Journal of Mathematical Research with Applications
Jul., 2013, Vol. 33, No. 4, pp. 493–504
DOI:10.3770/j.issn:2095-2651.2013.04.012
Http://jmre.dlut.edu.cn

Empirical Likelihood Based Variable Selection for Varying Coefficient Partially Linear Models with Censored Data

Peixin ZHAO
Department of Mathematics, Hechi University, Guangxi 546300*, P. R. China*

Abstract In this paper, we consider the variable selection for the parametric components of varying coefficient partially linear models with censored data. By constructing a penalized auxiliary vector ingeniously, we propose an empirical likelihood based variable selection procedure, and show that it is consistent and satisfies the sparsity. The simulation studies show that the proposed variable selection method is workable.

Keywords varying coefficient partially linear models; empirical likelihood; censored data; variable selection.

MR(2010) Subject Classification 62G08; 62G20

1. Introduction

The varying coefficient partially linear model has proved to be very useful as it combines the flexibility of nonparametric models and the interpretation of linear models. The model structure can be defined as follows

$$Y_i = X_i^{\mathrm{T}}\theta(U_i) + Z_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where β is a $q \times 1$ vector of unknown parameters, $\theta(\cdot)$ is a $p \times 1$ vector of unknown function, X_i, Z_i and U_i are covariates, and ϵ_i is the model error with $E(\epsilon_i|X_i, Z_i, U_i) = 0$.

Recently, a variety of methods have been proposed for the estimation and variable selection of model (1). Li et al. [1], Fan and Huang [2], and You and Zhou [3] considered the estimation of model (1) based on different methods. Li and Liang [4], and Zhao and Xue [5] considered the variable selection procedure for model (1) via nonconcave penalized likelihood method. An essential assumption in these papers is that all data can be observed directly. However, the presence of censoring causes major difficulties in the implementation of the existing approaches, because the value of Y_i is unknown for the censored observations.

Outcome censoring often occurs in many disciplines such as econometrics, biostatistics and bio-informatics. There have been many recent researches in the area of statistical inference for

Received March 15, 2012; Accepted September 3, 2012
Supported by the National Natural Science Foundation of China (Grant Nos. 11101119; 11126332), the National Social Science Foundation of China (Grant No. 11CTJ004), the Natural Science Foundation of Guangxi Province (Grant No. 2010GXNSFB013051) and the Philosophy and Social Sciences Foundation of Guangxi Province (Grant No. 11FTJ002).
E-mail address: zpx81@163.com

censored data. Based on the empirical likelihood method, Wang and Li [6] studied the estimation of the parametric components in a partly linear model with right censored data, and Yang et al. [7] considered the estimation for a partially linear single-index model with right censored data. However, the literature on empirical likelihood based variable selection procedure is relativey thin.

In this paper, we consider the variable selection for a varying coefficient partially linear model with right censored data based on a penalized empirical likelihood method. Specifically, we assume that the available data $\{(Y_i^*, \delta_i, X_i, Z_i, U_i),\ i = 1, \dots, n\}$, are independent and identically distributed, where $Y_i^* = \min\{Y_i,\ C_i\}$, $\delta_i = I(Y_i \leq C_i)$ and C_i is a censoring variable. We also assume that $\{C_i,\ i = 1, \dots, n\}$ are i.i.d. variables with a distribution function $G(\cdot)$, and Y_i is conditionally independent of C_i for the given prognostic variables (X_i, Z_i, U_i).

We propose an empirical likelihood based variable selection procedure for the parametric components of model (1) with right censored data, and investigate the asymptotic properties of the proposed variable selection method such as the consistence and sparsity. We also evaluate the performance of the proposed variable selection procedure by some simulation studies. The simulation results show that the empirical likelihood based variable selection procedure is workable.

Variable selection is an important topic in high-dimensional statistical modeling. Several variable selection procedures have been developed in the literature, including the sequential approach, prediction-error approach, and information-theoretic approach. But all of these variable selection methods are computationally expensive. Recently, a new method based on penalized likelihood has been lauded for its computational efficiency and stability. In this approach, the parametric likelihood is a crucial component, but in many situations a well-defined parametric likelihood is not easy to construct. The empirical likelihood based variable selection procedure, proposed in this paper, can overcome this problem, because it is constructed based on a set of estimating equations.

2. Methodology and results

As in Yang el al. [7], we define a synthetic variable $Y_{iG}^* = Y_i^* \delta_i / (1 - G(Y_i^*))$. It can be verified that $E(Y_{iG}^* | X_i, Z_i, U_i) = E(Y_i | X_i, Z_i, U_i)$. This implies that

$$Y_{iG}^* = X_i^{\mathrm{T}} \theta(U_i) + Z_i^{\mathrm{T}} \beta + \epsilon_i^*, \quad i = 1, \dots, n, \tag{2}$$

where $E(\epsilon_i^* | X_i, Z_i, U_i)$. When $G(\cdot)$ is known as well, model (2) is a standard varying coefficient partially linear model. For given β, using the same arguments as Fan and Huang [2], we can get the weighted local least-squares estimator of $\theta(u)$ by minimizing

$$\sum_{i=1}^{n} \Big\{ Y_{iG}^* - Z_i^{\mathrm{T}} \beta - \sum_{k=1}^{p} [a_k + b_k(U_i - u)] X_{ik} \Big\}^2 K_h(U_i - u), \tag{3}$$

where $K_h(\cdot) = h^{-1} K(\cdot / h)$, $K(\cdot)$ is a kernel function, h is a bandwidth and X_{ik} denotes the kth component of X_i.

Let $Z=(Z_1,\ldots,Z_n)^{\mathrm{T}}$, $Y_G^*=(Y_{1G}^*,\ldots,Y_{nG}^*)^{\mathrm{T}}$, I_p be $p\times p$ identity matrix, 0_p be $p\times p$ zero matrix, and $\Omega_u=\mathrm{diag}(K_h(u-U_1),\ldots,K_h(u-U_n))$ be $n\times n$ diagonal matrix. Then, the solution by minimizing (3) can be given by

$$\tilde{\theta}(u)=(I_p,0_p)(D_u^{\mathrm{T}}\Omega_u D_u)^{-1}D_u^{\mathrm{T}}\Omega_u(Y_G^*-Z\beta), \tag{4}$$

where $D_u=\begin{pmatrix} X_1 & \cdots & X_n \\ h^{-1}(U_1-u)X_1 & \cdots & h^{-1}(U_n-u)X_n\end{pmatrix}^{\mathrm{T}}$.

Let $(I_p,0_p)(D_u^{\mathrm{T}}\Omega_u D_u)^{-1}D_u^{\mathrm{T}}\Omega_u\equiv(S_1(u),\ldots,S_n(u))$, $\tilde{\mu}(u)=\sum_{k=1}^n S_k(u)Z_k^{\mathrm{T}}$, and $\tilde{g}(u)=\sum_{k=1}^n S_k(u)Y_{kG}^*$. Then we have $\tilde{\theta}(u)=\tilde{g}(u)-\tilde{\mu}(u)\beta$. Substituting this into (2), and by a simple calculation we have

$$\breve{Y}_{iG}^*=\breve{Z}_i^{\mathrm{T}}\beta+\epsilon_i, \tag{5}$$

where $\breve{Z}_i=Z_i-\tilde{\mu}(U_i)^{\mathrm{T}}X_i$, $\breve{Y}_{iG}^*=Y_{iG}^*-\tilde{g}(U_i)^{\mathrm{T}}X_i$. To give the empirical likelihood based variable selection procedure, we introduce the following penalized auxiliary random vector

$$\breve{\eta}_i(\beta)=\breve{Z}_i(\breve{Y}_{iG}^*-\breve{Z}_i^{\mathrm{T}}\beta)-b_\lambda(\beta),$$

where $b_\lambda(\beta)=(p'_\lambda(|\beta_1|)\mathrm{sgn}(\beta_1),\ldots,p'_\lambda(|\beta_q|)\mathrm{sgn}(\beta_q))^{\mathrm{T}}$, $\mathrm{sgn}(w)$ means the sign function for w, and $p'_\lambda(w)$ is the penalty function proposed by Fan and Li [8], which is defined as follows

$$p'_\lambda(w)=\lambda\{I(w\le\lambda)+\frac{(a\lambda-w)_+}{(a-1)\lambda}I(w>\lambda)\},$$

for some $a>2$, $w>0$ and $p_\lambda(0)=0$.

However, $G(\cdot)$ is usually unknown, and then $\breve{\eta}_i(\beta)$ cannot be used directly to make inference for β. To solve this problem, we replace $G(\cdot)$ in $\breve{\eta}_i(\beta)$ by its estimator. In this paper, we employ the Kaplan-Meier estimator

$$G_n(y)=1-\prod_{i=1}^n(\frac{N^+(Y_i^*)}{1+N^+(Y_i^*)})^{I(Y_i^*\le y,\delta_i=0)},$$

where $N^+(y)=\sum_{j=1}^n I(Y_j^*>y)$. Hence, an estimator of $\breve{\eta}_i(\beta)$ can be defined as

$$\hat{\eta}_i(\beta)=\breve{Z}_i(\breve{Y}_{iG_n}^*-\breve{Z}_i^{\mathrm{T}}\beta)-b_\lambda(\beta), \tag{6}$$

where $\breve{Y}_{iG_n}^*=Y_{iG_n}^*-\hat{g}(U_i)^{\mathrm{T}}X_i$ and $\hat{g}(u)=\sum_{k=1}^n S_k(u)Y_{kG_n}^*$. Then, a penalized empirical log-likelihood ratio function for β can be defined as

$$\hat{R}(\beta)=-2\max\Big\{\sum_{i=1}^n\log(np_i)|p_i\ge 0,\sum_{i=1}^n p_i=1,\sum_{i=1}^n p_i\hat{\eta}_i(\beta)=0\Big\}.$$

We can maximize $\{-\hat{R}(\beta)\}$ to obtain the maximum empirical likelihood estimator $\hat{\beta}_n$. The following Theorem 1 shows that $\hat{\beta}_n$ is $\sqrt{n}$-consistent and satisfies the sparsity. We first give some notations. Let β_0 be the true value of β with $\beta_{0j}\ne 0$ for $j\le d$ and $\beta_{0j}=0$ for $j>d$. The following theorem states the main theoretical results, including the existence of a $\sqrt{n}$-consistent solution and the sparsity of the solution $\hat{\beta}_n$.

Theorem 1 *Suppose that conditions $C1-C6$ in the Appendix hold. Then* *(i) There exists a $\sqrt{n}$-consistent solution to $\hat{R}(\beta)$, that is, $\hat{\beta}_n = \beta_0 + O_p(n^{-1/2})$.*

(ii) $\lim_{n\to\infty} P(\hat{\beta}_{nj} = 0) = 1$, $j = d+1, \ldots, q$.

Remark 1 Theorem 1 indicates that, with probability tending to 1, some components of the maximum empirical likelihood estimator $\hat{\beta}_n$ are set to be zero. Then, the corresponding covariates are removed from the final model. Hence, the penalized empirical likelihood procedure can be used for variable selection.

Next, we show that the penalized empirical likelihood ratio function $\hat{R}(\beta)$ has oracle property. That is, it works as well as the empirical likelihood ratio function that is constructed based on the correct submodel. We give some notations as follows:

$$\begin{aligned}
\Psi(u) &= E(XX^{\mathrm{T}}|U=u), \quad \Phi(u) = E(XZ^{\mathrm{T}}|U=u),\\
H(s) &= \frac{E\{[Z-\mu(U)^{\mathrm{T}}X]Y_G^* I(s<Y^*)\}}{(1-G(s))(1-F(s-))},\\
\mu(u) &= \Psi(u)^{-1}\Phi(u), \quad \Lambda^G(u) = \int_{-\infty}^{u} \frac{1}{1-G(s-)}\mathrm{d}G(s),\\
\Sigma_0(\beta) &= E\{[Z-\mu(U)^{\mathrm{T}}X][Z-\mu(U)^{\mathrm{T}}X]^{\mathrm{T}}[Y_G^* - Z^{\mathrm{T}}\beta - X^{\mathrm{T}}\theta(U)]^2\},\\
\Sigma_1 &= \int_{-\infty}^{\infty} H(s)H(s)^{\mathrm{T}}(1-F(s-))(1-\triangle\Lambda^G(s))\mathrm{d}G(s).
\end{aligned}$$

Furthermore, we let

$$\beta^{(1)} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}, \ \beta^{(2)} = (\beta_{d+1}, \ldots, \beta_q)^{\mathrm{T}}, \text{ and } \tilde{\eta}_i(\beta) = (\hat{\eta}_{i1}(\beta), \ldots, \hat{\eta}_{\mathrm{id}}(\beta))^{\mathrm{T}}.$$

Notice that $\beta_0^{(2)} = 0$, then it is easy to show that $\tilde{\eta}_i(\beta_0) = \tilde{\eta}_i(\beta_0^{(1)})$. Hence, we have

$$\tilde{R}(\beta_0) = \tilde{R}(\beta_0^{(1)}),$$

where $\tilde{R}(\beta_0)$ is the penalized empirical likelihood ratio function constructed by $\tilde{\eta}_i(\beta_0)$, and $\tilde{R}(\beta_0^{(1)})$ is the penalized empirical likelihood ratio function constructed by $\tilde{\eta}_i(\beta_0^{(1)})$. Let

$$\Sigma_0^{(1)}(\beta_0) = (\mathbf{I}_d, \mathbf{0})\Sigma_0(\beta_0)(\mathbf{I}_d, \mathbf{0})^{\mathrm{T}}, \quad \text{and} \quad \Sigma^{(1)}(\beta_0) = (\mathbf{I}_d, \mathbf{0})[\Sigma_0(\beta_0) - \Sigma_1](\mathbf{I}_d, \mathbf{0})^{\mathrm{T}}, \tag{7}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $\mathbf{0}$ is the $d \times (q-d)$ zero matrix. The following theorem presents the oracle property of the penalized empirical likelihood ratio function.

Theorem 2 *Suppose that conditions $C1 - C6$ in the Appendix hold. Then*

$$\tilde{R}(\beta_0^{(1)}) \xrightarrow{\mathcal{L}} w_1\chi_{1,1}^2 + w_2\chi_{1,2}^2 + \cdots + w_d\chi_{1,d}^2 \ ,$$

where $\{w_1, \ldots, w_d\}$ is the eigenvalues of $(\Sigma_0^{(1)}(\beta_0))^{-1}\Sigma^{(1)}(\beta_0)$, and $\chi_{1,1}^2, \ldots, \chi_{1,d}^2$ are independent standard chi-square random variables with 1 degree of freedom.

Remark 2 In fact, it is easy to show that $\tilde{R}(\cdot)$ is the the penalized empirical likelihood ratio function for $\beta_j, j = 1, \ldots, d$. Hence, the confidence region of $\beta_j, j = 1, \ldots, d$, can be constructed based on Theorem 2 if the unknown weights are estimated. We also can give an adjusted-penalized empirical likelihood ratio function as in Yang et al. [7], which can avoid the estimation for unknown weights.

Notice that $\{-\hat{R}(\beta)\}$ is irregular at the origin, then the common method is not applicable. Now, we develop an iterative algorithm based on local quadratic approximation of $p'_\lambda(|\beta_k|)$ as in Fan and Li [8]. More specifically, in a neighborhood of a given non-zero β_{0j}, an approximate of the penalty function at value β_{0j} can be given by $p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{0j}|)+1/2\{p'_\lambda(|\beta_{0j}|)/|\beta_{0j}|\}(\beta_j^2-\beta_{0j}^2)$, $j=1,\ldots,q$. Hence, the derivative of the penalty function can be well approximated by

$$p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{0j}|)/|\beta_{0j}|\}\beta_j. \tag{8}$$

Furthermore, with the similar arguments as in Xue [9], it can be shown that $\hat{\beta}_n$ is the solution of the estimating equation $\sum_{i=1}^n \hat{\eta}_i(\beta)=0$. Let $\Sigma_\lambda(\beta^{(0)})=\mathrm{diag}\{p'_\lambda(|\beta_{01}|)/|\beta_{01}|,\ldots,p'_\lambda(|\beta_{0q}|)/|\beta_{0q}|\}$. Then, from (6) and (8), we have that the solution to maximizing $\{-\tilde{R}(\beta)\}$ can be given by

$$\sum_{i=1}^n \breve{Z}_i(\breve{Y}^*_{iG_n}-\breve{Z}_i^{\mathrm{T}}\beta)-n\Sigma_\lambda(\beta^{(0)})\beta=0. \tag{9}$$

Hence, for a given initial estimator $\beta^{(0)}$, we obtain the following iterative algorithm,

(S1) Initialize $\beta^{(0)}$.

(S2) Set $\beta^{(0)}=\beta^{(k)}$, solve $\beta^{(k+1)}$ by penalty estimating equation (9).

(S3) Iterate step S2 until convergence of β, and denote the final estimator of β as $\hat{\beta}_n$.

In the initialization step, we can obtain an initial estimation of β by using the classic empirical likelihood method without penalty. To implement this method, the tuning parameter a and λ in the penalty function should be chosen. Fan and Li [8] showed that the choice of $a=3.7$ performs well in a variety of situations. Hence, we use their suggestion throughout this paper. Furthermore, similarly to [8], we can estimate λ by minimizing the following generalized cross-validation score

$$\mathrm{GCV}(\lambda)=\frac{\mathrm{RSS}(\lambda)/n}{(1-d(\lambda)/n)^2},$$

where $\mathrm{RRS}(\lambda)=\sum_{i=1}^n\|\breve{Y}^*_{iG_n}-\breve{Z}_i^{\mathrm{T}}\hat{\beta}_\lambda\|^2$ is the residual sum of squares, and $d(\lambda)=\mathrm{tr}\{\breve{Z}[\breve{Z}^{\mathrm{T}}\breve{Z}+n\Sigma_\lambda(\hat{\beta}_\lambda)]^{-1}\breve{Z}^{\mathrm{T}}\}$ is the effective number of parameters, where $\breve{Z}=(\breve{Z}_1,\ldots,\breve{Z}_n)^{\mathrm{T}}$.

3. Simulation study

We evaluate the performance of the proposed variable selection procedure through the following model

$$Y=X\theta(U)+Z^{\mathrm{T}}\beta+\epsilon,$$

where $U\sim U(0,1)$, $X\sim N(0,1)$, and $\theta(u)=0.8u(1-u)$. Furthermore, we take $\beta=(1.2,2.5,0,0,0,0,0.5,0,0,2)^{\mathrm{T}}$, and Z is a 10-dimensional normal distribution with zero mean and identity covariance matrix. Y is generated according to the model with $\epsilon\sim N(0,0.5)$. The censoring variable $C\sim N(\mu,1)$, where $\mu=20.5$, 14 and 10.3, respectively, such that the corresponding censoring rate (CR) is about $0.1, 0.3$ and 0.45. In the simulation, we generated $n=50$, 100 and 150 subjects, respectively. We use the Epanechnikov kernel function $K(u)=0.75(1-u^2)_+$, and the bandwidth h is given by $n^{-1/5}$. The average number of zero coefficients, with 1000 simulation runs, is reported in Table 1, in which the column labeled "C" gives the average number of coefficients of the six true zeros correctly set to zero, and the column labeled "I" gives the average number of the four true nonzeros incorrectly set to zero. Table 1 also presents the average false selection rate (FSR), which is defined as FSR = IN/TN, where "IN" is the average number of the six true zeros incorrectly set to nonzero, and "TN" is the average total number set to nonzero. In fact, FSR represents the proportion of falsely selected unimportant variables among the total variables selected in the variable selection procedure.

	$n=50$			$n=100$			$n=150$		
CR	I	C	FSR	I	C	FSR	I	C	FSR
0.1	0.003	5.578	0.095	0	5.879	0.030	0	5.975	0.006
0.3	0.012	5.519	0.107	0.008	5.854	0.035	0.007	5.953	0.011
0.45	0.030	4.991	0.192	0.016	5.811	0.045	0.008	5.947	0.013

Table 1 Simulation results for variable selection procedure by penalized empirical likelihood method

From Table 1, we can see that, under moderate sample size and censoring rate, the penalized empirical likelihood method has a smaller false selection rate and significantly reduces the model complexity. We also can see that the average false selection rate decreases as the the sample size increases. In addition, for given n, the results of variable selection for all cases of CR are similar. This implies that our adjustment scheme is workable. In general, the proposed variable selection method works well.

Appendix. Proof of Theorems

For convenience and simplicity, let C denote a positive constant which may be different value at each appearance throughout this paper. Before we prove the main theorems, we list some regularity conditions which are used in this paper.

(C1) The bandwidth $h=Cn^{-1/5}$, for some constant $C>0$. The kernel $K(\cdot)$ is a symmetric probability density function, and $\int u^4K(u)\mathrm{d}u<\infty$.

(C2) $\theta(u)$, $\sigma^2(u)$, $\Phi(u)$ and $\Psi(u)$ are twice continuously differentiable on $(0,1)$, where $\sigma^2(u)=E(\epsilon^2|U=u)$.

(C3) The density function of U, says $f(u)$, is bounded away from 0 and infinity on $[0,1]$, and is continuously differentiable on $(0,1)$.

(C4) For $s\le\tau_Q=\inf\{y:Q(y)=1\}$, $G(s)$ and $F(s)$ have no common jumps, where $Q(y)=P(Y^*\le y)$. Moreover, we assume that

$$\int_0^{\tau_Q}\|H(Y)\|^2(1-\Lambda^G(s))\mathrm{d}G(s)<\infty,\quad E\{\frac{\|Z-\mu(U)\||Y|}{[(1-G(Y))(1-F(Y))]^{1/2}}\}<\infty.$$

(C5) For given u, $\Psi(u)$ is positive definite matrix, and $E\{[Z-\mu(U)^{\mathrm{T}}X][Z-\mu(U)^{\mathrm{T}}X]^{\mathrm{T}}\}$ is nonsingular.

(C6) The penalty function $p_\lambda(\cdot)$ satisfies that

(i) $\lim\limits_{n\to\infty}\lambda=0$, and $\lim\limits_{n\to\infty}\sqrt{n}\lambda=\infty$.

(ii) For non-zero fixed w, $\lim_{n\to\infty}\sqrt{n}p'_\lambda(|w|)=0$, and $\lim_{n\to\infty}p''_\lambda(|w|)=0$.

(iii) $\lim_{n\to\infty}\sup_{|w|\le Cn^{-1/2}}p''_\lambda(|w|)=0$, and $\lim_{n\to\infty}\lambda^{-1}\inf_{|w|\le Cn^{-1/2}}p'_\lambda(|w|)>0$, for any $C>0$.

The proofs of theorems rely on the following lemmas.

Lemma 1 *Let $(X_1,\ Y_1),\dots,(X_n,\ Y_n)$ be i.i.d. random vectors, where Y_i is scalar random variables. Assume that $\sup_x\int|y|^s f(x,y)\mathrm{d}y<\infty$ and $E|Y_1|^s<\infty$, where $f(\cdot,\cdot)$ denotes the joint density of $(X,\ Y)$. Let $K(\cdot)$ be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_x\Big|\frac{1}{n}\sum_{i=1}^n\{K_h(X_i-x)Y_i-E[K_h(X_i-x)Y_i]\}\Big|=O_p\Big(\{\frac{\log(1/h)}{nh}\}^{1/2}\Big),$$

provided that $n^{2\delta-1}h\longrightarrow\infty$, for some $\delta<1-s^{-1}$.

Proof This follows immediately from the result that was obtained by Mack and Silverman [10]. □

Lemma 2 *Suppose that conditions $C1-C5$ hold. Then*

$$\sup_{0<u<1}\|\tilde{\mu}(u)-\mu(u)\|=O_p(C_n),\quad \sup_{0<u<1}\|\tilde{\theta}(u)-\theta(u)\|=O_p(C_n),$$

where $C_n=\{\dfrac{\log(1/h)}{nh}\}^{1/2}+h^2$.

Proof Let $S_{nl}=\sum_{i=1}^n X_iX_i^{\mathrm{T}}(\frac{U_i-u}{h})^lK_h(U_i-u),\ l=0,1,2.$ A simple calculation yields

$$E(S_{nl})=nf(u)\Psi(u)\int_0^1 s^lK(s)\mathrm{d}s+o(1),\quad l=0,1,2.$$

Notice that $D_u^{\mathrm{T}}\Omega_uD_u=\begin{pmatrix}S_{n0}&S_{n1}\\S_{nl}&S_{n2}\end{pmatrix}$, by Lemma 1, we obtain

$$D_u^{\mathrm{T}}\Omega_uD_u=nf(u)\Psi(u)\otimes\begin{pmatrix}1&0\\0&\int_0^1u^2K(u)\mathrm{d}u\end{pmatrix}\{1+O_p(C_n)\},\tag{10}$$

uniformly for $u\in(0,1)$, where $\otimes$ is the Kronecker product. Using the same argument, we have

$$D_u^{\mathrm{T}}\Omega_uZ=nf(u)\Phi(u)\otimes(1,0)^{\mathrm{T}}\{1+O_p(C_n)\},\tag{11}$$

uniformly for $u\in(0,1)$. Combining (10) and (11), we have that

$$\tilde{\mu}(u)=(I_p,0_p)(D_u^{\mathrm{T}}\Omega_uD_u)^{-1}D_u^{\mathrm{T}}\Omega_uZ=\mu(u)\{1+O_p(C_n)\},$$

uniformly for $u\in(0,1)$.

Invoking $E(Y_{iG}^*|X_i,Z_i,U_i)=E(Y_i|X_i,Z_i,U_i)$, and with the similar argument, we can prove $\sup_{0<u<1}\|\tilde{\theta}(u)-\theta(u)\|=O_p(C_n)$. This completes the proof of Lemma 2. □

Lemma 3 *Suppose that conditions $C1-C5$ hold. If β_0 is the true parameter, then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\breve{Z}_i(\breve{Y}_{iG_n}^*-\breve{Z}_i^{\mathrm{T}}\beta_0)\xrightarrow{\mathcal{L}}N(0,\Sigma(\beta_0)),$$

where $\Sigma(\beta_0)=\Sigma_0(\beta_0)-\Sigma_1$.

Proof Notice that

$$\breve{Z}_i(\breve{Y}^*_{iG_n}-\breve{Z}_i^{\rm T}\beta_0)=\breve{Z}_i(\breve{Y}^*_{iG}-\breve{Z}_i^{\rm T}\beta_0)+\breve{Z}_i(\breve{Y}^*_{iG_n}-\breve{Y}^*_{iG})\equiv J_{1i}(\beta_0)+J_{2i}(\beta_0). \tag{12}$$

A simple calculation yields

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}J_{1i}(\beta_0)=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Z_i-\mu(U_i)^{\rm T}X_i][Y^*_{iG}-X_i^{\rm T}\theta(U_i)-Z_i^{\rm T}\beta_0]+\\&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Z_i-\mu(U_i)^{\rm T}X_i]X_i^{\rm T}[\tilde{\theta}(U_i)-\theta(U_i)]+\\&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mu(U_i)-\tilde{\mu}(U_i)]^{\rm T}X_i[Y^*_{iG}-X_i^{\rm T}\theta(U_i)-Z_i^{\rm T}\beta_0]+\\&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mu(U_i)-\tilde{\mu}(U_i)]^{\rm T}X_iX_i^{\rm T}[\tilde{\theta}(U_i)-\theta(U_i)]\\\equiv&A_1+A_2+A_3+A_4.\end{aligned}$$

By the Central Limits Theorem, it is easy to prove $A_1\xrightarrow{\mathcal{L}}N(0,\Sigma_0(\beta_0))$. Next, we prove $A_\nu=o_p(1)$, $\nu=2,3,4$. Using Abel inequality, and invoking $E\{(Z_i-\mu(U_i)^{\rm T}X_i)X_i^{\rm T}\}=0$ and Lemma 2, we can prove that

$$\begin{aligned}\|A_2\|\le&\frac{C}{\sqrt{n}}\sup_{1\le i\le n}\left\|\tilde{\theta}(U_i)-\theta(U_i)\right\|\max_{1\le k\le n}\left\|\sum_{i=1}^{k}[Z_i-\mu(U_i)^{\rm T}X_i]X_i^{\rm T}\right\|\\=&\frac{C}{\sqrt{n}}O_p(C_n)O_p(\sqrt{n}\log n).\end{aligned}$$

That is $A_2=o_p(1)$. Invoking $E(Y^*_{iG}|X_i,Z_i,U_i)=E(Y_i|X_i,Z_i,U_i)$, and with a similar argument, we can prove $A_3=o_p(1)$. In addition, by Lemma 2, we have $\|A_4\|\le O_p(\sqrt{n}C_n^2)=o_p(1)$. Hence, we get that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}J_{1i}(\beta_0)\xrightarrow{\mathcal{L}}N(0,\Sigma_0(\beta_0)). \tag{13}$$

By Yang et al. [7], we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}J_{2i}(\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{[Z_i-\mu(U_i)^{\rm T}X_i]Y_i^*\delta_i}{1-G(Y_i^*)}\frac{G_n(Y_i^*)-G(Y_i^*)}{1-G(Y_i^*)}+o_p(1).$$

Then, using the similar argument to Wang and Li [6], we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}J_{2i}(\beta_0)\xrightarrow{\mathcal{L}}N(0,\Sigma_1),$$

and

$$\frac{1}{n}\sum_{i=1}^{n}E\{J_{1i}J_{2i}^{\rm T}\}\xrightarrow{P}-\Sigma_1.$$

Gathering these together with (13), we complete the proof of Lemma 3. □

Proof of Theorem 1 With the similar arguments as in [9], we have that the solution to maximizing $\{-\hat{R}(\beta)\}$ can be given by the following penalty estimating equation

$$\sum_{i=1}^{n} \breve{Z}_i(\breve{Y}^*_{iG_n} - \breve{Z}_i^{\mathrm{T}}\beta) - nb_\lambda(\beta) = 0.$$

Let $U(\beta) = \sum_{i=1}^{n} \breve{Z}_i(\breve{Y}^*_{iG_n} - \breve{Z}_i^{\mathrm{T}}\beta)$, $U^P(\beta) = U(\beta) - nb_\lambda(\beta)$ and $\beta = \beta_0 + n^{-1/2}\gamma$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that $\|\gamma\| = C$ and

$$P\{\min_{\|\beta_0-\beta\|=Cn^{-1/2}} (\beta_0 - \beta)^{\mathrm{T}}\Gamma^{\mathrm{T}}U^P(\beta) > 0\} > 1 - \varepsilon, \tag{14}$$

where $\Gamma = E\{[Z-\mu(U)^{\mathrm{T}}X][Z-\mu(U)^{\mathrm{T}}X]^{\mathrm{T}}\}$. Since Γ is nonsingular, (14) implies, with probability at least $1-\varepsilon$, that there exists a local solution to $U^P(\beta) = 0$ in the ball $\{\beta_0 + n^{-1/2}\gamma : \|\gamma\| \le C\}$. That is, there exists a local solution $\hat{\beta}_n$ of $U^P(\beta) = 0$ with $\hat{\beta}_n = \beta_0 + O_p(n^{-1/2})$.

Invoking Lemma 3, and by the law of large numbers, we can derive that $n^{-1}\sum_{i=1}^{n} \breve{Z}_i\breve{Z}_i^{\mathrm{T}} \xrightarrow{P} \Gamma$. Hence,

$$\begin{aligned}\frac{1}{\sqrt{n}}U^P(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \breve{Z}_i(\breve{Y}^*_{iG_n} - \breve{Z}_i^{\mathrm{T}}\beta_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \breve{Z}_i\breve{Z}_i^{\mathrm{T}}(\beta_0 - \beta) - \sqrt{n}b_\lambda(\beta) \\ &= \frac{1}{\sqrt{n}}U(\beta_0) + \sqrt{n}\Gamma(\beta_0 - \beta) - \sqrt{n}b_\lambda(\beta) + o_p(1). \end{aligned}\tag{15}$$

Notice that $p'_\lambda(0)\mathrm{sgn}(0) = 0$, then condition C6(ii) implies that $\sqrt{n}b_\lambda(\beta_0) \to 0$. Hence, it is easy to show that

$$\frac{1}{\sqrt{n}}U^P(\beta) = \frac{1}{\sqrt{n}}U(\beta_0) + \sqrt{n}\Gamma(\beta_0 - \beta) + \sqrt{n}\{b_\lambda(\beta_0) - b_\lambda(\beta)\} + o_p(1).$$

If $\beta_{0j} \neq 0$, then $\mathrm{sgn}(\beta_{0j}) = \mathrm{sgn}(\beta_j)$. Hence,

$$p'_\lambda(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j}) - p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) = \{p'_\lambda(|\beta_{0j}|) - p'_\lambda(|\beta_j|)\}\mathrm{sgn}(\beta_j).$$

If $\beta_{0j} = 0$, the above equation holds naturally. Then, a simple calculation yields

$$\frac{1}{\sqrt{n}}U^P(\beta) = \frac{1}{\sqrt{n}}U(\beta_0) + \sqrt{n}\Gamma(\beta_0 - \beta) + \sqrt{n}\Lambda_\lambda(\beta^*)(\beta_0 - \beta) + o_p(1),$$

where $\Lambda_\lambda(\beta^*) = \mathrm{diag}\{p''_\lambda(|\beta_1^*|)\mathrm{sgn}(\beta_1), \ldots, p''_\lambda(|\beta_q^*|)\mathrm{sgn}(\beta_q)\}$, and β_j^* lies between β_j and β_{0j}. From Lemma 3, we have $n^{-1/2}U(\beta_0) = O_p(1)$. Hence, we get that

$$\begin{aligned}\frac{1}{\sqrt{n}}(\beta_0 - \beta)^{\mathrm{T}}\Gamma^{\mathrm{T}}U^P(\beta) =& O_p(\|\beta_0 - \beta\|) + \sqrt{n}(\beta_0 - \beta)^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma(\beta_0 - \beta) + \\ & \sqrt{n}(\beta_0 - \beta)^{\mathrm{T}}\Gamma^{\mathrm{T}}\Lambda_\lambda(\beta^*)(\beta_0 - \beta).\end{aligned}$$

Notice that Γ is nonsingular, the second term on the right-hand side is larger than $a_0C^2n^{-1/2}$, where a_0 is the smallest eigenvalue of $\Gamma^{\mathrm{T}}\Gamma$. The first term is of order $O_p(n^{-1/2})$. By condition C6(iii), we have $\max_j p''_\lambda(|\beta_j^*|) \to 0$, so the third term is dominated by the second term. Therefore, for any $\varepsilon > 0$, if we choose C large enough so that, for large n, the probability that the absolute value of the first term is larger than the second term is less than ε, then (14) holds. This completes the proof of part (i).

For part (ii), it suffices to show that for any $\varepsilon > 0$, when n is large enough, $P(C_{nj}) < \varepsilon$, where $C_{nj} = \{\hat{\beta}_{nj} \neq 0\}$, $j = d+1, \ldots, q$. Since $\hat{\beta}_{nj} = O_p(n^{-1/2})$, when n is large enough, there exists some C such that

$$P(C_{nj}) < \varepsilon/2 + P\{\hat{\beta}_{nj} \neq 0, |\hat{\beta}_{nj}| < Cn^{-1/2}\}. \tag{16}$$

Using the jth component of (15), we can obtain that

$$\sqrt{n}p'_{\lambda_n}(|\hat{\beta}_{nj}|)\mathrm{sgn}(\hat{\beta}_{nj}) = \frac{1}{\sqrt{n}}U_j(\beta_0) + \sqrt{n}\Gamma_j(\hat{\beta}_n - \beta_0) + O_p(1).$$

The first two terms on the right-hand side are of order $O_p(1)$. Hence, for large n, there exists some C such that

$$P(\sqrt{n}p'_{\lambda}(|\hat{\beta}_{nj}|) > C) < \varepsilon/2. \tag{17}$$

By condition C6, we have that

$$\inf_{|\beta_{nj}| \leq Cn^{-1/2}} \sqrt{n}p'_{\lambda}(|\beta_{nj}|) = \sqrt{n}\lambda \inf_{|\beta_{nj}| \leq Cn^{-1/2}} \lambda^{-1}p'_{\lambda}(|\beta_{nj}|) \to \infty.$$

That is, $\hat{\beta}_{nj} \neq 0$ and $|\hat{\beta}_{nj}| < Cn^{-1/2}$ imply that $\sqrt{n}p'_{\lambda}(|\hat{\beta}_{nj}|) > C$ for large n. Then, invoking (16) and (17), we have that

$$P(C_{nj}) < \varepsilon/2 + P(\sqrt{n}p'_{\lambda}(|\hat{\beta}_{nj}|) > C) < \varepsilon.$$

This completes the proof of this theorem. □

Proof of Theorem 2 Note that $b_{\lambda}(\beta_0) = o(n^{-1/2})$, then from the proof of Lemma 3, it is easy to show that

$$\max_{1 \leq i \leq n} \|\hat{\eta}_i(\beta_0)\| = o_p(n^{1/2}).$$

Furthermore, by $\beta_0^{(2)} = 0$, it is easy to show that

$$\tilde{\eta}_i(\beta_0^{(1)}) = \tilde{\eta}_i(\beta_0) = (\mathbf{I}_d, \mathbf{0})\hat{\eta}_i(\beta_0), \tag{18}$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $\mathbf{0}$ is the $d \times (q-d)$ zero matrix. Then we have

$$\max_{1 \leq i \leq n} \|\tilde{\eta}_i(\beta_0^{(1)})\| = \max_{1 \leq i \leq n} \|(\mathbf{I}_d, \mathbf{0})\hat{\eta}_i(\beta_0)\| = o_p(n^{1/2}). \tag{19}$$

By the Lagrange multiplier method, $\tilde{R}(\beta_0^{(1)})$ can be represented as

$$\tilde{R}(\beta_0^{(1)}) = 2\sum_{i=1}^{n} \log\{1 + \delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})\}, \tag{20}$$

where δ is a $d \times 1$ vector given as the solution of

$$\sum_{i=1}^{n} \frac{\tilde{\eta}_i(\beta_0^{(1)})}{1 + \delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})} = 0. \tag{21}$$

Invoking (19), and using the arguments similar to Owen [11], we can obtain $\|\delta\| = O_p(n^{-1/2})$. Taking this together with (19) and applying the Taylor expansion to (20), we get that

$$\tilde{R}(\beta_0^{(1)}) = 2\sum_{i=1}^{n}\{\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)}) - [\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})]^2/2\} + o_p(1). \tag{22}$$

By (21), it follows that

$$0=\sum_{i=1}^{n}\frac{\tilde{\eta}_i(\beta_0^{(1)})}{1+\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})}$$
$$=\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})-\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\tilde{\eta}_i^{\mathrm{T}}(\beta_0^{(1)})\delta+\sum_{i=1}^{n}\frac{\tilde{\eta}_i(\beta_0^{(1)})[\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})]^2}{1+\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})}. \quad (23)$$

Then, it is easy to show that

$$\sum_{i=1}^{n}[\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})]^2=\sum_{i=1}^{n}\delta^{\mathrm{T}}\tilde{\eta}_i(\beta_0^{(1)})+o_p(1), \quad (24)$$

$$\delta=\Big\{\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\tilde{\eta}_i^{\mathrm{T}}(\beta_0^{(1)})\Big\}^{-1}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})+o_p(n^{-1/2}). \quad (25)$$

Invoking (22)–(25), by some algebra calculations, we have

$$\tilde{R}(\beta_0^{(1)})=\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\Big\}^{\mathrm{T}}(\hat{\Sigma}_0^{(1)}(\beta_0^{(1)}))^{-1}\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\Big\},$$

where $\hat{\Sigma}_0^{(1)}(\beta_0^{(1)})=n^{-1}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\tilde{\eta}_i^{\mathrm{T}}(\beta_0^{(1)})$. Invoking (18) and the proof of Lemma 3, we can obtain that $\hat{\Sigma}_0^{(1)}(\beta_0^{(1)})\xrightarrow{P}\Sigma_0^{(1)}(\beta_0)$, where $\Sigma_0^{(1)}(\beta_0)$ is defined by (7). Hence, we get

$$\tilde{R}(\beta_0^{(1)})=\Big([\Sigma^{(1)}(\beta_0)]^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\Big)^{\mathrm{T}}\Gamma(\beta_0)\Big([\Sigma^{(1)}(\beta_0)]^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0)\Big)+o_p(1),$$

where $\Gamma(\beta_0)=[\Sigma^{(1)}(\beta_0)]^{\frac{1}{2}}(\Sigma_0^{(1)}(\beta_0))^{-1}[\Sigma^{(1)}(\beta_0)]^{\frac{1}{2}}$. Let $D=\mathrm{diag}(w_1,\dots,w_d)$, where $w_1,\dots,w_q$ are the eigenvalues of $(\Sigma_0^{(1)}(\beta_0))^{-1}\Sigma^{(1)}(\beta_0)$. Notice that $(\Sigma_0^{(1)}(\beta_0))^{-1}\Sigma^{(1)}(\beta_0)$ has the same eigenvalues as $\Gamma(\beta_0)$. Hence, there exists orthogonal matrix Q such that $Q^{\mathrm{T}}DQ=\Gamma(\beta_0)$. Then, with a simple calculation we get that

$$\tilde{R}(\beta_0^{(1)})=\Big(Q[\Sigma^{(1)}(\beta_0)]^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\Big)^{\mathrm{T}}D\Big(Q[\Sigma^{(1)}(\beta_0)]^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\Big)+o_p(1). \quad (26)$$

Furthermore, invoking $b_\lambda(\beta_0)=o(n^{-1/2})$, $\tilde{\eta}_i(\beta_0^{(1)})=(\mathbf{I}_d,\mathbf{0})\hat{\eta}_i(\beta_0)$, and Lemma 3, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\eta}_i(\beta_0^{(1)})\xrightarrow{\mathcal{L}}N(0,\Sigma^{(1)}(\beta_0)).$$

Then, notice that Q is an orthogonal matrix, and by (26), the proof of Theorem 2 is completed. □

References

[1] Qi LI, C. J. HUANG, Dong LI, et al. *Semiparametric smooth coefficient models.* J. Bus. Econom. Statist., 2002, **20**(3): 412–422.

[2] Jianqing FAN, Tao HUANG. *Profile likelihood inferences on semiparametric varying-coefficient partially linear models.* Bernoulli, 2005, **11**(6): 1031–1057.

[3] Jinhong YOU, Yong ZHOU. *Empirical likelihood for semiparametric varying-coefficient partially linear regression models.* Statist. Probab. Lett., 2006, **76**(4): 412–422.

[4] Runze LI, Hua LIANG. *Variable selection in semiparametric regression modeling.* Ann. Statist., 2008, **36**(1): 261–286.

[5] Peixin ZHAO, Liugen XUE. *Variable selection for semiparametric varying coefficient partially linear models.* Statist. Probab. Lett., 2009, **79**(20): 2148–2157.

[6] Qihua WANG, Gang LI. *Empirical likelihood semiparametric regression analysis under random censorship.* J. Multivariate Anal., 2002, **83**(2): 469–486.

[7] Yiping YANG, Liugen XUE, Weihu CHENG. *An empirical likelihood method in a partially linear single-index model with right censored data.* Acta Math. Sin. (Engl. Ser.), 2012, **28**(5): 1041–1060.

[8] Jianqing FAN, Runze LI. *Variable selection via nonconcave penalized likelihood and its oracle properties.* J. Amer. Statist. Assoc., 2001, **96**(456): 1348–1360.

[9] Liugen XUE. *Empirical likelihood local polynomial regression analysis of clustered data.* Scand. J. Stat., 2010, **37**(4): 644–663.

[10] Y. P. MACK, B. W. SILVERMAN. *Weak and Strong uniform consistency of kernel regression estimates.* Z. Wahrsch. Verw. Gebiete, 1982, **61**(3): 4050–415.

[11] A. B. OWEN. *Empirical likelihood ratio confidence regions.* Ann. Statist., 1990, **18**(1): 90–120.