High-dimensional generalized semiparametric model for longitudinal data

Mozhgan Taavoni
Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology
Shahrood, Iran
taavonimozhgan@yahoo.com

Abstract This paper considers the problem of estimation in the generalized semiparametric model for longitudinal data when the number of parameters diverges with the sample size. A penalization type of generalized estimating equation method is proposed, while we use the regression spline to approximate the nonparametric component. Under some regularity conditions, the resulting estimators enjoy the oracle properties, under the high-dimensional regime. Simulation studies are carried out to assess the performance of proposed method, and a real data set are analyzed for procedure demonstration.

Keywords: High-dimension; Longitudinal data; SCAD penalty; Spline.

1 Introduction

High-dimensional longitudinal data, which consist of repeated measurements on a large number of covariates, have become increasingly common. Despite the large number of covariates, it often occurs that only a subset of them is relevant for modeling the response. Inclusion of redundant variables may hinder accuracy and efficiency for both estimation and inference. Thus, it is important to develop new statistical methodology and theory of variable selection and estimation for high-dimensional longitudinal data. The literature on variable selection for longitudinal data is rather limited due to the challenges imposed by incorporating the intracluster correlation and these works commonly apply to continuous outcome data. The use of penalization techniques for discrete longitudinal data in the framework of generalized linear models (GLM) is still in the beginning. Fu (2003) proposed a generalization of the bridge and Lasso penalties to the generalized estimating equations (GEE) model. Xu and Zhu (2010) extended the independence screening method to deal with the high dimensional longitudinal GLMs. Dziak (2006) generalized the Lasso and SCAD methods to the longitudinal GLMs. The SCAD-penalized selection procedures were illustrated in Xue et al (2010). However, the aforementioned work all assume that the dimension of predictors is fixed. Xu et al. (2012) proposed a weighted least squares type function to study the longitudinal GLMs with a diverging number of parameters. Wang et al. (2012) proposed the SCAD-penalized GEE for analyzing longitudinal data with high dimensional covariates. To the best of our knowledge, regularization in the generalized semiparametric mixed models (GSMM) is neglected.

In this paper, we focus on the GSMM with longitudinal data by allowing for non-Gaussian data and nonlinear link function. We consider the case where the number of variables p is allowed to increase with the number of sample size n. Similar to the work of Wang et al. (2012), we apply the penalty function to the estimating equation objective function. Our method is rather different from their work because of including random effects and a nonparametric component in the model. We adopt spline regression to estimate the nonparametric components. The proposed penalized estimation involves the specification of the posterior distribution of the random effects, which cannot be evaluated in a closed form. However, it is possible to approximate this posterior distribution by producing random draws from a distribution using the Metropolis algorithm, which does not require the specification of the posterior distribution. We establish the asymptotic theory for the proposed method in a high-dimensional framework where the number of covariates increases with the sample size.

The article is organized as follows. Section 2, formulates the model and considers the estimation under the GEE framework. Section 3 includes selection of the regularization parameters and the model selection procedure. Furthermore, asymptotic properties of the estimators are studied. Section 4 demonstrates the effectiveness of the proposed estimation method in the GSMM. through simulation studies and illustrates the method through an application to the real data. Some concluding remarks are given in Section 5.

2 GSMM and Estimation Procedure

Consider a longitudinal study with n subjects and n_i observations over time for the ith subject. Let $\boldsymbol{u}_i$ be a $q \times 1$ vector of random effects corresponding to the ith subject, y_{ij} be an observation of the ith subject measured at time t_{ij}, and $y_{i1}, \ldots, y_{in_i}$ given $\boldsymbol{u}_i$ are conditionally independent and each $y_{ij}|\boldsymbol{u}_i$ is distributed as an exponential family distribution whose p.d.f is given by

$$p(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}_n, \phi) = \exp\left[\phi^{-1}\{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, \phi)\right], \tag{2.1}$$

where ϕ is a scale parameter, $c(.,.)$ is a function only depending on y_{ij} and ϕ, and θ_{ij} is the (scalar) canonical parameter. The conditional expectations and variances of y_{ij} given $\boldsymbol{u}_i$ are given by $\mu_{ij} = E(y_{ij}|\boldsymbol{u}_i) = b^{\cdot}(\theta_{ij})$ and $\nu_{ij} = var(y_{ij}|\boldsymbol{u}_i) = \phi b^{\cdot\cdot}(\theta_{ij})$, respectively, where $b^{\cdot}(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ and $b^{\cdot\cdot}(\theta) =$

$\frac{\partial^2 b(\theta)}{\partial \theta^2}$. In this paper, we assume that the conditional mean μ_{ij} satisfies

$$g(\mu_{ij}) \triangleq \eta_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_n + \boldsymbol{Z}_{ij}^\top \boldsymbol{u}_i + f(t_{ij}), \quad i = 1, \ldots, n;\ j = 1, \ldots, n_i, \tag{2.2}$$

where $g(.)$ is a known monotonic link function, $\boldsymbol{X}_{ij}^\top$ is a $p_n \times 1$ vector of explanatory variables, $\boldsymbol{\beta}_n$ is a $p_n \times 1$ vector of unknown parameters of the fixed effects, $\boldsymbol{Z}_{ij}^\top$ is a $q \times 1$ vector relating to the random effects, $f(.)$ is an unknown smooth function which is continuous and twice differentiable function on some finite interval. The dimension of the covariates p_n is allowed to depend on the number of subjects n. To complete the specification, assume that the random effects $\boldsymbol{u} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q\}$ independently follow a distribution, depending on parameters $\boldsymbol{\Sigma}$ as

$$\boldsymbol{u}_i \sim f_u(\boldsymbol{u}_i | \boldsymbol{\Sigma}). \tag{2.3}$$

The model defined in Eqs. (2.1)–(2.3) is referred to as generalized semiparametric mixed model (GSMM). Specific assumptions will be considered for the number of variables p_n in section 3.2.

We approximate the unspecified smooth function using

$$f(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \ldots + \alpha_d t_{ij}^d + \sum_{l=1}^{L_n} \alpha_{(d+1)+l} (t_{ij} - t_i^{(l)})_+^d = \boldsymbol{B}(t_{ij})^\top \boldsymbol{\alpha}_n,$$

where d is the degree of the polynomial component, L_n is the number of interior knots (rate of L_n will be specified in Section 3.2), $t_i^{(l)}$ is referred as knots of the ith subject, $\boldsymbol{B}(t_{ij}) = \Big(1, t_{ij}, \ldots, t_{ij}^d, \big(t_{ij} - t_i^{(1)}\big)_+^d, \ldots, \big(t_{ij} - t_i^{(L_n)}\big)_+^d\Big)$ is a $h_n \times 1$ vector of basis functions, h_n is the number of basis functions used to approximate $f(t_{ij})$, $h_n = d + 1 + L_n$, $(a)_+ = \max(0, a)$, and $\boldsymbol{\alpha}_n = (\alpha_0, \ldots, \alpha_d, \alpha_{d+1}, \ldots, \alpha_{d+L_n})^\top$ is the spline coefficients vector of dimension h. Thus, we can represent the regression model (2.2) as $\eta_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}_n + \boldsymbol{Z}_{ij}^\top \boldsymbol{u}_i + \boldsymbol{B}(t_{ij})^\top \boldsymbol{\alpha}_n$. For convenience, it can take the form $\eta_{ij} = \boldsymbol{D}_{ij}^\top \boldsymbol{\theta}_n + \boldsymbol{Z}_{ij}^\top \boldsymbol{u}_i$, where $\boldsymbol{D}_{ij} = \big(\boldsymbol{X}_{ij}^\top, \boldsymbol{B}_j(\boldsymbol{t}_i)^\top\big)^\top$ being a $(p_n + h_n) \times 1$ design matrix combining the fixed-effects and spline-effects design matrices for the jth outcome of the ith subject, and $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n^\top, \boldsymbol{\alpha}_n^\top)^\top$ is a $(p_n + h_n) \times 1$ combined regression parameters vector.

For GSMM, the classical likelihood function can be defined as

$$L(\boldsymbol{\theta}_n, \boldsymbol{\Sigma}, \phi) = \prod_{i=1}^n \int p_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i | \boldsymbol{u}_i, \boldsymbol{\theta}_n, \phi) p_{\boldsymbol{u}}(\boldsymbol{u}_i | \boldsymbol{\Sigma}) d\boldsymbol{u}_i \tag{2.4}$$

where $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in_i})^\top$, $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, and $p_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i | \boldsymbol{u}_i, \boldsymbol{\theta}_n, \phi) = \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij} | \boldsymbol{u}_i, \boldsymbol{\theta}_n, \phi)$. The log-likelihood is given by $\ell(\boldsymbol{\theta}_n, \boldsymbol{\Sigma}, \phi) = \sum_{i=1}^n \ln p_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i | \boldsymbol{u}_i, \boldsymbol{\theta}_n, \phi) + \sum_{i=1}^n \ln p_{\boldsymbol{u}_i}(\boldsymbol{u}_i | \boldsymbol{\Sigma})$. Using the Monte Carlo Newton-Raphson (MCNR) algorithm of McCulloch (1997), the optimal estimating equation for

$\boldsymbol{\theta}_n$ is given by

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{y}}\Big[n^{-1}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)}{\partial\boldsymbol{\theta}_n^{\top}}\boldsymbol{V}_i^{-1}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big)\Big]=\mathbf{0}, \tag{2.5}$$

where $\boldsymbol{\mu}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)=(\mu_{i1},\dots,\mu_{in_i})^{\top}$ and $\boldsymbol{V}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)$ is the covariance matrix of $\boldsymbol{y}_i|\boldsymbol{u}_i$. In real applications the true intracluster covariance structure is often unknown. The GEE procedure adopts a working covariance matrix, which is specified through a working correlation matrix $\boldsymbol{R}(\boldsymbol{\rho})$: $\boldsymbol{V}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)=\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\boldsymbol{R}(\boldsymbol{\rho})\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)$, where $\boldsymbol{\rho}$ is a finite dimensional parameter and $\boldsymbol{A}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)=\mathrm{diag}(\nu_{i1},\dots,\nu_{in_i})$. With the estimated working correlation matrix $\widehat{\boldsymbol{R}}\equiv\boldsymbol{R}(\widehat{\boldsymbol{\rho}})$, the (2.5) reduces to $\mathbb{E}_{\boldsymbol{u}|\boldsymbol{y}}\Big[n^{-1}\sum_{i=1}^{n}\boldsymbol{D}_i^{\top}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)$ $\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big(\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big)\Big]=\mathbf{0}$, where $\boldsymbol{D}_i=(\boldsymbol{D}_{i1}^{\top},\dots,\boldsymbol{D}_{in_i}^{\top})^{\top}$. We formally define the estimator as the solution $\widehat{\boldsymbol{\theta}}_n$ of the above estimating equations.

3 Regularization in the GSMM

In order to select important covariates and estimate them simultaneously, the (2.5) is expanded to include the penalty term $\sum_{k=1}^{p_n}p_{\lambda_n}(|\beta_{nk}|)$ which yields the following penalized log likelihood

$$\ell^p(\boldsymbol{\beta}_n,\boldsymbol{\alpha}_n,\boldsymbol{D},\phi)=\sum_{i=1}^{n}\ln p_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{\theta}_n)+\sum_{i=1}^{n}p_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\Sigma})-n\sum_{k=1}^{p_n}p_{\lambda_n}(|\beta_{nk}|), \tag{3.6}$$

where $p_{\lambda}(|\beta_{nk}|)$ is any penalty function and λ_n is a tuning parameter. Since the coefficients $\boldsymbol{\theta}_n$ depends to the first and third terms of (3.6), we propose the penalized estimating equation

$$\boldsymbol{U}_n(\boldsymbol{\theta}_n)=\boldsymbol{S}_n(\boldsymbol{\theta}_n)-q_{\lambda_n}(|\boldsymbol{\beta}_n|)^{\top}\mathrm{sign}(\boldsymbol{\beta}_n),$$

where $\boldsymbol{S}_n(\boldsymbol{\theta}_n)$ is the left term in (2.5), with $q_{\lambda_n}(|\boldsymbol{\beta}_n|)=\big(q_{\lambda_n}(|\beta_{n1}|),\dots,q_{\lambda_n}(|\beta_{np_n}|)\big)$ is a $1\times p_n$ vector of penalty functions, $\mathrm{sign}(\boldsymbol{\beta}_n)=\big(\mathrm{sign}(\beta_{n1}),\dots,\mathrm{sign}(\beta_{np_n})\big)$ with $\mathrm{sign}(a)=I(a>0)-I(a<0)$ and $q_{\lambda_n}(|\beta_{nk}|)=p_{\lambda_n}^{'}(|\beta_{nk}|)$. We use the SCAD penalty proposed by Fan and Li (2001) defined by $q_{\lambda_n}(|\beta_n|)=p_{\lambda_n}^{'}(|\beta_n|)=\lambda_n\Big\{I(|\beta_n|\leqslant\lambda_n)+\frac{(a\lambda_n-|\beta_n|)_+}{(a-1)\lambda_n}I(|\beta_n|>\lambda_n)\Big\},\quad a>2$. Our proposed estimator for $\boldsymbol{\theta}_n$ is the solution of $\boldsymbol{U}_n(\boldsymbol{\theta}_n)=\mathbf{0}$. We apply the Newton-Raphson method to solve $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n)=o(a_n)$, and get the following updating formula

$$\widehat{\boldsymbol{\theta}}_n^{(m+1)}=\widehat{\boldsymbol{\theta}}_n^{(m)}+\Big\{\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})+n\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})\Big\}^{-1}\times\Big\{\boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})+n\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})\widehat{\boldsymbol{\theta}}_n^{(m)}\Big\}. \tag{3.7}$$

Where $\boldsymbol{H}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})=\mathbb{E}_{u|y}\Big[\sum_{i=1}^{n}\boldsymbol{D}_i^{\top}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\boldsymbol{D}_i\Big]$, and for a small numbers e.g. $\epsilon=10^{-6}$, $\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}_n^{(m)})=\mathrm{diag}\Big\{\frac{q_{\lambda_n}(|\beta_{n1}|)}{\epsilon+|\beta_{n1}|},\dots,\frac{q_{\lambda_n}(|\beta_{np_n}|)}{\epsilon+|\beta_{np_n}|},\mathbf{0}_{h_n}\Big\}$. $\mathbf{0}_{h_n}$ is a zero vector of dimension h_n.

Now we outline the computational procedure used for sample generation. Let $\boldsymbol{U}$ denote the previous draw from the conditional distribution of $\boldsymbol{U}|\boldsymbol{y}$, and generate a new value u_k^* for the jth component of $\boldsymbol{U}^* = (u_1, \ldots, u_{k-1}, u_k^*, u_{k+1}, \ldots, u_{nq})$ by using the candidate distribution $p_{\boldsymbol{u}}$, accept $\boldsymbol{U}^*$ as the new value with probability $\alpha_k(\boldsymbol{U}, \boldsymbol{U}_*) = \min\Big\{1, \frac{p_{u|y}(\boldsymbol{U}^*|\boldsymbol{y},\boldsymbol{\theta}_n,\boldsymbol{D})p_u(\boldsymbol{U}|\boldsymbol{D})}{p_{u|y}(\boldsymbol{U}|\boldsymbol{y},\boldsymbol{\theta}_n,\boldsymbol{D})p_u(\boldsymbol{U}^*|\boldsymbol{D})}\Big\}$. otherwise, reject it and retain the previous value $\boldsymbol{U}$. The second term in brace can be simplified to $\frac{p_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{U}^*|\boldsymbol{y},\boldsymbol{\theta}_n,\boldsymbol{D})p_u(\boldsymbol{U}|\boldsymbol{D})}{p_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{U}|\boldsymbol{y},\boldsymbol{\theta}_n,\boldsymbol{D})p_u(\boldsymbol{U}^*|\boldsymbol{D})} = \frac{p_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{U}^*,\boldsymbol{\theta}_n)}{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{U},\boldsymbol{\theta}_n)} = \frac{\prod_{i=1}^n p_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\boldsymbol{U}^*,\boldsymbol{\theta}_n)}{\prod_{i=1}^n f_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\boldsymbol{U},\boldsymbol{\theta}_n)}$. Note that, the calculation of the acceptance function $\alpha_k(\boldsymbol{U}, \boldsymbol{U}_*)$ here involves only the specification of the conditional distribution of $\boldsymbol{y}|\boldsymbol{u}$ which can be computed in a closed form.

3.1 Choice of regularization parameters

For computational convenience, we use equally spaced knots with the number of interior knots $L_n \approx n^{1/(2r+1)}$, where r is positive integer. Following Fan and Li (2001), we set $a = 3.7$ and to select the tuning parameter λ_n use the GCV given by $\text{GCV}_{\lambda_n} = \frac{\text{RSS}(\lambda_n)/n}{(1-d(\lambda_n)/n)^2}$, where $\text{RSS}(\lambda_n) = \frac{1}{N}\sum_{k=1}^N \Big[\sum_{i=1}^n \big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_n, U_i^{(k)})\big)^\top \boldsymbol{W}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}}_n, U_i^{(k)})\big)\Big]$ is the residual sum of squares, and $d(\lambda_n) = tr\Big[\Big\{\frac{1}{N}\sum_{k=1}^N \Big[\boldsymbol{H}_n\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{U}^{(k)}\big)\Big] + n\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}_n)\Big\}^{-1} \times \Big\{\frac{1}{N}\sum_{k=1}^N \Big[\boldsymbol{H}_n\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{U}^{(k)}\big)\Big]\Big\}\Big]$ is the effective number of parameters. Then, λ_{opt} is the minimizer of the GCV_{λ_n}. Note that $\boldsymbol{W}_i$ is an $n_i \times n_i$ covariance matrix of $\boldsymbol{y}_i$.

3.2 Asymptotic properties

Assume the true value of $\boldsymbol{\beta}_0$ is partitioned $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top)^\top$ and the corresponding design matrix into $\boldsymbol{X}_i = \big(\boldsymbol{X}_{i(1)}, \boldsymbol{X}_{i(2)}\big)$. In our study, the true regression coefficients are $\boldsymbol{\theta}_{n0} = (\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top, \boldsymbol{\alpha}_0^\top)^\top$, where $\boldsymbol{\alpha}_0$ is an h_n-dimensional vector depending on f_0. For technical convenience let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^\top, \boldsymbol{\theta}_{02}^\top)^\top$ where $\boldsymbol{\theta}_{01} = (\boldsymbol{\beta}_{01}^\top, \boldsymbol{\alpha}_0^\top)^\top$ is $(s = s^* + h_n)$-dimensional vector of true values that the elements are all nonzero and $\boldsymbol{\theta}_{02} = \boldsymbol{\beta}_{02} = \boldsymbol{0}$. Here, s^* is the dimension of $\boldsymbol{\theta}_{01}$ and assume that only a small number of covariates contribute to the response i.e. $\mathcal{S} = \{1 \leqslant j \leqslant p; \beta_j \neq 0\}$ has cardinal $|\mathcal{S}| = s^* < p$. Consequently, estimated values and the design matrix is repartitioned as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{n1}^\top, \widehat{\boldsymbol{\theta}}_{n2}^\top)^\top$, and $\boldsymbol{D}_i = \big(\boldsymbol{D}_{i(1)}^\top, \boldsymbol{D}_{i(2)}^\top\big)^\top$ which $\widehat{\boldsymbol{\theta}}_{n1} = (\widehat{\boldsymbol{\beta}}_{n1}^\top, \widehat{\boldsymbol{\alpha}}_n^\top)^\top$, $\boldsymbol{D}_{i(1)} = \big(\boldsymbol{X}_{i(1)}^\top, \boldsymbol{B}(\boldsymbol{t}_i)^\top\big)^\top$, $\widehat{\boldsymbol{\theta}}_{n2} = \widehat{\boldsymbol{\beta}}_{n2}$ and $\boldsymbol{D}_{i(2)} = \boldsymbol{X}_{i(2)}$. The following regularity conditions are required.

(A.5) The eigenvalues of matrix $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top|t_{ij} = t)$ are bounded away from 0 and infinity uniformly for all $t \in [a, b]$. The common true correlation matrix $\boldsymbol{R}_0$ has eigenvalues bounded away from zero and $+\infty$; the estimated working correlation matrix $\overline{\boldsymbol{R}}$ satisfies $\|\widehat{\boldsymbol{R}}^{-1} - \overline{\boldsymbol{R}}^{-1}\| = O_p(n^{-1/2})$, where $\overline{\boldsymbol{R}}$ is a constant positive definite matrix with eigenvalues bounded away from zero and $+\infty$; we do not require $\overline{\boldsymbol{R}}$ to be the true correlation matrix $\boldsymbol{R}_0$;

(A.6) Let $\boldsymbol{\epsilon}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i) = \big(\epsilon_{i1}(\boldsymbol{\theta}_n,\boldsymbol{u}_i),\ldots,\epsilon_{in_i}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big)^\top = \boldsymbol{A}_i^{-1/2}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\big)$. There exists a finite constant $M_1 > 0$ such that $\mathbb{E}(\|\boldsymbol{\epsilon}_i(\boldsymbol{\theta}_{n0},\boldsymbol{u}_i)\|^{2+\delta}) \leqslant M_1$, for all i and some $\delta > 0$; and there exist positive constants M_2 and M_3 such that $\mathbb{E}\Big[\exp\big(M_2|\epsilon_{ij}(\boldsymbol{\theta}_{n0},\boldsymbol{u}_i)|\big)\big|\boldsymbol{X}_i\Big] \leqslant M_3$ uniformly;

(A.7) Let $B_n = \{\boldsymbol{\theta}_n : \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n0}\| \leqslant \Delta\sqrt{p_n/n}$, then $\boldsymbol{\mu}^{\cdot}(\boldsymbol{D}_{ij}^\top\boldsymbol{\theta}_n)$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, are uniformly bounded away from 0 and ∞ on B_n; $\boldsymbol{\mu}^{\cdot\cdot}(\boldsymbol{D}_{ij}^\top\boldsymbol{\theta}_n)$ and $\boldsymbol{\mu}^{(3)}(\boldsymbol{D}_{ij}^\top\boldsymbol{\theta}_n)$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, are uniformly bounded by a finite positive constant M_2 on B_n;

(A.8) Assuming $\min_{1\leqslant k\leqslant s_n}|\theta_{n0k}|/\lambda_n \to \infty$ as $n \to \infty$ and $s_n^3 n^{-1} = o(1)$, $\lambda_n \to 0$, $s_n(\log n)^2 = o(n\lambda_n^2)$, $\log p_n = o(n\lambda_n^2/\log n)$, $p_n s_n^4(\log n)^6 = o(n^2\lambda_n^2)$, and $p_n s_n^3(\log n)^8 = o(n^2\lambda_n^4)$.

Theorems 1-3 below characterize the existency, consistency and normality of the proposed penalized estimator when $p_n \to \infty$.

Theorem 1. *(Existency). Assume (A.1)–(A.8). Let $S_{nk}(\widehat{\boldsymbol{\theta}}_n)$ denotes the kth element of $\boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n)$. Then, there exists an approximate penalized GEE solution $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^\top, \widehat{\boldsymbol{\theta}}_2^\top)^\top$ satisfies the*

$$(i)\ \ \mathbb{P}_n\Big(|U_{nk}(\widehat{\boldsymbol{\theta}}_n)| = 0,\ k = 1,\ldots,s_n^*,(s_n^*+1),\ldots,(s_n = s_n^* + h_n)\Big) \to 1,$$
$$(ii)\ \ \mathbb{P}_n\Big(|U_{nk}(\widehat{\boldsymbol{\theta}}_n)| \leqslant \frac{\lambda_n}{\log n},\ k = (s_n^* + h_n + 1),\ldots,p_n\Big) \to 1,$$

where

$$U_{nk}(\widehat{\boldsymbol{\theta}}_n) = \begin{cases} S_{nk}(\widehat{\boldsymbol{\theta}}_n) - n\frac{q_{\lambda_n}(|\widehat{\beta}_{nk}|)}{\epsilon+|\widehat{\beta}_{nk}|}\widehat{\beta}_{nk} & k = 1,\ldots,s_n, \\ S_{nk}(\widehat{\boldsymbol{\theta}}_n) & k = (s_n+1),\ldots,p_n \end{cases},$$

Theorem 2. *(Consistency). Assume (A1)–(A8). Then, $\boldsymbol{U}_n(\boldsymbol{\theta}_n) = o(1)$ has a root $\widehat{\boldsymbol{\theta}}_n$ that*

$$(i)\ \ \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\| = O_p(\sqrt{p_n/n}),$$
$$(ii)\ \ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big(\widehat{f}(t_{ij}) - f_0(t_{ij})\big)^2 = O_p(n^{-2r/(2r+1)}).$$

Theorem 3. *(Oracle properties). Assume (A.1)–(A.8). Then, we have*

$$(i)\ \ \mathbb{P}_n(\widehat{\boldsymbol{\beta}}_{n2} = \boldsymbol{0}) \to 1,$$
$$(ii)\ \ \boldsymbol{\xi}_n^\top\overline{\boldsymbol{M}}_n^{*^{-1/2}}(\boldsymbol{\beta}_{n0})\overline{\boldsymbol{H}}_n^*(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n01}) \xrightarrow{\mathcal{D}} N_{p_n}(0,1), \quad \forall\boldsymbol{\xi}_n \in \mathcal{R}^{p_n}\ \&\ \|\boldsymbol{\xi}_n\| = 1.$$

$\overline{\boldsymbol{M}}_n^* = \mathbb{E}_{u|y}\Big[\sum_{i=1}^n \boldsymbol{X}_i^{*^\top}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\overline{\boldsymbol{R}}^{-1}\boldsymbol{R}_0\overline{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\boldsymbol{X}_i^*\Big]$, $\boldsymbol{X}_i^* = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}_i$, $\boldsymbol{P} = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Omega}\boldsymbol{B})^{-1}\boldsymbol{B}^\top\boldsymbol{\Omega}$, $\boldsymbol{\Omega} = diag\{\boldsymbol{\Omega}_i\}$, $\boldsymbol{\Omega}_i = \mathbb{E}_{u|y}\Big[\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\overline{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n,\boldsymbol{u}_i)\Big]$, and

$\overline{\boldsymbol{H}}_n^* = \mathbb{E}_{u|y}\Big[\sum_{i=1}^n \boldsymbol{X}_i^{*^\top} \boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n, \boldsymbol{u}_i)\overline{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}_n, \boldsymbol{u}_i)\boldsymbol{X}_i^*\Big]$.

4 Numerical Studies

This section, first conduct simulation study to illustrate the consistency and the sure screening property of the proposed penalized procedure(P-GSMM) empirically, and compare its finite sample performance with some other different model settings, including the unpenalized one (GSMM), and the penalized generalized linear model (P-GLMM). We further apply our proposed method for analyzing two real data sets.

4.1 Simulation

We generated 100 data sets following $y_{ij}|b_i \sim \text{Pois}(\mu_{ij})$, with $\eta_{ij} = \log(\mu_{ij}) = \sum_{k=1}^p x_{ij}^{(k)}\beta_k + \sin(2\pi t_{ij}) + b_i$, where $i = 1, \dots, n$, and $j = 1, \dots, n_i$, $\boldsymbol{\beta} = (-1, -1, 2, 0, \dots, 0)$, $x_{ij}^{(k)} \sim U(-1, 1)$, $t_{ij} \sim U(0, 1)$, and $b_i \sim N(0, 0.25)$. The predictor dimension p_n is diverging but the dimension of the true model is fixed to be 3. Regarding the choice of the dimensionality of the parametric component, p_n, authors recommended many suggestions as a sensible choice. For example $p_n = [\frac{n}{2}]$, $p_n = [4.5n^{1/4}]$, and $p_n = [\frac{n}{b\log(n)}]$, where $b > 1$ and $[a]$ stands for the largest integer no larger than a. These only discuss the situation $p \to \infty$ as $n \to \infty$ with $p_n < n$. For case $p_n >> n$, we can mention to $\log(p_n) = o_p(n^b)$, where $0 < b < 1$. Ofcourse challenges arise when p is much larger than n, choosing a larger value of p_n increases the probability that variable selection methods will include all of the correct variables, but including more inactive variables will tend to have a slight detrimental effect on the performance of the final variable selection and parameter estimation method. We have found that this latter effect is most noticeable in models where the response provides less information. We therefore used the pairs of (n, p_n) as $(50, 11), (100, 14), (150, 16)$ and $(30, 100), (100, 500), (200, 2000)$ respectively for cases $p_n < n$ and $p_n >> n$. For evaluating estimation accuracy, we report the empirical mean square error (MSE), defined as $\sum_{k=1}^{100} \|\widehat{\boldsymbol{\beta}}_n^k - \boldsymbol{\beta}_{n0}\|/100$ where $\widehat{\boldsymbol{\beta}}_n^k$ is the estimator of $\boldsymbol{\beta}_{n0}$ obtained using the kth generated data set. The performance in variable selection is gauged by 'C' the mean over all 100 simulations of zero coefficients which are correctly estimated by zero, 'I' the mean over all 100 simulations of nonzero coefficients which are incorrectly estimated by zero, 'Under-fit' the proportion of excluding any true nonzero coefficients,'Correct-fit' the proportion of selecting the exact subset model, and 'Over-fit' the proportion of including all three important variables plus some noise variables. Table 1 summarize the results of the P-GSMM, GSMM, and the P-GLMM for the different values of (n, p_n). In terms of estimation accuracy the P-GSMM performs closely to the P-GLMM, whereas our proposed approach gives the smallest MSE, and consistently outperforms its P-GLMM counterpart. In terms of model selection we observe that the GSMM generally does not lead to a sparse model. Furthermore, the P-GSMM and the P-GLMM successfully selects all covariates with nonzero coefficients (i.e., I rates are zero), but it is obvious that the proposed approach has slightly stronger sparsity (i.e., a fairly higher number of Cs). For P-GSMM, The probability of identifying the exact underlying model is about 80% and this rate grows by increasing the sample size, confirming the good asymptotic properties of the penalized estimators. The results are the same in both cases of $p_n < n$ and $p_n >> n$, but when $p_n >> n$ zero coefficients tends to increasingly included in the model.

Table 1: Simulation results

method	case $p_n < n$						case $p_n >> n$					
	$(n, p) = (50, 11)$						$(n, p) = (30, 100)$					
	MSE	C(8)	I(0)	Under-fit	Correct-fit	Over-fit	MSE	C(97)	I(0)	Under-fit	Correct-fit	Over-fit
GPLMM	0.116	0.09	0.00	0.00	0.00	1.00	68.028	0.074	0.00	0.00	0.00	1.00
P-GLMM	0.060	6.54	0.00	0.00	0.13	0.87	0.435	96.48	0.00	0.00	0.55	0.45
P-GPLMM	0.052	7.59	0.00	0.00	0.64	0.36	0.391	96.41	0.00	0.00	0.60	0.40
	$(n, p) = (100, 14)$						$(n, p) = (100, 500)$					
	MSE	C(11)	I(0)	Under-fit	Correct-fit	Over-fit	MSE	C(497)	I(0)	Under-fit	Correct-fit	Over-fit
GPLMM	0.072	0.16	0.00	0.00	0.00	1.00	1499.136	47.02	0.03	0.03	0.00	1.00
P-GLMM	0.041	10.52	0.00	0.00	0.77	0.23	0.062	495.720	0.00	0.00	0.89	0.11
P-GPLMM	0.036	10.70	0.00	0.00	0.93	0.07	0.038	496.250	0.00	0.00	0.92	0.08
	$(n, p) = (150, 15)$						$(n, p) = (200, 2000)$					
	MSE	C(12)	I(0)	Under-fit	Correct-fit	Over-fit	MSE	C(1997)	I(0)	Under-fit	Correct-fit	Over-fit
GPLMM	0.060	0.26	0.00	0.00	0.00	1.00	125.406	1137.62	0.00	0.00	0.00	1.00
P-GLMM	0.044	11.25	0.00	0.00	0.92	0.08	0.018	1996.89	0.00	0.00	0.30	0.700
P-GPLMM	0.045	11.87	0.00	0.00	0.96	0.04	0.018	1996.93	0.00	0.00	0.54	0.46

4.2 AIDS data analysis

This dataset contains 2376 observations of CD4 cell counts on 369 men infected with the HIV virus. The first objective of this analysis is to characterize the population average time course of CD4 decay while accounting for the following additional predictor variables including AGE, SMOKE (smoking status measured by packs of cigarettes), DRUG (yes, 1; no, 0), SEXP (number of sex partners), DEPRESSION as measured by the CESD scale (larger values indicate increased depressive symptoms) and YEAR (the effect of time since seroconversion). This data analysed by many authors such as Wang et al. (2005), Huang et al. (2007) and **?**. Their analysis was conducted on square root transformed CD4 numbers whose distribution is more nearly Gaussian. In our analysis, we fit the data using an GSMM, without transforming the CD4 by adopting the Poisson regression. To take advantage of flexibility of partially linear models, we let 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. We further applied the proposed approach to select significant variables. To compare the performance of proposed P-GSMM method with GSMM and P-GLMM, we use the standard errors (SE). To best identify a model supported by the data, we adopt the Akaike information criterion (AIC) and the Bayesian information

Table 2: AIDS data results

	GSMM	P-GLMM	P-GSMM
Variabeles	$\widehat{\beta}$(SE)	$\widehat{\beta}$(SE)	$\widehat{\beta}$(SE)
AGE	0.073 (0.039)	-0.092 (0.051)	0 (0)
$SMOKE$	0.188 (0.179)	0.888 (0.192)	0.079 (0.045)
$DRUG$	0.130 (0.143)	6.068(0.125)	0.142 (0.074)
$SEXP$	-0.049 (0.031)	0.672 (0.030)	0.017 (0.012)
$CESD$	-0.001 (0.011)	0 (0)	0 (0)
$AGE * SMOKE$	0.002 (0.014)	0.014 (0.004)	0 (0)
$AGE * DRUG$	-0.034 (0.024)	0.032 (0.035)	0 (0)
$AGE * SEXP$	-0.009 (0.003)	0 (0)	0 (0)
$AGE * CESD$	0.001 (0.002)	0 (0)	0 (0)
$SMOKE * DRUG$	0.009 (0.054)	-0.584 (0.150)	-0.014 (0.038)
$SMOKE * SEXP$	-0.010 (0.012)	-0.034 (0.010)	0 (0)
$SMOKE * CESD$	-0.006 (0.009)	0 (0)	0 (0)
$DRUG * SEXP$	-0.025 (0.019)	-0.598 (0.041)	-0.022 (0.012)
$DRUG * CESD$	0.006 (0.006)	0 (0)	0 (0)
$SEXP * CESD$	0.001 (0.003)	0 (0)	0 (0)
$\ell_{\max}$	8463007	7529158	8624429
AIC	-16925983	-15058286	-17248827
BIC	-16925924	-15058228	-17248769

criterion (BIC). Table 2 presents the summary of the fitting results including the values of standard errors, together with $\ell_{\max}$, AIC, and BIC under the three models. Judging from Table 2, the P-GSMM tends to exhibit slightly SE compared to GSMM and P-GLMM, nevertheless this difference is not more dramatic. Meanwhile, the values of AIC, BIC of proposal are smaller than the two others, revealing that the P-GSMM can provide better fitting performance. Under P-GSMM, SMOKE, DRUGS, SEXP, $SOMKE * DRUG$ and $DRUG * SEXP$ are identifies as significant covariates. One notes some slight selection difference when P-GLMM is used, which suggests that $AGE * SMOKE$, $AGE * DRUG$, and $SMOKE * SEXP$ may also be significant. We also find some significant interactions among some covariates which may be ignored by Wang et al. (2005) and Huang et al. (2007).

5 Conclusions

We developed a general methodology for simultaneously selecting variables and estimating the unknown components in the semiparametric mixed-effects model for non-Gaussian longitudinal data when the number of parameters diverges with the sample size. Penalized estimating equation technique involves the specification of the posterior distribution of the random effects, which cannot be evaluated in a closed form, and we used a Metropolis algorithm, which does not require this specification. We further investigated some asymptotic properties of the estimates. To investigate the performance of our approach, it compared with the unpenalized generalized semiparametric mixed-effects model and penalized generalized linear mixed-effects model throw a simulation study and the analysis of two data sets. Results showed that the proposed model outperforms the linear counterparts on the provision of model selection and estimation. In addition, we found the estimation is more efficient when the partially part is taken into consideration. The results are consistent in both cases of $p_n < n$ and $p_n >> n$.

Acknowledgment

I would like to thank my supervisor Dr. Mohammad Arashi from Ferdowsi University of Mashhad for his constructive suggestions that largely improve the presentation of the paper.

References

Dziak, J. J. (2006). Penalized quadratic inference functions for variable selection in longitudinal research. Ph.D Thesis, the Pennsylvania State University.

Fan, J. Q., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59** 126–132.

Huang, J. Z., Zhang, L., and Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Stat.* **34** 451–477

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68** 353–360.

Wang, N., Carroll, R.J., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157

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.

Xu, P. R., Fu, W., and Zhu, L. X. (2012). Shrinkage estimation analysis of correlated binary data with a diverging number of parameters, To appear. *Science in China Series A: Mathematics.*

Xu, P. R., and Zhu, L. X. (2010). Sure independence screening for marginal longitudinal generalized linear models. *Unpublished manuscript.*