Bernoulli **19**(1), 2013, 252–274
DOI: 10.3150/11-BEJ386

Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data

SHUJIE MA[1], QIONGXIA SONG[2] and LI WANG[3]

[1]*Department of Statistics, UC-Riverside, Riverside, CA 92521, USA. E-mail: shujie.ma@ucr.edu*
[2]*Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: song@utdallas.edu*
[3]*Department of Statistics, The University of Georgia, Athens, GA 30602, USA. E-mail: lilywang@uga.edu*

We consider the problem of simultaneous variable selection and estimation in additive, partially linear models for longitudinal/clustered data. We propose an estimation procedure via polynomial splines to estimate the nonparametric components and apply proper penalty functions to achieve sparsity in the linear part. Under reasonable conditions, we obtain the asymptotic normality of the estimators for the linear components and the consistency of the estimators for the nonparametric components. We further demonstrate that, with proper choice of the regularization parameter, the penalized estimators of the non-zero coefficients achieve the asymptotic oracle property. The finite sample behavior of the penalized estimators is evaluated with simulation studies and illustrated by a longitudinal CD4 cell count data set.

Keywords: additive partially linear model; clustered data; longitudinal data; model selection; penalized least squares; spline

1. Introduction

In the past two decades, there has been a considerable amount of research to study additive, partially linear models (APLM); see Opsomer and Ruppert [27], Härdle, Liang and Gao [12], Li [15], Fan and Li [9], Liang *et al.* [18], Liu, Wang and Liang [21], Ma and Yang [24], among others. APLMs meet three fundamental aspects (Stone [29]) of statistical models: flexibility, dimensionality and interpretability. In this paper, we consider the APLMs for clustered and longitudinal data.

Let $\{(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}), 1 \le i \le n, 1 \le j \le m_i\}$ be the jth observation for the ith subject or cluster, where Y_{ij} is the response variable, $\mathbf{X}_{ij} = (1, X_{ij1}, \ldots, X_{ij(d_1-1)})^{\mathrm{T}}$ is a d_1-vector of covariates, and $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijd_2})^{\mathrm{T}}$ is a d_2-vector of covariates. An APLM for this kind of data is given by

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{l=1}^{d_2} \eta_l(Z_{ijl}) + \varepsilon_{ij}, \qquad j = 1, \ldots, m_i, i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is a d_1-dimensional regression parameter, and $\eta_l, l = 1, \ldots, d_2$, are unknown but smooth functions. We assume $\underline{\boldsymbol{\varepsilon}}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im_i})^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$. For identifiability, both the parametric

1350-7265

and nonparametric components must be centered, that is, $E\eta_l(Z_{ijl}) \equiv 0$, $l = 1, \ldots, d_2$, $EX_{ijk} = 0$, $k = 1, \ldots, d_1$. When $d_2 = 1$, model (1) is simplified to be the partially linear model (PLM) in Lin and Carroll [20]. Model (1) retains the merits of additive models, while it is more flexible than purely additive models by allowing a subset of the covariates to be discrete and/or unbounded. When m_is and $\mathbf{\Sigma}_i$s are the same for all individuals, Carroll *et al.* [3] considered the efficient estimation of $\boldsymbol{\beta}$ in model (1) using local linear smooth backfitting. In this paper we consider a more general scenario that both m_i and $\mathbf{\Sigma}_i$ may vary across subjects or experimental units to allow irregular measurements for individuals. Our goal is to simultaneously select significant variables and efficiently estimate the unknown components for model (1). This is challenging due to the issue of "curse of dimensionality" and the additional complexity of the correlation structures (Wang [34]) introduced by repeated measurements.

To alleviate the effect of the "curse of dimensionality," more parsimonious models become desirable in practice; see Fan [10], Hall, Müller and Wang [11] and Wang *et al.* [32]. Variable selection is fundamental to high-dimensional statistical modeling. In the absence of prior knowledge, a large number of variables may be included at the initial stage of modeling in order to reduce possible model bias. This may lead to a complicated model including many insignificant variables, resulting in less predictive powers and difficulty in interpretation. There is an extensive literature on variable selection via various approaches, for example, the classical information criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) in Yang [40], the least absolute shrinkage and selection operator (LASSO) proposed in Tibshirani [30,31], the non-negative garrote in Yuan and Liu [41], the difference convex algorithm in Wu and Liu [36], the combination of L_0 and L_1 penalties in Liu and Wu [22], and the nonparametric independence screening procedure in Fan, Feng and Song [6].

Many traditional variable selection procedures in use, including stepwise selection, AIC or BIC, can be expensive in computation and ignore stochastic errors inherited in the variable selection process. Penalized least squares approaches have gained popularity in recent years to automatically and simultaneously select significant variables; for example, Antoniadis [1] proposed the hard thresholding penalty which enables best subset selection and stepwise deletion in certain cases. The LASSO (Tibshirani [30,31]) is one of the most popular shrinkage estimators, but it has some deficiencies (Meinshausen and Bühlmann [26]). Fan and Li [7] proposed the smoothly clipped absolute deviation penalty (SCAD), which achieves an "oracle" property in the sense that it performs as well as if the subset of significant variables were known in advance. The SCAD-penalized selection procedures were illustrated in Fan and Li [7] for parametric models; Cai *et al.* [2] and Fan and Li [8] for survival models; Li and Liang [16] for generalized varying-coefficient models; Liang and Li [17] and Ma and Li [25] for measurement error models; Xue [37] for pure additive models; and Xue, Qu and Zhou [38] for generalized additive models with correlated data.

We propose a model selection method for APLMs with repeated measures by penalizing appropriate estimating functions. We approximate nonparametric components by spline functions and obtain asymptotic normality for the coefficient estimators via one step least squares. The proposed approach is computationally expedient and easy to implement, in contrast to the backfitting approach in Carroll *et al.* [3]. Moreover, it avoids the pitfall of the backfitting algorithms caused by dependence between covariates. Furthermore, we show that the estimator can correctly select the nonzero coefficients with probability converging to 1 and the $\sqrt{n}$-consistent estimators of the non-zero coefficients can perform as well as an oracle estimator in the sense of Fan and Li [7] with a suitable choice of penalty function.

The paper is organized as follows. In Section 2, we introduce the penalized polynomial spline estimating method. Section 3 provides the asymptotic properties of the proposed estimators, including the consistency and oracle property of the parametric components, as well as the rate of the L_2-convergence of the nonparametric components. In Section 4, we discuss some implementation issues of the proposed procedure. Simulation studies are presented in Section 5. Section 6 illustrates the application using longitudinal CD4 cell-count data. We conclude with a discussion in Section 7. Technical proofs are presented in the Appendix.

2. Penalized spline estimation

For simplicity, denote vectors $\underline{\mathbf{Y}}_i = (Y_{i1}, \dots, Y_{im_i})^{\mathrm{T}}$ and $\underline{\boldsymbol{\mu}}_i = (\mu_{i1}, \dots, \mu_{im_i})^{\mathrm{T}}$, $1 \le m_i \le M$, $1 \le i \le n$. Similarly, let $\underline{\mathbf{X}}_i = \{(\mathbf{X}_{i1}, \dots, \mathbf{X}_{im_i})^{\mathrm{T}}\}_{m_i \times d_1}$ and $\underline{\mathbf{Z}}_i = \{(\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{im_i})^{\mathrm{T}}\}_{m_i \times d_2}$. Assume that Z_{ijl} has the same distribution as Z_l, which is distributed on a compact interval $[a_l, b_l]$, $1 \le l \le d_2$, and, without loss of generality, we take all intervals $[a_l, b_l] = [0, 1]$, $1 \le l \le d_2$. Let $\eta_l(\mathbf{Z}_{il}) = \{\eta_l(Z_{i1l}), \dots, \eta_l(Z_{im_il})\}^{\mathrm{T}}$, for $l = 1, \dots, d_2$. The mean function in model (1) can be written in matrix notation as $\underline{\boldsymbol{\mu}}_i = \underline{\boldsymbol{X}}_i \boldsymbol{\beta} + \sum_{l=1}^{d_2} \eta_l(\mathbf{Z}_{il})$, which is a semiparametric extension of the marginal model in Liang and Zeger [19] with an identity link.

As in Wang, Carroll and Lin [35], we allow $\mathbf{X}$ and $\mathbf{Z}$ to be dependent. Let $\mathbf{V}_i = \mathbf{V}_i(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$ be the assumed "working" covariance of $\underline{\mathbf{Y}}_i$, where $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}$, $\mathbf{A}_i$ denotes a $m_i \times m_i$ diagonal matrix that contains the marginal variances of Y_{ij}, and $\mathbf{R}_i$ is an invertible working correlation matrix. Throughout, we assume that $\mathbf{V}_i$ depends on a nuisance finite dimensional parameter vector $\boldsymbol{\alpha}$.

Following Wang and Yang [33], we approximate the nonparametric functions η_l's by polynomial splines. Let G_n be the space of polynomial splines of degree $q \ge 1$. We introduce a sequence of spline knots

$$t_{-q} = \cdots = t_{-1} = t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1} = \cdots = t_{N+q+1},$$

where $N \equiv N_n$ is the number of interior knots, and N increases when sample size n increases with the precise order given in Assumption (A5). Then G_n consists of functions ϖ satisfying (i) ϖ is a polynomial of degree q on each of the subintervals $I_s = [t_s, t_{s+1})$, $s = 0, \dots, N_n - 1$, $I_{Nn} = [t_{N_n}, 1]$; (ii) for $q \ge 1$, ϖ is $(q-1)$ times continuously differentiable on [0, 1]. In the following, let $J_n = N_n + q + 1$, and we adopt the normalized B-spline space $G_n^0 = \{B_{s,l}\colon 1 \le l \le d_2, 1 \le s \le J_n\}^{\mathrm{T}}$ in Xue and Yang [39]. Equally spaced knots are used in this article for simplicity of proof. However, other regular knot sequences can also be used with similar asymptotic results.

Suppose that η_l can be approximated well by a spline function in G_n^0 so that

$$\eta_l(z_l) \approx \widetilde{\eta}_l(z_l) = \sum_{s=1}^{J_n} \gamma_{sl} B_{s,l}(z_l). \tag{2}$$

Let $\boldsymbol{\gamma} = (\gamma_{sl} : 1 \le s \le J_n, 1 \le l \le d_2)^{\mathrm{T}}$ be the collection of the coefficients in (2), and let

$$\mathbf{B}_{ijl} = [\{B_{s,l}(Z_{ijl}) \colon 1 \le s \le J_n\}^{\mathrm{T}}]_{J_n \times 1}, \qquad \mathbf{B}_{ij} = \{(\mathbf{B}_{ij1}^{\mathrm{T}}, \dots, \mathbf{B}_{ijd_2}^{\mathrm{T}})^{\mathrm{T}}\}_{d_2 J_n \times 1}; \tag{3}$$

then we have an approximation $\mu_{ij} \approx \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}$. We can also write the approximation in matrix notation as $\underline{\boldsymbol{\mu}}_i \approx \underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma}$, where $\underline{\mathbf{B}}_i = \{(\mathbf{B}_{i1}, \dots, \mathbf{B}_{im_i})^{\mathrm{T}}\}_{m_i \times d_2 J_n}$.

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \dots, \widehat{\beta}_{d_1})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\gamma}} = \{\widehat{\gamma}_{sl} \colon s = 1, \dots, J_n, l = 1, \dots, d_2\}^{\mathrm{T}}$ be the minimizer of

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\{\underline{\mathbf{Y}}_i - (\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma})\}^{\mathrm{T}}\mathbf{V}_i^{-1}\{\underline{\mathbf{Y}}_i - (\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma})\}, \tag{4}$$

which is corresponding to the class of working covariance matrices $\{\mathbf{V}_i, 1 \le i \le n\}$, or, equivalently, they solve the estimating equations

$$\sum_{i=1}^{n}\underline{\mathbf{X}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\{\underline{\mathbf{Y}}_i - (\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma})\} = \mathbf{0}, \tag{5}$$

$$\sum_{i=1}^{n}\underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\{\underline{\mathbf{Y}}_i - (\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma})\} = \mathbf{0}. \tag{6}$$

Solving (6) yields

$$\boldsymbol{\gamma} \equiv \boldsymbol{\gamma}(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n}\underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\underline{\mathbf{B}}_i\right)^{-1}\sum_{i=1}^{n}\underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\underline{\mathbf{Y}}_i - \underline{\mathbf{X}}_i\boldsymbol{\beta}). \tag{7}$$

Replacing $\boldsymbol{\gamma}$ by $\boldsymbol{\gamma}(\boldsymbol{\beta})$ in (4), we define

$$\begin{aligned} Q(\boldsymbol{\beta}) \equiv Q_n\{\boldsymbol{\beta}, \boldsymbol{\gamma}(\boldsymbol{\beta})\} = \frac{1}{2}\sum_{i=1}^{n}[\underline{\mathbf{Y}}_i - \{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\boldsymbol{\gamma}(\boldsymbol{\beta})\}]^{\mathrm{T}} \\ \times \mathbf{V}_i^{-1}[\underline{\mathbf{Y}}_i - \{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{B}}_i\boldsymbol{\gamma}(\boldsymbol{\beta})\}]. \end{aligned} \tag{8}$$

To select the significant parametric components, we add a penalty to $Q(\boldsymbol{\beta})$. Let $n_{\mathrm{T}} = \sum_{i=1}^{n} m_i$, and define the penalized version of $Q(\boldsymbol{\beta})$ as

$$Q_{\mathcal{P}}(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + n_{\mathrm{T}}\mathcal{P}(\boldsymbol{\beta}), \tag{9}$$

where $\mathcal{P}(\boldsymbol{\beta}) = \sum_{k=1}^{d_1} p_{\lambda_k}(|\beta_k|)$ for a pre-specified penalty function $p_\lambda(|\beta|)$ with a regularization parameter λ. Minimizing $Q_{\mathcal{P}}(\boldsymbol{\beta})$ in (9) yields a penalized estimator

$$\widehat{\boldsymbol{\beta}}^{\mathrm{P}} = \arg\min Q_{\mathcal{P}}(\boldsymbol{\beta}). \tag{10}$$

Various penalty functions can be used for $\mathcal{P}(\boldsymbol{\beta})$ in variable selection procedures. We consider two penalty functions, the hard thresholding penalty (Antoniadis [1]) $p_\lambda(\beta) = \lambda^2 - (|\beta| -$

$\lambda)^2 I(|\beta| < \lambda)$and the SCAD penalty (Fan and Li [7]), given by

$$p'_\lambda(\beta) = \lambda\left\{I(\beta \le \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda)\right\} \qquad \text{for some } a > 2 \text{ and } \beta > 0,$$

where $p_\lambda(0) = 0$, and λ and a are two tuning parameters. Justifying from a Bayesian statistical point of view, Fan and Li [7] suggested using $a = 3.7$, which will be used in our simulation studies.

The minimization problem in (10) is essentially a one-step least squares problem, which can be easily solved and implemented with many existing regression programs. The theorems established in Section 3.3 demonstrate that $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$ performs asymptotically as well as an oracle estimator in terms of selecting the correct model when the regularization parameter is appropriately chosen.

3. Asymptotic properties of the estimators

For positive numbers a_n and b_n, $n \ge 1$, let $a_n \sim b_n$ denote that $\lim_{n\to\infty} a_n/b_n = c$, where c is some non-zero constant. Let $|\phi|_{L_2} \equiv [\int_0^1 \{\phi(z)\}^2 \,\mathrm{d}z]^{1/2}$ denote the L_2 norm of any square integrable function $\phi(z)$ on $[0, 1]$. Denote the space of the pth order smooth functions as $C^{(p)}[0,1] = \{\phi \mid \phi^{(p)} \in C[0,1]\}$.

3.1. Assumptions

The assumptions for the asymptotic results are listed below:

(A1) The random variables Z_{ijl} are bounded, uniformly in $1 \le j \le m_i$, $1 \le i \le n$, $1 \le l \le d_2$. The marginal density $f_l(z_l)$ of Z_l has the uniform upper bound C_f and lower bound c_f on $[0, 1]$. The joint density $f_{ll'}(z_l, z_{l'})$ of $(Z_{ijl}, Z_{ijl'})$ satisfies that $c_f \le f_{ll'}(z_l, z_{l'}) \le C_f$, for all $(z_l, z_{l'}) \in [0,1]^2$, $1 \le l \ne l' \le d_2$.

(A2) The random variables X_{ijk} are bounded, uniformly in $1 \le j \le m_i$, $1 \le i \le n$, $1 \le k \le d_1$. The eigenvalues of $E\{\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|\mathbf{Z}_{ij}\}$ are bounded away from 0 and infinity, uniformly in $1 \le j \le m_i$, $1 \le i \le n$.

(A3) The eigenvalues of the true covariance matrices $\boldsymbol{\Sigma}_i$ are bounded away from 0 and infinity, uniformly in $1 \le i \le n$.

(A4) The eigenvalues of the working covariance matrices $\mathbf{V}_i$ are bounded away from 0 and infinity, uniformly in $1 \le i \le n$.

To make $\boldsymbol{\beta}$ estimable at the $\sqrt{n}$ rate, we need a condition to ensure that $\mathbf{X}$ and $\mathbf{Z}$ not functionally related. Define $\mathcal{H} = \{\psi(\mathbf{z}) = \sum_{l=1}^{d_2} \psi_l(z_l), E\psi_l(z_l) = 0, |\psi_l|_{L_2} < \infty\}$ the Hilbert space of theoretically centered L_2 additive functions on $[0,1]^{d_2}$. Let ψ_k^* be the function $\psi \in \mathcal{H}$ that minimizes

$$\sum_{i=1}^{n} E\big[\big\{\underline{\mathbf{X}}_i^{(k)} - \psi(\underline{\mathbf{Z}}_i)\big\}^{\mathrm{T}} \mathbf{V}_i^{-1} \big\{\underline{\mathbf{X}}_i^{(k)} - \psi(\underline{\mathbf{Z}}_i)\big\}\big],$$

where

$$\underline{\mathbf{X}}_i^{(k)} = (X_{i1k}, \dots, X_{im_ik})^{\mathrm{T}}, \qquad 1 \le k \le d_1. \tag{11}$$

Then

(A5) for $1 \le l \le d_2$, $1 \le k \le d_1$, assume that $\eta_l(z_l) \in C^{(p)}[0,1]$, $\psi_k^* \in C^{(p)}[0,1]$ for a given integer $p \ge 1$, and the spline degree satisfies $q + 1 \ge p$. The number of the spline basis functions $J_n \sim n^{1/(2p)} \log(n)$.

Assumptions (A1)–(A4) are identical with (C1)–(C4) in Huang, Zhang and Zhou [14], while Assumption (A5) is similar to (C1) and (C4) in Liu, Wang and Liang [21].

3.2. Asymptotic properties for the unpenalized estimators

According to the equations in (5) and (6), we have

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \left(\sum_{i=1}^{n} \underline{\mathbf{D}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{D}}_i \right)^{-1} \left(\sum_{i=1}^{n} \underline{\mathbf{D}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{Y}}_i \right), \tag{12}$$

where $\underline{\mathbf{D}}_i = (\underline{\mathbf{X}}_i, \underline{\mathbf{B}}_i)_{m_i \times (d_1 + d_2 J_n)}$. The centered additive component $\eta_l(z_l)$ is estimated by the empirically centered estimator

$$\widehat{\eta}_l(z_l) = \sum_{s=1}^{J_n} \widehat{\gamma}_{sl} B_{s,l}(z_l) - n_{\mathrm{T}}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} B_{s,l}(Z_{ijl}). \tag{13}$$

Next we derive the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}_l$. Let $\mathbb{X}$ and $\mathbb{Z}$ be the collections of all X_{ijk}s and Z_{ijl}s, respectively, that is, $\mathbb{X}_{n_{\mathrm{T}} \times d_1} = (\underline{\mathbf{X}}_1^{\mathrm{T}}, \dots, \underline{\mathbf{X}}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbb{Z}_{n_{\mathrm{T}} \times d_2} = (\underline{\mathbf{Z}}_1^{\mathrm{T}}, \dots, \underline{\mathbf{Z}}_n^{\mathrm{T}})^{\mathrm{T}}$. Define

$$\widetilde{\underline{\mathbf{X}}}_i^{(k)} = \underline{\mathbf{X}}_i^{(k)} - \psi_k^*(\underline{\mathbf{Z}}_i), \qquad 1 \le k \le d_1, \qquad \widetilde{\underline{\mathbf{X}}}_i = \big(\widetilde{\underline{X}}_i^{(1)}, \dots, \widetilde{\underline{\mathbf{X}}}_i^{(d_1)}\big)_{m_i \times d_1}, \tag{14}$$

for $1 \le i \le n$. Denote $\widetilde{\mathbb{X}} = \{(\widetilde{\underline{\mathbf{X}}}_1^{\mathrm{T}}, \dots, \widetilde{\underline{\mathbf{X}}}_n^{\mathrm{T}})^{\mathrm{T}}\}_{n_{\mathrm{T}} \times d_1}$,

$$\mathbb{V}^{-1} = \operatorname{diag}(\mathbf{V}_1^{-1}, \dots, \mathbf{V}_n^{-1})_{n_{\mathrm{T}} \times n_{\mathrm{T}}}, \qquad \mathbb{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_n)_{n_{\mathrm{T}} \times n_{\mathrm{T}}}.$$

Further define

$$\boldsymbol{\Omega}(\mathbb{V}, \mathbb{\Sigma}) = \{\widetilde{\mathbf{A}}(\mathbb{V})\}^{-1} \widetilde{\mathbf{B}}(\mathbb{V}, \mathbb{\Sigma}) \{\widetilde{\mathbf{A}}(\mathbb{V})\}^{-1} \tag{15}$$

with $\widetilde{\mathbf{A}}(\mathbb{V}) = E(n^{-1} \widetilde{\mathbb{X}}^{\mathrm{T}} \mathbb{V}^{-1} \widetilde{\mathbb{X}})$ and $\widetilde{\mathbf{B}}(\mathbb{V}, \mathbb{\Sigma}) = E(n^{-1} \widetilde{\mathbb{X}}^{\mathrm{T}} \mathbb{V}^{-1} \mathbb{\Sigma} \mathbb{V}^{-1} \widetilde{\mathbb{X}})$.

The following result gives the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ for general working covariance matrices.

Theorem 1. *Under Assumptions* (A1)–(A5), *as* $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathrm{N}(0, \boldsymbol{\Omega}(\mathbb{V}, \mathbb{\Sigma})).$$

Remark 1. It is easy to show that the covariance $\mathbf{\Omega}(\mathbb{V},\mathbb{\Sigma})$ in (15) is minimized by $\mathbb{V}=\mathbb{\Sigma}$, and in this case equals to $\{\widetilde{\mathbf{A}}(\mathbb{V})\}^{-1}$. To construct the confidence sets for $\boldsymbol{\beta}$, $\mathbf{\Omega}(\mathbb{V},\mathbb{\Sigma})$ is consistently estimated by

$$\widehat{\mathbf{\Omega}}(\mathbb{V},\widehat{\mathbb{\Sigma}})=n(\widehat{\mathbb{X}}^{\mathrm{T}}\mathbb{V}^{-1}\widehat{\mathbb{X}})^{-1}(\widehat{\mathbb{X}}^{\mathrm{T}}\mathbb{V}^{-1}\widehat{\mathbb{\Sigma}}\mathbb{V}^{-1}\widehat{\mathbb{X}})(\widehat{\mathbb{X}}^{\mathrm{T}}\mathbb{V}^{-1}\widehat{\mathbb{X}})^{-1},$$

where $\widehat{\mathbb{X}}=\{(\widehat{\underline{\mathbf{X}}}_1^{\mathrm{T}},\dots,\widehat{\underline{\mathbf{X}}}_n^{\mathrm{T}})^{\mathrm{T}}\}_{n_{\mathrm{T}}\times d_1}$, and

$$\widehat{\underline{\mathbf{X}}}_i=\underline{\mathbf{X}}_i-\mathrm{Proj}_{G_n^*}\underline{\mathbf{X}}_i,\qquad i=1,\dots,n, \tag{16}$$

in which $\mathrm{Proj}_{G_n^*}$ is the projection onto the empirically centered spline space.

Remark 2. The result of Proposition 2 with identity link in Wang, Carroll and Lin [35] is a special case of Theorem 1 with $\mathbf{V}_1=\cdots=\mathbf{V}_n=\mathbf{V}$, $m_1=\cdots=m_n=M$ and $d_2=1$.

The next theorem shows that the estimated function $\widehat{\eta}_l$ in (13) is L_2-consistent.

Theorem 2. *Under Assumptions* (A1)–(A5), $|\widehat{\eta}_l-\eta_l|_{L_2}^2=\mathrm{O}_P\{J_n^{1-2p}+(J_n/n)\}$, *for* $1\le l\le d_2$.

3.3. Sampling properties for the penalized estimators

We next show that with a proper choice of λ_k, the penalized estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$ has an oracle property. To avoid confusion, let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Let r be the number of non-zero components of $\boldsymbol{\beta}_0$. Let $\boldsymbol{\beta}_0=(\beta_{10},\dots,\beta_{d_10})^{\mathrm{T}}=(\boldsymbol{\beta}_{10}^{\mathrm{T}},\boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{10}$ is assumed to consist of all r non-zero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{20}=\mathbf{0}$ without loss of generality. In a similar fashion to $\boldsymbol{\beta}$, we can write the collections of all parametric components, $\mathbb{X}=(\mathbb{X}_1^{\mathrm{T}},\mathbb{X}_2^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\mathbb{X}}=(\widetilde{\mathbb{X}}_1^{\mathrm{T}},\widetilde{\mathbb{X}}_2^{\mathrm{T}})^{\mathrm{T}}$. Denote $a_n=\max_{1\le k\le d_1}\{|p'_{\lambda_k}(|\beta_{k0}|)|,\beta_{k0}\neq 0\}$, $w_n=\max_{1\le k\le d_1}\{|p''_{\lambda_k}(|\beta_{k0}|)|,\beta_{k0}\neq 0\}$.

Theorem 3. *Under Assumptions* (A1)–(A5), *and if* $a_n\to 0$ *and* $w_n\to 0$ *as* $n\to\infty$, *then there exists a local solution* $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$ *in* (10) *such that its rate of convergence is* $\mathrm{O}_P(n^{-1/2}+a_n)$.

Next define a vector $\boldsymbol{\kappa}_n=\{p'_{\lambda_1}(|\beta_{10}|)\,\mathrm{sgn}(\beta_{10}),\dots,p'_{\lambda_r}(|\beta_{r0}|)\,\mathrm{sgn}(\beta_{r0})\}^{\mathrm{T}}$ and a diagonal matrix $\mathbf{\Sigma}_\lambda=\mathrm{diag}\{p''_{\lambda_1}(|\beta_{10}|),\dots,p''_{\lambda_r}(|\beta_{r0}|)\}$. We further denote $\mathbf{\Sigma}_{1i}=\mathrm{Var}(\underline{\mathbf{Y}}_i|\underline{\mathbf{X}}_{1i},\underline{\mathbf{Z}}_i)$, $\mathbb{\Sigma}_1=\mathrm{diag}(\mathbf{\Sigma}_{11},\dots,\mathbf{\Sigma}_{1n})$, $\widetilde{\mathbf{A}}_1(\mathbb{V})=E(\widetilde{\mathbb{X}}_1^{\mathrm{T}}\mathbb{V}^{-1}\widetilde{\mathbb{X}}_1)$ and $\widetilde{\mathbf{B}}_1(\mathbb{V},\mathbb{\Sigma}_1)=E(\widetilde{\mathbb{X}}_1^{\mathrm{T}}\mathbb{V}^{-1}\mathbb{\Sigma}_1\mathbb{V}^{-1}\widetilde{\mathbb{X}}_1)$.

The theorem below shows that under regularity conditions, all the covariates with zero coefficients can be detected simultaneously with probability tending to 1, and the estimators of all the non-zero coefficients are asymptotically normally distributed.

Theorem 4. *Under Assumptions* (A1)–(A5), *if* $\lim_{n\to\infty}\sqrt{n}\lambda_{kn}\to\infty$ *and*

$$\liminf_{n\to\infty}\liminf_{\beta_k\to 0^+}\lambda_{kn}^{-1}p'_{\lambda_{kn}}(|\beta_k|)>0,$$

then the $\sqrt{n}$*-consistent estimator* $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$ *in Theorem* 3 *satisfies* $P(\widehat{\boldsymbol{\beta}}_2^{\mathrm{P}}=\mathbf{0})\to 1$, *as* $n\to\infty$, *and*

$$\sqrt{n}\{\widetilde{\mathbf{A}}_1(\mathbb{V})+\mathbf{\Sigma}_\lambda\}[\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}}-\boldsymbol{\beta}_{10}+\{\widetilde{\mathbf{A}}_1(\mathbb{V})+\mathbf{\Sigma}_\lambda\}^{-1}\boldsymbol{\kappa}_n]\to \mathrm{N}(\mathbf{0},\widetilde{\mathbf{B}}_1(\mathbb{V},\mathbb{\Sigma}_1)).$$

4. Implementation

In this section, we illustrate how to implement the proposed method in the semiparametric marginal estimation and variable selection. Let

$$\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \operatorname{diag}\left\{\frac{p'_{\lambda_1}(|\beta_1|)}{\epsilon + |\beta_1|}, \ldots, \frac{p'_{\lambda_{d_1}}(|\beta_{d_1}|)}{\epsilon + |\beta_{d_1}|}\right\}$$

for a small number ϵ ($\epsilon = 10^{-6}$ in our simulation studies). Applying the usual Taylor approximation, $Q_{\mathcal{P}}(\boldsymbol{\beta})$ can be locally approximated by

$$Q(\boldsymbol{\beta}) + \dot{Q}(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \ddot{Q}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tfrac{1}{2} n_{\mathrm{T}} \boldsymbol{\beta}^T \Sigma_{\boldsymbol{\lambda}}(\boldsymbol{\beta}_0)\boldsymbol{\beta}.$$

By the local quadratic approximations for penalty functions (Fan and Li [7], Section 3.3), the solution can be found iteratively,

$$\boldsymbol{\beta}^{(k+1)} = \left[\sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \{\mathbf{V}_i^{(k)}\}^{-1} \widehat{\underline{\mathbf{X}}}_i + n_{\mathrm{T}} \Sigma_{\boldsymbol{\lambda}}\{\boldsymbol{\beta}^{(k)}\}\right]^{-1} \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \{\mathbf{V}_i^{(k)}\}^{-1} \{\underline{\mathbf{Y}}_i - \widehat{\Pi}_n \underline{\mathbf{Y}}_i\},$$

where $\widehat{\Pi}_n \underline{\mathbf{Y}}_i$ is the projection of $\underline{\mathbf{Y}}_i$ onto the spline space G_n^0, and $\widehat{\underline{\mathbf{X}}}_i$ is given in (16).

Following Fan and Li [7], we derive a sandwich formula for the standard errors of the estimated covariates $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$

$$\begin{aligned}\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}^{\mathrm{P}}) &= \{\ddot{Q}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}}) + n_{\mathrm{T}} \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}})\}^{-1} \widehat{\mathrm{Cov}}\{\dot{Q}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}})\} \\ &\quad \times \{\ddot{Q}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}}) + n_{\mathrm{T}} \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}})\}^{-1},\end{aligned} \tag{17}$$

where $\ddot{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \widehat{\underline{\mathbf{X}}}_i$ and $\widehat{\mathrm{Cov}}\{\dot{Q}(\boldsymbol{\beta})\} = \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \widehat{\boldsymbol{\Sigma}}_i \mathbf{V}_i^{-1} \widehat{\underline{\mathbf{X}}}_i$. Applying conventional techniques that arise in the likelihood setting, we can show that the above sandwich formula is a consistent estimator and has good accuracy in our simulation study for moderate sample sizes.

We use BIC to select the tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_1})$. Let

$$e(\boldsymbol{\lambda}) = \mathrm{tr}\{[\ddot{Q}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}}) + n_{\mathrm{T}} \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}})]^{-1} \ddot{Q}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{\mathrm{P}})\}$$

be the effective number of parameters in the last step of the Newton–Raphson iteration. Then

$$\mathrm{BIC}(\boldsymbol{\lambda}) = \log\left\{\frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \mathbf{R}_i^{-1} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)\right\} + \frac{\log(n_{\mathrm{T}})}{n_{\mathrm{T}}} e(\boldsymbol{\lambda}).$$

The minimization problem over a d-dimensional space is difficult. However, Li and Liang [16] conjectured that the magnitude of λ_k should be proportional to the standard error of β_k. So we suggest taking $\lambda_k = \lambda * \mathrm{SE}(\widehat{\beta}_k)$, in practice, where $\mathrm{SE}(\widehat{\beta}_k)$ is the standard error of $\widehat{\beta}_k$, the unpenalized estimator defined above. Thus, the minimization problem can be reduced to a one-dimensional problem, and the tuning parameter can be estimated by a grid search.

5. Simulation

In this section, we discuss finite sample properties of the proposed estimators via simulation studies. We simulated 100 data sets of size $n = 100$, 200 and 400 from the model

$$Y_{ij} = \boldsymbol{\beta}^T \mathbf{X}_{ij} + \eta_1(Z_{ij1}) + \eta_2(Z_{ij2}) + \varepsilon_{ij}, \qquad i = 1, \ldots, n, j = 1, \ldots, 3 \tag{18}$$

where the coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$, function $\eta_1(z) = \sin 2\pi(z - 0.5)$ and function $\eta_2(z) = z - 0.5 + \sin\{2\pi(z - 0.5)\}$.

The 2-vector $\underline{\mathbf{Z}}_i$ was generated from a bivariate normal distribution with mean 0, a common marginal variance 0.25 with correlation 0.9, but truncated to the unit square $[0, 1]^2$. The covariates X_{ijk}, $k = 1, \ldots, 6$, were generated independently from N(0, 0.25). Covariate $X_{ij7} = 3(1 - 2Z_{ij1})(1 - 2Z_{ij2}) + u_{ij}$, where $u_{ij} \sim N(0, 0.25)$ and is independent of $\mathbf{Z}_{ij}$. Covariate X_{ij8} was generated as -0.5 and 0.5 with equal probability. We generated $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ from N(0, Σ_E), where $\Sigma_E = (1 - \alpha)\mathbf{I} + \alpha \mathbf{1}\mathbf{1}^{\mathrm{T}}$ with $\mathbf{1}$ being a vector with all "1" and $\alpha = 0.9$, that is, Σ_E is exchangeable.

Cubic B-splines were used to approximate the nonparametric functions as described in Section 2. We tried different numbers of knots (ranging from 2 to 10) and found that the choice of number of knots didn't make a significant difference in this simulation study. Our reported results in Tables 1 and 2 were based on using 4 equally spaced knots.

To the simulated data sets, we applied the proposed method for estimation and variable selection. To study how the structure of the working correlation could affect our estimation and variable selection results, we considered the following three correlation structures: the correct exchangeable working correlation structure (EX), working independence (WI) and AR (1) structures. Table 1 summarizes the estimation and variable selection results with two types of penalty functions: SCAD and HARD. The average number of zero coefficients is reported in Table 1, in which the column labeled "C" presents the average restricted only to the true zero coefficients, and the column labeled "I" shows the average of numbers erroneously set to zero. The rows with "SCAD" and "HARD" stand, respectively, for the penalized least squares with the SCAD and HARD penalties. The oracle estimates always identify the 5 zero coefficients and 3 non-zero coefficients correctly. The medians of relative model errors (MRME) as suggested in Fan and Li [7] and the root mean squared errors (RMSE) of the estimated coefficients over 100 simulated data sets are also reported in Table 1.

From Table 1, one sees that the choice of correlation structure has little impact on the results of variable selection: the number of correctly identified zero coefficients are all close to 5 regardless the correlation structure; and none of the nonzero coefficients were erroneously set to 0 in any scenario. Table 1 also shows that the estimators with correct working correlation have the smallest RMSEs, thus are more efficient than those estimators with misspecified working correlation structures. The efficiency of the estimators based on the AR(1) is close to those based on EX, but there seems to be some significant loss of efficiency for the estimators based on the WI structure which ignores the within subject/cluster correlation. In terms of choosing penalty functions, we find that both HARD and SCAD perform very well and the corresponding MRME and RMSE are comparable to those of the ORACLE.

We also tested the accuracy of our standard error formula based on (17). The median absolute deviation (MAD) divided by 0.6745 (denoted by SD in Table 2) of 100 estimated coefficients

Table 1. Model selection and estimation: the average number of correct (C) and incorrect (I) 0s, MRME (%) and RMSE

		EX				AR(1)				WI			
n	Penalty	C	I	MRME	RMSE	C	I	MRME	RMSE	C	I	MRME	RMSE
100	SCAD	4.67	0	80.63	0.1592	4.64	0	84.65	0.1727	4.64	0	82.38	0.5883
	HARD	4.80	0	85.90	0.1691	4.70	0	86.56	0.1916	4.85	0	85.81	0.4410
	ORACLE	5.00	0	77.23	0.1586	5.00	0	73.40	0.1723	5.00	0	70.71	0.4126
200	SCAD	4.72	0	76.30	0.1053	4.72	0	81.63	0.1127	4.70	0	79.99	0.3921
	HARD	4.79	0	82.81	0.1116	4.71	0	82.18	0.1252	4.98	0	86.15	0.2816
	ORACLE	5.00	0	66.96	0.1038	5.00	0	66.18	0.1110	5.00	0	70.86	0.2787
400	SCAD	4.92	0	84.91	0.0733	4.86	0	84.50	0.0864	4.88	0	85.78	0.2689
	HARD	4.93	0	91.23	0.0758	4.87	0	85.65	0.0924	4.90	0	91.08	0.2021
	ORACLE	5.00	0	68.33	0.0731	5.00	0	66.91	0.0860	5.00	0	71.52	0.1857

from the 100 simulations can be regarded as the true standard error. The median of the 100 estimated SDs (denoted by $\mathrm{SD_m}$) and the MAD error of the 100 estimated standard errors divided by 0.6745 (denoted by $\mathrm{SD_{mad}}$) gauge the overall performance of the standard error. Table 2 presents the standard errors for non-zero coefficients when the sample size $n = 100$, 200 and 400. It suggests that the sandwich formula performs satisfactorily for SCAD and HARD penalties. The standard errors based on the SCAD and HARD penalty functions are closer to those of the ORACLE as n increases. Similarly to the RMSE results shown in Table 1, Table 2 also shows that the estimation procedures with a correct EX working correlation are more efficient than their counterparts with WI working correlation. Estimation based on a misspecified AR(1) correlation structure will lead to some efficiency loss, but it is quite close to using the true EX structure.

6. Application

To illustrate our method, we considered the longitudinal CD4 cell count data among HIV seroconverters. This dataset contains 2376 observations of CD4 cell counts on 369 men infected with the HIV virus; see Zeger and Diggle [42] for a detailed description of this dataset. Both Wang, Carroll and Lin [35] and Huang, Zhang and Zhou [14] analyzed the same dataset using a PLM. Their analysis aimed to estimate the average time course of CD4 counts and the effects of other covariates. In our analysis, we fit the data using an APLM, with the square root transformed CD4 counts as the response, and covariates including AGE, SMOKE (smoking status measured by packs of cigarettes), DRUG (yes, 1; no, 0), SEXP (number of sex partners), DEPRESSION (measured by the CESD scale) and YEAR (the effect of time since seroconversion). To take advantage of flexibility of partially linear additive models, we let both DEPRESSION and YEAR be modeled nonparametrically, the remaining parametrically. It is of interest to examine whether there are any interaction effects between the parametric covariates, so we included all these interactions in the parametric part.

Table 2. Simulation results on standard error estimation for the non-zero coefficients ($\beta_1, \beta_2, \beta_5$)

		$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
n	Penalty	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
EX										
100	SCAD	0.0889	0.0906	0.0141	0.1082	0.0911	0.0116	0.1034	0.0894	0.0111
	HARD	0.0879	0.0907	0.0118	0.1102	0.0911	0.0113	0.0982	0.0897	0.0108
	ORACLE	0.0866	0.0988	0.0062	0.1066	0.0899	0.0112	0.1012	0.0903	0.0088
200	SCAD	0.0655	0.0638	0.0035	0.0616	0.0629	0.0036	0.0594	0.0633	0.0036
	HARD	0.0655	0.0637	0.0033	0.0627	0.0630	0.0033	0.0600	0.0632	0.0035
	ORACLE	0.0648	0.0699	0.0078	0.0614	0.0637	0.0034	0.0594	0.0629	0.0036
400	SCAD	0.0414	0.0445	0.0043	0.0379	0.0445	0.0041	0.0415	0.0449	0.0042
	HARD	0.0414	0.0446	0.0043	0.0373	0.0445	0.0041	0.0418	0.0449	0.0042
	ORACLE	0.0412	0.0485	0.0086	0.0368	0.0443	0.0049	0.0404	0.0445	0.0046
AR(1)										
100	SCAD	0.0983	0.0923	0.0141	0.1035	0.0940	0.0129	0.1153	0.0924	0.0132
	HARD	0.0996	0.0920	0.0160	0.1073	0.0939	0.0130	0.1117	0.0924	0.0128
	ORACLE	0.0976	0.0972	0.0097	0.0971	0.0915	0.0124	0.1173	0.0930	0.0122
200	SCAD	0.0635	0.0646	0.0041	0.0539	0.0634	0.0047	0.0647	0.0639	0.0045
	HARD	0.0632	0.0645	0.0045	0.0544	0.0634	0.0044	0.0626	0.0639	0.0045
	ORACLE	0.0624	0.0689	0.0073	0.0535	0.0646	0.0049	0.0657	0.0635	0.0048
400	SCAD	0.0452	0.0448	0.0056	0.0390	0.0451	0.0055	0.0535	0.0452	0.0052
	HARD	0.0451	0.0449	0.0057	0.0390	0.0451	0.0055	0.0537	0.0452	0.0053
	ORACLE	0.0454	0.0477	0.0061	0.0392	0.0448	0.0051	0.0539	0.0450	0.0056
WI										
100	SCAD	0.2177	0.2364	0.0164	0.2192	0.2381	0.0187	0.2341	0.2375	0.0189
	HARD	0.2235	0.2364	0.0151	0.2185	0.2396	0.0169	0.2393	0.2389	0.0193
	ORACLE	0.2239	0.0579	0.1623	0.2118	0.2341	0.0165	0.2218	0.2374	0.0181
200	SCAD	0.1864	0.1674	0.0204	0.1697	0.1677	0 .0188	0.1328	0.1671	0.0186
	HARD	0.1876	0.1675	0.0201	0.1676	0.1680	0.0180	0.1385	0.1676	0.0181
	ORACLE	0.1836	0.0415	0.1242	0.1656	0.1669	0.0162	0.1356	0.1671	0.0165
400	SCAD	0.0957	0.1162	0.0131	0.1055	0.1163	0.0117	0.1252	0.1164	0.0131
	HARD	0.0956	0.1162	0.0129	0.1042	0.1165	0.0121	0.1222	0.1163	0.0128
	ORACLE	0.0956	0.0289	0.0778	0.1069	0.1160	0.0110	0.1227	0.1162	0.0104

For the working variance, we considered the WI, the AR(1) and the "random intercept plus serial correlation and measurement error" covariance (RSM) in Zeger and Diggle [42]. One can obtain the RSM structure by fitting a full model to the data and inspecting the variogram of the residuals. Wang, Carroll and Lin [35] and Huang, Zhang and Zhou [14] also analyzed this data set using the RSM structure. More precisely, the working covariance matrices are specified by $\tau^2\mathbf{I} + \nu^2\mathbf{J} + \omega^2\mathbf{H}$, where $\mathbf{I}$ is an identity matrix, $\mathbf{J}$ is a matrix of 1s and $\mathbf{H}(j, j') = \exp(-\alpha|\mathrm{YEAR}_{ij} - \mathrm{YEAR}_{ij'}|)$. We used the covariance parameters $(\tau^2, \nu^2, \omega^2, \alpha^2) =$

(11.32, 3.26, 22.15, 0.23) calculated by Wang *et al.* [35]. Table 3 gives the estimates of the regression coefficients using WI, AR(1) and RSM covariance structures. The standard errors (SE) were all calculated using the sandwich method. We used cubic splines of 4 knots selected by the five-fold delete-subject-out cross-validation from the range of 0–20. We refer the reader to Huang, Wu, and Zhou [13] for the detail of the delete-subjects-out K-fold cross-validation. The left panel of Table 3 reports the estimation using full model, and the selection results are shown in the right panel.

We further applied the proposed approach to select significant variables. We used the SCAD penalty, the tuning parameter $\lambda = 0.4549, 0.2829, 0.3143$ for WI, AR(1) and RSM covariance structure, respectively. The results are also shown in Table 3. Under both WI and RSM structures, SMOKE, DRUGS, SEXP, SOMKE∗SEXP and DRUGS∗SEXP are identifies as significant covariates. One notes some slight selection difference when AR(1) structure is used, which suggests that SMOKE∗DRUGS may also be significant. Although the selection procedure is not sensitive to the choice of covariance structure as shown in our simulation study, different covariance structures may still lead to slight different results. Therefore, it is important for one to choose a covariance structure close to the true one. We also find some significant interactions among some covariates which may be ignored by Wang, Carroll and Lin [35] and Huang, Zhang and Zhou [14].

The nonparametric curve estimates using the WI (solid line), AR(1) (dotted line) and RSM (dashed line) estimators are plotted in Figure 1 for "DEPRESSION" and "YEAR." One can see that it is more reasonable to put "DEPRESSION" as a nonparametric component.

7. Discussion

We have developed a general methodology for simultaneously selecting variables and estimating the unknown components in APLMs for longitudinal and clustered data. We propose a one-step least squares approach to obtain the estimation of both the parametric and nonparametric components based on polynomial spline smoothing. This approach is flexible, computationally simple and very easy to implement in practice. We demonstrate that the asymptotic normality of the estimated coefficients for the linear part is retained. The proposed penalized regression method also achieves an "oracle" property in the sense that it performs as well as if the subset of significant parametric components were known in advance.

In this paper, our primary interest is the linear components, and we treat the nonparametric functions as nuisance components; thus we limit our discussions to estimation and variable selection for the linear part. Nonetheless, this may be extended to the nonparametric components using techniques similar to those in Xue [37]. An anonymous referee pointed out the feasibility of obtaining the asymptotic "oracle" property of the nonparametric components in Ma and Yang [24]. We believe that this property can be similarly obtained via a two-step spline backfitted kernel smoothing procedure (Ma and Yang [24]). However, the technical details deserve careful consideration, and this is an interesting topic of future research.

Table 3. Estimated coefficients for CD4 dataset

Variable	Full			Penalized		
	WI	AR(1)	RSM	WI	AR(1)	RSM
	$\hat{\beta}$ (SE($\hat{\beta}$))	$\hat{\beta}$ (SE($\hat{\beta}$))	$\hat{\beta}$ (SE($\hat{\beta}$))	$\hat{\beta}$ (SE($\hat{\beta}$))	$\hat{\beta}$ (SE($\hat{\beta}$))	$\hat{\beta}$ (SE($\hat{\beta}$))
INTERCEPT	24.365 (0.417)	24.540 (0.480)	24.819 (0.494)	24.487 (0.391)	24.454 (0.464)	24.793 (0.461)
AGE	−0.013 (0.035)	−0.023 (0.045)	−0.049 (0.049)	0 (0)	0 (0)	0 (0)
SMOKE	1.070 (0.234)	0.825 (0.259)	0.654 (0.264)	0.733 (0.148)	0.824 (0.247)	0.424 (0.176)
DRUG	2.671 (0.491)	1.958 (0.517)	1.468 (0.507)	2.486 (0.454)	2.025 (0.511)	1.340 (0.462)
SEXP	0.165 (0.082)	0.153 (0.084)	0.109 (0.082)	0.170 (0.076)	0.174 (0.080)	0.144 (0.078)
AGE*SMOKE	−0.014 (0.011)	0.000 (0.015)	0.002 (0.017)	0 (0)	0 (0)	0 (0)
AGE*DRUG	0.043 (0.036)	0.008 (0.042)	0.010 (0.044)	0 (0)	0 (0)	0 (0)
AGE*SEXP	0.001 (0.005)	0.005 (0.005)	0.008 (0.005)	0 (0)	0 (0)	0 (0)
SMOKE*DRUG	−0.402 (0.233)	−0.331 (0.246)	−0.288 (0.248)	0 (0)	−0.337 (0.241)	0 (0)
SMOKE*SEXP	0.058 (0.024)	0.043 (0.026)	0.047 (0.025)	0.051 (0.023)	0.045 (0.025)	0.046 (0.025)
DRUG*SEXP	−0.364 (0.087)	−0.251 (0.086)	−0.17 (0.083)	−0.355 (0.084)	−0.265 (0.084)	−0.186 (0.081)

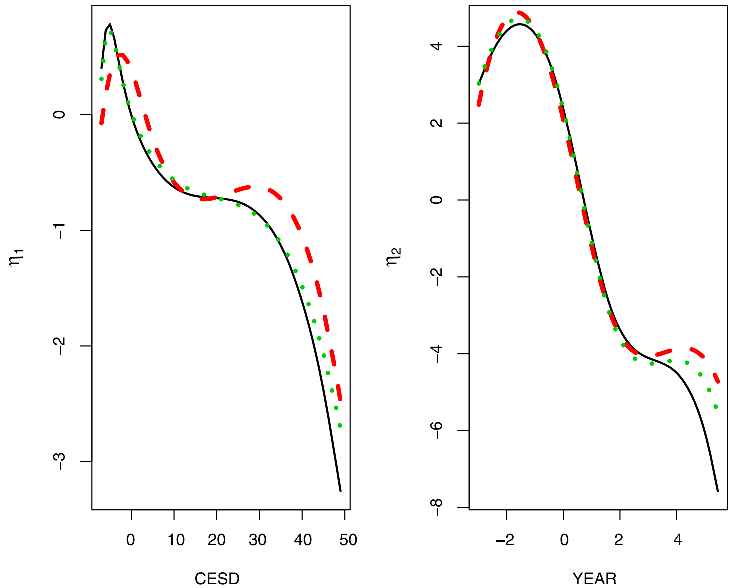

Figure 1. The estimates of the nonpararmetric components: $\widehat{\eta}_1$ and $\widehat{\eta}_2$. The solid, dotted and dashed curves correspond to the estimates under WI, AR(1) and RSM structures.

The simulation result indicates that the variable selection is consistent even if the correlation structure is misspecified. However, misspecification may lead to some efficiency loss. So, it would be desirable if one could choose an appropriate correlation structure based on available data in practice. The simulation results clearly show that there is marked improvement of efficiency when one uses the correct correlation structure though the variable selection seems to be consistent with misspecified structure. To select the correlation matrix, one might consider some resampling-based methods, such as the bootstrap and cross-validation methods in Pan and Connett [28] and other techniques in Diggle *et al.* [5]. There is, however, a clear need to formalize the procedures with solid theoretical justification. Instead of modeling the correlation through the "working" correlation matrix, one could also nonparametrically model the variance–covariance as some unknown smooth function (Chiou and Müller [4]). This is an excellent research problem for future study.

Appendix

For any vector $\mathbf{x} = (x_1, \dots, x_d)^{\mathrm{T}}$, we denote $\|\cdot\|$ the usual Euclidean norm, that is, $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{d} x_k^2}$, and $\|\cdot\|_\infty$ the sup norm, that is, $\|\mathbf{x}\|_\infty = \sup_{1\le k\le d} |x_k|$. For any functions ϕ, φ, let $\phi(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$ and $\varphi(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$ be m_i-vectors; then define the empirical inner product and the empirical norm as $\langle \phi, \varphi \rangle_n \equiv \langle \phi, \varphi \rangle_{n,\mathbf{V}} = n^{-1} \sum_{i=1}^{n} \phi(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)^{\mathrm{T}} \mathbf{V}_i^{-1} \varphi(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$, $\|\phi\|_n^2 = \langle \phi, \phi \rangle_n$, for the working covariance $\mathbf{V}_i$. Further denote $E_n(\phi) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{m_i}^{\mathrm{T}} \mathbf{V}_i^{-1} \phi(\underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$. If functions

ϕ, φ are L^2-integrable, we define the theoretical inner product and its corresponding theoretical L^2 norm as $\langle \phi, \varphi \rangle = E(\langle \phi, \varphi \rangle_n)$, $\|\phi\|^2 = E(\|\phi\|_n^2)$. Let $\widehat{\Pi}_n$ and Π_n denote, respectively, the projection onto G_n^0 relative to the empirical and theoretical inner products. For convenience, let $h = h_n \sim J_n^{-1}$ and $\mathbf{I}_d$ be the $d \times d$ identity matrix.

A.1. Proof of Theorem 1

Lemma A.1. *Define*

$$A_n = \sup_{g_1, g_2 \in G_n^0} |\langle g_1, g_2 \rangle_n - \langle g_1, g_2 \rangle| \|g_1\|^{-1} \|g_2\|^{-1},$$

$$B_n = \max_{1 \le k \le d_1} \sup_{g \in G_n^0} \left| \|x_k - g\|_n^2 / \|x_k - g\|^2 - 1 \right|,$$

then $A_n = \mathrm{O}_P\{\sqrt{\log(n)/(nh^2)}\}$ *and* $B_n = \mathrm{O}_P\{\sqrt{\log(n)/(nh^2)}\}$.

Lemma A.1 can be proved similarly to Lemmas A2 and A3 in Huang, Zhang and Zhou [14] and are thus omitted.

To obtain the closed-form expression of $\widehat{\boldsymbol{\beta}}$, we need the following block form of the inverse of $\sum_{i=1}^n \underline{\mathbf{D}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{D}}_i$:

$$\begin{pmatrix} \sum_{i=1}^n \underline{\mathbf{X}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{X}}_i & \sum_{i=1}^n \underline{\mathbf{X}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{B}}_i \\ \sum_{i=1}^n \underline{\mathbf{B}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{X}}_i & \sum_{i=1}^n \underline{\mathbf{B}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{B}}_i \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{H}_{\mathbf{XX}} & \mathbf{H}_{\mathbf{XB}} \\ \mathbf{H}_{\mathbf{BX}} & \mathbf{H}_{\mathbf{BB}} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{pmatrix}, \tag{A.1}$$

where $\mathbf{H}^{11} = (\mathbf{H}_{\mathbf{XX}} - \mathbf{H}_{\mathbf{XB}} \mathbf{H}_{\mathbf{BB}}^{-1} \mathbf{H}_{\mathbf{BX}})^{-1}$, $\mathbf{H}^{22} = (\mathbf{H}_{\mathbf{BB}} - \mathbf{H}_{\mathbf{BX}} \mathbf{H}_{\mathbf{XX}}^{-1} \mathbf{H}_{\mathbf{XB}})^{-1}$, $\mathbf{H}^{12} = -\mathbf{H}^{11} \mathbf{H}_{\mathbf{XB}} \mathbf{H}_{\mathbf{BB}}^{-1}$ and $\mathbf{H}^{21} = -\mathbf{H}^{22} \mathbf{H}_{\mathbf{BX}} \mathbf{H}_{\mathbf{XX}}^{-1}$. Consequently,

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{11} \left\{ \sum_{i=1}^n \underline{\mathbf{X}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{Y}}_i - \mathbf{H}_{\mathbf{XB}} \mathbf{H}_{\mathbf{BB}}^{-1} \sum_{i=1}^n \underline{\mathbf{B}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{Y}}_i \right\}. \tag{A.2}$$

Lemma A.2. *Under Assumptions* (A1)–(A5), *for* $\mathbf{H}_{\mathbf{BB}}$ *in* (A.1), *one has* (i) *there exist constants* $0 < c_H < C_H$, $C_H^* = c_H^{-1}$, $c_H^* = C_H^{-1}$ *such that*

$$c_H \mathbf{I}_{d_2 J_n} \le E(n^{-1} \mathbf{H}_{\mathbf{BB}}) \le C_H \mathbf{I}_{d_2 J_n}; \tag{A.3}$$

(ii) *with probability approaching* 1 *as* $n \to \infty$,

$$c_H \mathbf{I}_{d_2 J_n} \le n^{-1} \mathbf{H}_{\mathbf{BB}} \le C_H \mathbf{I}_{d_2 J_n}. \tag{A.4}$$

Since the proof of Lemma A.2 is a little complicated, we provide it in the supplemental article (Ma, Song and Wang [23]). The proofs of Lemmas A.3 to A.7 below are also provided in (Ma, Song and Wang [23]).

Lemma A.3. *Define* $\widehat{\mathbf{U}} = (\sum_{i=1}^{n} \underline{\mathbf{B}}_i^{\mathrm{T}} \underline{\mathbf{X}}_i)_{d_2 J_n \times d_1}$, *where* $\underline{\mathbf{B}}_i$ *is given in* (3). *Under Assumptions* (A1)–(A5), *there exist constants* $0 < c_U < C_U < \infty$, *such that with probability approaching* 1 *as* $n \to \infty$, $c_U \mathbf{I}_{d_1} \le (n^{-1}h)\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}} \le C_U \mathbf{I}_{d_1}$.

Lemma A.4. *Under Assumptions* (A1)–(A5), *there exist constants* $0 < c_{H_1} < C_{H_1} < \infty$, *such that with probability approaching* 1 *as* $n \to \infty$, $c_{H_1}\mathbf{I}_{d_1} \le n\mathbf{H}^{11} \le C_{H_1}\mathbf{I}_{d_1}$, *where* $\mathbf{H}^{11}$ *is given in* (A.1).

Let $\widetilde{\boldsymbol{\beta}}_{\mu}$ and $\widetilde{\boldsymbol{\beta}}_e$ be the solutions of (A.2) with $\underline{\mathbf{Y}}_i$ replaced by $\underline{\boldsymbol{\mu}}_i$ and $\underline{\mathbf{e}}_i = \underline{\mathbf{Y}}_i - \underline{\boldsymbol{\mu}}_i$, respectively. Then $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\widetilde{\boldsymbol{\beta}}_{\mu} - \boldsymbol{\beta}_0) + \widetilde{\boldsymbol{\beta}}_e$.

Lemma A.5. *Under Assumptions* (A1)–(A5), $\|\widetilde{\boldsymbol{\beta}}_{\mu} - \boldsymbol{\beta}_0\| = \mathrm{o}_P(n^{-1/2})$.

Note that $\widetilde{\boldsymbol{\beta}}_e = \mathbf{H}^{11}\{\sum_{i=1}^{n} \underline{\mathbf{X}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{e}}_i - \mathbf{H}_{\mathbf{XB}}\mathbf{H}_{\mathbf{BB}}^{-1} \sum_{i=1}^{n} \underline{\mathbf{B}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{e}}_i\}$; thus we can show that the conditional variance $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}_e | \mathbb{X}, \mathbb{Z})$ equals

$$\mathbf{H}^{11} \sum_{i=1}^{n} \{\underline{\mathbf{X}}_i - \underline{\mathbf{B}}_i \mathbf{H}_{\mathbf{BB}}^{-1} \mathbf{H}_{\mathbf{BX}}\}^{\mathrm{T}} \mathbf{V}_i^{-1} \boldsymbol{\Sigma}_i \mathbf{V}_i^{-1} \{\underline{\mathbf{X}}_i - \underline{\mathbf{B}}_i \mathbf{H}_{\mathbf{BB}}^{-1} \mathbf{H}_{\mathbf{BX}}\} \mathbf{H}^{11}. \tag{A.5}$$

Lemma A.6. *Under Assumptions* (A1)–(A5), *as* $n \to \infty$,

$$\{\mathrm{Var}(\widetilde{\boldsymbol{\beta}}_e | \mathbb{X}, \mathbb{Z})\}^{-1/2} (\widetilde{\boldsymbol{\beta}}_e) \longrightarrow N(0, \mathbf{I}_{d_1}).$$

Lemma A.7. *Under Assumptions* (A1)–(A5), *for the covariance matrix* $\boldsymbol{\Omega}(\mathbb{V}, \mathbb{Z})$ *defined in* (15), $c_V^* \mathbf{I}_{d_1} \le \boldsymbol{\Omega}(\mathbb{V}, \mathbb{Z}) \le C_V^* \mathbf{I}_{d_1}$ *and* $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}_e | \mathbb{X}, \mathbb{Z}) = n^{-1}\boldsymbol{\Omega}(\mathbb{V}, \mathbb{Z}) + \mathrm{O}_P(n^{-3/2} + n^{-1}h^{2p})$.

Theorem 1 follows from Lemmas A.5, A.6 and A.7.

A.2. Proof of Theorem 2

From (12) and (A.1), we obtain

$$\widehat{\boldsymbol{\gamma}} = \mathbf{H}^{22}\left(\sum_{i=1}^{n} \underline{\mathbf{B}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{Y}}_i - \mathbf{H}_{\mathbf{BX}} \mathbf{H}_{\mathbf{XX}}^{-1} \sum_{i=1}^{n} \underline{\mathbf{X}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \underline{\mathbf{Y}}_i\right). \tag{A.6}$$

Following the same idea as that in the proof of Lemma A.4, we have that there exist constants $0 < c_{H_2} < C_{H_2} < \infty$, such that with probability approaching 1 as $n \to \infty$, $c_{H_2}\mathbf{I}_{d_2 J_n} \le n\mathbf{H}^{22} \le C_{H_2}\mathbf{I}_{d_2 J_n}$. Letting $\widetilde{\boldsymbol{\gamma}}_{\mu}$ and $\widetilde{\boldsymbol{\gamma}}_e$ be the solutions of (A.6) with $\underline{\mathbf{Y}}_i$ replaced by $\underline{\boldsymbol{\mu}}_i$ and $\underline{\mathbf{e}}_i = \underline{\mathbf{Y}}_i - \underline{\boldsymbol{\mu}}_i$, respectively, $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = (\widetilde{\boldsymbol{\gamma}}_\mu - \boldsymbol{\gamma}) + \widetilde{\boldsymbol{\gamma}}_e$. Letting $\widehat{\Pi}_{n,\mathbf{X}}$ be the projection on $\{\underline{\mathbf{X}}_i\}_{i=1}^n$ to the empirical inner product, $\widetilde{\boldsymbol{\gamma}}_\mu - \boldsymbol{\gamma}$ equals

$$\begin{aligned}
&\mathbf{H}^{22}\left[\sum_{i=1}^n \underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il})\right\} - \mathbf{H}_{\mathbf{BX}}\mathbf{H}_{\mathbf{XX}}^{-1}\sum_{i=1}^n \underline{\mathbf{X}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il})\right\}\right] - \boldsymbol{\gamma} \\
&\quad = \mathbf{H}^{22}\sum_{i=1}^n \underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left[\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \underline{\mathbf{B}}_i\boldsymbol{\gamma}\right\} - \widehat{\Pi}_{n,\mathbf{X}}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \underline{\mathbf{B}}_i\boldsymbol{\gamma}\right\}\right] \\
&\quad = n\mathbf{H}^{22}\mathbf{S},
\end{aligned}$$

where $\mathbf{S} = (S_{11}, \ldots, S_{J_n d_2})$, with

$$S_{s,l} = n^{-1}\sum_{i=1}^n \big(\mathbf{B}_i^{(s,l)}\big)^{\mathrm{T}}\mathbf{V}_i^{-1}\left[\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \underline{\mathbf{B}}_i\boldsymbol{\gamma}\right\} - \widehat{\Pi}_{n,\mathbf{X}}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \underline{\mathbf{B}}_i\boldsymbol{\gamma}\right\}\right],$$

and $\mathbf{B}_i^{(s,l)} = [\{B_{s,l}(Z_{i1l}), \ldots, B_{s,l}(Z_{im_il})\}^{\mathrm{T}}]_{m_i\times 1}$. Let $\Delta\eta(\underline{\mathbf{Z}}_i) = \sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \underline{\mathbf{B}}_i\boldsymbol{\gamma}$, then the Cauchy–Schwarz inequality implies that

$$|S_{s,l}| \le \left\{n^{-1}\sum_{i=1}^n \big(\mathbf{B}_i^{(s,l)}\big)^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{B}_i^{(s,l)}\right\}^{1/2}\|\Delta\eta - \widehat{\Pi}_{n,\mathbf{X}}(\Delta\eta)\|_n = \mathrm{O}_P(h^p),$$

thus $\|\widetilde{\boldsymbol{\gamma}}_\mu - \boldsymbol{\gamma}\| = \mathrm{O}_P(J_n^{1/2}h^p)$. For any $\mathbf{c} \in \mathcal{R}^{J_n d_2}$ with $\|\mathbf{c}\| = 1$, we write $\mathbf{c}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_e = \sum_{i=1}^n a_i\varepsilon_i$, where ε_i are independent conditioning on $(\mathbb{X}, \mathbb{Z})$ and

$$a_i^2 = \mathbf{c}^{\mathrm{T}}\mathbf{H}^{22}\{\underline{\mathbf{B}}_i - \underline{\mathbf{X}}_i\mathbf{H}_{\mathbf{XX}}^{-1}\mathbf{H}_{\mathbf{XB}}\}^{\mathrm{T}}\mathbf{V}_i^{-1}\boldsymbol{\Sigma}_i\mathbf{V}_i^{-1}\{\underline{\mathbf{B}}_i - \underline{\mathbf{X}}_i\mathbf{H}_{\mathbf{XX}}^{-1}\mathbf{H}_{\mathbf{XB}}\}\mathbf{H}^{22}\mathbf{c}.$$

Following the same arguments as those in Lemma A.6, we have $\max_{1\le i\le n}|a_i| = \mathrm{O}_P(J_n^{1/2}n^{-1})$. Thus $\|\widetilde{\boldsymbol{\gamma}}_e\| \le J_n^{1/2}|\mathbf{c}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_e| = J_n^{1/2}|\sum_{i=1}^n a_i\varepsilon_i| = \mathrm{O}_P(J_n^{1/2}n^{-1/2})$. Therefore, $\|\widehat{\boldsymbol{\gamma}}_l - \boldsymbol{\gamma}_l\| = \mathrm{O}_P(J_n^{1/2}h^p + J_n^{1/2}n^{-1/2})$. Because $\widehat{\eta}_l(z_l) = \mathbf{B}_l^*(z_l)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_l$, $\widetilde{\eta}_l(z_l) = \mathbf{B}_l^*(z_l)^{\mathrm{T}}\boldsymbol{\gamma}_l$ and $|\widehat{\eta}_l - \widetilde{\eta}_l|_{L_2}^2 = \|\widehat{\boldsymbol{\gamma}}_l - \boldsymbol{\gamma}_l\|^2 \times \mathrm{O}_P(1) = \mathrm{O}_P(J_n h^{2p} + J_n n^{-1})$. Thus one has

$$|\widehat{\eta}_l - \eta_l|_{L_2}^2 \le 2(|\widehat{\eta}_l - \widetilde{\eta}_l|_{L_2}^2 + |\widetilde{\eta}_l - \eta_l|_{L_2}^2) = \mathrm{O}_P(J_n h^{2p} + J_n n^{-1}).$$

A.3. Proof of Theorem 3

Let $\tau_n = n^{-1/2} + a_n$. It suffices to show that for any given $\zeta > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q_{\mathcal{P}}(\boldsymbol{\beta}_0 + \tau_n\mathbf{u}) > Q_{\mathcal{P}}(\boldsymbol{\beta}_0)\right\} \ge 1 - \zeta. \tag{A.7}$$

Plugging $\boldsymbol{\gamma}(\boldsymbol{\beta})$ in (7) into $Q(\boldsymbol{\beta})$ defined in (8), we have

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{Y}_i - \left\{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\mathbf{H}_{\mathbf{BB}}^{-1}\sum_{i=1}^{n}\mathbf{B}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \underline{\mathbf{X}}_i\boldsymbol{\beta})\right\}\right]^{\mathrm{T}}$$
$$\times \mathbf{V}_i^{-1}\left[\mathbf{Y}_i - \left\{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{B}}_i\mathbf{H}_{\mathbf{BB}}^{-1}\sum_{i=1}^{n}\underline{\mathbf{B}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \underline{\mathbf{X}}_i\boldsymbol{\beta})\right\}\right].$$

Thus $Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\underline{\mathbf{X}}}_i\boldsymbol{\beta} - \widehat{\Pi}_n\mathbf{Y}_i)^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \widehat{\underline{\mathbf{X}}}_i\boldsymbol{\beta} - \widehat{\Pi}_n\mathbf{Y}_i)$. Let $U_{n,1} = Q(\boldsymbol{\beta}_0 + \tau_n\mathbf{u}) - Q(\boldsymbol{\beta}_0)$ and $U_{n,2} = n_{\mathrm{T}}\sum_{k=1}^{r}\{p_{\lambda_k}(|\beta_{k0} + \tau_n u_k|) - p_{\lambda_k}(|\beta_{k0}|)\}$, where r is the number of components of $\boldsymbol{\beta}_{10}$. Note that $p_{\lambda_k}(0) = 0$ and $p_{\lambda_k}(|\beta|) \geq 0$ for all β. Thus, $Q_{\mathcal{P}}(\boldsymbol{\beta}_0 + \tau_n\mathbf{u}) - Q_{\mathcal{P}}(\boldsymbol{\beta}_0) \geq U_{n,1} + U_{n,2}$.

For $U_{n,1}$, we have $Q(\boldsymbol{\beta}_0 + \tau_n\mathbf{u}) = Q(\boldsymbol{\beta}_0) + \tau_n\mathbf{u}^{\mathrm{T}}\dot{Q}(\boldsymbol{\beta}_0) + \frac{1}{2}\tau_n^2\mathbf{u}^{\mathrm{T}}\ddot{Q}(\boldsymbol{\beta}^*)\mathbf{u}$, where $\ddot{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\widehat{\underline{\mathbf{X}}}_i$, $\boldsymbol{\beta}^* = t(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) + (1-t)\boldsymbol{\beta}_0$, $t \in [0, 1]$. Note that

$$\begin{aligned}\dot{Q}(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \widehat{\underline{\mathbf{X}}}_i\boldsymbol{\beta}_0 - \widehat{\Pi}_n\mathbf{Y}_i)\\ &= \sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \widehat{\Pi}_n\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il})\right\} + \sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \widehat{\Pi}_n\underline{\mathbf{e}}_i),\end{aligned}$$

where $\underline{\mathbf{e}}_i = \underline{\mathbf{Y}}_i - \underline{\boldsymbol{\mu}}_i$. Mimicking the proof for Lemmas A.5 and A.6, we have

$$\sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left\{\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il}) - \widehat{\Pi}_n\sum_{l=1}^{d_2}\eta_l(\mathbf{Z}_{il})\right\} = \mathrm{o}_P(n^{1/2}),$$
$$\sum_{i=1}^{n}\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \widehat{\Pi}_n\underline{\mathbf{e}}_i) = \mathrm{O}_P(n^{1/2}).$$

Thus $\tau_n\mathbf{u}^{\mathrm{T}}\dot{Q}(\boldsymbol{\beta}_0) = \mathrm{O}_P(n^{1/2}\tau_n)\|\mathbf{u}\|$. By the proof of Lemma A.4, we obtain that $\frac{1}{2}\tau_n^2\mathbf{u}^{\mathrm{T}} \times \ddot{Q}(\boldsymbol{\beta}_0)\mathbf{u} = \mathrm{O}_P(n\tau_n^2) + \mathrm{o}_P(1)$. Thus

$$U_{n,1} = \mathrm{O}_P(n^{1/2}\tau_n) + \mathrm{O}_P(n\tau_n^2) + \mathrm{o}_P(1). \tag{A.8}$$

For $U_{n,2}$, by a Taylor expansion,

$$p_{\lambda_k}(|\beta_{k0} + \tau_n u_k|) = p_{\lambda_k}(|\beta_{k0}|) + \tau_n u_k p'_{\lambda_k}(|\beta_{k0}|)\mathrm{sgn}(\beta_{k0}) + \frac{1}{2}\tau_n^2 u_k^2 p''_{\lambda_k}(|\beta_k^*|),$$

where $\beta_k^* = (1-t)\beta_{k0} + t(\beta_{k0} + n^{-1/2}u_k)$, $t \in [0, 1]$ and

$$p_{\lambda_k}(|\beta_{k0} + \tau_n u_k|) = p_{\lambda_k}(|\beta_{k0}|) + \tau_n u_k p'_{\lambda_k}(|\beta_{k0}|)\mathrm{sgn}(\beta_{k0}) + \frac{1}{2}\tau_n^2 u_k^2 p''_{\lambda_k}(|\beta_{k0}|) + \mathrm{o}(n^{-1}).$$

Thus, by the Cauchy–Schwarz inequality,

$$n_{\mathrm{T}}^{-1}U_{n,2} = \tau_n \sum_{k=1}^{r} u_k p'_{\lambda_k}(|\beta_{k0}|)\mathrm{sgn}(\beta_{k0}) + \frac{1}{2}\tau_n^2 \sum_{k=1}^{r} u_k^2 p''_{\lambda_k}(|\beta_{k0}|)$$
$$\leq \sqrt{r}\tau_n a_n \|\mathbf{u}\| + \frac{1}{2}\tau_n^2 w_n \|\mathbf{u}\|^2 = C\tau_n^2\big(\sqrt{r} + w_n C\big).$$

As $w_n \to 0$, the first two terms on the right-hand side of (A.8) dominate $U_{n,2}$ by taking C sufficiently large. Hence (A.7) holds for sufficiently large C.

A.4. Proof of Theorem 4

We first show that the estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{P}}$ must possess the sparsity property $\widehat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma A.8. *Under the conditions of Theorem 4, with probability tending to* 1, *for any given* $\boldsymbol{\beta}_1$ *satisfying that* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = \mathrm{O}_P(n^{-1/2})$ *and any constant* C,

$$Q_{\mathcal{P}}\{(\boldsymbol{\beta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}\} = \min_{\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}} Q_{\mathcal{P}}\{(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})\}.$$

Proof. To prove that the maximizer is obtained at $\boldsymbol{\beta}_2 = 0$, it suffices to show that with probability tending to 1, as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = \mathrm{O}_P(n^{-1/2})$, and $\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}$, $\partial Q_{\mathcal{P}}(\boldsymbol{\beta})/\partial \beta_k$ and β_k have different signs for $\beta_k \in (-Cn^{-1/2}, Cn^{-1/2})$, for $k = r+1, \ldots, d_1$. Note that

$$\dot{Q}_{\mathcal{P},k}(\boldsymbol{\beta}) \equiv \frac{\partial Q_{\mathcal{P}}(\boldsymbol{\beta})}{\partial \beta_k} = \dot{Q}_k(\boldsymbol{\beta}) + n_{\mathrm{T}} p'_{\lambda_{kn}}(|\beta_k|)\,\mathrm{sgn}(\beta_k),$$

where $\dot{Q}_k(\boldsymbol{\beta}) = \dot{Q}_k(\boldsymbol{\beta}_0) + \sum_{k'=1}^{d_1} \ddot{Q}_{kk'}\{t\beta_{k'} + (1-t)\beta_{0k'}\}(\beta_{k'} - \beta_{0k'})$, $t \in [0, 1]$,

$$\dot{Q}_k(\boldsymbol{\beta}_0) = e_k^{\mathrm{T}} \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \widehat{\underline{\mathbf{X}}}_i \boldsymbol{\beta}_0 - \widehat{\Pi}_n \mathbf{Y}_i).$$

It follows by the similar arguments as given in the proofs of Theorems 1 and 3 that

$$\dot{Q}_k(\boldsymbol{\beta}_0) = e_k^{\mathrm{T}} \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \left\{ \sum_{l=1}^{d_2} \eta_l(\mathbf{Z}_{il}) - \widehat{\Pi}_n \sum_{l=1}^{d_2} \eta_l(\mathbf{Z}_{il}) \right\} + e_k^{\mathrm{T}} \sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \widehat{\Pi}_n \underline{\mathbf{e}}_i)$$
$$= n\left\{ n^{-1} \sum_{i=1}^{n} \Xi_k(\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i) + \mathrm{o}_P(n^{-1/2}) \right\},$$

where $\Xi_k(\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{Z}}_i)$ is the kth element of matrix $\widehat{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \widehat{\Pi}_n \underline{\mathbf{e}}_i)$. According to Lemma A.7, we have

$$n^{-1}\ddot{Q}(\boldsymbol{\beta}_0) = E\left(n^{-1}\sum_{i=1}^{n} \widetilde{\underline{\mathbf{X}}}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \widetilde{\underline{\mathbf{X}}}_i\right) + \mathrm{o}_P(1) = \mathbf{R} + \mathrm{o}_P(1),$$

$$\frac{1}{n}\sum_{k'=1}^{d_1} \ddot{Q}_{kk'}(\beta_{k'} - \beta_{0k'}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\big(R_k + \mathrm{o}_P(1)\big),$$

where R_k is the kth column of $\mathbf{R}$. Note that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \mathrm{O}_P(n^{-1/2})$ by the assumption. Thus, $n^{-1}\dot{Q}_k(\boldsymbol{\beta})$ is of the order $\mathrm{O}_P(n^{-1/2})$. Therefore, for any nonzero β_k and $k = r+1, \ldots, d_1$,

$$\dot{Q}_{\mathcal{P},k}(\boldsymbol{\beta}) = n\lambda_{kn}\left\{\lambda_{kn}^{-1} p'_{\lambda_{kn}}(|\beta_k|)\,\mathrm{sgn}(\beta_k) + \mathrm{O}_P\left(\frac{1}{\sqrt{n}\lambda_{kn}}\right)\right\}.$$

Since $\liminf_{n\to\infty} \liminf_{\beta_k\to 0^+} \lambda_{kn}^{-1} p'_{\lambda_{kn}}(|\beta_k|) > 0$ and $\sqrt{n}\lambda_{kn} \to \infty$, the sign of the derivative is determined by that of β_k. Thus the desired result is obtained. □

Proof of Theorem 4. From Lemma A.8, it follows that $\widehat{\boldsymbol{\beta}}_2^{\mathrm{P}} = \mathbf{0}$.

$$\begin{aligned}\dot{Q}_{\mathcal{P}}(\boldsymbol{\beta}) &= \dot{Q}(\boldsymbol{\beta}_0) + \ddot{Q}(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n_{\mathrm{T}}\{p'_{\lambda_{kn}}(|\beta_{k0}|)\,\mathrm{sign}(\beta_{k0})\}_{k=1}^{r} \\ &\quad + \left\{\sum_{k=1}^{r} p''_{\lambda_{kn}}(|\beta_{k0}|) + \mathrm{o}_P(1)\right\}(\widehat{\boldsymbol{\beta}}_{k1}^{\mathrm{P}} - \beta_{k0}),\end{aligned}$$

where $\boldsymbol{\beta}^* = t\boldsymbol{\beta}_0 + (1-t)\boldsymbol{\beta}$, $t \in [0,1]$. Using an argument similar to the proof of Theorem 3, it can be shown that there exists a $\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}}$ in Theorem 3 that is a root-n consistent local minimizer of $Q_{\mathcal{P}}\{(\boldsymbol{\beta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}\}$, satisfying the penalized least squares equations $\dot{Q}_{\mathcal{P}}[\{(\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\}^{\mathrm{T}}] = \mathbf{0}$. Mimicking the proofs for Lemmas A.5 and A.6 indicates that the left hand side of the above equation can be written as

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_{1i}^{\mathrm{T}} \mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \hat{\Pi}_n \underline{\mathbf{e}}_i) + \{p'_{\lambda_{kn}}(|\beta_{k0}|)\,\mathrm{sign}(\beta_{k0})\}_{k=1}^{r} + \mathrm{o}_P(n^{-1/2}) \\ &\quad + \left\{E\left(n^{-1}\sum_{i=1}^{n} \widetilde{\underline{\mathbf{X}}}_{1i}^{\mathrm{T}} \mathbf{V}_i^{-1} \widetilde{\underline{\mathbf{X}}}_{1i}\right) + \mathrm{o}_P(1)\right\}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}} - \boldsymbol{\beta}_{10}) + \left\{\sum_{k=1}^{r} p''_{\lambda_{kn}}(|\beta_{k0}|) + \mathrm{o}_P(1)\right\}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}} - \boldsymbol{\beta}_{10}).\end{aligned}$$

Thus we have

$$\begin{aligned}\mathbf{0} &= n^{-1}\sum_{i=1}^{n} \widehat{\underline{\mathbf{X}}}_{1i}^{\mathrm{T}} \mathbf{V}_i^{-1}(\underline{\mathbf{e}}_i - \hat{\Pi}_n \underline{\mathbf{e}}_i) + \kappa_n + \mathrm{o}_P(n^{-1/2}) \\ &\quad + \left\{E\left(n^{-1}\sum_{i=1}^{n} \widetilde{\underline{\mathbf{X}}}_{1i}^{\mathrm{T}} \mathbf{V}_i^{-1} \widetilde{\underline{\mathbf{X}}}_{1i}\right) + \boldsymbol{\Sigma}_\lambda + \mathrm{o}_P(1)\right\}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{P}} - \boldsymbol{\beta}_{10}).\end{aligned}$$

Similar arguments to Lemmas A.6 and A.7 yield the asymptotic normality. □

Acknowledgments

Ma's research was supported by a dissertation fellowship from Michigan State University. Wang's research was supported in part by NSF award DMS-0905730. The authors are grateful for the insightful comments from the editor, an associate editor and anonymous referees.

Supplementary Material

Supplement to "Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data" (DOI: 10.3150/11-BEJ386SUPP; .pdf). We provide detailed proofs of Lemmas A.2 to A.7 stated in the Appendix.

References

[1] Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Italian Jour. Statist.* **6**, 97–144.
[2] Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92** 303–316. MR2201361
[3] Carroll, R.J., Maity, A., Mammen, E. and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika* **96** 383–398. MR2507150
[4] Chiou, J.M. and Müller, H.G. (2005). Estimated estimating equations: Semiparametric inference for clustered and longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 531–553. MR2168203
[5] Diggle, P.J., Heagerty, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford: Oxford Univ. Press. MR2049007
[6] Fan, J., Feng, Y. and Song, R. (2010). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557.
[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581
[8] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656
[9] Fan, Y. and Li, Q. (2003). A kernel-based method for estimating additive partially linear models. *Statist. Sinica* **13** 739–762. MR1997172
[10] Fu, W.J. (2003). Penalized estimating equations. *Biometrics* **59** 126–132. MR1978479
[11] Hall, P., Müller, H.G. and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365
[12] Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*: *Contributions to Statistics*. Heidelberg: Physica-Verlag. MR1787637
[13] Huang, J.Z., Wu, C.O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972
[14] Huang, J.Z., Zhang, L. and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Statist.* **34** 451–477. MR2368793

[15] Li, Q. (2000). Efficient estimation of additive partially linear models. *Internat. Econom. Rev.* **41** 1073–1092. MR1790072
[16] Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36** 261–286. MR2387971
[17] Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104** 234–248. MR2504375
[18] Liang, H., Thurston, S.W., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95** 667–678. MR2443182
[19] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430
[20] Lin, X. and Carroll, R.J. (2001). Semiparametric regression for clustered data. *Biometrika* **88** 1179–1185. MR1872228
[21] Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statist. Sinica.* **21** 1225–1248.
[22] Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *J. Comput. Graph. Statist.* **16** 782–798. MR2412482
[23] Ma, S., Song, Q. and Wang, L. (2011). Supplement to "Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data". DOI:10.3150/11-BEJ386SUPP.
[24] Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *J. Statist. Plann. Inference* **141** 204–219. MR2719488
[25] Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli* **16** 274–300. MR2648758
[26] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[27] Opsomer, J.D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modelling. *J. Comput. Graph. Statist.* **8** 715–734.
[28] Pan, W. and Connett, J.E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statist. Sinica* **12** 475–490. MR1902720
[29] Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566
[30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[31] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.
[32] Wang, J.L., Xue, L., Zhu, L. and Chong, Y.S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322
[33] Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* **35** 2474–2503. MR2382655
[34] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90** 43–52. MR1966549
[35] Wang, N., Carroll, R.J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157. MR2156825
[36] Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19** 801–817. MR2514189
[37] Xue, L. (2009). Consistent variable selection in additive models. *Statist. Sinica* **19** 1281–1296. MR2536156
[38] Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *J. Amer. Statist. Assoc.* **105** 1518–1530. MR2796568

[39] Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16** 1423–1446. MR2327498

[40] Yang, Y. (2008). Localized model selection for regression. *Econometric Theory* **24** 472–492. MR2391618

[41] Yuan, M. and Lin, Y. (2007). On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 143–161. MR2325269

[42] Zeger, S.L. and Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50** 689–699.

Received August 2010 and revised June 2011