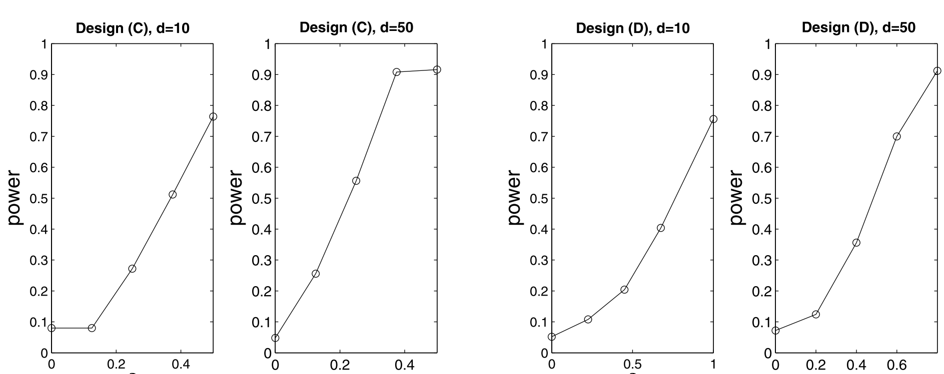

FIG. 3. *Simulation results for Example* 2. *The left graphs depict the case of Design* (C) *with parameter dimension being* 10 *and* 50. *The right graphs are for Design* (D).

3.3. *A real data example*. Income, to some extent, is considered as an index of a successful life. It is generally believed that demographic information, such as education level, relationship in the household, marital status, the fertility rate and gender, among others, has effects on amounts of income. For example, Murray (1997) illustrated that adults with higher intelligence have higher income. Kohavi (1996) predicted income using a Bayesian classifier offered by a machine learning algorithm. Madalozzo (2008) examined income differentials between married women and those who remain single or cohabit by using multivariate linear regression. Here we will use the single-index model to explore the relationship between income and some of its possible determinants.

We use the "Adult" database, which was extracted from the Census Bureau database and is available on website: http://archive.ics.uci.edu/ml/datasets/Adult. It was originally used to model income exceeds over USD 50,000/year based on census data. The purpose of using this example is to understand the personal income patterns and demonstrate the performance of the EFM method in real data analysis. After excluding a few missing data, the data set in our study includes 30,162 subjects. The selected explanatory variables are:

- *sex* (categorical): 1 = Male, 0 = Female.
- *native-country* (categorical): 1 = United-States, 0 = others.
- *work-class* (categorical): 1 = Federal-gov, 2 = Local-gov, 3 = Private, 4 = Self-emp-inc (self-employed, incorporated), 5 = Self-emp-not-inc (self-employed, not incorporated), 6 = State-gov.
- *marital-status* (categorical): 1 = Divorced, 2 = Married-AF-spouse (married, armed forces spouse present), 3 = Married-civ-spouse (married, civilian spouse present), 4 = Married-spouse-absent [married, spouse absent (exc. separated)], 5 = Never-married, 6 = Separated, 7 = Widowed.
- *occupation* (categorical): 1 = Adm-clerical (administrative support and clerical), 2 = Armed-Forces, 3 = Craft-repair, 4 = Exec-managerial (executive-managerial), 5 = Farming-fishing, 6 = Handlers-cleaners, 7 = Machine-op-inspct (machine operator inspection), 8 = Other-service, 9 = Priv-house-serv (private household services), 10 = Prof-specialty (professional specialty), 11 = Protective-serv, 12 = Sales, 13 = Tech-support, 14 = Transport-moving.
- *relationship* (categorical): 1 = Husband, 2 = Not-in-family, 3 = Other-relative, 4 = Own-child, 5 = Unmarried, 6 = Wife.
- *race* (categorical): 1 = Amer-Indian-Eskimo, 2 = Asian-Pac-Islander, 3 = Black, 4 = Other, 5 = White.
- *age* (integer): number of years of age and greater than or equal to 17.
- *fnlwgt* (continuous): The final sampling weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the United States.
- *education* (ordinal): 1 = Preschool (less than 1st Grade), 2 = 1st–4th, 3 = 5th–6th, 4 = 7th–8th, 5 = 9th, 6 = 10th, 7 = 11th, 8 = 12th (12th Grade no

Diploma), 9 = HS-grad (high school Grad-Diploma or Equiv), 10 = Some-college (some college but no degree), 11 = Assoc-voc (associate degree-occupational/vocational), 12 = Assoc-acdm (associate degree-academic program), 13 = Bachelors, 14 = Masters, 15 = Prof-school (professional school), 16 = Doctorate.

- *education-num* (continuous): Number of years of education.
- *capital-gain* (continuous): A profit that results from investments into a capital asset.
- *capital-loss* (continuous): A loss that results from investments into a capital asset.
- *hours-per-week* (continuous): Usual number of hours worked per week.

Note that all the explanatory variables up to "age" are categorical with more than two categories. As such, we use dummy variables to link up the corresponding categories. Specifically, for every original explanatory variable up to "age," we use dummy variables to indicate it in which the number of dummy variables is equal to the number of categories minus one. By doing so, we then have 41 explanatory variables, where the first 35 ones are dummy and the remaining ones are continuous. After a preliminary data check, we find that the explanatory variables X_{37} = "fnlwgt," X_{39} = "capital-gain" and X_{40} = "capital-loss" are very skewed to the left and the latter two often take zero value. Before fitting (3.8) we first make a logarithm transformation for these three variables to have log("fnlwgt"), log(1 + "capital-gain") and log(1 + "capital-loss"). To make the explanatory variables comparable in scale, we standardize each of them individually to obtain mean 0 and variance 1. Since "education" and "education-num" are correlated, "education" is dropped from the model and it results in a significantly smaller mean residual deviance.

The single-index model will be used to model the relationship between income and the relevant 43 predictors $\mathbf{X} = (X_1, \dots, X_{43})^\top$:

$$P(\text{“income”} > 50{,}000|\mathbf{X}) = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}], \tag{3.8}$$

where $Y = I(\text{“income”} > 50{,}000)$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{43})^\top$ and β_s represents the effect of the sth predictor. Formally, we are testing the effect of gender, that is,

$$H_0: \beta_1 = 0 \quad \longleftrightarrow \quad H_1: \beta_1 \neq 0. \tag{3.9}$$

The fixed-point iterative algorithm is employed to compute the estimate for $\boldsymbol{\beta}$. To illustrate further the practical implications of this approach, we compare our results to those obtained by using an ordinary logistic regression (LR). The coefficients of the two models are given in Table 6. To make the analyses presented in the table comparable, we consider two standardizations. First, we standardize every explanatory variable with mean 0 and variance 1 so that the coefficients can be used to compare the relative influence from different explanatory variables. However, such a standardization does not allow us to compare between the single-index

TABLE 6
Fitted coefficients for model (3.8) (estimated standard errors in parentheses)

Variables	**$\hat{\beta}$ of SIM**	**$\hat{\beta}$ of LR**
Sex	0.1102 (0.0028)	0.1975 (0.0181)
Native-country	0.0412 (0.0027)	0.0354 (0.0116)
Work-class		
Federal-gov	0.1237 (0.0059)	0.0739 (0.0108)
Local-gov	0.2044 (0.0065)	0.0155 (0.0135)
Private	−0.2603 (0.0075)	0.0775 (0.0200)
Self-em-inc	0.1252 (0.0068)	0.0520 (0.0112)
Self-emp-not-inc	0.1449 (0.0066)	−0.0157 (0.0147)
Marital-Status		
Divorced	−0.0353 (0.0061)	−0.0304 (0.0264)
Married-AF-spouse	0.0195 (0.0036)	0.0333 (0.0079)
Married-civ-spouse	0.3257 (0.0150)	0.4545 (0.0754)
Married-spouse-absent	−0.0115 (0.0029)	−0.0095 (0.0146)
Never-married	−0.1876 (0.0085)	−0.1452 (0.0370)
Separated	−0.0412 (0.0050)	−0.0221 (0.0179)
Occupation		
Adm-clerical	−0.0302 (0.0050)	0.0131 (0.0164)
Armed-Forces	−0.0086 (0.0031)	−0.0091 (0.0131)
Craft-repair	−0.0913 (0.0050)	0.0263 (0.0146)
Exec-managerial	0.1813 (0.0061)	0.1554 (0.0148)
Farming-fishing	−0.0370 (0.0036)	−0.0772 (0.0125)
Handlers-cleaners	−0.0947 (0.0033)	−0.0662 (0.0153)
Machine-op-inspct	−0.1067 (0.0038)	−0.0290 (0.0133)
Other-service	−0.1227 (0.0045)	−0.1192 (0.0195)
Priv-house-serv	−0.0501 (0.0020)	−0.0833 (0.0379)
Prof-specialty	0.2502 (0.0065)	0.1153 (0.0160)
Protective-serv	0.1954 (0.0061)	0.0508 (0.0095)
Sales	0.0316 (0.0050)	0.0615 (0.0147)
Tech-support	0.0181 (0.0037)	0.0619 (0.0102)
Relationship		
Husband	−0.1249 (0.0093)	−0.3264 (0.0254)
Not-in-family	−0.0932 (0.0093)	−0.2074 (0.0612)
Other-relative	−0.0958 (0.0038)	−0.1498 (0.0219)
Own-child	−0.2218 (0.0076)	−0.3769 (0.0498)
Unmarried	−0.1124 (0.0067)	−0.1739 (0.0446)
Race		
Amer-Indian-Eskimo	−0.0252 (0.0024)	−0.0226 (0.0109)
Asian-Pac-Islander	0.0114 (0.0030)	0.0062 (0.0101)
Black	−0.0300 (0.0024)	−0.0182 (0.0111)
Other	−0.0335 (0.0021)	−0.0286 (0.0129)

TABLE 6
(*Continued*)

Variables	$\hat{\beta}$ **of SIM**	$\hat{\beta}$ **of LR**
Age	0.2272 (0.0042)	0.1798 (0.0111)
Fnlwgt	0.0099 (0.0028)	0.0414 (0.0092)
Education-num	0.4485 (0.0045)	0.3732 (0.0122)
Capital-gain	0.2859 (0.0055)	0.2582 (0.0084)
Capital-loss	0.1401 (0.0042)	0.1210 (0.0078)
Hours-per-week	0.2097 (0.0035)	0.1823 (0.0101)

model and the ordinary logistic regression model. We then further normalize the coefficients to be with Euclidean norm 1, and then the estimates of their standard errors are also adjusted accordingly. The single-index model provides more reasonable results: $X_{38} =$ "education-num" has its strongest positive effect on income; those who got a bachelor's degree or higher seem to have much higher income than those with lower education level. In contrast, results derived from a logistic regression show that "married-civ-spouse" is the largest positive contributor.

Some other interesting conclusions could be obtained by looking at the output. Both "sex" and "native-country" have a positive effect. Persons who worked without pay in a family business, unpaid childcare and others earn a lower income than persons who worked for wages or for themselves. The "fnlwgt" attribute has a positive relation to income. Males are likely to make much more money than females. The expected sign for marital status except the *married* (married-AF-spouse, married-civ-spouse) is negative, given that the household production theory affirms that division of work is efficient when each member of a family dedicates his or her time to the more productive job. Men usually receive relatively better compensation for their time in the labor market than in home production. Thus, the expectation is that married women dedicate more time to home tasks and less to the labor market, and this would imply a different probability of working given the marital status choice.

Also "race" influences the income and Asian or Pacific Islanders seem to make more money than other races. And also, one's income significantly increases as working hours increase. Both "capital-gain" and "capital-loss" have positive effects, so we think that people make more money who can use more money to invest. The presence of young children has a negative influence on the income. "age" accounts for the experience effect and has a positive effect. Hence the conclusion based on the single-index model is consistent with what we expect.

To help with interpretation of the model, plots of $\boldsymbol{\beta}^\top \mathbf{X}$ versus predicted response probability and $\hat{g}(\boldsymbol{\beta}^\top \mathbf{X})$ are generated, respectively, and can be found on the right column in Figure 4. When the estimated single-index is greater than 0, $\hat{g}(\hat{\boldsymbol{\beta}}\mathbf{X})$ shows some degree of curvature. An alternative choice is to fit the data

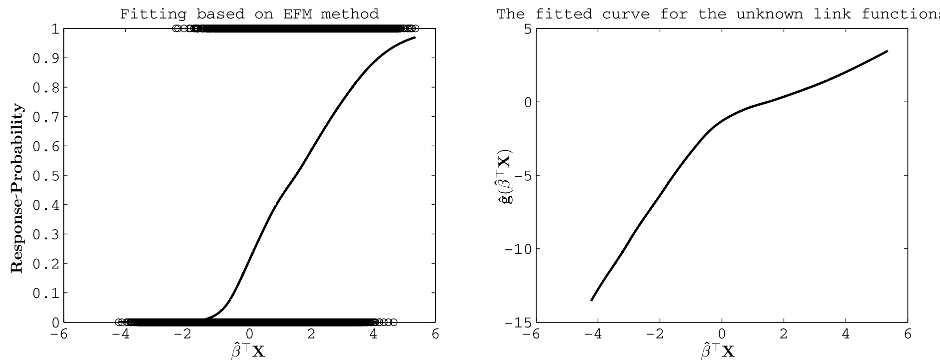

FIG. 4. *Adult data: The left graph is a plot of predicted response probability based on the single-index model. The right graph is the fitted curve for the unknown link function* $g(\cdot)$.

using generalized partially linear additive models (GPLAM) with nonparametric components of continuous explanatory variables. The relationships among "age," "fnlwgt," "capital-gain," "capital-loss" and "hours-per-week" all show nonlinearity. The mean residual deviances of SIM, LR and GPLAM are 0.7811, 0.6747 and 0.6240, respectively. SIM under study provides a slightly worse fit than the others. However, we note that LR is, up to a link function, linear about $\mathbf{X}$, and, according to the results of GPLAM, which is a more general model than LR, the actual relationship cannot have such a structure. SIM can reveal nonlinear structure. On the other hand, although the minimum mean residual deviance can be not surprisingly attained by GPLAM, this model has, respectively, $\approx$ 34 and 41 more degrees of freedom than SIM and LR have.

We now employ the quasi-likelihood ratio test to the test problem (3.9). The QLR test statistic is 166.52 with one degree of freedom, resulting in a P-value of $< 10^{-5}$. Hence this result provides strong evidence that gender has a significant influence on high income.

The Adult data set used in this paper is a rich data set. Existing work mainly focused on the prediction accuracy based on machine learning methods. We make an attempt to explore the semiparametric regression pattern suitable for the data. Model specification and variable selection merit further study.

APPENDIX: OUTLINE OF PROOFS

We first introduce some regularity conditions.

Regularity Conditions:

(a) $\mu(\cdot), V(\cdot), g(\cdot), \mathbf{h}(\cdot) = E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X} = \cdot)$ have two bounded and continuous derivatives. $V(\cdot)$ is uniformly bounded and bounded away from 0.

(b) Let $q(z, y) = \mu'(z)V^{-1}(z)\{y - \mu(z)\}$. Assume that $\partial q(z, y)/\partial z < 0$ for $z \in \mathbb{R}$ and y in the range of the response variable.

(c) The largest eigenvalue of $\mathbf{\Omega}_{22}$ is bounded away from infinity.

(d) The density function $f_{\boldsymbol{\beta}^\top \mathbf{x}}(\boldsymbol{\beta}^\top \mathbf{x})$ of random variable $\boldsymbol{\beta}^\top \mathbf{X}$ is bounded away from 0 on $T_{\boldsymbol{\beta}}$ and satisfies the Lipschitz condition of order 1 on $T_{\boldsymbol{\beta}}$, where $T_{\boldsymbol{\beta}} = \{\boldsymbol{\beta}^\top \mathbf{x} : \mathbf{x} \in T\}$ and T is a compact support set of $\mathbf{X}$.

(e) Let $Q^*[\boldsymbol{\beta}] = \int Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\}, y] f(y|\boldsymbol{\beta}^{0\top}\mathbf{x}) f(\boldsymbol{\beta}^{0\top}\mathbf{x})\, dy\, d(\boldsymbol{\beta}^{0\top}\mathbf{x})$ with $\boldsymbol{\beta}^0$ denoting the true parameter value and $Q[\mu, y] = \int_\mu^y \frac{s-y}{V\{\mu^{-1}(s)\}}\, ds$. Assume that $Q^*[\boldsymbol{\beta}]$ has a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, and

$$E\Big[\sup_{\boldsymbol{\beta}^{(1)}} \sup_{\boldsymbol{\beta}^\top \mathbf{X}} |\mu'\{g(\boldsymbol{\beta}^\top \mathbf{X})\} V^{-1}\{g(\boldsymbol{\beta}^\top \mathbf{X})\}[Y - \mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]|^2\Big] < \infty$$

and $E\|\mathbf{X}\|^2 < \infty$.

(f) The kernel K is a bounded and symmetric density function with a bounded derivative, and satisfies

$$\int_{-\infty}^{\infty} t^2 K(t)\, dt \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} |t|^j K(t)\, dt < \infty, \qquad j = 1, 2, \ldots.$$

Condition (a) is some mild smoothness conditions on the involved functions of the model. We impose condition (b) to guarantee that the solutions of (2.1), $\hat{g}(t)$ and $\hat{g}'(t)$, lie in a compact set. Condition (c) implies that the second moment of estimating equation (2.7), $\mathrm{tr}(\mathbf{J}^\top \mathbf{\Omega} \mathbf{J})$, is bounded. Then the CLT can be applied to $G(\boldsymbol{\beta})$. Condition (d) means that $\mathbf{X}$ may have discrete components and the density function of $\boldsymbol{\beta}^\top \mathbf{X}$ is positive, which ensures that the denominators involved in the nonparametric estimators, with high probability, are bounded away from 0. The uniqueness condition in condition (e) can be checked in the following case for example. Assume that Y is a Poisson variable with mean $\mu\{g(\boldsymbol{\beta}^\top \mathbf{x})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{x})\}$. The maximizer β_0 of $Q^*[\boldsymbol{\beta}]$ is equal to the solution of the equation $E[E\{[\exp\{g(\boldsymbol{\beta}^{0\top}\mathbf{X})\} - \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}]g'(\boldsymbol{\beta}^\top \mathbf{X})\}\mathbf{J}^\top \mathbf{X}|\boldsymbol{\beta}^{0\top}\mathbf{X}\}] = 0$. $\boldsymbol{\beta}_0$ is unique when $g'(\cdot)$ is not a zero-valued constant function and the matrix $\mathbf{J}^\top E(\mathbf{X}\mathbf{X}^\top)\mathbf{J}$ is not singular. Under the second part of condition (e), it is permissible to interchange differentiation and integration when differentiating $E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\}, Y]]$. Condition (f) is a commonly used smoothness condition, including the Gaussian kernel and the quadratic kernel. All of the conditions can be relaxed at the expense of longer proofs.

Throughout the Appendix, $Z_n = \mathcal{O}_P(a_n)$ denotes that $a_n^{-1} Z_n$ is bounded in probability and the derivation for the order of Z_n is based on the fact that $Z_n = \mathcal{O}_P\{\sqrt{E(Z_n^2)}\}$. Therefore, it allows to apply the Cauchy–Schwarz inequality to the quantity having stochastic order a_n.

A.1. Proof of Proposition 1. We outline the proof here, while the details are given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(i) Conditions (a), (b), (d) and (f) are essentially equivalent conditions given by Carroll, Ruppert and Welsh (1998), and as a consequence the derivation of bias and variance for $\hat{g}(\boldsymbol{\beta}^\top \mathbf{x})$ and $\hat{g}'(\boldsymbol{\beta}^\top \mathbf{x})$ is similar to that of Carroll, Ruppert and Welsh (1998).

(ii) The first equation of (2.1) is

$$0 = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\mu'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}$$
$$\times V^{-1}\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}[Y_j - \mu\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\}].$$

Taking derivatives with respect to $\boldsymbol{\beta}^{(1)}$ on both sides, direct observations lead to

$$\frac{\partial \hat{\alpha}_0}{\partial \boldsymbol{\beta}^{(1)}} = \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1}\{A_1(\boldsymbol{\beta}^\top \mathbf{x}) + A_2(\boldsymbol{\beta}^\top \mathbf{x}) + A_3(\boldsymbol{\beta}^\top \mathbf{x})\},$$

where

$$B(\boldsymbol{\beta}^\top \mathbf{x}) = -\sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\},$$

$$A_1(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\mathbf{J}^\top(\mathbf{X}_j - \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}\hat{\alpha}_1,$$

$$A_2(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}) q_z'\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$
$$\times (\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\frac{\partial \hat{\alpha}_1}{\partial \boldsymbol{\beta}^{(1)}},$$

$$A_3(\boldsymbol{\beta}^\top \mathbf{x}) = \sum_{j=1}^{n} h^{-1} K_h'(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x})\mathbf{J}^\top(\mathbf{X}_j - \mathbf{x}) q\{\hat{\alpha}_0 + \hat{\alpha}_1(\boldsymbol{\beta}^\top \mathbf{X}_j - \boldsymbol{\beta}^\top \mathbf{x}), Y_j\}$$

with $K_h'(\cdot) = h^{-1}K'(\cdot/h)$. Note that $\partial\hat{\alpha}_0/\partial\boldsymbol{\beta}^{(1)} = \partial\hat{g}(\boldsymbol{\beta}^\top \mathbf{x})/\partial\boldsymbol{\beta}^{(1)}$; then we have

$$\begin{aligned}\frac{\partial \hat{g}(\boldsymbol{\beta}^\top \mathbf{x})}{\partial \boldsymbol{\beta}^{(1)}} &= \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x}) \\ &\quad + \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_2(\boldsymbol{\beta}^\top \mathbf{x}) + \{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_3(\boldsymbol{\beta}^\top \mathbf{x}).\end{aligned} \tag{A.1}$$

We will prove that

$$\begin{aligned} &E\|\{B(\boldsymbol{\beta}^\top \mathbf{x})\}^{-1} A_1(\boldsymbol{\beta}^\top \mathbf{x}) - g'(\boldsymbol{\beta}^\top \mathbf{x})\mathbf{J}^\top\{\mathbf{x} - \mathbf{h}(\boldsymbol{\beta}^\top \mathbf{x})\}\|^2 \\ &\quad = \mathcal{O}_P(h^4 + n^{-1}h^{-3}),\end{aligned} \tag{A.2}$$

the second term in (A.1) is of order $\mathcal{O}_P(h^4 + n^{-1}h)$, and the third term is of order $\mathcal{O}_P(h^4 + n^{-1}h^{-3})$. The combination of (A.1) and these three results can directly lead to result (ii) of Proposition 1. The detailed proof is summarized in three steps and is given in the supplementary materials [Cui, Härdle and Zhu (2010)].

(iii) By mimicking the proof of (ii), we can show that (iii) holds. See supplementary materials for details.

A.2. Proofs of (2.6) and (2.7). It is proved in the supplementary materials [Cui, Härdle and Zhu (2010)].

A.3. Proof of Theorem 2.1. (i) Note that the estimating equation defined in (2.6) is just the gradient of the following quasi-likelihood:

$$\hat{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]$$

with $Q[\mu, y] = \int^{\mu} \frac{y-s}{V\{\mu^{-1}(s)\}}\, ds$ and $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. Then for $\boldsymbol{\beta}^{(1)}$ satisfying $(\sqrt{1-\|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\top})^\top \in \Theta$, we have

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\max_{\boldsymbol{\beta}^{(1)}} \hat{Q}(\boldsymbol{\beta}).$$

The proof is based on Theorem 5.1 in Ichimura (1993). In that theorem the consistency of $\boldsymbol{\beta}^{(1)}$ is proved by means of proving that

$$\text{(A.3)} \qquad \sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] \right| = o_P(1),$$

$$\text{(A.4)} \qquad \sup_{\boldsymbol{\beta}^{(1)}} \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1)$$

and

$$\text{(A.5)} \qquad \left| \frac{1}{n} \sum_{i=1}^{n} Q[\mu\{\hat{g}(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i] - \frac{1}{n} \sum_{i=1}^{n} E[Q[\mu\{g(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\}, Y_i]] \right| = o_P(1).$$

Regarding the validity of (A.5), this directly follows from (A.3) and (A.4). The type of uniform convergence result such as (A.4) has been well established in the literature; see, for example, Andrews (1987). We now verify the validity of (A.3), which reduces to showing the uniform convergence of the estimator $\hat{g}(t)$ under condition (e) [see Ichimura (1993)]. This can be obtained in a similar way as in Kong, Linton and Xia (2010), taking into account that the regularity conditions imposed in Theorem 2.1 are stronger than the corresponding ones in that paper.

(ii) Recall the notation $\mathbf{J}$, $\boldsymbol{\Omega}$ and $\mathbf{G}(\boldsymbol{\beta})$ introduced in Section 2. By (2.7), we have shown that

$$\text{(A.6)} \qquad \sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0}) = \frac{1}{\sqrt{n}}\{\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}\}^{+} \mathbf{G}(\boldsymbol{\beta}) + o_P(1).$$

Theorem 2.1 follows directly from the above asymptotic expansion and the fact that $E\{\mathbf{G}(\boldsymbol{\beta})\mathbf{G}^\top(\boldsymbol{\beta})\} = n\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J}$. □

A.4. Proof of Corollary 1. The asymptotic covariance of $\hat{\boldsymbol\beta}$ can be obtained by adjusting the asymptotic covariance of $\hat{\boldsymbol\beta}^{(1)}$ via the multivariate delta method, and is of form $\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top$. Next we will compare this asymptotic covariance with that (denoted by $\boldsymbol\Omega^+$) given in Carroll et al. (1997). Write $\boldsymbol\Omega$ as

$$\boldsymbol\Omega = \begin{pmatrix} \boldsymbol\Omega_{11} & \boldsymbol\Omega_{12} \\ \boldsymbol\Omega_{21} & \boldsymbol\Omega_{22} \end{pmatrix},$$

where $\boldsymbol\Omega_{22}$ is a $(d-1)\times(d-1)$ matrix. We will next investigate two cases, respectively: $\det(\boldsymbol\Omega_{22})\neq 0$ and $\det(\boldsymbol\Omega_{22})=0$. Let $\boldsymbol\alpha = -\boldsymbol\beta^{(1)}/\sqrt{1-\|\boldsymbol\beta^{(1)}\|^2} = -\boldsymbol\beta^{(1)}/\beta_1$.

Consider the case that $\det(\boldsymbol\Omega_{22})\neq 0$. Because $\operatorname{rank}(\boldsymbol\Omega)=d-1$, $\det(\boldsymbol\Omega_{11}\boldsymbol\Omega_{22}-\boldsymbol\Omega_{21}\boldsymbol\Omega_{12})=0$. Note that $\boldsymbol\Omega_{22}$ is nondegenerate; it can be easily shown that $\boldsymbol\Omega_{11}=\boldsymbol\Omega_{12}\boldsymbol\Omega_{22}^{-1}\boldsymbol\Omega_{21}$. Combining this with the following fact:

$$\begin{aligned}\mathbf{J}^\top\boldsymbol\Omega\mathbf{J} &= (\boldsymbol\alpha \quad \mathbf{I}_{d-1})\begin{pmatrix} \boldsymbol\Omega_{11} & \boldsymbol\Omega_{12} \\ \boldsymbol\Omega_{21} & \boldsymbol\Omega_{22} \end{pmatrix}\begin{pmatrix} \boldsymbol\alpha^\tau \\ \mathbf{I}_{d-1} \end{pmatrix} \\ &= \boldsymbol\Omega_{22} + (\boldsymbol\Omega_{21}/\sqrt{\boldsymbol\Omega_{11}} + \sqrt{\boldsymbol\Omega_{11}}\boldsymbol\alpha)(\boldsymbol\Omega_{12}/\sqrt{\boldsymbol\Omega_{11}} + \sqrt{\boldsymbol\Omega_{11}}\boldsymbol\alpha^\top) - \boldsymbol\Omega_{21}\boldsymbol\Omega_{12}/\boldsymbol\Omega_{11},\end{aligned}$$

we can get that $\mathbf{J}^\top\boldsymbol\Omega\mathbf{J}$ is nondegenerate. In this situation, its inverse $(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+$ is just the ordinary inverse $(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1}$. Then $\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top = \{\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1/2}\}\{(\mathbf{J}^\top\times\boldsymbol\Omega\mathbf{J})^{-1/2}\mathbf{J}^\top\}$, a full-rank decomposition. Then

$$\begin{aligned}\{\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top\}^+ &= \{\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1/2}\} \\ &\quad\times\{(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1/2}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1}\mathbf{J}^\top\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1/2}\}^{-1} \\ &\quad\times\{(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^{-1/2}\mathbf{J}^\top\} \\ &= \mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top\boldsymbol\Omega\mathbf{J}(\mathbf{J}^\top\mathbf{J})^{-1}\mathbf{J}^\top \\ &= \boldsymbol\Omega.\end{aligned}$$

This means that $\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top = \boldsymbol\Omega^+$.

When $\det(\boldsymbol\Omega_{22})=0$, we can obtain that

$$\boldsymbol\Omega^+ = \begin{pmatrix} 1/\boldsymbol\Omega_{11} + \boldsymbol\Omega_{12}\boldsymbol\Omega_{22.1}^+\boldsymbol\Omega_{21}/\boldsymbol\Omega_{11}^2 & -\boldsymbol\Omega_{12}\boldsymbol\Omega_{22.1}^+/\boldsymbol\Omega_{11} \\ -\boldsymbol\Omega_{22.1}^+\boldsymbol\Omega_{21}/\boldsymbol\Omega_{11} & \boldsymbol\Omega_{22.1}^+ \end{pmatrix}$$

with $\boldsymbol\Omega_{22.1} = \boldsymbol\Omega_{22} - \boldsymbol\Omega_{21}\boldsymbol\Omega_{12}/\boldsymbol\Omega_{11}$. Write $\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top$ as

$$\begin{pmatrix} \boldsymbol\alpha^\top(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\boldsymbol\alpha & \boldsymbol\alpha^\top(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+ \\ (\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\boldsymbol\alpha & (\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+ \end{pmatrix}.$$

Note that $\mathbf{J}^\top\boldsymbol\Omega\mathbf{J} = \boldsymbol\Omega_{22.1} + (\boldsymbol\Omega_{21}/\sqrt{\boldsymbol\Omega_{11}} + \sqrt{\boldsymbol\Omega_{11}}\boldsymbol\alpha)(\boldsymbol\Omega_{12}/\sqrt{\boldsymbol\Omega_{11}} + \sqrt{\boldsymbol\Omega_{11}}\boldsymbol\alpha^\top)$, so $\mathbf{J}^\top\boldsymbol\Omega\mathbf{J} \geq \boldsymbol\Omega_{22.1}$. Combining this with $\operatorname{rank}(\boldsymbol\Omega_{22}) = d-2$, we have that $(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+ \leq \boldsymbol\Omega_{22.1}^+$. It is easy to check that $\boldsymbol\alpha^\top\boldsymbol\Omega_{22.1}=0$, so $\boldsymbol\alpha\perp\operatorname{span}(\boldsymbol\Omega_{22.1})$ and $\boldsymbol\alpha^\top\boldsymbol\Omega_{22.1}^+\boldsymbol\alpha=0$, and then $\boldsymbol\alpha^\top(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+=0$. In this situation, $\mathbf{J}(\mathbf{J}^\top\boldsymbol\Omega\mathbf{J})^+\mathbf{J}^\top \leq \boldsymbol\Omega^+$ and the stick less-than sign holds since $\mathbf{J}^\top\boldsymbol\Omega\mathbf{J}\neq\boldsymbol\Omega_{22.1}$ and $1/\boldsymbol\Omega_{11}>0$. □

A.5. Proof of Theorem 2.2. Under H_0, we can rewrite the index vector as $\boldsymbol{\beta} = [\mathbf{e}\ \mathbf{B}]^\top(\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\tau})^\top$ where $\mathbf{e} = (1, 0, \dots, 0)^\top$ is an r-dimensional vector,

$$\mathbf{B} = \begin{pmatrix} \mathbf{0}^\top & 0 \\ \mathbf{I}_{r-1} & \mathbf{0} \end{pmatrix}$$

is an $r \times (d-1)$ matrix and $\boldsymbol{\omega}^{(1)} = (\beta_2, \dots, \beta_r)^\top$ is an $(r-1) \times 1$ vector. Let $\boldsymbol{\omega} = (\sqrt{1-\|\boldsymbol{\omega}^{(1)}\|^2}, \boldsymbol{\omega}^{(1)\top})^\top$. So under H_0 the estimator is also the local maximizer $\hat{\boldsymbol{\omega}}$ of the problem

$$\hat{Q}([\mathbf{e} \quad \mathbf{B}]^\top \hat{\boldsymbol{\omega}}) = \sup_{\|\boldsymbol{\omega}^{(1)}\|<1} \hat{Q}([\mathbf{e} \quad \mathbf{B}]^\top \boldsymbol{\omega}).$$

Expanding $\hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}})$ at $\hat{\boldsymbol{\beta}}^{(1)}$ by a Taylor's expansion and noting that $\partial \hat{Q}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{(1)}|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} = 0$, then $\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) = T_1 + T_2 + o_P(1)$, where

$$T_1 = -\frac{1}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \frac{\partial^2 \hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}\bigg|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}} (\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}),$$

$$T_2 = \frac{1}{6}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \times \frac{\partial\{(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})^\top \partial^2\hat{Q}(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}^{(1)}\,\partial\boldsymbol{\beta}^{(1)\tau})|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}^{(1)}}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}})\}}{\partial\boldsymbol{\beta}^{(1)}}.$$

Assuming the conditions in Theorem 2.1 and under the null hypothesis H_0, it is easy to show that

$$\sqrt{n}(\mathbf{B}^\top\hat{\boldsymbol{\omega}} - \mathbf{B}^\top\boldsymbol{\omega}) = \frac{1}{\sqrt{n}}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^+\mathbf{G}(\boldsymbol{\beta}) + o_P(1).$$

Combining this with (A.6), under the null hypothesis H_0,

$$\begin{aligned}\sqrt{n}&(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) \\ &= \frac{1}{\sqrt{n}}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\{\mathbf{I}_{d-1} - (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2}\mathbf{B}^\top\mathbf{B}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\} \\ &\quad \times (\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{A.7}$$

Since $\frac{1}{\sqrt{n}}\mathbf{G}(\boldsymbol{\beta}) = \mathcal{O}_P(1)$, $\frac{\partial^2\hat{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}\partial\boldsymbol{\beta}^{(1)\tau}}|_{\boldsymbol{\beta}^{(1)}} = -n\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J} + o_P(n)$ and matrix $\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}$ has eigenvalues uniformly bounded away from 0 and infinity, we have $\|\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}\| = \mathcal{O}_P(n^{-1/2})$ and then $|T_2| = o_P(1)$. Combining this and (A.7), we have

$$\begin{aligned}\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top\hat{\boldsymbol{\omega}}) &= \frac{n}{2}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)})^\top\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J}(\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{B}^\top\hat{\boldsymbol{\omega}}^{(1)}) \\ &= \frac{n}{2}\mathbf{G}^\top(\boldsymbol{\beta})(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{P}(\mathbf{J}^\top\boldsymbol{\Omega}\mathbf{J})^{1/2+}\mathbf{G}(\boldsymbol{\beta})\end{aligned}$$

with $\mathbf{P} = \mathbf{I}_{d-1} - (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2} \mathbf{B}^\top \mathbf{B} (\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+}$. Here $\mathbf{P}$ is idempotent having rank $d - r$, so it can be written as $\mathbf{P} = \mathbf{S}^\top \mathbf{S}$ where $\mathbf{S}$ ia a $(d - r) \times (d - 1)$ matrix satisfying $\mathbf{S}\mathbf{S}^\top = \mathbf{I}_{d-r}$. Consequently,

$$2\{\hat{Q}(\hat{\boldsymbol{\beta}}) - \hat{Q}(\mathbf{B}^\top \hat{\boldsymbol{\omega}})\} = \left(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\right)^\top \left(\sqrt{n}\mathbf{S}(\mathbf{J}^\top \boldsymbol{\Omega} \mathbf{J})^{1/2+} \mathbf{G}(\boldsymbol{\beta})\right) \xrightarrow{\mathcal{L}} \chi^2(d - r).$$

Acknowledgments. The authors thank the Associate Editor and two referees for their constructive comments and suggestions which led to a great improvement over an early manuscript.

SUPPLEMENTARY MATERIAL

Supplementary materials (DOI: 10.1214/10-AOS871SUPP; .pdf). Complete proofs of Proposition 1, (2.6) and (2.7).

REFERENCES

ANDREWS, D. W. K. (1987). Conssitency in nonlinear econometric models: A genetic uniform law of large numbers. *Econometrica* **55** 1465–1471. MR0923471

CARROLL, R. J., RUPPERT, D. and WELSH, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93** 214–227. MR1614624

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 447–489. MR1467842

CHANG, Z. Q., XUE, L. G. and ZHU, L. X. (2010). On an asymptotically more efficient estimation of the single-index model. *J. Multivariate Anal.* **101** 1898–1901. MR2651964

CUI, X., HÄRDLE, W. and ZHU, L. (2010). Supplementary materials for "The EFM approach for single-index models." DOI:10.1214/10-AOS871SUPP.

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

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150. MR1325121

FAN, J. and JIANG, J. (2007). Nonparametric inference with generalized likelihood ratio test. *Test* **16** 409–478. MR2365172

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

HÄRDLE, W. and MAMMEN, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21** 1926–1947. MR1245774

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models. *J. Amer. Statist. Assoc.* **93** 1461–1474. MR1666641

HÄRDLE, W., MAMMEN, E. and PROENCA, I. (2001). A bootstrap test for single index models. *Statistics* **35** 427–452. MR1880174

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HEYDE, C. C. (1997). *Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York. MR1461808

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of a single-index model with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104

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

HRISTACHE, M., JUDITSKY, A., POLZEHL, J. and SPOKOINY, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** 1537–1566. MR1891738

HUH, J. and PARK, B. U. (2002). Likelihood-based local polynomial fitting for single-index models. *J. Multivariate Anal.* **80** 302–321. MR1889778

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

KANE, M., HOLT, J. and ALLEN, B. (2004). Results concerning the generalized partially linear single-index model. *J. Stat. Comput. Simul.* **72** 897–912. MR2100843

KOHAVI, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 202–207. AAAI Press, Menlo Park, CA.

KONG, E., LINTON, O. and XIA, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* **26** 1529–1564. MR2684794

LIN, W. and KULASEKERA, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94** 496–501. MR2380574

MADALOZZO, R. C. (2008). An analysis of income differentials by marital status. *Estudos Econôicos* **38** 267–292.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Champman & Hall, London.

MURRAY, C. (1997). IQ and economic success. *The Public Interest* **128** 21–35.

POLZEHL, J. and SPERLICH, S. (2009). A note on structural adaptive dimension reduction. *J. Stat. Comput. Simul.* **79** 805–818. MR2751594

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117

WANG, H. and XIA, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103** 811–821. MR2524332

WANG, J. L., XUE, L. G., ZHU, L. X. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322

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

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

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

ZHOU, J. and HE, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451

ZHU, L. X. and XUE, L. G. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 549–570. MR2278341

ZHU, L. P. and ZHU, L. X. (2009a). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.* **100** 862–875. MR2498719

ZHU, L. P. and ZHU, L. X. (2009b). On distribution weighted partial least squares with diverging number of highly correlated predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 525–548. MR2649607

X. CUI
SCHOOL OF MATHEMATICS
AND COMPUTATIONAL SCIENCE
SUN YAT-SEN UNIVERSITY
GUANGZHOU
GUANGDONG PROVINCE, 510275
P.R. CHINA
E-MAIL: cuixia@mail.sysu.edu.cn

W. K. HÄRDLE
CASE-CENTER FOR APPLIED STATISTICS
AND ECONOMICS
HUMBOLDT-UNIVERSITÄT ZU BERLIN
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
SPANDAUER STR. 1
10178 BERLIN
GERMANY
E-MAIL: haerdle@wiwi.hu-berlin.de

L. ZHU
FSC1207, FONG SHU CHUEN BUILDING
DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST UNIVERSITY
KOWLOON TONG
HONG KONG
P.R. CHINA
E-MAIL: lzhu@hkbu.edu.hk