Ann Inst Stat Math (2012) 64:213–231
DOI 10.1007/s10463-010-0312-7

Variable selection in semiparametric regression analysis for longitudinal data

Peixin Zhao · Liugen Xue

Received: 18 June 2009 / Revised: 21 December 2009 / Published online: 31 July 2010

Abstract In this paper, we present a variable selection procedure by using basis function approximations and a partial group SCAD penalty for semiparametric varying coefficient partially linear models with longitudinal data. With appropriate selection of the tuning parameters, we establish the oracle property of this procedure. A simulation study is undertaken to assess the finite sample performance of the proposed variable selection procedure.

Keywords Semiparametric regression model · Longitudinal data · Variable selection

1 Introduction

Varying coefficient models are commonly used for analysis of data measured repeatedly over time, such as time series analysis, longitudinal data analysis and functional data analysis. In practice, however, only some of the coefficients vary with certain covariate, hence one useful extension of the varying coefficient model is the semiparametric varying coefficient partially linear model. Under longitudinal data, the semiparametric varying coefficient partially linear model has the following structure

$$Y(t) = X(t)^T\theta(t) + Z(t)^T\beta + \epsilon(t), \tag{1}$$

P. Zhao (✉)
Department of Mathematics, Hechi University,
Yizhou, Guangxi 546300, China
e-mail: zpx81@163.com

L. Xue
College of Applied Sciences, Beijing University of Technology,
Beijing 100124, China
e-mail: lgxue@bjut.edu.cn

where $\beta = (\beta_1, \ldots, \beta_q)^T$ is a $q \times 1$ vector of unknown parameters, $\theta(t) = (\theta_1(t), \ldots, \theta_p(t))^T$ is a $p \times 1$ vector of unknown functions, $X(t)$ and $Z(t)$ are covariates, $Y(t)$ is the response variable at time t, and $\epsilon(t)$ is a zero-mean stochastic process. Here, we assume that t ranges over a nondegenerate compact interval, without loss of generality, that is assumed to be the unit interval [0, 1].

Model (1) is a useful extension of the partially linear model (see Lin and Carroll 2001; Xue and Zhu 2007a) and varying coefficient model (see Xue and Zhu 2007b; Wang et al. 2008). Moreover, Model (1) has been considered by Li et al. (2002), Zhang et al. (2002), Fan and Huang (2005) and You and Zhou (2006) in the case of i.i.d observations. For longitudinal data, when the number of covariates is small, Sun and Wu (2005) and Fan et al. (2007) considered the estimation of the coefficients in Model (1). However, when the number of covariates in Model (1) is large, an important problem is to select the important variables in such model. Variable selection is a very important topic in modern statistical inference. Variable selection procedures that do not shrink coefficients include forward selection, backward, stepwise deletion, and all subsets regression methods. These methods generate a sequence of models and select the best submodel by hypothesis testing or goodness-of-fit testing. A criticism of these procedures is that these selection procedures 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. An alternative to such procedures is methods that simultaneously shrink regression coefficients and set some coefficients to zero, thereby, removing them from the final model. For linear models, Fan and Li (2001) proposed a family of variable selection procedures based on smoothly clipped absolute deviation penalty (SCAD), that include bridge regression (see Frank and Friedman 1993) and LASSO (see Tibshirani 1996). Fan and Li (2004) proposed to use the SCAD penalty for variable selection in longitudinal data analysis. Li and Liang (2008) adopted this methodology to select important variables in the parametric components of semiparametric regression models. In addition, Wang et al. (2008) proposed a group SCAD procedure for variable selection of pure varying coefficient models with longitudinal data.

This paper extends the group SCAD variable selection procedure to the semiparametric varying coefficient partially linear regression model. We propose a partial group SCAD variable selection procedure that can select significant variables in the parametric components and nonparametric components simultaneously. Furthermore, this procedure can simultaneously select significant variables and estimate unknown coefficients. We also study the asymptotic properties of the resulting estimators. With proper choice of regularization parameters, we show that the variable selection procedure is consistent, and the estimators of coefficients have oracle property. Here the oracle property means that the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the nonzero coefficients in the parametric components have the same asymptotic distribution as that based on the correct submodel. This indicates that the penalized estimators work as well as if the subset of true zero coefficients were already known.

Our method offers the following improvements over existing methods. Firstly, we regard the observation times as realizations from an arbitrary counting process that can handle the inter-series dependence of the longitudinal data. In contrast, although

Wang et al. (2008) proposed a group SCAD procedure of variable selection for varying coefficient models with longitudinal data, the model they considered is just a special case of Model (1). In addition, they used the same constant tuning parameter for all coefficients, that implies that all coefficients are equally penalized. This is somewhat unfair, because we expect that the tuning parameter for zero coefficient should be larger than that for nonzero coefficient. Thus, we can simultaneously unbiasedly estimate large coefficients, and shrink the small coefficients toward zero. In this paper, the adaptive tuning parameters are used, and our simulation studies indicate that the variable selection procedure based on the adaptive tuning parameters performs better than that based on the constant tuning parameter. Secondly, although Li and Liang (2008) have considered the problem of variable selection for Model (1), they proposed a series of generalized likelihood ratio tests (GLRT) for selecting significant variables in the nonparametric components. This procedure poses great challenges because, for each submodel, it is necessary to estimate the varying coefficient functions. This will dramatically increase the computational burden. In addition, the limiting null distribution of the GLRT is a Chi-square distribution with diverging degrees of freedom. Then, it is inconvenient to obtain critical values for the GLRT. In contrast, our method can select significant variables in the parametric components and nonparametric components simultaneously, as well as simultaneously select significant variables and estimate unknown coefficients. This implies that our method can avoid the heavy computational burden, which is the essential improvement over Li and Liang (2008). In addition, our current statistical data set is longitudinal data, that is different from Li and Liang (2008).

The rest of this paper is organized as follows. In Sect. 2, we first propose the partial group SCAD variable selection procedure. Then, we present theoretical properties of our variable selection procedure, including the consistency of the variable selection procedure and the oracle property of the penalized estimators. In Sect. 3, based on local quadratic approximations, we propose an iterative algorithm for finding the penalized estimators. In Sect. 4, some simulations are carried out to assess the performance of the proposed methods. Finally, in Sect. 5, we present a brief discussion of the results and methods. The technical proofs of all asymptotic results are provided in the Appendix.

2 Variable selection via partial group SCAD

Suppose that we have a random sample of n subjects. For the ith subject, the response variable $Y_i(t)$ and the covariate vectors $X_i(t)$, $Z_i(t)$ are collected at time points $t = t_{i1}, \ldots, t_{in_i}, i = 1, \ldots, n$, where n_i is the total number of observations on the ith subject. Thus,

$$Y_i(t_{ij}) = X_i(t_{ij})^T \theta(t_{ij}) + Z_i(t_{ij})^T \beta + \epsilon_i(t_{ij}), \tag{2}$$

for $i = 1, \ldots, n$, and $j = 1, \ldots, n_i$. We assume that $X_i(t_{ij})$, $Z_i(t_{ij})$ and $\epsilon_i(t_{ij})$ from different subjects are independent, and $E\{\epsilon_i(t_{ij})|X_i(t_{ij}), Z_i(t_{ij})\} = 0$. As in Xue and Zhu (2007a), we also assume, in our asymptotic study, that n_i is bounded but the number of subjects n goes to infinity. In this paper, the time points, at which

the observations on the ith subject are made, are characterized by a counting process $N_i(t) \equiv \sum_{j=1}^{n_i} I(t_{ij} \le t), i = 1, \ldots, n$, that is a random sample from a certain population, where $I(\cdot)$ is the indicator function. $X_i(t)$, $Z_i(t)$ and $Y_i(t)$ are observed at the jump points of $N_i(t)$. In this paper, we assume that the observation times are independent of the covariates. That is,

$$E\{\mathrm{d}N_i(t)|X_i(t), Z_i(t)\} = \mathrm{d}\Lambda(t), \quad i = 1, \ldots, n,$$

where $\Lambda(t)$ is an arbitrary nondecreasing function. Although this assumption is not the weakest possible condition, it is imposed to facilitate the technical proofs, and it can be satisfied in many applications.

Let $B(t) = (B_1(t), \ldots, B_L(t))^T$ be B-spline basis functions with the order of M, where $L = K + M + 1$, and K is the number of interior knots. Then, $\theta_k(t)$ can be approximated by

$$\theta_k(t) \approx B(t)^T \gamma_k, \quad k = 1, \ldots, p.$$

Substituting this into Model (2), we can get

$$Y_i(t_{ij}) = W_i(t_{ij})^T \gamma + Z_i(t_{ij})^T \beta + \epsilon_i(t_{ij}), \tag{3}$$

where $W_i(t_{ij}) = I_p \otimes B(t_{ij}) \cdot X_i(t_{ij})$ and $\gamma = (\gamma_1^T, \ldots, \gamma_p^T)^T$. Model (3) is a standard linear regression model. Note that each function $\theta_k(t)$ in Model (2) is characterized by γ_k in Model (3). This motivate us to adopt the following partial group version of the SCAD regularized estimation for Model (3). That is, we estimate γ and β by minimizing

$$\begin{aligned} Q(\gamma, \beta) = & \sum_{i=1}^{n} \int_0^1 \left\{ Y_i(t) - W_i(t)^T \gamma - Z_i(t)^T \beta \right\}^2 \mathrm{d}N_i(t) \\ & + n \sum_{k=1}^{p} p_{\lambda_{1k}}(\|\gamma_k\|_H) + n \sum_{l=1}^{q} p_{\lambda_{2l}}(|\beta_l|), \end{aligned} \tag{4}$$

where $\|\gamma_k\|_H = (\gamma_k^T H \gamma_k)^{1/2}$, $H = (h_{ij})_{L \times L}$ is a matrix with $h_{ij} = \int B_i(t) B_j(t) \mathrm{d}t$, and $p_\lambda(\cdot)$ is the SCAD penalty function with λ as a tuning parameter (see Fan and Li 2001), defined as

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

with $a > 2$, $w > 0$ and $p_\lambda(0) = 0$. Here, the tuning parameter λ is not necessarily the same for all $\theta_k(\cdot)$ and β_l, and we denote λ as λ_{1k} and λ_{2l}, respectively in (4).

Although the longitudinal measurements are independent between different subjects, they are likely to be correlated within each subject. We use the counting process $N_i(t)$ to divide the data into n groups in (4). This method can handle the inter-series

dependency of longitudinal data. Let $\hat{\beta}$ and $\hat{\gamma} = (\hat{\gamma}_1^T, \ldots, \hat{\gamma}_p^T)^T$ be the solution by minimizing (4). Then, $\hat{\beta}$ is the penalized least squares estimator of β, and the estimator of $\theta_k(t)$ can be obtained by $\hat{\theta}_k(t) = B(t)^T \hat{\gamma}_k$.

Next, we study the asymptotic properties of the resulting penalized least squares estimators. Let $\theta_0(\cdot)$ and β_0 be the true values of $\theta(\cdot)$ and β, respectively. Without loss of generality, we assume that $\beta_{l0} = 0, l = s+1, \ldots, q$, and $\beta_{l0}, l = 1, \ldots, s$ are all nonzero components of β_0. Furthermore, we assume that $\theta_{k0}(\cdot) = 0$, $k = d+1, \ldots, p$, 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 least squares estimators.

Theorem 1 *Suppose that the regularity conditions C1–C6 in the Appendix hold and the number of knots* $K = O_p(n^{1/(2r+1)})$, *where* r *is defined in Condition* $C1$ *in the Appendix. Then,*

(i) $\|\hat{\beta} - \beta_0\| = O_p(n^{\frac{-r}{2r+1}} + a_n)$,
(ii) $\|\hat{\theta}_k(t) - \theta_{k0}(t)\| = O_p(n^{\frac{-r}{2r+1}} + a_n), \quad k = 1, \ldots, p,$

where $a_n = \max_{k,l} \left\{ |p'_{\lambda_{2l}}(|\beta_{l0}|)|, |p'_{\lambda_{1k}}(\|\gamma_{k0}\|_H)| : \beta_{l0} \neq 0, \gamma_{k0} \neq 0 \right\}$.

Furthermore, under some conditions, we show that such consistent estimators must possess the sparsity property, that is stated as follows

Theorem 2 *Suppose that the regularity conditions C1–C6 in the Appendix hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Let* $\lambda_{\max} = \max_{k,l}\{\lambda_{1k}, \lambda_{2l}\}$, *and* $\lambda_{\min} = \min_{k,l}\{\lambda_{1k}, \lambda_{2l}\}$. *If* $\lambda_{\max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{\min} \to \infty$, *as* $n \to \infty$. *Then, with probability tending to 1,* $\hat{\beta}$ *and* $\hat{\theta}(t)$ *must satisfy*

(i) $\hat{\beta}_l = 0, l = s+1, \ldots, q$,
(ii) $\hat{\theta}_k(t) = 0, k = d+1, \ldots, p$.

By Remark 1 in Fan and Li (2001), we have that, if $\lambda_{\max} \to 0$ as $n \to \infty$, then $a_n = 0$. Hence from Theorems 1 and 2, it is clear that, by choosing proper tuning parameters, our variable selection method is consistent and the estimators of nonparametric components achieve the optimal convergence rate as if the subset of true zero coefficients were already known (see Schumaker 1981). Next, we show that the estimators for nonzero coefficients in the parametric components have the same asymptotic distribution as that based on the correct submodel. To demonstrate this, we need more notations to present the asymptotic property of the resulting estimators. Let $\beta^* = (\beta_1, \ldots, \beta_s)^T$ and $\theta^*(\cdot) = (\theta_1(\cdot), \ldots, \theta_d(\cdot))^T$, and β_0^* and $\theta_0^*(\cdot)$ be the true values of β^* and $\theta^*(\cdot)$, respectively. Corresponding covariates are denoted by Z_i^*, and $X_i^*, i = 1, \ldots, n$. In addition, let

$$\Gamma = E\left\{ \int_0^1 [Z^*(t) - \mu(t)^T X^*(t)]^{\otimes 2} \mathrm{d}N(t) \right\},$$

$$B = E\left\{ \int_0^1 [Z^*(t) - \mu(t)^T X^*(t)] \epsilon(t) \mathrm{d}N(t) \right\}^{\otimes 2},$$

where $A^{\otimes 2} = AA^T$, $\mu(t) = \Phi(t)^{-1}\Psi(t)$, $\Phi(t) = E\{X^*(t)X^*(t)^T|t\}$ and $\Psi(t) = E\{X^*(t)Z^*(t)^T|t\}$. The following result states the asymptotic normality of $\hat{\beta}^*$.

Theorem 3 *Suppose that the regularity conditions C1–C6 in the Appendix hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Then,*

$$\sqrt{n}(\hat{\beta}^* - \beta^*) \to N(0, \Sigma),$$

where $\Sigma = \Gamma^{-1}B\Gamma^{-1}$.

Remark 1 Theorem 2 indicates that our variable selection procedure is consistent, and Theorems 1 and 3 indicate that the penalized estimators have the oracle property. That is, the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the nonzero coefficients in the parametric components have the same asymptotic distribution as that based on the correct submodel.

3 Algorithm

Because $Q(\gamma, \beta)$ is irregular at the origin, the commonly used gradient method is not applicable. Now, we develop an iterative algorithm based on local quadratic approximation of the penalty function $p_\lambda(\cdot)$ as in Fan and Li (2001). 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

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

Hence, for given the initial value β_{l0} with $|\beta_{l0}| > 0, k = 1, \ldots, q$, and γ_{k0} with $\|\gamma_{k0}\|_H > 0, l = 1, \ldots, p$, we can obtain

$$p_{\lambda_{2l}}(|\beta_l|) \approx p_{\lambda_{2l}}(|\beta_{l0}|) + \frac{1}{2}\frac{p'_{\lambda_{2l}}(|\beta_{l0}|)}{|\beta_{l0}|}\left(|\beta_l|^2 - |\beta_{l0}|^2\right),$$

$$p_{\lambda_{1k}}(\|\gamma_k\|_H) \approx p_{\lambda_{1k}}(\|\gamma_{k0}\|_H) + \frac{1}{2}\frac{p'_{\lambda_{1k}}(\|\gamma_{k0}\|_H)}{\|\gamma_{k0}\|_H}\left(\|\gamma_k\|_H^2 - \|\gamma_{k0}\|_H^2\right).$$

Let $U_i(t) = (Z_i(t)^T, W_i(t)^T)^T$, $\alpha = (\beta^T, \gamma^T)^T$, and

$$\Sigma_\lambda(\alpha_0) = \mathrm{diag}\left\{\frac{p'_{\lambda_{21}}(|\beta_{10}|)}{|\beta_{10}|}, \ldots, \frac{p'_{\lambda_{2q}}(|\beta_{q0}|)}{|\beta_{q0}|}, \frac{p'_{\lambda_{11}}(\|\gamma_{10}\|_H)}{\|\gamma_{10}\|_H}H, \ldots, \frac{p'_{\lambda_{1p}}(\|\gamma_{p0}\|_H)}{\|\gamma_{p0}\|_H}H\right\}.$$

As a consequence, except for a constant term, (4) becomes

$$Q(\alpha) = \sum_{i=1}^{n}\int_0^1 \left\{Y_i(t) - U_i(t)^T\alpha\right\}^2 \mathrm{d}N_i(t) + \frac{n}{2}\alpha^T\Sigma_\lambda(\alpha_0)\alpha.$$

This is a quadratic form and can be solved by

$$\left(\sum_{i=1}^{n}\int_0^1 U_i(t)U_i(t)^T \mathrm{d}N_i(t) + \frac{n}{2}\Sigma_\lambda(\alpha_0)\right)\alpha = \sum_{i=1}^{n}\int_0^1 U_i(t)Y_i(t)\mathrm{d}N_i(t). \quad (5)$$

Hence, we obtain the following iterative algorithm

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

Step 2. Set $\alpha^{(0)} = \alpha^{(k)}$, solve $\alpha^{(k+1)}$ by (5).

Step 3. Iterate Step 2 until convergence, and denote the final estimator of α as $\hat{\alpha}$. Then $\hat{\beta} = (I_{q\times q}, 0_{q\times pL})\hat{\alpha}$, and $\hat{\gamma} = (0_{pL\times q}, I_{pL\times pL})\hat{\alpha}$, where $I_{q\times q}$ and $I_{pL\times pL}$ are $q \times q$ and $pL \times pL$ identity matrices, respectively, and $0_{q\times pL}$ and $0_{pL\times q}$ are zero matrices.

In the initialization step, we obtain an initial estimator of α by using ordinary least squares method based on Model (3). To implement this method, the number of interior knots K, and the tuning parameters a and λ in the penalty function should be chosen. Fan and Li (2001) showed that the choice of $a = 3.7$ performs well in a variety of situations. Hence, we use their suggestion throughout this paper. In our simulations, we can see that this choice works well, although $a = 3.7$ maybe is not the optimal tuning parameter any more for the semiparametric models that we considered in this paper.

In addition, from our simulation studies in Sect. 4, we can see that the performance of the variable selection for the parametric components does not depend sensitively on the choice of the number of interior knots (see Table 2). Hence, we can use the similar method to choose the tuning parameters as in Wang et al. (2008) for pure varying coefficient models. More specifically, we can estimate λ_{1k}'s, λ_{2l}'s and K by minimizing the following cross-validation score

$$CV(K, \lambda_{11}, \ldots, \lambda_{1p}, \lambda_{21}, \ldots, \lambda_{2q}) = \sum_{i=1}^{n}\left\{Y_i - \tilde{Z}_i^T\hat{\alpha}^{(-i)}\right\}^2,$$

where $\hat{\alpha}^{(-i)}$ is the solution of (4) after deleting the ith subject. The minimization problem over a $p+q+1$-dimensional space is very difficult. However, it is expected that the choice of λ_{1k} and λ_{2l} should satisfy that the tuning parameter for zero coefficient is larger than that for nonzero coefficient. Thus, we can simultaneously unbiasedly estimate the large coefficients, and shrink the small coefficients toward zero. Hence, in practice, we suggest taking the following adaptive tuning parameters

$$\lambda_{1k} = \lambda/\left\|\hat{\gamma}_k^{(0)}\right\|_H, \quad \lambda_{2l} = \lambda/\left|\hat{\beta}_l^{(0)}\right|, \qquad k = 1, \ldots, p, \quad l = 1, \ldots, q, \quad (6)$$

where $\hat{\gamma}_k^{(0)}$ and $\hat{\beta}_l^{(0)}$ are initial estimators of γ_k and β_l, respectively, by using ordinary least squares method based on Model (3). Then, we can minimize the following two-dimensional minimization problem

$$CV(K,\lambda) = \sum_{i=1}^{n} \left\{ Y_i - \tilde{Z}_i^T \hat{\alpha}^{(-i)} \right\}^2. \tag{7}$$

In fact, such a choice of tuning parameters, in some sense, is the same rationale behind the adaptive lasso (see Zou 2006), and from our simulation experience, we found that this method works well.

4 Numerical results

In this section, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed method. And as in Li and Liang (2008), the performance of estimator $\hat{\beta}$ will be assessed by using the generalized mean square error (GMSE), defined as

$$\text{GMSE} = (\hat{\beta} - \beta)^T E(ZZ^T)(\hat{\beta} - \beta).$$

The performance of estimator $\hat{\theta}(\cdot)$ will be assessed by using the square root of average square errors (RASE)

$$\text{RASE} = \left\{ \frac{1}{M} \sum_{s=1}^{M} \sum_{k=1}^{p} \left[\hat{\theta}_k(t_s) - \theta_k(t_s) \right]^2 \right\}^{1/2},$$

where $t_s, s = 1, \ldots, M$ are the grid points at which the function $\hat{\theta}(t)$ are evaluated. In our simulation, $M = 200$ is used.

We simulate data from Model (1) with $n = 100, 150$ and 200, respectively, and make 1,000 simulation runs in each case. Furthermore, in each simulation, we generated covariates randomly according to the model $Z_l(t) = T(t) + e_l, l = 1, \ldots, 10$, and $X_k(t) = T(t) + e_k, k = 1, \ldots, 10$, where $e_l \sim N(0, 1.5), e_k \sim N(0, 1)$, and $T(t) \sim U(-0.5t, 0.5t)$ for given t. The counting process $N(t)$ for the observation times is set to be a random-effects Poisson process with intensity rate ζ, where $\zeta \sim U(0, 1)$. $Y(t)$ is generated according to Model (1), where $\epsilon(t)$ is a Gaussian process with zero mean and covariance function $E\{\epsilon(s)\epsilon(t)\} = \exp(-2|t - s|)$. This set up allows the data to be correlated in each subject and independent between different subjects.

To perform this simulation, we take $\beta = (\beta_1, \ldots, \beta_{10})^T$ with $\beta_1 = 1.2, \beta_2 = 2.5, \beta_3 = 0.5$ and $\beta_4 = 2$ to represent different important levels of the parametric components, and $\theta(t) = (\theta_1(t), \ldots, \theta_{10}(t))^T$ with $\theta_1(t) = 5.5 + 0.1\exp(2t - 1), \theta_2(t) = 0.8 + t(2 - t)$ and $\theta_6(t) = 2 - 3\sin(\pi t)$ to represent different types of nonparametric functions. While the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. Furthermore, in the following simulations, we use the cubic B-splines, and the number of interior knots and the tuning parameters are obtained by (7).

We first compare the performance of the variable selection procedure based on the adaptive tuning parameters (ATP) λ_{1k} and λ_{2l} defined by (6) with that based on the constant tuning parameter (CTP) used in Wang et al. (2008). The latter is taking the

Table 1 Variable selections by using the adaptive tuning parameters (ATP) and the constant tuning parameter (CTP)

	$n = 100$			$n = 150$			$n = 200$		
	C	I	GMSE	C	I	GMSE	C	I	GMSE
β									
ATP	5.661	0.003	0.015	5.775	0	0.012	5.874	0	0.008
CTP	4.756	0.001	0.050	5.555	0	0.016	5.821	0	0.009
$\theta(\cdot)$									
ATP	6.650	0	0.073	6.893	0	0.016	6.977	0	0.011
CTP	6.091	0	0.101	6.584	0	0.027	6.971	0	0.012

same tuning parameter for all parametric components and nonparametric components in the variable selection procedure, that is,

$$\lambda_{1k} = \lambda_{2l} \equiv \lambda, \qquad k = 1, \ldots, p, \quad l = 1, \ldots, q,$$

where λ can be obtained using cross-validation score function as in Wang et al. (2008). The average numbers of the estimated zero coefficients for the parametric components and the nonparametric components are reported in Table 1. The column labeled "C" gives the average number of coefficients of the true zeros correctly set to zero, and the column labeled "I" gives the average number of the true nonzeros incorrectly set to zero. Furthermore, Table 1 also presents the median of GMSE for the parametric components and the median of RASE for the nonparametric components over the 1,000 simulations.

From Table 1, we can see that the variable selections based on ATP and CTP both become better in terms of model error and model complexity as n increases. We also can see that, for given n, the performance of the variable selection based on ATP performs better than that based on CTP. The latter cannot eliminate some unimportant variables and gives larger model errors, and this is specially true when n is small. For large n, the both variable selection procedures perform very similar.

Next, we evaluate the sensitivity of the partial group SCAD variable selection procedure (gSCAD), proposed by this paper, for the parametric components and the nonparametric components on the choice of the number of interior knots. In this simulation, the number of interior knots is fixed at $K = \lfloor 0.5K_0 \rfloor$, K_0, $2K_0$ and $3K_0$, respectively, where K_0 is the number of interior knots obtained by (7). Furthermore, the tuning parameter λ is obtained by (7) for given K in each case. The average numbers of the estimated zero coefficients for the parametric components and the nonparametric components are reported in Tables 2 and 3, respectively. In Tables 2 and 3, the row labeled "Oracle" means the oracle estimators computed by using the true model when the zero coefficients are known.

From Tables 2 and 3, we can see that the gSCAD variable selection for the parametric components and the nonparametric components becomes more and more closer to the oracle procedure in terms of model error and model complexity as n increases

Table 2 Variable selections for the parametric components with different numbers of interior knots by gSCAD, where $K_1 = \lfloor 0.5K_0 \rfloor$, $K_2 = 2K_0$ and $K_3 = 3K_0$

	$n = 100$			$n = 150$			$n = 200$		
K	C	I	GMSE	C	I	GMSE	C	I	GMSE
K_0	5.665	0.003	0.016	5.776	0	0.012	5.877	0	0.008
K_1	5.620	0.004	0.020	5.760	0	0.013	5.874	0	0.008
K_2	5.617	0.036	0.025	5.754	0	0.014	5.875	0	0.008
K_3	5.615	0.040	0.034	5.752	0	0.014	5.870	0	0.009
Oracle	6	0	0.011	6	0	0.007	6	0	0.005

Table 3 Variable selections for the nonparametric components with different numbers of interior knots by gSCAD, where $K_1 = \lfloor 0.5K_0 \rfloor$, $K_2 = 2K_0$ and $K_3 = 3K_0$

	$n = 100$			$n = 150$			$n = 200$		
K	C	I	GMSE	C	I	GMSE	C	I	GMSE
K_0	6.650	0	0.074	6.854	0	0.038	6.976	0	0.011
K_1	6.197	0.001	0.165	6.561	0	0.116	6.782	0	0.078
K_2	5.499	0.034	0.234	5.923	0.012	0.131	6.417	0	0.095
K_3	4.627	0.079	0.686	5.156	0.026	0.255	5.998	0	0.193
Oracle	7	0	0.018	7	0	0.015	7	0	0.010

in each case. We also can see that, for given n, the different number of interior knots will affect the variable selection of gSCAD for the nonparametric components. It is mainly because the different number of interior knots will affect the estimator of the nonparametric components significantly. While the performance of gSCAD for the parametric components does not depend sensitively on the choice of the number of interior knots, and this is especially true when n is large.

Lastly, we compare the performance of the gSCAD variable selection with some existing variable selection methods. For the parametric components, we compare the performance of gSCAD with the variable selection procedure, says profile SCAD (pSCAD), proposed by Li and Liang (2008). The basic idea of pSCAD variable selection procedure is to transform Model (1) into the following linear model

$$Y(t)^* = Z(t)^T \beta + \epsilon(t), \tag{8}$$

where $Y(t)^* = Y(t) - X(t)^T \tilde{\theta}(t)$, and $\tilde{\theta}(\cdot)$ is a consistent estimator of $\theta(\cdot)$ for given β. Then, the SCAD procedure is used to select significant variables in the parametric components and obtain the regularized estimator $\tilde{\beta}$ based on Model (8). The simulation results are summarized in Table 4.

For the nonparametric components, we compared the gSCAD method with the generalized likelihood ratio tests (GLRT) method proposed by Li and Liang (2008).

Table 4 Variable selections for the parametric components with different variable selection methods

	$n = 100$			$n = 150$			$n = 200$		
	C	I	GMSE	C	I	GMSE	C	I	GMSE
Methods									
gSCAD	5.667	0.003	0.015	5.773	0	0.012	5.879	0	0.007
pSCAD	5.780	0.001	0.039	5.357	0	0.018	5.805	0	0.011
Oracle	6	0	0.012	6	0	0.007	6	0	0.005

Table 5 Variable selections for the nonparametric components with different variable selection methods

	$n = 100$				$n = 200$			
	C	I	RASE	Time (s)	C	I	RASE	Time (s)
Methods								
gSCAD	6.650	0	0.074	2.35	6.976	0	0.011	4.994
GLRT	6.646	0	0.076	4.51	6.974	0	0.017	15.728
Oracle	7	0	0.008	0.515	7	0	0.004	0.645

Here, the basic idea of GLRT method is transform Model (1) into the following varying coefficient model

$$Y(t)^{**} = X(t)^T \theta(t) + \epsilon(t), \tag{9}$$

where $Y(t)^{**} = Y(t) - Z(t)^T \tilde{\beta}$, and $\tilde{\beta}$ is obtained by (8) based on pSCAD. Then, a series of generalized likelihood ratio tests are used for selecting significant variables in the nonparametric components based on Model (9). In this simulation, we only generate $n = 100$ and 200 subjects, respectively. The simulation results are summarized in Table 5. In Table 5, we also represent the average computing time of each variable selection procedure. From Tables 4 and 5, we can make the following observations:

(1) For the parametric and nonparametric components, the performances of all variable selection procedures become more and more closer to the oracle procedure in terms of model error and model complexity as n increases.
(2) For the parametric components, when n is large, the results based on gSCAD are similar to that based on pSCAD. However, when n is small, the gSCAD method outperforms pSCAD method. In addition, pSCAD method is less discriminate, and can not eliminate some unimportant variables when n is small. This mainly because we can not give a workable estimator $\tilde{\theta}(\cdot)$ when n is small, that may affect the variable selection for β based on Model (8).
(3) For the nonparametric components, we can see that the performances of the gSCAD method are similar to the GLRT method in terms of model error and model complexity. However, the GLRT method needs much more computing time than that the gSCAD method is used.

5 Conclusion and discussion

We have proposed a variable selection procedure for semiparametric varying coefficient partially linear models with longitudinal data. This procedure can select significant variables in the parametric components and the nonparametric components simultaneously. Furthermore, this procedure can simultaneously select significant variables and estimate unknown coefficients. Our method extends the group SCAD penalty of Wang et al. (2008) from nonparametric setting to a semiparametric setting. We have shown that the proposed method is consistent in variable selection, and has the oracle property of the regularized coefficient estimators. Simulation studies indicated that the proposed procedure was very effective in selecting significant variables and estimating the regression coefficients.

In this paper, we assume that the dimensions of the covariates $X(t)$ and $Z(t)$ are fixed. However, if the dimensions p and q go to infinity as $n \to \infty$, the variable selection procedure proposed by this paper will not work any more. For such high-dimensional problems, some work has been done for the variable selection in linear models (see Fan and Lv 2008) and partially linear models (see Xie and Huang 2009). As a future research topic, it is interest to consider the variable selection for the semiparametric varying coefficient partially linear models with high-dimensional covariates. In addition, in this paper, we assume the total number of observations on the ith subject, says n_i, is bounded. Another interesting topic of further research is investigating the case that the number of observations on each subject goes to infinity.

Appendix: Proof of Theorems

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

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

C2 The intensity function of $N_i(t)$, says $f(t)$, is bounded away from 0 and infinity on $[0, 1]$. Furthermore, we assume that $f(t)$ is continuously differentiable on $(0, 1)$.

C3 Let $G_1(t) = E\{Z(t)Z(t)^T|t\}$, $G_2(t) = E\{X(t)X(t)^T|t\}$. Then, $G_1(t)$, $G_2(t)$ and $E\{\epsilon(t)^2|t\}$ are continuous with respect to t. Furthermore, for given t, $G_1(t)$ and $G_2(t)$ are positive definite matrix, and the eigenvalues of $G_1(t)$ and $G_2(t)$ are bounded.

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

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

C5 Let $b_n = \max_{k,l}\{|p''_{\lambda_{2l}}(|\beta_{l0}|)|, |p''_{\lambda_{1k}}(\|\gamma_{k0}\|_H)| : \beta_{l0} \neq 0, \gamma_{k0} \neq 0\}$, then $b_n \to 0$, as $n \to \infty$.

C6 The penalty function satisfies

$$\liminf_{n\to\infty}\liminf_{\beta_l\to 0^+}\lambda_{2l}^{-1}p'_{\lambda_{2l}}(|\beta_l|)>0,\quad l=s+1,\ldots,q,$$
$$\liminf_{n\to\infty}\liminf_{\|\gamma_k\|_H\to 0}\lambda_{1k}^{-1}p'_{\lambda_{1k}}(\|\gamma_k\|_H)>0,\quad k=d+1,\ldots,p.$$

These conditions are commonly adopted in the nonparametric literature and variable selection methodology. Conditions C1–C3 are similar to those used in Xue and Zhu (2007a) and Fan and Li (2004). Condition C4 implies that $c_0,\ldots,c_{K+1}$ is a C_0-quasi-uniform sequence of partitions of [0, 1] (see Schumaker 1981, p.216). Conditions C5 and C6 are assumptions on the penalty function, that are similar to that used in Wang et al. (2008), Fan and Li (2001), and Li and Liang (2008).

Proof of Theorem 1 Let $\delta=n^{-r/(2r+1)}+a_n$, $\beta=\beta_0+\delta U_1$, $\gamma=\gamma_0+\delta U_2$ and $U=(U_1^T,\ U_2^T)^T$. For part (i), we first show that, for any given $\varepsilon>0$, there exists a large constant C such that

$$P\left\{\inf_{\|U\|=C}Q(\gamma,\beta)>Q(\gamma_0,\beta_0)\right\}\geq 1-\varepsilon. \tag{10}$$

Let $\Delta(\gamma,\beta)=K^{-1}\{Q(\gamma,\beta)-Q(\gamma_0,\beta_0)\}$, then, invoking $\beta_{l0}=0, l=s+1,\ldots,q, \gamma_{k0}=0, k=d+1,\ldots,p$, and $p_\lambda(0)=0$, with a simple calculation, we have that

$$\begin{aligned}\Delta(\gamma,\beta)\geq&-\frac{2\delta}{K}\sum_{i=1}^n\int_0^1\left\{\epsilon_i(t)+X_i(t)^TR(t)\right\}\left\{Z_i(t)^TU_1+W_i(t)^TU_2\right\}\mathrm{d}N_i(t)\\&+\frac{\delta^2}{K}\sum_{i=1}^n\int_0^1\left\{Z_i(t)^TU_1+W_i(t)^TU_2\right\}^2\mathrm{d}N_i(t)\\&+\frac{n}{K}\sum_{l=1}^s[p_{\lambda_{2l}}(|\beta_l|)-p_{\lambda_{2l}}(|\beta_{l0}|)]\\&+\frac{n}{K}\sum_{k=1}^d[p_{\lambda_{1k}}(\|\gamma_k\|_H)-p_{\lambda_{1k}}(\|\gamma_{k0}\|_H)]\\\equiv&\ I_1+I_2+I_3+I_4,\end{aligned}$$

where $R(t)=(R_1(t),\ldots,R_p(t))^T$, and $R_k(t)=\theta_k(t)-B(t)^T\gamma_k, k=1,\ldots,p$. From conditions C1, C2 and Corollary 6.21 in Schumaker (1981), we get that $\|R(t)\|=O(K^{-r})$. Then, invoking condition C3, a simple calculation yields

$$\sum_{i=1}^n\int_0^1X_i(t)^TR(t)\left\{Z_i(t)^TU_1+W_i(t)^TU_2\right\}\mathrm{d}N_i(t)=O_p(nK^{-r}\|U\|). \tag{11}$$

Then, notice that $E\{\epsilon_i(t)|Z_i(t), X_i(t)\} = 0$, we can prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \epsilon_i(t)\left\{Z_i(t)^T U_1 + W_i(t)^T U_2\right\} \mathrm{d}N_i(t) = O_p(\|U\|).$$

Together this with (11), it is easy to show that

$$I_1 = O_p(\sqrt{n}K^{-1}\delta)\|U\| + O_p(nK^{-1-r}\delta)\|U\| = O_p\left(1 + n^{\frac{r}{2r+1}}a_n\right)\|U\|.$$

Similarly, we can prove that

$$I_2 = O_p(nK^{-1}\delta^2)\|U\|^2 = O_p\left(1 + 2n^{\frac{r}{2r+1}}a_n\right)\|U\|^2.$$

Hence, by choosing a sufficiently large C, I_2 dominates I_1 uniformly in $\|U\| = C$. Furthermore, by the standard argument of the Taylor expansion, we get that

$$\begin{aligned} I_3 &= K^{-1}n\sum_{l=1}^{s}[p_{\lambda_{2l}}(|\beta_l|) - p_{\lambda_{2l}}(|\beta_{l0}|)] \\ &= \sum_{l=1}^{s} K^{-1}n\delta p'_{\lambda_{2l}}(|\beta_{l0}|)\mathrm{sgn}(\beta_{l0})|U_{1l}| \\ &\quad + \sum_{l=1}^{s} K^{-1}n\delta^2 p''_{\lambda_{2l}}(|\beta_{l0}|)|U_{1l}|^2\{1 + o(1)\} \\ &\le \sqrt{s}K^{-1}n\delta a_n\|U\| + K^{-1}n\delta^2 b_n\|U\|^2. \end{aligned}$$

Then, it is easy to show that I_3 is dominated by I_2 uniformly in $\|U\| = C$. With the same argument, we can prove that I_4 is also dominated by I_2 uniformly in $\|U\| = C$. Hence, by choosing a sufficiently large C, (10) holds. Then, by the convexity of $Q(\cdot,\cdot)$, we have that

$$P\left\{\inf_{\|U\|\le C} Q(\gamma,\beta) > Q(\gamma_0,\beta_0)\right\} \ge 1 - \varepsilon.$$

This implies, with probability at least $1-\varepsilon$, that there exists a local minimizer $\hat{\beta}$ such that $\|\hat{\beta} - \beta_0\| = O_p(\delta)$, that completes the proof of part (i).

Next, we prove part (ii). Note that

$$\begin{aligned} \|\hat{\theta}_k(t) - \theta_{k0}(t)\|^2 &= \int_0^1 \left\{\hat{\theta}_k(t) - \theta_{k0}(t)\right\}^2 \mathrm{d}t \\ &= \int_0^1 \left\{B(t)^T\hat{\gamma}_k - B(t)^T\gamma_k + R_k(t)\right\}^2 \mathrm{d}t \end{aligned}$$

$$\leq 2\int_0^1 \left\{B(t)^T\hat{\gamma}_k - B(t)^T\gamma_k\right\}^2 \mathrm{d}t + 2\int_0^1 R_k(t)^2\mathrm{d}t$$
$$= 2\int_0^1 (\hat{\gamma}_k - \gamma_k)^T B(t)B(t)^T(\hat{\gamma}_k - \gamma_k)\mathrm{d}t + 2\int_0^1 R_k(t)^2\mathrm{d}t.$$

With the same arguments as the proof of part (i), we can get that $\|\hat{\gamma} - \gamma\| = O_p(n^{-r/(2r+1)} + a_n)$. Then, a simple calculation yields

$$\int_0^1 (\hat{\gamma}_k - \gamma_k)^T B(t)B(t)^T(\hat{\gamma}_k - \gamma_k)\mathrm{d}t = O_p\left\{(n^{\frac{-r}{2r+1}} + a_n)^2\right\}. \tag{12}$$

In addition, it is easy to show that

$$\int_0^1 R_k(t)^2\mathrm{d}t = O_p\left(n^{\frac{-2r}{2r+1}}\right). \tag{13}$$

Invoking (12) and (13), we complete the proof of part (ii). □

Proof of Theorem 2 We first prove part (i). From $\lambda \to 0$, it is easy to show that $a_n = 0$ for large n. Then by Theorem 1, it is sufficient to show that, for any γ that satisfies $\|\gamma - \gamma_0\| = O_p(n^{-r/(2r+1)})$, β_l that satisfies $\|\beta_l - \beta_{l0}\| = O_p(n^{-r/(2r+1)})$, $l = 1, \ldots, s$, and some given small $\varepsilon = Cn^{-r/(2r+1)}$, when $n \to \infty$, with probability tending to 1 we have

$$\frac{\partial Q(\gamma, \beta)}{\partial \beta_l} > 0, \quad \text{for } 0 < \beta_l < \varepsilon, \quad l = s+1, \ldots, q, \tag{14}$$

and

$$\frac{\partial Q(\gamma, \beta)}{\partial \beta_l} < 0, \quad \text{for } -\varepsilon < \beta_l < 0, \quad l = s+1, \ldots, q. \tag{15}$$

Thus, (14) and (15) imply that the minimizer of $Q(\gamma, \beta)$ attains at $\beta_l = 0, l = s+1, \ldots, q$.

By a similar the proof of Theorem 1, we have that

$$\frac{\partial Q(\gamma, \beta)}{\partial \beta_l} = \sum_{i=1}^n \int_0^1 Z_{il}(t)\left\{Y_i(t) - Z_i(t)^T\beta - W_i(t)^T\gamma\right\}\mathrm{d}N_i(t)$$
$$+ np'_{\lambda 2l}(|\beta_l|)\mathrm{sgn}(\beta_l)$$
$$= -2\sum_{i=1}^n \int_0^1 Z_{il}(t)\left\{\epsilon_i(t) + X_i(t)^T R(t)\right\}\mathrm{d}N_i(t)$$
$$-2\sum_{i=1}^n \int_0^1 Z_{il}(t)Z_i(t)^T(\beta_0 - \beta)\mathrm{d}N_i(t)$$

$$-2\sum_{i=1}^{n}\int_0^1 Z_{il}(t)W_i(t)^T(\gamma_0-\gamma)\mathrm{d}N_i(t)+np'_{\lambda_{2l}}(|\beta_l|)\mathrm{sgn}(\beta_l)$$
$$=n\lambda_{2l}\left\{\lambda_{2l}^{-1}p'_{\lambda_{2l}}(|\beta_l|)\mathrm{sgn}(\beta_l)+O_p\left(\lambda_{2l}^{-1}n^{-\frac{r}{2r+1}}\right)\right\}.$$

In addition, condition C6 implies that $\lim_{n\to\infty}\liminf_{\beta_l\to 0}\lambda_{2l}^{-1}p'_{\lambda_{2l}}(|\beta_l|)>0$, and note that $\lambda_{2l}n^{\frac{r}{2r+1}}>\lambda_{min}n^{\frac{r}{2r+1}}\to\infty$, it is clear that the sign of the derivative is completely determined by that of β_l, then (14) and (15) hold. This completes the proof of part (i).

Applying the similar techniques as in the analysis of part (i) in this theorem, we have, with probability tending to 1, that $\hat{\gamma}_k=0, k=d+1,\ldots,p$. Then, the result of this theorem is immediately achieved from $\hat{\theta}_k(t)=B^T(t)\hat{\gamma}_k$. □

Proof of Theorem 3 Let $\gamma^*=(\gamma_1^T,\ldots,\gamma_d^T)^T$, and γ_0^* be the true value of γ^*. Corresponding covariates are denoted by W_i^*, $i=1,\ldots,n$. Then, Theorems 1 and 2 imply that, as $n\to\infty$, with probability tending to 1, $Q(\gamma,\beta)$ attains the minimal value at $(\hat{\beta}^{*T},0)^T$ and $(\hat{\gamma}^{*T},0)^T$. Let $Q_{1n}(\gamma,\beta)=\partial Q(\gamma,\beta)/\partial\beta^*$ and $Q_{2n}(\gamma,\beta)=\partial Q(\gamma,\beta)/\partial\gamma^*$, then, $(\hat{\beta}^{*T},0)^T$ and $(\hat{\gamma}^{*T},0)^T$ must satisfy

$$\frac{1}{n}Q_{1n}((\hat{\gamma}^{*T},0)^T,(\hat{\beta}^{*T},0)^T)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Z_i^*(t)\left\{Y_i(t)-W_i^*(t)^T\hat{\gamma}^*-Z_i^*(t)^T\hat{\beta}^*\right\}\mathrm{d}N_i(t)$$
$$+\sum_{l=1}^{s}p'_{\lambda_{2l}}(|\hat{\beta}_l|)\mathrm{sgn}(\hat{\beta}_l)=0. \tag{16}$$
$$\frac{1}{n}Q_{2n}((\hat{\gamma}^{*T},0)^T,(\hat{\beta}^{*T},0)^T)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 W_i^*(t)\left\{Y_i(t)-W_i^*(t)^T\hat{\gamma}^*-Z_i^*(t)\hat{\beta}^*\right\}\mathrm{d}N_i(t)$$
$$+\sum_{k=1}^{d}p'_{\lambda_{1k}}(\|\hat{\gamma}_k\|_H)\frac{H\hat{\gamma}_k}{\|\hat{\gamma}_k\|_H}=0. \tag{17}$$

Applying the Taylor expansion to $p'_{\lambda_{2l}}(|\hat{\beta}_l|)$, we get that

$$p'_{\lambda_{2l}}(|\hat{\beta}_l|)=p'_{\lambda_{2l}}(|\beta_{l0}|)+\left\{p''_{\lambda_{2l}}(|\beta_{l0}|)+o_p(1)\right\}(\hat{\beta}_l-\beta_{l0}).$$

Furthermore, Condition C5 implies that $p''_{\lambda_{2l}}(|\beta_{l0}|)=o_p(1)$, and note that $p'_{\lambda_{2l}}(|\beta_{l0}|)=0$ as $\lambda_{max}\to 0$, then, $\sum_{l=1}^{s}p'_{\lambda_{2l}}(|\hat{\beta}_l|)\mathrm{sgn}(\hat{\beta}_l)=o_p(\hat{\beta}^*-\beta_0^*)$. Hence, by (16), a

simple calculation yields

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Z_i^*(t)\left\{Z_i^*(t)^T(\beta_0^*-\hat{\beta}^*)+W_i^*(t)^T(\gamma_0^*-\hat{\gamma}^*)+X_i^*(t)^T R^*(t)\right.$$
$$\left.+\epsilon_i(t)\right\}\mathrm{d}N_i(t)+o_p(\hat{\beta}^*-\beta_0^*)=0, \tag{18}$$

where $R^*(t)=(R_1(t),\ldots,R_d(t))^T$. Invoking (17), and using the similar arguments to (18), we can prove that

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 W_i^*(t)\left\{Z_i^*(t)^T\left(\beta_0^*-\hat{\beta}^*\right)+W_i^*(t)^T\left(\gamma_0^*-\hat{\gamma}^*\right)+X_i^*(t)^T R^*(t)\right.$$
$$\left.+\epsilon_i(t)\right\}\mathrm{d}N_i(t)+o_p\left(\hat{\gamma}^*-\gamma_0^*\right)=0. \tag{19}$$

Let $\Phi_n=\frac{1}{n}\sum_{i=1}^n\int_0^1 W_i^*(t)W_i^*(t)^T\mathrm{d}N_i(t)$, and $\Psi_n=\frac{1}{n}\sum_{i=1}^n\int_0^1 W_i^*(t)Z_i^*(t)^T\mathrm{d}N_i(t)$, then, by (19), we have that

$$\hat{\gamma}^*-\gamma_0^*=[\Phi_n+o_p(1)]^{-1}\frac{1}{n}\sum_{i=1}^{n}\int_0^1 W_i^*(t)\left[X_i^*(t)^T R^*(t)+\epsilon_i(t)\right]\mathrm{d}N_i(t)$$
$$+[\Phi_n+o_p(1)]^{-1}\Psi_n\left(\beta_0^*-\hat{\beta}^*\right).$$

Substituting this into (18), and a simple calculation yields

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Z_i^*(t)\left\{Z_i^*(t)-\Psi_n\Phi_n^{-1}W_i^*(t)\right\}^T\mathrm{d}N_i(t)\left(\hat{\beta}^*-\beta_0^*\right)+o_p(1)\left(\hat{\beta}^*-\beta_0^*\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Z_i^*(t)\left\{\epsilon_i(t)+X_i^*(t)^T R^*(t)-W_i^*(t)^T\left[\Phi_n^{-1}+o_p(1)\right]\Lambda_n\right\}\mathrm{d}N_i(t), \tag{20}$$

where $\Lambda_n=\frac{1}{n}\sum_{i=1}^n\int_0^1 W_i^*(t)[\epsilon_i(t)+X_i^*(t)^T R^*(t)]\mathrm{d}N_i(t)$. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Psi_n^T\Phi_n^{-1}W_i^*(t)\left\{Z_i^*(t)^T-W_i^*(t)^T\Phi_n^{-1}\Psi_n\right\}\mathrm{d}N_i(t)=0,$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Psi_n^T\Phi_n^{-1}W_i^*(t)\left\{\epsilon_i(t)+X_i^*(t)^T R^*(t)-W_i^*(t)^T\Phi_n^{-1}\Lambda_n\right\}\mathrm{d}N_i(t)=0.$$

Then, by (20), it is easy to show that

$$\begin{aligned}
&\left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \check{Z}_i^*(t)\check{Z}_i^*(t)^T \mathrm{d}N_i(t) + o_p(1)\right\}\sqrt{n}\left(\hat{\beta}^* - \beta_0^*\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \check{Z}_i^*(t)\epsilon_i(t)\mathrm{d}N_i(t) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \check{Z}_i^*(t)X_i^*(t)^T R^*(t)\mathrm{d}N_i(t) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \check{Z}_i^*(t)W_i^*(t)^T\left[\Phi_n^{-1} + o_p(1)\right]\Lambda_n \mathrm{d}N_i(t) \\
&\equiv I_1 + I_2 + I_3, \qquad (21)
\end{aligned}$$

where $\check{Z}_i^*(t) = Z_i^*(t) - \Psi_n^T\Phi_n^{-1}W_i^*(t)$. Using the Central Limits Theorem, we can obtain

$$I_1 \xrightarrow{\mathcal{L}} N(0, B), \qquad (22)$$

where $\xrightarrow{\mathcal{L}}$ means the convergence in distribution. In addition, note that

$$\sum_{i}^{n}\int_0^1 \check{Z}_i^*(t)W_i^*(t)^T \mathrm{d}N_i(t) = 0,$$

we have that $I_3 = 0$. Furthermore, a simple calculation yields

$$\begin{aligned}
I_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \left\{Z_i^*(t) - \mu(t)^T X_i^*(t)\right\} X_i^*(t)^T R^*(t)\mathrm{d}N_i(t) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \left\{\mu(t)^T X_i^*(t) - \Psi_n^T\Phi_n^{-1}W_i^*(t)\right\} X_i^*(t)^T R^*(t)\mathrm{d}N_i(t) \\
&\equiv I_{21} + I_{22}.
\end{aligned}$$

Invoking $E\{[Z_i^*(t) - \mu(t)^T W_i^*]X_i^*(t)^T\} = 0$, we can prove

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \left\{Z_i^*(t) - \mu(t)^T X_i^*(t)\right\} X_i^*(t)^T \mathrm{d}N_i(t) = O_p(1).$$

Together this with $\|R(u)\| = o(1)$, it is clear that $I_{21} = o_p(1)$. Similarly, we can prove that $I_{22} = o_p(1)$. Hence, we have that $I_2 = o_p(1)$. In addition, by the law of large numbers, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \check{Z}_i^*(t)\check{Z}_i^*(t)^T \mathrm{d}N_i(t) \xrightarrow{P} \Gamma, \qquad (23)$$

where $\xrightarrow{P}$ means the convergence in probability. Then, invoking (21)–(23), and using the Slutsky Theorem, we completes the proof of Theorem 3. □

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No. 10871013), the National Natural Science Foundation of Guangxi (Grant No. 2010GXNSFB013051), the National Natural Science Foundation of Beijing (Grant No. 1102008), the Graduate Student Foundation of Hechi University (Grant No. 2008QS-N014), and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (IHLB).

References

Fan, J. Q., Huang, T. (2005). Profile likelihood inference on semiparametric varying-coefficient partially linear models. *Bernoulli, 11*, 1031–1057.

Fan, J. Q., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J. Q., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of American Statistical Association, 99*, 710–723.

Fan, J. Q., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B, 70*, 849–911.

Fan, J. Q., Huang, T., Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association, 102*, 632–641.

Frank, I. E., Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*, 109–148.

Li, Q., Huang, C. J., Li, D., Fu, T. T. (2002). Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics, 20*, 412–422.

Li, R., Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics, 36*, 261–286.

Lin, X. H., Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association, 96*, 1045–1056.

Schumaker, L. L. (1981). *Spline functions*. New York: Wiley.

Sun, Y., Wu, H. (2005). Semiparametric time-varying coefficients regression model for longitudinal data. *Scandinavian Journal of Statistics, 32*, 21–47.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B, 58*, 267–288.

Wang, L., Li, H., Huang, J. Z. (2008). Variable selection in nonparametric varying coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association, 103*, 1556–1569.

Xie, H., Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics, 37*, 673–696.

Xue, L. G., Zhu, L. X. (2007a). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika, 94*, 921–937.

Xue, L. G., Zhu, L. X. (2007b). Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association, 102*, 642–654.

You, J. H., Zhou, Y. (2006). Empirical likelihood for semiparametric varying-coefficient partially linear regression models. *Statistics & Probability Letters, 76*, 412–422.

Zhang, W., Lee, S. Y., Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis, 82*, 166–188.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418–1429.