Bernoulli **18**(3), 2012, 836–856
DOI: 10.3150/11-BEJ365

Empirical likelihood for single-index varying-coefficient models

LIUGEN XUE[1] and QIHUA WANG[2,3]

[1]*College of Applied Sciences, Beijing University of Technology, Beijing 100124, China. E-mail: lgxue@bjut.edu.cn*
[2]*Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100080, China*
[3]*School of Mathematics and Statistics, Yunnan University, Kunming 650091, China. E-mail: qhwang@amss.ac.cn*

In this paper, we develop statistical inference techniques for the unknown coefficient functions and single-index parameters in single-index varying-coefficient models. We first estimate the nonparametric component via the local linear fitting, then construct an estimated empirical likelihood ratio function and hence obtain a maximum empirical likelihood estimator for the parametric component. Our estimator for parametric component is asymptotically efficient, and the estimator of nonparametric component has an optimal convergence rate. Our results provide ways to construct the confidence region for the involved unknown parameter. We also develop an adjusted empirical likelihood ratio for constructing the confidence regions of parameters of interest. A simulation study is conducted to evaluate the finite sample behaviors of the proposed methods.

Keywords: confidence region; empirical likelihood; nonparametric component; parametric component; single-index varying-coefficient model

1. Introduction

Consider a single-index varying-coefficient model of the form

$$Y = \mathrm{g}_0^T(\beta_0^T X)Z + \varepsilon, \tag{1.1}$$

where $(X, Z) \in R^p \times R^q$ is a vector of covariates, Y is the response variable, β_0 is an $p \times 1$ vector of unknown parameters, $\mathrm{g}_0(\cdot)$ is an $q \times 1$ vector of unknown functions and ε is a random error with mean 0 and finite variance σ^2. Assume that ε and (X, Z) are independent. For the sake of identifiability, it is often assumed that $\|\beta_0\| = 1$, and the first non-zero element is positive, where $\|\cdot\|$ denotes the Euclidean metric.

Model (1.1) includes a class of important statistical models. For example, if $q = 1$ and $Z = 1$, (1.1) reduces to the single-index model (see, e.g., Härdle, Hall and Ichimura [11], Weisberg and Welsh [24], Zhu and Fang [33], Chiou and Müller [6], Hristache, Juditsky and Spokoiny [13], Xue and Zhu [31]). If $p = 1$ and $\beta_0 = 1$, (1.1) is the varying-coefficient model (see, e.g., Chen and Tsay [5], Hastie and Tibshirani [12], Wu, Chiang and Hoover [25], Fan and Zhang [10], Cai, Fan and Li [2], Cai, Fan and Yao [3], Xue and Zhu [29]). If the last component of β_0 to be non-zero and $Z = (1, X^{*T})^T$ where X^* is the remaining vector of X with its pth component deleted, (1.1) becomes the adaptive varying-coefficient linear model (see, e.g., Fan, Yao and Cai [9], Lu, Tjøstheim and Yao [15]).

1350-7265

Model (1.1) is easily interpreted in real applications because it has the features of the single-index model and the varying-coefficient model. In addition, model (1.1) may include cross-product terms of some components of X and Z. Hence it has considerable flexibility to cater for complex multivariate nonlinear structure. Xia and Li [26] investigated a class of single-index coefficient regression models, which include model (1.1) as a special example. When it is used as a nonparametric time series model, Xia and Li [26] obtained the estimator of $g(\cdot)$ by kernel smoothing and then derived the estimator of β_0 by the least squares method and proved that the corresponding estimators are consistent and asymptotically normal.

In this paper, we develop statistical inference techniques of $g_0(\cdot)$ and β_0 with independent observations of (Y, X, Z). We can construct an empirical likelihood ratio function for β_0 by assuming $g_0(\cdot)$ and its derivative to be known functions. In practice, however, they are unknown, and hence the empirical likelihood ratio function cannot be used to make inference on β. This motivates us to estimate the unknown $g_0(\cdot)$ and $\dot{g}_0(\cdot)$ via the local linear smoother, and then obtain an estimated empirical likelihood ratio of β_0. The estimated empirical log-likelihood ratio is asymptotically distributed as a weighted sum of independent χ_1^2 variables with unknown weights. This result cannot be applied directly to construct confidence region for β_0. To solve this issue, two methods may be used (see Wang and Rao [22]). The first method is to estimate the unknown weights consistently so that the distribution of the estimated weighted sum of chi-squared variables can be estimated from the data. The second method is to adjust the estimated empirical log-likelihood ratio so that the resulting adjusted empirical log-likelihood ratio is asymptotically chi-squared. Also, we obtain a maximum empirical likelihood estimator of β_0, by maximizing the estimated empirical likelihood ratio function, and investigate its asymptotic property. In addition, we obtain the convergence rate of the estimator of σ^2 and define the consistent estimator of asymptotic variance; this allows us to construct a confidence region for β_0.

Comparing with the existing methods, our estimating method has the following advantage: The asymptotic variance of our estimator for β_0 is the same as those of Härdle *et al.* [11] and Xia and Li [26] when the model reduces to the single-index model; this shows that our estimator for β_0 is the same efficient as than those of Härdle *et al.* [11] and Xia and Li [26]. The difference between the proposed estimating approaches and the existing estimating approaches is that we use an empirical likelihood ratio to define the estimator of β_0 while the existing work uses the least squares techniques (see, e.g., Härdle *et al.* [11], Xia and Li [26]). Also, we develop an empirical likelihood inference for constructing a confidence region of β. The empirical likelihood method, introduced by Owen [17], has many advantages for constructing confidence regions or intervals. For example, it does not impose prior constraints on region shape, and it does not require the construction of a pivotal quantity. The empirical likelihood has been studied by many authors. The related works are Wang and Rao [22], Wang, Linton and Härdle [21], Xue and Zhu [29–31], Zhu and Xue [32], Qin and Zhang [18], Stute, Xue and Zhu [20], Xue [27,28], Wang and Xue [23], among others.

The rest of the paper is organized as follows. In Section 2, we define an estimated empirical likelihood ratio, and then obtain a maximum empirical likelihood estimator of β_0 by maximizing the empirical likelihood ratio function; the asymptotic properties of the proposed estimators are also investigated. In Section 3, we define an adjusted empirical log-likelihood and derive its asymptotic distribution. Section 4 reports a simulation study. Proofs of theorems are relegated to the Appendix. It should be pointed that some special techniques are used in the proofs.

2. Estimated empirical likelihood

2.1. Methodology

Suppose that $\{(Y_i, X_i, Z_i); 1 \le i \le n\}$ is an independent and identically distributed (i.i.d.) sample from (1.1), that is

$$Y_i = \mathrm{g}_0^T(\beta_0^T X_i)Z_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where ε_is are i.i.d. random errors with mean 0 and finite variance σ^2. Assume that $\{\varepsilon_i; 1 \le i \le n\}$ are independent of $\{(X_i, Z_i); 1 \le i \le n\}$.

To construct an empirical likelihood ratio function for β_0, we introduce an auxiliary random vector

$$\eta_i(\beta) = \{Y_i - \mathrm{g}_0^T(\beta^T X_i)Z_i\}\dot{\mathrm{g}}_0^T(\beta^T X_i)Z_i X_i w(\beta^T X_i), \tag{2.1}$$

where $\dot{\mathrm{g}}_0(\cdot)$ stands for the derivative of the function vector $\mathrm{g}_0(\cdot)$, and $w(\cdot)$ is a bounded weight function with a bounded support $\mathcal{U}_w$, which is introduced to control the boundary effect in the estimations of $\mathrm{g}_0(\cdot)$ and $\dot{\mathrm{g}}_0(\cdot)$. To convenience, we take that $w(\cdot)$ is the indicator function of the set $\mathcal{U}_w$. Note that $E\{\eta_i(\beta)\} = 0$ if $\beta = \beta_0$. Hence, the problem of testing whether β is the true parameter is equivalent to testing whether $E\{\eta_i(\beta)\} = 0$ for $i = 1, 2, \dots, n$. By Owen [17], this can be done by using the empirical likelihood. That is, we can define the profile empirical likelihood ratio function

$$L_n(\beta) = \max\left\{\prod_{i=1}^n (np_i) \,\middle|\, p_i \ge 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \eta_i(\beta) = 0\right\}.$$

It can be shown that $-2\log L_n(\beta_0)$ is asymptotically chi-squared with p degrees of freedom. However, $L_n(\beta)$ cannot be directly used to make statistical inference on β_0 because $L_n(\beta)$ contains the unknowns $\mathrm{g}_0(\cdot)$ and $\dot{\mathrm{g}}_0(\cdot)$. A natural way is to replace $\mathrm{g}_0(\cdot)$ and $\dot{\mathrm{g}}_0(\cdot)$ in $L_n(\beta)$ by their estimators and define an estimated empirical likelihood function. In this paper, we estimate the vector functions $\mathrm{g}_0(\cdot)$ and $\dot{\mathrm{g}}_0(\cdot)$ via the local linear regression technique (see, e.g., Fan and Gijbels [8]). The local linear estimators for $\mathrm{g}_0(u)$ and $\dot{\mathrm{g}}_0(u)$ are defined as $\hat{\mathrm{g}}(u; \beta_0) = \hat{\mathrm{a}}$ and $\hat{\dot{\mathrm{g}}}(u; \beta_0) = \hat{\mathrm{b}}$ at the fixed point β_0, where $\hat{\mathrm{a}}$ and $\hat{\mathrm{b}}$ minimize the sum of weighted squares

$$\sum_{i=1}^n [Y_i - \{\mathrm{a} + \mathrm{b}(\beta_0^T X_i - u)\}^T Z_i]^2 K_h(\beta_0^T X_i - u),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function, and $h = h_n$ is a bandwidth sequence that decreases to 0 as n increases to ∞. It follows from the least squares theory that

$$(\hat{\mathrm{g}}^T(u; \beta_0), h\hat{\dot{\mathrm{g}}}^T(u; \beta_0))^T = S_n^{-1}(u; \beta_0)\xi_n(u; \beta_0),$$

where

$$S_n(u; \beta_0) = \begin{pmatrix} S_{n,0}(u; \beta_0) & S_{n,1}(u; \beta_0) \\ S_{n,1}(u; \beta_0) & S_{n,2}(u; \beta_0) \end{pmatrix} \quad \text{and} \quad \xi_n(u; \beta_0) = \begin{pmatrix} \xi_{n,0}(u; \beta_0) \\ \xi_{n,1}(u; \beta_0) \end{pmatrix}$$

with

$$S_{n,j}(u;\beta_0) = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T \left(\frac{\beta_0^T X_i - u}{h}\right)^j K_h(\beta_0^T X_i - u)$$

and

$$\xi_{n,j}(u;\beta_0) = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i \left(\frac{\beta_0^T X_i - u}{h}\right)^j K_h(\beta_0^T X_i - u).$$

Since the convergence rate of the estimator of $\dot{g}_0'(u)$ is slower than that of the estimator of $g_0(u)$ if the same bandwidth is used, this leads to a slower convergence rate for the estimator $\hat{\beta}$ of β_0 than $\sqrt{n}$. To increase the convergence rate of the estimator of $\dot{g}_0'(u)$, we introduce the another bandwidth h_1 to replace h in $\hat{\dot{g}}(u;\beta)$, and define as $\hat{\dot{g}}_{h_1}(u;\beta)$.

Let $\hat{\eta}_i(\beta)$ be $\eta_i(\beta)$, with $g_0(\beta^T X_i)$ and $\dot{g}_0(\beta^T X_i)$ replaced by $\hat{g}(\beta^T X_i;\beta)$ and $\hat{\dot{g}}_{h_1}(\beta^T X_i;\beta)$, respectively, for $i = 1,\dots,n$. Then an estimated empirical likelihood ratio function is defined by

$$\hat{L}(\beta) = \max\left\{\prod_{i=1}^{n}(np_i)\,\middle|\, p_i \ge 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\hat{\eta}_i(\beta) = 0\right\}.$$

By the Lagrange multiplier method, $\log \hat{L}(\beta)$ can be represented as

$$\log \hat{L}(\beta) = -\sum_{i=1}^{n}\log\bigl(1 + \lambda^T\hat{\eta}_i(\beta)\bigr), \tag{2.2}$$

where λ is determined by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\eta}_i(\beta)}{1+\lambda^T\hat{\eta}_i(\beta)} = 0. \tag{2.3}$$

Let $\mathcal{B} = \{\beta \in R^p\colon \|\beta\| = 1$, and the first non-zero element is positive. Then β_0 is an inner point of the set $\mathcal{B}$. Therefore we need only search for β_0 over $\mathcal{B}$. A maximum empirical likelihood estimator for β_0 is given by

$$\hat{\beta} = \arg\sup_{\beta\in\mathcal{B}} \hat{L}(\beta). \tag{2.4}$$

With $\hat{\beta}$, we define the estimate of $g(u)$ by $\hat{g}(u) = \hat{g}(u,\hat{\beta})$, and the estimate of σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{g}^T(\hat{\beta}^T X_i;\hat{\beta})Z_i\}^2. \tag{2.5}$$

It is well known that if β is known, the optimal bandwidth h for $\hat{g}(u)$ is of order $O(n^{-1/5})$. However, if β is unknown, in order to ensure that the estimator $\hat{\beta}$ is root-n consistent, the bandwidth h should be smaller than $O(n^{-1/5})$, if we only assume $g(\cdot)$ are second-order differentiable (see Theorem 2 below). Note that once the estimator $\hat{\beta}$ is available, an optimal bandwidth of order $O(n^{-1/5})$ can be used in the final estimator for $g(\cdot)$.

2.2. Asymptotic properties

In order to obtain the asymptotic behaviors of our estimators, we first give the following conditions:

- (C1) The density function of $\beta^T X$, $f(u)$, is bounded away from zero for $u \in \mathcal{U}_w$ and β near β_0, and satisfies the Lipschitz condition of order 1 on $\mathcal{U}_w$, where $\mathcal{U}_w$ is the support of $w(u)$.
- (C2) The functions $g_j(u)$, $1 \le j \le q$, have continuous second derivatives on $\mathcal{U}_w$, where $g_j(u)$ are the jth components of $\mathrm{g}_0(u)$.
- (C3) $E(\|X\|^6) < \infty$, $E(\|Z\|^6) < \infty$ and $E(|\varepsilon|^6) < \infty$.
- (C4) $nh^2/\log^2 n \to \infty$, $nh^4 \log n \to 0$; $nhh_1^3/\log^2 n \to \infty$, $nh_1^5 = \mathrm{O}(1)$.
- (C5) The kernel $K(\cdot)$ is a symmetric probability density function with a bounded support and satisfies the Lipschitz condition of order 1 and $\int u^2 K(u)\,\mathrm{d}u \ne 0$.
- (C6) The matrix $D(u) = E(ZZ^T | \beta_0^T X = u)$ is positive definite, and each entry of $D(u)$ and $C(u) = E(VZ^T | \beta_0^T X = u)$ satisfies the Lipschitz condition of order 1 on $\mathcal{U}_w$, where $V = X\dot{\mathrm{g}}_0^T(\beta_0^T X) Z w(\beta_0^T X)$, and $\mathcal{U}_w$ is defined in (C1).
- (C7) The matrices $B(\beta_0) = E(VV^T)$ and $B_*(\beta_0) = B(\beta_0) - E\{C(\beta_0^T X)\dot{\mathrm{g}}_0(\beta_0^T X)E(X^T | \beta_0^T X)\}$ are positive definite, where V is defined in (C6).

Remark 1. Condition (C1) is used to bound the density function of $\beta^T X$ away from zero. This ensures that the denominators of $\hat{\mathrm{g}}(u;\beta)$ and $\hat{\dot{\mathrm{g}}}(u;\beta)$ are, in probability one, bounded away from 0 for $u \in \mathcal{U}_w$. The second derivatives in (C2) are standard smoothness conditions. (C3)–(C5) are necessary conditions for the asymptotic normality or the uniform consistency of the estimators. Conditions (C6) and (C7) ensure that the asymptotic variance for the estimator of β_0 exists.

Let $\mathcal{B}_n = \{\beta \in \mathcal{B}\colon \|\beta - \beta_0\| \le c_0 n^{-1/2}\}$ for some positive constant c_0. This is motivated by the fact that, since we anticipate that $\hat{\beta}$ is root-n consistent, we should look for a maximum of $\hat{L}(\beta)$ which involves β distant from β_0 by order $n^{-1/2}$. Similar restrictions were also made by Härdle, Hall and Ichimura [11], Xia and Li [26] and Wang and Xue [23].

The following theorem shows that $-2\log \hat{L}(\beta_0)$ is asymptotically distributed as a weighted sum of independent χ_1^2 variables.

Theorem 1. *Suppose that conditions* (C1)–(C7) *hold. Then*

$$-2\log \hat{L}(\beta_0) \xrightarrow{D} w_1\chi_{1,1}^2 + \cdots + w_p\chi_{1,p}^2,$$

where $\xrightarrow{D}$ represents convergence in distribution, $\chi_{1,1}^2, \ldots, \chi_{1,p}^2$ are independent χ_1^2 variables and the weights w_j, for $1 \le j \le p$, are the eigenvalues of $G(\beta_0) = B^{-1}(\beta_0)A(\beta_0)$. Here $B(\beta_0)$ is defined in condition (C7),

$$A(\beta_0) = B(\beta_0) - E\{C(\beta_0^T X)D^{-1}(\beta_0^T X)C^T(\beta_0^T X)\}, \tag{2.6}$$

and $C(u)$ and $D(u)$ are defined in condition (C6).

To apply Theorem 1 to construct a confidence region or interval for β_0, we need to consistently estimate the unknown weights w_j. By the "plug-in" method, $A(\beta_0)$ and $B(\beta_0)$ can be consistently estimated by

$$\hat{A}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{V}_i\hat{V}_i^T - \hat{C}(\hat{\beta}^T X_i)\hat{D}^{-1}(\hat{\beta}^T X_i)\hat{C}^T(\hat{\beta}^T X_i)\} \tag{2.7}$$

and

$$\hat{B}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{V}_i\hat{V}_i^T, \tag{2.8}$$

respectively, where $\hat{\beta}$ is the maximum empirical likelihood estimator of β_0 defined by (2.4), $\hat{V}_i = X_i\hat{\dot{g}}^T(\hat{\beta}^T X_i;\hat{\beta})Z_i w(\hat{\beta}^T X_i)$, $\hat{C}(\cdot) = \sum_{i=1}^n W_{ni}(\cdot)\hat{V}_i Z_i^T$ and $\hat{D}(\cdot) = \sum_{i=1}^n W_{ni}(\cdot)Z_i Z_i^T$ with

$$W_{ni}(\cdot) = K_1\left(\frac{\hat{\beta}^T X_i - \cdot}{b_n}\right)\Big/\sum_{k=1}^{n}K_1\left(\frac{\hat{\beta}^T X_k - \cdot}{b_n}\right),$$

where $K_1(\cdot)$ is a kernel function, and b_n is a bandwidth with $0 < b_n \to 0$.

This implies that the eigenvalues of $\hat{G}(\hat{\beta}) = \hat{B}^{-1}(\hat{\beta})\hat{A}(\hat{\beta})$, say $\hat{w}_j$, consistently estimate w_j for $j = 1, \dots, p$. Let $\hat{c}_{1-\alpha}$ be the $1-\alpha$ quantile of the conditional distribution of the weighted sum $\hat{s} = \hat{w}_1\chi^2_{1,1} + \cdots + \hat{w}_p\chi^2_{1,p}$ given the data. Then an approximate $1-\alpha$ confidence region for β_0 can be defined as

$$\mathcal{R}_{\mathrm{eel}}(\alpha) = \{\beta \in \mathcal{B}\colon -2\log\hat{L}(\beta) \le \hat{c}_{1-\alpha}\}.$$

In practice, the conditional distribution of the weighted sum $\hat{s}$, given the sample $\{(Y_i, X_i, Z_i), 1 \le i \le n\}$, can be calculated using Monte Carlo simulations by repeatedly generating independent samples $\chi^2_{1,1}, \dots, \chi^2_{1,p}$ from the χ^2_1 distribution.

The following theorem states an interesting result about $\hat{\beta}$. The asymptotic variance of $\hat{\beta}$ is smaller than that of Härdle *et al.* [11] when our model reduces to a single-index model.

Theorem 2. *Suppose that conditions* (C1)–(C7) *hold. Then*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \sigma^2 B_*^{-1}(\beta_0)A(\beta_0)B_*^{-1}(\beta_0)),$$

where $B_*(\beta_0)$ and $A(\beta_0)$ are defined in condition (C7) *and* (2.6), *respectively.*

In model (1.1), if $q = 1$ and $Z = 1$, then (1.1) reduces to the single-index model. By Theorem 2, we derive the following result.

Corollary 1. *Suppose that the conditions of Theorem 2 hold. If $q = 1$ and $Z = 1$ in model* (1.1), *then*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \sigma^2 A_1^{-}(\beta_0)),$$

where $A_1(\beta_0) = E[\{X - E(X|\beta_0^T X)\}\{X - E(X|\beta_0^T X)\}^T \dot{g}_0^2(\beta_0^T X) w(\beta_0^T X)]$ *and* A_1^- *represents a generalized inverse of the matrix* A_1^-.

Corollary 1 is the same as the results of Härdle et al. [11] and Xia and Li [26] for the single-index model.

For the estimator of the variance of error, $\hat{\sigma}^2$, we have the following result.

Theorem 3. *Suppose that conditions* (C1)–(C7) *hold. Then*,

$$\hat{\sigma}^2 - \sigma^2 = \mathrm{O}_P(n^{-1/2}).$$

To apply Theorem 2 to construction of the confidence region of β_0, we use the estimators $\hat{\sigma}^2$ and $\hat{A}(\hat{\beta})$ defined in (2.5) and (2.7), and define the estimator of $B_*(\beta_0)$ as follows

$$\hat{B}_*(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{V}_i\hat{V}_i^T - \hat{C}(\hat{\beta}^T X_i)\hat{g}(\hat{\beta}^T X_i; \hat{\beta})\hat{\mu}^T(\hat{\beta}^T X_i)\},$$

where $\hat{\mu}(\cdot) = \sum_{i=1}^{n} W_{ni}(\cdot)X_i$ is the estimator of $\mu(u) = E(X|\beta_0^T X = u)$. It can be shown that $\hat{A}(\hat{\beta}) \xrightarrow{P} A(\beta_0)$ and $\hat{B}_*(\hat{\beta}) \xrightarrow{P} B_*(\beta_0)$, where $\xrightarrow{P}$ denotes convergence in probability. By Theorems 3 and 4, we have

$$\{\hat{\sigma}^2\hat{B}_*^{-1}(\hat{\beta})\hat{A}(\hat{\beta})\hat{B}_*^{-1}(\hat{\beta})\}^{-1/2}\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, I_p).$$

Using Theorem 10.2d in Arnold [1], we obtain

$$(\hat{\beta} - \beta_0)^T\{n^{-1}\hat{\sigma}^2\hat{B}_*^{-1}(\hat{\beta})\hat{A}(\hat{\beta})\hat{B}_*^{-1}(\hat{\beta})\}^-(\hat{\beta} - \beta_0) \xrightarrow{D} \chi_p^2.$$

Let $\chi_p^2(1-\alpha)$ be the $1-\alpha$ quantile of χ_p^2 for $0 < \alpha < 1$. Then

$$\{\beta\colon\ (\hat{\beta} - \beta)^T(n^{-1}\hat{\sigma}^2\hat{B}_*^{-1}(\hat{\beta})\hat{A}(\hat{\beta})\hat{B}_*^{-1}(\hat{\beta}))^-(\hat{\beta} - \beta) \le \chi_p^2(1-\alpha)\}$$

gives an approximate $1-\alpha$ confidence region for β_0.

3. Adjusted empirical likelihood

In addition to the above, direct way of approximating the asymptotic distributions, we can also consider the following alternative. The alternative is motivated by the results of Rao and Scott [19]. By Rao and Scott [19] the distribution of $\rho(\beta_0)\sum_{i=1}^{p} w_i\chi_{1,i}^2$ can be approximated by χ_p^2, where $\rho(\beta_0) = p/\operatorname{tr}\{G(\beta_0)\}$. Let $\hat{\rho}(\hat{\beta}) = p/\operatorname{tr}\{\hat{G}(\hat{\beta})\}$ with $\hat{G}(\hat{\beta}) = \hat{A}^{1/2}(\hat{\beta})\hat{B}^{-1}(\hat{\beta})\hat{A}^{1/2}(\hat{\beta})$, where $\hat{A}(\hat{\beta})$ and $\hat{B}(\hat{\beta})$ are defined in (2.7) and (2.8). Invoking Theorem 1 and the consistency of $\hat{G}(\hat{\beta})$, the asymptotic distribution of $\hat{\rho}(\hat{\beta})\{-2\log\hat{L}(\beta)\}$ can be approximated by χ_p^2. Clearly, $\hat{\beta}$

in $\hat{\rho}(\cdot)$ can be replaced by β. Therefore, an improved Rao–Scott adjusted empirical log-likelihood can be defined as

$$\tilde{l}(\beta) = \hat{\rho}(\beta)\{-2\log \hat{L}(\beta)\}.$$

However, the accuracy of this approximation still depends on the values of the w_is. Now, we propose another adjusted empirical log-likelihood, whose asymptotic distribution is chi-squared with p degrees of freedom. The adjustment technique is developed by Wang and Rao [22] by using an approximate result in Rao and Scott [19]. Note that $\hat{\rho}(\beta)$ can be written as

$$\hat{\rho}(\beta) = \frac{\mathrm{tr}\{\hat{A}^{-}(\beta)\hat{A}(\beta)\}}{\mathrm{tr}\{\hat{B}^{-1}(\beta)\hat{A}(\beta)\}}.$$

By examining the asymptotic expansion of $-2\log \hat{L}(\beta)$, which is specified in the proof of Theorem 4 below, we define an adjustment factor

$$\hat{r}(\beta) = \frac{\mathrm{tr}\{\hat{A}^{-}(\beta)\hat{\Sigma}(\beta)\}}{\mathrm{tr}\{\hat{B}^{-1}(\beta)\hat{\Sigma}(\beta)\}},$$

by replacing $\hat{A}(\beta)$ in $\hat{\rho}(\beta)$ by $\hat{\Sigma}(\beta)$, where $\hat{\Sigma}(\beta) = \{\sum_{i=1}^{n}\hat{\eta}_i(\beta)\}\{\sum_{i=1}^{n}\hat{\eta}_i(\beta)\}^T$. The adjusted empirical log-likelihood ratio is defined by

$$\hat{l}_{\mathrm{ael}}(\beta) = \hat{r}(\beta)\{-2\log \hat{L}(\beta)\}, \tag{3.1}$$

where $\log \hat{L}(\beta)$ is defined in (2.2).

Theorem 4. *Suppose that conditions* (C1)–(C6) *hold. Then* $\hat{l}_{\mathrm{ael}}(\beta_0) \xrightarrow{D} \chi_p^2$.

According to Theorem 4, $\hat{l}_{\mathrm{ael}}(\beta)$ can be used to construct an approximate confidence region for β_0. Let

$$\mathcal{R}_{\mathrm{ael}}(\alpha) = \{\beta \in \mathcal{B}\colon\ \hat{l}_{\mathrm{ael}}(\beta) \le \chi_p^2(1-\alpha)\}.$$

Then, $\mathcal{R}_{\mathrm{ael}}(\alpha)$ gives a confidence region for β_0 with asymptotically correct coverage probability $1-\alpha$.

4. Numerical results

4.1. Bandwidth selection

Various existing bandwidth selection techniques for nonparametric regression, such as the cross-validation and generalized cross-validation, can be adapted for the estimation $\hat{g}(\cdot)$. But we, in our simulation, use the modified multi-fold cross-validation (MMCV) criterion proposed by Cai, Fan and Yao [3] to select the optimal bandwidth because the algorithm is simple and quick. Let m and Q be two given positive integers and $n > mQ$. The basic idea is first to use Q sub-series of lengths $n - km$ $(k = 1, \dots, Q)$ to estimate the unknown coefficient functions and then to compute the one-step forecasting error of the next section of the sample of lengths m based on the estimated models. More precisely, we choose h which minimizes

$$\mathrm{AMS}(h) = \sum_{k=1}^{Q} \mathrm{AMS}_k(h), \tag{4.1}$$

where, for $k = 1, \dots, Q$,

$$\mathrm{AMS}_k(h) = \frac{1}{m} \sum_{i=n-km+1}^{n-km+m} \left\{ Y_i - \sum_{j=1}^{q} \hat{g}_{j,k}(U_i) Z_{ij} \right\}^2,$$

and $\{\hat{g}_{j,k}(\cdot)\}$ are computed from the sample $\{(Y_i, U_i, Z_i), 1 \le i \le n - km\}$ with bandwidth equal $h(\frac{n}{n-km})^{1/5}$. Note that for different sample size, we re-scale bandwidth according to its optimal rate, that is, $h \propto n^{-1/5}$. Since the selected bandwidth does not depend critically on the choice of m and Q, to computation expediency, we take $m = [0.1n]$ and $Q = 4$ in our simulation.

Let h_{opt} be the bandwidth obtained by minimizing (4.1) with respect to $h > 0$; that is, $h_{\mathrm{opt}} = \inf_{h>0} \mathrm{AMS}(h)$. Then h_{opt} is the optimal bandwidth for estimating $\hat{g}(\cdot)$. When calculating the empirical likelihood ratios and estimator of β_0, we use the approximation bandwidth

$$h = h_{\mathrm{opt}} n^{-1/20} (\log n)^{-1/2}, \qquad h_1 = h_{\mathrm{opt}},$$

because this insures that the required bandwidth has correct order of magnitude for the optimal asymptotic performance (see, e.g., Carroll *et al.* [4]), and the bandwidth $\hat{h}$ satisfies condition (C4).

4.2. Simulation study

We now examine the performance of the procedures described in Sections 2 and 3. Consider the regression model

$$Y_i = g_0(\beta_0^T X_i) + g_1(\beta_0^T X_i) Z_{i1} + g_2(\beta_0^T X_i) Z_{i2} + \varepsilon_i, \tag{4.2}$$

where $\beta_0 = (1/\sqrt{5}, 2/\sqrt{5})^T$ and the ε_is are independent $N(0, 0.8^2)$ random variables. The sample $\{X_i = (X_{i1}, X_{i2})^T; 1 \le i \le n\}$ was generated from a bivariate uniform distribution on $[-1, 1]^2$ with independent components, $\{Z_i = (Z_{i1}, Z_{i2})^T; 1 \le i \le n\}$ was generated from a bivariate normal distribution $N(0, \Sigma)$ with $\mathrm{var}(Z_{i1}) = \mathrm{var}(Z_{i2}) = 1$ and the correlation coefficient between Z_{i1} and Z_{i2} is $\rho = 0.6$. In model (4.2), the coefficient functions are $g_0(u) = 12\exp(-2u^2)$, $g_1(u) = 10u^2$ and $g_2(u) = 16\sin(\pi u)$.

For the smoother, we used a local linear smoother with a Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ with a MMCV bandwidth throughout all smoothing steps. We take the weight function $w(u) = I_{[-3/\sqrt{5}, 3/\sqrt{5}]}(u)$. The sample size for the simulated data is 100, and the run is 500 times in all simulations.

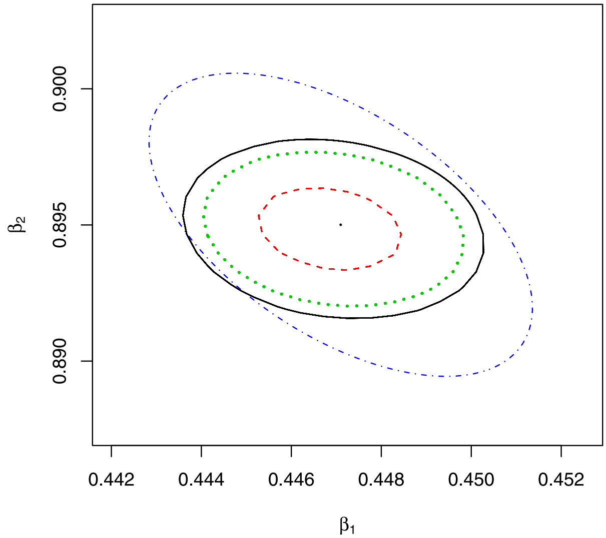


Figure 1. Averages of 95% confidence regions of (β_1, β_2), based on EEL (solid curve), AEL (dashed curve), IRSAEL (doted curve) and NA (dot-dashed curves) when $n = 100$.

The confidence regions of β_0 and their coverage probabilities, with nominal level $1-\alpha = 0.95$, were computed from 500 runs. Four methods were used to construct the confidence regions: the estimated empirical likelihood (EEL) with a conditional approximation, the adjusted empirical likelihood (AEL), the improved Rao–Scott adjusted empirical likelihood (IRSAEL) and the normal approximation (NA). A comparison among three methods was made through coverage accuracies and coverage areas of the confidence regions. The simulated results are given in Figure 1.

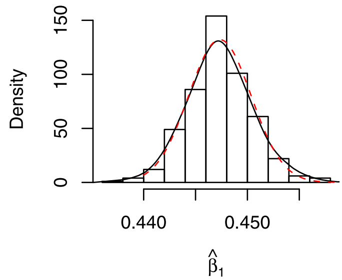
From Figure 1 we can see that EEL, AEL and IRSAEL give smaller confidence regions than NA, and the region obtained by AEL is much smaller than the others. Thus, AEL is the best of the four algorithms.

The histograms of the 500 estimators of the parameter β_1 and β_2 are in Figures 2(a) and (b), respectively. The Q–Q plots of the 500 estimators of the parameter β_1 and β_2 are in Figures 2(c) and (d), respectively.

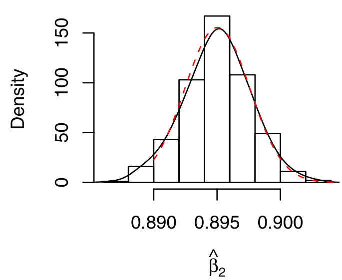
Figure 2 shows empirically that these estimators are asymptotically normal. The means of the estimates of the unknown parameters β_1 and β_2 are 0.44734 and 0.89502, respectively, and their biases (standard deviations) are 0.000131 (0.00302) and 0.000596 (0.00257), respectively.

We also consider the average estimates of the coefficient functions $g_0(u)$, $g_1(u)$ and $g_2(u)$ over the 500 replicates. The estimators $\hat{g}_j(\cdot)$ are assessed via the root mean squared errors (RMSE); that is, $\mathrm{RMSE} = \sum_{j=0}^{2} \mathrm{RMSE}_j$, where

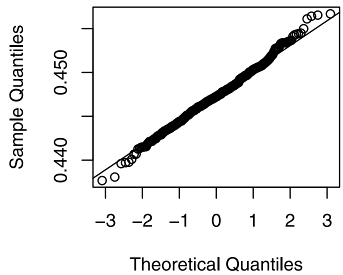
$$\mathrm{RMSE}_j = \left[n_{\mathrm{grid}}^{-1} \sum_{k=1}^{n_{\mathrm{grid}}} \{\hat{g}_j(u_k) - g_j(u_k)\}^2 \right]^{1/2},$$



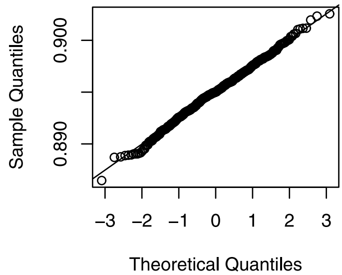
(a) Histogram



(b) Histogram



(c) Normal Q–Q Plot



(d) Normal Q–Q Plot

Figure 2. (a) for β_1 and (b) for β_2: the histograms of the 500 estimators of every parameter, the estimated curve of density (solid curve) and the curve of normal density (dashed curve); (c) for β_1 and (d) for β_2: the Q–Q plot of the 500 estimators of every parameter.

and $\{u_k, k = 1, \ldots, n_{\text{grid}}\}$ are regular grid points. The boxplot for the 500 RMSEs is given in Figure 3.

From Figures 3(a)–(c) we see every estimated curve agrees with the true function curve very closely. Figure 3(d) shows that all RMSEs of estimates for the unknown functions are very small.

Appendices

We divide the appendices into Appendix A and Appendix B. The proofs of Theorems 1–4 are presented in Appendix A, and the proofs of Lemmas 2 and 3 are presented in Appendix B. We use c to represent any positive constant which may take a different value for each appearance.

Appendix A: Proofs of theorems

The following lemma gives uniformly convergent rates of $\hat{g}(u;\beta)$ and $\hat{\hat{g}}(u;\beta)$. This lemma is a straightforward extension of known results in nonparametric function estimation; for its proof, the reader may refer to Theorem 2 in Wang and Xue [23], we hence omit the proof.

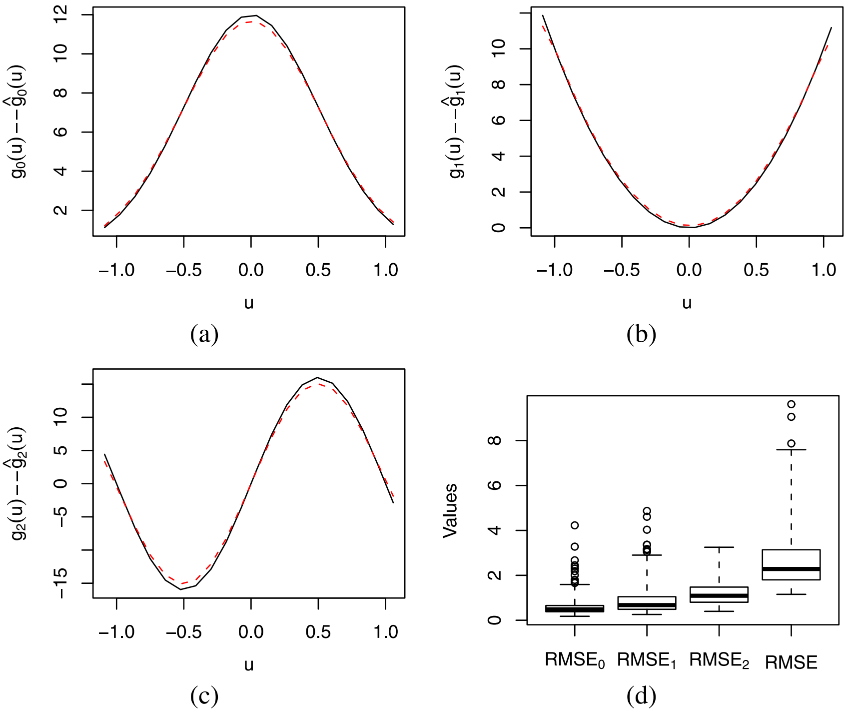


Figure 3. The true cure (solid curve) and the estimated curve (dashed curve). (a) for $g_0(\cdot)$, (b) for $g_1(\cdot)$, (c) for $g_2(\cdot)$; (d) the boxplots of the 500 RMSE values in estimations of $g_0(\cdot)$, $g_1(\cdot)$, $g_2(\cdot)$ and the sum of the three RMSEs.

Lemma 1. *Suppose that conditions* (C1)–(C3), (C5) *and* (C6) *hold. Then*

$$\sup_{u\in\mathcal{U}_w,\beta\in\mathcal{B}_n} \|\hat{\mathrm{g}}(u;\beta)-\mathrm{g}_0(u)\| = \mathrm{O}_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}+h^2\right)$$

and

$$\sup_{u\in\mathcal{U}_w,\beta\in\mathcal{B}_n} \|\hat{\dot{\mathrm{g}}}(u;\beta)-\dot{\mathrm{g}}_0(u)\| = \mathrm{O}_P\left(\left\{\frac{\log(1/h)}{nh^3}\right\}^{1/2}+h\right).$$

Denote $\mathcal{G}=\{\mathrm{g}\colon \mathcal{U}_w\times\mathcal{B}\mapsto R^q\}$, $\|\mathrm{g}\|_{\mathcal{G}}=\sup_{u\in\mathcal{U}_w,\beta\in\mathcal{B}_n}\|\mathrm{g}(u;\beta)\|$. From Lemma 1, we have $\|\hat{\mathrm{g}}-\mathrm{g}_0\|_{\mathcal{G}}=\mathrm{o}_P(1)$ and $\|\hat{\dot{\mathrm{g}}}-\dot{\mathrm{g}}_0\|_{\mathcal{G}}=\mathrm{o}_P(1)$; hence we can assume that g lies in $\mathcal{G}_\delta$ with $\delta=\delta_n\to 0$ and $\delta>0$, where

$$\mathcal{G}_\delta=\{\mathrm{g}\in\mathcal{G}\colon \|\mathrm{g}-\mathrm{g}_0\|_{\mathcal{G}}\le\delta,\ \|\dot{\mathrm{g}}-\dot{\mathrm{g}}_0\|_{\mathcal{G}}\le\delta\}. \tag{A.1}$$

Let $\mathrm{g}_0(\beta^TX;\beta)=E\{\mathrm{g}_0(\beta_0^TX)|\beta^TX\}$ and $\dot{\mathrm{g}}_0(\beta^TX;\beta)=E\{\dot{\mathrm{g}}_0(\beta_0^TX)|\beta^TX\}$,

$$Q(\mathrm{g},\beta)=E[\{Y-\mathrm{g}^T(\beta^TX;\beta)Z\}\dot{\mathrm{g}}^T(\beta^TX;\beta)ZXw(\beta^TX)], \tag{A.2}$$

$$Q_n(\mathrm{g},\beta)=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\mathrm{g}^T(\beta^TX_i;\beta)Z_i\}\dot{\mathrm{g}}^T(\beta^TX_i;\beta)Z_iX_iw(\beta^TX_i). \tag{A.3}$$

The following two lemmas are required for obtaining the proofs of the theorems; their proofs can be found in Appendix B.

Lemma 2. *Suppose that conditions* (C1)–(C6) *hold. Then*

$$\sup_{(\mathrm{g},\beta)\in\mathcal{G}_\delta\times\mathcal{B}_n}\|J_1(\mathrm{g},\beta)\| = \mathrm{o}_P(n^{-1/2}), \tag{A.4}$$

$$\sup_{\beta\in\mathcal{B}_n}\|J_2(\hat{\mathrm{g}},\beta)\| = \mathrm{o}_P(n^{-1/2}), \tag{A.5}$$

$$\sup_{(\mathrm{g},\beta)\in\mathcal{G}_\delta\times\mathcal{B}_n}\|J_3(\mathrm{g},\beta)\| = \mathrm{o}(n^{-1/2}), \tag{A.6}$$

$$\sqrt{n}J_4(\hat{\mathrm{g}},\beta_0)\xrightarrow{D}N(0,\sigma^2A(\beta_0)), \tag{A.7}$$

where $A(\beta_0)$ is defined in (2.6),

$$\begin{aligned}
J_1(\mathrm{g},\beta)&=Q_n(\mathrm{g},\beta)-Q(\mathrm{g},\beta)-Q_n(\mathrm{g}_0,\beta_0),\\
J_2(\mathrm{g},\beta)&=Q(\mathrm{g},\beta)-Q(\mathrm{g}_0,\beta)\\
&\quad-\varpi(\mathrm{g}_0(\beta^TX;\beta);\beta)\{\mathrm{g}(\beta^TX;\beta)-\mathrm{g}_0(\beta^TX;\beta)\},\\
J_3(\mathrm{g},\beta)&=\varpi(\mathrm{g}_0(\beta^TX),\beta)\{\mathrm{g}(\beta^TX;\beta)-\mathrm{g}_0(\beta^TX)\}\\
&\quad-\varpi(\mathrm{g}_0(\beta_0^TX;\beta),\beta_0)\{\mathrm{g}(\beta_0^TX;\beta_0)-\mathrm{g}_0(\beta_0^TX;\beta)\}
\end{aligned}$$

and

$$J_4(\beta_0,\mathrm{g})=Q_n(\mathrm{g}_0,\beta_0)+\varpi(\mathrm{g}_0(\beta_0^TX),\beta_0)\{\mathrm{g}(\beta_0^TX;\beta_0)-\mathrm{g}_0(\beta_0^TX)\}.$$

Lemma 3. *Suppose that conditions* (C1)–(C6) *hold. Then*

$$\sup_{\beta\in\mathcal{B}_n}\|Q_n(\hat{\mathrm{g}},\beta)\|=\mathrm{O}_P(n^{-1/2}), \tag{A.8}$$

$$\sup_{\beta\in\mathcal{B}_n}\|R_n(\beta)-\sigma^2B(\beta_0)\|=\mathrm{o}_P(1), \tag{A.9}$$

$$\sup_{\beta\in\mathcal{B}_n}\max_{1\le i\le n}\|\hat{\eta}_i(\beta)\|=\mathrm{o}_P(n^{1/2}), \tag{A.10}$$

$$\sup_{\beta\in\mathcal{B}_n}\|\lambda(\beta)\|=\mathrm{o}_P(n^{-1/2}), \tag{A.11}$$

where $Q_n(\hat{\mathrm{g}},\beta)$ is defined in (A.3), *$R_n(\beta)=n^{-1}\sum_{i=1}^n\hat{\eta}_i(\beta)\hat{\eta}_i^T(\beta)$, $B(\beta_0)$ is defined in condition* (C7) *and $\hat{\eta}_i(\beta)$ is defined in* (2.2).

Proof of Theorem 1. Note that, when $\beta = \beta_0$, Lemma 3 also holds. Applying the Taylor expansion to (2.2) and invoking Lemma 3, we can obtain

$$-2\log \hat{L}(\beta_0) = -\sum_{i=1}^{n}\left[\lambda^T \hat{\eta}_i(\beta_0) - \frac{1}{2}\{\lambda^T \hat{\eta}_i(\beta_0)\}^2\right] + o_P(1). \tag{A.12}$$

By (2.3) and Lemma 3, we have

$$\sum_{i=1}^{n}\{\lambda^T \hat{\eta}_i(\beta_0)\}^2 = \sum_{i=1}^{n}\lambda^T \hat{\eta}_i(\beta_0) + o_P(1)$$

and

$$\lambda = \left\{\sum_{i=1}^{n}\hat{\eta}_i(\beta_0)\hat{\eta}_i^T(\beta_0)\right\}^{-1}\sum_{i=1}^{n}\hat{\eta}_i(\beta_0) + o_P(n^{-1/2}).$$

This together with (A.12) proves that

$$-2\log \hat{L}(\beta_0) = nQ_n^T(\hat{g},\beta_0)R_n^{-1}(\beta_0)Q_n(\hat{g},\beta_0) + o_P(1), \tag{A.13}$$

where $Q_n(\hat{g},\beta_0)$ and $R_n(\beta_0)$ are defined in (A.3) and (A.9), respectively. From (A.9) of Lemma 3 and (A.13), we obtain

$$-2\log \hat{L}(\beta_0) = \left\{(\sigma^2 A)^{-1/2}\sqrt{n}Q_n(\hat{g},\beta_0)\right\}^T G(\beta_0)\left\{(\sigma^2 A)^{-1/2}\sqrt{n}Q_n(\hat{g},\beta_0)\right\} + o_P(1), \tag{A.14}$$

where $G(\beta_0) = A^{1/2}(\beta_0)B^{-1}(\beta_0)A^{1/2}(\beta_0)$. Let $G_0 = \operatorname{diag}(w_1,\ldots,w_p)$, where w_i, $1 \le i \le p$, are the eigenvalues of $G(\beta_0)$. Then there exists an orthogonal matrix H such that $H^T G_0 H = G(\beta_0)$. Using the notations of Lemma 2, we have

$$Q_n(\hat{g},\beta) = J_1(\hat{g},\beta) + J_2(\hat{g},\beta) + J_3(\hat{g},\beta) + J_4(\hat{g},\beta_0) + Q(g_0,\beta). \tag{A.15}$$

Noting that $Q(g_0,\beta_0) = 0$, from the above equation and Lemma 2, we have

$$Q_n(\hat{g},\beta_0) = J_4(\hat{g},\beta_0) + o_P(n^{-1/2}).$$

Hence, by (A.7) of Lemma 2, we have

$$H\{\sigma^{-2}A^{-}(\beta_0)\}^{1/2}\sqrt{n}Q_n(\hat{g},\beta_0) \xrightarrow{D} N(0, I_p),$$

where I_p is the $p \times p$ identity matrix. This together with (A.14) proves Theorem 1. □

Proof of Theorem 2. Under the conditions of Theorem 2, we can follow similar arguments to those used by Wang and Xue [23] and show that $\hat{\beta}$ is a root-n consistent estimator of β_0. Because the proof is straightforward, we do not present it here. We next demonstrate the asymptotic normality of $\hat{\beta}$. By Lemma 3 and, similarly to the proof of (A.13), we can obtain

$$\log \hat{L}(\beta) = -\frac{n}{2} Q_n^T(\hat{g}, \beta)\{\sigma^2 B(\beta)\}^{-1} Q_n(\hat{g}, \beta) + o_P(1), \tag{A.16}$$

uniformly for $\beta \in \mathcal{B}_n$, where $o_P(1)$ tends to 0 in probability uniformly for $\beta \in \mathcal{B}_n$. Since the estimator $\hat{\beta}$ is a maximum of $\log \hat{L}(\beta)$, and $B(\beta_0)$ is a positive definite matrix, the resulting estimator $\hat{\beta}$ is equivalent to solving the estimation equation $Q_n(\hat{g}, \beta) = 0$; that is, $Q_n(\hat{g}, \hat{\beta}) = 0$. Note that $Q(g_0, \beta_0) = 0$, and we then have, by Taylor's expansion, that

$$Q(g_0, \beta) = -B_*(\beta_0)(\beta - \beta_0) + o(n^{-1/2}), \tag{A.17}$$

uniformly for $\beta \in \mathcal{B}_n$, where $B_*(\beta_0)$ is the same as that in (A.9). By (A.15), (A.17) and (A.4)–(A.6) of Lemma 2, we have

$$Q_n(\hat{g}, \hat{\beta}) = J_4(\hat{g}, \beta_0) - B_*(\beta_0)(\hat{\beta} - \beta_0) + o_P(n^{-1/2}).$$

Noting that $Q_n(\hat{g}, \hat{\beta}) = 0$, we get

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sqrt{n} B_*^{-1}(\beta_0) J_4(\hat{g}, \beta_0) + o_P(1).$$

This together with (A.7) of Lemma 2 proves Theorem 2. □

Proof of Theorem 3. Decomposing $\hat{\sigma}^2$ into several parts, we get

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n [g_0(X_i^T\beta_0) - \hat{g}(X_i^T\hat{\beta}; \hat{\beta})\}^T Z_i]^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n \varepsilon_i \{g_0(X_i^T\beta_0) - \hat{g}(X_i^T\hat{\beta}; \hat{\beta})\}^T Z_i \\
&\equiv I_1 + I_2 + I_3.
\end{aligned}$$

Using the central limit theorem, we have

$$\sqrt{n}(I_1 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) \xrightarrow{D} N(0, \operatorname{var}(\varepsilon^2)).$$

By Lemma 1, we can obtain

$$|I_2| \le \frac{1}{n}\sum_{i=1}^n \|Z_i\|^2 \Big\{ \sup_{(u,\beta)\in(\mathcal{U}_w, \mathcal{B}_n)} \|\hat{g}(u; \beta) - g_0(u)\| \Big\}^2 = o_P(n^{-1/2}).$$

For I_3, we have

$$\begin{aligned}
I_3 &= \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\{\mathrm{g}_0(X_i^T\beta_0)-\hat{\mathrm{g}}(X_i^T\beta_0;\beta_0)\}^T Z_i \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\{\hat{\mathrm{g}}(X_i^T\beta_0;\beta_0)-\hat{\mathrm{g}}(X_i^T\hat{\beta};\hat{\beta})\}^T Z_i \\
&\equiv I_{31}+I_{32}.
\end{aligned}$$

It is not hard to show that $I_{31}=\mathrm{O}_P(n^{-1/2})$. By Theorems 1 and 3, we obtain

$$|I_{32}| \le \frac{2}{n}\sum_{i=1}^{n}\bigl(\|Z_i\||\varepsilon_i|\|X_i-E(X_i|\beta_0^T X_i)\|\bigr)\|\hat{\beta}-\beta_0\|\mathrm{O}_P(1)=\mathrm{O}_P(n^{-1/2}).$$

This together with above results proves Theorem 3. □

Proof of Theorem 4. Note that $\hat{A}(\beta_0)\xrightarrow{P}A(\beta_0)$ and $\hat{B}(\beta_0)\xrightarrow{P}B(\beta_0)$. By the expansion of $\hat{l}_{\mathrm{ael}}(\beta_0)$, defined in (3.1) and (A.16), we get

$$\hat{l}_{\mathrm{ael}}(\beta_0)=nQ_n^T(\hat{\mathrm{g}},\beta_0)\{\sigma^{-2}A^{-}(\beta_0)\}Q_n(\hat{\mathrm{g}},\beta_0)+\mathrm{o}_P(1). \tag{A.18}$$

This together with (A.15) and (A.18) proves Theorem 4. □

Appendix B: Proofs of lemmas

Proof of Lemma 2. We first prove (A.4). Denote $r_n(\mathrm{g},\beta)=\sqrt{n}\{Q_n(\mathrm{g},\beta)-Q(\mathrm{g},\beta)\}$. Noting that $Q(\mathrm{g}_0,\beta_0)=0$, we clearly have

$$J_1(\mathrm{g},\beta)=n^{-1/2}\{r_n(\mathrm{g},\beta)-r_n(\mathrm{g}_0,\beta_0)\}. \tag{B.1}$$

It can be shown that the empirical process $\{r_n(\mathrm{g},\beta)\colon \mathrm{g}\in\mathcal{G}_1,\beta\in\mathcal{B}_1\}$ has the stochastic equicontinuity, where $\mathcal{B}_1=\{\beta\in\mathcal{B}\colon \|\beta-\beta_0\|\le 1\}$ and $\mathcal{G}_1$ are defined in (A.1) with $\delta=1$, which are subsets of $\mathcal{B}$ and $\mathcal{G}$, respectively. The equicontinuity is sufficient for proof of (A.4) since $\delta<1$ for large enough n. This stochastic equicontinuity follows by checking the conditions of Theorem 1 in Doukhan, Massart and Rio [7]. Therefore, we have $r_n(\mathrm{g},\beta)-r_n(\mathrm{g}_0,\beta_0)=\mathrm{o}_P(1)$, uniformly for $\beta\in\mathcal{B}_1$ and $\mathrm{g}\in\mathcal{G}_1$. This together with (B.1) proves (A.4).

We now prove (A.5). Define the functional derivative $\varpi(\mathrm{g}_0(\cdot;\beta),\beta)$ of $Q(\mathrm{g},\beta)$ with respect to $\mathrm{g}(\cdot;\beta)$ at $\mathrm{g}_0(\cdot;\beta)$ at the direction $\mathrm{g}(\cdot;\beta)-\mathrm{g}_0(\cdot;\beta)$ by

$$\begin{aligned}
&\varpi(\mathrm{g}_0(\cdot;\beta),\beta)\{\mathrm{g}(\cdot;\beta)-\mathrm{g}_0(\cdot;\beta)\} \\
&\quad=\lim_{\tau\to 0}\bigl[Q\bigl(\mathrm{g}_0(\cdot;\beta)+\tau\bigl(\mathrm{g}(\cdot;\beta)-\mathrm{g}_0(\cdot;\beta)\bigr),\beta\bigr)-Q(\mathrm{g}_0(\cdot;\beta),\beta)\bigr]\cdot\frac{1}{\tau},
\end{aligned}$$

where $Q(\mathrm{g},\beta)$ is defined in (A.2). We have

$$\begin{aligned}&\varpi(\mathrm{g}_0(\beta^T X;\beta),\beta)\{\mathrm{g}(\beta^T X;\beta)-\mathrm{g}_0(\beta^T X;\beta)\}\\&\quad=-E[\{\mathrm{g}(\beta^T X;\beta)-\mathrm{g}_0(\beta^T X;\beta)\}^T Z\dot{\mathrm{g}}_0^T(\beta^T X;\beta)ZXw(\beta^T X)].\end{aligned}\tag{B.2}$$

It follows from (B.2) that

$$\begin{aligned}J_2(\mathrm{g},\beta)=-E[&\{\mathrm{g}(\beta^T X;\beta)-\mathrm{g}_0(\beta_0^T X)\}^T ZXZ^T\\&\times\{\dot{\mathrm{g}}(\beta^T X;\beta)-\dot{\mathrm{g}}_0(\beta^T X;\beta)\}w(\beta^T X)],\end{aligned}$$

and hence we have

$$\begin{aligned}\omega^T J_2(\hat{\mathrm{g}},\beta)=-\int&\{\hat{\mathrm{g}}(u;\beta)-\mathrm{g}_0(u)\}^T\mu_\omega(u)\\&\times\{\hat{\dot{\mathrm{g}}}(u;\beta)-\dot{\mathrm{g}}_0(u)\}w(u)f(u)\,\mathrm{d}u+\mathrm{o}_P(n^{-1/2})\end{aligned}\tag{B.3}$$

for any p-dimension vector ω, where $\mu_\omega(u)=E\{Z\omega^T XZ^T|\beta^T X=u\}$, and $f(u)$ is the probability density of $\beta^T X$. Using the standard argument of nonparametric estimation, we can prove

$$\hat{\mathrm{g}}(u;\beta)-\mathrm{g}_0(u)=D^{-1}(u)\{f(u)\}^{-1}\xi_n(u;\beta)+\mathrm{O}_P(c_n),\tag{B.4}$$

uniformly for $u\in\mathcal{U}_w$ and $\beta\in\mathcal{B}_n$, where $c_n=n^{-1/2}+h^2$ and $D(u)$ is defined in condition (C6).

$$\xi_n(u;\beta)=\frac{1}{n}\sum_{i=1}^n Z_i\{Y_i-\mathrm{g}_0^T(\beta^T X_i)Z_i\}K_h(\beta^T X_i-u).$$

This together with (B.3) derives that

$$\begin{aligned}\omega^T J_2(\hat{\mathrm{g}},\beta)&=-\int\{D^{-1}(u)\xi_n(u;\beta)\}^T\mu_\omega(u)\{\hat{\dot{\mathrm{g}}}(u;\beta)-\dot{\mathrm{g}}_0(u)\}\,\mathrm{d}u+\mathrm{O}_P(c_n)\\&=-n^{-1/2}\{\gamma_n(\hat{\dot{\mathrm{g}}},\beta)-\gamma_n(\dot{\mathrm{g}}_0,\beta)\}+\mathrm{O}_P(c_n),\end{aligned}$$

where $\gamma_n(\dot{\mathrm{g}},\beta)=n^{-1/2}\sum_{i=1}^n\varepsilon_i w(\beta^T X_i)Z_i^T D^{-1}(\beta^T X_i)\mu_\omega(\beta^T X_i)\dot{\mathrm{g}}(\beta^T X_i;\beta)$. Using the empirical process techniques, and similarly to the proof of (A.4), we can show that the stochastic equicontinuity of $\gamma_n(\dot{\mathrm{g}},\beta)$, and hence $\|\gamma_n(\hat{\dot{\mathrm{g}}},\beta)-\gamma_n(\dot{\mathrm{g}}_0,\beta)\|=\mathrm{o}_P(1)$. Also, $nh^4=\mathrm{O}(1)$ implies $h^2=\mathrm{O}(n^{-1/2})$, and hence $c_n=\mathrm{O}(n^{-1/2})$. Thus, the proof of (A.5) is complete.

We now prove (A.6). Denote $\psi(\dot{\mathrm{g}}_0,\beta)=\dot{\mathrm{g}}_0^T(\beta^T X;\beta)ZXw(\beta^T X)$ and $\varphi(\mathrm{g},\beta)=\{\mathrm{g}(\beta^T X;\beta)-\mathrm{g}_0(\beta^T X;\beta)\}^T Z$. It follows from (B.2) that

$$\begin{aligned}J_3(\mathrm{g},\beta)&=-E\{\varphi(\mathrm{g},\beta)\psi(\dot{\mathrm{g}}_0,\beta)\}+E\{\varphi(\mathrm{g},\beta_0)\psi(\dot{\mathrm{g}}_0,\beta_0)\}\\&=-E[\{\varphi(\mathrm{g},\beta)-\varphi(\mathrm{g},\beta_0)\}\psi(\dot{\mathrm{g}}_0,\beta)]\\&\quad-E[\varphi(\mathrm{g},\beta_0)\{\psi(\dot{\mathrm{g}}_0,\beta)-\psi(\dot{\mathrm{g}}_0,\beta_0)\}]\\&\equiv J_{31}(\mathrm{g},\beta)+J_{32}(\mathrm{g},\beta).\end{aligned}$$

By condition (C2), we get

$$\begin{aligned}
&\|\varphi(\mathrm{g},\beta)-\varphi(\mathrm{g},\beta_0)\| \\
&\quad=\|[\{\mathrm{g}(\beta^T X;\beta)-\mathrm{g}(\beta_0^T X;\beta_0)\}-\{\mathrm{g}_0(\beta^T X;\beta)-\mathrm{g}_0(\beta_0^T X)\}]^T Z\| \\
&\quad=\|[\{\dot{\mathrm{g}}(\beta_1^T X;\beta_1)-\dot{\mathrm{g}}_0(\beta_2^T X)\}(\beta-\beta_0)^T\{X-E(X|\beta_0^T X)\}]^T Z\| \\
&\quad\le c\|\dot{\mathrm{g}}-\dot{\mathrm{g}}_0\|_{\mathcal{G}}\|\beta-\beta_0\|(\|X-E(X|\beta_0^T X)\|)(\|Z\|),
\end{aligned}$$

where β_1 and β_2 are between β and β_0, and $\|\psi(\dot{\mathrm{g}}_0,\beta)\|\le c(\|Z\|)(\|X\|)$. Therefore, we have $\|J_{31}(\mathrm{g},\beta)\|=\mathrm{o}(n^{-1/2})$, uniformly for $\mathrm{g}\in\mathcal{G}_\delta$ and $\beta\in\mathcal{B}_n$. Similarly, we can prove $\|J_{32}(\mathrm{g},\beta)\|=\mathrm{o}(n^{-1/2})$, uniformly for $\mathrm{g}\in\mathcal{G}_\delta$ and $\beta\in\mathcal{B}_n$, and hence (A.6) follows.

Finally, we prove (A.7). Let $f_0(u)$ denote the density function of $\beta_0^T X$. By (B.2) and (B.4), and using the dominated convergence theorem (Loève [14]), we can obtain

$$\begin{aligned}
&\varpi(\mathrm{g}_0(\beta_0^T X),\beta_0)\{\hat{\mathrm{g}}(\beta_0^T X;\beta_0)-\mathrm{g}_0(\beta_0^T X)\} \\
&\quad=-\int C(u)\{\hat{\mathrm{g}}(u;\beta_0)-\mathrm{g}_0(u)\}f_0(u)\,\mathrm{d}u \\
&\quad=-\frac{1}{n}\sum_{i=1}^n \varepsilon_i C(\beta_0^T X_i)D^{-1}(\beta_0^T X_i)Z_i+\mathrm{o}_P(c_n).
\end{aligned}$$

This together with (A.3) proves that

$$J_4(\hat{\mathrm{g}},\beta_0)=\frac{1}{n}\sum_{i=1}^n \varepsilon_i\zeta_i+\mathrm{o}_P(c_n),$$

where $\zeta_i=V_i-C(\beta_0^T X_i)D^{-1}(\beta_0^T X_i)Z_i$ and $V_i=X_i\dot{\mathrm{g}}_0^T(\beta_0^T X_i)Z_i w(\beta_0^T X_i)$. Therefore, by the central limit theorem and Slutsky's theorem, we get

$$\sqrt{n}J_4(\hat{\mathrm{g}},\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\zeta_i+\mathrm{o}_P(1)\xrightarrow{D}N(0,\sigma^2 A(\beta_0)).$$

This proves (A.7). The proof of Lemma 2 is complete. □

Proof of Lemma 3. By (A.15), (A.17) and Lemma 2, we can prove (A.8). We now prove (A.9). Let

$$\begin{aligned}
R_{ni}(\beta)&=\varepsilon_i\dot{\mathrm{g}}_0^T(\beta_0^T X_i)Z_i X_i\{w(\beta^T X_i)-w(\beta_0^T X_i)\} \\
&\quad+\varepsilon_i\{\hat{\dot{\mathrm{g}}}(\beta^T X_i;\beta)-\dot{\mathrm{g}}_0(\beta_0^T X_i)\}^T Z_i X_i w(\beta^T X_i) \\
&\quad+\{\mathrm{g}_0(\beta_0^T X_i)-\hat{\mathrm{g}}(\beta^T X_i;\beta)\}^T Z_i Z_i^T\dot{\mathrm{g}}_0(\beta_0^T X_i)X_i w(\beta^T X_i) \\
&\quad+\{\mathrm{g}_0(\beta_0^T X_i)-\hat{\mathrm{g}}(\beta^T X_i;\beta)\}^T Z_i Z_i^T \\
&\quad\times\{\hat{\dot{\mathrm{g}}}(\beta^T X_i;\beta)-\dot{\mathrm{g}}_0^T(\beta_0^T X_i)\}X_i w(\beta^T X_i).
\end{aligned}$$

Then we have $\hat{\eta}_i(\beta) = \eta_i(\beta_0) + R_{ni}(\beta)$, where $\eta_i(\cdot)$ is defined in (2.1), and hence

$$\begin{aligned} R_n(\beta) &= \frac{1}{n}\sum_{i=1}^n \eta_i(\beta_0)\eta_i^T(\beta_0) + \frac{1}{n}\sum_{i=1}^n R_{ni}(\beta)R_{ni}^T(\beta) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \eta_i(\beta_0)R_{ni}^T(\beta) + \frac{1}{n}\sum_{i=1}^n R_{ni}(\beta)\eta_i^T(\beta_0) \\ &\equiv M_1(\beta_0) + M_2(\beta) + M_3(\beta) + M_4(\beta). \end{aligned} \tag{B.5}$$

By the law of large numbers, we have $M_1(\beta_0) \xrightarrow{P} \sigma^2 B(\beta_0)$. Therefore, to prove (A.9), we only need to show that $M_k(\beta) \xrightarrow{P} 0$ uniformly for β, $k = 2, 3, 4$.

Let $M_{2,st}(\beta)$ denote the (s,t) element of $M_2(\beta)$, and $R_{ni,s}(\beta)$ denote the sth component of $R_{ni}(\beta)$. Then by the Cauchy–Schwarz inequality, we have

$$|M_{2,st}(\beta)| \le \left(\frac{1}{n}\sum_{i=1}^n R_{ni,s}^2(\beta)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n R_{ni,t}^2(\beta)\right)^{1/2}. \tag{B.6}$$

It can be shown by a direct calculation that

$$\frac{1}{n}\sum_{i=1}^n R_{ni,s}^2(\beta) \xrightarrow{P} 0,$$

uniformly for $\beta \in \mathcal{B}_n$. This together with (B.6) proves that $M_2(\beta) \xrightarrow{P} 0$, uniformly for $\beta \in \mathcal{B}_n$. Similarly, it can be shown that $M_3(\beta) \xrightarrow{P} 0$ and $M_4(\beta) \xrightarrow{P} 0$, uniformly for $\beta \in \mathcal{B}_n$. This together with (B.5) proves (A.9).

Similarly to above proof, we can derive (A.10). (A.11) can be shown by using (A.8)–(A.10), and employing the same arguments used in the proof of (2.14) in Owen [16]. □

Acknowledgements

We are grateful for the many detailed suggestions of the editor, the associate editor and the referees, which led to significant improvements of the paper. Liugen Xue's research was supported by the National Natural Science Foundation of China (10871013,11171012), the Beijing Natural Science Foundation (1102008), the Beijing Municipal Education Commission Foundation (KM201110005029) and the PHR(IHLB). Qihua Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), National Natural Science Foundation of China (10671198), the National Science Fund for Creative Research Groups in China and a grant from the Key Lab of Random Complex Structure and Data Science, CAS and the Key grant from Yunnan Province (2010CC003).

References

[1] Arnold, S.F. (1981). *The Theory of Linear Models and Multivariate Analysis*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0606011
[2] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446
[3] Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95** 941–956. MR1804449
[4] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842
[5] Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88** 298–308. MR1212492
[6] Chiou, J.M. and Müller, H.G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* **93** 1376–1387. MR1666634
[7] Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **31** 393–427. MR1324814
[8] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. London: Chapman & Hall. MR1383587
[9] Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 57–80. MR1959093
[10] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497
[11] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171
[12] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion and a reply by the authors). *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881
[13] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29** 595–623. MR1865333
[14] Loève, M. (2000). *Probability Theory I*, 4th ed. New York: Springer.
[15] Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: Asymptotic theory. *Statist. Sinica* **17** 177–197. MR2352508
[16] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387
[17] Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049
[18] Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 101–122. MR2301502
[19] Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76** 221–230. MR0624328
[20] Stute, W., Xue, L. and Zhu, L. (2007). Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. *J. Amer. Statist. Assoc.* **102** 332–346. MR2345546
[21] Wang, Q., Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.* **99** 334–345. MR2062820
[22] Wang, Q. and Rao, J.N.K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika* **89** 345–358. MR1913963
[23] Wang, Q. and Xue, L. (2011). Statistical inference in partially-varying-coefficient single-index model. *J. Multivariate Anal.* **102** 1–19. MR2729416

[24] Weisberg, S. and Welsh, A.H. (1994). Adapting for the missing link. *Ann. Statist.* **22** 1674–1700. MR1329165
[25] Wu, C.O., Chiang, C.T. and Hoover, D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **93** 1388–1402. MR1666635
[26] Xia, Y. and Li, W.K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94** 1275–1285. MR1731489
[27] Xue, L. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scand. J. Stat.* **36** 671–685. MR2573302
[28] Xue, L. (2009). Empirical likelihood for linear models with missing responses. *J. Multivariate Anal.* **100** 1353–1366. MR2514134
[29] Xue, L. and Zhu, L. (2007). Empirical likelihood for a varying coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **102** 642–654. MR2370858
[30] Xue, L. and Zhu, L. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika* **94** 921–937. MR2416799
[31] Xue, L. and Zhu, L. (2006). Empirical likelihood for single-index models. *J. Multivariate Anal.* **97** 1295–1312. MR2279674
[32] Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 549–570. MR2278341
[33] Zhu, L. and Fang, K.T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24** 1053–1068. MR1401836

Received August 2010