Received: 23 July 2021 | Accepted: 28 March 2022

DOI: 10.1111/biom.13672

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Model-based clustering of high-dimensional longitudinal data via regularization

Luoying Yang | **Tong Tong Wu**

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York, USA

Correspondence
Tong Tong Wu, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA.
Email: tongtong_wu@urmc.rochester.edu

Funding information
National Science Foundation, Grant/Award Number: NSF-CCF-1934962; National Heart, Lung, and Blood Institute, Grant/Award Number: NIH/R01HL119058

Abstract

We propose a model-based clustering method for high-dimensional longitudinal data via regularization in this paper. This study was motivated by the Trial of Activity in Adolescent Girls (TAAG), which aimed to examine multilevel factors related to the change of physical activity by following up a cohort of 783 girls over 10 years from adolescence to early adulthood. Our goal is to identify the intrinsic grouping of subjects with similar patterns of physical activity trajectories and the most relevant predictors within each group. The previous analyses conducted clustering and variable selection in two steps, while our new method can perform the tasks simultaneously. Within each cluster, a linear mixed-effects model (LMM) is fitted with a doubly penalized likelihood to induce sparsity for parameter estimation and effect selection. The large-sample joint properties are established, allowing the dimensions of both fixed and random effects to increase at an exponential rate of the sample size, with a general class of penalty functions. Assuming subjects are drawn from a Gaussian mixture distribution, model effects and cluster labels are estimated via a coordinate descent algorithm nested inside the Expectation-Maximization (EM) algorithm. Bayesian Information Criterion (BIC) is used to determine the optimal number of clusters and the values of tuning parameters. Our numerical studies show that the new method has satisfactory performance and is able to accommodate complex data with multilevel and/or longitudinal effects.

KEYWORDS
exponentially growing number of variables, linear mixed-effects models, nonconcave penalty functions, simultaneous effects selection

1 | INTRODUCTION

Longitudinal data are commonly encountered in various research fields. Very often, subjects recruited in the same study have different characteristics and therefore investigators would like to address the subgroup heterogeneity by grouping those subjects. Specifically, clustering of longitudinal data has drawn attention recently. Some methods were developed to cluster the longitudinal trajectories of the outcome variable (Genolini and Falissard, 2010; McNicholas and Murphy, 2010). Such methods have drawbacks such as the same length of trajectories is required, which means missing data are not allowed and need to be removed from the analysis. A more serious problem is that those methods cannot model the relationship between the outcome variable and covariates. An alternative approach is to assume a mixture of regression models, such as linear or generalized linear mixed-effects models

(LMM, Komárek and Komárková, 2013; Proust-Lima *et al.*, 2015) or semiparametric mixed-effects models (Arribas-Gil *et al.*, 2015). However, challenges arise with the increasing dimensionality of longitudinal data. The estimation and inference for those data become much more difficult when the number of covariates increases.

In this paper, we will introduce a novel clustering method for high-dimensional longitudinal data. It can perform clustering and variable selection simultaneously, allowing the dimensions of fixed and random effects to grow at an exponential rate of the sample size under a general class of penalty functions with the proved joint properties. This idea was motivated by the Trial of Activity in Adolescent Girls (TAAG) (Young *et al.*, 2018), which followed up a cohort of girls at three waves: in 2006 at their middle school, in 2009 at their high school, and in 2015 in their early adulthood. Holding rich insights to physical health of adolescent girls, this high-dimensional longitudinal data set has been analyzed in previous papers (Young *et al.*, 2018; LaLonde *et al.*, 2019; Young *et al.*, 2019). However, all these papers conducted the analysis in two separate steps, that is, LMM with fixed effects selection for the whole sample first and then clustering on the outcome trajectories or random effects. Those analyses failed to model the cluster-specific relationship between the outcome and the covariates. Second, no random effects selection was enabled. Third, although a total of 730 subjects were enrolled at the first wave of the study, only 428 had data available at all three waves. The subjects with missing waves were discarded in the previous analyses.

Although not much available for clustering of high-dimensional longitudinal data, there is a considerable amount of literature in the shrinkage estimation for LMM that are fitted for longitudinal data under the sparsity assumption; see Müller *et al.* (2013) for a thorough review. Recently, Li *et al.* (2018) proposed a method that selects both fixed effects and random effects via the LASSO penalty with diverging p and q. However, the asymptotic properties were proved in two separate steps—assuming the fixed-effects components to be fixed when proving the consistency of random effects estimates and vice versa. Pointed out by one reviewer, there is one paper by Du *et al.* (2013) worked on a mixture of LMM with variable selection that can handle high-dimensional covariates. However, there are some limitations: (1) The selection of random effects in Du *et al.* (2013)'s paper is through the penalization of the diagonal elements of the random effects. If the diagonal element is estimated to be zero, then all the corresponding off-diagonal elements are assumed to be zero. Our model considers the overall effects of individual random components and penalizes the whole vector with some group penalty to achieve "all-in-or-all-out." (2) This paper considers the same penalty function for the fixed effects and the diagonal elements of the random effects (e.g., LASSO penalty on both), while our method allows for different penalty functions, which is important since a group penalty is the most appropriate to penalize the random effects in an "all-in-or-all-out" fashion. (3) The oracle properties of Du *et al.*'s model were established for fixed p and q. In this paper, we aim to investigate a general case by considering exponentially growing p and q for a general class of penalty functions, which is a much harder problem, and close the gap by proving the joint properties in one step.

The contributions of this paper are the following. First, it is the first clustering method for high-dimensional longitudinal data with diverging p and q. Clusters can be characterized with different sets of fixed and random effects. The dimension of fixed and random effects can grow at an exponential rate of the sample size. In addition, a general class of penalty functions that satisfy some regularity conditions can be used for the selection of effects. The convergence rate and sparsistency (Lam and Fan, 2009) was proved for the fixed and random effects jointly in one step.

The reminder of the paper is organized as follows. Section 2 concerns a simultaneous selection of exponentially growing numbers of fixed effects and random effects in LMM within clusters. Section 3 introduces the clustering method via the mixture of the penalized LMM from the previous section. The model performance will be evaluated in Section 4 using simulated data. The method will be applied to the TAAG data set in Section 5. A discussion will be provided in Section 6.

2 | LMM WITH DIVERGING NUMBERS OF FIXED AND RANDOM EFFECTS

Suppose that we have a sample of n subjects. For the ith subject, we collect the data at m_i time points such that at time t_{ij}, $j = 1, \dots, m_i$, we observe $(y_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})$, where y_{ij} is the response variable, $\boldsymbol{X}_{ij} \in \mathbb{R}^{p_n}$ is the fixed effects covariates, and $\boldsymbol{Z}_{ij} \in \mathbb{R}^{q_n}$ is the random effects covariates. The number of repeated measures m_i can differ from subject to subject. The use of p_n and q_n is to emphasize the dimension of covariates can increase as n at a certain rate. Using the matrix notation $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im_i})^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^T, \dots, \boldsymbol{X}_{im_i}^T)^T$, and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^T, \dots, \boldsymbol{Z}_{im_i}^T)^T$, the LMM for the ith subject is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\beta}$ is the constant fixed effects coefficients of size p_n, and $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$ is the subject-specific random effects coefficients of size q_n, and $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{I}_{m_i})$ is the i.i.d.

random error. The covariance matrix of the random effects $\boldsymbol{D}$ specifies the within-subject correlation.

In model (1), conditioning on $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, $\boldsymbol{y}_i$ follows a Gaussian distribution, that is, $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_i)$, with $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{I}_{m_i}$. Removing the constant terms, the full log likelihood of the whole data set under model (1) is

$$l_n(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\log|\sigma^2\boldsymbol{V}_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \quad (2)$$

The maximum likelihood estimates (MLE) of $\tilde{\boldsymbol{\Theta}} = (\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)$ can be obtained by maximizing (2). To simplify the computation, we write σ^2 as a function of $\boldsymbol{\beta}$ and $\boldsymbol{D}$ such that

$$\hat{\sigma}_{MLE}(\boldsymbol{\beta}, \boldsymbol{D}) = \frac{1}{N}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \quad (3)$$

where $N = \sum_{i=1}^{n} m_i$. Substituting Equation (3) into (2), the profile log likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{D}$ is given by

$$p_F(\boldsymbol{\beta}, \boldsymbol{D}) = \frac{1}{2}\sum_{i=1}^{n}\log|\boldsymbol{V}_i| + \frac{N}{2}\log\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}, \quad (4)$$

which can be minimized, if appropriately penalized with some penalty functions on $\boldsymbol{\beta}$ and $\boldsymbol{D}$, to select the fixed and random effects under the sparsity assumption.

2.1 | Penalized likelihood estimation with fixed and random effects selection

The selection of fixed and random effects can be done through the simultaneous regularization on $\boldsymbol{\beta}$ and $\boldsymbol{D}$. If any component β_j is estimated as zero, the corresponding fixed effect will be removed from the model. For the estimation and selection of random effects, we employ the Cholesky decomposition such that $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. Such a decomposition is helpful since it transforms the constrained optimization problem to an unconstrained problem and ensures the estimates of $\boldsymbol{D}$ to be positive semidefinite. According to Lindstrom and Bates (1988), it can dramatically improve the algorithm convergence. Consequently, $\boldsymbol{L}$ will be penalized and estimated to obtain the estimates of $\boldsymbol{D}$. The correspondence between the sparsity in $\boldsymbol{L}$ and that in $\boldsymbol{D}$ is shown by Li *et al.* (2018) such that for any $\boldsymbol{L}_{(k)}$ where $\boldsymbol{L}_{(k)}$ denotes the kth row of $\boldsymbol{L}$, if $\boldsymbol{L}_{(k)} = \boldsymbol{0}$, then the variance of the kth random effect and its covariance to all other random effects are zero.Thus, this random effect will be removed from the model. Given this fact, a group penalty on $\boldsymbol{L}_{(k)}$ for each k will be used so that the whole vector will be selected all or none ("all-in-or-all-out").

To achieve simultaneous selection, penalties will be imposed to both the fixed and random effects in the profile log likelihood (4) and we will get the objective function:

$$Q_n(\boldsymbol{\beta}, \boldsymbol{L}) = p_F(\boldsymbol{\beta}, \boldsymbol{L}) + \sum_{j=1}^{p_n} P_{\lambda_{1n}}(|\beta_j|) + \sum_{k=2}^{q_n} P_{\lambda_{2n}}(||\boldsymbol{L}_{(k)}||), \quad (5)$$

where $P_{\lambda_{1n}}$ and $P_{\lambda_{2n}}$ are penalty functions depending on the nonnegative tuning parameters λ_{1n} and λ_{2n}, and $||\cdot||$ represents the L_2 norm of a vector such that $||\boldsymbol{L}_{(k)}|| = \sqrt{L_{k1}^2 + L_{k2}^2 + \cdots + L_{kq_n}^2}$. The first penalty function $P_{\lambda_{1n}}(|\beta_j|)$ regulates the sparsity of $\boldsymbol{\beta}$. For the penalty on the random effects, note that it starts from $k = 2$ (i.e., the second row) since we intend to keep the random intercepts in the model. Despite the long history of use and the popularity, LASSO (Tibshirani, 1996) is known to yield estimation bias, therefore we also consider nonconcave penalty functions, such as Smoothly Clipped Absolute Deviatio (SCAD) Penalty (Fan and Li, 2001), adaptive LASSO (Zou, 2006) and minimax concave penalty (MCP) (Zhang, 2010), for more flexible choices. The second penalty function $P_{\lambda_{2n}}(||\boldsymbol{L}_{(k)}||)$ regulates the sparsity of $\boldsymbol{L}$ and thus $\boldsymbol{D}$. Specifically, we use a group penalty on $\boldsymbol{L}_{(k)}$ to shrink the norm of the whole vector to achieve "all-in-or-all-out." The paper Bakin (1999) introduces the group LASSO as an extension of LASSO for grouped variables and a computational algorithm. Like LASSO, group LASSO also produces bias in estimation and tends to overly shrink large coefficients, therefore some nonconcave penalties, such as SCAD and MCP, have been extended for group selection (Wang *et al.*, 2007; Huang *et al.*, 2012).

2.2 | Coordinate gradient descent for simultaneous selection

To estimate the high-dimensional $\boldsymbol{\beta}$ and $\boldsymbol{L}$ in the objective function (5), we use an algorithm based on the Block Coordinate Gradient Descent in Tseng and Yun (2009). The main idea of coordinate descent algorithm is to cycle through all the coordinates or prespecified blocks in turn and minimize the objective function in that single coordinate or block while keeping the others fixed

(Friedman *et al.*, 2007; Meier *et al.*, 2008; Wu and Lange, 2008). In the related work done by Schelldorfer *et al.* (2011), coordinate descent was used to select the fixed effects by penalizing the likelihood function with a single LASSO penalty. We extend their work to implement the simultaneous fixed and random effects with double penalties as shown in (5), which is nontrivial since the selection of random effects is generally more difficult.

Let $\boldsymbol{\Theta} = (\boldsymbol{\beta}, vec(\boldsymbol{L})) \in \mathbb{R}^{p_n+q_n^*}$, where $q_n^* = q_n(q_n+1)/2$ is the number of parameters in the lower triangular matrix $\boldsymbol{L}$. For a large q_n, its squared order with no doubt will cause drastic difficulty in computation. Block coordinate descent therefore becomes an important and handy tool. As the two penalty functions are separable, the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{L}$ only involves its own penalty. For the ease of presentation, we denote the jth parameter in $\boldsymbol{\Theta} = (\beta_1, \ldots, \beta_{p_n}, L_{11}, L_{21}, L_{22}, L_{31}, \ldots, L_{q_nq_n})$ as Θ_j, that is, $\Theta_j = \boldsymbol{\beta}_j$ for $j \le p_n$ and $\Theta_j = vec(\boldsymbol{L})_{j-p_n}$ for $j > p_n$. Accordingly, if $j \le p_n$ we have $\lambda_j = \lambda_{1n}$ and $P_n(\Theta_j) = P_{\lambda_j}(|\beta_j|)$; if $j > p_n$, we have $\lambda_j = \lambda_{2n}$ and $P_n(\Theta_j) = P_{\lambda_j}(\|\boldsymbol{L}_{(\lceil\sqrt{2(j-p_n)+\frac{1}{4}}-\frac{1}{2}\rceil)}\|)$.

Using the notation defined above, we rewrite and minimize the objective function (5) as

$$Q_n(\boldsymbol{\Theta}) = p_F(\boldsymbol{\Theta}) + P_n(\boldsymbol{\Theta}), \tag{6}$$

with respect to a single Θ_j while keeping the others fixed using coordinate descent. For a more stable and faster computation, we first derived the second derivative of $Q_n(\boldsymbol{\Theta})$ and then approximated it with some easily computed form h, which would be used to find the descent direction to minimize the objective function in the next step. For example, Tseng and Yun (2009) suggested using a diagonal Hessian approximation $h \approx \partial^2 Q_n(\boldsymbol{\Theta})/\partial(\Theta_j)^2$. An inexact line search was employed to ensure a decrease in the objective function.

An outline of the algorithm is given in Section 3.3 and more computational details are provided in Section 1 of the Supporting Information. In addition, we have developed an R package *splmm*, which is available at The comprehensive R Archive Network (CRAN), for the implementation of LMM with simultaneous selection of fixed and random effects.

2.3 | Asymptotic properties

While many papers have studied the rates of p_n in linear regression models (Lan, 2006; Schelldorfer *et al.*, 2011), only a few studies the rates of q_n. Bickel and Levina (2008) established the large sample properties of regularized estimation of large covariance matrices using the banding operator and proved the estimation is consistent if $\log(q_n)/n \to 0$. Lam and Fan (2009) examined the rate of convergence for large covariance matrix estimation regularized by a class of penalty functions. Bondell *et al.* (2010) studied the asymptotic normality for joint selection of both fixed and random effects by adaptive LASSO with fixed p and q. The recent work by Li *et al.* (2018) established the large sample properties for the fixed and random effects with LASSO regularization with diverging p_n and q_n. However, their proof was done for the fixed and random effects separately without considering the joint properties.

Here, we consider the joint properties for $\boldsymbol{\beta}$ and $\boldsymbol{L}$ which has never been done in the literature. The growing rate of p_n and q_n can be as high as the exponential rate of n. In addition, a general class of penalty functions are allowed as long as they satisfy the regularity conditions (B1)–(B4) provided in Section 2 of the Supporting Information.

For $\boldsymbol{\Theta} = (\boldsymbol{\beta}, vec(\boldsymbol{L})) = (\beta_1, \ldots, \beta_{p_n}, L_{11}, L_{12}, \ldots, L_{q_nq_n})$, denote its true value as $\boldsymbol{\Theta}_0$, that is, the true value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$ and the true value of $\boldsymbol{L}$ is $\boldsymbol{L}_0$. Define the sets of true nonzero coefficient indicators as $\mathcal{J} = \{j : \beta_{0j} \neq 0\}$, $\mathcal{K} = \{k : \boldsymbol{L}_{0(k)} \neq \boldsymbol{0}\}$, and $\mathcal{S} = \{s : \Theta_{0s} \neq 0\}$, with the cardinality $|\mathcal{J}| = s_n$, $|\mathcal{K}| = d_n$, and $|\mathcal{S}| = s_n + d_n$. We proved the following two large-sample theorems for $\boldsymbol{\Theta}$ that concern the rate of estimation convergence and sparsistency, respectively. The required regularity conditions are listed in Section 2 of the Supporting Information.

Theorem 1 (Rate of convergence). *Assume the observations satisfy conditions* $(A1)-(A6)$*, and the penalty functions satisfy conditions* $(B1)-(B4)$*. If* $\log(p_n)/n = O(\lambda_{1n}^2)$*,* $\log(q_n)/n = O(\lambda_{2n}^2)$*,* $s_n = O(1)$*, and* $d_n = O(1)$*, then there exists a local minimizer* $\hat{\boldsymbol{\Theta}}$ *for* $\boldsymbol{\Theta}$ *such that* $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0||^2 = O_p(\log(p_n)/n + \log(q_n)/n)$.

Theorem 1 states how the number of nonzero elements and the dimensions of design matrices affect the rate of convergence. Under the sparsity assumption on both the fixed effects and random effects, if λ_{1n} and λ_{2n} satisfy the regularity conditions (B1) and (B2), the estimator $\hat{\boldsymbol{\Theta}}$ is consistent even if p_n and q_n increase at an exponential rate of the sample size n. To prove this theorem, we showed that there exists a local minimum in $Q_n(\boldsymbol{\Theta})$ in the neighborhood of $\boldsymbol{\Theta}_0$ and hence a local minimizer $\hat{\boldsymbol{\Theta}}$. We then used Taylor expansion to approximate the difference between $Q_n(\boldsymbol{\Theta})$ and $Q_n(\boldsymbol{\Theta}_0)$ and showed that the quadratic term dominate the first-order term. The proof can be found in Section 3.1 of the Supporting Information.

Theorem 2 (Sparsistency). *Under the conditions given in the previous theorem, for any local optimizer satisfying* $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0||^2 = O_p(\log(p_n)/n + \log(q_n)/n)$*, with probability tending to 1,* $\hat{\Theta}_s = 0$ *for all* $s \in \mathcal{S}^c$.

The term "sparsistency" was first introduced by Lam and Fan (2009). It is the property that all true zero parameters are actually estimated as zero with probability tending to one. The proved sparsistency property ensures the selection consistency for both fixed effects and random effects. The proof of this theorem (Section 3.2 of the Supporting Information) is to show that the sign of the partial derivative of $Q_n(\boldsymbol{\Theta})$ with respect to each Θ_{0s}, $s \in \mathcal{S}$, depends on the sign of the Θ_{0s} with probability tending to 1. Therefore, a zero value in Θ_{0s} will result in a zero value in the partial derivative of $Q_n(\boldsymbol{\Theta})$, which is the optimum one looks for.

3 | CLUSTERING OF LONGITUDINAL DATA VIA MIXTURE OF LMM WITH SIMULTANEOUS SELECTION

In the previous section, we have introduced the LMM with simultaneous selection of fixed and random effects for exponentially growing p_n and q_n within clusters. Now let us turn to the clustering problem. We consider the clustering of longitudinal data through the mixture of such models.

3.1 | Mixture of LMM

The goal of clustering is to classify a sample of subjects into one of the G groups, where the number of clusters G is either known or unknown, based on some defined rule of similarities of their observed patterns. One natural approach is to assume that the observed data $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ follow a mixture of LMM in G groups, and we can then interpret each mixture component as one cluster. We further assume that each mixture component $g \in \{1, \ldots, G\}$ has cluster-specific parameters $\tilde{\boldsymbol{\Theta}}_g = (\boldsymbol{\beta}_g, \boldsymbol{D}_g, \sigma_g^2)$. Denote the mixing probability that any subject i belongs to cluster g is $P(\tau_i = g) \equiv \pi_g$ and $\sum_{g=1}^{G} \pi_g = 1$. Let $w_{ig} = \mathbf{1}_{\{\tau_i = g\}}$ be the binary indicator for whether subject i belongs to cluster g, and let $\boldsymbol{w}_g = (w_{1g}, \ldots, w_{ng})$ and $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G)$.

Now suppose that we also observe $\boldsymbol{w}$ in addition to $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$. We call $\{\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\}$ the "complete" data. Then, the complete log likelihood is given by

$$\begin{aligned} &\log \left\{ \prod_{i=1}^{n} \prod_{g=1}^{G} \pi_g^{w_{ig}} \phi(\boldsymbol{y}_i | \boldsymbol{X}_i \boldsymbol{\beta}_g, \sigma_g^2 \boldsymbol{V}_{ig})^{w_{ig}} \right\} \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \left\{ w_{ig} \log(\pi_g) + w_{ig} \log \phi(\boldsymbol{y}_i | \boldsymbol{X}_i \boldsymbol{\beta}_g, \sigma_g^2 \boldsymbol{V}_{ig}) \right\} \end{aligned} \tag{7}$$

where ϕ is the multivariate Gaussian function with mean $\boldsymbol{X}_i \boldsymbol{\beta}_g$ and covariance $\sigma_g^2 \boldsymbol{V}_{ig} = \sigma_g^2 (\boldsymbol{Z}_i^T \boldsymbol{D}_g \boldsymbol{Z}_i + \boldsymbol{I}_{mi})$. Like before, we write σ_g^2 as a function of $(\boldsymbol{\beta}_g, \boldsymbol{D}_g)$ and rewrite the full log likelihood as the profile likelihood. Let $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_G)$ with $\boldsymbol{\Theta}_g = (\boldsymbol{\beta}_g, \boldsymbol{L}_g, \boldsymbol{w}_g, \pi_g)$ being the parameters of interest associated with cluster g. The above complete log likelihood function can be written as

$$\begin{aligned} l_c(\boldsymbol{\Theta}) &= \sum_{g=1}^{G} \sum_{i=1}^{n} w_{ig} \log \pi_g \\ &\quad - \sum_{g=1}^{G} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_{ig} \log |\boldsymbol{V}_{ig}| \right. \\ &\quad \left. + \frac{N}{2} \log \left\{ \sum_{i=1}^{n} w_{ig} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_g)^T \boldsymbol{V}_{ig}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_g) \right\} \right\} \\ &= \sum_{g=1}^{G} \sum_{i=1}^{n} w_{ig} \log \pi_g - \sum_{g=1}^{G} p_F(\boldsymbol{\Theta}_g). \end{aligned} \tag{8}$$

If $\boldsymbol{w}$ is known, the estimate of $\boldsymbol{\Theta}_g$ can be easily obtained by minimizing the profile log likelihood $p_F(\boldsymbol{\Theta}_g)$ with respect to $\boldsymbol{\Theta}_g = (\boldsymbol{\beta}_g, \boldsymbol{D}_g)$. Unfortunately, $\boldsymbol{w}$ is usually unobserved, that is, we do not know which cluster each subject belongs to. Hence, the problem can be cast into missing data framework. A widely used approach is the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977).

3.2 | Penalized model-based clustering

Using similar notation as in Section 2 by adding subscript g to the parameters in cluster g, we propose a novel penalized model-based clustering method with fixed and random effects selection by minimizing the following objective function:

$$\begin{aligned} Q_n(\boldsymbol{\Theta}) &= - \sum_{g=1}^{G} \sum_{i=1}^{n} w_{ig} \log \pi_g + \sum_{g=1}^{G} p_F(\boldsymbol{\Theta}_g) \\ &\quad + \sum_{g=1}^{G} \left\{ \sum_{j=1}^{p_n} P_{\lambda_{g1n}}(|\beta_{gj}|) + \sum_{k=2}^{q_n} P_{\lambda_{g2n}}(||\boldsymbol{L}_{g(k)}||) \right\} \\ &= - \sum_{g=1}^{G} \sum_{i=1}^{n} w_{ig} \log \pi_g + \sum_{g=1}^{G} Q_n(\boldsymbol{\Theta}_g), \end{aligned} \tag{9}$$

where $Q_n(\boldsymbol{\Theta}_g)$ is the within-cluster objective function defined in (5).

The estimation and selection of the regression coefficients depends on the values of regularization parameters, λ_{g1n} for $\boldsymbol{\beta}_g$ and λ_{g2n} for $\boldsymbol{L}_g$. The number of clusters is

also need to be estimated from the data. Cross-validation is good criterion for high-dimensional data. However, given the complexity of the computation and number of parameters, the computational burden is way too high. We therefore choose different versions of Bayesian Information Criterion (BIC) for the determination of G and optimal values of $(\boldsymbol{\lambda}_{1n}, \boldsymbol{\lambda}_{2n})$ with $\boldsymbol{\lambda}_{1n} = (\lambda_{1,1n}, \dots, \lambda_{G,1n}) \in \mathbb{R}^G$ and $\boldsymbol{\lambda}_{2n} = (\lambda_{1,2n}, \dots, \lambda_{G,2n}) \in \mathbb{R}^G$. The basic form of BIC is given by

$$BIC = -2l_c(\boldsymbol{\Theta}) + d \log N, \tag{10}$$

where d is the total number of nonzero parameters in the model. We have also included other variations of BIC such as BICC (Wang *et al.*, 2009) and Extended Bayesian Information Criterion (Chen and Chen, 2008) in our R package cluster-splmm (2022).

3.3 | Algorithm

Given the task of performing clustering and simultaneous selection within clusters, we use a coordinate descent algorithm nested inside the EM algorithm to minimize the objective function (9). The EM algorithm estimates the parameters by alternating two steps. In the Expectation step (E-step), we update the subject-specific and overall clustering probabilities $\hat{w}_{ig}$ and $\hat{\pi}_g$ with the current estimate $\hat{\boldsymbol{\Theta}}_g$; and in the Maximization step (M-step), we update the value of $\hat{\boldsymbol{\Theta}}_g$ by minimizing the conditional expectation of the penalized log likelihood via a coordinate descent algorithm.

M-step

Treat the estimate $\hat{w}_{ig}^{(r)}$ and $\hat{\pi}_g^{(r)}$ from the rth iteration as known. The parameter estimate $\hat{\boldsymbol{\Theta}}$ can be updated by minimizing the conditional expectation of (9) given by

$$\begin{aligned} Q_n(\boldsymbol{\Theta}_g|\boldsymbol{\Theta}^{(r)}) = &-\sum_{i=1}^{n} \hat{w}_{ig}^{(r)} \log \hat{\pi}_g^{(r)} + \frac{1}{2}\sum_{i=1}^{n} \hat{w}_{ig}^{(r)} \log |\boldsymbol{V}_{ig}| \\ &+ \frac{N}{2} \log \left\{ \sum_{i=1}^{n} \hat{w}_{ig}^{(r)} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g)^T \boldsymbol{V}_{ig}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_g) \right\} \Bigg) \\ &+ \sum_{j=1}^{p_n} P_{\lambda_{g1n}}(|\beta_{g_j}|) + \sum_{k=2}^{q_n} P_{\lambda_{g2n}}(||\boldsymbol{L}_{g(k)}||) \end{aligned} \tag{11}$$

via the coordinate gradient descent algorithm such that

$$\hat{\Theta}_{gj}^{(r+1)} = \arg\min_{\Theta_{gj}} Q_n(\boldsymbol{\Theta}_g|\boldsymbol{\Theta}^{(r)}) \tag{12}$$

for $j = 1, \dots, p_n + q_n(q_n+1)/2$.

E-step

Given the updated estimate $\hat{\boldsymbol{\Theta}}^{(r+1)}$ from the M-step, we can update the values of $\hat{w}_{ig}$ and $\hat{\pi}_g$ by

$$\begin{aligned} \hat{w}_{ig}^{(r+1)} &= \frac{\hat{\pi}_g^{(r)} \phi_g\left(\boldsymbol{y}_i|\boldsymbol{X}_i, \hat{\boldsymbol{\Theta}}_g^{(r+1)}\right)}{\sum_{l=1}^{G} \hat{\pi}_l^{(r)} \phi_l\left(\boldsymbol{y}_i|\boldsymbol{X}_i, \hat{\boldsymbol{\Theta}}_l^{(r+1)}\right)}, \\ \hat{\pi}_g^{(r+1)} &= \frac{\sum_{i=1}^{n} \hat{w}_{ig}^{(r+1)}}{n}. \end{aligned} \tag{13}$$

We summarize our nested algorithm in the following pseudo code. The clusters and parameter $\boldsymbol{\Theta}$ can be initialized with some plausible methods or values. For example, the initial clusters can be assigned using clustering methods based on trajectories only (e.g., *kml* (Genolini and Falissard, 2010) or *longclust* (McNicholas and Murphy, 2010)) with estimated $\hat{w}_{ig}^{(0)}$ and $\hat{\pi}_g^{(0)}$. Then $\hat{\boldsymbol{\beta}}_g^{(0)}$ in the initial cluster g can be initialized as the ordinary Lasso estimates in linear regression models by ignoring the within-subject correlation; $\hat{\boldsymbol{L}}_g^{(0)}$ can be estimated as a diagonal matrix with empirical variance of each random effect. The algorithm is made available at GitHub with a simulated data example in our R package cluster-splmm (2022) (Algorithm 1).

4 | SIMULATIONS

4.1 | LMM with fixed and random effects selection within clusters

We first evaluated the empirical performance of our proposed LMM with fixed and random effects selection in one cluster using simulated data in this section. Data were generated for $n = 100$ subjects with $m_i = 5$ repeated measurements using model (1) with the fixed effects coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p_n}) = (1, 1.5, 0, 1.8, 2, 0, \dots, 0)$, and the random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1}, \dots, b_{iq_n}) \sim N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D} = diag(1, 2^2, 3^2, 0, \dots, 0)$. Note that the intercepts β_0 and b_{0i} were not penalized in the estimation, and p_n and q_n did not count for the intercepts in all of our numerical studies. In this setting, there were three true fixed effects and two true random effects, both excluding the intercepts. The covariates $\boldsymbol{X}_i$ were generated from a multivariate Gaussian distribution $N(0, \sigma^2\boldsymbol{\Sigma})$, where $\sigma^2 = 1$ and $\Sigma_{jk} = \rho^{|j-k|}$ (i.e., AR(1) structure) for $\rho = 0, 0.5$. We let $\boldsymbol{Z}_i$ be the first q_n columns of $\boldsymbol{X}_i$. The i.i.d. random errors were generated as $\boldsymbol{\epsilon}_i \sim N(0, I_{m_i})$. BIC was used to tune the parameters $(\lambda_{1n}, \lambda_{2n})$ using a grid search.

We examined three combinations of $(p_n, q_n) = (100, 10), (600, 10), (600, 20)$ with LASSO or SCAD penalty and repeated each setting for 100 times. The results were summarized in Table 1. Columns 5–8 report the average

ALGORITHM 1 Coordinate Gradient Descent Nested Inside EM Algorithm

Initialization: Initialize clusters $g = 1, \ldots, G$, $w_{ig}^{(0)}, \pi_g^{(0)}, \boldsymbol{\beta}_g^{(0)}, \boldsymbol{L}_g^{(0)}$;

repeat

M-step:

for $g = 1, \ldots, G$ **do**

update $\boldsymbol{\beta}_g$ and $\boldsymbol{L}_g$ via the coordinate gradient descent algorithm:

repeat

for $j = 1, \ldots, p_n + q_n^*$ **do**

Let $\Theta_{gj}^{(r)}$ be the parameter at the r-th iteration.

- Approximate the second derivative (see Section 1.1 in Supporting Information) $\frac{\partial^2}{\partial(\Theta_{gj}^{(r)})^2} Q_n(\boldsymbol{\Theta}_g^{(r)})$ with $h^{(r)}$;
- Calculate the descent direction (see Section 1.2 in Supporting Information)

$$d^{(r+1)} := \arg\min_{d \in \mathbb{R}} \left\{ p_F(\boldsymbol{\Theta}_g^{(r)}) + \frac{\partial p_F(\boldsymbol{\Theta}_g^{(r)})}{\partial \Theta_{gj}^{(r)}} \cdot d + \frac{1}{2} h^{(r)} \cdot d^2 + \lambda_{gn} P_n\big(\Theta_{gj}^{(r)} + \boldsymbol{e}_{\Theta_{gj}^{(r)}} \cdot d\big) \right\}$$

where $\boldsymbol{e}_{\Theta_{gj}^{(r)}}$ is the unit vector corresponding to $\Theta_{gj}^{(r)}$.

- Choose a stepsize $\alpha^{(r+1)} > 0$ (see Section 1.3 in Supporting Information) and update

$$\Theta_{gj}^{(r+1)} = \Theta_{gj}^{(r)} + \alpha^{(r+1)} d^{(r+1)} \boldsymbol{e}_{\Theta_{gj}^{(r)}}$$

to ensure an increase in $Q_n(\boldsymbol{\Theta}_g)$.

end

until

$\boldsymbol{\Theta}_g$ has converged

end

E-step: update w_{ig} and π_g by

$$\hat{w}_{ig}^{(r+1)} = \frac{\hat{\pi}_g^{(r)} \phi_g(\boldsymbol{y}_i | \boldsymbol{X}_i, \hat{\boldsymbol{\Theta}}_g^{(r+1)})}{\sum_{l=1}^{G} \hat{\pi}_l^{(r)} \phi_l(\boldsymbol{y}_i | \boldsymbol{X}_i, \hat{\boldsymbol{\Theta}}_l^{(r+1)})}, \quad \hat{\pi}_g^{(r+1)} = \frac{\sum_{i=1}^{n} \hat{w}_{ig}^{(r+1)}}{n};$$

until

$\boldsymbol{\Theta}$ has converged or the objective function $Q_n(\boldsymbol{\Theta})$ starts to decrease.

TABLE 1 Variable selection results of LMM in Section 4.1 for $n = 100$ subjects with $m_i = 5$ repeated measurements

(p_n, q_n)	β penalty	L penalty	ρ	β_N(true)	Bias	Model error	D_N(true)	Time (s)
(100,10)	LASSO	LASSO	0	3.25(2.93)	0.32	1.08	2.74(1.99)	3.20
			0.5	3.23(2.93)	0.31	1.18	2.46(2.00)	3.15
	SCAD	SCAD	0	3.07(2.91)	0.27	0.87	2.37(2.00)	2.61
			0.5	2.99(2.87)	0.29	0.88	2.24(2.00)	3.07
(600,10)	LASSO	LASSO	0	4.08(2.93)	0.32	1.13	2.14(1.91)	7.93
			0.5	3.93(2.88)	0.33	1.39	2.05(1.94)	7.81
	SCAD	SCAD	0	3.28(2.85)	0.3	1.12	2.06(1.97)	7.49
			0.5	3.20(2.82)	0.31	1.10	2.02(1.94)	8.04
(600,20)	LASSO	LASSO	0	3.33(2.92)	0.25	0.79	1.89(1.81)	14.79
			0.5	3.3(2.89)	0.32	1.04	1.90(1.85)	16.63
	SCAD	SCAD	0	3.29(2.92)	0.25	0.79	1.89(1.81)	14.36
			0.5	3.26(2.89)	0.31	1.03	1.86(1.83)	16.54

Note. The columns β_N(true) and D_N(true) report the average number of selected fixed and random effects, respectively, with the average number of true effects being selected in the parenthesis; bias is calculated as $\frac{\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|}{\|\boldsymbol{\beta}_k\|}$; model error is calculated as $(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$; and the last column reports the averaged computing time in seconds on a personal lap.

number of selected fixed effects (with the average number of selected true fixed effects in the parentheses), the average estimation bias of fixed effects coefficients, model error, and the average number of selected random effects (with the average number of selected true random effects in the parentheses).

It is clear from Table 1 that the three nonintercept true fixed effects and the two nonintercept true random effects can be identified almost every time using either LASSO or SCAD penalty, with only a few false positives. LMM with SCAD penalty appears to have fewer false positives than LASSO and its model error is also lower. The selection of random effects is impressive since the task is usually quite difficult. The computing time (in seconds) on a personal laptop with the selected tuning parameters was reported in the last column. The time increased when p_n and/or q_n increased, but it was still acceptable in all situations. In conclusion, our proposed LMM with simultaneous effects selection is effective at identifying true effects and removing irrelevant ones.

4.2 | Comparison with existing model-based clustering methods

To our best knowledge, there is no clustering method with variable selection available for high-dimensional longitudinal data. Therefore, we first assessed the clustering performance of our method in this section by comparing with the existing model-based clustering methods without variable selection for longitudinal data in low-dimensional settings, including *mixAK* (Komárek and Komárková, 2013) and *lcmm* (Proust-Lima *et al.*, 2015). The performance of clustering and variable selection will be evaluated in the next section.

We considered a two-cluster setting. The data were generated using the following LMM in cluster g, $g = 1, 2$:

$$\boldsymbol{y}_{ig} = \boldsymbol{X}_i \boldsymbol{\beta}_g + \boldsymbol{Z}_i \boldsymbol{b}_{ig} + \boldsymbol{\epsilon}_{ig}. \tag{14}$$

In each experiment, we set $n = 100$ with $m_i = 3$ and ran 100 replicates. The BIC criterion defined in (10) was used to determine the optimal number of clusters with the best (λ_1, λ_2) value using a grid search and for each simulated data set. In about 90% simulations, BIC could identify the correct number of clusters.

In cluster 1, we set $\boldsymbol{\beta}_1 = (2, 1, 2, 0, 0, 3, 0, \dots, 0)$ and $b_{1i} \sim N(\mathbf{0}, 0.4^2 \cdot \text{diag}(1, 1, 1, 1, 0, \dots, 0))$; in cluster 2, we set $\boldsymbol{\beta}_2 = (-2, -2, -1.5, 0, 0, 1, 0, \dots, 0)$ and $b_{2i} \sim N(\mathbf{0}, 0.8^2 \cdot \text{diag}(1, 1, 1, 1, 0, \dots, 0))$. Thus, the numbers of true fixed and random effects were both three. The random errors were generated as $\boldsymbol{\epsilon}_{ig} \sim N(0, 0.1^2 \cdot \boldsymbol{I}_{m_i})$. In both clusters, the design matrix $\boldsymbol{X}_{gi}$ included a time indicator $x_{gij1} = t$ for $t = 1, \dots, m_i$, a Bernoulli variable $x_{gij6} \sim$ Bernoulli(0.5), and an ordinal variable x_{gij7} randomly drawn from $\{1, \dots, 10\}$. The rest of $\boldsymbol{X}_{gi}$ were generated as $\boldsymbol{X}_{gij} \sim N_{(p-3)}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \sim \text{AR}(1)$ with $\Sigma_{jk} = \rho^{|j-k|}$ for $\rho = 0.5$. The random effects design matrix $\boldsymbol{Z}_{gi}$ included a random variable $Z_{gi1} \sim \text{Unif}(1, 5)$ and the first $q_n - 1$ columns of $\boldsymbol{X}_{gi}$.

The comparison results were summarized in Table 2. Our proposed method was named as "cluster-splmm" in the table. We reported the overall and cluster-specific clustering rates and computing time (in seconds). The overall clustering rate was calculated as the number of correctly

TABLE 2 Clustering rates for the comparison between mixAK, lcmm, and cluster-splmm (proposed method) using 100 simulated data sets for $n = 100$ and $m_i = 3$ in Section 4.2

	Overall	Cluster 1	Cluster 2	Time (s)
($p_n = 3, q_n = 1$, oracle)				
mixAK	0.98	0.97	0.99	6.96
lcmm	1.00	1.00	1.00	0.24
cluster-splmm	0.98	0.99	0.97	0.72
($p_n = 10, q_n = 1$)				
mixAK	0.98	0.97	0.99	6.93
lcmm	1.00	1.00	1.00	1.3
cluster-splmm	0.98	0.99	0.98	0.76
($p_n = 20, q_n = 1$)				
mixAK	0.98	0.97	1.00	7.27
lcmm	1.00	1.00	1.00	7.83
cluster-splmm	0.98	0.98	0.97	1.67
($p_n = 3, q_n = 3$, oracle)				
mixAK	NA			
lcmm	0.99	0.99	0.99	0.92
cluster-splmm	0.95	0.96	0.95	1.98
($p_n = 10, q_n = 5$)				
mixAK	NA			
lcmm	0.97	0.94	1.00	15.65
cluster-splmm	0.98	0.99	0.97	3.21
($p_n = 10, q_n = 10$)				
mixAK	NA			
lcmm	0.65	0.72	0.59	106.31
cluster-splmm	0.92	0.97	0.87	5.24
($p_n = 20, q_n = 10$)				
mixAK	NA			
lcmm	0.67	0.76	0.58	295.87
cluster-splmm	0.92	0.97	0.87	4.79
($p_n = 20, q_n = 20$)				
mixAK	NA			
lcmm	0.56	0.28	0.85	1944.77
cluster-splmm	0.91	0.97	0.86	11.23

Note. The term "oracle" means only the true effects were included in the model.

clustered subjects divided by the total number of subjects, and the cluster-specific rate was calculated as the number of correctly clustered subjects in the cluster divided by the true cluster size. In the "oracle" settings reported in the table, only the true effects were included in the analysis. The results indicate that when $q_n = 1$ and p_n is low, all three methods have similar performance. The method of *mixAK* only allows for one random effect besides the random intercepts and is not applicable when $q_n > 1$. A moderate size of q_n (e.g., $q_n = 10$) yields a much worse performance for *lcmm*. Our model had stable and consistently high clustering rates in all settings. We also observed a dramatic increase in computing time for *lcmm* when q_n gets larger while our method still took seconds.

4.3 | Clustering and variable selection performance evaluation

We evaluated the performance of our proposed method in high-dimensional settings in this section from two aspects: clustering performance and variable selection. In all settings, data were generated for 100 times using model (14) for $m_i = 3$ repeated measurements for each subject. The design matrix $\boldsymbol{X}_i$ were generated from a multivariate Gaussian distribution $N(\mathbf{1}, \sigma^2\boldsymbol{\Sigma})$ such that $\sigma^2 = 0.5^2$ and $\Sigma_{jk} = \rho^{|j-k|}$ (i.e., AR(1) structure). We let $\boldsymbol{Z}_i$ be the first q_n columns of $\boldsymbol{X}_i$.

- **Experiment 1: Two clusters with the same variables**

In cluster 1, we set $\boldsymbol{\beta}_1 = (2, 1.5, 0, 1.8, 2, 0, 0, \dots, 0)$, and $b_{2i} \sim N(\mathbf{0}, \boldsymbol{D}_1)$ with $\boldsymbol{D}_1 = \text{diag}(0.5^2, 1, 0.8^2, 0, \dots, 0)$; in cluster 2, $\boldsymbol{\beta}_2 = (-2, -1.5, 0, -1.8, -2, 0, 0, \dots, 0)$, and $b_{2i} \sim N(\mathbf{0}, \boldsymbol{D}_2)$ with $\boldsymbol{D}_2 = \text{diag}(0.5^2, 0.8^2, 1, 0, \dots, 0)$. There were three true fixed effects (i.e., X_1, X_3, X_4), and two true random effects (i.e., Z_1, Z_2) in both clusters. We examined two scenarios, which represent an unbalanced design with $p \leq n$ and a balanced design with $p > n$, respectively. The results were reported in Table 3.

- **Experiment 2: Three balanced clusters with the same variables**

In Experiment 2, the data in first two clusters were generated in the same ways as described in the Experiment 1. In addition, we added the third cluster such that $\boldsymbol{\beta}_3 = (22, 2, 0, -2, 1, 0, 0, \dots, 0)$, and $b_{3i} \sim N(\mathbf{0}, \boldsymbol{D}_3)$ with $\boldsymbol{D}_3 = \text{diag}(0.5^2, 0.5^2, 0.7^2, 0, \dots, 0)$. We used a balanced design by setting $n_1 = n_2 = n_3 = 50$ and summarized the results in Table 3.

- **Experiment 3: Three unbalanced clusters with the same or different variables**

We considered a setting with three unbalanced clusters with time effect in this section, which is similar to the real data analysis in the next section. In this setting, we let $(n_1, n_2, n_3) = (50, 150, 250)$, and $(p_n, q_n) = (50, 10)$. We generated an additional time indicator $x_{ij1} = t$, $t = 1, \dots, m_i$, for each subject, and the rest of the fixed effects as well as the random effects were generated in the same way as in Experiment 2.

TABLE 3 Simulation results of Experiments 1, 2, and 3 in Section 4.3

		$(p_n = 50, q_n = 10)$		$(p_n = 200, q_n = 10)$	
		$(n_1 = 100, n_2 = 50)$		$(n_1 = 50, n_2 = 50)$	
		$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
Experiment 1	**Overall**	**0.95**	**0.96**	**0.92**	**0.93**
Cluster 1	correct	0.96	0.97	0.91	0.90
	β_N(true)	3.06 (3.00)	3.38 (3.00)	3.22 (2.99)	4.32 (2.97)
	bias	0.18	0.37	0.35	0.41
	D_N(true)	3.20 (2.00)	3.44 (1.93)	3.49 (1.99)	3.42 (1.90)
Cluster 2	correct	0.92	0.92	0.93	0.96
	β_N(true)	3.06 (3.00)	2.82 (2.65)	2.76 (2.74)	4.38 (2.91)
	bias	0.53	0.60	0.68	0.54
	D_N(true)	3.67 (1.95)	3.76 (1.88)	3.42 (1.91)	3.05 (1.87)
Time (s)		4.83	4.88	3.82	4.56
Experiment 2		$(p_n = 50, q_n = 10)$		$(p_n = 200, q_n = 10)$	
		$(n_1 = 50, n_2 = 50, n_3 = 50)$		$(n_1 = 50, n_2 = 50, n_3 = 50)$	
		$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
	overall	0.94	0.90	0.95	0.92
Cluster 1	correct	0.85	0.80	0.91	0.86
	β_N(true)	3.57 (2.98)	4.23 (2.69)	4.31 (3.00)	2.90 (2.71)
	bias	0.35	0.67	0.23	0.59
	D_N(true)	4.31 (1.99)	4.56 (1.90)	4.36 (2.00)	4.37 (1.93)
Cluster 2	correct	0.98	0.91	0.95	0.91
	β_N(true)	3.10 (2.99)	2.79 (2.66)	2.93 (2.86)	2.57 (2.23)
	bias	0.32	0.67	0.62	0.70
	D_N(true)	2.96 (1.99)	4.56 (1.90)	3.14 (1.90)	3.66 (1.88)
Cluster 3	correct	0.99	0.98	0.99	0.98
	β_N(true)	3.39 (2.69)	5.77 (2.63)	2.66 (2.42)	5.86 (2.79)
	bias	0.05	0.06	0.21	0.05
	D_N(true)	2.63 (1.92)	2.52 (1.80)	2.86 (1.95)	2.45 (1.81)
Time (s)		5.84	5.83	6.98	5.77
Experiment 3		$(p_n = 50, q_n = 10)$		$(p_n = 50, q_n = 10)$	
		$(n_1 = 50, n_2 = 150, n_3 = 250)$		$(n_1 = 50, n_2 = 150, n_3 = 250)$	
		$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
	overall	0.98	0.96	0.98	0.97
Cluster 1	correct	0.96	0.96	0.94	0.95
	β_N(true)	3.02 (2.99)	3.38 (2.95)	2.83 (2.80)	3.02 (2.92)
	bias	0.20	0.23	0.29	0.23
	D_N(true)	3.22 (1.90)	3.28 (1.84)	3.21 (1.87)	3.26 (1.87)
Cluster 2	correct	0.96	0.95	0.97	0.96
	β_N(true)	4.22 (3.00)	5.09 (2.99)	3.24 (3.00)	4.34 (2.98)
	bias	0.34	0.27	0.28	0.23
	D_N(true)	4.00 (1.98)	3.91 (1.96)	3.96 (1.98)	3.80 (1.94)
Cluster 3	correct	0.99	0.97	0.99	0.98
	β_N(true)	2.99 (2.99)	3.98 (2.67)	2.73 (2.73)	3.75 (2.59)
	bias	0.10	0.12	0.11	0.12
	D_N(true)	3.00 (1.98)	3.10 (1.94)	3.03 (1.97)	3.06 (1.95)
Time (s)		10.33	10.61	10.46	10.81

Note. Here "overall" and "correct" are the percentage of correctly assigned subjects overall and in the corresponding cluster, respectively; β_N(true) and D_N(true) are the average number of selected fixed and random effects, respectively, with the averaged number of true effects being selected in the parenthesis; bias is calculated as $\frac{||\hat{\beta}_g - \beta_g||}{||\beta_g||}$.

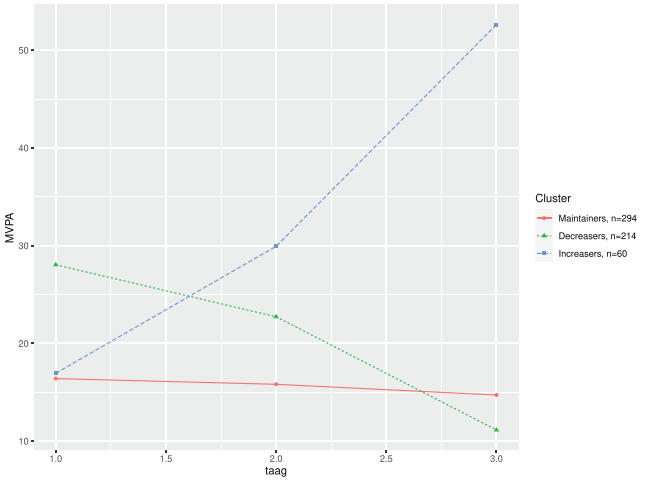

FIGURE 1 Mean MVPA trajectories of subjects in the three clusters identified by the method. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

(1) Scenario 1 (same variables): $\boldsymbol{\beta}_1 = (-5, 6, 0, 2, 1.5, 0, \dots, 0)$, $\boldsymbol{\beta}_2 = (5, -6, 0, -2, -1.5, 0, \dots, 0)$, $\boldsymbol{\beta}_3 = (-22, 1, 0, 2, 1, 0, \dots, 0)$;
(2) Scenario 2 (different variables): $\boldsymbol{\beta}_1 = (-5, 6, 2, 1.5, 0, 0, \dots, 0)$, $\boldsymbol{\beta}_2 = (5, -6, 0, 0, -2, -1.5, 0, \dots, 0)$, $\boldsymbol{\beta}_3 = (-22, 1, 0, 0, 0, 0, 2, 1, 0, \dots, 0)$.

Scenario 2 was actually designed to mimic the real TAAG data in Section 5. The three clusters of size 50, 150, 250 correspond to the Increasers (the smallest cluster), the Decreasers (the medium cluster), and Maintainers (the largest cluster), respectively. The sample response curves of the three clusters in the Supporting Information Figure 1 show similar trends over time as the TAAG example. In addition, the variables in clusters are cluster-specific, the same as TAAG. The results summarized in Table 3 are satisfying.

In summary, the results in all three experiments were very promising. As seen, our method had very satisfactory performance in terms of both clustering and variable selection in all the simulated settings. It was able to identify the true fixed and random effects nearly perfectly with a very low false positive rate. The increase in cluster sizes improved the clustering rate, as expected. The increase in p_n did not impact the clustering and variable selection much and the results were still very good with large p_n.

5 | APPLICATION TO TAAG

The proposed clustering method was applied to the TAAG (Young *et al.*, 2018) mentioned in Section 1. The previous papers (Young *et al.*, 2018; LaLonde *et al.*, 2019; Young *et al.*, 2019) employed a two-step procedure to select the variables and perform clustering separately for a cohort of Baltimore girls at age 14 ($n = 730$) in 2006, at age 17 ($n = 589$) in 2009, and at age 23 ($n = 460$) in 2015. We ran the model with 35 covariates at individual-, social-, and neighborhood levels on the moderate-to-vigorous physical activity (MVPA) minutes. All the 35 variables were used for the fixed effects selection and the 13 neighborhood-level variables were used for random effects selection. The data file is available upon request.

Removing missing records from the data set, there were 1561 records from 568 participants. Based on the BIC criterion (10), we identified three clusters of girls with similar longitudinal physical activity patterns (Figure 1). This figure appears in color in the electronic version of this article, and any mention of color refers to that version. The largest cluster ("maintainers") contained 294 participants whose daily MVPA was consistently low over time (age 14: 16.41 ± 6.34 min/day, age 17: 15.82 ± 6.81 min/day, age 23: 14.72 ± 8.82 min/day); the second cluster ("decreasers") contained 214 participants with decreasing MVPA over time (age 14: 28.05 ± 12.27 min/day, age 17: 22.73 ± 13.59 min/day, age 23: 8.99 ± 9.39 min/day); and the smallest cluster ("increasers") had 60 participants with daily MVPA increasing over time (age 14: 16.95 ± 7.57 min/day, age 17: 29.94 ± 13.46 min/day, age 23: 52.63 ± 22.67 min/day). The MVPA patterns in the three clusters were associated with different variables, as displayed in Table 4.

The variables identified for each cluster were summarized in Table 4. The intercept for each cluster was also listed since it represents the mean MVPA at baseline. The variable *TAAG*, which is a time indicator for wave 1, 2, and 3, was selected for all clusters as an important indicator of changing in MVPA over time in all three groups.

In the "maintainer" group, the 294 girls started with a mean 12.62 min of MVPA per day at baseline, which was significantly lower than the daily 60 min or more of MVPA, recommended in the Physical Activity Guidelines for Americans (2nd edition) by the US Department of Health and Human Services (Piercy *et al.*, 2018). The coefficient −1.95 for *TAAG* indicates a further decrease in MVPA over time. In addition, six fixed effects and five random effects were selected into the model of this cluster. Greater *friend support* and more *parks within 1 mile of participant's homes*, lower *BMI*, better *self-management strategy*, lower *self-efficacy*, lower *perceived barriers* were associated with higher MVPA within this cluster. Meanwhile, population heterogeneity seems to exist in the neighborhood factors, including the *sidewalks*, *walking/biking trails*, *safety to walk/jog*, *walkers/bikers*, and *kids playing in neighborhood*.

In the "decreaser" group, the 214 girls started with a mean 38.9 min of daily MVPA, which was higher than the increaser group although still lower than the daily physical

TABLE 4 Fixed effects and random effects selection results in each cluster of the model with neighborhood-level random effects

	Maintainers $n = 294$ (**51.8**%)	**Decreasers** $n = 214$ (**37.7**%)	**Increasers** $n = 60$ (**10.5**%)
	Fixed effects		
	$N_\beta = 7$	$N_\beta = 7$	$N_\beta = 7$
Intercept	12.62	38.90	18.92
TAAG	−1.95	−17.06	29.44
Sidewalks in neighborhood			−2.75
Kids playing in neighborhood			2.75
BMI	−0.00086		−1.72
Self-management strategy	0.14	0.82	0.07
Self-efficacy	−0.026	0.48	
Enjoyment			0.69
Perceived barriers	−0.025		
Outcome-expectancy (belief)		−0.68	
Outcome-expectancy (importance)		−0.014	
Social support		−2.29	3.86
Friend support	0.14		
Family support		−0.29	
Number of parks within 1 mile	0.51		
	Random effects		
	$N_D = 5$	$N_D = 8$	$N_D = 1$
Sidewalks in neighborhood	1.68×10^{-4}	1.14×10^{-3}	
Trails in neighborhood	4.47×10^{-4}		
Safe to walk/jog in neighborhood	1.85×10^{-4}	8.13×10^{-4}	1.11×10^{-4}
Walkers/Bikers in neighborhood	1.05×10^{-4}	6.18×10^{-3}	
Traffic in neighborhood		2.87×10^{-4}	
Crime in neighborhood		3.17×10^{-4}	
Kids playing in neighborhood	1.56×10^{-4}	1.55×10^{-3}	
Interesting things in neighborhood		2.48×10^{-4}	
Neighborhood street well lit		5.98×10^{-3}	

activity guideline, but had a significant decrease of 17.06 min in MVPA after each wave. Better *self-management strategy*, higher *self-efficacy*, lower *outcome-expectancy*, less *social support*, and less *family support* were associated with higher MVPA in this cluster. Also, *number of sidewalks*, *safety to walk/jog*, *walkers/bikers*, *traffic*, *crime*, *kids playing*, *interesting things to look in neighborhoods*, and *how well neighborhoods were lit* makes a variation on physical activity level among participants in this group.

The "increaser" group, although only composes 10.5% of the study sample, is the most interesting group. The 60 girls started with a mean MVPA of 18.92 min per day, their MVPA increased by 29.44 min after each wave on average and exceeded the guideline at the third wave. Fewer *sidewalks in neighborhood*, more *kids playing in neighborhood*, lower *BMI*, better *self-management strategy*, more *enjoyment of physical activity*, and greater *social support* were associated with higher MVPA within this cluster. In addition, *safety to walk/jog in neighborhood* contributes to the variation on individual's physical activity level in this group.

Our findings provide potentially useful information on how to encourage physical activity for girls from adolescence to early adulthood. Compared to previous analysis, our results identified distinct patterns of MVPA trajectories as well as cluster-specific predictors. It implies that for girls in different clusters, one might want to consider different strategies to improve their MVPA. It is also of great interest to further study the increaser group and understand their characteristics in order to pump other girls into this group. The individual MVPA trajectories can be found in Figure 2 in the online Supporting Information.

6 | CONCLUSION

In this paper, we proposed a novel clustering method for longitudinal data in ultra high-dimensions via a mixture of LMM with double penalties from a general class. We proved that the proposed method works for the situation with the dimensions of fixed and random effects growing in an exponential rate of the sample size. The performance of the method was evaluated from different aspects using simulated data. It was applied to the real problem which motivated this study.

Our method has made two major contributions. First, we developed an LMM for simultaneous fixed and random effects selection and studied the joint asymptotic properties for exponentially growing p_n and q_n within clusters. It extended the previous work by Li *et al.* (2018) by closing the gap in their proof and also generalized penalty function from LASSO to a more general class including nonconcave functions. We developed an R package *splmm* using an efficient and stable computing algorithm for the proposed LMM. It is the first and so far the only R package that can perform fixed effects and random effects selection for repeated measures.

Second, this is the first clustering method for high-dimensional longitudinal data with diverging p and q, which has great potential in biomedical and social studies. It allows missing data, contrasted with the approaches using response trajectories only such as *kml* (Genolini and Falissard, 2010) and *longclust* (McNicholas and Murphy, 2010). Our method models the relationship between the response trajectories and covariates and characterizes the clusters with different sets of selected covariates.

The possible directions for future research based on the current work include: (1) extension to generalized LMM to incorporate other common types of outcomes such as binary, ordinal, and counts; (2) modeling of multiple outcomes, similar to the paper of Komárek and Komárková (2013). This extension will allow us to model the MVPA trajectories and longitudinal sedentary behaviors together in the TAAG project; (3) one can consider other distributions rather than Gaussian, for example, t distribution is a good choice for data with heavy tail.

ACKNOWLEDGMENTS

Wu's work was partly supported by grants from the National Institutes of Health NIH/R01HL119058 and National Science Foundation NSF-CCF-1934962.

DATA AVAILABILITY STATEMENT

The Trial of Activity in Adolescent Girls (TAAG) data are available from Dr Deborah Rohm Young (Deborah.R.Young@kp.org) at Kaiser Permanente Southern California with the permission of Kaiser Permanente and University of Maryland.

ORCID

Luoying Yang https://orcid.org/0000-0003-1653-8350
Tong Tong Wu https://orcid.org/0000-0002-1175-9923

REFERENCES

Arribas-Gil, A., De la Cruz, R., Lebarbier, E. and Meza, C. (2015) Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. *Biometrics*, 71, 333–343.

Bakin, S. (1999) Adaptive regression and model selection in data mining problems. Thesis, Australian National University.

Bickel, P.J. and Levina, E. (2008) Regularized estimation of large covariance matrices. *Annals of Statistics*, 36, 199–227.

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.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

cluster-splmm. (2022) R package for clustering of high-dimensional longitudinal data based on simultaneously penalized linear mixed-effects models. Available at: https://github.com/lyang19/cluster-splmm. Accessed March 29, 2022.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39, 1–38.

Du, Y., Khalili, A., Nešlehová, J.G. and Steele, R.J. (2013) Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models. *Canadian Journal of Statistics*, 41, 596–616.

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

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302–332.

Genolini, C. and Falissard, B. (2010) KmL: k-Means for longitudinal data. *Computational Statistics*, 25, 317–328.

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

Komárek, A. and Komárková, L. (2013) Clustering for multivariate continuous and discrete longitudinal data. *Annals of Applied Statistics*, 7, 177–200.

LaLonde, A., Love, T., Young, D. and Wu, T.T. (2019) Clustering adolescent female physical activity levels with an infinite mixture model on random effects. Manuscript.

Lam, C. and Fan, J. (2009) Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37, 4254–4278.

Lan, L. (2006) Variable selection in linear mixed model for longitudinal data. PhD thesis, Department of Statistics, North Carolina State University.

Li, Y., Wang, S., Song, P.X.-K., Wang, N., Zhou, L. and Zhu, J. (2018) Doubly regularized estimation and selection in linear

mixed-effects models for high-dimensional longitudinal data. *Statistics and Its Interface*, 11, 721–737.

Lindstrom, M.J. and Bates, D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014–1022.

McNicholas, P.D. and Murphy, T.B. (2010) Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38, 153–168.

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 53–71.

Müller, S., Scealy, J.L. and Welsh, A.H. (2013) Model selection in linear mixed models. *Statistical Science*, 28, 135–167.

Piercy, K.L., Troiano, R.P., Ballard, R.M., Carlson, S.A., Fulton, J.E., Galuska, D.A. et al. (2018) The physical activity guidelines for Americans. *Journal of the American Medical Association*, 320, 2020–2028.

Proust-Lima, C., Philipps, V. and Liquet, B. (2015) Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *arXiv preprint arXiv:1503.00890* .

Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011) Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38, 197–214.

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

Tseng, P. and Yun, S. (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117, 387–423.

Wang, H., Li, B. and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 671–683.

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

Wu, T.T. and Lange, K. (2008) Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2, 224–244.

Young, D.R., Cohen, D., Koebnick, C., Mohan, Y., Saksvig, B.I., Sidell, M. et al. (2018) Longitudinal associations of physical activity among females from adolescence to young adulthood. *Journal of Adolescent Health*, 63, 466–473.

Young, D.R., Sidell, M.A., Koebnick, C., Saksvig, B.I., Mohan, Y., Cohen, D.A. et al. (2019) Longitudinal sedentary time among females aged 17 to 23 years. *American Journal of Preventive Medicine*, 56, 540–547.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38, 894–942.

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

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2.2, 2.3, 3.3, 4.3, 5 are available with this paper at the Biometrics website on Wiley Online Library. We have developed two R packages in this paper. The R package *splmm*, which is available at CRAN (https://cran.r-project.org/web/packages/splmm/index.html), is for the implementation of LMM with simultaneous selection of fixed and random effects for one homogeneous population. The R package *cluster-splmm*, available at GitHub with a simulated data example (https://github.com/lyang19/cluster-splmm), performs clustering of high-dimensional longitudinal data based on simultaneously penalized LMM. The code of *cluster-splmm* (the zip file) is posted online with this paper.

How to cite this article: Yang, L., and Wu, T.T. (2023) Model-based clustering of high-dimensional longitudinal data via regularization. *Biometrics*. 79:761–774. https://doi.org/10.1111/biom.13672