Journal of Multivariate Analysis 102 (2011) 372–385

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A profile-type smoothed score function for a varying coefficient partially linear model

Gaorong Li[a], Sanying Feng[b], Heng Peng[c,*]

[a] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China
[b] College of Mathematics and Science, Luoyang Normal University, Luoyang 471022, China
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 12 April 2010
Available online 15 October 2010

AMS 2000 subject classifications:
primary 62G08
62H12
secondary 62G20

Keywords:
Varying coefficient partially linear model
Local likelihood
Profile-type smoothed score function
Confidence region
Curse of dimensionality

ABSTRACT

The varying coefficient partially linear model is considered in this paper. When the plug-in estimators of coefficient functions are used, the resulting smoothing score function becomes biased due to the slow convergence rate of nonparametric estimations. To reduce the bias of the resulting smoothing score function, a profile-type smoothed score function is proposed to draw inferences on the parameters of interest without using the quasi-likelihood framework, the least favorable curve, a higher order kernel or under-smoothing. The resulting profile-type statistic is still asymptotically Chi-squared under some regularity conditions. The results are then used to construct confidence regions for the parameters of interest. A simulation study is carried out to assess the performance of the proposed method and to compare it with the profile least-squares method. A real dataset is analyzed for illustration.

1. Introduction

Semiparametric models with a large number of predictors have frequently appeared in contemporary statistical studies. Various semiparametric models have been proposed to achieve a balance between modeling bias and the "curse of dimensionality". One model of importance is the varying coefficient partially linear model (VCPLM). Suppose that Y is a response variable and $(U, \mathbf{X}^T, \mathbf{Z}^T)$ are the associated covariates. The VCPLM takes the form:

$$Y = \mathbf{X}^T\alpha(U) + \mathbf{Z}^T\beta + \varepsilon, \tag{1.1}$$

where $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^T$ is a q-dimensional vector of unknown regression functions, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional vector of unknown regression coefficients and ε is an independent random error with $E(\varepsilon|\mathbf{X}, \mathbf{Z}, U) = 0$ almost certain. Model (1.1) consists of the unknown regression parameter vector β that serves as the parameter of interest and the unknown coefficient function $\alpha(\cdot)$ that is taken as the nonparametric nuisance component.

The VCPLM is, of course, an extension of the partially linear model and the varying coefficient model [6] and has attracted much attention. Examples can be found in the studies of Li et al. [9], Xia et al. [21], Ahmad et al. [1], Fan and Huang [5], You and Chen [22], Li and Liang [10], and Zhou and Liang [23]. Lam and Fan [8] investigated the generalized varying coefficient partially linear model (GVCPLM) when the number p of the parameters β grows with the sample size. They consider the

[*] Corresponding author.
E-mail address: hpeng@math.hkbu.edu.hk (H. Peng).

doi:10.1016/j.jmva.2010.10.007

profile likelihood ratio inference for the GVCPLM with a growing number of parameters. Li et al. [11] have also employed the empirical likelihood to study this model with a growing dimension p of parameters of interest.

In this paper, we are interested in inferring the parameter vector β. To investigate the accuracy of the estimator of β, the confidence region of β often needs to be constructed. One classical method is to use a normal approximation and sandwich formula and the other typical nonparametric approach is to apply the empirical likelihood method. If the normal approximation is used, we need to estimate the limiting variance or covariance matrix of the estimate of the parameters of interest. However, in nonparametric and semiparametric regression settings, the estimate of the variance or the covariance matrix is often complicated and inaccurate. In addition, the confidence region derived from the limiting normal distribution is predetermined to be symmetric, which may not be adequate when the underlying distribution is typically asymmetric. To avoid estimating the variance or the covariance matrix, the likelihood ratio or empirical likelihood is often used as an alternative. When the model has a known likelihood (or a known quasi-likelihood framework), the profile likelihood can be applied; see, for example, [15,17,16,3]. In this case, if the "least favorable cure" for the nonparametric function can be well defined and consistently estimated, the limit of the likelihood ratio is tractable. In addition, although the empirical likelihood method avoids estimating the variance or the covariance estimate of the parameters of interest and has many advantages in the construction of confidence regions, the limit of the empirical log-likelihood ratio is no longer a chi-square variable while a weighted sum of chi-square variables with unknown weights when there are infinite-dimensional nuisance functions. Thus, there is a need to investigate bias-corrected techniques (see [25,12,24]). Furthermore, when the semiparametric model cannot be expressed as a regression framework, one has to estimate a conditional expectation, making the bias-correction procedures complicated and likely to be inefficient. This motivates us to propose a new method for drawing inferences for β.

To overcome the problems discussed above, an alternative approach via the smoothed score function was proposed by Manski [13,14]. However, the smoothed score function cannot be used to draw inferences for β due to the slow convergence rate of the nonparametric estimation. In this paper, we investigate a profile-type smoothed score function that is different from those proposed by Severini and Staniswalis [16] and Severini and Wong [17]. This general methodology can be employed in many semiparametric models in which nonparametric nuisance functions (infinite-dimensional nuisance parameters), such as the varying coefficient partially linear model (1.1) are included. Unlike prior analysis of relevant models which have an infinite-dimensional nuisance parameter (e.g., [17,16,8]), the following conditions need not be assumed in investigating this profile-type smoothed score function: the distributions of the variables involved are given, a quasi-likelihood framework applies, a higher order kernel and under-smoothing are used to reduce the bias, and the least favorable curve applies.

The rest of the paper is organized as follows. In Section 2, the varying coefficient partially linear model is considered and the profile-type smoothed score function for parameters of interest is proposed. In Section 3, the theoretical results of the resulting profile-type statistic are investigated under some regularity conditions. In Section 4, simulation studies are carried out to assess the performance of the proposed method. A real data example is given for illustration. The technical proofs of the main theoretical results are relegated to the Appendix.

2. Model and methodology

Let $\{(Y_i; \mathbf{X}_i^T, \mathbf{Z}_i^T, U_i), 1 \le i \le n\}$ be an independent identically distributed (i.i.d.) random sample which comes from model (1.1). That is,

$$Y_i = \mathbf{X}_i^T \alpha(U_i) + \mathbf{Z}_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) random errors with $E(\varepsilon_i|\mathbf{X}_i, \mathbf{Z}_i, U_i) = 0$. Note that in this model, the distribution and variance of ε_i are not specified, and hence the profile (quasi) likelihood method is unavailable. Let $g(\cdot)$ be the unknown density function of ε. To draw inferences for the unknown parameter vector β, we first introduce the score function as an auxiliary vector

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log g(\varepsilon_i(\beta)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\beta}g(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\mathbf{Z}_i,\end{aligned} \tag{2.2}$$

where $g'(\cdot)$ denotes the derivative of $g(\cdot)$ with regard to a parameter vector β and

$$\eta_i(\beta) = \frac{\partial}{\partial\beta}\log g(\varepsilon_i(\beta)) = -\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\mathbf{Z}_i.$$

Note that $\{\eta_i(\beta), 1 \le i \le n\}$ are independent and $E[\eta_i(\beta)] = 0$ when $\alpha(u)$ and β are respectively the true coefficient function and the parameter vector. Thus, we can construct an estimating equation $\sum_{i=1}^{n}\eta_i(\beta) = 0$.

Because $\eta_i(\beta)$ in fact contains the unknown $g(\cdot), g'(\cdot), \alpha(\cdot)$, a natural way of solving this problem is to replace them with their consistent estimators. To this end, we first estimate the coefficient function using a local polynomial smoother (see [4]). Rewrite model (2.1) as

$$Y_i - \mathbf{Z}_i^T\beta = \mathbf{X}_i^T\alpha(U_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\alpha(u) = (\alpha_1(u), \ldots, \alpha_q(u))^T$. Obviously, if β is known, (2.3) is a version of the usual varying coefficient model. Thus, the local linear regression approximation can be used to estimate the coefficient functions $\{\alpha_j(\cdot),\ j = 1, \ldots, q\}$. For v in a neighborhood of u, approximate each $\alpha_j(v)$ by

$$\alpha_j(v) \approx \alpha_j(u) + \alpha_j'(u)(v-u) \equiv a_j + b_j(v-u), \quad j = 1, \ldots, q. \tag{2.4}$$

Denote $\mathbf{a} = (a_1, \ldots, a_q)^T$ and $\mathbf{b} = (b_1, \ldots, b_q)^T$. For any fixed β, a local linear fit is defined as the following solution of the weighted least-squares problem: finding $\mathbf{a}$ and $\mathbf{b}$ to minimize

$$\sum_{i=1}^n \{Y_i - \mathbf{X}_i^T(\mathbf{a} + \mathbf{b}(U_i - u)) - \mathbf{Z}_i^T\beta\}^2 K_h(U_i - u), \tag{2.5}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h represents the size of the local neighborhood called a bandwidth. The kernel function is introduced to reflect the fact that the local model (2.4) is only applied to the data around u. It gives a larger weight to the data closer to the point u. Let $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ be the solutions of the minimization of (2.5). Then

$$[\hat{\mathbf{a}}^T, h\hat{\mathbf{b}}^T]^T = (\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u(Y - \mathbf{Z}\beta), \tag{2.6}$$

where

$$\mathbf{D}_u = \begin{pmatrix} \mathbf{X}_1^T & \dfrac{U_1 - u}{h}\mathbf{X}_1^T \\ \vdots & \vdots \\ \mathbf{X}_n^T & \dfrac{U_n - u}{h}\mathbf{X}_n^T \end{pmatrix}, \qquad \mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T = \begin{pmatrix} Z_{11} & \cdots & Z_{1p} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \cdots & Z_{np} \end{pmatrix},$$
$$Y = (Y_1, \ldots, Y_n)^T, \qquad \mathbf{W}_u = \mathrm{diag}(K_h(U_1 - u), \ldots, K_h(U_n - u)).$$

Note that $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ thereby typically depend on β. If β is given, we can estimate $\alpha(u)$ by

$$\hat{\alpha}(u, \beta) = (I_q, \mathbf{0}_q)(\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u)^{-1}\mathbf{D}_u^T\mathbf{W}_u(Y - \mathbf{Z}\beta), \tag{2.7}$$

where I_q denotes a q-dimensional identity matrix and $\mathbf{0}_q$ is the $q \times q$ matrix with all the entries being zero. Substituting the estimate $\hat{\alpha}(u, \beta)$ into (2.3), we then have the following approximate residuals:

$$\begin{aligned} \hat{\varepsilon}_i(\beta) &= Y_i - \mathbf{Z}_i^T\beta - \mathbf{X}_i^T\hat{\alpha}(U_i, \beta) \\ &= Y_i - \mathbf{Z}_i^T\beta - \sum_{k=1}^n S_{ik}(Y_k - \mathbf{Z}_k^T\beta) \\ &= Y_i - \hat{Y}_i - \beta^T(\mathbf{Z}_i - \hat{\mathbf{Z}}_i), \quad i = 1, \ldots, n, \end{aligned} \tag{2.8}$$

where $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T = \mathbf{S}Y$, $\hat{\mathbf{Z}} = (\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_n)^T = \mathbf{SZ}$, S_{ik} is the (i, k)th element of the smoothing matrix $\mathbf{S}$, which depends only on the observations $\{(U_i, \mathbf{X}_i),\ i = 1, \ldots, n\}$, with

$$\mathbf{S} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{0}^T)(\mathbf{D}_{u_1}^T\mathbf{W}_{u_1}\mathbf{D}_{u_1})^{-1}\mathbf{D}_{u_1}^T\mathbf{W}_{u_1} \\ \vdots \\ (\mathbf{X}_n^T\mathbf{0}^T)(\mathbf{D}_{u_n}^T\mathbf{W}_{u_n}\mathbf{D}_{u_n})^{-1}\mathbf{D}_{u_n}^T\mathbf{W}_{u_n} \end{pmatrix}.$$

From (2.7) and (2.8), we can estimate $g(\varepsilon_i(\beta))$ and $g'(\varepsilon_i(\beta))$ respectively by

$$\hat{g}(\hat{\varepsilon}_i(\beta)) = \frac{1}{n}\sum_{j=1}^n L_h(\hat{\varepsilon}_i(\beta) - \hat{\varepsilon}_j(\beta)), \tag{2.9}$$

$$\hat{g}'(\hat{\varepsilon}_i(\beta)) = \frac{1}{n}\sum_{j=1}^n L_h'(\hat{\varepsilon}_i(\beta) - \hat{\varepsilon}_j(\beta)), \tag{2.10}$$

where $L_h(\cdot) = h^{-1}L(\cdot/h)$ is a kernel function and h is a bandwidth that depends on the sample size n. We plug the estimators $\hat{\alpha}(u, \beta), \hat{g}(\hat{\varepsilon}_i(\beta))$ and $\hat{g}'(\hat{\varepsilon}_i(\beta))$ into $\eta_i(\beta)$ in (2.2), respectively. A plug-in estimating smoothed score function can be defined as follows

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{\eta}_i(\beta) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))}{\hat{g}(\hat{\varepsilon}_i(\beta))}\mathbf{Z}_i. \tag{2.11}$$

Although each component of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{\eta}_i(\beta)$ is asymptotically zero when β is the true parameter vector, such replacement slows the convergence rate and results in non-negligible bias and enhanced variance because the convergence rates of the plug-in estimators $\hat{\alpha}(u, \beta), \hat{g}(\hat{\varepsilon}_i(\beta))$ and $\hat{g}'(\hat{\varepsilon}_i(\beta))$ are slower than $n^{-1/2}$ when an optimal bandwidth is adopted (see [4]). Thus, we cannot draw inferences for β directly from (2.11). Furthermore, such replacement ignores the functional dependence of $\hat{\alpha}(U_i, \beta)$ in (2.7) on β implicitly. From (2.8), note that $\hat{\varepsilon}_i(\beta)$ contains the estimated coefficient function $\hat{\alpha}(U_i, \beta)$, which is dependent on the unknown parameter vector β. To make inferences for β, we define the profile-type smoothing score function (PSSF) as follows:

$$\begin{aligned}\mathrm{PSSF}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\eta}_i(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log\hat{g}(\hat{\varepsilon}_i(\beta))\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\beta}\hat{g}(\hat{\varepsilon}_i(\beta))}{\hat{g}(\hat{\varepsilon}_i(\beta))} = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))}{\hat{g}(\hat{\varepsilon}_i(\beta))}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i),\end{aligned} \tag{2.12}$$

where $\hat{\mathbf{Z}}_i = \sum_{k=1}^{n} S_{ik}\mathbf{Z}_k$. Note that the second component in (2.12) vanishes in a solely parametric framework or if the estimate $\hat{\alpha}(\cdot)$ does not depend on β. In general, the profile-type score function in semiparametric models has two components. Wang et al. [19] propose a similar procedure for a marginal generalized semiparametric partially linear model with longitudinal/clustered data. This modification turns out to be the key to constructing the confidence regions for β. As shown in Section 4 and later illustrated in the simulation, the asymptotic property of PSSF(β) is insensitive to the choice of bandwidth. For example, the optimal bandwidth h of order $O(n^{-1/5})$ can be used.

3. Theoretical results

Throughout the paper, let $\Gamma(u) = E(\mathbf{X}\mathbf{X}^T|U = u)$, $\Phi(u) = E(\mathbf{X}\mathbf{Z}^T|U = u)$ and $\mu(u) = \Phi^T(u)\Gamma^{-1}(u)$. In this section, we establish the asymptotic result for PSSF(β) defined in (2.12). We need the following conditions to derive the main results.

(C1) The random variable U has a compact support Ω. The density function $f_U(u)$ of U has a continuous second derivative and is uniformly bounded away from zero.
(C2) The density function $g(\cdot)$ of ε is bounded away from zero on $\mathcal{T}$ and satisfies the Lipschitz condition of order 1 on $\mathcal{T}$, where $\mathcal{T}$ is a bounded support set of $\mathbb{R}$.
(C3) The $q \times q$ matrix $\Gamma(u)$ is non-singular for each $u \in \Omega$. $\Gamma(u)$, $\Gamma(u)^{-1}$ and $\Phi(u)$ are all Lipschitz continuous.
(C4) $\{\alpha_i(\cdot),\ i = 1, \dots, q\}$ have continuous second derivatives in $u \in \Omega$.
(C5) The kernel functions $K(\cdot)$ and $L(\cdot)$ are the bounded symmetric density functions with bounded support.
(C6) The bandwidth h satisfies that $nh^8 \to 0$ and $nh^3/(\log n)^3 \to \infty$.
(C7) There is an $s > 2$ such that $E\|\mathbf{Z}\|^{2s} < \infty$, $E\|\mathbf{X}\|^{2s} < \infty$ and $E\|\mu(U)\mathbf{X}\|^{2s} \le \infty$ for some $\delta < 2 - s^{-1}$ such that $n^{2\delta-1}h \to \infty$.
(C8) $V(\beta)$ and $\Sigma = E\{\varepsilon(\mathbf{Z} - \Phi^T(U)\Gamma^{-1}(U)\mathbf{X})\}^{\otimes 2}$ are the positive definite matrices, where $A^{\otimes 2} = AA^T$.

Note that the above conditions are assumed to hold uniformly in $u \in \Omega$. Conditions (C1)–(C5) are also found in the study of Fan and Huang [5]. These conditions are actually quite mild and can be easily satisfied. Condition (C6) gives a range of bandwidths from $O(n^{-1/3}\log n)$ to $O(n^{-1/8})$ that includes the order of optimal bandwidth. Condition (C8) ensures that there exists an asymptotic variance for the estimator of β.

For the sake of convenience, let $\hat{V}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i(\beta)\hat{\eta}_i^T(\beta)$, where

$$\hat{\eta}_i(\beta) = -\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))}{\hat{g}(\hat{\varepsilon}_i(\beta))}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)$$

is defined in (2.12). We first give the following Proposition 1 which is a crucial theoretical result of this article and the key to investigation of the asymptotic properties of PSSF(β).

Proposition 1. *Assume that conditions* (C1)–(C8) *hold. If β is the true value of the parameter vector, we have*

$$\mathrm{PSSF}(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(\beta) + o_P(1), \tag{3.1}$$

and

$$\hat{V}(\beta) = V(\beta) + o_P(1), \tag{3.2}$$

where $\xi_i(\beta) = -\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}(\mathbf{Z}_i - \Phi^T(U_i)\Gamma^{-1}(U_i)\mathbf{X}_i)$ *and* $V(\beta) = \mathrm{Cov}(\xi(\beta))$ *is a positive matrix.*

Theorem 1. *Assume that conditions* (C1)–(C8) *hold. If* β *is the true value of the parameter vector, we have*

$$\mathrm{PSSF}(\beta) \xrightarrow{d} N(0, V(\beta)), \tag{3.3}$$

where "$\xrightarrow{d}$" *denotes the convergence in distribution and* $V(\beta)$ *is defined in Proposition 1.*

To construct the confidence regions of β, we first define the following asymptotic profile-type statistic:

$$\hat{\mathcal{L}}(\beta) = \mathrm{PSSF}^T(\beta)\hat{V}^{-1}(\hat{\beta})\mathrm{PSSF}(\beta), \tag{3.4}$$

where $\hat{V}(\hat{\beta})$ can be obtained by using the plug-in method, then

$$\hat{V}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{g}'(\hat{\varepsilon}_i(\hat{\beta}))}{\hat{g}(\hat{\varepsilon}_i(\hat{\beta}))}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)\left(\frac{\hat{g}'(\hat{\varepsilon}_i(\hat{\beta}))}{\hat{g}(\hat{\varepsilon}_i(\hat{\beta}))}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)\right)^T \tag{3.5}$$

with $\hat{\varepsilon}_i(\hat{\beta}) = Y_i - \mathbf{Z}_i^T\hat{\beta} - \mathbf{X}_i^T\hat{\alpha}(U_i, \hat{\beta})$. From (2.1) and (2.8), $\hat{\beta}$ is the profile least-squares estimator of β defined by

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mathbf{X}_i^T\hat{\alpha}(U_i, \beta) - \mathbf{Z}_i^T\beta\}^2\right) \\ &= \left\{\sum_{i=1}^{n}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)^T\right\}^{-1}\sum_{i=1}^{n}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)(Y_i - \hat{Y}_i).\end{aligned} \tag{3.6}$$

Theorem 2. *Assume that conditions* (C1)–(C8) *hold. If* β *is the true value of the parameter vector, we have*

$$\hat{\mathcal{L}}(\beta) \xrightarrow{d} \chi_p^2 \quad as \quad n \to \infty, \tag{3.7}$$

where χ_p^2 *denotes the chi-square distribution with p degrees of freedom.*

As a conclusion of Theorem 2, $\hat{\mathcal{L}}(\beta)$ can be used to construct a confidence region for β. More precisely, for any $0 < \alpha < 1$, let c_α be such that $P(\chi_p^2 > c_\alpha) \le 1 - \alpha$. Then

$$I_\alpha(\beta) = \left\{\beta \in \mathbb{R}^p : \hat{\mathcal{L}}(\beta) \le c_\alpha\right\}$$

constitutes a confidence region for β with asymptotic coverage $1 - \alpha$.

4. Numerical studies

Throughout this section, we use the quartic kernel $K(u) = \frac{15}{16}(1 - u^2)_+^2$, and use a "leave-one-sample-out" method to select the bandwidth h. This method has been widely applied in practice (see, for example, [2,5,23]). We define the cross-validation score for h as $\mathrm{CV}(h) = n^{-1}\sum_{i=1}^{n}\{Y_i - \mathbf{X}_i^T\hat{\alpha}_{h,-i}(U_i) - \mathbf{Z}_i^T\hat{\beta}_{-i}\}^2$, where $\hat{\beta}_{-i}$ is the profile least-squares estimator defined by (3.6) and is computed from data with measurements of the ith observation deleted and $\hat{\alpha}_{h,-i}(\cdot)$ is the estimator defined in (2.7) with β replaced by $\hat{\beta}_{-i}$. Then cross-validation smoothing parameter h_{cv} is then the minimizer of $\mathrm{CV}(h)$. That is, $h_{cv} = \mathrm{argmin}_h\mathrm{CV}(h)$. Next, we use the same quartic kernel $K(u)$ to estimate the density function $g(\varepsilon(\beta))$. Theoretically, in the argument in (2.9), the bandwidth for the density estimator $\hat{g}(\hat{\varepsilon}_i(\beta))$ can be taken as the same as the bandwidth for the local polynomial smoother $\hat{\alpha}(\cdot)$ and this does not impact our theoretical result. We therefore use the same bandwidth h to estimate the density function $g(\varepsilon(\beta))$.

4.1. Simulation study

In this subsection, we present the results of Monte Carlo simulations to illustrate the finite sample performance of the proposed method. Our simulated data were generated from the following model:

$$Y_i = \mathbf{X}_i^T\alpha(U_i) + \mathbf{Z}_i^T\beta + \varepsilon_i. \tag{4.1}$$

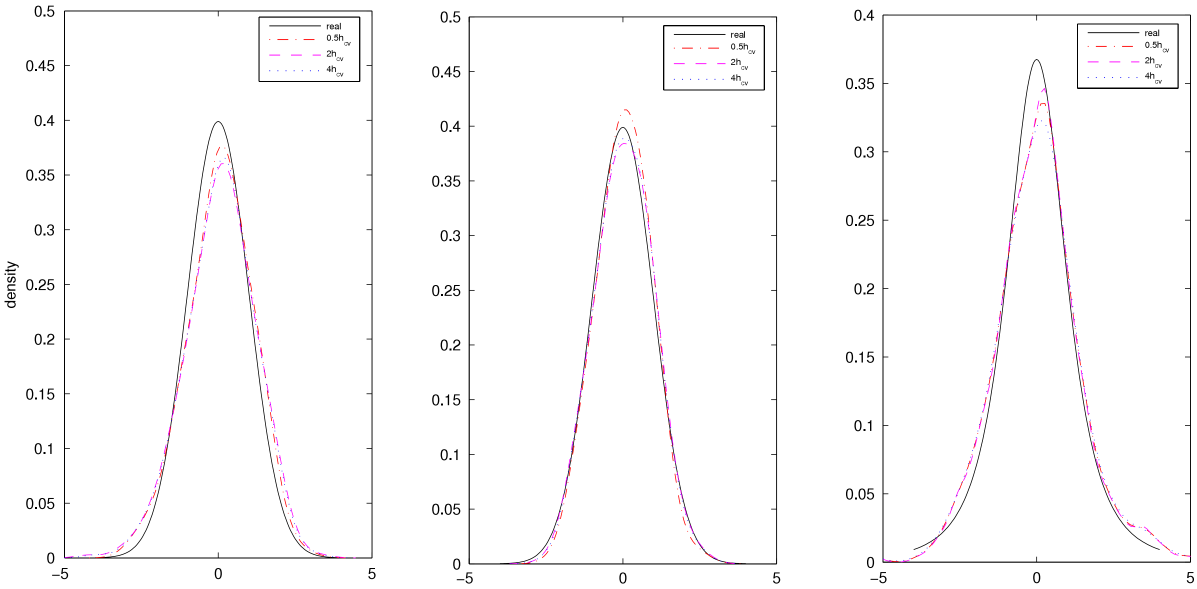

Fig. 1. The plot of the true density and the estimated densities for three bandwidths. The left panel shows a standard normal distribution with $h_{cv} = 0.3319$. The middle panel shows a mixture normal distribution with $h_{cv} = 0.3621$. The right panel shows a standard t-distribution with three degrees of freedom and $h_{cv} = 0.4022$.

In our simulation study, the covariate U_i is uniformly distributed on [0, 1], the nonparametric component $\alpha(u) = (\alpha_1(u), \alpha_2(u))^T$ with $q = 2$ in which X_{i1} and X_{i2} are independent and normal with mean zero and variance 0.8. Because we take two-dimensional parametric components $\beta = [1.5, -2.5]^T$, the covariable $\mathbf{Z}_i$ is a two-dimensional normal random vector with mean zero and a covariance matrix (σ_{ij}) with $\sigma_{ij} = 0.5^{|i-j|}$. Furthermore, the varying coefficient functions are given as

$$\alpha_1(u) = \exp(-u^2) + \sin(2\pi u) \quad \text{and} \quad \alpha_2(u) = 2u(1-u) - \cos(\pi u).$$

In the simulations, we draw 1000 random samples of sizes $n = 200$, 400 and 600 from the above model, respectively.

The first aim of these simulations is to study the performance of the proposed method for three model noises. The three model noises ε are generated from the following three different distributions, respectively: the standard normal distribution, the mixture normal distribution $2/3 * N(0, 1/2) + 1/3 * N(0, 2)$, and the standard t-distribution with three degrees of freedom. To demonstrate the insensitive bandwidth for the density estimator $\hat{g}(\hat{\varepsilon}_i(\beta))$, we take the smoothing parameter at three values $h = 0.5 * h_{cv}$, $h = 2 * h_{cv}$ and $h = 4 * h_{cv}$ under the three different distributions. The plots of the true density function and the estimated density function with the sample size $n = 400$ at the three different levels of bandwidth are presented in Fig. 1. The optimal bandwidths are about $h_{cv} = 0.3319$, 0.3621 and 0.4022 for the above three model noises, respectively. Fig. 1 shows that the proposed method is insensitive to the choice of bandwidth.

The second aim of this simulation study is to construct the confidence regions of parameters β of interest. We consider two approaches for comparison: the profile-type smoothed score function (PSSF) approach and normal approximation by the profile least-squares estimator (PLS). Similar to the results of Fan and Huang [5], the profile least-squares estimator defined by (3.6) can be proved to be asymptotically normal as follows:

$$\sqrt{n}B\Sigma^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, 1),$$

where $B = E(\mathbf{Z}_1\mathbf{Z}_1^T) - E[\Phi^T(U_1)\Gamma^{-1}(U_1)\Phi(U_1)]$ and Σ is defined in condition (C6). To construct confidence regions, we also need to estimate the asymptotic variance of the form:

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)^T, \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)(\mathbf{Z}_i - \hat{\mathbf{Z}}_i)^T,$$

where $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i^T\alpha(U_i, \hat{\beta}) - \mathbf{Z}_i^T\hat{\beta}$. The confidence regions and their coverage probabilities with the nominal level $1 - \alpha = 0.95$, are computed. The comparison is made through the average coverage probabilities and the average area of the confidence regions for three sample sizes $n = 200$, 400 and 600. Because the two methods are similar in terms of coverage accuracy, we present only the confidence region plots for the sample size $n = 400$ here.

Table 1 reports the simulation results for the average coverage probabilities and Fig. 2 reports the confidence regions. From Table 1 and Fig. 2, we can see that the average coverage probabilities and the sizes of the confidence regions estimated by each method are about the same when the model noise ε is generated from the standard normal distribution. This shows

Table 1
Coverage probabilities of (β_1, β_2) with 95% nominal level.

n	PSSF			PLS		
	$N(0, 1)$	N_{mix}[a]	$t(3)$	$N(0, 1)$	N_{mix}[a]	$t(3)$
200	0.9340	0.9370	0.9360	0.9330	0.9310	0.9320
400	0.9420	0.9450	0.9450	0.9430	0.9420	0.9390
600	0.9480	0.9490	0.9480	0.9470	0.9460	0.9460

[a] N_{mix} denotes the mixture normal distribution.

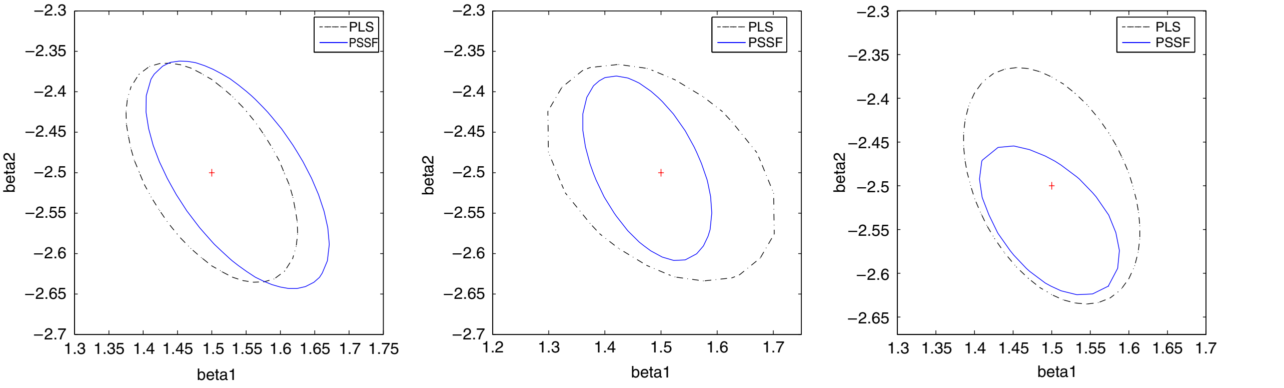

Fig. 2. The confidence regions for three model noises. The left panel shows a standard normal distribution. The middle panel shows a mixture normal distribution. The right panel shows a standard t-distribution with three degrees of freedom.

Table 2
Coverage probabilities for β_1 and β_2 with 95% nominal level.

β	n	PSSF			PLS		
		$N(0, 1)$	N_{mix}^{a}	$t(3)$	$N(0, 1)$	N_{mix}^{a}	$t(3)$
β_1	200	0.9280	0.9320	0.9290	0.9290	0.9310	0.9270
	400	0.9390	0.9430	0.9410	0.9380	0.9420	0.9380
	600	0.9470	0.9490	0.9470	0.9470	0.9480	0.9460
β_2	200	0.9320	0.9350	0.9310	0.9290	0.9330	0.9300
	400	0.9410	0.9450	0.9420	0.9390	0.9440	0.9380
	600	0.9470	0.9490	0.9460	0.9450	0.9470	0.9450

[a] N_{mix} denotes the mixture normal distribution.

Table 3
Average lengths of the confidence intervals for β_1 and β_2 with 95% nominal level.

β	n	PSSF			PLS		
		$N(0, 1)$	N_{mix}^{a}	$t(3)$	$N(0, 1)$	N_{mix}^{a}	$t(3)$
β_1	200	0.1261	0.1183	0.1231	0.1255	0.1202	0.1247
	400	0.1108	0.0994	0.1102	0.1129	0.1026	0.1128
	600	0.0955	0.0857	0.0947	0.0984	0.0901	0.0979
β_2	200	0.1259	0.1136	0.1261	0.1332	0.1311	0.1348
	400	0.1098	0.1025	0.1107	0.1164	0.1103	0.1162
	600	0.0946	0.0853	0.0945	0.1009	0.0908	0.1021

[a] N_{mix} denotes the mixture normal distribution.

that the PSSF method achieves slightly higher coverage levels than the PLS does, while the PSSF-based confidence regions are smaller than those obtained by the PLS when the model noises are the mixture normal distribution and the standard t-distribution with three degrees of freedom.

The proposed method can also be used to construct the confidence interval for one component of the regression parameter. For example, we can construct the confidence interval of β_1 while we replace β_2 by its consistent estimator (such as PLS estimator). The average coverage probabilities and average lengths of the confidence intervals for β_1 and β_2 are given in the following Tables 2 and 3, respectively. From Tables 2 and 3, we further see that the PSSF method performs better than the PLS method.

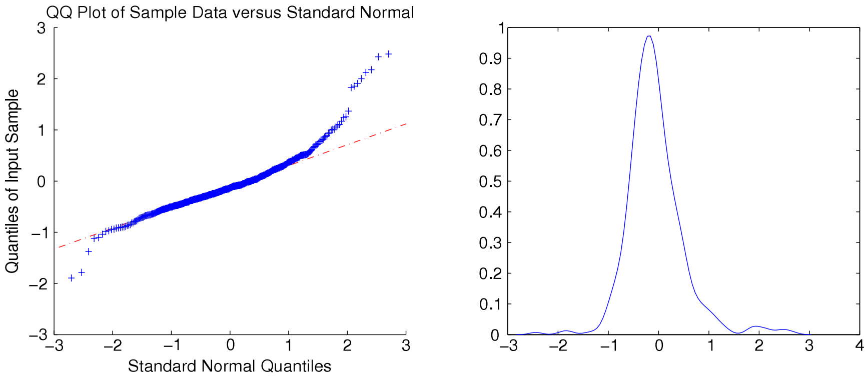

Fig. 3. Q–Q plot and density function curve of the residual $\hat{\varepsilon}$.

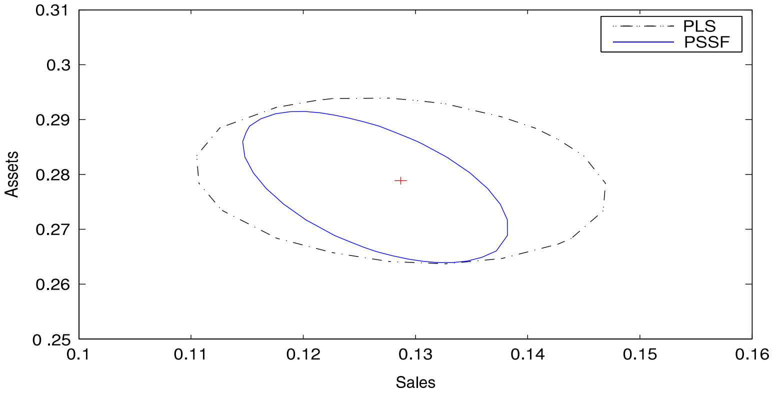

Fig. 4. 95% confidence regions for the coefficients (β_1, β_2) of the sales and assets variables.

4.2. A real data example

We now illustrate the proposed method through its application to CEO data and compare it with the profile least-squares (PLS) method. The CEO dataset was collected from Forbes' 1999 list of Corporate America's Most Powerful People. The CEO sample contains 447 observations and 7-non-constant independent variables: salary (1999 salary + bonuses), totcomp (1999 CEO total compensation), tenure (number of years as CEO, tenure equals 0 if less than 6 months), age (age of CEO), sales (total 1998 sales revenue of firm), profits (1998 profits of firm), assets (total assets of firm in 1998). For simplicity of notation, the covariates of age, tenure, sales, assets, totcomp, and profits are denoted respectively by $U, X_2, Z_1, \ldots, Z_4$. The dataset is standardized to zero mean and unit variance to ensure the comparability of the different variables. The following model is then considered

$$Y = \alpha_1(U) + \alpha_2(U)X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \varepsilon, \tag{4.2}$$

where response variable Y denotes the CEO's salary. The quartic kernel is employed to estimate the coefficient functions and the density function of ε, and the cross-validation (CV) method is used to select the bandwidth $h_{cv} = 3.7593$. We obtain the profile least-squares estimators (0.1292, 0.2801, 0.2553 and 0.0961) of $(\beta_1, \ldots, \beta_4)$. The Q–Q plot and the density function curve (Fig. 3) of the residual $\hat{\varepsilon}$ show that the model error does not have a normal distribution and the distribution has very heavy tail.

To compare the PSSF method with the PLS method, we obtain the 95% confidence regions for the coefficients (β_1, β_2) of the sales and assets variables that are shown in Fig. 4. We also obtain similar results for other coefficients but we omit them from this paper. From Fig. 4, we find that the PSSF-based confidence region is smaller than that based on the PLS.

Acknowledgments

Gaorong Li's research was supported by Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR20110822) and Training Programme Foundation for the Beijing Municipal Excellent Talents (2010D005015000002). Sanying Feng's research was supported by the Natural

Science Research Project of Henan (2008B110009). Heng Peng's research was supported by CERG grants from the Hong Kong Research Grants Council (HKBU 201707, HKBU 201809, and HKBU 201610), FRG grants from Hong Kong Baptist University (FRG/08-09/II-33), and a grant from National Natural Science Foundation of China (NNSF 10871054). The authors thank the editor, the associate editor, and the referees for their constructive comments that improved the article.

Appendix. Proofs of the main results

For the sake of convenience, let $c\,(0 < c < \infty)$ denote a constant that does not depend on n but takes a different value at each appearance. The following notations will be used in the proof of the lemmas and theorems. Let $\mu_i = \int u^i K(u)\mathrm{d}u$, $\nu_i = \int u^i K^2(u)\mathrm{d}u$, and $i = 1, 2, 3, 4$.

Lemma 1. *Suppose that conditions* (C1)–(C6) *hold. If* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$, *then letting* $c_n = \left\{\frac{\log n}{nh}\right\}^{1/2} + h^2$ *and* $d_n = \left(\frac{\log n}{nh}\right)^{1/2}$,

$$\sup_{u\in\Omega} \frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^l X_{ij}\varepsilon_i = O_P(d_n),$$

$$\sup_{u\in\Omega} \left|\frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^l X_{ij_1}X_{ij_2} - f_U(u)\mu_l\Gamma_{j_1j_2}(u)\right| = O_P(c_n),$$

$$\sup_{u\in\Omega} \left|\frac{1}{n}\sum_{i=1}^n K_h(U_i - u)\left(\frac{U_i - u}{h}\right)^l X_{ij}Z_{ik} - f_U(u)\Phi_{jk}(u)\right| = O_P(c_n),$$

where $j_1, j_2,\ j = 1, \ldots, q,\ k = 1, \ldots, p,\ l = 0, 1, 2, 4,\ \Gamma_{j_1j_2}(u)$ *is the* (j_1, j_2)*th element of* $\Gamma(u)$ *and* $\Phi_{jk}(u)$ *is the* (j, k)*th element of* $\Phi(u)$.

Because the proof of Lemma 1 is similar to that of Lemma A.2 of [20], we omit the details here.

For any given parametric component β, the following lemma provides the consistency rate of the estimators of nonparametric functions. Let $\alpha_j(u)$ denote the jth component of $\alpha(u)$, $j = 1, \ldots, q$.

Lemma 2. *Under the conditions of Lemma 1, as* $n \to \infty$, *we have*

$$\|\hat{\alpha}(u, \beta) - \alpha(u)\| = O_P(c_n), \tag{A.1}$$

and

$$\max_{1\le j\le q}\sup_{u\in\Omega} |\hat{\alpha}_j(u, \beta) - \alpha_j(u)| = O_P(c_n) \tag{A.2}$$

holds uniformly in $u \in \Omega$, *the support of* U.

Proof. We first present the proof of Eq. (A.1). Let

$$S_{n,l} = \sum_{i=1}^n K_h(U_i - u)\mathbf{X}_i\mathbf{X}_i^T\left(\frac{U_i - u}{h}\right)^l, \quad l = 0, 1, 2.$$

Note that

$$\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u = \begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2}\end{pmatrix}.$$

Each element of the above matrix is in the form of a kernel regression. By Lemma 1 and some elementary calculations, we find that

$$S_{n,l} = nf_U(u)\mu_l\Gamma(u)(1 + O_P(c_n)) \tag{A.3}$$

holds uniformly in $u \in \Omega$. By (A.3), we also see that

$$\mathbf{D}_u^T\mathbf{W}_u\mathbf{D}_u = nf_U(u)\Gamma(u) \otimes \begin{pmatrix} 1 & 0 \\ 0 & \mu_2\end{pmatrix}\{1 + O_P(c_n)\} \tag{A.4}$$

holds uniformly in $u \in \Omega$. By (2.7) and (A.4), we have

$$\begin{aligned}\hat{\alpha}(u, \beta) &= [nf_U(u)\Gamma(u)]^{-1}\sum_{i=1}^n K_h(U_i - u)\mathbf{X}_i(Y_i - \mathbf{Z}_i^T\beta) + O_P(c_n)\\ &= [nf_U(u)\Gamma(u)]^{-1}\sum_{i=1}^n K_h(U_i - u)\mathbf{X}_i\left\{\mathbf{X}_i^T\alpha(U_i) + \varepsilon_i\right\} + O_P(c_n).\end{aligned} \tag{A.5}$$

Applying Lemma 1, similar to the calculation of (A.3), we can easily show that

$$\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\mathbf{X}_i\mathbf{X}_i^T\alpha(U_i)=f_U(u)\Gamma(u)\alpha(u)\{1+O_P(c_n)\} \tag{A.6}$$

and that

$$\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)\mathbf{X}_i\varepsilon_i=o_P(1) \tag{A.7}$$

holds uniformly in $u\in\Omega$. By (A.5)–(A.7), $\hat{\alpha}(u,\beta)=\alpha(u)+O_P(c_n)$ holds uniformly in $u\in\Omega$. This completes the proof of Eq. (A.1).

To prove (A.2), similar to Xia and Li [20], we further decompose $\hat{\alpha}_j(u,\beta),\ j=1,\dots,q$. Without loss of generality, we only consider $\hat{\alpha}_1(u,\beta)$. For convenience, let $K_{ih}(u)=K_h(U_i-u)$, $S_i=(X_{i2},\dots,X_{iq})$, $T_i=(X_{i1},\dots,X_{iq})$. Without confusion, we let $V_i=(S_i,(U_i-u)T_i)$ although it relates to u. Following Lemma 3 of [7], we have

$$\begin{aligned}\hat{\alpha}_1(u,\beta)=\alpha_1(u)\ +\ &\frac{\sum_{i=1}^{n}K_{ih}(u)(X_{i1}-J_nH_n^{-1}V_i^T)\mathbf{X}_i^T(\alpha(U_i)-\alpha(u)-\alpha'(u)(U_i-u))}{\sum_{i=1}^{n}K_{ih}(u)(X_{i1}-J_nH_n^{-1}V_i^T)^2}+\frac{\sum_{i=1}^{n}K_{ih}(u)(X_{i1}-J_nH_n^{-1}V_i^T)\varepsilon_i}{\sum_{i=1}^{n}K_{ih}(u)(X_{i1}-J_nH_n^{-1}V_i^T)^2}\\ =:\ &\alpha_1(u)+I_1+I_2,\end{aligned}$$

where

$$\begin{aligned}H_n\ &=\ \sum_{i=1}^{n}K_{ih}(u)V_i^TV_i=\begin{pmatrix}\sum_{i=1}^{n}K_{ih}(u)S_i^TS_i & h\sum_{i=1}^{n}K_{ih}(u)\left(\frac{U_i-u}{h}\right)S_i^TT_i\\ h\sum_{i=1}^{n}K_{ih}(u)\left(\frac{U_i-u}{h}\right)T_i^TS_i & h^2\sum_{i=1}^{n}K_{ih}(u)\left(\frac{U_i-u}{h}\right)^2T_i^TT_i\end{pmatrix}\\ &=:\ \begin{pmatrix}P_n & hR_n\\ hR_n^T & h^2Q_n\end{pmatrix},\\ J_n\ &=\ \sum_{i=1}^{n}K_{ih}(u)X_{i1}V_i=\left(\sum_{i=1}^{n}K_{ih}(u)X_{i1}S_i,\quad h\sum_{i=1}^{n}K_{ih}(u)\left(\frac{U_i-u}{h}\right)X_{i1}T_i\right)\\ &=:\ (A_n,\quad hB_n).\end{aligned}$$

Let $A(u)=(\Gamma_{12}(u),\Gamma_{13}(u),\dots,\Gamma_{1q}(u))$, $P(u)=(\Gamma_{ij}(u))_{i,\,j=2,\dots,q}$ and $Q(u)=(\Gamma_{ij}(u))_{i,j=1,\dots,q}$. By Lemma 1 and condition (C5), it is easy to show that

$$\begin{aligned}&\frac{1}{n}A_n=f_U(u)A(u)+O_P(c_n),\qquad \frac{1}{n}B_n=O_P(c_n)\mathbf{1}_q^T,\qquad \frac{1}{n}R_n=O_P(c_n)\mathbf{1}_{q-1}\mathbf{1}_q^T,\\ &\frac{1}{n}Q_n=f_U(u)\mu_2Q(u)+O_P(c_n),\qquad \frac{1}{n}P_n=f_U(u)P(u)+O_P(c_n).\end{aligned} \tag{A.8}$$

Here $\mathbf{1}_q$ is the $q\times 1$ vector with 1 as all the elements. It can be seen that P_n is a symmetric matrix and its inverse exists, then

$$H_n^{-1}=\begin{pmatrix}P_n^{-1}+h^2P_n^{-1}R_n\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1} & -hP_n^{-1}R_n\boldsymbol{\mathcal{K}}_n^{-1}\\ -h\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1} & \boldsymbol{\mathcal{K}}_n^{-1}\end{pmatrix},$$

where $\boldsymbol{\mathcal{K}}_n=h^2(Q_n-R_n^TP_n^{-1}R_n)$, and

$$\begin{aligned}&J_nH_n^{-1}=(A_nP_n^{-1}+h^2A_nP_n^{-1}R_n\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1}-h^2B_n\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1},-hA_nP_n^{-1}R_n\boldsymbol{\mathcal{K}}_n^{-1}+hB_n\boldsymbol{\mathcal{K}}_n^{-1}),\\ &J_nH_n^{-1}V_i^T=A_nP_n^{-1}S_i^T+h^2A_nP_n^{-1}R_n\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1}S_i^T-h^2B_n\boldsymbol{\mathcal{K}}_n^{-1}R_n^TP_n^{-1}S_i^T-h(U_i-u)(A_nP_n^{-1}R_n\mathcal{K}_n^{-1}T_i^T-hB_n\boldsymbol{\mathcal{K}}_n^{-1}T_i^T).\end{aligned}$$

From (A.8), we have

$$\begin{aligned}&J_nH_n^{-1}=(A(u)(P(u))^{-1}+O_P(c_n),\ O_P(c_n)\mathbf{1}_q^T),\\ &J_nH_n^{-1}J_n^T=nA(u)(P(u))^{-1}A^T(u)f_U(u)+O_P(c_n).\end{aligned}$$

To deal with I_i for $i = 1, 2$, we consider their denominator first. Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(X_{i1} - J_n H_n^{-1} V_i^T)^2 &= \frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)X_{i1}^2 - \frac{1}{n}J_n H_n^{-1} J_n^T \\ &= \Gamma_{11}(u)f_U(u) - A(u)(P(u))^{-1}A^T(u)f_U(u) + O_P(c_n) \\ &= f_U(u)\det(Q(u))/\det(P(u)) + O_P(c_n) \qquad (A.9)\end{aligned}$$

holds uniformly in $u \in \Omega$. Now we are in the position to handle I_1. Using the Taylor expansion, $\alpha_j(U_i) - \alpha_j(u) - \alpha_j'(u)(U_i - u) = \frac{1}{2}\alpha_j''(u^*)(U_i - u)^2$, $j = 1, \ldots, q$, where u^* is a point between U_i and u. By the Cauchy–Schwarz inequality, Lemma 1 and condition (C4), uniformly over $1 \le j \le q$, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(X_{i1} - J_n H_n^{-1} V_i^T)X_{ij}^T(\alpha_j(U_i) - \alpha_j(u))\right| \\ &\le \left\{\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(X_{i1} - J_n H_n^{-1} V_i^T)^2 \frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)X_{ij}^2(\alpha_j(U_i) - \alpha_j(u))^2\right\}^{1/2} \\ &= \left\{\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(X_{i1} - J_n H_n^{-1} V_i^T)^2 \frac{1}{4n}\sum_{i=1}^{n} K_{ih}(u)X_{ij}^2\alpha_j''^2(u^*)(U_i - u)^4\right\}^{1/2} \\ &= c\{[f_U(u)\det(Q(u))/\det(P(u)) + O_P(c_n)] \cdot h^4[\mu_4 f_U(u)\Gamma_{11}(u) + O_P(c_n)]\}^{1/2} \\ &= O_P(h^2). \qquad (A.10)\end{aligned}$$

From (A.9) and (A.10), we have $|I_1| = O_P(h^2)$. For I_2, we again apply Lemma 1 to obtain

$$\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)T_i^T\varepsilon_i = O_P(d_n), \qquad \frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)\left(\frac{U_i - u}{h}\right)T_i^T\varepsilon_i = O_P(d_n),$$

where d_n is defined in Lemma 1. Therefore, we can obtain that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(X_{i1} - J_n H_n^{-1} V_i^T)\varepsilon_i &= (1, -J_n H_n^{-1})\left(\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)T_i\varepsilon_i, \frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)(U_i - u)T_i\varepsilon_i\right)^T \\ &= (1, -A(u)(P(u))^{-1})\frac{1}{n}\sum_{i=1}^{n} K_{ih}(u)T_i^T\varepsilon_i + O_P(hc_n d_n). \qquad (A.11)\end{aligned}$$

Combining (A.9) with (A.11) and invoking Lemma 1 again, we have $|I_2| = O_P(d_n)$. Thus, this completes the proof of (A.2). □

Proof of Proposition 1. Assume that β is known. From (2.8) and (2.9), we have

$$\begin{aligned}&\hat{\varepsilon}_i(\beta) = Y_i - \mathbf{Z}_i^T\beta - \mathbf{X}_i^T\hat{\alpha}(U_i, \beta), \\ &\hat{g}(\hat{\varepsilon}_i(\beta)) = \frac{1}{n}\sum_{j=1}^{n} L_h(\hat{\varepsilon}_i(\beta) - \hat{\varepsilon}_j(\beta)).\end{aligned}$$

For simplicity, let $H_i = \mathbf{X}_i^T(\alpha(U_i) - \hat{\alpha}(U_i, \beta))$. By Lemma 1 and condition (C5), and using the Taylor expansion, we have

$$\begin{aligned}\hat{g}(\hat{\varepsilon}_i(\beta)) &= \frac{1}{n}\sum_{j=1}^{n} L_h(\hat{\varepsilon}_i(\beta) - \hat{\varepsilon}_j(\beta)) \\ &= \frac{1}{n}\sum_{j=1}^{n} L_h(\varepsilon_i(\beta) - \varepsilon_j(\beta) + H_i - H_j) \\ &= \frac{1}{n}\sum_{j=1}^{n} L_h(\varepsilon_i(\beta) - \varepsilon_j(\beta)) + \frac{1}{n}\sum_{j=1}^{n} L_h'(\varepsilon_i(\beta) - \varepsilon_j(\beta))(H_i - H_j) + O_P(c_n) \\ &=: g(\varepsilon_i(\beta)) + R_{1i} + O_P(c_n),\end{aligned}$$

where

$$R_{1i} = \hat{g}(\varepsilon_i(\beta)) - g(\varepsilon_i(\beta)) + \frac{1}{n}\sum_{j=1}^{n} L_h'(\varepsilon_i(\beta) - \varepsilon_j(\beta))(H_i - H_j).$$

Then

$$\begin{aligned}\frac{1}{\hat{g}(\hat{\varepsilon}_i(\beta))} &= \frac{1}{g(\varepsilon_i(\beta))}\frac{1}{1+\frac{1}{g(\varepsilon_i(\beta))}(R_{1i}+O_P(c_n))}\\ &= \frac{1}{g(\varepsilon_i(\beta))}-\frac{R_{1i}}{g^2(\varepsilon_i(\beta))}+O_P(c_n).\end{aligned}\tag{A.12}$$

From (A.12) and the definition of $\hat{\eta}_i(\beta)$ in (2.12), and through some elementary calculation, we have

$$\begin{aligned}\hat{\eta}_i(\beta) &= \xi_i(\beta)-\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}(\Phi^T(U_i)\Gamma^{-1}(U_i)\mathbf{X}_i-\hat{\mathbf{Z}}_i)-\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))-g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}(\mathbf{Z}_i-\hat{\mathbf{Z}}_i)\\ &\quad+\hat{g}'(\hat{\varepsilon}_i(\beta))(\mathbf{Z}_i-\hat{\mathbf{Z}}_i)\left[\frac{R_{1i}}{g^2(\varepsilon_i(\beta))}+O_P(c_n)\right]\\ &=: \xi_i(\beta)+M_{i,1}+M_{i,2}+M_{i,3}.\end{aligned}\tag{A.13}$$

From (2.12) and (A.13), we obtain

$$\mathrm{PSSF}(\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(\beta)+\sum_{k=1}^{3}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}M_{i,k}\right)=:\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(\beta)+\sum_{k=1}^{3}M_k.$$

Thus, to prove (3.1), we need only to show that M_1, M_2 and M_3 are of the order $o_P(1)$. We first deal with M_1, a p-dimensional column vector. Invoking Lemma 1 and the proof of Lemma 2, it is easy to show that

$$\begin{aligned}M_1 &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}(\Phi^T(U_i)\Gamma^{-1}(U_i)\mathbf{X}_i-\hat{\mathbf{Z}}_i)\\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\mathbf{1}_pO_P(c_n).\end{aligned}$$

Here $\mathbf{1}_p$ is the $p\times 1$ vector with 1 as all the elements. Given that $E\left(\frac{g'(\varepsilon_i(\beta))^2}{g(\varepsilon_i(\beta))^2}\,\middle|\,\mathbf{X}_i,U_i,\mathbf{Z}_i\right)\le\infty$, we have

$$\|M_1\|=\left(\frac{p}{n}\sum_{i=1}^{n}\left(\frac{g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\right)^2O_P(c_n^2)\right)^{1/2}=O_P(c_n),$$

which implies that $M_1=o_P(1)$. Using similar arguments for M_1, we obtain

$$\begin{aligned}M_2 &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))-g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}(\mathbf{Z}_i-\hat{\mathbf{Z}}_i)\\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))-g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}([\mathbf{Z}_i-\Phi^T(U_i)\Gamma^{-1}(U_i)\mathbf{X}_i]+\mathbf{1}_pO_P(c_n))\\ &=: M_{2,1}+M_{2,2}.\end{aligned}$$

Basic algebraic calculation makes it easy to check that $M_{2,1}$ has a slower rate of convergence than $M_{2,2}$. Therefore, we need only to control the rate of $M_{2,1}$. For simplicity, let $\mu(U)=\Phi^T(U)\Gamma^{-1}(U)$ be a $p\times q$ matrix and let $\mu_k(U)$ denote the kth row of $\mu(U)$, and $\widetilde{Z}_{ik}=Z_{ik}-\mu_k(U_i)\mathbf{X}_i$ be the kth component of $\mathbf{Z}_i-\mu(U_i)\mathbf{X}_i$. By conditions (C1)–(C3) and (C6)–(C7) and Theorem C of [18], and by invoking independence from the other sample, we have

$$\begin{aligned}\|M_{2,1}\| &= \left(\sum_{k=1}^{p}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{g}'(\hat{\varepsilon}_i(\beta))-g'(\varepsilon_i(\beta))}{g(\varepsilon_i(\beta))}\right)^2\widetilde{Z}_{ik}^2\right)^{1/2}\\ &= O_P(h^2+(1/nh^3)^{1/2})=o_P(1).\end{aligned}$$

By noting that $\|M_2\|\le\|M_{2,1}\|+\|M_{2,2}\|$ and that $M_{2,2}$ has a faster rate than $M_{2,1}$, we can derive that $M_2=o_P(1)$. In addition, by employing Lemmas 1 and 2, and invoking similar arguments for M_1 and M_2, we can see that $M_3=o_P(1)$. This completes the proof of (3.1).

We now prove (3.2). Let $M_i^*=M_{i,1}+M_{i,2}+M_{i,3}$. From (A.13) and $\hat{V}(\beta)=\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i(\beta)\hat{\eta}_i^T(\beta)$, we have

$$\begin{aligned}\hat{V}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)\xi_i^T(\beta)+\frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)M_i^{*T}+\frac{1}{n}\sum_{i=1}^{n}M_i^*\xi_i^T(\beta)+\frac{1}{n}\sum_{i=1}^{n}M_i^*M_i^{*T}\\ &=: J_1+J_2+J_3+J_4.\end{aligned}$$

By the law of large numbers, it is easy to show that $J_1 \xrightarrow{P} V(\beta)$. We now need to prove that $J_i = o_P(1), i = 2, 3, 4$. For J_2, note that

$$\begin{aligned} J_2 &= \frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)M_{i,1}^T + \frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)M_{i,2}^T + \frac{1}{n}\sum_{i=1}^{n}\xi_i(\beta)M_{i,3}^T \\ &=: J_{21} + J_{22} + J_{23}. \end{aligned}$$

Let $J_{21,rs}$ denote the (r, s) component of J_{21} and let $\xi_{i,r}(\beta)$ and $M_{i,1s}$ denote the rth and sth components of $\xi_i(\beta)$ and $M_{i,1}$, respectively. By the Cauchy–Schwarz inequality, we have

$$|J_{21,rs}| \le \left(\frac{1}{n}\sum_{i=1}^{n}\xi_{i,r}^2(\beta)\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}M_{i,1s}^2\right)^{1/2}. \tag{A.14}$$

By condition (C8) and $E(\xi_i(\beta)) = 0$ and using a similar proof to that of M_1, we find that $\frac{1}{n}\sum_{i=1}^{n}\xi_{i,r}^2(\beta) = O_P(1)$ and $\frac{1}{n}\sum_{i=1}^{n}M_{i,1s}^2 = o_P(1)$. Together with (A.14), these equations prove that $J_{21} = o_P(1)$. Similarly, we can show that $J_{22} = o_P(1)$ and $J_{23} = o_P(1)$. Thus, we have $J_2 = o_P(1)$. Looking at the structure of J_3 and J_4, we can see that they are very similar to that of J_2. Thus, using similar arguments to those employed for J_2, we find that $J_3 = o_P(1)$ and $J_4 = o_P(1)$. (3.2) then follows. □

Proof of Theorem 1. When β is the true value of the parameter vector, it is easy to see that $\xi(\beta) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(\beta)$ is a sum of independent and identically distributed random variables. Note that

$$\begin{aligned} E(\xi_1(\beta)) &= -E\left[\frac{g'(\varepsilon_1(\beta))}{g(\varepsilon_1(\beta))}(\mathbf{Z}_1 - \Phi^T(U_1)\Gamma^{-1}(U_1)\mathbf{X}_1)\right] \\ &= -E\left\{E\left[\frac{g'(\varepsilon_1(\beta))}{g(\varepsilon_1(\beta))}(\mathbf{Z}_1 - \Phi^T(U_1)\Gamma^{-1}(U_1)\mathbf{X}_1)|\mathbf{X}_1, \mathbf{Z}_1, U_1\right]\right\} \\ &= -E\left\{(\mathbf{Z}_1 - \Phi^T(U_1)\Gamma^{-1}(U_1)\mathbf{X}_1)E\left(\frac{g'(\varepsilon_1(\beta))}{g(\varepsilon_1(\beta))}\right)\right\}. \end{aligned}$$

Note that for any given U, $\Phi^T(U)\Gamma^{-1}(U)\mathbf{X}$ is the projection of $\mathbf{Z}$ onto the space spanned by $\mathbf{X}$. By this, it is easy to show that $E\{(\mathbf{Z}_1 - \Phi^T(U_1)\Gamma^{-1}(U_1)\mathbf{X}_1)\} = o(1)$. In addition, by condition (C2), we have

$$\begin{aligned} E\left(\frac{g'(\varepsilon_1(\beta))}{g(\varepsilon_1(\beta))}\right) &= \int_{\mathcal{T}}\frac{g'(\varepsilon_1(\beta))}{g(\varepsilon_1(\beta))}g(\varepsilon_1(\beta))\mathrm{d}\varepsilon_1(\beta) = \int_{\mathcal{T}}g'(\varepsilon_1(\beta))\mathrm{d}\varepsilon_1(\beta) \\ &= \frac{\partial}{\partial\varepsilon_1(\beta)}\int_{\mathcal{T}}g(\varepsilon_1(\beta))\mathrm{d}\varepsilon_1(\beta) = 0. \end{aligned}$$

Therefore, we obtain that $E(\xi(\beta)) = 0$. By the central limit theorem, we can show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i(\beta) \xrightarrow{d} N(0, V(\beta)), \tag{A.15}$$

where $V(\beta)$ is as defined in Proposition 1. By (3.1) and (A.15) and the Slutsky theorem, the result holds. □

Proof of Theorem 2. By Theorem 4.1 of [5], it is easy to see that the profile least-squares estimator $\hat{\beta}$ defined by (3.6) is a $\sqrt{n}$-consistent estimator of β. Together with (3.2), we then have $\hat{V}(\hat{\beta}) \xrightarrow{P} V(\beta)$. Therefore, Theorem 2 directly holds from Theorem 1 and Proposition 1. We omit the details from this paper. □

References

[1] I. Ahmad, S. Leelahanon, Q. Li, Efficient estimation of a semiparametric partially linear varying coefficient model, Ann. Statist. 33 (2005) 258–283.
[2] Z. Cai, J. Fan, R. Li, Efficient estimation and inference for varying-coefficient models, J. Amer. Statist. Assoc. 95 (2000) 888–902.
[3] Q. Chen, L. Lin, L.X. Zhu, Bias-corrected smoothed score function for single-index models, Metrika 71 (2010) 45–58.
[4] J. Fan, I. Gijbels, Local Polynomial Modeling and its Applications, Chapman and Hall, London, 1996.
[5] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[6] T.J. Hastie, R. Tibshirani, Varying-coefficient models, J. R. Stat. Soc. Ser. B 55 (1993) 757–796.
[7] T.L. Lai, H. Robbins, C.Z. Wei, Strong consistency of least squares estimates in multiple regression II, J. Multivariate Anal. 9 (1979) 343–361.
[8] C. Lam, J. Fan, Profile-kernel likelihood inference with diverging number of parameters, Ann. Statist. 36 (2008) 2232–2260.
[9] Q. Li, C.J. Huang, D. Li, T.-T. Fu, Semiparametric smooth coefficient models, J. Bus. Econom. Statist. 20 (2002) 412–422.
[10] R. Li, H. Liang, Variable selection in semiparametric regression modeling, Ann. Statist. 36 (2008) 261–286.
[11] G.R. Li, L.X. Zhu, L. Lin, Empirical likelihood for a varying coefficient partially linear model with diverging number of parameters, Tech. Report, in: IMS-China International Conference on Statistics and Probability, 2008.
[12] G.R. Li, L.X. Zhu, L.G. Xue, S.Y. Feng, Empirical likelihood inference in partially linear single-index models for longitudinal data, J. Multivariate Anal. 101 (2010) 718–732.
[13] C.F. Manski, Maximum score estimation of the stochastic utility model of choice, J. Econometrics 3 (1975) 205–228.

[14] C.F. Manski, Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator, J. Econometrics 27 (1985) 313–333.
[15] S.A. Murphy, A.W. van der Vaart, On profile likelihood, J. Amer. Statist. Assoc. 95 (2000) 449–465.
[16] T.A. Severini, J.G. Staniswalis, Quasi-likelihood estimation in semiparametric models, J. Amer. Statist. Assoc. 89 (1994) 501–511.
[17] T.A. Severini, W.H. Wong, Profile likelihood and conditionally parametric models, Ann. Statist. 20 (1992) 1768–1802.
[18] B.W. Silverman, Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, Ann. Statist. 6 (1978) 177–184.
[19] N. Wang, R.J. Carroll, X. Lin, Efficient semiparametric marginal estimation for longitudinal/clustered data, J. Amer. Statist. Assoc. 100 (2005) 147–157.
[20] Y. Xia, W.K. Li, On the estimation and testing of functional-coefficient linear models, Statist. Sinica 9 (1999) 737–757.
[21] Y. Xia, W. Zhang, H. Tong, Efficient estimation for semivarying-coefficient models, Biometrika 91 (2004) 661–681.
[22] J.H. You, G.M. Chen, Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model, J. Multivariate Anal. 97 (2006) 324–341.
[23] Y. Zhou, H. Liang, Statistical inference for semiparametric varying-coefficient partially linear models with generated regressors, Ann. Statist. 37 (2009) 427–458.
[24] L.X. Zhu, L. Lin, X. Cui, G.R. Li, Bias-corrected empirical likelihood in a multi-link semiparametric model, J. Multivariate Anal. 101 (2010) 850–868.
[25] L.X. Zhu, L.G. Xue, Empirical likelihood confidence regions in a partially linear single-index model, J. R. Stat. Soc. Ser. B 68 (2006) 549–570.