Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 3505306, 11 pages
https://doi.org/10.1155/2020/3505306

*Research Article*

# Penalized Quadratic Inference Function-Based Variable Selection for Generalized Partially Linear Varying Coefficient Models with Longitudinal Data

**Jinghua Zhang**[1,2] **and Liugen Xue**[2]

[1] *Department of Information Engineering, Jingdezhen Ceramic Institute, Jiangxi, China*
[2] *College of Applied Sciences, Beijing University of Technology, Beijing, China*

Correspondence should be addressed to Jinghua Zhang; zhangjinghua@jci.edu.cn

Received 5 April 2020; Revised 17 August 2020; Accepted 15 September 2020; Published 5 October 2020

Academic Editor: Luca Faes

Semiparametric generalized varying coefficient partially linear models with longitudinal data arise in contemporary biology, medicine, and life science. In this paper, we consider a variable selection procedure based on the combination of the basis function approximations and quadratic inference functions with SCAD penalty. The proposed procedure simultaneously selects significant variables in the parametric components and the nonparametric components. With appropriate selection of the tuning parameters, we establish the consistency, sparsity, and asymptotic normality of the resulting estimators. The finite sample performance of the proposed methods is evaluated through extensive simulation studies and a real data analysis.

## 1. Introduction

Identifying the significant variables is of great significance in all regression analysis. In practice, a number of variables are available for an initial analysis, but many of them may not be significant and should be excluded from the final model in order to increase the accuracy of prediction. Various procedures and criteria, such as stepwise selection and subset selection with Akaike information criterion (AIC), Mallows Cp, and Bayesian information criterion (BIC), have been developed. Nevertheless, these selection methods suffer from expensive computational costs. Many shrinkage methods have been developed for the purpose of computational efficiency, e.g., the nonnegative garrote [1], the LASSO [2], the bridge regression [3], the SCAD [4], and the one-step sparse estimator [5]. Among those, the SCAD possesses the virtues of continuity, unbiasedness, and sparsity. There are a number of works on the SCAD estimation methods in various regression models, e.g., [6–9]. Zhao and Xue [8] proposed a variable selection method to select significant variables in the parametric components and the nonparametric components simultaneously for the varying coefficient partially linear models (VCPLMs).

On the other hand, longitudinal data occurs frequently in biology, medicine, and life science, in which it is often necessary to make repeated measurements of subjects over time. The responses from different subjects are independent, but the responses from the same subject are very likely to be correlated. This feature is called "within-cluster correlation". Qu et al. [10] proposed a method of quadratic inference functions (QIFs) to treat the longitudinal data. The QIF can efficiently take the within-cluster correlation into account and is more efficient than the generalized estimating equation (GEE) [11] approach when the working correlation is misspecified. The QIF approach has been applied to many models, including varying coefficient models (VCM) [12, 13], partially linear models (PLM) [14], varying coefficient partially linear models (VCPLMs) [15], and generalized partially linear models (GPLM) [16]. Wang et al. [13] proposed a group SCAD procedure for variable selection of VCM with longitudinal data. More recently, Tian et al. [15] proposed a QIF-based SCAD penalty for the variable selection for VCPLM with longitudinal data.

As introduced in Li and Liang [17], the generalized partially linear varying coefficient model (GPLVCM) possesses the great flexibility of a nonparametric regression model and provides the explanatory power of a generalized linear regression model, which arises naturally due to categorical covariates. Many models are the special case of GPLVCM, e.g., VCM, VCPLM, PLM, and GLM. Li and Liang [17] studied variable selection for GPLVCM, where the parametric components are identified via the SCAD but the nonparametric components are selected via a generalized likelihood ratio test instead of shrinkage. In this paper, we extend the QIF-based group SCAD variable selection procedure to GPLVCM with longitudinal data, and the B-spline methods are adopted to approximate the nonparametric component in the model. With suitable chosen tuning parameters, the proposed variable selection procedure is consistent, and the estimators of regression coefficients have oracle property, i.e., the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the parametric components have the same asymptotic distribution as that based on the correct submodel.

The rest of this paper is organized as follows. In Section 2, we propose a variable selection procedure for the GPLVCM with longitudinal data. Asymptotic properties of the resulting estimators and an iteration algorithm are presented in Section 3. In Section 4, we carry out simulation studies to assess the finite sample performance of the method. A real data analysis is given in Section 5 to illustrate the proposed methodology. The details of proofs are provided in the appendix.

## 2. Methodology

*2.1. GPLVCM with Longitudinal Data.* In this article, we consider a longitudinal study with $n$ subjects and $m_i$ observations over time for the $i$th subject ($i = 1, \cdots, n$) for a total of $N = \sum_{i=1}^{n} m_i$ observations. Each observation consists of a response variable $Y_{ij}$ and the predicator variables $(X_{ij}, Z_{ij}, U(ij))$, where $X_{ij} \in R^p$, $Z_{ij} \in R^q$ and $U_{ij}$ is a scalar. We assume that the observations from different subjects are independent, but those within the same subject are dependent. The generalized varying coefficient partially linear model (GPLVM) with longitudinal data takes the form

$$\mu_{ij} = E\left(Y_{ij} \mid X_{ij}, Z_{ij}, U_{ij}\right) = h\left(X_{ij}^T\beta + Z_{ij}^T\alpha\left(U_{ij}\right)\right), \quad (1)$$

where $\mu_{ij}$ is the expectation of $Y_{ij}$ when $X_{ij}$, $Z_{ij}$, and $U_{ij}$ are given, $\beta = (\beta_1,\cdots,\beta_p)^T$ is an unknown $p \times 1$ regression coefficient vector, $h(\cdot)$ is a known smooth link function, and $\alpha(u) = (\alpha_1(u), \alpha_2(u),\cdots,\alpha_q(u))^T$ is a $q \times 1$ unknown monotonic smooth function vector. Without loss of generality, we assume $U \sim U[0, 1]$.

We approximate $\alpha(\cdot)$ by B-spline basis functions $B(u) = (B_1(u),\cdots,B_L(u))^T$ with the order of $M$, where $L = K + M + 1$ and $K$ is the number of interior knots, i.e.,

$$\alpha_k(u) \approx \alpha_k^*(u) = B(u)^T\gamma_k, \quad k = 1, \cdots, q, \quad (2)$$

where $\gamma_k = (\gamma_{k1},\cdots,\gamma_{kL})^T$ is a $L \times 1$ vector of unknown regression coefficients. Accordingly, $\mu_{ij}$ is approximated by

$$\mu_{ij} = E\left(Y_{ij} \mid X_{ij}, Z_{ij}, U_{ij}\right) = h\left(X_{ij}^T\beta + Z_{ij}^T \cdot I_q \otimes B\left(U_{ij}\right)^T\gamma\right), \quad (3)$$

where $\gamma = (\gamma_1^T, \cdots, \gamma_q^T)^T$ and "$\otimes$" is the Kronecker product. We use the B-spline basis functions because they are numerically stable and have bounded support [18]. The spline approach also treats a nonparametric function as a linear function with the basis functions as pseudodesign variables, and thus, any computational algorithm for the generalized linear models can be used for the GPLVCMs.

To incorporate the within-cluster correlation, we apply the QIFs to estimate $\beta$ and $\gamma$, respectively. Denote $\theta = (\beta^T, \gamma^T)^T$, we define the extended score $g_N(\theta)$ as follows:

$$g_N(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_i^T A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}}(Y_i - \mu_i) \\ \vdots \\ \dot{\mu}_i^T A_i^{-\frac{1}{2}} M_s A_i^{-\frac{1}{2}}(Y_i - \mu_i) \end{pmatrix}, \quad (4)$$

where $\dot{\mu}_i = \partial\mu_i/\partial\theta$, $A_i = \text{diag}\left(\text{Var}(Y_{i1}),\cdots,\text{Var}(Y_{im})\right)$ is the marginal variance matrix of subject $Y_i$, and $M_1, \cdots, M_s$ are the base matrices to represent the inverse of the working correlation matrix $R$ in GEE approach. Following Qu et al. [10], we define the quadratic inference functions to be

$$\mathcal{Q}_n(\theta) = n g_N^T(\theta)\Omega_n(\theta)^{-1} g_N(\theta), \quad (5)$$

where $\Omega_n(\theta) = (1/n)\sum_{i=1}^{n} g_i(\theta) g_i(\theta)^T$. Note that $\Omega_n$ depends on $\theta$. The QIF estimate $\widetilde{\theta}$ is then given by

$$\widetilde{\theta} = \text{argmin}_\theta \mathcal{Q}_n(\theta). \quad (6)$$

*2.2. Penalized QIF.* In real data analysis, the true regression model is always unknown. An overfitted model lowers the efficiency of estimation while an underfitted one leads to a biased estimator. A popular approach to identify the relevant predictors while estimating the nonzero parameters and functions in model (1) simultaneously is to exert some kind of "penalty" on the original objective function. Here, we choose the smoothly clipped absolute deviation (SCAD) penalty because it has several advantages such as unbiasedness, sparsity, and continuity. The SCAD-penalized quadratic inference function (PQIF) is defined as follows:

$$\mathcal{Q}_n^p(\theta) = \mathcal{Q}_n(\theta) + n\sum_{l=1}^{p} p_{\lambda_1}(|\beta_l|) + n\sum_{k=1}^{q} p_{\lambda_2}(\|\gamma_k\|_H), \quad (7)$$

where $\|\gamma_k\|_H = (\gamma_k^T H \gamma_k)^{1/2}$, $H = (h_{ij})_{L\times L}$, $h_{ij} = \int_0^1 B_i(u) B_j^T(u) du$ and $p_\lambda$ is the SCAD penalty function, where the derivative is defined as

$$p'_l(\omega) = \lambda\left\{I(\omega \leqslant l) + \frac{(a\lambda - \omega)_+}{(a-1)\lambda} I\{\omega > \lambda\}\right\}, \tag{8}$$

where $a > 2$, $\omega > 0$, $p_\lambda(0) = 0$; here, we choose $a = 3.7$ as in [4].

Note that

$$\|\gamma_k\|_H = \left(\int_0^1 \gamma_k^T B(u) B^T(U) \gamma_k du\right)^{1/2} = \left(\int_0^1 [\alpha^*(u)]^2 du\right)^{1/2}. \tag{9}$$

This group-wised penalization ensures that the spline coefficient vector of the same nonparametric component is treated as an entire group in model selection.

Denote $\widehat{\theta}$ to be the penalized estimator obtained by minimizing the penalized objective function of (7). Then, $\widehat{\beta} = (\theta\wedge_1,\cdots,\theta\wedge_p)^T$ is the estimator of the parameter $\beta$ and the estimator of the nonparametric function $\alpha(u)$ is calculated by $\widehat{\alpha}(u) = B(u)^T \widehat{\gamma}$, where $\widehat{\gamma} = (\gamma\wedge_1^T,\cdots,\gamma\wedge_q^T)^T = (\theta\wedge_{p+1},\cdots,\theta\wedge_{p+L},\theta\wedge_{p+L+1},\cdots,\theta\wedge_{p+qL})^T$.

## 3. Asymptotic Properties

*3.1. Oracle Property.* We next establish the asymptotic properties of the resulting penalized QIF estimators. We first introduce some notations. Let $\beta_0$ and $\alpha_0(\cdot)$ denote the true values of $\beta(\cdot)$ and $\alpha(\cdot)$. In addition, $\gamma_0$ is the spline coefficient vector from the spline approximation to $\alpha_0(\cdot)$. Without loss of generality, we assume that $\beta_{0l} \neq 0, l = 1,\cdots,p_1$ and $\beta_{0l} = 0, l = p_1 + 1,\cdots,p$, i.e., only the first $p_1$ component of $\beta_0$ is nonzero. Similarly, we assume that $\alpha_{0k}(\cdot) \neq 0, k = 1,\cdots,q_1$ and $\alpha_{0k}(\cdot) = 0, k = q_1 + 1,\cdots,q$, i.e., only the first $q_1$ component of $\alpha_0(\cdot)$ is nonzero. For convenience and simplicity, let $C$ denote a positive constant that may have different values at each appearance throughout this paper and $\|A\|$ denote the modulus of the largest singular value of matrix or vector $A$. Before the proof of our main theorems, we list some regularity conditions used in this paper.

*Assumption* (A1). The spline regression parameter $\gamma$ is identifiable, that is, $\gamma_0$ is the spline coefficient vector from the spline approximation to $\alpha_0(\cdot)$. In addition, there is a unique $\theta_0 = (\beta_0, \gamma_0) \in S$ satisfying $E\{g_N(\theta_0)\} = 0$, where $S$ is the parameter space.

*Assumption* (A2). The weight matrix $\Omega_n = (1/n)\sum_{i=1}^n g_i(\theta) g_i^T(\theta)$ converges almost surely to a constant matrix $\Omega_0$, where $\Omega_0$ is invertible.

*Assumption* (A3). The covariate matrices $X_i$ and $Z_i$, $i = 1,\cdots,n$, satisfy $\sup_i E\|X_i\|^4 < \infty$ and $\sup_i E\|Z_i\|^4 < \infty$.

*Assumption* (A4). The error $\varepsilon_i = Y_i - \mu_i$ satisfies $E(\varepsilon_i \varepsilon_i^T) = V_i$, $\sup_i\|V_i\| < \infty$, and there exists a positive constant $\delta$ such that $\sup_i E\|\varepsilon_i\|^{2+\delta} < \infty$.

*Assumption* (A5). All marginal variances $A_i \geq 0$ and $\sup_i\|A_i\| < \infty$.

*Assumption* (A6). $\{m_i\}$ is a bounded sequence of positive integers.

*Assumption* (A7). $\alpha_i(u), i = 1, 2,\cdots,q$ is $r$th continuous differentiable on $(0, 1)$, where $r \geq 2$.

*Assumption* (A8). The inner knots $\{c_i, i = 1,\cdots,K\}$ satisfy

$$\max_{1\leq i\leq K}|h_{i+1} - h_i| = o\left(K^{-1}\right), \qquad \frac{\max h_i}{\min h_i} \leq C_0, \tag{10}$$

where $h_i = c_i - c_{i-1}$.

*Assumption* (A9). The link function $h(\cdot)$ is 2th continuous differentiable and $E\{h^{2+\delta}\} < \infty$ for some $\delta > 2$.

*Assumption* (A10). $a_n = O(n^{-1/2})$; $b_n \longrightarrow 0$ as $n \longrightarrow \infty$, where

$$a_n = \max_{k,l}\left\{\left|p'_{\lambda_1}(|\beta_{0l}|)\right|, \left|p'_{\lambda_2}(\|\gamma_{0k}\|_H)\right| : \beta_{ol} \neq 0, \gamma_{0k} \neq 0\right\},$$
$$b_n = \max_{k,l}\left\{\left|p''_{\lambda_1}(|\beta_{0l}|)\right|, \left|p''_{\lambda_2}(\|\gamma_{0k}\|_H)\right| : \beta_{ol} \neq 0, \gamma_{0k} \neq 0\right\}. \tag{11}$$

Theorem 1 indicates that the estimator of nonparametric components achieve the optimal convergence rate.

**Theorem 1.** *Assume that Assumptions (A.1)–(A.10) hold and the number of knots* $K = O(N^{1/(2r+1)})$*, then*

$$\|\widehat{\alpha}_k(\cdot) - \alpha_{0k}(\cdot)\| = O_p\left(n^{-r/(2r+1)}\right), \quad k = 1,\cdots,q. \tag{12}$$

*Furthermore, under suitable condition, Theorem 1 shows that the penalized QIF estimator has the sparsity property.*

**Theorem 2.** *Assume that the conditions in Theorem 1 hold and* $\lambda_{\max} \longrightarrow 0$, $\sqrt{n}\lambda_{\min} \longrightarrow \infty$ *as* $n \longrightarrow \infty$*, with probability approaching 1,*

$$\widehat{\beta}_l = 0, \quad l = p_1 + 1,\cdots,p, \qquad \widehat{\alpha}_k(\cdot) \equiv 0, \quad k = q_1 + 1,\cdots,q, \tag{13}$$

*where* $\lambda_{\max} = \max\{\lambda_1, \lambda_2\}$, $\lambda_{\min} = \min\{\lambda_1, \lambda_2\}$.

*Theorems 1 and 2 indicate that with the tune parameter* $\lambda$ *being suitably chosen, the proposed selection method possesses* *model selection consistency. Next, we establish the asymptotic property for the estimator of the nonzero parametric components. Let* $\beta^* = (\beta_1, \cdots, \beta_{p_1})^T, \alpha^*(\cdot) = (\alpha_1^*(\cdot), \cdots, \alpha_{q_1}^*(\cdot))^T$ *and let* $\beta_0^*$ *and* $\alpha_0^*(\cdot)$ *denote their true value, respectively. In addition, let* $\gamma^* = (\gamma_1^T, \cdots, \gamma_{q_1}^T)^T$ *and* $\gamma_0^* = (\gamma_{01}^T, \cdots, \gamma_{0q_1}^T)^T$ *denote the spline coefficient vector of* $\alpha^*(\cdot)$ *and* $\alpha_0^*(\cdot)$*, respectively, and let* $X_i^*$ *and* $Z_i^*, i = 1, \cdots, n$ *denote their correspondent covariate. Let* $\tilde{X}_i = H'(\eta_i)X_i^*, \tilde{X} = (\tilde{X}_1^T, \cdots, \tilde{X}_n^T), \tilde{W}_i = H'(\eta_i) W_i^*, \tilde{W} = (\tilde{W}_1^T, \cdots, \tilde{W}_n^T)$*, and*

$$\Gamma = E\left\{\tilde{X}^T \tau \tilde{X} - E\left[\tilde{X}^T \tau \tilde{W} | u\right] E\left[\tilde{W}^T \tau \tilde{W} | u\right]^{-1} \tilde{W}^T \tau \tilde{X}\right\},$$
$$\Delta = E\left\{\left[\tau - E\left[\tilde{X}^T \tau \tilde{W} | u\right] E\left[\tilde{W}^T \tau \tilde{W} | u\right]^{-1} E\left[\tilde{W}^T \tau | u\right]\right] \varepsilon\right\}^{\otimes 2}, \tag{14}$$

*where* $\Delta^{\otimes 2} = \Delta\Delta^T$, $\tau = (\tau_{ij})_{n\times n}$ *is a* $n \times n$ *block matrix with its* $(i, j)$ *block taking the form*

$$\tau_{ij} = \sum_{k=1}^{s} \sum_{l=1}^{s} A_i^{-1/2} M_k A_i^{-1/2} H'(\eta_i) P_i^* \Omega_{lk}^{-1} P_j^{*T} H'\left(\eta_j\right) A_j^{-1/2} M_l A_j^{-1/2}. \tag{15}$$

*Theorem 3 states that* $\beta\wedge^*$ *is asymptotically normally distributed.*

**Theorem 3.** *Suppose that Assumptions (A.1)–(A.9) hold and the number of knots* $K = O(N^{1/(2r+1)})$*, then*

$$\sqrt{n}(\beta\wedge^* - \beta_0^*) \xrightarrow{L} N(0, \Sigma), \tag{16}$$

*where* $\Sigma = (\Gamma\Delta^{-1}\Gamma)^{-1}$ *and* $\xrightarrow{L}$ *represents the convergence in distribution.*

*3.2. Selection of Tuning Parameters.* Theorems 1–3 imply that the proposed variable selection procedure possessed the oracle property. However, this attractive feature relies on the choice of tuning parameters $\lambda_i$. The popular criteria to choose $\lambda_i$ include cross-validation, generalized cross-validation, AIC, and BIC. Wang et al. [19] suggested using BIC for the SCAD estimator in linear models and partially linear models and proved its model selection consistency property, i.e., the optimal parameter chosen by BIC can identify the true model with probability tending to one. Tian proved that for partially linear models. Hence, we adopt BIC to choose the optimal $\{\lambda_1, \lambda_2\}$. Following [19–21], we simplify the tuning parameters as

$$\lambda_1 = \frac{\lambda_0}{\left\|\tilde{\gamma}_k^{(0)}\right\|_H},$$
$$\lambda_2 = \frac{\lambda_0}{\left|\tilde{\beta}_k^{(0)}\right|}, \tag{17}$$

where $\tilde{\beta}_k^{(0)}$ and $\tilde{\gamma}_k^{(0)}$ are the unpenalized QIF estimates. Consequently, the original two-dimensional problem becomes a univariate problem about $\lambda_0$, which can be selected according to the following BIC-type criterion:

$$\text{BIC}_\lambda = \mathcal{Q}_n^p\left(\widehat{\theta}_\lambda\right) + df_\lambda \times \log(n), \tag{18}$$

where $\widehat{\theta}_\lambda = (\widehat{\beta}_\lambda, \widehat{\gamma}_{1\lambda}^T, \cdots, \widehat{\gamma}_{q\lambda}^T)$ is the regression coefficient estimated by minimizing the penalized QIF in (2.8) for a given $\lambda$ and $df_\lambda$ is the number of nonzero coefficients of $\widehat{\beta}_\lambda$ and $\|\widehat{\gamma}_{1\lambda}\|_H, \cdots, \|\widehat{\gamma}_{q\lambda}\|_H$. Thus, the tuning parameter $\lambda$ is obtained by

$$\widehat{\lambda} = \arg \min_{\lambda} \text{BIC}_\lambda. \tag{19}$$

From Theorem 4 of Tian et al. [15], the BIC tuning parameter selector enables us to select the true model consistently.

*3.3. An Algorithm Using Local Quadratic Approximation.* Based on Fan and Li's local quadratic approximating approach [4], we propose an iterative algorithm to minimize the PQIF (7). Similar with Tian et al. [15], we choose the unpenalized QIF estimator $\widetilde{\boldsymbol{\theta}}$ as the initial estimator. Let $\boldsymbol{\theta}^k = (\beta_1^k, \cdots, \beta_p^k, \boldsymbol{\gamma}_1^{kT}, \cdots, \boldsymbol{\gamma}_q^{kT})^T$ be the value of $\boldsymbol{\theta}$ at the $k$th iteration. If $\beta_l^k$ (or $\boldsymbol{\gamma}_l^k$) is close to 0 (or $\mathbf{0}$), i.e., $|\beta_l^k| \leqslant \epsilon$ (or $\|\boldsymbol{\gamma}_l^k\|_H \leqslant \epsilon$) with some small threshold value $\epsilon$, then we set $\beta_l^k = 0$ (or $\boldsymbol{\gamma}_l^k = \mathbf{0}$). We consider $\epsilon = 10^{-6}$ in our simulations.

Suppose $\beta_l^{k+1} = 0$, for $l = p_k + 1, \cdots, p$, and $\boldsymbol{\gamma}_l^{k+1} = \mathbf{0}$, for $l = q_k + 1, \cdots, q$, and $\boldsymbol{\beta}^{k+1} = (\beta_1^{k+1}, \cdots, \beta_{p_k}^{k+1}, \beta_{p_k+1}^{k+1}, \cdots, \beta_p^{k+1})^T = ((\boldsymbol{\beta}_N^{k+1})^T, (\boldsymbol{\beta}_Z^{k+1})^T)^T$, where $\boldsymbol{\beta}_N^{k+1} = (\beta_1^{k+1}, \cdots, \beta_{p_k}^{k+1})^T$ are the nonzero parametric components and $\boldsymbol{\beta}_Z^{k+1} = (\beta_{p_k+1}^{k+1}, \cdots, \beta_p^{k+1})^T = \mathbf{0}$. Similarly, let $\boldsymbol{\gamma}^{k+1} = ((\boldsymbol{\gamma}_1^{k+1})^T, \cdots, (\boldsymbol{\gamma}_{q_k}^{k+1})^T, (\boldsymbol{\gamma}_{q_k+1}^{k+1})^T, \cdots, (\boldsymbol{\gamma}_q^{k+1})^T)^T = ((\boldsymbol{\gamma}_N^{k+1})^T, (\boldsymbol{\gamma}_Z^{k+1})^T)^T$, where $\boldsymbol{\gamma}_N^{k+1} = ((\boldsymbol{\gamma}_1^{k+1})^T, \cdots, (\boldsymbol{\gamma}_{q_k}^{k+1})^T)^T$ and $\boldsymbol{\gamma}_Z^{k+1} = ((\boldsymbol{\gamma}_{q_k+1}^{k+1})^T, \cdots, (\boldsymbol{\gamma}_q^{k+1})^T)^T$ correspond to $q_k$ zero functions and $q - q_K$ zero functions, respectively. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_N^T, \boldsymbol{\beta}_Z^T, \boldsymbol{\gamma}_N^T, \boldsymbol{\gamma}_Z^T)^T$ denote a vector which has the same length and same partition with $\boldsymbol{\theta}^{k+1}$.

For the parametric term, if $|\beta_l^k| > \epsilon$, the penalty function at $\beta_l \approx \beta_l^k$ is approximated by

$$p_\lambda(\beta_l) \approx p_\lambda\left(\beta_l^k\right) + \frac{1}{2}\frac{p_\lambda'\left(\left|\beta_l^k\right|\right)}{\left|\beta_l^k\right|}\left(\beta_l^2 - \left(\beta_l^k\right)^2\right). \tag{20}$$

Similarly, to the nonparametric component, if $\|\boldsymbol{\gamma}_l^k\|_H > \epsilon$, the penalty function at $\boldsymbol{\gamma}_l \approx \boldsymbol{\gamma}_l^k$ is approximated by

$$\begin{aligned} p_\lambda(\|\boldsymbol{\gamma}_l\|_H) &\approx p_\lambda\left(\left\|\boldsymbol{\gamma}_l^k\right\|_H\right) + \frac{1}{2}\frac{p'_\lambda\left(\|\boldsymbol{\gamma}_l^k\|_H\right)}{\|\boldsymbol{\gamma}_l^k\|_H}\left(\|\boldsymbol{\gamma}_l\|_H{}^2 - \left\|\boldsymbol{\gamma}_l^k\right\|_H{}^2\right) \\ &= p_\lambda\left(\left\|\boldsymbol{\gamma}_l^k\right\|_H\right) + \frac{1}{2}\frac{p'{}_\lambda\left(\|\boldsymbol{\gamma}_l^k\|_H\right)}{\|\boldsymbol{\gamma}_l^k\|_H}\left(\boldsymbol{\beta}_l^T H\boldsymbol{\beta}_l - \boldsymbol{\beta}_l^{kT} H\boldsymbol{\beta}_l^k\right), \end{aligned} \tag{21}$$

where $p'{}_\lambda$ is the first-order derivative of the penalty function $p_\lambda$. This leads to the local approximation of the PQIF $\mathcal{Q}_n^p(\boldsymbol{\theta})$ by a quadratic function:

$$\begin{aligned} &\mathcal{Q}_n\left(\boldsymbol{\theta}^k\right) + \dot{\mathcal{Q}}_n\left(\boldsymbol{\theta}^k\right)^T\left(\boldsymbol{\omega}_{11} - \boldsymbol{\omega}_{11}^k\right) \\ &\quad + \frac{1}{2}\left(\boldsymbol{\omega}_{11} - \boldsymbol{\omega}_{11}^k\right)^T\ddot{\mathcal{Q}}_n\left(\boldsymbol{\theta}^k\right)\left(\boldsymbol{\omega}_{11} - \boldsymbol{\omega}_{11}^k\right) + \frac{n}{2}\boldsymbol{\omega}_{11}^t\Lambda\left(\theta^k\right)\boldsymbol{\omega}_{11}, \end{aligned} \tag{22}$$

where $\dot{\mathcal{Q}}_n(\boldsymbol{\theta}^k) = \partial\mathcal{Q}_n(\boldsymbol{\theta}^k)/\partial\boldsymbol{\omega}_{11}$, $\ddot{\mathcal{Q}}_n(\boldsymbol{\theta}^k) = \partial^2\mathcal{Q}_n(\boldsymbol{\theta}^k)/\partial\boldsymbol{\omega}_{11}\partial\boldsymbol{\omega}_{11}^T$, with $\boldsymbol{\omega}_{11} = (\boldsymbol{\beta}_N^T, \boldsymbol{\gamma}_Z^T)^T$, and

$$\Lambda\left(\boldsymbol{\theta}^k\right) = \operatorname{diag}\left\{\frac{p'{}_{\lambda_2}\left(\left|\beta_1^k\right|\right)}{\left|\beta_1^k\right|}, \cdots, \frac{p'{}_{\lambda_2}\left(\beta_{p_k}^k\right)}{\left|\beta_{p_k}^k\right|}, \frac{p'{}_{\lambda_1}\left(\|\boldsymbol{\gamma}_1^k\|_H\right)}{\|\boldsymbol{\gamma}_1^k\|_H}H, \cdots, \frac{p'{}_{\lambda_1}\left(\left\|\boldsymbol{\gamma}_{q_k}^k\right\|_H\right)}{\left\|\boldsymbol{\gamma}_{q_k}^k\right\|_H}H\right\}. \tag{23}$$

Minimizing the quadratic function (22), we obtain $\boldsymbol{\omega}_{11}^{k+1}$. The Newton-Raphson method then iterates the following process to convergence:

$$\begin{aligned} \boldsymbol{\omega}_{11}^{k+1} = \boldsymbol{\omega}_{11}^k &- \left\{\ddot{\mathcal{Q}}_n\left(\boldsymbol{\omega}_{11}^k\right) + n\Lambda\left(\boldsymbol{\omega}_{11}^k\right)\right\}^{-1} \\ &\cdot \left\{\dot{\mathcal{Q}}_n\left(\boldsymbol{\omega}_{11}^k\right) + n\Lambda\left(\boldsymbol{\omega}_{11}^k\right)\boldsymbol{\omega}_{11}^k\right\}. \end{aligned} \tag{24}$$

## 4. Simulation Studies

*4.1. Assessing Rule.* In this section, we conduct a simulation study to assess the finite sample performance of the proposed procedures. Following [17], the performance of estimator $\widehat{\beta}$ will be assessed by the generalized mean square error (GMSE), which is defined as

$$\mathrm{GMSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta} - \beta\right)X_i^* X_i^{*T}\left(\widehat{\beta} - \beta\right). \tag{25}$$

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

$$\mathrm{RASE} = \left\{\frac{1}{M}\sum_{v=1}^{M}\sum_{k=1}^{q}\left[\alpha\wedge_k(u_v) - \alpha_k(u_v)\right]^2\right\}^{1/2}, \tag{26}$$

where $u_v$, $v = 1, \cdots, M$ are the grid points where the function $\widehat{\alpha}(u)$ is evaluated. In our simulation, $M = 300$ is used.

To assess the performance of the variable selection, we use "$C$" to denote the average number of zero regression coefficients that are correctly estimated as zero and use "IC" to denote the average number of nonzero regression coefficients that are erroneously set to zero. The more closer the value of "$C$" to the number of true zero coefficient in the model and the more closer the value of "IC" to zero, the better the performance of the variable selection procedure is.

In our simulations, we use the sample quantiles of $U_{ij}$ as knots and take the number of internal knots to be 3, that is, $O(N^{1/5})$. This particular choice is consistent with the asymptotic theory in Section 3 and performs well in the simulations. For each simulated dataset, the proposed estimation procedures for finding out penalized QIF estimators with SCAD and LASSO penalty functions are considered. The tuning parameters $\lambda_1, \lambda_2$ for the penalty functions are chosen by BIC from 50 equispaced grid points in $[-15, 5]$. For each of these methods, the average of zero coefficients over the 500 simulated datasets is reported.

*4.2. Study 1 (Partial Penalty).* Consider a Bernoulli response

$$\operatorname{logit}\left\{Y_{ij}\right\} = X_{ij}^T\beta + \alpha\left(U_{ij}\right), \tag{27}$$

where $\beta = (2{,}1.5{,}0.7, \mathbf{0}_{17}^T)^T$, $m = 6$, $X_{ij} \sim N(0, I_{20})$, $\alpha(U_{ij}) = 0.4\cos((\pi/2)U_{ij})$, and $U_{ij}$ are drawn independently from $U[0, 1]$. Response variable $Y_{ij}$ with compound symmetry correlation structure (CS) is generated according to Oman [22]. In our simulation study, we consider $\rho = 0.25$ and 0.75, representing weak and strong correlations, respectively. In some situations, we prefer not to shrink some certain components in the variable selection procedure when some kind of prior information is available. Partial penalty arises naturally for such case. In this example, we only exert penalty on the parametric component, i.e., coefficient $\beta$. In this situation, the PQIF (7) becomes

$$\mathcal{Q}_n^p(\theta) = \mathcal{Q}_n(\theta) + n\sum_{l=1}^{p} p_{\lambda_1}(\beta_l). \tag{28}$$

Table 1: Variable selection for the parametric components under different methods.

| | Method | $n = 150$ | | | $n = 200$ | | | $n = 300$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GMSE | $C$ | IC | GMSE | $C$ | IC | GMSE | $C$ | IC |
| $\rho = 0.75$ | SCAD | 0.0011 | 15.83 | 0 | 0.0006 | 16.246 | 0 | 0.0005 | 16.746 | 0 |
| | LASSO | 0.0006 | 14.81 | 0 | 0.0005 | 15.346 | 0 | 0.0004 | 15.574 | 0 |
| $\rho = 0.25$ | SCAD | 0.0011 | 15.75 | 0 | 0.0006 | 16.70 | 0 | 0.0004 | 16.846 | 0 |
| | LASSO | 0.0007 | 14.82 | 0 | 0.0006 | 14.96 | 0 | 0.0005 | 15.35 | 0 |

The variable selection result is reported in Tables 1 and 2. Tables 1 and 2 show that the performance of the proposed variable selection approach improves as $n$ increases, e.g., the number of correctly recognized zero coefficient increases to the number of true zero coefficient in the model and the GMSE of $\widehat{\beta}$ decreases as $n$ increases. In addition, the RASE of $\widehat{\alpha}(u)$ also decreases as $n$ increases, which means the estimated curve of $\widehat{\alpha}(u)$ fits better to the true line of $\alpha(u)$ when the sample size increases. Moreover, the SCAD penalty method outperforms the LASSO penalty ones in the sense of correct variable selection rate, which significantly reduces the model uncertainty and complexity.

*4.3. Study 2 (Fixed-Dimensional Setup).* In this example, we generate data from the following model:

$$\operatorname{logit}\left\{Y_{ij} = 1 \mid X_{ij}, U_{ij}\right\} = X_{ij}^T\beta + Z_{ij}^T\alpha\left(U_{ij}\right), \qquad (29)$$

where $\beta = (2,1.5,0.7,\mathbf{0}_7^T)$ and $\alpha(u) = (\alpha_1(u), \alpha_2(u), \mathbf{0}_5^T)^T$ with $\alpha_1(u) = 0.8\cos((\pi/2)u), \alpha_2(u) = 1.5 + u^2$, $X_{ij}$ and $Z_{ij}$ $(j = 1,\cdots,6)$ come from a multivariate normal distribution with mean zero, marginal variance 1 and correlation coefficient 0.5, and $u \sim U(0,1)$. Response variable $Y_{ij}$ with compound symmetry correlation structure (CS) is generated by the same method as study 1 and we also consider $\rho = 0.25$ and 0.75, representing weak and strong correlations, respectively. We generated 500 datasets for each pair of $(N, \rho)$. The results are also reported in Tables 3 and 4.

Table 3 reports the variable selection for the parametric components; it shows that the performances become better and better as $n$ increases, e.g., the number of correctly recognized zero coefficients, which is denoted as values in the column labeled "$C$," becomes more and more closer to the true number of zero regression coefficients in the model. At the same time, the GMSE decreases steadily as $n$ increases. Table 4 shows that, for the nonparametric components, the performances of the proposed variable selection method are similar to those of the method for the parametric components. As $n$ increases, the RASE of the estimated nonparametric function also becomes smaller and smaller. This reflects that the estimate curves fit better to the corresponding true line as the sample size increases. Moreover, the SCAD penalty method outperforms the LASSO penalty ones in the sense of correct variable selection rate, which significantly reduces the model uncertainty and complexity.

Table 2: RASE of $\widehat{\alpha}(u)$ under different methods.

| | Method | $n = 150$ | $n = 200$ | $n = 300$ |
|---|---|---|---|---|
| $\rho = 0.75$ | SCAD | 0.1920 | 0.2051 | 0.1054 |
| | LASSO | 0.0999 | 0.0840 | 0.1064 |
| $\rho = 0.25$ | SCAD | 0.2449 | 0.2460 | 0.0694 |
| | LASSO | 0.1399 | 0.1205 | 0.1033 |

To study the influence of misspecified correlation structure to the proposed approach, we perform variable selection when the working correlation structure is specified to be CS and first-order autoregressive (AR-1), respectively. The result is listed in Table 5. It is known that the QIF estimator is insensitive to misspecification in correlation structure. Table 5 shows that the proposed variable selection procedure gives similar results even when the correlation structure is misspecified. This indicates that our method is robust.

*4.4. Study 3 (High-Dimensional Setup).* In this example, we discuss how the proposed variable selection procedure can be applied to the "large $n$, diverging $p/q$" setup for longitudinal models. We consider the high-dimensional setup of study 2. In this simulation, we take $n = 300, m = 6, p = 20 = O(N^{1/4}), q = 10 = O(N^{1/4})$. The true coefficient vector is $\beta = (2,1.5,0.7, 0_{17}^T)^T, \alpha(u) = (\alpha_1(u), \alpha_2(u), 0_{10}^T)^T$, where $\alpha_1(u)$ and $\alpha_2(u)$ are defined in study 2. The other settings are the same with study 2. The results are reported in Table 6. It is easy to see that the proposed variable selection procedure is able to correctly identify the true model and works well in the "large $n$, diverging $p/q$" setup.

## 5. Application to Infectious Disease Data

We apply the proposed method to analyze an infectious disease data (indon.data), which has been well analyzed by many authors, such as [16, 23–27]. In this study, a total of 275 preschool children were examined every three months for 18 months. The response is the presence of respiratory infection (1 = yes, 0 = no). The primary interest is in studying the relationship between the risk of respiratory infection and vitamin A deficiency (1 = yes, 0 = no).

In our study, we consider the following GPLVCM model

$$\operatorname{logit}\left\{\mu_{ij} \mid X_{ij}, t_{ij}\right\} = \sum_{k=1}^{6} \beta_i x_{ij} + \alpha_0\left(t_{ij}\right) + z_{ij}\alpha_1\left(t_{ij}\right), \qquad (30)$$

Table 3: Variable selection for the parametric components under different methods.

| | Method | | $n = 150$ | | | $n = 200$ | | | $n = 300$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GMSE | $C$ | IC | GMSE | $C$ | IC | GMSE | $C$ | IC |
| $\rho = 0.75$ | SCAD | 0.0048 | 6.76 | 0 | 0.0036 | 6.846 | 0 | 0.0030 | 6.864 | 0 |
| | LASSO | 0.0039 | 4.694 | 0 | 0.0033 | 4.766 | 0 | 0.0028 | 5.074 | 0 |
| $\rho = 0.25$ | SCAD | 0.0047 | 6.76 | 0 | 0.0035 | 6.718 | 0 | 0.0028 | 6.846 | 0 |
| | LASSO | 0.0038 | 4.814 | 0 | 0.0035 | 4.98 | 0 | 0.0029 | 5.048 | 0 |

Table 4: Variable selection for the nonparametric components under different methods.

| | Method | | $n = 150$ | | | $n = 200$ | | | $n = 300$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GMSE | $C$ | IC | GMSE | $C$ | IC | GMSE | $C$ | IC |
| $\rho = 0.75$ | SCAD | 0.1696 | 4.35 | 0 | 0.1221 | 4.66 | 0 | 0.0812 | 4.83 | 0 |
| | LASSO | 0.1932 | 4.38 | 0 | 0.1540 | 4.36 | 0 | 0.1235 | 4.57 | 0 |
| $\rho = 0.25$ | SCAD | 0.1636 | 4.42 | 0 | 0.1076 | 4.72 | 0 | 0.0344 | 4.85 | 0 |
| | LASSO | 0.1982 | 4.40 | 0 | 0.1160 | 4.68 | 0 | 0.0398 | 4.76 | 0 |

Table 5: Variable selection when the true $R$ is CS when $n = 300$.

| Working $R$ | Method | | $\beta$ | | | $\alpha(\cdot)$ | |
|---|---|---|---|---|---|---|---|
| | | GMSE | $C$ | IC | RASE | $C$ | IC |
| $\rho = 0.75$ | | | | | | | |
| CS | SCAD | 0.0030 | 6.864 | 0 | 0.0812 | 4.83 | 0 |
| | LASSO | 0.0028 | 5.074 | 0 | 0.1235 | 4.57 | 0 |
| AR-1 | SCAD | 0.0033 | 6.856 | 0 | 0.0935 | 4.82 | 0 |
| | LASSO | 0.0034 | 4.924 | 0 | 0.1230 | 4.57 | 0 |
| $\rho = 0.25$ | | | | | | | |
| CS | SCAD | 0.0028 | 6.846 | 0 | 0.0344 | 4.85 | 0 |
| | LASSO | 0.0029 | 5.048 | 0 | 0.0398 | 4.76 | 0 |
| AR-1 | SCAD | 0.0030 | 6.846 | 0 | 0.0354 | 4.86 | 0 |
| | LASSO | 0.0031 | 5.048 | 0 | 0.0411 | 4.75 | 0 |

Table 6: Variable selection under high-dimensional setup.

| | Method | | $\beta$ | | | $\alpha(\cdot)$ | |
|---|---|---|---|---|---|---|---|
| | | GMSE | $C$ | IC | RASE | $C$ | IC |
| $\rho = 0.75$ | SCAD | 0.0036 | 16.664 | 0 | 0.1148 | 9.656 | 0 |
| | LASSO | 0.0033 | 15.574 | 0 | 0.1239 | 9.546 | 0 |
| $\rho = 0.25$ | SCAD | 0.0034 | 16.846 | 0 | 0.1047 | 9.875 | 0 |
| | LASSO | 0.0039 | 15.35 | 0 | 0.1138 | 9.802 | 0 |

where $t$ is age, $X_1$ is vitamin A deficiency, $X_2, X_3$ are the seasonal cosine and seasonal sine variables, respectively, which indicate the season when those examinations took place, $X_4$ is gender (1 = female, 0 = male), $X_5$ is height, $X_6$ is stunting status (1 = yes, 0 = no), and $Z_1 = X_5^2$ is the square of height. The with-cluster correlation structure is assumed to be exchangeable, i.e., compound symmetric. This structure is also used in [16, 26, 27].

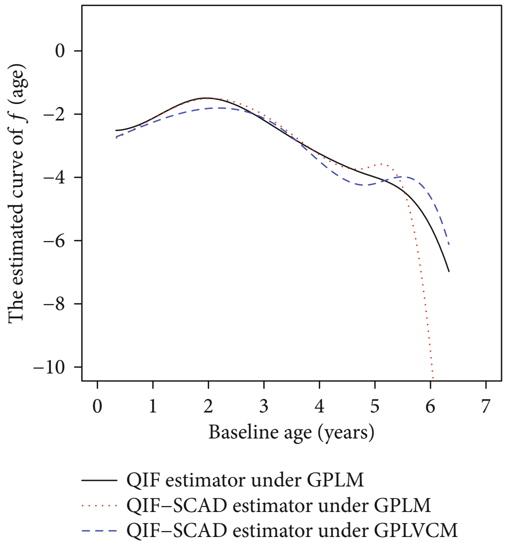

Figure 1: The estimated function on age for the infectious disease data.

We apply the proposed QIF-based group SCAD variable selection procedure to the above model and recognize five nonzero coefficients and one nonzero function $\alpha_0(t)$, where $\beta_1 = 0.842$, $\beta_2 = -0.685$, $\beta_3 = -0.309$, $\beta_4 = -0.554$, and $\beta_6 = 0.966$. The results are generally consistent with those previous studies, but our results show that the height has no significant impact on the infectious rate and can be removed from the model. Figure 1 reports the curve of baseline age function $\alpha_0(t)$ estimated by QIF-based group SCAD that is estimated by QIF and that is estimated by QIF-based SCAD partial penalty to $\beta$ in [16], where the GPLM without the varying coefficient term is used. Figure 1 implies that the probability of having respiratory infection increases at the very early stage, then decreases steadily, and declines dramatically when the age is over 5.5 years old. This also coincides with previous results [16, 26, 27].

## 6. Conclusion and Discussion

We proposed a QIF-based group SCAD variable selection procedure for the generalized partially linear varying coefficient models with longitudinal data. This procedure can select significant variables in the parametric components and nonparametric components simultaneously. Under mild conditions, the estimators of regression coefficients have oracle property. Simulation studies indicate that the proposed procedure is very effective in selecting significant variables and estimating the regression coefficients.

In this paper, we assume that the dimensions of the covariates $X$ and $Z$ are fixed. Study 3 in simulations shows that the proposed approach still have desired results when the dimensions $p$ and $q$ go to infinity as $n \longrightarrow \infty$. However, when in ultrahigh-dimensional case, the proposed variable selection procedure may not work well anymore. As a future research topic, it is interesting to consider the variable selection for the generalized partially linear varying coefficient models with ultrahigh-dimensional covariates.

## Appendix

## A. Proofs of the Main Results

For convenience and simplicity, let $C$ denote a positive constant that may have different values at each appearance throughout this paper and $\|A\|$ denote the modulus of the largest singular value of matrix or vector $A$.

Let $\eta_{ij} = X_{ij}^T\beta + Z_{ij}^T \cdot I_q \otimes B(U_{ij})^T\gamma$, then $\mu_{ij} = h(\eta_{ij})$. Let $\eta_i = (\eta_{i1}, \cdots, \eta_{im})^T$, $\mu_i = (\mu_{i1}, \cdots, \mu_{im})^T$, and $\theta = (\beta^T, \gamma^T)^T$, $Y_i = (Y_{i1}, \cdots, Y_{im})^T, X_i = (X_{i1}, \cdots, X_{im})^T$.

Similarly, let $W_{ij} = B(U_{ij}) \otimes I_q \cdot Z_{ij}, P_{ij} = (X_{ij}^T, W_{ij})^T$, and $W_i = (W_{i1}, \cdots, W_{im}))^T, P_i = (P_{i1}, \cdots, P_{im})^T = (X_i, W(U_i))$; then, $\eta_{ij} = P_{ij}^T\theta, \eta_i = P_i\theta$, and $\partial\eta_{ij}/\partial\theta = P_{ij}, \partial\eta_i/\partial\theta = P_i^T$.

Let $h'(t) = dh(t)/dt$, then $\partial\mu_{ij}/\partial\theta = h'(\eta_{ij})P_{ij}$. Let

$$H'(\eta_i) \triangleq \begin{pmatrix} h'(\eta_{i1}) & & \\ & \ddots & \\ & & h'(\eta_{im}) \\ & & \end{pmatrix}, H''(\eta_i) \triangleq \begin{pmatrix} h''(\eta_{i1}) & & \\ & \ddots & \\ & & h''(\eta_{i,}) \\ & & \end{pmatrix}. \tag{A.1}$$

Then,

$$\dot{\mu}_i = \begin{pmatrix} \dfrac{\partial\mu_{i1}}{\partial\beta_1} & \cdots & \dfrac{\partial\mu_{i1}}{\partial\gamma_{qL}} \\ \vdots & \cdots & \vdots \\ \dfrac{\partial\mu_{im}}{\partial\beta_1} & \cdots & \dfrac{\partial\mu_{im}}{\partial\gamma_{qL}} \end{pmatrix} = \begin{pmatrix} \left(\dfrac{\partial\mu_{i1}}{\partial\theta}\right)^T \\ \vdots \\ \left(\dfrac{\partial\mu_{im}}{\partial\theta}\right)^T \end{pmatrix} = \begin{pmatrix} P_{i1}^T h'(\eta_{i1}) \\ \vdots \\ P_{im}^T h'(\eta_{im}) \end{pmatrix}$$
$$= H'(\eta_i)P_i. \tag{A.2}$$

*Proof of Theorem 1.* Let $\delta = n^{-1/2}$, $\beta = \beta_0 + \delta D_1$, $\gamma = \gamma_0 + \delta D_2$, and $D = (D_1^T, D_2^T)^T$. We first show that for any given $\varepsilon > 0$, there exists a large constant $C$ such that

$$P\left\{\inf_{\|D\|=C} \mathcal{Q}_n^P(\beta, \gamma) > \mathcal{Q}_n^P(\beta_0, \gamma_0)\right\} \geq 1 - \varepsilon. \tag{A.3}$$

Note that $\beta_{0l} = 0$, for all $l = P_1 + 1, \cdots, p$, and $\gamma_{0k} = 0$, for all $k = q_1, \cdots, q$, together with Assumption (A1) and $p_\lambda(0) = 0$, we have

$$\begin{aligned} \mathcal{Q}_n^p(\theta) - \mathcal{Q}_n^p(\theta_0) \geqslant {} & [\mathcal{Q}_n(\theta) - \mathcal{Q}_n(\theta_0)] \\ & + n\sum_{l=1}^{p_1} \left[p_{\lambda_2}(|\beta_l|) - p_{\lambda_2}(|\beta_{0l}|)\right] \\ & + n\sum_{k=1}^{q_1} \left[p_{\lambda_1}(\|\gamma_k\|_H) - p_{\lambda_1}(\|\gamma_{0k}\|_H)\right] \\ \triangleq {} & I_1 + I_2 + I_3. \end{aligned} \tag{A.4}$$

By Taylor expansion and Assumption (A4), we have

$$\begin{aligned} I_2 &= n\sum_{l=1}^{p_1} \Big[\delta p'_{\lambda_2}(|\beta_{0l}|) \text{ sgn } (\beta_{0l})|D_{1l}| \\ &\quad + \delta p''_{\lambda_2}(|\beta_{0l}|) \text{ sgn } (\beta_{0l})|D_{1l}|^2\{1 + o(1)\}\Big] \\ &\leqslant \sqrt{p_1}a_n\|D\|O\left(n^{1/2}\right) + b_n\|D\|^2 O(1) \\ &= \sqrt{p_1}\|D\|O\left(n^{-1/2}\right) + \|D\|^2 o(1). \end{aligned} \tag{A.5}$$

Invoking the proof of Theorem 2 in Zhang and Xue [16],

$$\begin{aligned} I_1 &= \mathcal{Q}_n(\theta) - \mathcal{Q}_n(\theta_0) \\ &= D^T\dot{g}_N^T(\theta_0)\Omega_n^{-1}(\theta_0)\dot{g}_N(\theta_0)D \\ &\quad + \|D\|^2 o_p(1) + \|D\|O_p(1). \end{aligned} \tag{A.6}$$

By choosing a sufficient large $C$, $I_1$ dominates $I_2$. Similarly, $I_1$ dominates $I_3$ for a sufficient large $C$. Thus (A.3) holds, i.e., with probability at least $1 - \varepsilon$, there exists a local minimizer $\widehat{\theta}$ that satisfies $\|\widehat{\theta} - \theta_0\| = O_p(\delta)$. Therefore, $\|\widehat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$ and $\|\widehat{\beta} - \beta_0\| = O_p(n^{-1/2})$. Let $R_k(u) = \alpha_k(u) - B(U)^T\gamma_k$ and $\gamma_{ok}$ denote the spline coefficient vector from the spline approximation to $\alpha_k(\cdot)$. From Assumptions (A7) and (A8) and Theorem 12.7 in [18], we get that $\|R_k(u)\| = O(K^{-r})$. Therefore,

$$\begin{aligned}
&\|\alpha\wedge_k(u) - \alpha_{0k}(u)\|^2 \\
&\quad = \int_0^1 \{\alpha\wedge_k(u) - \alpha_{0k}(u)\}^2\,du \\
&\quad = \int_0^1 \left\{B(u)^T\gamma\wedge_k - B(u)^T\gamma_{0k} + R_k(u)\right\}^2 du \\
&\quad \le 2\int_0^1 \left\{B(u)^T\gamma\wedge_k - B(u)^T\gamma_{0k}\right\}^2 du + 2\int_0^1 R_k(u)^2\,du \\
&\quad = 2(\gamma\wedge_k - \gamma_{0k})^T \int_0^1 B(u)B^T(u)\,du\,(\widehat{\gamma}_k - \gamma_{0k}) \\
&\qquad + 2\int_0^1 R_k(u)^2\,du = O_p\left(n^{-2r/(2r+1)}\right).
\end{aligned} \tag{A.7}$$

Thus, we complete the proof of Theorem 1.

*Proof of Theorem 2.* According to Theorem 2, in order to prove the first part of Theorem 2, we need only to prove that, for any $\gamma$ satisfying $\|\gamma - \gamma_0\| = O_p(n^{-1/2})$ and for any $\beta_l$ satisfying $\|\beta_l - \beta_{0l}\| = O_p(n^{-1/2}), l = 1, \cdots, p_1$, there exists a certain $\epsilon = Cn^{-1/2}$ that satisfies, as $n \longrightarrow \infty$, with probability tending to 1:

$$\frac{\partial \mathscr{Q}_n^p(\beta,\gamma)}{\partial \beta_l} > 0, \text{for} \quad 0 < \beta_l < \epsilon, l = p_1 + 1, \cdots, p, \tag{A.8}$$

$$\frac{\partial \mathscr{Q}_n^p(\beta,\gamma)}{\partial \beta_l} < 0, \text{for} \quad -\epsilon < \beta_l < 0, l = p_1 + 1, \cdots, p. \tag{A.9}$$

These imply that the PQIF $\mathscr{Q}_n^p(\beta,\gamma)$ reaches its minimum at $\beta_l = 0, l = p_1 + 1, \cdots, p$.

Following Lemmas 3 and 4 of [16], we have

$$\begin{aligned}
\frac{\partial \mathscr{Q}_n^p(\beta,\gamma)}{\partial \beta_l} &= \frac{\partial g_n^T(\beta,\gamma)}{\partial \beta_l}\Omega_n^{-1}(\beta,\gamma) g_n(\beta,\gamma) + O_p(1) \\
&\quad + np'_{\lambda_2(|\beta_l|)}\ \text{sgn}\ (\beta_l) \\
&= -2\sum_{i=1}^n \begin{pmatrix} \dot{\mu}_i^T A_i^{-1/2} M_1 A_i^{-1/2} \dfrac{\partial \mu_i}{\partial \beta_l} \\ \vdots \\ \dot{\mu}_i^T A_i^{-1/2} M_s A_i^{-1/2} \dfrac{\partial \mu_i}{\partial \beta_l} \end{pmatrix}^T \Omega_n^{-1}(\beta,\gamma) g_n(\beta,\gamma) \\
&\quad + np'_{\lambda_2(|\beta_l|)}\ \text{sgn}\ (\beta_l) + O_p(1) \\
&= n^{1/2}\left[n^{1/2}\lambda_2\left\{\lambda_2^{-1} p'_{\lambda_2}(|\beta_l|)\ \text{sgn}\ (\beta_l)\right\} + O_p(1)\right].
\end{aligned} \tag{A.10}$$

According to (8), the expression of the derivative of SCAD-penalized function, it is easy to see that $\lim_{n\to\infty} \liminf_{\beta_l \to 0} \lambda_2^{-1} p'_{\lambda_2}(|\beta_l|) = 1$. Together with Assumption (A10), $\lambda_2 n^{1/2} > \lambda_{\min} n^{1/2} \longrightarrow \infty$, it is clear that the sign of (A.10) is decided by that of $\beta_l$. This implies (A.8) and (A.9) hold. Thus, we complete the proof of the first part.

Similarly, we can prove that with probability tending to 1, $\widehat{\gamma}_k = 0, k = q_1 + 1, \cdots, q$. Note that $\|B(u)\| = O(1)$ and $\widehat{\alpha}_k(u) = B^T(u)\widehat{\gamma}_k$; the second part of Theorem 2 is proved. Thus, we complete the proof of Theorem 2.

*Proof of Theorem 3.* Let $\theta^* = (\beta^{*T}, \gamma^{*T})^T$ and let $P_i^* = (X_i^{*T}, W_i^{*T})^T, i = 1, \cdots, n$ denote the covariates corresponding to $\theta^*$. Denote $\dot{\mathscr{Q}}_{1n}(\beta,\gamma)$ and $\dot{\mathscr{Q}}_{2n}(\beta,\gamma)$ to be the first derivatives of the PQIF $\mathscr{Q}_n^p$ with respect to $\beta$ and $\gamma$, respectively, i.e.,

$$\begin{aligned}
\dot{\mathscr{Q}}_{1n}(\beta,\gamma) &= \frac{\partial \mathscr{Q}_n^p(\beta,\gamma)}{\partial \beta}, \\
\dot{\mathscr{Q}}_{2n}(\beta,\gamma) &= \frac{\partial \mathscr{Q}_n^p(\beta,\gamma)}{\partial \gamma}.
\end{aligned} \tag{A.11}$$

By Theorems 1 and 2, $(\beta\wedge^{*T}, \mathbf{0}^T)^T$ and $(\gamma\wedge^{*T}, \mathbf{0}^T)^T$ satisfies that

$$\begin{aligned}
\dot{\mathscr{Q}}_{1n}\left(\left(\beta\wedge^{*T}, \mathbf{0}^T\right)^T, \left(\gamma\wedge^{*T}, \mathbf{0}^T\right)^T\right) &= \mathbf{0}^T, \\
\dot{\mathscr{Q}}_{2n}\left(\left(\beta\wedge^{*T}, \mathbf{0}^T\right)^T, \left(\gamma\wedge^{*T}, \mathbf{0}^T\right)^T\right) &= \mathbf{0}^T.
\end{aligned} \tag{A.12}$$

By the Taylor expansion, we have

$$\begin{aligned}
&\mathscr{Q}_{1n}\Big|_{\left(\left(\beta\wedge^{*T},\mathbf{0}^T\right)^T,\left(\gamma\wedge^{*T},\mathbf{0}^T\right)^T\right)} \\
&\quad = \mathscr{Q}_{1n}\Big|_{\left(\left(\beta_0^{*T},\mathbf{0}^T\right)\left(\beta_0^{*T},\mathbf{0}^T\right)^T,\left(\gamma_0^{*T},\mathbf{0}^T\right)^T\right)} \\
&\qquad + \frac{\partial \mathscr{Q}_{1n}}{\partial \beta}\Big|_{\theta=\widetilde{\theta}}\left\{\left(\beta\wedge^{*T}, \mathbf{0}^T\right)^T - \left(\beta_0^{*T}, \mathbf{0}^T\right)^T\right\} \\
&\qquad + \frac{\partial \mathscr{Q}_{1n}}{\partial \gamma}\Big|_{\theta=\widetilde{\theta}}\left\{\left(\gamma\wedge^{*T}, \mathbf{0}^T\right)^T - \left(\gamma_0^{*T}, \mathbf{0}^T\right)^T\right\} \\
&\qquad + \sum_{i=1}^{p_1} np'_{\lambda_2}\left(\widehat{\beta}_l\right)\ \text{sgn}\ \left(\widehat{\beta}_l\right),
\end{aligned} \tag{A.13}$$

where $\widetilde{\theta}$ is between $((\beta_0^{*T}, \mathbf{0}^T)^T, (\gamma_0^{*T}, \mathbf{0}^T)^T)$ and $((\beta\wedge^{*T}, \mathbf{0}^T)^T, (\gamma\wedge^{*T}, \mathbf{0}^T)^T)$. Apply the Taylor expansion to $p'_{\lambda_2}(|\widehat{\beta}_l|)$, we obtain

$$p'_{\lambda_2}\left(\left|\widehat{\beta}_l\right|\right) = p'_{\lambda_2}(|\beta_{0l}|) + \left\{p''_{\lambda_{2l}}(|\beta_{0l}| + o_p(1)\right\}\left(\widehat{\beta}_l - \beta_{0l}\right). \tag{A.14}$$

By Assumption (A10), $p''_{\lambda_2}(|\beta_{0l}|) = o_p(1)$. Note that $p'_{\lambda_{2l}}(|\beta_{0l}|) = 0$ as $\lambda_{\max} \longrightarrow 0$; therefore, by Lemma 4 of [16] and through some calculation, we have

$$\begin{aligned}
\frac{1}{n}\mathcal{Q}_{1n}\Big|_{\left(\left(\beta_0^{*T},\mathbf{0}^T\right)^T,\left(\gamma_0^{*T},\mathbf{0}^T\right)^T\right)}
&= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{s}\sum_{l=1}^{s}\left\{X_i^{*T}H'(\eta_i)A_i^{-1/2}M_kA_i^{-1/2}H'(\eta_i)P_i^{*}\right. \\
&\quad \left.\cdot\,\Omega_{kl}^{-1}P_j^{*T}H'\left(\eta_j\right)A_j^{-1/2}M_lA_j^{-1/2}\left(Y_j-\mu_{0j}\right)\right\} + o_p\left(n^{-1/2}\right) \\
&= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}X_i^{*T}H'(\eta_i)\tau_{ij}\left(Y_j-\mu_{0j}\right) + o_p\left(n^{-1/2}\right) \\
&= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{X}_i^{T}\tau_{ij}\left(\tilde{R}\left(U_j\right)+\varepsilon_j\right) + o_p\left(n^{-1/2}\right),
\end{aligned} \tag{A.15}$$

where $\tilde{X}_i = H'(\eta_i)X_i^{*}$, $\tilde{R}(U_i) = H'(\eta_i)R(U_i)$, $\Omega_{kl}^{-1}$ is the $(l, k)$ block of $\Omega^{-1}$ and

$$\tau_{ij} = \sum_{k=1}^{s}\sum_{l=1}^{s}A_i^{-1/2}M_kA_i^{-1/2}H'(\eta_i)P_i^{*}\Omega_{kl}^{-1}P_j^{*T}H'\left(\eta_j\right)A_j^{-1/2}M_lA_j^{-1/2}. \tag{A.16}$$

Similarly, we have

$$\begin{aligned}
\frac{1}{n}\frac{\partial\mathcal{Q}_{1n}}{\partial\beta}\Big|_{\theta=\tilde{\theta}} &= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{X}_i^{T}\tau_{ij}\tilde{X}_j + o_p\left(n^{-1/2}\right), \\
\frac{1}{n}\frac{\partial\mathcal{Q}_{1n}}{\partial\gamma}\Big|_{\theta=\widetilde{\theta}} &= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{X}_i^{T}\tau_{ij}\tilde{W}\left(U_j\right) + o_p\left(n^{-1/2}\right),
\end{aligned} \tag{A.17}$$

where $\tilde{W}(U_j) = H'(\eta_j)W^{*}(U_j)$, $W^{*}(U_j) = (W_{j1}^{*}, \cdots, W_{jm}^{*})^T$, $W_{ij}^{*} = B(U_{ij}) \otimes I_q \cdot Z_{ij}^{*}$. Hence,

$$\begin{aligned}
\frac{1}{n}\mathcal{Q}_{1n}\Big|_{\left(\left(\beta_0^{*T},\mathbf{0}^T0\right)^T,\left(\gamma_0^{*T},\mathbf{0}^T0\right)^T\right)}
&= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{X}_i^{T}\tau_{ij}\{\tilde{X}_j(\beta_0^{*}-\beta\wedge^{*}) + \tilde{W}\left(U_j\right)\cdot(\gamma_0^{*}-\gamma\wedge^{*}) \\
&\quad + \tilde{R}\left(U_j\right) + \varepsilon_j\} + o_p(\beta\wedge^{*}-\beta_0^{*}), \\
\frac{1}{n}\mathcal{Q}_{2n}\Big|_{\left(\left(\beta_0^{*T},\mathbf{0}^T0\right)^T,\left(\gamma_0^{*T},\mathbf{0}^T0\right)^T\right)}
&= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{W}(U_i)^T\tau_{ij}\{\tilde{X}_j(\beta_0^{*}-\beta\wedge^{*}) \\
&\quad + \tilde{W}\left(U_j\right)\cdot(\gamma_0^{*}-\gamma\wedge^{*}) + \tilde{R}\left(U_j\right) + \varepsilon_j\} + o_p(\gamma\wedge^{*}-\gamma_0^{*}).
\end{aligned} \tag{A.18}$$

Following the proof of Theorem 2 in [16], we prove (16). Thus, we complete the proof of Theorem 3.

## Data Availability

The data can be downloaded from https://content.sph.harvard.edu/xlin/dat/indon.dat.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The research is funded by the National Natural Science Foundation of China (11571025) and the Beijing Natural Science Foundation (1182008). This support is greatly appreciated.

## Supplementary Materials

The R code presented in Word format for the real data analysis is included in the supplementary file. *(Supplementary Materials)*

## References

[1] L. Breiman, “Better subset regression using the nonnegative garrote,” *Techonometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[2] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” *Journal of Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[3] W. J. Fu, “Penalized Regressions: the bridge versus the LASSO,” *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[4] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[5] H. Zhou and R. Li, “One-step sparse estimates in nonconcave penalized likelihood models,” *Annals of Statistics*, vol. 36, pp. 1509–1533, 2007.

[6] J. FAN and R. Li, “New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis,” *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 710–723, 2004.

[7] J. FAN and W. Zhang, “Statistical methods with varying coefficient models,” *Statistics and Its Interface*, vol. 1, no. 1, pp. 179–195, 2008.

[8] P. X. Zhao and L. G. Xue, “Variable selection for semiparametric varying coefficient partially linear models,” *Statistics & Probability Letters*, vol. 79, no. 20, pp. 2148–2157, 2009.

[9] L. Xue, A. Qu, and J. Zhou, “Consistent model selection for marginal generalized additive model for correlated data,” *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1518–1530, 2010.

[10] A. Qu, B. G. Lindsay, and B. Li, “Improving generalised estimating equations using quadratic inference functions,” *Biometrika*, vol. 87, no. 4, pp. 823–836, 2000.

[11] K. L. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[12] A. Qu and R. Li, "Quadratic inference functions for varying coefficient models with longitudinal data," *Biometrics*, vol. 62, no. 2, pp. 379–391, 2006.

[13] L. Wang, H. Li, and J. Z. Huang, "Variable selection in nonparametric varying coefficient models for analysis of repeated measurements," *Journal of American Statistical Association*, vol. 103, pp. 1556–1569, 2008.

[14] Y. Bai, Z. Y. Zhu, and W. K. Fung, "Partial linear models for longitudinal data based on quadratic inference functions," *Scandinavian Journal of Statistics*, vol. 35, no. 1, pp. 104–118, 2008.

[15] R. Q. Tian, L. G. Xue, and C. L. Liu, "Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data," *Journal of Multivariate Analysis*, vol. 132, pp. 94–110, 2014.

[16] J. H. Zhang and L. G. Xue, "Quadratic inference functions for generalized partially models with longitudinal data," *Chinese Journal of Applied Probability and Statistics*, vol. 33, pp. 417–432, 2017.

[17] R. Li and H. Liang, "Variable selection in semiparametric regression modeling," *The Annals of Statistics*, vol. 36, no. 1, pp. 261–286, 2008.

[18] G. Schumaker, *Spline Function*, Wiley, New York, NY, USA, 1981.

[19] H. Wang, R. Li, and C. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[20] H. Zhou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[21] H. S. Wang and Y. C. Xia, "Shrinkage estimation of the varying coefficient model," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 747–757, 2009.

[22] S. D. Oman, "Easily simulated multivariate binary distributions with given positive and negative correlations," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 999–1005, 2009.

[23] S. L. Zeger and M. R. Karim, "Generalized linear models with random effects: a Gibbs sampling approach," *Journal of the American Statistical Association*, vol. 86, pp. 79–86, 1991.

[24] P. J. Diggle, K. Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, Oxford University Press, Oxford, England, 1994.

[25] X. H. Lin and R. J. Carroll, "Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of the American Statistical Association*, vol. 95, pp. 520–534, 2000.

[26] X. H. Lin and R. J. Carroll, "Semiparametric regression for clustered data using generalized estimating equations," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1045–1056, 2001.

[27] X. He, W. Fung, and Z. Zhu, "Robust estimation in generalized Partial linear Models for Clustered data," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1176–1184, 2005.