Journal of Mathematical Research & Exposition
Mar., 2011, Vol. 31, No. 2, pp. 251–260
DOI:10.3770/j.issn:1000-341X.2011.02.008
Http://jmre.dlut.edu.cn

Variable Selection for Varying-Coefficient Models with Missing Response at Random

Pei Xin ZHAO[1,2,*], **Liu Gen XUE**[1]

1. *College of Applied Sciences, Beijing University of Technology, Beijing* 100124*, P. R. China;*
2. *Department of Mathematics, Hechi University, Guangxi* 546300*, P. R. China*

Abstract In this paper, we present a variable selection procedure by combining basis function approximations with penalized estimating equations for varying-coefficient models with missing response at random. With appropriate selection of the tuning parameters, we establish the consistency of the variable selection procedure and the optimal convergence rate of the regularized estimators. A simulation study is undertaken to assess the finite sample performance of the proposed variable selection procedure.

Keywords varying-coefficient model; variable selection; missing data.

Document code A
MR(2010) Subject Classification 62G08; 62G20
Chinese Library Classification O212.7

1. Introduction

Consider the following varying-coefficient model

$$Y = X^{\mathrm{T}}\theta(U) + \epsilon, \tag{1}$$

where $\theta(u) = (\theta_1(u), \ldots, \theta_p(u))^{\mathrm{T}}$ is a $p \times 1$ vector of unknown functions, X and U are covariates, Y is the response variable, and ϵ is the model error with $E(\epsilon|X, U) = 0$. In addition, we assume that the variable U ranges over a nondegenerate compact interval, without loss of generality, that is assumed to be the unit interval $[0, 1]$.

When the dimension of the covariate X is small, many methods have been developed for estimating the coefficients in model (1) and its extensions [1–5]. However, when the dimension of the covariate in model (1) is large, an important problem is to select the important variables in such models. Traditional methods for variable selection, such as stepwise deletion method, ignore stochastic errors inherited in the stages of variable selections. Hence, the accuracy of the regression coefficient estimators in the final model is somewhat difficult to understand. To avoid this drawback, for linear models, Fan and Li [6] proposed a family of variable selection procedures,

Received April 15, 2009; Accepted October 14, 2009
Supported by the National Natural Science Foundation of China (Grant No. 10871013), the Natural Science Foundation of Beijing (Grant No. 1072004), the Natural Science Foundation of Guangxi (Grant No. 2010GXNSFB013051) and the Graduate Student Foundation of Hechi University (Grant No. 2008QS-N014).
* Corresponding author
E-mail address: zpx81@163.com (P. X. ZHAO)

called SCAD, via a nonconcave penalized likelihood function. This procedure can simultaneously estimate regression coefficients and shrink some coefficients to zero, thereby, removing them from the final model. Li and Liang [7] adopted this methodology to select important variables in the parametric components of semiparametric regression modeling. Wang et al. [8] considered the variable selection for model (1) with the group SCAD penalty, which is an extension of the SCAD penalty. Wang et al. [9] adopted the group SCAD penalty to select important variables of varying coefficient models with longitudinal data. An essential assumption in their papers is that all data can be observed, but missing data frequently occurs in many applications such as health and social science studies. For such models with missing data, the variable selection procedures proposed by [8, 9] cannot be used directly any more. Recently, some work has been done on the inferences of missing data. Wang and Sun [10] proposed an estimation method for regression coefficients of partially linear models with missing responses at random based on inverse marginal probability weighted method. Sun et al. [11] considered the model checking for partially linear models with missing responses at random. Zhou et al. [12] considered the inferences of missing data by imputation-based estimating equations. However, the variable selection for such varying coefficient models with missing data seems to be missing. Taking this issue into account, we propose a variable selection procedure for the varying-coefficient models with missing response at random based on basis function approximations and penalized estimating equations. Furthermore, with proper choice of regularization parameters, we show that this variable selection procedure is consistent, and the estimators of the coefficient functions achieve the convergence rate as if the subset of true zero coefficients were already known.

The rest of this paper is organized as follows. In Section 2, we first propose the variable selection procedure. Then, we present theoretical properties of our variable selection procedure including the consistency of the variable selections and the convergence rates of the regularized estimators. In Section 3, based on local quadratic approximations, we propose an iterative algorithm for finding regularized estimators. In Section 4, some simulations are carried out to assess the performance of the proposed methods. The technical proofs of all asymptotic results are provided in the Appendix.

2. Variable selection via penalized estimating equations

Suppose that we have an incomplete random sample $(Y_i, \delta_i, X_i, U_i)$, $i = 1, \ldots, n$, from model (1), where X_i and U_i are observed, and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. In this paper, we assume that Y is missing at random (MAR). That is, $P(\delta = 1|Y, X, U) = P(\delta = 1|X, U)$. Thus,

$$\delta_i Y_i = \delta_i X_i^{\mathrm{T}} \theta(U_i) + \delta_i \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

Let $B(u) = (B_1(u), \ldots, B_L(u))^{\mathrm{T}}$ be B-spline basis functions with the order of M, where $L = K + M + 1$, and K is the number of interior knots. Then, $\theta_k(u)$ can be approximated by $\theta_k(u) \approx B(u)^{\mathrm{T}} \beta_k$, $k = 1, \ldots, p$. Substituting this into model (2), we can get

$$\delta_i Y_i = \delta_i W_i^{\mathrm{T}} \beta + \delta_i \epsilon_i, \tag{3}$$

where $W_i = I_p \otimes B(U_i) \cdot X_i$ and $\beta = (\beta_1^{\mathrm{T}}, \ldots, \beta_p^{\mathrm{T}})^{\mathrm{T}}$. Model (3) is a standard linear regression model. Let $\psi_i(\beta) = W_i(Y_i - W_i^{\mathrm{T}}\beta)$, $H = \int_0^1 B(u)B(u)^{\mathrm{T}}\mathrm{d}u$, and $\|\beta_k\|_H = (\beta_k^{\mathrm{T}} H \beta_k)^{1/2}$. Furthermore, let $p_\lambda(\cdot)$ be a penalty function with λ as a tuning parameter. Then, it is easy to show that

$$\frac{\partial p_\lambda(\|\beta_k\|_H)}{\partial \beta_k} = p'_\lambda(\|\beta_k\|_H)\frac{H\beta_k}{\|\beta_k\|_H}.$$

Invoking this, we note that $E\{\psi_i(\beta)\} = o(1)$, and each function $\theta_k(u)$ in model (2) is characterized by β_k in model (3). This motivates us to adopt the following penalized estimating equations, that is

$$\eta(\beta) = \sum_{i=1}^{n} \delta_i \psi_i(\beta) - n\mathbf{b}(\beta), \tag{4}$$

where

$$\mathbf{b}(\beta) = \Big(p'_{\lambda_1}(\|\beta_1\|_H)\frac{\beta_1^{\mathrm{T}} H}{\|\beta_1\|_H}, \ldots, p'_{\lambda_p}(\|\beta_p\|_H)\frac{\beta_p^{\mathrm{T}} H}{\|\beta_p\|_H}\Big)^{\mathrm{T}}$$

is a $pL \times 1$ penalized vector. There are many ways to specify the penalty function $p_\lambda(\cdot)$. Throughout this paper, we take $p_\lambda(\cdot)$ as the SCAD penalty function, proposed by Fan and Li [6], which is defined as

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

with $a > 2$, $w > 0$ and $p_\lambda(0) = 0$. Here, the tuning parameter λ is not necessarily the same for all $\theta_k(\cdot)$. Hence, we denote λ, that corresponds to $\theta_k(\cdot)$, as λ_k in $\mathbf{b}(\beta)$, respectively. However, the subjects with missing data are discarded and only the complete data are used in $\eta(\beta)$. To improve efficiency, we adopt the strategy of Zhou et al. [12], and propose an imputation-based penalized estimating equation as

$$\tilde{\eta}(\beta) = \sum_{i=1}^{n} \tilde{\psi}_i(\beta) - n\mathbf{b}(\beta), \tag{5}$$

where $\tilde{\psi}_i(\beta) = \delta_i\psi_i(\beta) + (1-\delta_i)\tilde{m}_\psi(U_i, \beta)$, and $\tilde{m}_\psi(u, \beta)$ is the kernel estimator of $E\{\psi_i(\beta)|U_i = u\}$, which can be defined as

$$\tilde{m}_\psi(u, \beta) = \sum_{i=1}^{n} \delta_i \omega_{ni}(u)\psi_i(\beta),$$

where $\omega_{ni}(u) = K_h(u - U_i)/\sum_{j=1}^{n} \delta_j K_h(u - U_j)$, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function, and h is a band width.

Let $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \ldots, \hat{\beta}_p^{\mathrm{T}})^{\mathrm{T}}$ be the solution of $\tilde{\eta}(\beta) = 0$. Then, the estimator of $\theta_k(u)$ can be obtained by $\hat{\theta}_k(u) = B(u)^{\mathrm{T}}\hat{\beta}_k$. Next, we study the asymptotic properties of the resulting penalized estimators. Let $\theta_0(\cdot)$ be the true value of $\theta(\cdot)$, and corresonding true value of β is denoted by β_0. Without loss of generality, we assume that $\theta_{k0}(\cdot) \equiv 0$, $k = d+1, \ldots, p$, and $\theta_{k0}(\cdot)$, $k = 1, \ldots, d$ are all nonzero components of $\theta_0(\cdot)$. The following theorem gives the consistency of the penalized estimators.

Theorem 1 *Suppose that the regularity conditions C1–C6 in the Appendix hold and the number*

of knots $K = O_p(n^{1/(2r+1)})$. *Let* $a_n = \max_k\{|p'_{\lambda_k}(\|\beta_{k0}\|_H)| : \beta_{k0} \neq 0\}$. *Then,*

$$\|\hat{\theta}_k(\cdot) - \theta_{k0}(\cdot)\| = O_p(n^{\frac{-r}{2r+1}} + a_n), \quad k = 1, \ldots, p,$$

where r *is defined in condition* $C1$ *in the Appendix.*

By Remark 1 in [6], we have that, if $\max\{\lambda_k\} \to 0$ as $n \to \infty$, then $a_n = 0$. Hence, Theorem 1 implies that, by choosing proper tuning parameters, the estimators of coefficient functions achieve the optimal convergence rate as if the subset of true zero coefficients were already known [13]. Furthermore, under some conditions, we show that such consistent estimators must possess the sparsity property, which is stated as follows

Theorem 2 *Suppose that the regularity conditions C1–C7 in the Appendix hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Let* $\lambda_{max} = \max\{\lambda_k : k = 1, \ldots, p\}$, *and* $\lambda_{\min} = \min\{\lambda_k : k = 1, \ldots, p\}$. *If* $\lambda_{max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{\min} \to \infty$, *as* $n \to \infty$. *Then, with probability tending to 1,* $\hat{\theta}(\cdot)$ *must satisfy* $\hat{\theta}_k(\cdot) \equiv 0$, $k = d+1, \ldots, p$.

Remark 1 Invoking Theorems 1 and 2, it is clear that by choosing proper tuning parameters, our variable selection procedure is consistent, and the estimators of coefficient functions achieve the optimal convergence rate as if the subset of true zero coefficients were already known.

3. Algorithm

Because $\tilde{\eta}(\beta)$ is irregular at the origin, the common method is not applicable to solve equation (5). Now, we develop an iterative algorithm based on local quadratic approximation of the penalty function $p_\lambda(\cdot)$ as in [6]. More specifically, in a neighborhood of a given non-zero w_0, an approximation of the penalty function at value w_0 can be given by

$$p_\lambda(w) \approx p_\lambda(w_0) + \frac{1}{2}\frac{p'_\lambda(w_0)}{w_0}(w^2 - w_0^2).$$

Hence, for the given initial value $\beta_k^{(0)}$ with $\|\beta_k^{(0)}\|_H > 0$, $k = 1, \ldots, p$, we have

$$p_{\lambda_k}(\|\beta_k\|_H) \approx p_{\lambda_k}(\|\beta_k^{(0)}\|_H) + \frac{1}{2}\frac{p'_{\lambda_k}(\|\beta_k^{(0)}\|_H)}{\|\beta_k^{(0)}\|_H}(\beta_k^{\mathrm{T}} H \beta_k - \beta_k^{(0)\mathrm{T}} H \beta_k^{(0)}).$$

Then,

$$\frac{\partial p_{\lambda_k}(\|\beta_k\|_H)}{\partial \beta_k} \approx \frac{p'_{\lambda_k}(\|\beta_k^{(0)}\|_H)}{\|\beta_k^{(0)}\|_H} H\beta_k. \tag{6}$$

By combining (5) with (6), it is easy to show that

$$\tilde{\eta}(\beta) \approx \sum_{i=1}^{n} \tilde{\psi}_i(\beta) - n\Sigma(\beta^{(0)})\beta, \tag{7}$$

where $\Sigma(\beta^{(0)}) = \mathrm{diag}\{\frac{p'_{\lambda_1}(\|\beta_1^{(0)}\|_H)}{\|\beta_1^{(0)}\|_H} H, \ldots, \frac{p'_{\lambda_p}(\|\beta_p^{(0)}\|_H)}{\|\beta_p^{(0)}\|_H} H\}$. Based on (7) and the definition of $\tilde{m}_\psi(u, \beta)$, let $\tilde{\eta}(\beta) = 0$. Then a simple calculation yields

$$\Big\{\sum_{i=1}^{n}(\delta_i W_i W_i^{\mathrm{T}} + (1-\delta_i)\tilde{\mu}(U_i)) + n\Sigma(\beta^{(0)})\Big\}\beta = \sum_{i=1}^{n}(\delta_i W_i Y_i + (1-\delta_i)\tilde{g}(U_i)), \tag{8}$$

where $\tilde{\mu}(u) = \sum_{j=1}^{n} \delta_j \omega_{nj}(u) W_j W_j^{\mathrm{T}}$, and $\tilde{g}(u) = \sum_{j=1}^{n} \delta_j \omega_{nj}(u) W_j Y_j$. Hence, we obtain the following iterative algorithm

Step 1. Initialize $\beta^{(0)}$.

Step 2. Set $\beta^{(0)} = \beta^{(k)}$, solve $\beta^{(k+1)}$ by equation (8).

Step 3. Iterate Step 2 until convergence, and denote the final estimator of β as $\hat{\beta}$.

In the initialization step, we obtain an initial estimator of β by using traditional imputation-based estimating equations based on the first term of the right side in (5). To implement this method, the number of interior knots K, the tuning parameters a and λ_k's in the penalty functions, and the bandwidth h in the kernel estimator $\tilde{m}_{\psi}(u, \beta)$ should be chosen. As in [12], we take $h = Cn^{-1/3}$ in this paper, where C is the sample deviation of U. In addition, Fan and Li [6] showed that the choice of $a = 3.7$ performs well in a variety of situations. Hence, we use their suggestion throughout this paper. Furthermore, we can use the similar method to choose the tuning parameters as in [8, 9]. More specifically, we can estimate λ_k's and K by minimizing the following cross-validation score

$$CV(K, \lambda_1, \ldots, \lambda_p) = \sum_{i=1}^{n} \{Y_i - W_i^{\mathrm{T}} \hat{\beta}^{(-i)}\}^2, \tag{9}$$

where $\hat{\beta}^{(-i)}$ is the solution of $\tilde{\eta}(\beta) = 0$ based on (5) after deleting the ith subject. The minimization problem over a $p+1$-dimensional space is very difficult. However, it is expected that the choice of λ_k should satisfy that the tuning parameter for zero coefficient is larger than that for nonzero coefficient. Thus we can simultaneously unbiasedly estimate large coefficients, and shrink the small coefficients toward to zero. Hence, in practice, we suggest taking $\lambda_{_k} = \lambda / \|\hat{\beta}_k^{(0)}\|_H$, where $\hat{\beta}_k^{(0)}$ is initial estimators of β_k. Then (9) will reduce to the following two-dimensional minimization problem

$$CV(K, \lambda) = \sum_{i=1}^{n} \{Y_i - W_i^{\mathrm{T}} \hat{\beta}^{(-i)}\}^2. \tag{10}$$

In fact, such a choice of tuning parameters, in some sense, is the same rationale behind adaptive lasso [14], and from our simulation experience, we found that this method works well.

4. Simulation study

In this section, we conduct some simulations to evaluate finite sample performance of the proposed method. We simulate data from model (1), where $\theta(u) = (\theta_1(u), \ldots, \theta_{10}(u))^{\mathrm{T}}$ with $\theta_1(u) = 5.5 + 0.5\exp(2u-1)$, $\theta_2(u) = 2 - \sin(2\pi u)$ and $\theta_3(u) = 0.5 + u(2-u)$. While the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. To perform this simulation, we take the covariates $X_k \sim N(0, 1.5)$, $k = 1, \ldots, 10$, and $U \sim U(0, 1)$. Y is generated according to the model with $\epsilon \sim N(0, 0.5)$. Furthermore, we consider the following functions of selection probability $\Delta(u) = P(\delta = 1 | U = u)$: 1) $\Delta(u) = 0.9 + 0.2(u - 0.5)$; 2) $\Delta(u) = \frac{e^u}{0.7 + e^u}$; 3) $\Delta(u) = 0.5$ for all u. Then, the missing rates corresponding to the three scenarios are approximately 0.1, 0.3 and 0.5, respectively.

	$1-E(\Delta(U))=0.1$			$1-E(\Delta(U))=0.3$			$1-E(\Delta(U))=0.5$		
Methods	C	I	RASE	C	I	RASE	C	I	RASE
gSCAD	6.382	0	0.015	6.036	0.002	0.028	6.244	0.013	0.079
cSCAD	6.405	0	0.015	5.041	0.009	0.074	1.869	0.011	0.115
fSCAD	6.677	0	0.014	6.672	0	0.014	6.672	0	0.014
NE	0	0	0.412	0	0	0.539	0	0	0.798

Table 1 Simulation results of variable selections with different variable selection procedures

We compare the performance of the variable selection procedure proposed in this paper, called gSCAD, based on imputed estimating equations with that based on the full data set (i.e., no missing data), called fSCAD, and the complete data set (i.e., ignoring the missing data), called nSCAD. For comparison, we also consider the naive estimation procedure (NE) for β in our simulation. That is, we delete the penalty in the objective function, and obtain the estimators of β by using traditional imputation-based estimating equations based on the first term of the right side in (5). In this simulation, we generate $n = 200$ subjects, and use the cubic B-splines. The number of interior knots K and the tuning parameter λ are obtained by (10). The average number of the estimated zero coefficients for coefficient functions, with 1000 simulation runs, is reported in Table 1, in which the column labeled "C" gives the average number of the true zero coefficients correctly set to zero, and the column labeled "I" gives the average number of the true nonzero coefficients incorrectly set to zero. Furthermore, Table 1 also presents the median of RASE over the 1000 simulations. Here RASE is the square root of average square errors, which is defined as

$$\text{RASE} = \Big\{\frac{1}{N}\sum_{s=1}^{N}\sum_{k=1}^{p}[\hat{\theta}_k(u_s)-\theta_k(u_s)]^2\Big\}^{1/2},$$

where u_s, $s = 1,\ldots,N$ are the grid points at which the function $\hat{\theta}(u)$ is evaluated. In our simulation, $N = 200$ is used. From Table 1, we can make the following observations:

(i) The performances of both gSCAD and cSCAD procedures become better in terms of model error and model complexity as the missing rate decreases. Furthermore, when the missing rate is large, the performance of gSCAD is significantly better than that of cSCAD. This implies that our imputation scheme is workable.

(ii) As expected, the performance of fSCAD is the best in all the cases for selecting significant variables. In addition, the performance of gSCAD becomes closer and closer to that based on fSCAD as the missing rate decreases.

(iii) The naive estimation procedure performs the worst in terms of model complexity and model error. This estimation procedure cannot eliminate some unimportant variables. Then, this procedure results in largest model errors. Hence, our penalized estimation procedure outperforms the naive estimation procedure.

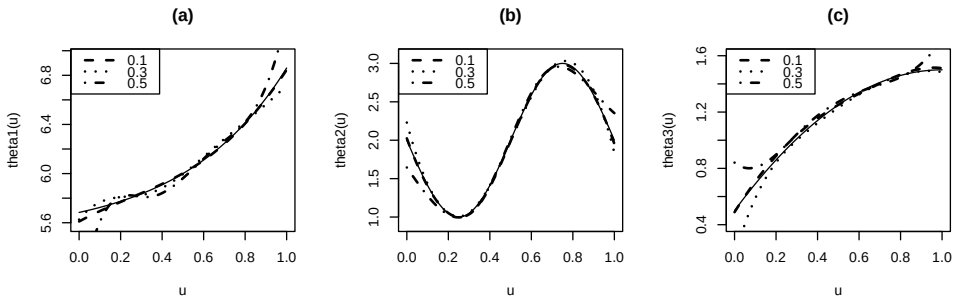

Figure 1 The averages of estimated coefficients based on gSCAD procedure over 1000 replications

Figure 1 shows the estimators of coefficients $\theta_1(u)$ (a), $\theta_2(u)$ (b)and $\theta_3(u)$ (c) based on the gSCAD method proposed in this paper, where the solid curve represents the real curve of $\theta(u)$. From Figure 1, it is clear that the estimators fit the true functions very well in all cases, and do not depend sensitively on the choice of the selection probability function $\Delta(u)$. This implies that the imputation scheme is workable.

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 our main theorems, we list some regularity conditions which are used in this paper.

C1. $\theta(u)$ is rth continuously differentiable on $(0,1)$, where $r > 1/2$.

C2. The density function of U, say $f(u)$, is bounded away from 0 and infinity on $[0,1]$. Furthermore, we assume that $f(u)$ is continuously differentiable on $(0,1)$.

C3. Let $\sigma^2(u) = E\{\epsilon^2|U=u\}$, and $G(u) = E\{XX^T|U=u\}$. Then, $\sigma^2(u)$ and $G(u)$ are continuous with respect to u. Furthermore, for given u, $G(u)$ is a positive definite matrix, and the eigenvalues of $G(u)$ are bounded.

C4. Let $c_1,\ldots,c_K$ be the interior knots of $[0,1]$. Furthermore, let $c_0 = 0$, $c_{K+1} = 1$, $h_i = c_i - c_{i-1}$. Then, there exists a constant C_0 such that

$$\frac{\max\{h_i\}}{\min\{h_i\}} \le C_0, \quad \max\{|h_{i+1}-h_i|\} = o\left(K^{-1}\right).$$

C5. The bandwidth in $\tilde{m}_\psi(u,\beta)$ satisfies $nh^2 \to \infty$, and $nh^4 \to 0$ as $n \to \infty$.

C6. For any given nonzero w, we have $\lim_{n\to\infty} n^{\frac{r}{2r+1}} p'_\lambda(|w|) = 0$.

C7. $\lim_{n\to\infty} \inf_{|w|\le Cn^{-r/(2r+1)}} \lambda^{-1} p'_\lambda(|w|) > 0$, and $\lim_{n\to\infty} \sup_{|w|\le Cn^{-r/(2r+1)}} p''_\lambda(|w|) = 0$.

These conditions are commonly adopted in the nonparametric literature and variable selection methodology. Conditions C1–C3 are similar to those used in [1–5]. Condition C4 implies that $c_0,\ldots,c_{K+1}$ is a C_0-quasi-uniform sequence of partitions of $[0,1]$ (see [15, p.216]). Condition C5 is common in the nonparametric literature. Condition C6 and C7 are assumptions on the penalty function, which were used in [16] and [17]. Furthermore these conditions on the penalty function are similar to that used in [6, 7].

Proof of Theorem 1 Let $\tau = n^{-r/(2r+1)} + a_n$, and $\beta = \beta_0 + \tau T$. We first show that, for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\{\inf_{\|T\|=C}(\beta_0 - \beta)^{\mathrm{T}}\tilde{\eta}(\beta) > 0\} \geq 1 - \varepsilon. \tag{11}$$

Let $\Delta(\beta) = K^{-1}\left(\beta_0^{\mathrm{T}} - \beta^{\mathrm{T}}\right)\tilde{\eta}(\beta)$. Then a simple calculation yields

$$\begin{aligned}\Delta(\beta) &= \frac{-\tau}{K}\sum_{i=1}^{n}\delta_i T^{\mathrm{T}}\psi_i(\beta) + \frac{-\tau}{K}\sum_{i=1}^{n}(1-\delta_i)T^{\mathrm{T}}\tilde{m}_{\psi}(U_i,\beta) + \frac{\tau n}{K}T^{\mathrm{T}}\mathbf{b}(\beta)\\ &\equiv I_1 + I_2 + I_3.\end{aligned}$$

Hence, invoking the definition of $\psi_i(\beta)$, we have that

$$\begin{aligned}I_1 &= -\frac{\tau}{K}\sum_{i=1}^{n}\delta_i W_i^{\mathrm{T}}TX_i^{\mathrm{T}}R(U_i) - \frac{\tau}{K}\sum_{i=1}^{n}\delta_i W_i^{\mathrm{T}}T\epsilon_i + \frac{\tau^2}{K}\sum_{i=1}^{n}\delta_i(W_i^{\mathrm{T}}T)^2\\ &\equiv I_{11} + I_{12} + I_{13},\end{aligned}$$

where $R(u) = (R_1(u), \ldots, R_p(u))^{\mathrm{T}}$, and $R_k(u) = \theta_{k0}(u) - B(u)^{\mathrm{T}}\beta_{k0}$, $k = 1, \ldots, p$. From condition C1, C4 and Corollary 6.21 in [15], we get that $\|R_k(u)\| = O(K^{-r})$. Then, invoking condition C3, a simple calculation yields

$$I_{11} = O_p(\tau n K^{-1-r})\|T\| = O_p(1 + n^{\frac{r}{2r+1}}a_n)\|T\|.$$

Then, notice that $E\{\epsilon_i | X_i, U_{_i}\} = 0$, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i W_i^{\mathrm{T}}T\epsilon_i = O_p(\|T\|).$$

Hence, it is easy to show that

$$I_{12} = O_p(\tau\sqrt{n}K^{-1})\|T\| = O_p(n^{\frac{-1}{2(2r+1)}} + n^{\frac{2r-1}{2(2r+1)}}a_n)\|T\|.$$

Similarly, we can prove that $I_{13} = O_p(1 + 2n^{\frac{r}{2r+1}}a_n)\|T\|^2$. Hence, by choosing a sufficiently large C, I_{13} dominates I_{11} and I_{12} uniformly in $\|T\| = C$. This implies that for any given $\varepsilon > 0$, if we choose C large enough, then

$$P\{\inf_{\|T\|=C} I_1 > 0\} \geq 1 - \varepsilon. \tag{12}$$

With the similar arguments, we can prove that for any given $\varepsilon > 0$, if C is sufficiently large, then

$$P\{\inf_{\|T\|=C} I_2 > 0\} \geq 1 - \varepsilon. \tag{13}$$

Next, we deal with I_3. Notice that $p'_{\lambda}(0)\mathrm{sgn}(0) = 0$, Condition C6 implies that

$$\frac{\tau n}{K}T^{\mathrm{T}}\mathbf{b}(\beta_0) = \tau n K^{-1}n^{\frac{-r}{2r+1}}T^{\mathrm{T}}n^{\frac{r}{2r+1}}\mathbf{b}(\beta_0) = o_p(\|T\|).$$

Then, it is easy to show that $I_3 = \tau n K^{-1}T^{\mathrm{T}}\{\mathbf{b}(\beta) - \mathbf{b}(\beta_0)\} + o_p(\|T\|)$. Furthermore, invoking condition C7, and using the standard argument of the Taylor expansion, we can prove that

$$I_3 = \tau^2 n K^{-1}\|T\|^2 o_p(1) = o_p(1 + n^{r/(2r+1)}a_n)\|T\|^2.$$

Hence, invoking I_{13}, it is clear that I_3 is dominated by I_1 uniformly in $\|T\| = C$. Then, by choosing a sufficiently large C, (11) holds. This implies, with probability at least $1 - \varepsilon$, that

there exists a local minimizer $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\tau)$. Hence, we have

$$(\hat{\beta}_k - \beta_k)^{\mathrm{T}} H(\hat{\beta}_k - \beta_k) = O_p\{n^{\frac{-2r}{2r+1}} + 2n^{\frac{-r}{2r+1}} a_n + a_n^2\}. \tag{14}$$

In addition, it is easy to show that

$$\int_0^1 R_k(u)^2 \mathrm{d}u = O_p(n^{\frac{-2r}{2r+1}}). \tag{15}$$

Then, note that

$$\begin{aligned}
\|\hat{\theta}_k(\cdot) - \theta_{k0}(\cdot)\|^2 &= \int_0^1 \{\hat{\theta}_k(u) - \theta_{k0}(u)\}^2 \mathrm{d}u \\
&= \int_0^1 \{B^{\mathrm{T}}(u)\hat{\beta}_k - B^{\mathrm{T}}(u)\beta_k + R_k(u)\}^2 \mathrm{d}u \\
&\le 2\int_0^1 \{B^{\mathrm{T}}(u)\hat{\beta}_k - B^{\mathrm{T}}(u)\beta_k\}^2 \mathrm{d}u + 2\int_0^1 R_k(u)^2 \mathrm{d}u \\
&= 2(\hat{\beta}_k - \beta_k)^{\mathrm{T}} H(\hat{\beta}_k - \beta_k) + 2\int_0^1 R_k(u)^2 \mathrm{d}u.
\end{aligned}$$

Combining this with (14) and (15), we complete the proof of Theorem 1. □

Proof of Theorem 2 Note that $\sup_u B(u) = O(1)$, and $\hat{\theta}_k(u) = B(u)^{\mathrm{T}}\hat{\beta}_k$. It suffices to show that for any given $\varepsilon > 0$, when n is large enough, $P(C_{nk}) < \varepsilon$, where $C_{nk} = \{\hat{\beta}_k \neq 0\}$, $k = d+1, \ldots, p$. Since $\hat{\beta}_k = O_p(n^{-r/(2r+1)})$, when n is large enough, there exists some C such that

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

If $\hat{\beta}_k \neq 0$, then by condition C6, we can prove that, when n is large enough, there exists some C such that

$$P(n^{\frac{r}{2r+1}} p'_{\lambda_k}(\|\hat{\beta}_k\|) > C) < \varepsilon/2. \tag{17}$$

In addition, by condition C7, we have

$$\inf_{\|\beta_k\| \le Cn^{-r/(2r+1)}} n^{\frac{r}{2r+1}} p'_{\lambda_k}(\|\beta_k\|) = n^{\frac{r}{2r+1}} \lambda_k \inf_{\|\beta_k\| \le Cn^{-r/(2r+1)}} \lambda_k^{-1} p'_{\lambda_k}(\|\beta_k\|) \to \infty.$$

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

$$P(C_{nk}) < \varepsilon/2 + P(n^{r/(2r+1)} p'_{\lambda_k}(\|\hat{\beta}_k\|) > C) < \varepsilon.$$

This completes the proof of Theorem 2. □

References

[1] CAI Zongwu, FAN Jianqing, LI Runze. *Efficient estimation and inferences for varying-coefficient models* [J]. J. Amer. Statist. Assoc., 2000, **95**(451): 888–902.

[2] HUANG Jianhua, WU C O, ZHOU Lan. *Polynomial spline estimation and inference for varying coefficient models with longitudinal data* [J]. Statist. Sinica, 2004, **14**(3): 763–788.

[3] XUE Liugen, ZHU Lixing. *Empirical likelihood for a varying coefficient model with longitudinal data* [J]. J. Amer. Statist. Assoc., 2007, **102**(478): 642–654.

[4] TANG Qingguo, CHENG Longsheng. *M-estimation and B-spline approximation for varying coefficient models with longitudinal data* [J]. J. Nonparametr. Stat., 2008, **20**(7): 611–625.
[5] WEI Chuanhua, WU Xizhi. *Asymptotic normality of estimators in partially linear varying coefficient models* [J]. J. Math. Res. Exposition, 2008, **28**(4): 877–885.
[6] FAN Jianqing, LI Runze. *Variable selection via nonconcave penalized likelihood and its oracle properties* [J]. J. Amer. Statist. Assoc., 2001, **96**(456): 1348–1360.
[7] LI Runze, LIANG Hua. *Variable selection in semiparametric regression modeling* [J]. Ann. Statist., 2008, **36**(1): 261–286.
[8] WANG Lifeng, CHEN Guang, LI Hongzhe. *Group SCAD regression analysis for microarray time course gene expression data* [J]. Bioinformatics, 2007, **23**(12): 1486–1494.
[9] WANG Lifeng, LI Hongzhe, HUANG Jianhua. *Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements* [J]. J. Amer. Statist. Assoc., 2008, **103**(484): 1556–1569.
[10] WANG Qihua, SUN Zhihua. *Estimation in partially linear models with missing responses at random* [J]. J. Multivariate Anal., 2007, **98**(7): 1470–1493.
[11] SUN Zhihua, WANG Qihua, DAI Pengjie. *Model checking for partially linear models with missing responses at random* [J]. J. Multivariate Anal., 2009, **100**(4): 636–651.
[12] ZHOU Yong, WAN A T K, WANG Xiaojing. *Estimating equations inference with missing data* [J]. J. Amer. Statist. Assoc., 2008, **103**(483): 1187–1199.
[13] STONE C J. *Optimal global rates of convergence for nonparametric regression* [J]. Ann. Statist., 1982, **10**(4): 1040–1053.
[14] ZOU Hui. *The adaptive lasso and its oracle properties* [J]. J. Amer. Statist. Assoc., 2006, **101**(476): 1418–1429.
[15] SCHUMAKER L L. *Spline Functions: Basic Theory* [M]. John Wiley & Sons, Inc., New York, 1981.
[16] JOHNSON B A. *Variable selection in semiparametric linear regression with censored data* [J]. J. R. Stat. Soc. Ser. B Stat. Methodol., 2008, **70**(2): 351–370.
[17] JOHNSON B A, LIN D Y, ZENG Donglin. *Penalized estimating functions and variable selection in semiparametric regression models* [J]. J. Amer. Statist. Assoc., 2008, **103**(482): 672–680.