Statistics and Its Interface Volume 8 (2015) 355–365

Variable selection in strong hierarchical semiparametric models for longitudinal data

Xianbin Zeng[†], Shuangge Ma[‡], Yichen Qin[†], and Yang Li[*,†,‡]

In this paper, we consider the variable selection problem in semiparametric additive partially linear models for longitudinal data. Our goal is to identify relevant main effects and corresponding interactions associated with the response variable. Meanwhile, we enforce the strong hierarchical restriction on the model, that is, an interaction can be included in the model only if both the associated main effects are included. Based on B-splines basis approximation for the nonparametric components, we propose an iterative estimation procedure for the model by penalizing the likelihood with a partial group minimax concave penalty (MCP), and use BIC to select the tuning parameter. To further improve the estimation efficiency, we specify the working covariance matrix by maximum likelihood estimation. Simulation studies indicate that the proposed method tends to consistently select the true model and works efficiently in estimation and prediction with finite samples, especially when the true model obeys the strong hierarchy. Finally, the China Stock Market data are fitted with the proposed model to illustrate its effectiveness.

Keywords and phrases: Variable selection, Interaction, Semiparametric additive partially linear model, Strong hierarchy, Longitudinal data.

1. INTRODUCTION

Longitudinal data arise frequently in many research areas such as biology, medicine, economics, and social science. The common feature of these kinds of studies is the repeated measurements on the same subjects. Various parametric models have been developed for longitudinal data analysis [20, 10], but they may have the risk of introducing biases when the relationship between the response and covariates is complex and cannot be featured adequately by parametric forms [12]. To relax the assumptions on parametric forms, we consider the semiparametric additive partially linear model (APLM) in this paper, which combines the parsimony of parametric regression and flexibility of nonparametric regression, and thus provides a nice trade-off between model interpretability and flexibility. Estimation for APLMs has been receiving increasing attention, and there is a considerable amount of relevant studies (see, e.g., [7, 22, 2, 8]).

With the emerging of high-dimensional longitudinal data, variable selection for longitudinal data becomes a fundamental issue. Including only the relevant covariates in the model will often enhance predictability and give a parsimonious and interpretable model. Different penalization-based variable selection methods have been proposed in linear regression analysis, such as the Lasso [25, 26], Bridge [18], SCAD [11], Elastic Net [35], Adaptive Lasso [36], MCP [32] and Group Lasso [31, 19], etc. However, variable selection for APLMs with longitudinal data is rather challenging due to the semiparametric relationship and the within-subject correlation structure among the repeated observations of the same subjects, and thus has not received sufficient attention. Taking account of the within-subject correlation correctly is essential for obtaining efficient estimators for longitudinal data [27, 7].

In addition, in many practical problems, significant joint effects, or interactions, may exist among the covariates. In modern genetic association studies, for example, the risks of multifactorial traits, such as cancer, are often determined by complex interactions between genetic and environmental exposures. In order to discover the underlying susceptibility genes, the heterogeneity in genetic effects due to the gene-environment or gene-gene interactions cannot be ignored [23]. When interactions exist, it is generally assumed that there is a natural hierarchy among the covariates, that is, an interaction cannot be chosen until a split has been made on its associated main effects [21]. A general variable selection approach ignoring the hierarchical constraints may select an interaction without the corresponding main effects. Such models can be difficult to interpret in practice. Some recent studies have extended the penalized variable selection methods to regression models with pairwise interactions terms [33, 9, 24, 4, 5, 21]. However, all of these studies are concentrated on purely parametric or nonparametric models for sectional data. In this paper, we consider the APLM of longitudinal data with all possible two-way interactions between parametric terms and between parametric

*Corresponding to: Yang Li, an associate professor in the School of Statistics, Renmin University of China, yang.li@ruc.edu.cn.
[†]National Natural Science Foundation of China under Grant 71301162.
[‡]National Social Science Foundation of China under Grant 13CTJ001, NIH award CA165923.
[§]Scientific Research Foundation for the Returned Overseas Chinese Scholars, Fundamental Research Funds for the Central Universities (the Research Funds of Renmin University of China State Education Ministry) under Grant 13XNF058.

and nonparametric terms, and study its variable selection problem.

The rest of the paper is organized as follows. In Section 2 we introduce the APLM with strong hierarchical interactions, which we call Strong Hierarchical APLM (SHAPLM) in this paper, and provide the B-splines basis approximation of the nonparametric components. We propose the penalized likelihood criterion and iterative estimation approach for SHAPLM in Section 3 to select the relevant main effects and interactions while simultaneously enforcing the strong hierarchy. A set of simulation studies with the truth obeying or disobeying the strong hierarchy are surveyed to assess performance of the proposed method in Section 4. An application example is presented in Section 5. We conclude the paper with a discussion in Section 6.

2. STRONG HIERARCHICAL APLM

Suppose that the data consist of m subjects, with each subject $i(i = 1, \ldots, m)$ having n_i observations. Let $\{(y_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}), 1 \le i \le m, 1 \le j \le n_i\}$ be the jth observation for subject i, where y_{ij} is the response variable, $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is a $p \times 1$ covariate vector corresponding to the parametric components, and $\boldsymbol{U}_{ij} = (u_{ij1}, \ldots, u_{ijq})'$ is a $q \times 1$ covariate vector corresponding to the nonparametric components. Then the APLM is of the form

$$y_{ij} = \mu + \sum_{k=1}^{p} x_{ijk}\beta_k + \sum_{l=1}^{q} f_l(u_{ijl}) + \epsilon_{ij} \tag{1}$$

where $f_l(\cdot)$, $l = 1, \ldots, q$ are unknown smooth functions and represent the nonlinear effects of $\boldsymbol{U}_{ij}$. We assume that the errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})'$ are independently normally distributed as $N(0, \boldsymbol{\Sigma}_i)$.

Furthermore, by including all possible pairwise interactions between parametric terms and between parametric and nonparametric terms in the APLM (1), we get the following model:

$$\begin{aligned} y_{ij} = \mu &+ \sum_{k=1}^{p} x_{ijk}\beta_k + \sum_{l=1}^{q} f_l(u_{ijl}) \\ &+ \sum_{k=1}^{p-1}\sum_{k'=k+1}^{p} x_{ijk}x_{ijk'}\alpha_{kk'} \\ &+ \sum_{k=1}^{p}\sum_{l=1}^{q} x_{ijk} f_l(u_{ijl})\pi_{kl} + \epsilon_{ij} \end{aligned} \tag{2}$$

In model (2), the interactions between the parametric and nonparametric terms are also assumed to be in linear product forms. It is the same as the interaction terms in Maity et al. [23] which are linear products of the parametric genetic effects and nonparametric environmental effects. Interactions among the nonparametric terms are of complex and unintelligible nonparametric forms. For example, the interaction between $f_1(u_1)$ and $f_2(u_2)$ is $f_{12}(u_1, u_2)$, which is an unknown two-dimensional nonparametric function and would also cause "curse of dimensionality". Therefore such interactions are not considered in our model.

Our goal here is to fit model (2) while simultaneously enforce the strong hierarchy, which means that an interaction can be present only if both of its associated main effects are present in the model. Another type of hierarchy is weak hierarchy, which is obeyed as long as either associated main effects of the selected interaction are present [21], and is not investigated in this paper.

To impose the strong hierarchical constraints on the interactions in model (2), we reparameterize their coefficients $\alpha_{kk'}$ and π_{kl} as $\alpha_{kk'} = \gamma_{kk'}\beta_k\beta_{k'}$ and $\pi_{kl} = \eta_{kl}\beta_k$, and thus produce the Strong Hierarchical APLM (SHAPLM):

$$\begin{aligned} y_{ij} = \mu &+ \sum_{k=1}^{p} x_{ijk}\beta_k + \sum_{l=1}^{q} f_l(u_{ijl}) \\ &+ \sum_{k=1}^{p-1}\sum_{k'=k+1}^{p} x_{ijk}x_{ijk'}\gamma_{kk'}\beta_k\beta_{k'} \\ &+ \sum_{k=1}^{p}\sum_{l=1}^{q} x_{ijk} f_l(u_{ijl})\eta_{kl}\beta_k + \epsilon_{ij} \end{aligned} \tag{3}$$

By expressing the interaction coefficient $\alpha_{kk'}$ or π_{kl} as the product of a separate parameter ($\gamma_{kk'}$ or η_{kl}) and the associated main effects coefficients, model (3) naturally enforces the strong hierarchy. The separate parameters $\boldsymbol{\gamma} = (\gamma_{12}, \ldots, \gamma_{p-1,p})'$ or $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_q')'$ are used to capture interactions, where $\boldsymbol{\eta}_l = (\eta_{1l}, \ldots, \eta_{pl})'$, $l = 1, \ldots, q$. It is obvious that if a main effect $\boldsymbol{X}_{(k)} = (x_{11k}, \ldots, x_{1n_1k}, \ldots, x_{m1k}, \ldots, x_{mn_mk})'$ or $\boldsymbol{U}_{(l)} = (u_{11l}, \ldots, u_{1n_1l}, \ldots, u_{m1l}, \ldots, u_{mn_ml})'$ is excluded from the model, that is $\beta_k = 0$ or $f_l(\cdot) \equiv 0$, then all the associated interactions, $\boldsymbol{X}_{(k)} :: \boldsymbol{X}_{(k')} = (x_{11k}x_{11k'}, \ldots, x_{mn_mk}x_{mn_mk'})'$ $(\forall k' \neq k)$ or $\boldsymbol{X}_{(k)} :: f_l(\boldsymbol{U}_{(l)}) = (x_{11k}f_l(u_{11l}), \ldots, x_{mn_mk}f_l(u_{mn_ml}))'$ $(\forall k)$, are certainly excluded from the model. Identical forms of the interactions between parametric components and between parametric and nonparametric components in model (3) are proposed in Choi, Li and Zhu [9] and Maity et al. [23], respectively.

The continuous covariate u_{ijl} in model (3) is often assumed to be distributed on a compact interval $[a_l, b_l]$. Without loss of generality, we take all $[a_l, b_l] = [a, b]$ for $l = 1, \ldots, q$. For identifiability, we assume $\int_a^b f_l(u)du = 0$ for each l. Under the smoothness assumptions of $f_l(\cdot)$, we can approximate $f_l(\cdot)$ by a basis expansion [15, 16], i.e.,

$$f_l(u) \approx \sum_{d=1}^{D_l} \phi_l^{(d)} B_l^{(d)}(u) = \boldsymbol{B}_l'(u)\boldsymbol{\phi_l} \tag{4}$$

where D_l is the number of basis functions in approximating $f_l(\cdot)$, $\boldsymbol{B}_l(\cdot) = (B_l^{(1)}(\cdot), \ldots, B_l^{(D_l)}(\cdot))'$ is the vector of known basis functions, and $\boldsymbol{\phi}_l = (\phi_l^{(1)}, \ldots, \phi_l^{(D_l)})'$ is the vector of regression coefficients, $l = 1, \ldots, q$. For simplicity, we apply the same number of basis functions for each nonparametric component with $D_l = D$ for $l = 1, \ldots, q$. With this expansion, one nonparametric covariate effect corresponds to a group of multiple regression coefficients for the basis

functions. In this paper, cubic B-splines basis functions are used to estimate the nonparametric components for their good approximation properties and desirable computational speed [1, 28].

We further define some matrix notations for convenience. For subject i, let $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the $n_i \times 1$ response, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i})'$ be the $n_i \times p$ covariates matrix, and $\boldsymbol{F}_i = (\boldsymbol{f}_{i1}, \ldots, \boldsymbol{f}_{iq})$ be the $n_i \times q$ smooth functions matrix, where $\boldsymbol{f}_{il} = (f_l(u_{i1l}), \ldots, f_l(u_{in_il}))', l = 1, \ldots, q$. We denote the interactions matrices as

$$
\begin{aligned}
\boldsymbol{X}_i :: \boldsymbol{X}_i &= \begin{pmatrix} x_{i11}x_{i12} & \cdots & x_{i1,p-1}x_{i1p} \\ \vdots & \ddots & \vdots \\ x_{in_i1}x_{in_i2} & \cdots & x_{in_i,p-1}x_{in_ip} \end{pmatrix} \text{ and} \\
\boldsymbol{X}_i :: \boldsymbol{F}_i &= \begin{pmatrix} x_{i11}f_1(u_{i11}) & \cdots & x_{i1p}f_q(u_{i1q}) \\ \vdots & \ddots & \vdots \\ x_{in_i1}f_1(u_{in_i1}) & \cdots & x_{in_ip}f_q(u_{in_iq}) \end{pmatrix}
\end{aligned} \tag{5}
$$

respectively. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{\alpha} = (\gamma_{12}\beta_1\beta_2, \ldots, \gamma_{p-1,p}\beta_{p-1}\beta_p)'$, $\boldsymbol{\pi}_l = (\eta_{1l}\beta_1, \ldots, \eta_{pl}\beta_p)'$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_q')'$, then model (3) can be written in matrix form as

$$
\begin{aligned}
\boldsymbol{Y}_i = \mu\mathbf{1}_{n_i} + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{F}_i\mathbf{1}_q + (\boldsymbol{X}_i :: \boldsymbol{X}_i)\boldsymbol{\alpha} \\
+ (\boldsymbol{X}_i :: \boldsymbol{F}_i)\boldsymbol{\pi} + \boldsymbol{\epsilon_i}
\end{aligned} \tag{6}
$$

where $\mathbf{1}_q$ denotes a column vector of length q with all elements 1. It follows from (4) that

$$
\begin{aligned}
&\boldsymbol{f}_{il} = (\boldsymbol{B}_l'(u_{i1l})\boldsymbol{\phi}_l, \ldots, \boldsymbol{B}_l'(u_{in_il})\boldsymbol{\phi}_l)' = \boldsymbol{Z}_{il}\boldsymbol{\phi}_l \\
&\textit{and} \quad \boldsymbol{F}_i\mathbf{1}_q = \boldsymbol{Z}_i\boldsymbol{\phi}
\end{aligned} \tag{7}
$$

where $\boldsymbol{Z}_{il} = (\boldsymbol{B}_l(u_{i1l}), \ldots, \boldsymbol{B}_l(u_{in_il}))'$ is the $n_i \times D$ B-splines basis matrix corresponding to the lth expansion, $l = 1, \ldots, q$, $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{iq})$ is a $n_i \times qD$ matrix, and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_q')'$.

For a prespecified B-splines basis matrix $\boldsymbol{Z}_i$, estimating the main effects and interactions coefficients $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ and the nonparametric components $\boldsymbol{F}_i$ are converted to estimating the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\gamma}', \boldsymbol{\eta}')'$.

Before presenting the optimization and estimation procedure, we provide some basic assumptions for the SHAPLM as follows:

(A1) The covariates u_{ijl} are uniformly bound with $u_{ijl} \in [a, b]$ for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ and $l = 1, \ldots, q$.

(A2) The marginal density function $p_{ijl}(\cdot)$ of u_{ijl} is bounded away from 0 and ∞ on its support $[a, b]$ for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ and $l = 1, \ldots, q$.

(A3) The joint density function of any pair of u_{ijl} and $u_{ij'l'}$, $p_{ijj'll'}(\cdot, \cdot)$, is bounded away from 0 and ∞ on its support $[a, b]^2$ and has continuous partial derivatives for $i = 1, \ldots, m$, $j, j' = 1, \ldots, n_i$ and $l, l' = 1, \ldots, q$.

(A4) The nonparametric functions $f_l(\cdot)$ are twice continuously differentiable, and $\int_a^b \{f_l''(u)\}^2 du$ is finite and $Ef_l(u_{ijl}) = \int_a^b f_l(u)du = 0$ for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ and $l = 1, \ldots, q$.

(A5) The eigenvalues of the true covariance matrices $\boldsymbol{\Sigma}_i$ are uniformly bounded away from 0 and ∞ for $i = 1, \ldots, m$.

Similar assumptions have been made in Huang, Zhang and Zhou [16] and Cheng, Zhou and Huang [8] when considering the (additive) partially linear model for longitudinal data. Assumptions ***A1*−*A4*** are standard smoothness conditions for the B-splines basis approximation and ensure identifiability of the additive nonparametric components. ***A5*** is a practical condition for the within-subject correlation structure of longitudinal data.

3. OPTIMIZATION AND ESTIMATION

3.1 Penalized likelihood

Denote the norm $\|\boldsymbol{v}\|_2 = (\boldsymbol{v}'\boldsymbol{v})^{1/2}$ and $\|\boldsymbol{v}\|_{\boldsymbol{R}} = (\boldsymbol{v}'\boldsymbol{R}\boldsymbol{v})^{1/2}$ for a column vector $\boldsymbol{v} \in \mathbb{R}^d$ with $d \geq 1$ and a positive definite matrix $\boldsymbol{R}$. To achieve the purpose of variable selection for SHAPLM (6), we consider the penalized likelihood criterion of minimizing the penalized log-likelihood function

$$
\begin{aligned}
\sum_{i=1}^{m} \|\boldsymbol{Y}_i - \mu\mathbf{1}_{n_i} - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{F}_i\mathbf{1}_q - (\boldsymbol{X}_i :: \boldsymbol{X}_i)\boldsymbol{\alpha} \\
- (\boldsymbol{X}_i :: \boldsymbol{F}_i)\boldsymbol{\pi}\|_{\boldsymbol{W}_i}^2 + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\eta})
\end{aligned} \tag{8}
$$

where $\boldsymbol{W}_i = \boldsymbol{V}_i^{-1}$ with $\boldsymbol{V}_i$ being the working covariance matrix, a substitute of the true unknown covariance matrix $\boldsymbol{\Sigma}_i$, and $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\eta}) = \sum_k p_{\lambda_\beta, r}(|\beta_k|) + \sum_l p_{\lambda_\phi, r}(\|\boldsymbol{\phi}_l\|_2) + \sum_{k<k'} p_{\lambda_\gamma, r}(|\gamma_{k,k'}|) + \sum_{k,l} p_{\lambda_\eta, r}(|\eta_{k,l}|)$ is the summation of penalties with the tuning parameters $\boldsymbol{\lambda} = (\lambda_\beta, \lambda_\phi, \lambda_\gamma, \lambda_\eta)'$. In theory, different data-dependent tunings for $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ can be applied, but for computational simplicity, we set the tuning parameters equal and select a single $\lambda = \lambda_\beta = \lambda_\phi = \lambda_\gamma = \lambda_\eta$. Note that there is a grouping structure in $\boldsymbol{\phi}_l$ because $f_l(\cdot) \equiv 0$ is equivalent to all elements in $\boldsymbol{\phi}_l$ are 0, $l = 1, \ldots, q$. In this paper we use the minimax concave penalty, or MCP [32, 6], for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and Group MCP [13] for $\boldsymbol{\phi}$, resulting (8) to be a partial group MCP problem. The MCP penalty function is defined on $[0, +\infty)$ by

$$
p_{\lambda,r}(\theta) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2r}, & if\ \theta \leq r\lambda \\ \dfrac{1}{2}r\lambda^2, & if\ \theta > r\lambda \end{cases} \tag{9}
$$

with the tuning parameter $\lambda \geq 0$ and regularization parameter $r > 1$ which is used to control the concavity of $p_{\lambda,r}(\theta)$. We use $r = 3$ for simplicity in our simulation studies as suggested in Breheny and Huang [6]. The Group MCP for $\boldsymbol{\phi}$ is actually imposing the MCP on the norm $\|\boldsymbol{\phi}_l\|_2$, $l = 1, \ldots, q$, which selects the relevant groups of B-splines bases, and thus identifies the relevant nonparametric components. The MCP and group MCP problems can be well solved by the coordinate descent and group coordinate descent algorithms [6, 13, 14].

3.2 Iterative estimation procedures

Choi, Li and Zhu [9] considered a purely parametric regression model with main effects and all two-way interactions, and proposed an iterative estimation approach between the parameters β_k and $\gamma_{kk'}$. Similarly, we proposed to iteratively estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ for SHAPLM (6) based on the B-splines basis approximation of the nonparametric components. For a given $\boldsymbol{W}_i$ and the prespecified B-splines basis matrix $\boldsymbol{Z}_i$, the algorithm for solving the penalized likelihood problem (8) is described as follows:

Step 1: Initialization. Set $\hat{\boldsymbol{\gamma}}^{(0)} = 0$ and $\hat{\boldsymbol{\eta}}^{(0)} = 0$, and then estimate $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\phi}}^{(0)}$ using the ordinary least square. The intercept is estimated as $\hat{\mu}^{(0)} = \frac{1}{N}\sum_{i=1}^{m} \|\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{Z}_i\hat{\boldsymbol{\phi}}^{(0)}\|_1$ with $N = \sum_{i=1}^{m} n_i$. Let the iteration index $t = 1$.

Step 2: Update $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\eta}}$. Obtain $\hat{\boldsymbol{F}}_i^{(t)}$ from $\hat{\boldsymbol{\phi}}^{(t)}$ and (7), and let

$$\tilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \hat{\mu}^{(t)}\mathbf{1}_{n_i} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{Z}_i\hat{\boldsymbol{\phi}}^{(t)},$$
$$\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i :: \boldsymbol{X}_i)diag(\widehat{\beta_1}^{(t)}\widehat{\beta_2}^{(t)}, \ldots, \widehat{\beta_{p-1}}^{(t)}\widehat{\beta_p}^{(t)}),$$
$$\tilde{\boldsymbol{Z}}_i = (\boldsymbol{X}_i :: \hat{\boldsymbol{F}}_i^{(t)})diag(\hat{\boldsymbol{\beta}}^{(t)'}, \ldots, \hat{\boldsymbol{\beta}}^{(t)'}),$$

where $diag(a_1, \ldots, a_n)$ represents a $n \times n$ diagonal matrix with the diagonal elements $a_1, \ldots, a_n$. Then we obtain $\hat{\boldsymbol{\gamma}}^{(t+1)}$ and $\hat{\boldsymbol{\eta}}^{(t+1)}$ by minimizing

$$\sum_{i=1}^{m} \|\tilde{\boldsymbol{Y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\gamma} - \tilde{\boldsymbol{Z}}_i\boldsymbol{\eta}\|^2_{\boldsymbol{W}_i} + \sum_{k<k'} p_{\lambda,r}(|\gamma_{k,k'}|) + \sum_{k,l} p_{\lambda,r}(|\eta_{k,l}|),$$

which is a general penalized least square problem with response $\tilde{\boldsymbol{Y}}_i$ and covariates $(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Z}}_i)$ and can be solved by MCP.

Step 3: Update $\hat{\boldsymbol{\phi}}$. Let

$$\tilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \hat{\mu}^{(t)}\mathbf{1}_{n_i} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{X}_i :: \boldsymbol{X}_i)(\widehat{\gamma_{1,2}}^{(t)}\widehat{\beta_1}^{(t)}\widehat{\beta_2}^{(t)}, \ldots, \widehat{\gamma_{p-1,p}}^{(t)}\widehat{\beta_{p-1}}^{(t)}\widehat{\beta_p}^{(t)})'$$

and rearrange $(\boldsymbol{X}_i :: \boldsymbol{F}_i)\boldsymbol{\pi}$ as $(\boldsymbol{X}_i :: \boldsymbol{F}_i)\boldsymbol{\pi} = (\boldsymbol{X}_i :: \boldsymbol{Z}_{i1}\boldsymbol{\phi}_1, \ldots, \boldsymbol{X}_i :: \boldsymbol{Z}_{iq}\boldsymbol{\phi}_q)(\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_q')' = \boldsymbol{A}_i\boldsymbol{\phi}$ where $\boldsymbol{A}_i = \boldsymbol{A}_i(\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is a $n_i \times D$ matrix and the function of $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Then we obtain $\boldsymbol{\phi}^{(t+1)}$ by minimizing

$$\sum_{i=1}^{m} \|\tilde{\boldsymbol{Y}}_i - \boldsymbol{Z}_i\boldsymbol{\phi} - \boldsymbol{A}_i(\boldsymbol{X}_i, \boldsymbol{Z}_i, \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\gamma}}^{(t+1)})\boldsymbol{\phi}\|^2_{\boldsymbol{W}_i} + \sum_{l} p_{\lambda,r}(\|\boldsymbol{\phi}_l\|_2)$$

which is a general group variable selection problem with response $\tilde{\boldsymbol{Y}}_i$ and covariates $(\boldsymbol{Z}_i, \boldsymbol{A}_i(\boldsymbol{X}_i, \boldsymbol{Z}_i, \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\gamma}}^{(t+1)}))$ and can be solved by Group MCP.

Step 4: Update $\hat{\boldsymbol{\beta}}$. We get $\hat{\boldsymbol{f}}_{il}^{(t+1)}$, $l = 1, \ldots, q$ and $\hat{\boldsymbol{F}}_i^{(t+1)}$ from $\hat{\boldsymbol{\phi}}^{(t+1)}$, and let $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)}$. For each $k \in \{1, \ldots, p\}$, let

$$\tilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \hat{\mu}^{(t)}\mathbf{1}_{n_i} - \sum_{k' \neq k} \boldsymbol{X}_{i(k')}\widehat{\beta_{k'}}^{(t+1)} - \boldsymbol{Z}_i\hat{\boldsymbol{\phi}}^{(t+1)} - \sum_{k'<k'',k',k'' \neq k} (\boldsymbol{X}_{i(k')} :: \boldsymbol{X}_{i(k'')})\widehat{\gamma_{k',k''}}^{(t+1)}\widehat{\beta_{k'}}^{(t+1)}\widehat{\beta_{k''}}^{(t+1)} - (\boldsymbol{X}_{i(-k)} :: \hat{\boldsymbol{F}}_i^{(t+1)})\tilde{\boldsymbol{\pi}},$$
$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_{i(k)} + \sum_{k'<k} (\boldsymbol{X}_{i(k')} :: \boldsymbol{X}_{i(k)})\widehat{\gamma_{k',k}}^{(t+1)}\widehat{\beta_{k'}}^{(t+1)} + \sum_{k'>k} (\boldsymbol{X}_{i(k)} :: \boldsymbol{X}_{i(k')})\widehat{\gamma_{k,k'}}^{(t+1)}\widehat{\beta_{k'}}^{(t+1)} + \sum_{l=1}^{q} (\boldsymbol{X}_{i(k)} :: \hat{\boldsymbol{f}}_{il}^{(t+1)})\hat{\eta}_{kl}^{(t+1)}$$

where $(a_1, \ldots, a_n)' :: (b_1, \ldots, b_n)' = (a_1b_1, \ldots, a_nb_n)'$, $\boldsymbol{X}_{i(k)}$ represents the kth column of $\boldsymbol{X}_i$, $\boldsymbol{X}_{i(-k)}$ represents the matrix of removing the kth column of $\boldsymbol{X}_i$, and $\tilde{\boldsymbol{\pi}} = (\tilde{\boldsymbol{\pi}}_1', \ldots, \tilde{\boldsymbol{\pi}}_q')'$ with $\tilde{\boldsymbol{\pi}}_l$ representing the vector of removing the kth element of $(\hat{\eta}_{1l}^{(t+1)}\widehat{\beta_1}^{(t+1)}, \ldots, \hat{\eta}_{pl}^{(t+1)}\widehat{\beta_p}^{(t+1)})'$, $l = 1, \ldots, q$. Then we have

$$\widehat{\beta_k}^{(t+1)} = argmin_{\beta_k} \sum_{i=1}^{m} \|\tilde{\boldsymbol{Y}}_i - \tilde{\boldsymbol{X}}_i\beta_k\|^2_{\boldsymbol{W}_i} + p_{\lambda,r}(|\beta_k|),$$

which is a simple MCP problem with only one parameter β_k.

Step 5: Update $\hat{\mu}$.

$$\hat{\mu}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{m} \|\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{Z}_i\hat{\boldsymbol{\phi}}^{(t+1)} - (\boldsymbol{X}_i :: \boldsymbol{X}_i)\hat{\boldsymbol{\alpha}}^{(t+1)} - (\boldsymbol{X}_i :: \hat{\boldsymbol{F}}_i^{(t+1)})\hat{\boldsymbol{\pi}}^{(t+1)}\|_1$$

Step 6: Let $t = t+1$, and iterate Step 2 through Step 5 until convergence to obtain the final estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\eta}}')'$. The convergence criterion is $\|\hat{\boldsymbol{\theta}}^{(t+1)} - \hat{\boldsymbol{\theta}}^{(t)}\|_2 < 10^{-3}$.

The convergence of the iterative processes is guaranteed as discussed in Choi, Li and Zhu [9].

3.3 Choice of tuning parameter

To implement the proposed approach, we need to choose the tuning parameter $\lambda = \lambda_\beta = \lambda_\phi = \lambda_\gamma = \lambda_\eta$ appropriately. Various criterions such as AIC, BIC, Generalized

Cross-Validation (GCV), and K-fold Cross-Validation have been proposed to select the tuning parameter. It is known that under general conditions, BIC is consistent for model selection while AIC is not when the true model belongs to the class of models considered [17, 29]. In this paper, we use the BIC criterion defined as

$$BIC(\lambda) = Nlog(RSS_\lambda) + logN \times (df_\lambda) \tag{10}$$

where RSS_λ is the residual sum of squares of the selected model and df_λ is the degrees of freedom for a given tuning parameter λ, and $N = \sum_{i=1}^{m} n_i$ is the total sample size. df_λ is often taken as the number of nonzero coefficients of the fitted model, and in our approach, it is the total numbers of nonzero coefficients in $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. We search in a grid of λ's and obtain the solution of (8) for each λ, and then we select the λ minimizing the BIC criterion (10) and get the corresponding solution.

3.4 Specification of working covariance matrix

In (8), the most efficient estimation of the working covariance matrix $\boldsymbol{V}_i$ is the true covariance matrix $\boldsymbol{\Sigma}_i$, which is usually unknown in practice. When the working covariance matrix is misspecified, the resulting estimate is still consistent, but not efficient [20, 12]. In order to improve the estimation efficiency of the regression parameters and reduce the bias of the semiparametric estimate for longitudinal data, it's essential to specify the working covariance matrix correctly [27, 7]. Here we first assume balanced data with $n_i = n$ for all i and propose the maximum likelihood estimation for $\boldsymbol{V}_i$, and then present the implementation with unbalanced data later.

3.4.1 Balanced data

Under the assumption of balanced data, we assume $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_0$ for $i = 1, 2, \ldots, m$. The log-likelihood function is given by

$$\begin{aligned} &-\frac{m}{2}log|\boldsymbol{V}_0| - \frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{Y}_i - \mu\mathbf{1}_{n_i} - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{F}_i\mathbf{1}_q \\ &\quad - (\boldsymbol{X}_i :: \boldsymbol{X}_i)\boldsymbol{\alpha} - (\boldsymbol{X}_i :: \boldsymbol{F}_i)\boldsymbol{\pi}\|^2_{\boldsymbol{V}_0^{-1}} \end{aligned} \tag{11}$$

after dropping constant terms. Then from Anderson [3], the maximum likelihood estimation of $\boldsymbol{V}_i = \boldsymbol{V}_0$ is obtained by maximizing (11) as

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{Y}_i - \hat{\mu}\mathbf{1}_{n_i} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{F}}\mathbf{1}_q - (\boldsymbol{X}_i :: \boldsymbol{X}_i)\hat{\boldsymbol{\alpha}} \\ &\quad - (\boldsymbol{X}_i :: \hat{\boldsymbol{F}}_i)\hat{\boldsymbol{\pi}})(\boldsymbol{Y}_i - \hat{\mu}\mathbf{1}_{n_i} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{F}}\mathbf{1}_q \\ &\quad - (\boldsymbol{X}_i :: \boldsymbol{X}_i)\hat{\boldsymbol{\alpha}} - (\boldsymbol{X}_i :: \hat{\boldsymbol{F}}_i)\hat{\boldsymbol{\pi}})' \end{aligned} \tag{12}$$

To gradually improve the estimation efficiency, we iterate the penalized likelihood estimation procedure (8) of parameters and the maximum likelihood estimation procedure (12) of the working covariance matrix until convergence. The working independence covariance matrix is used as the initial estimate of $\boldsymbol{V}_i$, that is, $\hat{\boldsymbol{V}}_i^{(0)} = \boldsymbol{I}_{n_i}$.

3.4.2 Unbalanced data

In practice, longitudinal data may be highly unbalanced or irregular due to missing observations. We assume the completely observed time is n, and define the $n \times n_i$ transformation matrix $\boldsymbol{T}_i$ for the ith subject by removing the columns corresponding to the missing observations of the identity matrix $\boldsymbol{I}_n$. Let $\boldsymbol{Y}_i^* = \boldsymbol{T}_i\boldsymbol{Y}_i$, $\boldsymbol{X}_i^* = \boldsymbol{T}_i\boldsymbol{X}_i$ and $\boldsymbol{F}_i^* = \boldsymbol{T}_i\boldsymbol{F}_i$, thus the unbalanced data can be transformed to balanced data. Then we can estimate the working covariance matrix corresponding to the fully observed subjects, $\boldsymbol{V}_0$, from (12), and let $\hat{\boldsymbol{V}}_i = \boldsymbol{T}_i'\hat{\boldsymbol{V}}_0\boldsymbol{T}_i$. A similar solution for implementation with unbalanced data can be seen in Zhou and Qu [34].

4. SIMULATION STUDIES

In this section, we conduct simulation studies to investigate the finite sample performance of our proposed estimation process for SHAPLM. Similar to Bien, Taylor and Tibshirani [5] and Lim and Hastie [21], we consider the following four different setups:

Case I: Truth is hierarchical and has both main effects and interactions: $\beta_k = 0 \to \alpha_{kk'} = 0, \pi_{kl} = 0$ for all k, k' and l.

Case II: Truth is hierarchical and has only main effects: $\alpha_{kk'} = 0$ and $\pi_{kl} = 0$ for all k, k' and l.

Case III: Truth is anti-hierarchical and has both main effects and interactions: $\beta_k = 0 \to \alpha_{kk'} \neq 0, \pi_{kl} \neq 0$ for some k, k' and l.

Case IV: Truth is anti-hierarchical and has only interactions: $\beta_k = 0$ and $f_l(\cdot) \equiv 0$ for all k and l.

In all cases, we assumed there are $p = 15$ parametric covariates and $q = 10$ nonparametric covariates, and the last 10 parametric covariates and last 7 nonparametric covariates have no effects on the response in the true model, that is, $\beta_k = 0$ for $6 \leq k \leq 15$ and $f_l(\cdot) \equiv 0$ for $4 \leq l \leq 10$. The covariates x_{ijk}, $k = 1, \ldots, 15$ were generated independently from a normal distribution $N(0, 3)$. u_{ijl}, $l = 1, \ldots, 10$ were generated from a uniform distribution on $[-3, 3]$. Several forms of Gaussian processes of ϵ_{ij} were considered for modeling various within-subject correlation in previous studies, and we generated ϵ_{ij} from a normal distribution with mean 0, variance 1 and exponentially decaying correlation $corr(\epsilon_{ij_1}, \epsilon_{ij_2}) = exp(-|j_1 - j_2|)$ as suggested in Diggle et al. [10] and Fan and Li [12]. Particularly, for the nonparametric part, we define three smooth functions on $[-3, 3]$ as $g_1(u) = -3sin(\pi u - \pi)$, $g_2(u) = 8.85(\frac{u}{3} + 0.2)^2 - exp(-\frac{2}{3}u + 0.6)$ and $g_3(u) = -(u-2)^2 + 2u + 7$ which satisfy $\int_{-3}^{3} g_i(u)du = 0, i = 1, 2, 3$. Cubic B-splines were used to approximate the nonparametric functions with the number of basis functions D

Table 1. Simulation study: coefficients and nonparametric components of the true model

	β_1	β_2	β_3	β_4	β_5	f_1	f_2	f_3	α_{12}	α_{13}	α_{14}	α_{23}	α_{24}	π_{31}	π_{41}	π_{22}	π_{42}	π_{52}
Case I	2	3	3	4	4	g_1	g_2	g_3	0.8	1	2	1	1	1	2	0.8	1	1
Case II	2	3	3	4	4	g_1	g_2	g_3	0	0	0	0	0	0	0	0	0	0
	β_1	β_2	β_3	β_4	β_5	f_1	f_2	f_3	α_{12}	α_{15}	α_{23}	α_{25}	π_{11}	π_{21}	π_{31}	π_{51}	π_{33}	π_{43}
Case III	0	3	3	4	0	g_1	g_2	0	1	0.8	1	0.6	1	0.8	0.5	1	0.6	1
Case IV	0	0	0	0	0	0	0	0	1	0.8	1	0.6	1	0.8	0.5	1	0.6	1

chosen from $8, 10, 12, \ldots, 20$, and we selected the one giving the minimal prediction error (PE). The main effects and interactions coefficients and the nonparametric components of the 4 cases are shown in Table 1, and the effects excluded from Table 1 are exactly zeros.

For each simulation setup, we generated $K = 100$ datasets consisting of both balanced and unbalanced situations with $m = 100$ or 300 subjects and $n = 6$ observations for the fully observed subjects. The unbalanced datasets were created from the balanced ones by keeping 30% of the subjects with complete observations and the other 70% subjects with a probability of 0.3 to be missing in each observation.

4.1 Variable selection performance

To assess the variable selection performance of our proposed method, we consider sensitivity, which is the proportion of number of selected relevant terms to total number of relevant terms, and specificity, which is the proportion of number of unselected irrelevant terms to total number of irrelevant terms.

The average sensitivity and specificity computed based on the 100 datasets corresponding to the parametric main effects, the nonparametric main effects, the interactions between parametric terms, the interactions between parametric and nonparametric terms, and the overall effects, respectively, represented by "Main. X", "Main. U", "Inter. X", "Inter. X&U" and "Overall", are presented in Table 2.

We can see from Table 2 that our approach has a fairly good variable selection performance for each case, and the unbalanced situation is a little less effective but almost comparable since over 20% of the observations are missing. When the sample size increases, such as when $m = 300$, the proposed approach shows a consistent model selection trend, especially when the truth satisfies the strong hierarchy. Even when the true model does not obey the hierarchical constraint, the sensitivity and specificity are still considerably high even though the selected model is always wrong. Under this circumstance, our approach either estimates some irrelevant main effects with small nonzero coefficients in order to include the significant interactions, or estimates the relevant interactions with zero coefficients in order to exclude the associated irrelevant main effects, resulting in some reduction of sensitivity of the interactions and specificity of the main effects, but sensitivity of the main effects and specificity of the interactions are still quite high.

4.2 Prediction and estimation performance

The overall prediction accuracy is measured by the prediction error (PE), which is defined as

$$PE = \frac{1}{N}\sum_{i=1}^{m} \|\hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i\|_2^2$$

We assess performance of the estimators of main effects coefficients $\boldsymbol{\beta}$ and and interactions coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ by mean squares error (MSE). The MSE of $\hat{\boldsymbol{\beta}}$ based on the $K = 100$ simulated datasets is estimated by

$$MSE(\hat{\boldsymbol{\beta}}) = E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$$

and $MSE(\hat{\boldsymbol{\alpha}})$ and $MSE(\hat{\boldsymbol{\pi}})$ defined similarly. The finite sample performance of the estimators of nonparametric components is evaluated by the square root of average squared errors (RASE)

$$RASE(\hat{\boldsymbol{f}}) = \sqrt{\frac{1}{N}\sum_{l=1}^{q}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \|\hat{f}_l(u_{ijl}) - f_l(u_{ijl})\|_2^2}$$

as suggested in Fan and Li [12] and Ai, You and Zhou [2]. The $MSE(\hat{\boldsymbol{\beta}})$, $MSE(\hat{\boldsymbol{\alpha}})$, $MSE(\hat{\boldsymbol{\pi}})$, and means and standard deviations of PE and $RASE(\hat{\boldsymbol{f}})$ computed based on the 100 datasets are shown in Table 3.

Table 3 shows that in case I and II where the truth obeys the strong hierarchy, the overall prediction accuracy is pretty high, and the estimations for both main effect and interaction coefficients and nonparametric components are really efficient, especially when the sample size is comparatively large. However, when the true model is anti-hierarchical, the prediction and estimation accuracy reduces. We can find that $MSE(\hat{\boldsymbol{\beta}})$, $MSE(\hat{\boldsymbol{\alpha}})$ and $RASE(\hat{\boldsymbol{f}})$ in case III and IV are still low even though much bigger compared to case I and II, but $MSE(\hat{\boldsymbol{\pi}})$ is extremely high, which contributes the most to the high PE. As shown in Table 2, the sensitivity of interactions between parametric and nonparametric terms is relatively low in case III and IV, thus the underestimate of $\boldsymbol{\pi}$ and strong effects of the

Table 2. Simulation results: sensitivity and specificity of the fitted models in each case

			Balance					Unbalance				
Case	m		**Overall**	**Main. X**	**Main. U**	**Inter. X**	**Inter. X&U**	**Overall**	**Main. X**	**Main. U**	**Inter. X**	**Inter. X&U**
I	100	Sensitivity	0.998	1	0.997	1	0.996	0.995	1	0.990	1	0.988
		Specificity	0.987	0.988	1	1	0.978	0.981	0.970	0.999	1	0.968
	300	Sensitivity	1	1	1	1	1	1	1	1	1	1
		Specificity	1	0.993	1	1	1	0.999	0.990	1	1	0.998
II	100	Sensitivity	1	1	1	-	-	1	1	1	-	-
		Specificity	0.955	0.779	1	0.964	0.957	0.959	0.795	1	0.965	0.964
	300	Sensitivity	1	1	1	-	-	1	1	1	-	-
		Specificity	0.968	0.780	1	0.969	0.978	0.968	0.772	1	0.968	0.979
III	100	Sensitivity	0.847	1	0.985	0.678	0.838	0.820	1	0.955	0.653	0.797
		Specificity	0.961	0.826	0.875	1	0.950	0.966	0.840	0.875	1	0.958
	300	Sensitivity	0.901	1	1	0.820	0.872	0.895	1	1	0.790	0.878
		Specificity	0.969	0.832	0.875	1	0.965	0.964	0.832	0.875	1	0.954
IV	100	Sensitivity	0.719	-	-	0.733	0.710	0.708	-	-	0.745	0.683
		Specificity	0.940	0.669	0.838	1	0.933	0.943	0.671	0.843	1	0.939
	300	Sensitivity	0.789	-	-	0.805	0.778	0.808	-	-	0.818	0.802
		Specificity	0.950	0.666	0.821	1	0.953	0.944	0.661	0.818	1	0.944

Table 3. Simulation results: PE, MSE and RASE in each case

		Balance					Unbalance				
Case	m	**PE**	$\boldsymbol{MSE}(\hat{\boldsymbol{\beta}})$	$\boldsymbol{MSE}(\hat{\boldsymbol{\alpha}})$	$\boldsymbol{MSE}(\hat{\boldsymbol{\pi}})$	$\boldsymbol{RASE}(\hat{\boldsymbol{f}})$	**PE**	$\boldsymbol{MSE}(\hat{\boldsymbol{\beta}})$	$\boldsymbol{MSE}(\hat{\boldsymbol{\alpha}})$	$\boldsymbol{MSE}(\hat{\boldsymbol{\pi}})$	$\boldsymbol{RASE}(\hat{\boldsymbol{f}})$
I	100	5.625(2.376)	0.535	0.049	1.006	1.569(0.438)	6.285(2.389)	0.776	0.075	0.980	1.689(0.439)
	300	2.105(0.382)	0.141	0.008	0.342	0.871(0.166)	2.239(0.500)	0.183	0.010	0.372	0.893(0.151)
II	100	1.160(0.361)	0.009	0.001	0.039	0.554(0.195)	1.166(0.470)	0.011	0.002	0.051	0.615(0.221)
	300	1.042(0.126)	0.003	0.000	0.026	0.307(0.115)	1.060(0.164)	0.003	0.001	0.010	0.356(0.135)
III	100	38.58(11.11)	1.257	1.565	16.09	3.949(0.736)	41.61(13.33)	1.314	1.641	16.24	4.108(0.748)
	300	32.51(7.077)	1.014	1.091	13.25	3.472(0.742)	34.09(7.436)	1.090	1.134	12.11	3.710(0.689)
IV	100	75.18(14.75)	1.245	1.416	17.61	3.185(0.655)	75.96(14.39)	1.208	1.408	19.52	2.998(0.699)
	300	67.77(11.31)	1.142	1.125	15.08	3.107(0.637)	67.16(11.86)	1.183	1.134	18.09	3.024(0.524)

Table 4. Details of the predictors

Aspect	Variable	Definition	Aspect	Variable	Definition
Cash flow capacity	CF_1	the cash flow ratio	**Short-term debt-paying ability**	SD_1	the current ratio
	CF_2	the revenue cash ratio		SD_2	the quick ratio
	CF_3	the sales receive cash ratio		SD_3	the working capital ratio
	CF_4	the surplus cash coverage ratio		SD_4	the working capital
	CF_5	the net cash flow per share			
Operation capacity	OP_1	the accounts receivable turnover	**Profitability**	PR_1	the operating margin
	OP_2	the inventory turnover		PR_2	the net profit to sales ratio
	OP_3	the accounts payable turnover		PR_3	the assets return rate
	OP_4	the working capital turnover		PR_4	the return on assets (ROA)
	OP_5	the current assets turnover		PR_5	the return on current assets
	OP_6	the fixed assets turnover		PR_6	the return on fixed assets
	OP_7	the long-term asset turnover		PR_7	the return on equity (ROE)
	OP_8	the total assets turnover		PR_8	the earnings before interest and tax
	OP_9	the stockholders equity turnover		PR_9	the earnings per share (EPS)
	OP_{10}	the operating income per share			
Development ability	DE_1	the rate of capital accumulation	**Share structure**	SS_1	the top five shareholding ratio
	DE_2	the growth rate of fixed assets		SS_2	the gap between the first and the second shareholding ratios
	DE_3	the growth rate of total assets			
	DE_4	the growth rate of operating income		SS_3	the supervisor shareholding ratio
Long-term debt-paying ability	LD_1	the asset-liability ratio	**Board structure**	BS_1	the total number of directors
	LD_2	the current assets ratio		BS_2	the CEO duality
	LD_3	the fixed assets ratio		BS_3	the proportion of independent directors
	LD_4	the current debt ratio	**Risk factor**	RK_1	the consolidated leverage
	LD_5	the long-term debt ratio			
	LD_6	the equity to debt ratio			
	LD_7	the tangible net debt ratio			
	LD_8	the debt to equity price ratio			

nonparametric components due to their high variances work together and lead to a high PE. It demonstrates that our proposed method for SHAPLM is to be improved in estimating the interactions between parametric and nonparametric terms when they dissatisfy the hierarchical constraint and the nonparametric effects are strong.

5. APPLICATION

To illustrate the effectiveness of proposed approach on real longitudinal data, we considered the financial indicator data of the China Stock Market from the China Stock Market and Accounting Research Database (CSMAR), published by GTA Information Technology Company (http://www.gtarsc.com/). We aim to explore which financial indicators and interactions have influential effects on the price-to-earnings ratio (P/E ratio), which is defined as the market price per share divided by annual earnings per share. After excluding those abnormal stocks with P/E ratios exceeding 500, we obtain the dataset containing 149 stocks of the mechanical listed companies from 2009 to 2011. We have 47 predictors of 9 aspects detailed in Table 4, and consider the natural logarithm of P/E ratio as the response. Among the 47 predictors, BS_2 is binary and the others are continuous, so BS_2 is not suitable to be modelled nonparametricly. The predictors in the first 6 aspects in Table 4 reflect the financial statements of a listed company and are often considered linearly correlated with the logarithmic P/E ratio, and thus are modelled as parametric components. The effects of share structure and board structure on P/E ratio were rarely studied and unclear, so we modeled SS_1, SS_2, SS_3, BS_1 and BS_3 as nonparametric components. The Risk factor was often considered to affect logarithmic P/E ratio

Table 5. Selected main effects and interactions in the application

Selected terms based on all data				Selected terms based on 100 bootstrap samples			
Main effects	**Coefficients**	**Interactions**	**Coefficients**	**Main effects**	**Frequency (%)**	**Interactions**	**Frequency (%)**
CF_5	0.065	$CF_5 :: BS_2$	0.230	LD_8	100	$PR_9 :: LD_8$	82
OP_5	-0.139	$OP_5 :: PR_7$	0.137	LD_7	92	$LD_8 :: BS_2$	74
PR_7	-0.481	$PR_7 :: BS_2$	0.067	PR_9	91	$CF_5 :: BS_2$	57
PR_9	-0.307	$PR_9 :: LD_8$	0.172	DE_4	86	$LD_7 :: BS_2$	51
LD_7	0.315	$LD_7 :: BS_2$	0.195	BS_2	76	$PR_9 :: BS_2$	43
LD_8	-0.563	$LD_8 :: BS_2$	-0.443	CF_5	69	$PR_5 :: LD_8$	41
DE_4	0.115	$PR_9 :: SS_3$	-1.392	PR_7	67	$OP_5 :: PR_7$	40
BS_2	-0.121	$LD_8 :: SS_3$	2.964	OP_8	52	$PR_7 :: BS_2$	33
SS_3	-	$OP_5 :: RK_1$	0.991	CF_4	49	$LD_8 :: SS_3$	77
RK_1	-	$PR_7 :: RK_1$	-2.444	OP_5	48	$PR_9 :: SS_3$	64
				PR_5	40	$PR_9 :: RK_1$	58
				SS_3	100	$LD_8 :: RK_1$	53
				RK_1	100	$PR_7 :: RK_1$	44
						$OP_5 :: RK_1$	35

nonlinearly, so RK_1 is also modeled nonparametricly. Thus we have $m = 149$, $n_i = n = 3$ for all $1 \le i \le m$, $p = 41$ and $q = 6$ in our application.

Before fitting the model, we standardize the response and all continuous predictors in the parametric parts to have mean 0 and variance 1, and rescale all predictors in the nonparametric parts to range between 0 and 1. As in the simulation, cubic B-splines are used. The number of basis functions D is chosen from $\{8, 10, 12, \ldots, 20\}$ by minimizing PE. The selected number of basis functions is 10. The normalized mean square error (NMSE) is $NMSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y})^2 = 0.162$. Thus the goodness-of-fit measure $R^2 = 1 - NMSE = 0.838$, indicating that the model fits fairly well.

The selected main effects and interactions and their coefficients are as shown in the left part of Table 5. We can see that the most important financial indicators include the net cash flow per share, current assets turnover, ROE, EPS, tangible net debt ratio, debt to equity price ratio, growth rate of operating income, CEO duality, supervisor shareholding ratio, and consolidated leverage, and some interactions among these indicators.

To further assess the stability of variable selection results, we carried out bootstrap analysis on the stock level, that is, either none or all three years' data of a stock will appear in a bootstrap sample, and thus the within-subject correlation is preserved. Based on 100 bootstrap samples, the main effects with selection frequency higher than 40% and interactions with selection frequency higher than 30% are summarized in the right part of Table 5. The mean of R^2s of bootstrap samples is 0.859 with a standard deviation of 0.034. We can see that the terms we select based on all data (i.e., terms appear in the left part of Table 5) have high selection frequencies (in right part of Table 5), demonstrating that our variable selection results are fairly stable.

6. DISCUSSION

In this paper, we have extended the penalization variable selection methods to fit the semiparametric additive partially linear model with strong hierarchical interactions (SHAPLM) for longitudinal data. By reparameterizing the interaction coefficients, the strong hierarchy is naturally guaranteed in our model. However, this reparameterization formulates a nonconvex optimization problem, which may lead to heavy computational burdens and cannot guarantee convergence to the global maximum, especially when the dimension of covariates is high [9]. Yuan, Joseph and Zou [30] and Bien, Taylor and Tibshirani [5] proposed to enforce hierarchy by adding linear inequality constraints in the penalties, and other studies have used nested or overlapped group penalties to achieve hierarchical sparsity, such as Zhao, Rocha and Yu [33], Radchenko and James [24] and Lim and Hastie [21]. These approaches lead to convex optimization problems, and therefore are computationally efficient, but their penalty structures are severely constrained, and their optimization processes are extremely complicated for our semiparametric models. In this paper, our penalty structure and optimization procedures are straightforward, and simulation studies indicate that our approach achieves satisfactory variable selection performance and seldom gets stuck in local maximums, similar to Choi, Li and Zhu [9].

In addition, to improve estimation efficiency which is considerably affected by the within-subject correlation in the longitudinal data, we specify the working covariance matrix using the maximum likelihood estimation rather than assuming a certain correlation structure, such as working independent, exchangeable, or autoregressive correlation structure, as in most previous studies. The working covariance matrix is adjusted gradually through the iterative estimation procedure. So our approach is data-driven and more flexible in practice.

Finally, the simulations show that when the true model seriously violates the strong hierarchy, our method still performs excellently in variable selection. However, the prediction accuracy is relatively low. It is mainly caused by underestimating the interactions between parametric and nonparametric terms and also by the high variances of nonparametric components. This may be improved by a better basis approximation for the nonparametric components. More preferable solutions are left to future work.

Received 27 February 2014

REFERENCES

[1] Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *The Annals of Statistics* **31**(1) 258–283. MR2157803

[2] Ai, C., You, J. and Zhou, Y. (2014). Estimation of fixed effects panel data partially linear additive regression models. *The Econometrics Journal* **10.1111/ectj.12011**. MR3171213

[3] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis, 3rd edition*. Wiley, New York. MR1990662

[4] Bien, J., Simon, N. and Tibshirani, R. (2012). A lasso for hierarchical testing of interactions. *arXiv preprint* **arXiv:1211.1344**.

[5] Bien, J., Taylor, J. and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Tha Annals of Statistics* **41**(3) 1111–1141. MR3113805

[6] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5**(1) 232–253. MR2810396

[7] Carroll, R., Maity, A. and Mammen, E. (2009). Efficient semiparametric marginal estimation for the partially linear additive model for longitudinal/clustered data. *Statistics in biosciences* **1**(1) 10–31.

[8] Cheng, G., Zhou, L. and Huang, J. Z. (2014). Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data. *Bernoulli* **20**(1) 141–163. MR3160576

[9] Choi, N. H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**(489) 354–364. MR2656056

[10] Diggle, P., Heagerty, P., Liang K. Y. and et al. (2002). *Analysis of longitudinal data, 2nd edition*. Oxford University Press, Oxford.

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456) 1348–1360. MR1946581

[12] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**(467) 710–723. MR2090905

[13] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**(4) 481–499. MR3025130

[14] Huang, J., Wei, F. and Ma, S. (2012). Semiparametric regression pursuit. *Statistica Sinica* **22**(4) 1403–1426. MR3027093

[15] Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**(1) 111–128.MR1888349

[16] Huang, J. Z., Zhang, L. and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scandinavian Journal of Statistics* **34**(3) 451–477. MR2368793

[17] Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association* **99**(465) 279–290. MR2061890

[18] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**(5) 1356–1378. MR1805787

[19] Li, Y., Yu, C., Qin, Y., Wang, L., Chen, J., Yi, D., Shia, B.-C. and Ma, S. (2014). Regularized receiver operating characteristic based logistic regression for grouped variable selection with composite criterion. *Journal of Statistical Computation and Simulation* **DOI:10.1080/00949655.2014.899362**.

[20] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1) 13–22. MR0836430

[21] Lim, M. and Hastie, T. (2013). Learning interactions through hierarchical group-lasso regularization. *arXiv preprint* **arXiv: 1308.2719**.

[22] Ma, S., Song, Q. and Wang, L. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli* **19**(1) 252–274. MR3019494

[23] Maity, A., Carroll, R. J., Mammen, E. and et al. (2009). Testing in semiparametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society: Series B* **71**(1) 75–96. MR2655524

[24] Radchenko, P. and James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association* **105**(492) 1541–1553. MR2796570

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**(1) 267–288. MR1379242

[26] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B* **73**(3) 273–282. MR2815776

[27] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data. *Journal of the American Statistical Association* **100**(469) 147–157. MR2156825

[28] Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statistica Sinica* **16**(4) 1423–1446. MR2327498

[29] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92**(4) 937–950. MR2234196

[30] Yuan, M., Joseph, V. R. and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics* **3**(4) 1738–1757. MR2752156

[31] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68**(1) 49–67. MR2212574

[32] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**(2) 894–942. MR2604701

[33] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Applied Statistics* **37**(6A) 3468–3497. MR2549566

[34] Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* **107**(498) 701–710. MR2980078

[35] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**(2) 301–320. MR2137327

[36] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476) 1418–1429. MR2279469

Xianbin Zeng
School of Statistics
Statistical Consulting Center
Renmin University of China
Beijing 100872
P.R. China
E-mail address: xbinzeng@163.com

Shuangge Ma
School of Statistics
Renmin University of China
Beijing 100872
P.R. China

Department of Biostatistics
Yale University
New Haven 06511
USA
E-mail address: shuangge.ma@yale.edu

Yichen Qin
Department of Operations
Business Analytics and Information Systems
University of Cincinnati
Cincinnati 45221
USA
E-mail address: qinyn@ucmail.uc.edu

Yang Li
Center for Applied Statistics
School of Statistics
Statistical Consulting Center
Renmin University of China
Beijing 100872
P.R. China
E-mail address: yang.li@ruc.edu.cn