Regularization in Generalized Semiparametric Mixed Effect Model for Longitudinal Data

M. Taavoni* and M. Arashi†

September 19, 2019

Abstract

This paper considers the problem of simultaneous variable selection and estimation in the generalized semiparametric mixed effects 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 using regression spline to approximate the nonparametric component. Our approach applies SCAD to the estimating equation objective function in order to simultaneously estimate parameters and select the important variables. The proposed procedure 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 the distribution using a Metropolis algorithm, which does not require the specification of the posterior distribution. For practical implementation, we develop an appropriate iterative algorithm to select the significant variables and estimate the nonzero coefficient functions. Under some regularity conditions, the resulting estimators enjoy the oracle properties, under a high-dimensional regime. Simulation studies are carried out to assess the performance of our proposed method, and a real data set is analyzed to illustrate the procedure.

Keywords: Generalized estimating equations; High dimension; Longitudinal data; Smoothing spline; Variable selection.

*Department of Statistic, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran (e-mail: taavonimozhgan@yahoo.com)
†Corresponding author
‡Department of Statistic, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran (e-mail: m_arashi_stat@yahoo.com)
1 Introduction

Longitudinal studies are often conducted in epidemiology, social science and other biomedical research areas. A challenge in the analysis of longitudinal data is that the repeated measurements from the same subjects are correlated over time. A popular way for incorporating this correlation within the likelihood framework is to use the linear mixed effects model (LMM; [Laird and Ware (1982)]) to analyze continuous longitudinal data and the generalized linear mixed effects model (GLMM; [Zeger and Karim (1991)]) to analyse discrete longitudinal data, where the random component takes care the correlation among observations from the same subjects. However, the traditional GLMM assumes parametric fixed effects that may be too restrictive to account complex covariate effects, especially when the variety of response over time is in a complicated manner. To eliminate the limitation of the GLMMs for modeling non linear time trend, a generalized semiparametric mixed effects model (GSMM), a natural extension of the GLMMs and semiparametric mixed models (SMM; [Zeger and Diggle (1994)]), is widely used to analyze longitudinal data by incorporating the within subject correlation using random effects and an arbitrary smooth function to model the time effect. Further developments along this line in the framework of GSMM can be found in [Fan et al. (2007)], [Qin and Zhu (2007)], [Qin and Zhu (2009)], [Liang (2009)], and [Kurum et al. (2016)] to mention afew.

High dimensional longitudinal data, which consist of repeated measurements on a large number of covariates, have become increasingly common. 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. There is a large body of variable selection methods for cross-sectional data. Among all, we refer to bridge regression ([Frank and Friedman (1993)], Lasso ([Tibshirani (1996)], adaptive
Lasso (Zou, 2006), Elastic-net (Zou and Hastie, 2005), and SCAD (Fan and Li, 2001). The literature on variable selection for longitudinal data is rather limited due to the challenges imposed by incorporating the intracluster correlation. Fan and Li (2004) extended the SCAD procedure to the semiparametric model for longitudinal data. Bondell et al. (2010) proposed simultaneous selection of the fixed and random factors using a penalized joint log likelihood for the LMM. Ni et al. (2010) proposed a double-penalized likelihood approach for simultaneous model selection and estimation for the SMM. Ma et al. (2013) applied proper penalty functions in the additive semiparametric model. Chu et al. (2016) developed a screening procedure for ultrahigh dimensional longitudinal data. In contrast to extensive attention on model selection for Gaussian longitudinal data, research on model selection for non-Gaussian longitudinal data in the framework of the GLM remains largely unexplored.

To do variable selection, Pan (2001) developed a quasi-likelihood information criterion (QIC) which is analogous to AIC; Cantoni et al. (2005) generalized Mallow’s $C_p$ criterion, and Wang and Qu (2009) proposed a BIC criterion based on the quadratic inference function. These are best subset type model selection procedures which become computationally intensive when the number of parameters is moderately large.

Regarding regularization methods for longitudinal data, 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) for the generalized additive model with correlated data. In all aforementioned studies on the penalized GLM for longitudinal data, the dimension of predictors is fixed. Xu et al. (2012) proposed a weighted least squares type function to study the longitudinal GLMs with a diverg-
ing number of parameters. For correlated discrete outcome data, the joint likelihood function does not have a closed form if the correlation information is taken into account. When the dimension of parameters diverges, numerical approximation to the joint likelihood function tends to be computationally infeasible as it often involves high-dimensional integration. This motivated [Liang and Zeger (1986)] to develop an approach of the GEE which is a multivariate analogue of the quasi-likelihood. [Johnson et al. (2008)] recently derived the asymptotic theory for the penalized estimating equations for independent data. [Wang et al. (2012)] employed rather different techniques than those in [Johnson et al. (2008)] and proposed the SCAD-penalized GEE for analyzing longitudinal data with high dimensional covariates. To the best of our knowledge, regularization in the GSMM is neglected.

In this paper, we focus on the GSMM with longitudinal data by allowing for non-Gaussian data and nonlinear link functions. We consider the case where the number of variables $p$ is allowed to increase with the number of sample size $n$ under a high dimensional regime. Similar to the work of [Wang et al. (2012)], we apply the penalty functions to the estimating equation objective function such that the proposed procedure can simultaneously estimate parameters and select the important variables. 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 [Tanner (1993)], 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. To estimate the parameters, a computationally flexible iterative algorithm is developed. Furthermore, we propose a sandwich formula to estimate the asymptotic covariance matrix.

The rest of this paper is organized as follows. Section 2 formulates the model, discusses the approximation of the nonparametric function using splines and considers the estimation under the GEE framework. Section 3 include selection of the regularization parameters and the model selection procedure for assessing the fits of candidate models is also addressed. Moreover, asymptotic properties of the estimators and a Monte Carlo Newton-Raphson algorithm to implement the procedures is given. In Section 4 we apply a number of simulations to assess the finite sample performance of the proposed estimation method in the GSMM. A real data analysis is also presented in this section to augment the theoretical results. Some concluding remarks are given in Section 5.

Further, the proofs of the main results as well as some instrumental lemmas are provided in a separate supplementary file.

2 Generalized Semiparametric Mixed Effects Model

2.1 Model specification

Consider a longitudinal study with \( n \) subjects and \( n_i \) observations over time for the \( i \)th subject \((i = 1, \ldots, n)\). Let \( \mathbf{u}_i \) be a \( q \times 1 \) vector of random effects corresponding to the \( i \)th subject, and \( y_{ij} \) be an observation of the \( i \)th subject measured at time \( t_{ij} \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, n_i \). Suppose that \( y_{i1}, \ldots, y_{in_i} \) given \( \mathbf{u}_i \) are conditionally independent and each \( y_{ij} | \mathbf{u}_i \) is distributed as an exponential family distribution whose
probability density function is given by

\[ p(y_{ij}|\mathbf{u}_i, \beta_n, \phi) = \exp \left[ \phi^{-1}\{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, \phi) \right], \tag{1} \]

where \( \phi \) is a scale parameter, \( c(., .) \) is a function only depending on \( y_{ij} \) and \( \phi \), and \( \theta_{ij} \) is the (scalar) canonical parameter. The conditional expectations and variances of \( y_{ij} \) given \( \mathbf{u}_i \) are given by \( \mu_{ij} = E(y_{ij}|\mathbf{u}_i) = b'(\theta_{ij}) \) and \( \nu_{ij} = \text{var}(y_{ij}|\mathbf{u}_i) = \phi b''(\theta_{ij}) \), respectively, where \( b'(\theta) = \frac{\partial b(\theta)}{\partial \theta} \) and \( b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2} \). In this paper, we assume that the conditional mean \( \mu_{ij} \) satisfies

\[ g(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}^\top \beta_n + \mathbf{Z}_{ij}^\top \mathbf{u}_i + f(t_{ij}), \quad i = 1, \ldots, n; \quad j = 1, \ldots, n_i, \tag{2} \]

where \( g(.) \) is a known monotonic link function, \( \mathbf{X}_{ij}^\top \) is a \( p_n \times 1 \) vector of explanatory variables, \( \beta_n \) is a \( p_n \times 1 \) vector of unknown parameters of the fixed effects, \( \mathbf{Z}_{ij}^\top \) is a \( q \times 1 \) vector of explanatory variables 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 \( \mathbf{u} = \{\mathbf{u}_1, \ldots, \mathbf{u}_q\} \) independently follow the same distribution, depending on parameters \( \Sigma \) as

\[ \mathbf{u}_i \sim f_u(\mathbf{u}_i|\Sigma). \tag{3} \]

The model defined in Eqs. 1–3 is referred to as generalized semiparametric mixed model (GSMM). Specific assumptions will be considered for the number of variables \( p_n \) in section 3.3.
2.2 Smoothing spline approximation

Following the most literature such as He et al. (2005) and Qin and Zhu (2009), the unspecified smooth function can be approximated sufficiently well by the following polynomial spline

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

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 3.3), $t_{ij}^{(l)}$ is referred as knots of the $i$th subject, $B(t_{ij}) = (1, t_{ij}, \ldots, t_{ij}^d, (t_{ij} - t_{ij}^{(1)})_+^d, \ldots, (t_{ij} - t_{ij}^{(L_n)})_+^d)$ 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 $\alpha_n = (\alpha_0, \ldots, \alpha_d, \alpha_{d+1}, \ldots, \alpha_{d+1+L_n})^\top$ is the spline coefficients vector of dimension $h$. Regression splines have some desirable properties in approximating a smooth function. It often provides good approximations with a small number of knots. The spline approach also treats a nonparametric function as a linear function with the basis functions as pseudo design variables, and thus this linearizes our regression model (2) so that our regression problem becomes

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^\top\beta_n + Z_{ij}^\top u_i + B(t_{ij})\alpha_n, \quad i = 1, \ldots, n; j = 1, \ldots, n_i. \quad (4)$$

For convenience, model (4) can take the form

$$g(\mu_{ij}) = \eta_{ij} = D_{ij}^\top\theta_n + Z_{ij}^\top u_i, \quad i = 1, \ldots, n; j = 1, \ldots, n_i,$$

where $D_{ij} = (X_{ij}^\top, B_j(t_i)^\top)^\top$ being a $(p_n + h_n) \times 1$ design matrix combining the fixed-effects and spline-effects design matrices for the $j$th outcome of the $i$th subject, and $\theta_n = (\beta_n^\top, \alpha_n^\top)^\top$ is a $(p_n + h_n) \times 1$ combined regression parameters vector that
must be estimated.

According to the linearization of the GSMM [2] using spline approach, any computational algorithm developed for the GMM can be used for the GSMM. We now formulate the linearization of the GSMM in the seamless form

\[
p(y_{ij}|u_i, \theta_n, \phi) = \exp\left[\phi^{-1}\{y_{ij}\theta - b(\theta_{ij})\} + c(y_{ij}, \phi)\right],
\]

\[
u_i \sim f_u(u_i|\Sigma),
\]

\[
\mu_{ij} = \mathbb{E}(y_{ij}|u_i),
\]

\[
g(\mu_{ij}) = \eta_{ij} = D_{ij}^\top\theta_n + Z_{ij}^\top u_i, \quad i = 1, \ldots, n; j = 1, \ldots, n_i.
\]

\[
2.3 \text{ Estimation procedure}
\]

For linearization of the GSMM defined in (5), the classical likelihood function can be defined as

\[
L(\theta_n, \Sigma, \phi) = \prod_{i=1}^{n} \int p_{y_i|u_i}(y_i|u_i, \theta_n, \phi) p_u(u_i|\Sigma) du_i
\]

where \(y_i = (y_{i1}, \ldots, y_{in_i})^\top\), \(u = (u_1, \ldots, u_n)\), and

\[
p_{y_i|u_i}(y_i|u_i, \theta_n, \phi) = \prod_{j=1}^{n_i} p(y_{ij}|u_i, \theta_n, \phi).
\]

For the maximum likelihood (ML) estimates of the parameters \(\theta_n, \phi\) and \(\Sigma\), one can maximize this likelihood function by using suitable numerical techniques. The EM algorithm is an attractive method to obtain the ML estimates, in presence of incomplete data, which avoids explicit calculation of the observed data log-likelihood.

To set up the EM algorithm we consider the random effects, \(u_i\), to be the missing data. The complete data, is then \((y_i, u_i)\) and the complete data log-likelihood is
given by
\[ \ell(\ell_n, \Sigma, \phi) = \sum_{i=1}^{n} \ln p_{y|u_i}(y_i|u_i, \ell_n, \phi) + \sum_{i=1}^{n} \ln p_{u_i}(u_i|\Sigma). \] (7)

Considering \( u_i \) to be the missing has an advantage that in the M-step, maximization can be accomplished with respect to the parameters \( \ell_n \) and \( \phi \) only in the first term of (7). Thus, the M-step with respect to \( \ell_n \) and \( \phi \) uses only the GLM part of the likelihood function. Hence, the procedure is similar to a standard GLM computation assuming \( u_i \) is known. Therefore, maximizing with respect to \( \Sigma \), in the second term, can be handled by the maximum likelihood using the distribution of \( p_{u_i}(u_i|\Sigma) \) after replacing sufficient statistics with the conditional expected values. It thus makes sense to develop a simulation analogous to the Newton-Raphson approach for fitting the GPLMM. Using this separation as in (7), the ML equations for \( \ell_n \) and \( \Sigma \) take the following forms
\[
\mathbb{E}\left[ \frac{\partial \ln p_{y_{ij}|u_i}(y_{ij}|u_i, \ell_n, \phi)}{\partial \ell_n} | y_{ij} \right] = 0,
\]
\[
\mathbb{E}\left[ \frac{\partial \ln p_{u_i}(u_i|\Sigma)}{\partial \Sigma} | y_{ij} \right] = 0.
\]

The ML estimates of \( \ell_n \) and \( \Sigma \) can be obtained by solving the preceding equations numerically. McCulloch (1997) developed a Monte Carlo Newton-Raphson (MCNR) algorithm for solving these estimating equations, and obtained approximate ML estimates of the parameters for the GMM. Motivated by his work and combining it with the GEE of Liang and Zeger (1986), the optimal estimating equation for \( \ell_n \) is given by
\[
\mathbb{E}_{u|y}\left[ n^{-1} \sum_{i=1}^{n} \frac{\partial \mu_i(\ell_n, u_i)}{\partial \ell_n} V_{i^{-1}}(\ell_n, u_i)(y_i - \mu_i(\ell_n, u_i)) \right] = 0, \quad (8)
\]
where $\mu_i(\theta_n, u_i) = (\mu_{i1}, \ldots, \mu_{in_i})^\top$ and $V_i(\theta_n, u_i)$ is the covariance matrix of $y_i | u_i$. 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 $R(\rho) : V_i(\theta_n, u_i) = A_i^\frac{1}{2}(\theta_n, u_i)R(\rho)A_i^\frac{1}{2}(\theta_n, u_i)$, where $\rho$ is a finite dimensional parameter and $A_i(\theta_n, u_i) = \text{diag}(\nu_{i1}, \ldots, \nu_{in_i})$. Some commonly used working correlation structures include independence, autocorrelation (AR)-1, equally correlated (also called compound symmetry), or unstructured correlation, among others. For a given working correlation structure, $\rho$ can be estimated using the residual-based method of moments. With the estimated working correlation matrix $\hat{R} \equiv R(\hat{\rho})$, the estimating equations in (8) reduces to

$$E_{u | y} \left[ n^{-1} \sum_{i=1}^{n} D_i^\top A_i^\frac{1}{2}(\theta_n, u_i)\hat{R}^{-1} A_i^\frac{1}{2}(\theta_n, u_i) (y_i - \mu_i(\theta_n, u_i)) \right] = 0, \quad (9)$$

where $D_i = (D_{i1}^\top, \ldots, D_{in_i}^\top)^\top$. We formally define the estimator as the solution $\hat{\theta}_n$ of the above estimating equations. For ease of exposition, we assume $\phi = 1$ and $n_i = m < \infty$ in the rest of the article. Extension of the methodology to the cases of unequal $n_i$ is straightforward. We vary the dimension of $A_i$ and replace $\hat{R}$ by $\hat{R}_i$, which is the $n_i \times n_i$ matrix using the specified working correlation structure and the corresponding initial parameter $\rho$ estimator.

### 3 Regularization in the GPLMM

In order to select important covariates and estimate them simultaneously, the log likelihood (8) is expanded to include the penalty term $\sum_{k=1}^{p_{\phi}} \lambda_k |\beta_{nk}|$ which yields
the following penalized log likelihood

\[ \ell^p(\beta_n, \alpha_n, D, \phi) = \sum_{i=1}^{n} \ln p_{y_i|u_i}(y_i|u_i, \theta_n) + \sum_{i=1}^{n} p_{u_i}(u_i|\Sigma) - n \sum_{k=1}^{p_n} p_{\lambda_n}(|\beta_{nk}|), \]

where \( p_{\lambda_n}(|\beta_{nk}|) \) is any penalty function and \( \lambda_n \) is a tuning parameter that determines the amount of shrinkage. Since the coefficients \( \theta_n \) depends to the first and third terms of (10), we propose the penalized estimating equation

\[ U_n(\theta_n) = S_n(\theta_n) - q_{\lambda_n}(|\beta_n|) \text{sign}(\beta_n), \]

where

\[ S_n(\theta_n) = \mathbb{E}_{u_i|y_i} \left[ \sum_{i=1}^{n} D_i^\top A_i^\frac{1}{2} (\theta_n, u_i) \hat{R}^{-1} A_i^{-\frac{1}{2}} (\theta_n, u_i) (y_i - \mu_i(\theta_n, u_i)) \right], \]

with \( q_{\lambda_n}(|\beta_n|) = (q_{\lambda_n}(|\beta_{n1}|), \ldots, q_{\lambda_n}(|\beta_{np_n}|))^\top \) is a \( p_n \times 1 \) vector of penalty functions, \( \text{sign}(\beta_n) = (\text{sign}(\beta_{n1}), \ldots, \text{sign}(\beta_{np_n}))^\top \) with \( \text{sign}(a) = I(a > 0) - I(a < 0) \) and \( q_{\lambda_n}(|\beta_{nk}|) = p_{\lambda_n}(|\beta_{nk}|). \)

Note that we assume the semiparametric part contains significant contribution in the model and the proposed penalized estimating equation has been defined to shrink small components of the coefficient \( \beta_n \) to zero not \( \alpha_n \). Thus, the method performing variable selection for fixed effects, produces estimators of the nonzero components and the nonparametric component.

Among all penalty functions, the smoothing clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) can be used to retain the good features of both subset selection and ridge regression, for producing sparse solutions, and to ensure continuity of the selected models. Therefore, we will use the SCAD penalty.
in our simulation and application studies. The SCAD penalty function is defined by

\[ q_{\lambda_n}(|\beta_n|) = p'_{\lambda_n}(|\beta_n|) = \lambda_n \left\{ I(|\beta_n| \leq \lambda_n) + \frac{(a\lambda_n - |\beta_n|) + I(|\beta_n| > \lambda_n)}{(a-1)\lambda_n} \right\}; \quad a > 2, \]

where the notation \((.)_+\) stands for the positive part of \((.)\).

Our proposed estimator for \(\theta_n\) is the solution of \(U_n(\theta_n) = 0\). Because \(U_n(\theta_n)\) has discontinuous points, an exact solution to \(U_n(\theta_n) = 0\) may not exist. We formally define the estimator \(\hat{\theta}_n\) to be an approximate solution, i.e., \(U_n(\hat{\theta}_n) = o(a_n)\) for a sequence \(a_n \to 0\). Alternatively, since the penalty function is singular at the origin, it is challenging to obtain the estimator of \(\theta_n\) by solving \(U_n(\theta_n) = 0\). Following Fan and Li (2001) we locally approximate the penalty function by a quadratic function. In the neighborhoods of the true parameter values \(\beta_{n0k}, |\beta_{n0k}| > 0\), the derivative of the penalty function is well approximated by

\[ q_{\lambda_n}(|\beta_{nk}|) \text{sign}(\beta_{nk}) \approx \frac{q_{\lambda_n}(|\beta_{n0k}|)}{|\beta_{n0k}|} \beta_{nk}. \]

With the local quadratic approximation, we apply the Newton-Raphson method to solve \(U_n(\hat{\theta}_n) = o(a_n)\), and get the following updating formula

\[ \hat{\theta}_n^{(m+1)} = \hat{\theta}_n^{(m)} + \left\{ H_n(\hat{\theta}_n^{(m)}) + nE_n(\hat{\theta}_n^{(m)}) \right\}^{-1} \times \left\{ S_n(\hat{\theta}_n^{(m)}) + nE_n(\hat{\theta}_n^{(m)})\hat{\theta}_n^{(m)} \right\}, \quad (11) \]

where

\[ H_n(\hat{\theta}_n^{(m)}) = \mathbb{E}_{u|y} \left[ \sum_{i=1}^{n} D_i^\top A_\beta^{(1/2)}(\theta, u_i) \hat{R}^{-1} A_\beta^{(1/2)}(\theta, u_i) D_i \right], \]

\[ E_n(\hat{\theta}_n^{(m)}) = \text{diag} \left\{ \frac{q_{\lambda_n}(|\beta_{n1}|)}{\epsilon + |\beta_{n1}|}, \ldots, \frac{q_{\lambda_n}(|\beta_{np1}|)}{\epsilon + |\beta_{np1}|}, 0_{hn} \right\}, \]
for a small numbers e.g. $\epsilon = 10^{-6}$. Here, and $0_{hn}$ denotes a zero vector of dimension $hn$.

Note that, in general, the expectations in (11) cannot be computed in a closed form as the conditional distribution of $u_i|y_i$ involves the marginal distribution of $y_i$, which is not easy to be computed explicitly. Similar to McCulloch (1997), here, we use an alternative method that produces random observations from the conditional distribution of $u_i|y_i$ by using a Metropolis algorithm where the specification of the density of $y_i$ is not required.

In the Metropolis algorithm, $p_u$ is chosen as the candidate distribution from which potential new draws are made; then we specify the acceptance function that provides the probability of accepting the new value (as opposed to retaining the previous value). In the forthcoming section we outline the computational procedure used for sample generation.

### 3.1 MCNR algorithm

Let $U$ denote the previous draw from the conditional distribution of $U|y$, and generate a new value $u^*_k$ for the $j$th component of $U^* = (u_1, \ldots, u_{k-1}, u^*_k, u_{k+1}, \ldots, u_{nq})$ by using the candidate distribution $p_u$, accept $U^*$ as the new value with probability

$$\alpha_k(U, U^*) = \min\left\{1, \frac{p_{u|y}(U^*|y, \theta_n, D)p_u(U|D)}{p_{u|y}(U|y, \theta_n, D)p_u(U^*|D)}\right\}.$$  \hspace{1cm} (12)

otherwise, reject it and retain the previous value $U$. The second term in brace in (11) can be simplified to

$$\frac{p_{u|y}(U^*|y, \theta_n, D)p_u(U|D)}{p_{u|y}(U|y, \theta_n, D)p_u(U^*|D)} = \frac{p_{y|u}(y|U^*, \theta_n)}{f_{y|u}(y|U, \theta_n)} = \frac{\prod_{i=1}^n p_{y_i|u}(y_i|U^*, \theta_n)}{\prod_{i=1}^n f_{y_i|u}(y_i|U, \theta_n)}.$$
Note that, the calculation of the acceptance function \( \alpha_k(U, U_\ast) \) here involves only the specification of the conditional distribution of \( y|u \) which can be computed in a closed form.

Algorithm 1 describes the Metropolis step into the Newton-Raphson iterative equation (11) for the Monte Carlo estimates of expected values.

**Algorithm 1** Monte Carlo Newton-Raphson (MCNR) algorithm

**step 1.** Set \( m_k = 0 \). Choose initial values \( \theta_n^0 \) and \( \Sigma^0 \).

**step 2.** Generate \( N \) observations \( U^{(1)}, \ldots, U^{(N)} \) from the distribution \( p_{u|y}(u|\theta_n^{(m_k)}, \Sigma^{(m_k)}) \) using the Metropolis algorithm. Use these observations to find the Monte Carlo estimates of the expectations. Specially,

a) Compute \( \theta_n^{(m_k + 1)} \) from the expression

\[
\theta_n^{(m_k + 1)} = \theta_n^{(m_k)} + \left\{ \frac{1}{N} \sum_{k=1}^{N} \left[ H_n(\hat{\theta}_n^{(m_k)}, U^{(k)}) + nE_n(\hat{\beta}_n^{(m_k)}) \right] - \frac{1}{N} \sum_{k=1}^{N} \left[ S_n(\hat{\theta}_n^{(m_k)}, U^{(k)}) - nE_n(\hat{\beta}_n^{(m_k)}) \hat{\beta}_n^{(m_k)} \right] \right\}^{-1}
\]

where

\[
H_n(\hat{\theta}_n^{(m_k)}, U^{(k)}) = \sum_{i=1}^{n} D_i^T A_i^{\frac{1}{2}}(\theta_n^{(m_k)}, U_i^{(k)}) \hat{R}^{-1} A_i^{\frac{1}{2}}(\theta_n^{(m_k)}, U_i^{(k)}) D_i,
\]

\[
S_n(\hat{\theta}_n^{(m_k)}, U^{(k)}) = \sum_{i=1}^{n} D_i^T A_i^{\frac{1}{2}}(\theta_n^{(m_k)}, U_i^{(k)}) \hat{R}^{-1} A_i^{\frac{1}{2}}(\theta_n^{(m_k)}, U_i^{(k)}) (y_i - \mu_i(\theta_n^{(m_k)}, U_i^{(k)})).
\]

b) Compute \( \Sigma^{(m_k + 1)} \) by maximizing

\[
\frac{1}{N} \sum_{k=1}^{N} \ln f_u(U^{(k)}|\Sigma).
\]

c) Set \( m_k = m_k + 1 \).

**step 3.** Go to step 2 until convergence is achieved. Choose \( \theta_n^{(m_k + 1)} \) and \( \Sigma^{(m_k + 1)} \) to be the MCNR estimates of \( \theta_n \) and \( \Sigma \).
3.2 Choice of regularization parameters

To implement the proposed method, several parameters need to be chosen appropriately. One needs to choose the knot sequence in the polynomial spline approximation, \( \lambda_n \), and \( a \) in the SCAD penalty function. It is important that the number of distinct knots \( h \), must increase with the sample size \( n \). On the other hand, too many knots would increase the variance of our estimators. Therefore, the number of knots must be properly chosen to balance the bias and variance. For computational convenience, we use equally spaced knots with the number of interior knots \( L_n \approx n^{1/(2p+1)} \). A similar strategy for knot selection can also be found in He et al. (2002); Qin and Zhu (2007); Sinha and Sattar (2015). To reduce the computational burden, we follow Fan and Li (2001) and set \( a = 3.7 \). Finally we need to choose \( \lambda_n \).

Due to the lack of joint likelihood in the generalized model, to select the tuning parameter \( \lambda_n \) we use generalized cross validation (GCV) suggested by Fan and Li (2001) as defined by

\[
\text{GCV}_{\lambda_n} = \frac{\text{RSS}(\lambda_n)/n}{(1-d(\lambda_n)/n)^2},
\]

where

\[
\text{RSS}(\lambda_n) = \frac{1}{N} \sum_{k=1}^{N} \left[ \sum_{i=1}^{n} \left( y_i - \mu_i(\hat{\theta}_n, U_i^{(k)}) \right)^T W^{-1}_i \left( y_i - \mu_i(\hat{\theta}_n, U_i^{(k)}) \right) \right]
\]

is the residual sum of squares, and

\[
d(\lambda_n) = \text{tr} \left\{ \frac{1}{N} \sum_{k=1}^{N} \left[ H_n(\hat{\theta}_n, U^{(k)}) \right] + nE_n(\hat{\theta}_n) \right\}^{-1} \times \left\{ \frac{1}{N} \sum_{k=1}^{N} \left[ H_n(\hat{\theta}_n, U^{(k)}) \right] \right\}
\]

is the effective number of parameters. Then, \( \lambda_{opt} \) is the minimizer of the \( \text{GCV}_{\lambda_n} \).

Note that \( W_i \) in (13) is an \( n_i \times n_i \) covariance matrix of \( y_i \), that can be computed
as \( W_i = \mathbb{E}_u(y_i | u_i) + \text{var}_u(y_i | u_i) \), where

\[
\begin{align*}
\mathbb{E}_u(y_i | u_i) &= \frac{1}{N} \sum_{k=1}^{N} \left[ V_i(\hat{\theta}_n, U^{(k)}_i) \right], \\
\text{var}_u(y_i | u_i) &= \frac{1}{N} \sum_{k=1}^{N} \left[ \mu_i(\hat{\theta}_n, U_i^{(k)}) \right]^2 - \left[ \frac{1}{N} \sum_{k=1}^{N} \mu_i(\hat{\theta}_n, U_i^{(k)}) \right]^2.
\end{align*}
\]

3.3 Asymptotic properties

In this section, we study the asymptotic properties of the estimators \( \hat{\beta} \) under the proposed penalized estimating equations. We assume that \( n_i \) is bounded but the number of subjects \( n \) goes to infinity in our asymptotic study. Also consider the framework where the true value of \( \beta_0 \) is partitioned \( \beta_0 = (\beta_{01}^T, \beta_{02}^T)^T \) and the corresponding design matrix into \( X_i = (X_{i(1)}, X_{i(2)}) \). In our study, the true regression coefficients are \( \theta_0 = (\beta_{01}, \beta_{02}, \alpha_0)^T \) where \( \alpha_0 \) is an \( h_n \)-dimensional vector depending on \( f_0 \). For technical convenience let \( \theta_n = (\theta_{01}, \theta_{02})^T \) where \( \theta_{01} = (\beta_{01}^T, \alpha_0^T)^T \) is \( (s = s^* + h_n) \)-dimensional vector of true values that the elements are all nonzero and \( \theta_{02} = \beta_{02} = 0 \). Here, \( s^* \) is the dimension of \( \theta_{01} \) and assume that only a small number of covariates contribute to the response i.e. \( f^* = \{1 \leq j \leq p; \beta_j \neq 0\} \) has cardinality \( |s| = s^* < p \). Consequently, estimated values and the design matrix is re-partitioned as \( \hat{\theta}_n = (\hat{\theta}_{n1}^T, \hat{\theta}_{n2}^T)^T \), and \( D_i = (D_{i(1)}, D_{i(2)})^T \) which \( \hat{\theta}_{n1} = (\hat{\beta}_{n1}^T, \hat{\alpha}_n^T)^T \), \( D_{i(1)} = (X_{i(1)}, B(t_i)^T)^T \), \( \hat{\theta}_{n2} = \hat{\beta}_{n2} \) and \( D_{i(2)} = X_{i(2)} \).

Meanwhile, If Eq. (9) has multiple solutions, only a sequence of consistent estimator \( \hat{\theta}_n \) is considered. A sequence \( \hat{\theta}_n \) is said to be a consistent sequence, if \( \hat{\beta}_n - \beta_{n0} \to 0 \) and \( \sup_t |B^T(t)\hat{\alpha}_n - f_0(t)| \to 0 \) in probability as \( n \to \infty \).

The following regularity conditions are required for the main results.

(A.1) \( n_i \) is a bounded sequence of positive integers, and the distinct values of \( t_{ij} \)
form a quasi-uniform sequence that grows dense on \([0, 1]\), and the \(k\)th derivative of \(f_0(t)\) is bounded for some \(k \geq 2\);

(A.2) \(X_{ij}, 1 \leq i \leq n, 1 \leq j \leq m\) are uniformly bounded;

(A.3) The unknown parameter \(\beta_n\) belongs to a compact subset \(B \subseteq \mathbb{R}^p\), the true parameter value \(\beta_{n0}\) lies in the interior of \(B\);

(A.4) There exist two positive constants, \(b_1\) and \(b_2\), such that

\[
\frac{b_2}{\lambda_{\min}} \leq \left( n^{-1} \sum_{i=1}^{n} X_i^\top X_i \right) \leq \frac{b_3}{\lambda_{\max}} \left( n^{-1} \sum_{i=1}^{n} X_i^\top X_i \right),
\]

where \(\lambda_{\min}\) (resp. \(\lambda_{\max}\)) denotes the minimum (resp. maximum) eigenvalue of a matrix;

(A.5) The common true correlation matrix \(R_0\) has eigenvalues bounded away from zero and \(+\infty\); the estimated working correlation matrix \(\hat{R}\) satisfies \(\|\hat{R}^{-1} - R^{-1}\| = O_p(n^{-1/2})\), where \(R\) is a constant positive definite matrix with eigenvalues bounded away from zero and \(+\infty\); we do not require \(\hat{R}\) to be the true correlation matrix \(R_0\);

(A.6) Let \(e_i(\theta_n, u_i) = (e_{i1}(\theta_n, u_i), \ldots, e_{i_{m_i}}(\theta_n, u_i))^\top = A_i^{-1/2}(\theta_n, u_i)(Y_i - \mu_i(\theta_n, u_i))\). There exists a finite constant \(M_1 > 0\) such that \(\mathbb{E}(\|e_i(\theta_{n0}, u_i)\|^{2+\delta}) \leq M_1\), for all \(i\) and some \(\delta > 0\); and there exist positive constants \(M_2\) and \(M_3\) such that \(\mathbb{E}\left[\exp(M_2|e_{ij}(\theta_{n0}, u_i)||X_i)\right] \leq M_3\), uniformly in \(i = 1, \ldots, n\), \(j = 1, \ldots, m\);

(A.7) Let \(B_n = \{\theta_n : \|\theta_n - \theta_{n0}\| \leq \Delta \sqrt{p_n/n}, \text{ then } \mu(\boldsymbol{D}^\top_{ij} \theta_n), 1 \leq i \leq n, 1 \leq j \leq m, \text{ are uniformly bounded away from } 0 \text{ and } \infty \text{ on } B_n\}; \mu^3(\boldsymbol{D}^\top_{ij} \theta_n)\) and \(\mu^{(3)}(\boldsymbol{D}^\top_{ij} \theta_n), 1 \leq i \leq n, 1 \leq j \leq m\), are uniformly bounded by a finite positive constant \(M_2\) on \(B_n\);
(A.8) Assuming $\min_{1 \leq k \leq s_n} |\theta_{n0k}|/\lambda_n \to \infty$ 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 $p_n s_n^2 (\log n)^8 = o(n^2 \lambda_n^4)$.

Under (A.1), the total sample size $n$ is of the same order as the number of subjects $m$ and there exists only local dependence in the sample. The smoothness condition on $f_0$ given by (A.1) determines the rate of convergence of the spline estimate $\hat{f}(t) = B(t)^\top \hat{\alpha}_n$. Condition (A.2) is a common assumption in the literature on M-estimators with diverging dimension. When $m = 1$ (i.e., each subject has only one observation), condition (A.4) is also popularly adopted in the literature on regression for independent data. Condition (A.5) is a similar assumption in the Liang and Zeger (1986), assumes that the estimator of the working correlation matrix parameter $\hat{\tau}$ satisfies $\sqrt{n}(\hat{\tau} - \tau_0) = O_p(1)$ for some $\tau_0$. Note that when a nonparametric moment estimator is used for the working correlation matrix, we have $\bar{R} = R_0$. The first part of condition (A.6) is similar to the condition in Lemma 2 of Xie and Yang (2003) and condition (A.5) in Balan and Schiopu-Kratina (2005); the second part is satisfied for Gaussian distribution, sub-Gaussian distribution, and Poisson distribution, etc. Condition (A.7) requires $\mu_{ij}^{(k)}(D_{ij}^\top \theta_n)$, which denotes the $k$th derivative of $\mu_{ij}^{(k)}(t)$ evaluated at $D_{ij}^\top \theta_n$, to be uniformly bounded when $\theta_n$ is in a local neighborhood around $\theta_{n0}, k = 1, 2, 3$. This condition is generally satisfied for the GEE. For example, when the marginal model follows a Poisson distribution, $\mu(t) = \exp(t)$; thus $\mu_{ij}^{(k)}(D_{ij}^\top \theta_n) = \exp(D_{ij}^\top \theta_n), k = 1, 2, 3$, are uniformly bounded around $\theta_{n0}$ on $B_n$.

Now, consider the following estimating equation

$$\bar{S}_n(\theta) = \mathbb{E}_{u_i} \left[ \sum_{i=1}^n D_{ij}^\top A_i^\frac{1}{2} (\theta, u_i) \bar{R}^{-1} A_i^{-\frac{1}{2}} (\theta, u_i) (y_i - \mu_i (\theta, u_i)) \right].$$
Let $\overline{M}_n(\theta_n)$ to be the covariance matrix of $\overline{S}_n(\theta)$, then

$$
\overline{M}_n(\theta) = \mathbb{E}_{u_i|x}\left[ \sum_{i=1}^n D_i^\top A_i^\top(\theta, u_i)R^{-1}R_0R^{-1}A_i^\top(\theta, u_i)D_i \right].
$$

However, for the estimating equation approach, alternative techniques are necessary to establish the asymptotic theory, in the supplementary file we present some useful lemmas. By Lemma 1, we approximate $f_0(t)$ by $B(t)\alpha_0$, then have

$$
\eta_{ij}(\theta) = g(\mu_{ij}(\theta_0)) = X_{ij}^\top \beta_0 + B(t_{ij})\alpha_0 + Z_{ij}^\top u_i, \quad \theta_0 = (\beta_0^\top, \alpha_0^\top)_{(p_n+N)\times 1}.
$$

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

**Theorem 1. (Existency).** Assume the conditions (A.1)–(A.8). Then, there exists an approximate penalized GEE solution $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$ which satisfies the following properties

$$
(i) \quad \mathbb{P}_n\left(|U_{nk}(\hat{\theta}_n)| = 0, \; k = 1, \ldots, s_n^*, (s_n^* + 1), \ldots, (s_n = s_n^* + h_n)\right) \to 1,
$$

$$
(ii) \quad \mathbb{P}_n\left(|U_{nk}(\hat{\theta}_n)| \leq \frac{\lambda_n}{\log n}, \; k = (s_n^* + h_n + 1), \ldots, p_n\right) \to 1,
$$

where

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

and $S_{nk}(\hat{\theta}_n)$ denotes the $k$th element of $S_n(\hat{\theta}_n)$.

Theorem 2 provides a more precise characterization for the approximate solution of the penalized estimating equation procedure.
Theorem 2. (Consistency). Assume conditions (A1)–(A8) and that \( n^{-1}p_n^2 = o(1) \). Then, \( U_n(\theta_n) = o(1) \) has a root \( \hat{\theta}_n \) such that

\[
\begin{align*}
(i) & \quad \|\hat{\theta}_n - \theta_{n0}\| = O_p(\sqrt{p_n/n}), \\
(ii) & \quad \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (\hat{f}(t_{ij}) - f_0(t_{ij}))^2 = O_p(n^{-2r/(2r+1)}).
\end{align*}
\]

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

Theorem 3. (Asymptotic normality). Assume (A.1)–(A.8). If the number of knots \( L_n = O_p(n^{1/(2r+1)}) \), and \( n^{-1}p_n^3 = o(1) \), then \( \forall \xi_n \in \mathbb{R}^{p_n} \) such that \( \|\xi_n\| = 1 \), we have

\[
\begin{align*}
(i) & \quad \mathbb{P}_n(\hat{\beta}_{n2} = 0) \to 1, \\
(ii) & \quad \xi_n^\top \overline{M}_n^{-1/2} \overline{H}_n^{*} (\beta_{n0}) (\hat{\beta}_{n1} - \beta_{n01}) \xrightarrow{D} N_{p_n}(0, 1),
\end{align*}
\]

where

\[
\begin{align*}
\overline{M}_n^{*} &= \mathbb{E}_{u|y} \left[ \sum_{i=1}^n X_i^* \mathbb{A}_i^{3/2} (\theta_n, u_i) \overline{R}^{-1} R_0 \overline{R}^{-1} \mathbb{A}_i^{3/2} (\theta_n, u_i) X_i^* \right], \\
\overline{H}_n^{*} &= \mathbb{E}_{u|y} \left[ \sum_{i=1}^n X_i^* \mathbb{A}_i^{3/2} (\theta_n, u_i) \overline{R}^{-1} \mathbb{A}_i^{3/2} (\theta_n, u_i) X_i^* \right],
\end{align*}
\]

\( X_i^* = (I - P) X_i, \ P = B(B^\top \Omega B)^{-1} B^\top \Omega, \ \Omega = \text{diag}\{\Omega_i\} \) and \( \Omega_i = \mathbb{E}_{u|y} \left[ \mathbb{A}_i^{3/2} (\theta_n, u_i) \overline{R}^{-1} \mathbb{A}_i^{3/2} (\theta_n, u_i) \right] \).

Theorem 3 is often referred to as the oracle properties of variable selection, that is, the procedure estimates the true zero coefficient as zero with probability approaching
one and estimates the nonzero coefficients as efficiently as if the true model is known in advance.

From the Algorithm 1, we obtain the following sandwich formula to estimate the asymptotic covariance matrix of $\hat{\theta}_n$:

$$\text{Cov}(\hat{\theta}_n) \approx [H_n(\hat{\theta}_n, u_i) + nE_n(\hat{\theta}_n)]^{-1}M_n(\hat{\theta}_n, u_i)[H_n(\hat{\theta}_n, u_i) + nE_n(\hat{\theta}_n, u_i)]^{-1},$$

where $H_n$ and $E_n$ are defined in Section 3 and

$$M_n(\hat{\theta}_n, u_i) = \sum_{i=1}^n D_i^\top A_i^{1/2}(\hat{\theta}_n, u_i) \hat{R}^{-1} [\epsilon_i(\hat{\theta}_n, u_i) \epsilon_i^\top(\hat{\theta}_n, u_i)] \hat{R}^{-1} A_i^{1/2}(\hat{\theta}_n, u_i) D_i^\top.$$

## 4 Numerical Studies

In this section, we first conduct simulation study to illustrate the consistency and the sure screening property of the proposed procedure empirically, and compare its finite sample performance with some other different model settings. We further apply our proposed method for analyzing two real data sets.

### 4.1 Simulation studies

We simulated 100 data sets from the GSMM for longitudinal data. The underlying model is the random intercept Poisson model so that the nonlinear function were set to a sinusoidal function

$$y_{ij} | b_i \sim \text{Pois}(\mu_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

$$\eta_{ij} = \log(\mu_{ij}) = \sum_{k=1}^p x_{ij}^{(k)} \beta_k + \sin(2\pi t_{ij}) + b_i.$$
where $i = 1, \ldots, n$ ($n = 50, 100$ and $150$), and $j = 1, \ldots, n_i$ which the number of observations per subjects assumed to be fixed at $n_i = 5$. The true regression coefficients are $\beta = (-1, -1, 2, 0, \ldots, 0)$ with the mutually independent covariates $X_{ij}^T = (x_{ij}^{(1)}, \ldots, x_{ij}^{(p)})$ are drawn independently from uniform distribution on $(-1, 1)$. The measurement time points $t_{ij}$ are drawn from uniform distribution on $(0, 1)$. The random effect process $b_i$ is taken to be a Gaussian process with mean $0$, variance $\sigma^2 = 0.25$. 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 $p_n = \lceil \frac{n}{2} \rceil$, $p_n = [4.5n^{1/4}]$, and $p_n = \lceil \frac{n}{b \log(n)} \rceil$, 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 $p_n \gg n$, we can mention to $\log(p_n) = o_p(n^b)$, where $0 < b < 1$. Of course 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. In particular, the binary response of a logistic regression model and, to a lesser extent, the integer valued response in a Poisson regression model are less informative than the real valued response in a linear model. 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 $p_n \gg n$. The choices seem to be satisfactory and consist of all suggestions, as the performance of the procedures is quite robust to different choices of $p_n$.

The performance of estimators was evaluated separately for the structural com-
ponents and the variance. By averaging

Performance of the proposed penalized procedure compared with the unpenalized one and the penalized GLMM where each simulated data set was fitted under these three methods. For evaluating estimation accuracy, we report the empirical mean square error (MSE), defined as \( \sum_{k=1}^{100} \| \hat{\beta}^k_n - \beta_{n0} \| / 100 \) where \( \hat{\beta}^k_n \) is the estimator of \( \beta_{n0} \) obtained using the \( k \)th generated data set. The performance in variable selection is gauged by \((C, I)\), where 'C' is the mean over all 100 simulations of zero coefficients which are correctly estimated by zero and 'I' is the mean over all 100 simulations of nonzero coefficients which are incorrectly estimated by zero. To present a more comprehensive picture, we also use other criteria for variable selection performance evaluation. 'Under-fit' corresponds to the proportion of excluding any true nonzero coefficients. Similarly, we report the proportion of selecting the exact subset model as 'Correct-fit' and the proportion of including all three important variables plus some noise variables as 'Over-fit'.

The results of Table 1 summarize the estimation accuracy and model selection properties of the penalized GSMM (P-GSMM), the unpenalized GSMM, and the penalized GLMM (P-GLMM) for the different values of \( (n, p_n) \). In terms of estimation accuracy the penalized GSMM procedure performs closely to the penalized GLMM, whereas our proposed approach gives the smallest MSE, and consistently outperforms its penalized GLMM counterpart. In terms of model selection we observe that the unpenalized GSMM generally does not lead to a sparse model. Furthermore, the penalized GSMM and the penalized 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) than the penalized GLMM. For penalized 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.

To further investigate the performance of the proposed method, Table 2 reports its bias, the estimated standard deviation (calculated from the sandwich variance formula), the empirical standard deviation, and the empirical coverage probability of 95% confidence interval for estimating $\beta_1$, $\beta_2$, and $\beta_3$. The estimated standard deviation is close to the empirical standard deviation, and the empirical coverage probability is close to 95%. This indicated good performance of the sandwich variance formula.

These observations suggest that considering partial part is important to modify the estimation accuracy and model selection when the growth curves of the data exhibit a nonlinear fashion over time, especially in a complicated manner. On the other hand, penalized GSMM allows us to make systematic inference on all model parameters by representing a partially model as a modified penalized GLMM.

Table 1: Model selection results for poisson responses: comparision of P-GSMM, GSMM, and P-GLMM with the cases of $p_n < n$ and $p_n >> n$

| method     | case $p_n < n$ | case $p_n >> n$ |
|------------|----------------|-----------------|
|            | $(n, p) = (50, 11)$ | $(n, p) = (50, 11)$ | $(n, p) = (30, 100)$ | $(n, p) = (50, 100)$ | $(n, p) = (100, 500)$ | $(n, p) = (200, 2000)$ |
|            | MSE | C(8) | I(0) | Under-fit | Correct-fit | Over-fit | MSE | C(8) | I(0) | Under-fit | Correct-fit | Over-fit | MSE | C(8) | I(0) | Under-fit | Correct-fit | Over-fit | MSE | C(8) | 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 |
|            | 0.072 | 0.16 | 0.00 | 0.00 | 0.00 | 1.00 | 1499.136 | 47.02 | 0.00 | 0.00 | 0.00 | 1.00 |
|            | 0.043 | 10.52 | 0.00 | 0.00 | 0.77 | 0.23 | 0.062 | 495.720 | 0.00 | 0.00 | 0.89 | 0.11 |
|            | 0.036 | 10.70 | 0.00 | 0.00 | 0.93 | 0.07 | 0.038 | 496.250 | 0.00 | 0.00 | 0.92 | 0.08 |
|            | 0.060 | 0.26 | 0.00 | 0.00 | 0.00 | 1.00 | 125.406 | 1137.62 | 0.00 | 0.00 | 0.00 | 1.00 |
|            | 0.044 | 11.25 | 0.00 | 0.00 | 0.92 | 0.08 | 0.018 | 1996.89 | 0.00 | 0.00 | 0.30 | 0.79 |
|            | 0.045 | 11.87 | 0.00 | 0.00 | 0.96 | 0.04 | 0.018 | 1996.93 | 0.00 | 0.00 | 0.54 | 0.46 |
Table 2: Estimation results for Poisson response: performance of the P-GSMM with the cases of $p_n < n$ and $p_n >> n$. Bias: absolute value of the empirical bias; SD1: estimated standard deviation using the sandwich variance estimator; SD2: sample standard deviation; CP: denotes the empirical coverage probability of the 95% confidence interval.

| Case $p_n < n$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | Case $p_n >> n$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|----------------|-----------|-----------|-----------|----------------|-----------|-----------|-----------|
| $(n, p_n)$     |           |           |           | $(n, p_n)$     |           |           |           |
| (50, 11)       | Bias      | 0.047     | 0.096     | 0.069         | Bias      | 0.147     | 0.344     | 0.365     |
|                | SD1       | 0.092     | 0.085     | 0.113         | SD1       | 0.084     | 0.071     | 0.094     |
|                | SD2       | 0.097     | 0.094     | 0.097         | SD2       | 0.162     | 0.178     | 0.148     |
|                | CP        | 0.96      | 0.95      | 0.92          | CP        | 0.95      | 0.96      | 0.97      |
| (100, 14)      | Bias      | 0.076     | 0.103     | 0.053         | Bias      | 0.078     | 0.061     | 0.017     |
|                | SD1       | 0.071     | 0.067     | 0.084         | SD1       | 0.049     | 0.044     | 0.066     |
|                | SD2       | 0.072     | 0.067     | 0.078         | SD2       | 0.070     | 0.072     | 0.086     |
|                | CP        | 0.96      | 0.95      | 0.96          | CP        | 0.94      | 0.95      | 0.96      |
| (150, 15)      | Bias      | 0.101     | 0.124     | 0.099         | Bias      | 0.049     | 0.072     | 0.028     |
|                | SD1       | 0.060     | 0.059     | 0.070         | SD1       | 0.039     | 0.040     | 0.054     |
|                | SD2       | 0.052     | 0.059     | 0.059         | SD2       | 0.051     | 0.052     | 0.058     |
|                | CP        | 0.97      | 0.96      | 0.95          | CP        | 0.94      | 0.92      | 0.94      |

For proposed method, the estimated baseline function $f(t)$ is also evaluated through visualization. We plot and compare the estimated $f(t)$ and pointwise biases, for the cases of $p_n < n$ and $p_n >> n$ by two sample size $n = 50$ and 100. We also plot the pointwise standard deviations (calculated from the sandwich variance formula), and coverage probability of 95% confidence intervals. Figures 1 shows that the $p_n < n$ case, our approach yields smaller overall biases and standard deviations than the $p_n >> n$ case. Also, it can be seen that the empirical coverage probability for $f(t)$ is close to 95% for two cases. Figure 2 depicts the results for $n = 100$. As shown, larger sample size modified the biases and the differences between two cases. Nevertheless, the case of $p_n < n$ has smaller standard deviation.
Figure 1: Plots for estimated $f(t)$ in the $p_n < n$ and $p_n >> n$ cases ($n=50$) based on 100 samples. Plots top-left and top-right show the averaged fit and pointwise bias; plot bottom-left shows the standard deviation; and plot bottom-right plots the averaged coverage probability rates for 95% confidence intervals.

4.2 Real data analyses

4.2.1 AIDS data

In this section, to illustrate our method, we considered the longitudinal CD4 cell count data among HIV seroconverters. This dataset contains 2376 observations of CD4 cell counts on 369 men infected with the HIV virus; see (Zeger and Diggle, 1994) for a detailed description of this dataset. Figure 3 (top-left) display the trajectories of 369 men for exploring the evolution of CD4 cell counts.

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
Figure 2: Plots for estimated $f(t)$ in the $p_n < n$ and $p_n >> n$ cases ($n=100$) based on 100 samples. Plots top-left and top-right show the averaged fit and pointwise bias; plot bottom-left shows the standard deviation; and plot bottom-right plots the averaged coverage probability rates for 95% confidence intervals.

CESD scale (larger values indicate increased depressive symptoms) and YEAR (the effect of time since seroconversion). Since there seems to exist a positive correlation among responses from the same patient, we need to incorporate a correlation structure into the estimation scheme. Zeger and Diggle (1994) found that the compound symmetry covariance matrix fitted the data reasonably well. This data analysed by many authors such as Wang et al. (2005), Huang et al. (2007) and Ma et al. (2013).

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 nonpara-
metrical, 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. We used the SCAD penalty, and the tuning parameter $\lambda = 0.45$. To compare the performance of our proposed method (P-SMM) with other two existing scenarios, including the unpenalized GSMM, and the penalized GLMM (P-GLMM), we use the standard errors (SE) were all calculated using the sandwich method. To best identify a model supported by the data, we adopt the Akaike information criterion (AIC; Akaike (1973)) and the Bayesian information criterion (BIC; Schwarz (1978)). They are defined as

$$AIC = 2m - 2\ell_{\text{max}}, \quad BIC = m \log n - 2\ell_{\text{max}}$$

where $\ell_{\text{max}}$ is the maximized log-likelihood value, $m$ is the number of free parameters in the model. Table 3 presents the summary of the fitting results including the values of standard errors, together with $\ell_{\text{max}}$, AIC, and BIC under the three models.

Judging from Table 3, the P-GSMM tends to exhibit slightly standard errors compared to GSMM and P-GLMM, nevertheless this difference is not more dramatic. Meanwhile, the values of AIC, BIC of our proposed model are smaller than those for the other two competing models, revealing that the P-GSMM can provide better fitting performance. Under P-GSMM, SMOKE, DRUGS, SEXP, SMOKE * 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). The Results for nonparametric curve estimates using the P-GSMM estimators are plotted in Figure 3 for \textit{YEAR}. It shows the
Table 3: Summary of parameter estimates along with standard errors (in parentheses) under the three fitted models for the AIDS data.

| Variables  | GSMM  | P-GLMM | P-GSMM |
|------------|-------|--------|--------|
| 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) | 6.065 (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 |

The estimated nonparametric function $f(t)$, its 95% pointwise confidence bands, standard deviation, and 95% coverage probability given by the empirical and sandwich formula variance. We can see that the baseline function $f(t)$ has decreasing effect as time passing. Therefore, one can see that it is more reasonable to put it as a nonparametric component. We notice the disparity between the empirical and the sandwich formula standard deviation in the boundary positions and the sandwich formula standard deviations are smaller in which case the coverage probability recede from 95%.
Figure 3: Plots for estimated $f(t)$ for AIDS data based on P-GSMM. Plot top-left shows the trajectories plot for CD4 data. Observed evolution (in gray) of CD4 cell counts for 369 men against time (in YEAR). Solid (in thick blue) line show the smoothed mean profile of men. Plot top-right shows the estimated baseline function $f(t)$ (in thick blue) in the selected model of P-GSMM and the 95% confidence interval (dashed line) corresponding to the robust confidence interval. Plots bottom-left and bottom-right respectively, show the standard deviation and coverage probability rates for 95% confidence intervals based on empirical variance and sandwich formula.

4.2.2 Yeast Cell-Cycle Gene Expression data

A yeast cell-cycle gene expression data collected in the CDC15 experiment of Spellman et al. (1998) where genome-wide mRNA levels of 6178 yeast ORFs (abbreviation for open reading frames, which are DNA sequences that can determine which amino acids will be encoded by a gene) at 7 minute intervals for 119 minutes, which covers two cell-cycle periods for a total of 18 time points, measured. The cell cycle is a tightly regulated life process where cells grow, replicate their DNA, segregate their chromosomes, and divide into as many daughter cells as the environment allows.
The cell-cycle process is commonly divided into M/G1-G1-S-G2-M stages. Refer to Wang et al. (2012), for more detailed description of this data set.

Transcription factors (TFs) have been observed to play critical roles in gene expression regulation. A TF (sometimes called a sequence-specific DNA-binding factor) is a protein that binds to specific DNA sequences, thereby controlling the flow (or transcription) of genetic information from DNA to mRNA. To better understand the phenomenon underlying cell-cycle process, it is important to identify TFs that regulate the gene expression levels of cell cycle-regulated genes. It is not clear where these TFs regulate all cell cycle genes, however.

We applied our methods to identify the key TFs. The dataset that we use present a subset of 283 cell-cycled-regularized genes observed over 4 time points at G1 stage. The response variable $Y_{ij}$ is the log-transformed gene expression level of gene $i$ measured at time point $j$, for $i = 1, \ldots, 283$ and $t = 1, \ldots, 4$. We use the following semiparametric mixed model

$$y_{ij} = \sum_{k=1}^{96} x_{ij}^{(k)} + f(t_{ij}) + b_i,$$

where the covariates $x_{ij}^{(k)}$, $k = 1, \ldots, 96$, is the matching score of the binding probability of the $k$th TF on the promoter region of the $i$th gene. The binding probability is computed using a mixture modeling approach based on data from a ChIP binding experiment; see Wang et al. (2007) for details. Covariates $x_{ij}^{(k)}$ is standardized to have mean zero and variance 1. $t_{ij}$ denotes time, $f(t_{ij})$ models the nonparametric time effect, and $b_i$ is the random intercept. Our goal is to identify the TFs that might be related to the expression patterns of these 283 cellcycleregulated genes. Therefore we apply a penalization procedure by the proposal P-GSMM and also by ignoring the nonparametric component $f(t_{ij})$ using P-GLMM.
Table 4 summarizes the TFs identified when p-GSMM and p-GLMM are adopted. Our analysis reveals that a total of 13 and 16 TFs related to yeast cell-cycle processes are identified respectively by the P-GSMM and P-GLMM. The sets of TFs selected at different methods have only small overlaps. These common TFs are GAT3, MBP1, MSN4, PHD1, SMP1, SWI4, and SWI6. For stage G1, MBP1, SWI4, and SWI6 are three TFs that have been proved important in the aforementioned biological experiments and they have been selected by the two methods. However, model selection criteria, including the values of standard errors, together with \( \ell_{\text{max}} \), AIC, and BIC confirm the superiority of our proposed model.
5 Concluding Remarks

In general, when the number of covariates is large, identifying the exact underlying model is a challenging task, in particular when some of the nonzero signals are relatively weak. Here, we have developed a general methodology for simultaneously selecting variables and estimating the unknown components in semiparametric mixed effects model for non-Gaussian longitudinal data when the number of parameters diverges with the sample size. We approximate the nonparametric components based on polynomial B-spline smoothing, propose a penalized estimating equation approach to do variable selection, and the estimation of the both parametric and nonparametric components obtained in the framework of maximum likelihood estimation. We apply the penalty functions to the estimating equation objective function such that the proposed procedure can simultaneously estimate parameters and select the important variables. The procedure involves the specification of the posterior distribution of the random effects, which cannot be evaluated in a closed form, and we use a Metropolis algorithm, which does not require the specification of the posterior distribution. To implement the procedure in practice, a computationally flexible iterative algorithm is developed. We obtain asymptotic normality of the resulting estimators and demonstrate that the asymptotic normality of the estimated coefficients for the linear part is retained. To investigate the performance of our approach, we compared it 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 set including AIDS data and Yeast Cell-Cycle Gene Expression data. Results show that the proposal works well, and can estimate the nonzero coefficients and select model simultaneously, the model outperforms the penalized generalized linear mixed effects counterparts on the provision of likelihood-based model selection and estimating the regression coefficients. In addition, we can find that estimation is
more efficient when the partially part is taken into consideration. The results are consistent in both cases of \( p_n < n \) and \( p_n \gg n \). Our primary interest is the linear components, and we treat the nonparametric functions as nuisance components; thus we limit our discussions to estimation and variable selection for the linear part. Nonetheless, this may be extended to the nonparametric components using techniques similar to those in [Xue (2009)]. However, the technical details deserve careful consideration, and this is an interesting topic of future research.

**SUPPLEMENTARY MATERIAL**

**Web-based Supplementary Materials:** The proofs of the main results as well as some instrumental lemmas are provided in a separate supplementary file. (.pdf file)

**References**

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", *In 2nd Int. Symp. on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267–281. Akademiai Kiado, Budapest.

Balan, R. M., and Schiopu-Kratina, I. (2005), "Asymptotic results with generalized estimating equations for longitudinal data", *The Annals of Statistics*, 32, 522–541.

Bondell, H. D., Krishna, A., and Ghosh S. K. (2010), "Joint variable selection for fixed and random effects in linear mixed-effects models", *Biometrics*, 66, 1069–1077.

Cantoni, E., Mills, F. J., and Ronchetti, E. (2005), "Variable selection for marginal longitudinal generalized linear models", *Biometrics*, 61(2), 507–514.

Chiou, J. M., and Muller, H. G. (2005), "Estimated estimating equations: semi-parametric inference for clustered and longitudinal data", *Journal of the Royal Statistical Society* B, 67(4), 531–553.

Chu, W., Li, R., and Reimherr, M. (2016), "Featurescreening for time-varying coefficient models with ultrahigh dimensional longitudinal data", *Annals of Applied Statistics*, 10, 596–617
Dziak, J. J. (2006), ”Penalized quadratic inference functions for variable selection in longitudinal research”, Ph.D Thesis, the Pennsylvania State University. (https://etda.libraries.psu.edu/paper/7084/)

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

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

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

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004), *Applied longitudinal analysis*, John Wiley & Sons, Hoboken.

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

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

He, X. M., Fung, W. K., and Zhu, Z. Y. (2005), ”Robust estimation in generalized partial linear models for clustered data”, *Journal of the American Statistical Association*, 100, 1176–1184.

He, X. M., Zhu, Z. Y., and Fung, W. K. (2002), ”Estimation in a semiparametric model for longitudinal data with unspecified dependence structure”, *Biometrika*, 89, 579–590.

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, 451–477.

Johnson, B., Lin, D. Y., and Zeng, D. (2008), ”Penalized estimating functions and variable selection in semiparametric regression models”, *Journal of the American Statistical Association*, 103, 672–680.

Kurum, E., Li, R., Shiffman, S., and Yao, W. (2016), ”Time- varying coefficient models for joint modeling binary and continues outcome in longitudinal data”, *Statistica sinica*, 29, 979–1000.

Laird, N. M., and Ware, J. H. (1982), ”Random effects models for longitudinal data”, *Biometrics*, 38, 963–974.
Liang, H. (2009), "Generalized partially linear mixed-effects models incorporating mismeasured covariates", *Annals of the Institute of Statistical Mathematics*, 61, 27–46.

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

Li, Z., and Zhu, L. (2010), "On variance components in semiparametric mixed models for longitudinal data", *Scandinavian Journal of Statistics*, 37, 442–457.

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.

McCulloch, C.E. (1997), "Maximum likelihood algorithms for generalized linear mixed models", *Journal of the American Statistical Association*, 92, 162–170.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*. Chapman and Hall, London, second edition.

Ni, X., Zhang, D., and Zhang, H. H. (2010), "Variable selection for semiparametric mixed models in longitudinal studies", *Biometrics*, 66, 79–88.

Pan, W. (2001), "Akaikes information criterion in generalized estimating equations", *Biometrics*, 57, 120–125.

Pan, W. (2002), "Goodness-of-fit tests for GEE with correlated binary data", *Scandinavian Journal of Statistics*, 29, 101–110.

Qin, G.Y., and Zhu, Z.Y., (2007), "Robust estimation in generalized semiparametric mixed models for longitudinal data", *Journal of Multivariate Analysis*, 98, 1658–1683.

Qin, G.Y., and Zhu, Z.Y., (2009), "Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data", *Biometrics*, 65, 52–59.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving generalised estimating equations using quadratic inference functions", *Biometrika*, 87, 823–836.

Schwarz, G. (1978), "Estimating the dimension of a model", *The Annals of Statistics*, 6, 461–464.

Sinha, S.K., and Sattar, A. (2015), "Inference in semi-parametric spline mixed models for longitudinal data”, *METRON*, 73, 377–395.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", *Molecular Biology of Cell*, 9, 3273–3297.
Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.), New York: Springer-Verlag.

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

Ueki, M. (2009), "A note on automatic variable selection using smooth-threshold estimating equations", *Biometrika*, 96(4), 1005–1011.

Wang, L. (2011), "GEE analysis of clustered binary data with diverging number of covariates", *The Annals of Statistics*, 39, 389–417.

Wang, L., Chen, G., and Li, H. (2007), "Group SCAD regression analysis for microarray time course gene expression data", *Bioinformatics*, 23, 1486–1494.

Wang, L., and Qu, A. (2009), "Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach", *Journal of the Royal Statistical Society*, Ser. B, 71, 177–190.

Wang, L., Zhou, J., and Qu, A. (2012), "Penalized generalized estimating equations for high-dimensional longitudinal data analysis", *Biometrics*, 68(2), 353–360.

Zeger, S. L., and Karim, M. R. (1991), "Generalized linear models with random effects: A Gibbs sampling approach", *Journal of the American Statistical Association*, 86, 79–86.
Zeger, S. L., and Diggle, P. J. (1994), "Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters", *Biometrics*, 50, 689–99.

Zhang, D. (2004), "Generalized linear mixed models with varying coefficients for longitudinal data", *Biometrics*, 60, 8–15.

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

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elasticnet", *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.