Statistics and Its Interface Volume 11 (2018) 721–737

Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data

Yun Li[*,†,§], Sijian Wang[‡,§], Peter X.-K. Song[¶,‖], Naisyin Wang[*], Ling Zhou[¶], and Ji Zhu[†]

The linear mixed-effects model (LMM) is widely used in the analysis of clustered or longitudinal data. This paper aims to address analytic challenges arising from estimation and selection in the application of the LMM to high-dimensional longitudinal data. We develop a doubly regularized approach in the LMM to simultaneously select fixed and random effects. On the theoretical front, we establish large sample properties for the proposed method under the high-dimensional setting, allowing both numbers of fixed effects and random effects to be much larger than the sample size. We present new regularity conditions for the diverging rates, under which the proposed method achieves both estimation and selection consistency. In addition, we propose a new algorithm that solves the related optimization problem effectively so that its computational cost is comparable with that of the Newton-Raphson algorithm for maximum likelihood estimator in the LMM. Through simulation studies we assess performances of the proposed regularized LMM in both aspects of variable selection and estimation. We also illustrate the proposed method by two data analysis examples.

AMS 2000 subject classifications: Primary 62J05, 62J07; secondary 62F12.

Keywords and phrases: Diverging rate, Regularization, Random-effects selection, Variable selection.

1. INTRODUCTION

In this paper, we consider estimation and variable selection in the analysis of high-dimensional clustered or longitudinal data. Such data are becoming increasingly popular in many subject-matter areas, especially in life sciences, social sciences, and medical and health sciences. Linear mixed-effects models (LMM; Laird and Ware, 1982), being one of the most widely used models in the analysis of repeated measurements, are greatly challenged by data with the number of covariates diverging to infinity along with the sample size. This paper focuses on the development of a novel and effective variable selection procedure in the LMM that extracts important predictors from a vast pool of candidates.

When the number of predictors is large, a variable selection method enables us to achieve parsimonious models that include most important predictors. A parsimonious model is easier to interpret and implement in practice. Information criteria, such as AIC (Akaike, 1973), BIC (Schwarz, 1978) and conditional AIC (Vaida and Blanchard, 2005; Liang, Wu and Zou, 2008; Greven and Kneib, 2010) are among the most popular model selection tools in the LMM. However, these selection procedures are known to be inefficient or even infeasible when the number of possible mixed-effects models is large.

Extending from the recent variable selection literature and assuming a fixed structure of random effects, Lan (2006) and Schelldorfer et al. (2011) developed penalized likelihood-based approaches to selecting fixed effects. However, neither of the work considered the selection of random effects. In practice, the selection of random effects is equally important to the selection of fixed effects, as the configuration of the random effects component not only determines the marginal covariance structure of the correlated data, but also steers the interpretation of subject-specific effects of covariates. Though a misspecified covariance structure may not affect the consistency of fixed effects estimators (e.g. Verbeke and Lesaffre, 1997), it does affect the estimates of random effects and the asymptotic covariance matrix. For example, Lange and Laird (1989) showed that an under-specified random-effects component would lead to biased estimation for the variance of fixed effects. On the other hand, an over-parameterized covariance structure may lead to unstable algorithms and loss of estimation efficiency. Thus, an appropriate composition of the random-effects component is critically important in the LMM.

In the situation where both numbers of fixed and random effects are fixed as constant, there are several works that have contributed to the selection of the random effects component in the LMM. Stram and Lee (1994) discussed the

[*] The research of Y. Li and N. Wang is partially supported by an NCI grant CA74552.

[†] The research of Y. Li and J. Zhu is partially supported by NIH R01-GM-096194 and NSF DMS-1407698.

[‡] The research of S. Wang is partially supported by NIH R01-HG-007377.

[§] Y. Li and S. Wang are co-first author of this paper.

[¶] The research of P. Song and L. Zhou is supported in part by NIH R01-ES-024732 and NSF DMS-1513595.

[‖] Corresponding author.

asymptotic behavior of a likelihood ratio test for nonzero random effects variances. For the special case where one is interested in whether any random effects should be included, Commenges and Jacqmin-Gadda (1997), Lin (1997) and Hall and Praestgaard (2001) proposed score tests. Foster et al. (2009) proposed a LASSO random effects models with no fixed effects, where random effects were assumed to follow a double exponential distribution, and the Laplace approximation was used to obtain the marginal likelihood function. Albert and Chib (1997) and Chen and Dunson (2003) also tackled the problem of random-effects selection using Bayesian approaches. Sinharay and Stern (2001) used Bayes factors to compare variance components in the LMM. Similar to Foster et al. (2009), these papers did not consider the fixed effects selection.

Several recent papers have investigated simultaneous selection of fixed and random effects in the LMM. Jiang et al. (2008) developed a "fence" method for variable selection in a general mixed-effects model. Bondell et al. (2010) developed a penalized joint log-likelihood approach with an adaptive penalty. Two Bayesian approaches were proposed by Kinney and Dunson (2007) and Ibrahim et al. (2011), respectively. The former considered a prior with mass at zero, while the latter also considered a regularized likelihood-based method. Fan and Li (2012) recently proposed a two-step method for selecting both the fixed and random effects. In the first step, the method selects random effects using group-lasso via regularizing the mode of the posterior distribution of random effects. In the second step, the method focuses on fixed-effects selection with given random effects, and the authors have concentrated on the scenario that the number of fixed effects is smaller than the sample size. Ahn et al. (2012) proposes a moment-based method for random effects selection in linear mixed models. The theoretical results with fixed p are established. Lai et al. (2012) considered fixed and random effects selection in nonparametric additive mixed models. Yang (2012) proposed Bayesian variable selection for logistic mixed model with nonparametric random effects. Du et al. (2013) considered the fixed and random effects selection in a finite mixture of linear mixed-effects models. Lin et al. (2013) proposed a two-stage model selection procedure for the linear mixed-effects models. The procedure consists of two steps: First, penalized restricted log-likelihood is used to select the random effects. Next, the penalized log-likelihood is used to select the fixed effects. The theoretical results with fixed p are established. Pan and Huang (2014) considered the selection of random effects with a fixed dimension with no theoretical justification. Based on a reparametrization of the covariance matrix of random effects by a modified Cholesky decomposition, they added a LASSO penalty function on the variances of the random effects, resulting in a non-convex constrained optimization that was numerically of great difficulty.

In this paper, we consider a new regularization approach that performs estimation and variable selection simultaneously for both fixed and random effects. Our development differs from previous methods in two aspects. First, our method allows both numbers of fixed and random effects to diverge to infinity as the sample size increases, while previous methods have restricted their attention to finite dimensions of fixed and/or random effects. Furthermore, in the scenario we consider, large-sample properties have not been studied previously. The reach of estimation and selection consistency requires a delicate control of signal-to-noise ratio in the model, which involves an inter-play between the strength of signals (fixed effects) and the amount of variations (random effects and random errors). One of our new contributions is to establish a set of regularity conditions concerning the diverging rates for tuning parameters, under which the proposed regularization method achieves both estimation and selection sparsistency. Second, our method is implemented by an efficient optimization algorithm, whereas previous methods are based on the EM or Monte Carlo algorithm which is known to be computationally intensive, particularly when the dimensions of fixed and random effects are large. In contrast, our new optimization algorithm is as effective as the Newton-Raphson algorithm for computing the maximum likelihood estimator (MLE) in the LMM. Finally, using the Cholesky decomposition of the selected covariance matrix of random effects, we ensure it to be positive-definite. Similar techniques have been considered in the literature including, for example, Pourahmadi (1999, 2000), Pan and MacKenzie (2003) and Ye and Pan (2006). Further, the resulting random effects selection is invariant with respect to the ordering of predictors appearing in the Cholesky decomposition.

The rest of the paper is organized as follows. In Section 2, we introduce our new method: the doubly regularized MLE. In Section 3, we discuss a new algorithm to carry out the related optimization. In Section 4, we study the asymptotic behavior of the proposed method under some mild regularity conditions, including the classical assumption of restricted eigenvalues for covariates associated with fixed effects and a regularity condition of similar flavor for covariates associated with random effects. In Sections 5 and 6, we demonstrate the use of our method via simulations and two data examples, respectively. We conclude the paper with Section 7. All technical proofs are given in the Appendix section.

2. METHODOLOGY

2.1 Model

Suppose there are n subjects under study, and there are m_i repeated observations recorded for subject i, $i = 1, \ldots, n$; throughout the paper, we consider bounded m_i. There are p_n covariates associated with the fixed effects, denoted by $X_1, \ldots, X_{p_n}$, while q_n covariates associated with the random effects, denoted by $Z_1, \ldots, Z_{q_n}$. Usually, the q_n random-effects covariates are a subset of the p_n fixed-effects covariates. In this paper, we allow both $p_n \to \infty$ and $q_n \to \infty$ as $n \to \infty$. For the ease of exposition, in the following presentation we use simple notation of p and q unless the subscript n is necessary. For subject i at observation j, let Y_{ij} denote the response variable, $\mathbf{x}_{ij}$ be the vector of p predictors in the fixed effects component, and $\mathbf{z}_{ij}$ be the vector of q predictors in the random-effects component. The linear mixed-effects model is then written as follows:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}, \tag{1}$$

where errors ϵ_{ij}'s are assumed $i.i.d.$ $N(0, \sigma^2)$, and the random effects, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^T$, are $i.i.d.$ according to a multivariate normal distribution $\mathrm{MVN}_q(0, \sigma^2\mathbf{D})$. Denote the set of parameters to be estimated by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$. Without loss of generality, we assume each covariate X_j or Z_k is standardized to have zero mean and unit Euclidean norm. Thus, the fixed intercept can be removed from the model. However, we will always keep the random intercept, denoted by b_1, in the model to account for the minimal level of within-subject correlation.

For notational simplicity, we rewrite (1) in a matrix format:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \epsilon_i,$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T, \mathbf{X}_i^T = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}), \mathbf{Z}_i^T = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{im_i})$, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^T$. The first two moments of $\mathbf{Y}_i$ are then given by

$$\begin{aligned} E(\mathbf{Y}_i) &= \mathbf{X}_i \boldsymbol{\beta}, \\ Var(\mathbf{Y}_i) &= \sigma^2 \Big(\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{I}_{m_i} \Big). \end{aligned}$$

Clearly, the component of fixed effects, i.e. $\mathbf{X}_i$, affects the mean model, and the component of random effects, i.e. $\mathbf{Z}_i$, affects the covariance structure. Our goal is to jointly select and estimate both fixed and random effects.

2.2 Maximum likelihood estimation

Our variable selection method is built upon a modified maximum likelihood (ML) estimation in the LMM (e.g., Laird and Ware, 1982; Jennrich and Schluchter, 1986; Lindstrom and Bates, 1988), which is detailed as follows.

Under model (1), the marginal distribution of $\mathbf{Y}_i$ is given by

$$\mathbf{Y}_i \sim \mathrm{MVN}_{m_i}(\mathbf{X}_i \boldsymbol{\beta}, \sigma^2 \mathbf{V}_i),$$

where $\mathbf{V}_i = \mathbf{I}_{m_i} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T$. Subject to a constant, the (full) log-likelihood for the data is

$$\begin{aligned} \ell_{nF}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) = &-\frac{1}{2} \sum_{i=1}^n \log \Big| \sigma^2 \mathbf{V}_i \Big| \\ &- \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}), \end{aligned} \tag{2}$$

and the ML estimates of parameters $\boldsymbol{\beta}, \mathbf{D}$ and σ^2 can be obtained by maximizing the log-likelihood function (2). Note that if $\mathbf{D}$ were known, the MLE for $\boldsymbol{\beta}$ would be given by

$$\hat{\boldsymbol{\beta}}(\mathbf{D}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \Big(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \Big)^T \mathbf{V}_i^{-1} \Big(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \Big). \tag{3}$$

One well-known criticism on the ML estimation is that for the variance components (i.e. $\mathbf{D}$), there is a downward finite-sample bias due to the fact that the ML method does not take into account the loss in degrees of freedom from the estimation of $\boldsymbol{\beta}$. The restricted maximum likelihood estimate (REML) corrects for this bias by deriving estimates of the variance components as the maximizers of the log-likelihood based on $N - p$ linearly independent error contrasts, where N is the total number of observations from all individuals, i.e., $N = \sum_{i=1}^n m_i$. This restricted log-likelihood, according to Harville (1974), is

$$\begin{aligned} &\ell_R(\mathbf{D}, \sigma^2) \\ &\quad = -\frac{1}{2} \sum_{i=1}^n \log \Big| \sigma^2 \mathbf{V}_i \Big| - \frac{1}{2} \log \Big| \sigma^{-2} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \Big| \\ &\quad - \frac{1}{2\sigma^2} \sum_{i=1}^n \Big\{ \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\mathbf{D}) \Big\}^T \mathbf{V}_i^{-1} \Big\{ \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\mathbf{D}) \Big\}, \end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\beta}}(\mathbf{D}))$ is given by (3).

Joining the estimator (3) and the REML (4), we may write a modified log-likelihood as

$$\begin{aligned} \ell_{nM}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) = &-\frac{1}{2} \sum_{i=1}^n \log \Big| \sigma^2 \mathbf{V}_i \Big| \\ &- \frac{1}{2} \log \Big| \sigma^{-2} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \Big| \\ &- \frac{1}{2\sigma^2} \sum_{i=1}^n \Big(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \Big)^T \mathbf{V}_i^{-1} \Big(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta} \Big), \end{aligned} \tag{5}$$

provided that all the determinants in (5) are positive. The estimates of $\boldsymbol{\beta}$ and $\mathbf{D}$ can then be obtained by jointly maximizing (5).

2.3 Doubly regularized likelihood estimation

The selection of fixed- and random-effects (SOFARE) can be realized through the selection of nonzero elements in $\boldsymbol{\beta}$ and $\mathbf{D}$. If $\beta_j = 0$, the corresponding predictor X_j (a fixed effect) will be excluded from the model. If a diagonal element $D_{kk} = 0$, which means the variance of the kth random effect is zero, then the random effect b_k will be removed from the model. In order to obtain the desired sparsity in the final estimates, we propose to regularize the estimation of both $\boldsymbol{\beta}$ and $\mathbf{D}$ simultaneously. Specifically, we consider the following two cases.

(I) **The $N < p$ case:** When the total number of observations N is smaller than the total number of candidate fixed effects p, the modified log-likelihood (5) is not applicable as it relies on $N - p$ (< 0) linearly independent error contrasts. In this case, we propose to use double regularization on the log-likelihood (2), that is, we wish to find $\boldsymbol{\beta}$, $\mathbf{D}$, and σ^2 that maximize

$$Q_n(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \ell_{nF}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) - \lambda_1 J_1(\boldsymbol{\beta}) - \lambda_2 J_2(\mathbf{D}),$$

where $\ell_{nF}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$ is given in (2).

(II) **The $N > p$ case:** When the total number of observations N is greater than the total number of candidate fixed effects p, as discussed in Section 2.2, to correct for the bias in the variance component estimation, we propose to use the modified log-likelihood (5) to carry out regularized estimation; that is, we wish to find $\boldsymbol{\beta}$, $\mathbf{D}$, and σ^2 that maximize

$$Q_n(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \ell_{nM}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) - \lambda_1 J_1(\boldsymbol{\beta} - \lambda_2 J_2(\mathbf{D}),$$

where $\ell_{nM}(\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$ is from (5).

Note that in the above formulations, λ_1 and λ_2 are two nonnegative tuning parameters. The first penalty function $J_1(\boldsymbol{\beta})$ controls the sparsity of final estimation of $\boldsymbol{\beta}$, and hence navigates the selection of fixed effects. The second penalty function $J_2(\mathbf{D})$ dictates the sparsity of the final estimation of $\mathbf{D}$, and hence rules the selection of random effects. The reason that we propose two versions of regularized objective functions is to take the advantage that the modified likelihood (5) has better finite-sample performances, which however becomes ill-defined in the case of large p small N due to singular covariance matrices in (5).

Specifically, we adopt the L_1-norm penalty for $J_1(\boldsymbol{\beta})$ (Tibshirani, 1996),

$$J_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

It is well-known that due to the singularity of $|\beta_j|$ at 0, some estimates of $\hat{\beta}_j, j = 1, \ldots, p$ will be exactly zero.

For the random-effects selection, to ensure the positive definiteness of the estimated $\mathbf{D}$, we invoke the Cholesky decomposition, i.e., $\mathbf{D} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements. This decomposition converts a constrained optimization into an unconstrained problem, and the resulting computation is stable and fast. Consequently, the selection procedure will target on $\mathbf{L}$, rather than on $\mathbf{D}$. The relation between the sparsity of $\mathbf{D}$ and the sparsity of $\mathbf{L}$ is given by the following Lemma.

Lemma 2.1. *Denote $\mathbf{L} = (\mathbf{L}_{(1)}^T, \ldots, \mathbf{L}_{(q)}^T)^T$, where $\mathbf{L}_{(k)}$ is the kth row of $\mathbf{L}$. Then for any given k, we have*

$$\mathbf{L}_{(k)} = \mathbf{0} \Longleftrightarrow D_{kk} = 0 \textit{ and } D_{kj} = D_{jk} = 0, \forall j.$$

The proof is straightforward and is omitted. Lemma 2.1 indicates that if the vector $\mathbf{L}_{(k)} = \mathbf{0}$, then the diagonal element D_{kk}, known as the variance of the random effect b_k, is zero. Furthermore, for any $j \neq k$, the off-diagonal elements D_{kj} are also equal to 0, which implies that the covariances between b_k and all the other random effects are estimated as zero. Thus, the random effect b_k can be excluded from the model. The above observation motivates us to shrink the entire vector $\mathbf{L}_{(k)}$ towards a zero vector. For this, we adopt the L_2-norm penalty (Yuan and Lin, 2006) for $J_2(\mathbf{D})$,

$$J_2(\mathbf{L}) = \sum_{k=2}^{q} \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}. \tag{7}$$

Note that the summation starts from $k = 2$, for we intend to keep the random intercept in the model, which generates a minimal within-cluster correlation. Like the L_1-norm penalty, the L_2-norm penalty is singular at the point $\mathbf{L}_{(k)} = \mathbf{0}$, which encourages $\mathbf{L}_{(k)}$ to be estimated as an exact zero vector.

Furthermore, noting that $D_{kk} = L_{k1}^2 + \cdots + L_{kq}^2$, we may rewrite the J_2 penalty as $J_2(\mathbf{D}) = \sum_{k=2}^{q} \sqrt{D_{kk}}$. The fact that the value of $J_2(\mathbf{D})$ remains unchanged regardless the ordering of D_{kk} (or random effects) appearing in the model implies that the estimation for $\mathbf{D}$ is invariant with respect to the ordering of random effects in the Cholesky decomposition.

3. ALGORITHM

We aim to estimate $\boldsymbol{\beta}$ and $\mathbf{L}$ ($\mathbf{D} = \mathbf{L}\mathbf{L}^T$) by maximizing the following doubly regularized objective function:

$$Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2) = \ell_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\mathbf{L}_{(k)}\|_2,$$

where $\ell_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ takes (2) or (5) depending on whether $N < p$ or $N > p$, and $\|\mathbf{L}_{(k)}\|_2 = \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}$.

To simplify the computation, following Lindstrom and Bates (1988), we estimate σ^2 by, if $N > p$

$$\hat{\sigma}^2(\boldsymbol{\beta}, \mathbf{L}) = \frac{1}{N - p} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{8}$$

and if $N < p$,

$$\hat{\sigma}^2(\boldsymbol{\beta}, \mathbf{L}) = \frac{1}{N} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{9}$$

Substituting the expression (8) into $\ell_{nM}(\beta, \mathbf{L}, \sigma^2)$ or the expression (9) into $\ell_{nF}(\beta, \mathbf{L}, \sigma^2)$, we obtain the doubly regularized profile log-likelihood of the form:

$$Q_R(\boldsymbol{\beta}, \mathbf{L}) = P_R(\boldsymbol{\beta}, \mathbf{L}) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\mathbf{L}_{(k)}\|_2, \tag{10}$$

where in the case of $N > p$,

$$P_R(\boldsymbol{\beta}, \mathbf{L}) = -\frac{1}{2} \sum_{i=1}^{n} \log\left|\mathbf{V}_i\right| - \frac{1}{2} \log\left|\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i\right|$$

$$-\frac{N-p}{2}\log\left\{\sum_{i=1}^{n}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)^T\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)\right\},$$

or in the case of $N<p$,

$$P_R(\boldsymbol{\beta},\mathbf{L}) = -\frac{1}{2}\sum_{i=1}^{n}\log\Big|\mathbf{V}_i\Big| \\ -\frac{N}{2}\log\left\{\sum_{i=1}^{n}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)^T\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)\right\}.$$

The estimation of $\boldsymbol{\beta}$ and $\mathbf{L}$ can be obtained through an iterative algorithm: We first fix $\mathbf{L}$ and estimate $\boldsymbol{\beta}$, then fix $\boldsymbol{\beta}$ and estimate $\mathbf{L}$. Since the penalty function in (10) is separable, we iterate between the two steps above until the algorithm converges. Note that in both the loop of updating the fixed effects $\boldsymbol{\beta}$ and the loop of updating the parameters in $\mathbf{L}$, the algorithm calculates the estimate of only one parameter while holding other parameters fixed. This type of coordinate descent algorithm and some of its modified versions have been widely used for optimization in the literature, and the algorithmic convergence has been shown in Tseng (2001) and Tseng and Yun (2009), among others. In our case, at each update, the corresponding objective function is strictly convex, which guarantees the existence of a stationary point corresponding to the minimum of the objective function. The proof of this property can be given in a very similar way to that outlined for Theorem 3 in Schelldorfer et al. (2011), and is omitted in this paper. It is worth noting that since the objective function (10) may not be jointly convex, the convergent value from our algorithm is not guaranteed to be the globally optimal solution.

When $\mathbf{L}$ is fixed, maximizing (10) with respect to $\boldsymbol{\beta}$ is similar to a LASSO type optimization; hence we can apply either the LARS/LASSO algorithm (Efron et al., 2004) or a quadratic programming package to efficiently solve for $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is fixed, directly maximizing (10) with respect to $\mathbf{L}$ is challenging. Following the same spirit as Lin and Zhang (2006), we transform the optimization to an equivalent problem that is easily solvable. The following proposition warrants the validity and feasibility of our new algorithm.

Proposition 3.1. *For any given $\boldsymbol{\beta}$ and λ_2, consider the following two objective functions:*

$$\text{(11)}\qquad Q_{1,\hat{\boldsymbol{\beta}}}(\mathbf{L}) = P_R(\hat{\boldsymbol{\beta}},\mathbf{L}) - \lambda_2\sum_{k=2}^{q}\sqrt{L_{k1}^2+\cdots+L_{kq}^2},$$

$$\text{(12)}\qquad Q_{2,\hat{\boldsymbol{\beta}}}(\mathbf{L},\boldsymbol{\gamma}) = P_R(\hat{\boldsymbol{\beta}},\mathbf{L}) - \sum_{k=2}^{q}\gamma_k^2 \\ -\frac{\lambda_2^2}{4}\sum_{k=2}^{q}\frac{1}{\gamma_k^2}\Big(\sum_{j=1}^{q}L_{kj}^2\Big).$$

Let $\hat{L}_{kj}$ be the local maximizer of (11), and $(\tilde{\gamma}_k,\tilde{L}_{kj})$ be the local maximizer of (12), $k=2,\ldots,q, j=1,\ldots,q$. Then we have

$$\text{(13)}\qquad \hat{L}_{kj} = \tilde{L}_{kj},\ k=2,\ldots,q, j=1,\ldots,q;$$

$$\text{(14)}\qquad \tilde{\gamma}_k = \sqrt{\frac{\lambda_2}{2}\|\tilde{\mathbf{L}}_{(k)}\|_2},\ k=2,\ldots,q.$$

The proof of Proposition 3.1 is given in the appendix. This proposition suggests that, instead of maximizing (11) with respect to $\mathbf{L}$ directly, one can maximize (12) iteratively between γ_k and L_{kj}. Note that when γ_k is fixed, the objective function (12) resembles a generalized ridge regression, which can be solved via the Newton-Raphson algorithm. When L_{kj}'s are fixed, γ_k can be easily computed using formula (14). Overall, our proposed algorithm iteratively updates $\boldsymbol{\beta},\gamma_k$ and L_{kj}, and proceeds as follows:

1. Initialization: Initialize $\boldsymbol{\beta}^{(0)},\gamma_k^{(0)}$ and $L_{kj}^{(0)}$ with some plausible values. For example, $\boldsymbol{\beta}^{(0)}$ can be initialized by the least squares regression results for the $N>p$ case or the ridge regression results for the $N\le p$ case. $\mathbf{L}^{(0)}$ can be simply initialized by the identity matrix and $\gamma_k^{(0)}$ can be obtained based on $\mathbf{L}^{(0)}$.
2. Update L_{kj}: For iteration r, let
$$L_{kj}^{(r)} = \arg\max_{L_{kj}} P_R(\boldsymbol{\beta}^{(r-1)},\mathbf{D}) \\ -\frac{\lambda_2^2}{4}\sum_{k=1}^{q}\frac{1}{\Big(\gamma_k^{(r-1)}\Big)^2}\Big(\sum_{j=1}^{k}L_{kj}^2\Big).$$
3. Update γ_k:
$$\gamma_k^{(r)} = \sqrt{\frac{\lambda_2}{2}\|\mathbf{L}_{(k)}^{(r)}\|_2}.$$
4. Update $\boldsymbol{\beta}$ by LASSO:
$$\boldsymbol{\beta}^{(r)} = \arg\min_{\boldsymbol{\beta}}\frac{1}{2}\sum_{i=1}^{n}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)^T\mathbf{V}_i^{(r)^{-1}}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big) \\ +\lambda_1\sum_{j=1}^{p}|\beta_j|.$$
5. If both $\max_{k,j}\{|L_{kj}^{(r)}-L_{kj}^{(r-1)}|\}$ and $\max_j|\beta_j^{(r)}-\beta_j^{(r-1)}|$ are small enough, stop the algorithm. Otherwise, let $r=r+1$ and go back to step 2.

4. ASYMPTOTIC THEORY

In this section we present the large-sample properties for the proposed method. For clarity, we use notation of p_n and q_n to reflect the fact that the dimensions of both fixed and random effects diverge to infinity. As shown in the following two main theorems, both the diverging rates for p_n and q_n can be faster than n, a scenario referred to as $p_n\gg n$ and $q_n\gg n$. Similar rates for p_n have been studied in a vast literature. For example, Bickel et al. (2009) derived the large sample properties for the LASSO and Danzig selector in the case of $p_n \gg n$ in the linear model. On the other hand, our results regarding q_n are new. For example, Lam and Fan (2009) established asymptotic properties for a penalized maximum likelihood method to estimate the covariance matrix, and they found that in order to achieve a desirable convergence rate, the dimension of the covariance matrix cannot grow faster than the number of observations. Our proposed method concerns a similar problem, i.e. selecting the random effects through a regularized estimation of covariance matrix $\mathbf{D}$. We found that due to the sparsity of the random effects, the diverging rate for q_n (similar to that for p_n) can be established. As a result, q_n will grow at a faster rate than n, which differs from the covariance estimation result given by Lam and Fan (2009).

There is little literature available concerning the large-sample theory on the simultaneous regularization for the regression mean model and the covariance matrix with both divergent dimensions of p_n and q_n. Our proofs of the main theorems (Theorems 4.1 and 4.2) are laid out from a non-trivial integration of analytics established by Bickel et al. (2009) and Lam and Fan (2009). The analytic complexity pertains to the non-diagonal covariance matrix with a divergent dimension of variance parameters. When both dimensions of fixed and random effects grow along with the sample size, the simultaneous selection method requires to reconcile between the variance of signal and the variance of noise in the model, which demands subtle controls on their diverging rates.

In the following presentation, we focus on the case where p_n is allowed to be larger than N, noting that the $N > p$ case can be handled in a relatively straightforward fashion based on the results in this section. Thus, we wish to maximize the following objective function:

$$(15)\quad Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2) = \frac{1}{n}\ell_{nF}(\boldsymbol{\beta}, \mathbf{L}\sigma^2) - \lambda_{1n}\sum_{j=1}^{p_n}|\beta_j| - \lambda_{2n}\sum_{k=2}^{q_n}\|\mathbf{L}_{(k)}\|_2,$$

where $\ell_{nF}(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ is the full log-likelihood in (2). For the convenience of discussion, we assume that each subject i consists of equal m observations; we also absorb σ^2 into $\mathbf{D}$ and rewrite $\sigma^2\mathbf{V}_i = \mathbf{Z}_i^T\mathbf{D}\mathbf{Z}_i + \sigma^2\mathbf{I}_m$.

The tuning parameters λ_{1n} and λ_{2n} in (15) vary with the sample size n and the dimensions, p_n and q_n. Denote the true vector of fixed effects by $\boldsymbol{\beta}^*$, the true Cholesky decomposition $\mathbf{L}$ of $\mathbf{D}$ as $\mathbf{L}^*$ and the true standard deviation of the observation error by σ^*.

We define the following notations:

$$\begin{aligned}
\mathcal{J} &= \{j : \beta_j^* \neq 0\} \text{ and } \mathcal{J}^c = \{j : \beta_j^* = 0\};\\
\mathcal{S} &= \{(k,j) : D_{kj}^* \neq 0\} \text{ and } \mathcal{S}^c = \{(k,j) : D_{kj}^* = 0\};\\
\mathcal{S}_D &= \{k : D_{kk}^* \neq 0\} \text{ and } \mathcal{S}_D^c = \{k : D_{kk}^* = 0\}.
\end{aligned}$$

For the fixed-effects parameters, we let $\mathcal{J}$ and $\mathcal{J}^c$ contain the indices of coefficients which are truly non-zero and truly zero, respectively. Equivalently defined for variances of random effects are $\mathcal{S}_D$ and $\mathcal{S}_D^c$. Finally, $\mathcal{S}$ and $\mathcal{S}^c$ contain the indices of elements in $\mathbf{D}$ which are truly non-zero and zero, respectively. For a vector $\boldsymbol{\beta} \in \mathtt{R}^{p_n}$ and a subset $J \subseteq \{1, \cdots, p_n\}$, we denote by $\boldsymbol{\delta}_J$ the vector in $\mathtt{R}^{p_n}$ that has the same coordinates as $\boldsymbol{\delta}$ on J and zero coordinates on the complement J^c of J. Denote the cardinality of a set J as $|J|$, and write $s_n = |\mathcal{J}|$ and $d_n = |\mathcal{S}|$. Without loss of generality, we assume that the first $\sqrt{d_n}$ random effects are in the true model. Let $m = \max_{i=1,\ldots,n} m_i$.

We developed two large-sample theorems in this section. Theorem 4.1 concerns the rate of estimation convergence, and Theorem 4.2 is devoted to the property of sparsistency. All regularity conditions required by the two theorems are stated in the appendix. It is worth noting that a condition in Assumption A.3 concerning the restricted eigenvalue on the random-effects covariates $\mathbf{Z}$ is critical for the scenario of $q_n \gg n$ in the main theorems. (Theorems 4.1 and 4.2)

Theorem 4.1. *(Rate of convergence) Under regularity conditions Assumption A.1 - Assumption A.5 in the appendix and sparse assumptions that both d_n and s_n are $O(1)$, if $\log p_n/n = O_p(\lambda_{1n}^2)$ and $\log q_n/n = O_p(\lambda_{2n}^2)$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{L}}$ and $\hat{\sigma}^2$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 = O_p(\log p_n/n)$, $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 = O_p(\log q_n/n)$ and $|\hat{\sigma}^2 - \sigma^{*2}|^2 = O_p(\log m/n)$. Here $\|A\|_F^2$ denotes the Frobenius norm.*

Theorem 4.1 implies that if p_n and n satisfy the rate of $\log p_n/n = o_p(1)$, or equivalently p_n diverges at an exponential rate with the sample size, the regularized ML approach provides a consistent estimator for the vector of regression coefficients, $\boldsymbol{\beta}$. Thus, our method allows the fixed-effects dimension p_n to be much larger than the sample size n. Similarly, the random-effects dimension q_n can also be much larger than the sample size n if certain regularity conditions related to $\mathbf{X}$ and $\mathbf{Z}$ are satisfied.

Theorem 4.2. *(Sparsistency) Under the conditions given in Theorem 4.1, for any local maximizer of (15) satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 = O_p(\log p_n/n)$, $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 = O_p(\log q_n/n)$ and $|\hat{\sigma}^2 - \sigma^{*2}|^2 = O_p(\log m/n)$, with probability tending to 1, $\hat{\beta}_j = 0$ for all $j \in \mathcal{J}^c$ and $\hat{D}_{kk} = 0$ for $k \in \mathcal{S}_D^c$.*

This sparsistency property ensures the selection consistency for the true signals on both fixed and random effects.

5. SIMULATION EXPERIMENT

We have conducted four simulation studies to explore the performances of the proposed method. In the first two examples, we studied the case of $N > p$ and generated longitudinal outcomes of 200 subjects, each consisting of 8 repeated observations with 100 predictors, i.e., $N = 200 \times 8$. The true model is based on 4 important predictors, three of which are subject-specific and have non-zero random effects. In the other two examples, we conducted simulation studies for the case of $N < p$ and generated 100 subjects with 5

Table 1. Simulation results of Examples 5.1 - 5.4 for non-adaptive regularization (NAR) and adaptive regularization (AR) methods; reported values are sensitivity, specificity, averaged model size (AMS), means (std. dev.) of the estimates over 200 repetitions.

Ex.	Method	Variable Selection						Parameter Estimation						
		Fixed effect			Random effect			Fixed effect				Variance component		
		Sen.	Spec.	AMS	Sen.	Spec.	AMS	β_1	β_2	β_5	β_{10}	$\sqrt{D_{11}}$	$\sqrt{D_{22}}$	$\sqrt{D_{10,10}}$
5.1	NAR	100	92.9	10.8	100	88.7	8.3	2.74	1.22	1.94	1.70	0.64	0.63	0.64
	std dev	(0)	(2.8)	(2.7)	(0)	(5.1)	(2.4)	(.09)	(.08)	(.03)	(.08)	(.08)	(.07)	(.07)
	AR	100	98.5	5.4	100	98.5	3.7	2.98	1.47	2.00	1.97	0.74	0.71	0.73
	std dev	(0)	(1.2)	(1.2)	(0)	(1.7)	(.80)	(.07)	(.06)	(.03)	(.07)	(.09)	(.09)	(.07)
5.2	NAR	100	90.4	13.2	100	92.6	8.3	2.73	1.20	1.94	1.70	0.61	0.61	0.62
	std dev	(0)	(3.4)	(3.3)	(0)	(3.8)	(2.4)	(.08)	(.07)	(.03)	(.07)	(.08)	(.08)	(.08)
	AR	100	98.7	5.3	100	97.9	3.7	2.98	1.47	2.00	1.97	0.73	0.72	0.73
	std dev	(0)	(1.0)	(1.0)	(0)	(2.0)	(.80)	(.07)	(.06)	(.03)	(.07)	(.09)	(.09)	(.09)
5.3	NAR	100	96.2	26.6	100	81.4	11.6	2.71	1.23	1.80	1.71	0.61	0.60	0.62
	std dev	(0)	(.97)	(5.8)	(0)	(6.9)	(3.6)	(.15)	(.13)	(.13)	(.14)	(.13)	(.12)	(.13)
	AR	100	98.6	12.2	100	93.3	5.9	2.96	1.48	1.95	1.97	0.71	0.73	0.73
	std dev	(0)	(.76)	(3.5)	(0)	(3.9)	(2.1)	(.10)	(.11)	(.10)	(.10)	(.10)	(.12)	(.11)
5.4	NAR	99.9	95.8	31.0	99.8	80.3	12.2	2.69	1.18	1.82	1.69	0.59	0.60	0.55
	std dev	(1.8)	(1.1)	(6.5)	(2.4)	(6.8)	(3.5)	(.13)	(.20)	(.15)	(.13)	(.12)	(.13)	(.19)
	AR	100	98.2	14.9	100	93.0	6.3	2.93	1.43	1.96	1.93	0.70	0.72	0.69
	std dev	(0)	(.71)	(4.1)	(0)	(4.0)	(1.9)	(.11)	(.13)	(.10)	(.09)	(.11)	(.12)	(.15)

repeated observations in each subject. The predictor size, p, is 600, which is greater than $N = 100 \times 5$.

Example 5.1. The true model used to simulate data is given by

$$\begin{aligned} y_{it} &= (1+b_{i0}) + (3+b_{i1})x_{it1} + (1.5+b_{i2})x_{it2} \\ &\quad +(2+0)x_{it5} + (2+b_{i,10})x_{it,10} + \varepsilon_{it}, \\ &\quad (b_{i0}, b_{i1}, b_{i2}, b_{i,10})^T \sim \text{MVN}(0, 0.8^2\mathbf{R}), \end{aligned}$$

where $x_{itj} \sim N(0,1)$ and $\text{Corr}(x_{itj}, x_{itj'}) = 0.5^{|j-j'|}$, and errors ε_{ij} are *i.i.d.* $N(0,1)$. The correlation matrix in the random-effects distribution is

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.5 & 0.3 & 0.2 \\ 0.5 & 1.0 & 0.5 & 0.3 \\ 0.3 & 0.5 & 1.0 & 0.5 \\ 0.2 & 0.3 & 0.5 & 1.0 \end{bmatrix}.$$

Example 5.2. The LMM is the same as that in Example 5.1, except that we have an equal-correlation structure among the predictors, $\text{Corr}(x_{itj}, x_{itj'}) = 0.5$.

Example 5.3. The LMM is the same as that in Example 5.1, except for the sample size $N = 100 \times 5$ and the predictor size $p = 600$.

Example 5.4. The LMM is the same as that in Example 5.2, except for the sample size $N = 100 \times 5$ and the predictor size $p = 600$.

When fitting the model, we included all predictors in the fixed effects component ($p = 100$ in Examples 1 and 2, $p = 600$ in Examples 3 and 4) and the first 50 predictors in the random-effects component ($q = 50$). Both fixed and random intercepts are always included in all models and are not subject to the selection process. The true size of the fixed-effects component is 4 and that of the random-effects component is 3.

Regarding the penalty, we considered two alternatives in controlling the tuning parameters. One is referred to as non-adaptive regularization (NAR), which is based on a simple grid search for the tuning parameters. The other is termed as adaptive regularization (AR), which allocates different penalty weights on different parameters. The idea of adaptive regularization has been extensively discussed in the literature, for example, Zou (2006), Wang et al. (2007), Zhang and Lu (2007), among others. Specifically, for the adaptive version, we first set λ_1 and λ_2 at some small values and obtain $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{L}}$. Then the adaptive weights for the two penalties are set by the corresponding reciprocals, i.e., $1/|\tilde{\boldsymbol{\beta}}_j|$ and $1/\|\tilde{\mathbf{L}}_{(k)}\|_2$. Further, following Wang et al. (2007), we selected tuning parameters λ_{1n} and λ_{2n} by minimizing the following BIC criterion:

$$\text{BIC} = -2P_R(\boldsymbol{\beta}, \mathbf{L}) + \left[d_\beta + \frac{(1+d_D)d_D}{2}\right]\log(n), \tag{16}$$

where d_β and d_D are the the total number of nonzero estimates in $\boldsymbol{\beta}$ and that in the diagonal elements of $\mathbf{D}$, respectively. It is known that BIC is computationally convenient and enables us to detect important covariates at a low rate of reporting false signals.

For each example, we repeated the analysis over 200 simulations. Table 1 summarizes the results of four examples (Ex. 5.1-5.4). We reported the selection sensitivity and specificity

Table 2. Results for the analysis of psychiatric symptom data using both non-adaptive and adaptive versions of the propsed regularized LMM as well as the LMM-EM of Bondell et al. (2010). The reported values are estimated fixed effects $\hat{\beta}_j$'s and the estimated variance components of random effects $\sqrt{\hat{D}_{kk}}$'s.

	Non-Adaptive Method		Adaptive Method		LMM-EM	
	Fix. Eff.	Var. Comp.	Fix. Eff.	Var. Comp.	Fix. Eff.	Var. Comp.
Age at baseline	0	0	0	0	0	0.052
Gender	0	0	0	0	0	0.080
Hispanic	0.010	0	0	0	0.046	0.069
Time	−0.060	0.142	−0.069	0.182	−0.029	0
Summer	−0.045	0.014	−0.033	0.033	−0.037	0.100
Winter	−0.041	0.010	−0.029	0.029	−0.035	0.004
Treatment	0	0	0	0	0	0.003
Time*Trt	−0.027	0.002	−0.006	0.005	−0.007	0.118
Gender*Trt	0.075	0	0.065	0	0.142	0
Hispanic*Trt	0	0	0	0	−0.025	0

for both fixed and random effects. Here, Sensitivity is defined as the number of correctly selected variables divided by the number of important variables, and specificity is defined as the number of correctly deleted variables divided by the number of noise variables. We also listed average selected model size and average point estimates over 200 repetitions as well as the corresponding empirical standard deviation. Average model size for the fixed effects is the arithmetic mean of the number of non-zero fixed effects over simulation runs. Average model size for the random effects is defined similarly. Since the random intercept was always included in the model, we omit it in the calculations of summary statistics.

The simulation results in Table 1 appear very encouraging. As seen, both non-adaptive regularization (NAR) and adaptive regularization (AR) methods identified important fixed and random effects perfectly, and were very effective in removing unimportant predictors. The selected model size was reasonably close to the true model size. We also see that the AR method enjoyed substantially smaller estimation bias for the variance component than the NAR method, in comparison to the true $\sqrt{D_{jj}} = 0.8$. In conclusion, both versions of the proposed regularization approaches were effective on identifying signals and useful in building prediction models. For the purpose of discovery, the AR version is recommended, since it appeared to have a slightly better control of false discovery rate than the NAR version.

6. DATA EXAMPLES

In this section, we apply the proposed method to two data examples.

6.1 Data example I

The data of the first example was collected from a longitudinal randomized controlled intervention trial on 423 adolescent children (11–21 years old) with HIV+ parents in a Hispanic population in New York City (Rotheram-Borus et al., 2004). The primary outcome of interest was a certain psychiatric symptom, specifically, a negative state of mind measured repeatedly by a Basic Symptoms Inventory (BSI) over a period of six years (with an average of 11.5 visits per person). Interested readers may refer to Weiss (2005) for detailed definition and normalization of the BSI score variable.

There were six covariates, including treatment (1 for the treatment group and 0 for the control group), age at baseline, gender, indicator for Hispanic race (1 if the subject is Hispanic and 0 otherwise), time of visit (logarithm of year), and season of visit. Seasonality was coded into three categories, with Winter, Spring and Summer corresponding to the periods of November through February, March through June, and July through October, respectively. In our analysis, we used Spring as the reference level and created two dummy variables for Summer and Winter. We also included two-way interactions between treatment and other covariates including time, gender and Hispanic. Thus, the LMM for the data analysis takes the following form:

$$\begin{aligned} \mathrm{BSI} \sim\ & \mathrm{Age\_at\_Baseline} + \mathrm{Gender} + \mathrm{Hispanic} + \mathrm{Summer} \\ & + \mathrm{Winter} + \mathrm{Time} + \mathrm{Treatment} + \mathrm{Time} * \mathrm{Treatment} \\ & + \mathrm{Gender} * \mathrm{Treatment} + \mathrm{Hispanic} * \mathrm{Treatment}, \end{aligned}$$

where these 10 predictors were included in both $\mathbf{X}_i$ for fixed effects and $\mathbf{Z}_i$ for random effects, that is, $p = 10$ and $q = 11$ (the extra one being the random intercept). The interaction between Time and Treatment allows us to assess whether there is a difference in the trend of changes of BSI in control and treatment groups.

We first applied the non-adaptive version of the proposed regularization method in the analysis, in which we selected tuning parameters using the BIC in (16). The results are summarized in the left part of Table 2. As we can see, the non-adaptive method selected Hispanic, Time, Summer, Winter, Time*Treatment and Gender*Treatment for nonzero fixed effects, and Time, Summer, Winter and Time*Treatment for nonzero random effects.

Table 3. Summary of bootstrap results in the psychiatric symptom data analysis. "Sel. Freq." represents the selection frequency over 200 bootstrap samples. Averaged estimates over 200 bootstrap samples and the corresponding standard errors (numbers in the parentheses) are also reported.

	Fixed Effect		Variance Component	
	Sel. Freq. (%)	Averaged Estimate	Sel. Freq. (%)	Averaged Estimate
Non-Adaptive Method				
Age at baseline	21	0.005 (0.009)	7	0.001 (0.009)
Gender	37	0.014 (0.030)	6	0.018 (0.155)
Hispanic	34	0.021 (0.044)	15	0.009 (0.067)
Time	99	−0.062 (0.017)	100	0.125 (0.048)
Summer	97	−0.043 (0.017)	93	0.014 (0.020)
Winter	98	−0.039 (0.016)	87	0.010 (0.011)
Treatment	11	−0.003 (0.018)	2	0.003 (0.020)
Time*Trt	64	−0.021 (0.020)	81	0.006 (0.011)
Gender*Trt	72	0.065 (0.064)	10	0.009 (0.084)
Hispanic*Trt	10	−0.005 (0.047)	8	0.016 (0.114)
Adaptive Method				
Age at baseline	23	0.006 (0.011)	16	0.001 (0.003)
Gender	16	0.010 (0.031)	18	0.041 (0.221)
Hispanic	45	0.028 (0.043)	27	0.024 (0.081)
Time	98	−0.063 (0.020)	100	0.160 (0.070)
Summer	84	−0.035 (0.021)	83	0.026 (0.031)
Winter	77	−0.030 (0.021)	77	0.020 (0.031)
Treatment	20	−0.011 (0.036)	12	0.002 (0.013)
Time*Trt	43	−0.018 (0.026)	78	0.027 (0.064)
Gender*Trt	74	0.093 (0.077)	27	0.004 (0.015)
Hispanic*Trt	22	−0.012 (0.044)	23	0.010 (0.051)

To assess this selection, we drew 100 bootstrap samples from the original dataset. Each bootstrap sample was then analyzed in the same way as done for the original dataset. The selection frequency and average estimates of the regression coefficients and variance components are reported in the upper part of Table 3.

We see that, among the fixed effects, Time, Summer, Winter, Time*Treatment and Gender*Treatment had high selection frequencies while Hispanic was selected in a much lower rate. As for the random effects, Time, Summer, Winter and Time*Treatment had high frequencies of being selected.

We also applied the adaptive doubly regularized LMM regression on the BSI dataset. For the construction of adaptive weights, we used the inverse of the estimates from ridge-penalized LMM. The results are also summarized in Table 2 (middle part). Similar as the non-adaptive method, the adaptive method also selected Time, Summer, Winter, Time*Treatment and Gender*Treatment for nonzero fixed effects, and Time, Summer, Winter and Time*Treatment for nonzero random effects. However, unlike the non-adaptive method, the adaptive method did not select Hispanic, which agrees with the low selection frequency from the 100 bootstrap sample analysis. In terms of the magnitude of the estimates, the non-adaptive and adaptive methods provided similar estimates for the fixed effects, while the estimates for the variance components from the adaptive method are slightly larger than those from the non-adaptive method.

Similar as the assessment done for the non-adaptive method, we also used bootstrap to evaluate the selection of the adaptive method. The results are reported in the lower part of Table 3, and they are similar to those from the non-adaptive method.

Overall, it seems that there were strong time effects and season effects on the psychiatric symptom in the study. There was also evidence that the treatment program was effective and the program worked better for boys than for girls, due to the nonzero interaction effects between Time and Treatment and between Gender and Treatment. The negative coefficient for Time indicates that the average symptom score decreased over time. The estimated coefficients for Winter and Summer also indicated that symptoms were more severe in spring than in winter or summer, while the summer and winter were not much different from each other.

Furthermore, some population heterogeneity seemed to exist in the time effect, season effects (Summer and Winter), and treatment effect (interaction between Time and Treatment) indicated by the corresponding nonzero variance components. This implies that subject-specific effects are imperative to interpret the relationship between the symptom and the four predictors. For example, the expected psychiatric symptom in the summer is different among the subjects, conditional on the other predictors being fixed.

For the purpose of comparison, we also applied the regularization method proposed by Bondell et al. (2010), in which an EM algorithm was implemented with a single tuning parameter for both fixed and random effects. For the rest of the section, we refer to this method as LMM-EM. The results are summarized in the right part of Table 2.

Similar to our methods, the LMM-EM also selected Time, Summer, Winter, Time*Treatment and Gender*Treatment for nonzero fixed effects, and Summer, Winter and Time*Treatment for nonzero random effects. However, there are two noticeable differences between our methods and the LMM-EM. First, the LMM-EM selected much more nonzero variance components than our methods, including Age_at_Baseline, Gender, Hispanic and Treatment. Note that these components, especially Age_at_Baseline, Gender and Treatment, were selected with very low frequencies in the bootstrap analysis of our methods. Second, the Time covariate was selected as a nonzero variance component with 100 percent frequency in the boostrap analysis of our methods; however, it was not selected by the LMM-EM. We note that the code for LMM-EM was kindly provided by Dr. Bondell. Due to the nature of the EM, the algorithm is not computationally efficient. In our case, it took about 4 full days to finish the model fitting on the BSI dataset, where the best tuning parameter was selected from 8 values on a grid. Thus, this greatly limited us to consider some additional analyses within a reasonable period of time, such as to carry out a bootstrap analysis for the LMM-EM method.

To further assess the covariates selected by the LMM-EM and our methods, we fit the regular linear mixed-effects model (without any penalty) using the selected fixed and random effects, respectively, by the non-adaptive method, the adaptive method and the LMM-EM. The model based on the effects selected by the adaptive method obtained the smallest AIC (3220.9) and BIC (3235.2) values, in comparison to the AIC (3223.1) and BIC (3237.4) values for the model selected by the non-adaptive method, and the AIC (3242.3) and BIC (3263.7) values for the model selected by LMM-EM. Overall, all evidence indicate that for this BSI dataset, the fixed and random effects selected by our methods are probably more reasonable than those by the LMM-EM.

6.2 Data example II

The data of the second example was gathered from a clinical study that aimed to identify protein signatures associated with post-transplant renal function of patients who underwent kidney transplant. Cibrik et al. (2013) reported a set of 17 proteins as biomarkers to predict transplant patients experiencing acute allograft rejection. The authors also envisioned a study to identify protein signatures for the prediction of long-term post-transplant outcomes. In this analysis, the outcome of interest was longitudinal trajectory of renal function measured by glomerular filtration rate (GFR) from each of 95 renal transplant patients. The goal was to identify proteins that were significantly associated with longitudinal GFR. Twenty-eight proteins were measured at the baseline on each patient, in which two proteins were removed from the analysis due to very low variation across subjects. To reduce the false discovery rate, we further removed those proteins that were found nonlinearly associated with GFR. The analysis of nonlinear relationship was performed by using the additive mixed effects model with a linear time effect and a smoothing spline function of each baseline protein marker. This resulted in 22 proteins to be included in our final analysis. Table 4 reports the results produced by both non-adaptive and adaptive versions of regularization methods, as well as the summary statistics drawn from 100 bootstrap replicates. To assess the selection stability at a similar level of sparsity, we fixed the tuning parameters at their optimal values determined by BIC from the analysis of the original dataset.

Estimates of the fixed effects and the variance components as well the selection frequency were shown in Table 4. Interestingly, the adaptive regularization method was able to detect multiple proteins associated with longitudinal GFR, whereas the non-adaptive regularization did not select any protein. Some of the detected proteins by the adaptive regularization had 100% selection rate, which were recommended to our collaborators for validation. Using the proposed regularization methods we also found that some of proteins had non-zero estimates of variance components, suggesting subject-specific effects among these 95 transplant patients.

7. CONCLUDING REMARKS

We have proposed a doubly regularized likelihood approach to select important fixed effects and random effects simultaneously. We have also established a large-sample theory for the rate of convergence and sparsistency under the situation where both dimensions of fixed and random effects can increase much faster than the sample size. Numerical results indicate that the proposed regularized methods work well in the selection of fixed and random effects, as well as the determination of the true model size. We have tested the proposed method for several high-dimensional cases in simulation studies as well as by two data examples of relatively low dimensionality that already presented great challenges to existing methods in the linear mixed-effects methods. There is a downward bias in the estimation of the variance components by the non-adaptive regularization method; however, it can be reduced by the adaptive regularization.

The new algorithm proposed for optimization is effective, as it is built upon two quadratic optimization recipes. In general, fast computational algorithms are critical to enhance the use of the regularized LMM regression in practice. Currently existing methods based on the EM algorithm may be disadvantageous in this aspect (e.g. Linstrom and Bates, 1988); the slow convergence rate of the EM algorithm will limit its capability for handling a large number of predictors. For example, in the analysis of the BSI data, we have already experienced the advantage of com-

Table 4. Summary of bootstrap results with the tuning parameters fixed for protein data analysis. "Sel. Freq." represents the selection frequency over 100 bootstrap samples. Averaged estimates over 100 bootstrap samples and the corresponding standard errors (numbers in the parentheses) are also reported

	Non-Adaptive			
	Fixed Effect		Variance Component	
	Sel. Freq. (%)	Averaged Estimate	Sel. Freq. (%)	Averaged Estimate
Time	100	4.89 (0.03)	76	0.03 (0.003)
BCAM	13	0	0	0
CD30	0	0	0	0
E.Cadherin	2	0	0	0
GRO.alpha	0	0	0	0
IFN.gamma	0	0	0	0
IL.13	2	0	0	0
IL.2	0	0	0	0
IL.4	0	0	0	0
IL.5	4	0	0	0
IL.6	0	0	0	0
IL.1.beta	3	0	1	0
IL.12.p70	0	0	0	0
KIM.1	3	0	0	0
Lactoferrin	0	0	0	0
MCP.1	1	0	0	0
MCP.2	0	0	0	0
MIP.1.alpha	0	0	0	0
Osteopontin	0	0	0	0
TIMP.4	2	0	0	0
TNF.alpha	0	0	0	0
VEGF	4	0	0	0
VEGF.R2	0	0	0	0
	Adaptive			
	Fixed Effect		Variance Component	
	Sel. Freq. (%)	Averaged Estimate	Sel. Freq. (%)	Averaged Estimate
Time	100	5.190 (0.034)	100	0.115 (0.003)
BCAM	98	0.015 (0.001)	0	0
CD30	47	0.001 (0.000)	0	0
E.Cadherin	8	0.001 (0.000)	0	0
GRO.alpha	93	-0.005 (0.001)	0	0
IFN.gamma	100	-0.088 (0.041)	85	0.033 (0.002)
IL.13	85	-0.008 (0.001)	95	0.047 (0.004)
IL.2	95	0.236 (0.050)	15	0.003 (0.001)
IL.4	98	0.007 (0.010)	76	0.021 (0.002)
IL.5	100	-0.315 (0.030)	91	0.027 (0.002)
IL.6	100	-0.026 (0.028)	100	0.027 (0.002)
IL.1.beta	98	-1.298 (0.184)	100	0.175 (0.009)
IL.12.p70	100	-0.007 (0.001)	0	0
KIM.1	92	0.007 (0.001)	0	0
Lactoferrin	0	0	0	0
MCP.1	95	-0.013 (0.003)	0	0
MCP.2	98	-0.026 (0.007)	14	0.002 (0.000)
MIP.1.alpha	76	0	0	0
Osteopontin	0	0	0	0
TIMP.4	21	0	0	0
TNF.alpha	100	0.164 (0.016)	28	0.006 (0.001)
VEGF	99	-0.035 (0.005)	0	0
VEGF.R2	44	0.001 (0.000)	0	0

puting speed enjoyed by our method over the LMM-EM method.

A useful extension arises from possible hierarchy between fixed effects and random effects. For example, one may prefer the composition of random effects be a subset of the included fixed effects. In other words, if a predictor is identified to have a subject-specific effect, then the corresponding fixed effect should also be included in the model. The proposed method can be easily generalized to handle this constraint.

Without loss of generality, suppose $\mathbf{Z}_i$ is the first q columns of $\mathbf{X}_i$, for $i = 1, \ldots, n$. Now consider a reparameterized Cholesky decomposition

$$\mathbf{D} = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_q \end{pmatrix} \mathbf{L}\mathbf{L}^T \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_q \end{pmatrix}, \tag{17}$$

where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements. Clearly, if $\beta_j = 0$, the jth row and the jth column of $\mathbf{D}$ are also zero, regardless of the value of $\mathbf{L}_{(j)}$.

For regularization, we may then consider the following optimization problem:

$$(\hat{\boldsymbol{\beta}}, \hat{L}_{ij}) = \arg\max_{\boldsymbol{\beta}, \mathbf{L}} P_R - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\mathbf{L}_{(k)}\|_2. \tag{18}$$

As pointed out above, if $\hat{\beta}_j = 0$, from (17) the penalty on $\mathbf{L}$ will guarantee that $\hat{\mathbf{L}}_{(j)}$ is also estimated as zero. As a result, when a fixed effect β_j is shrunk to zero, the corresponding random effect will be automatically excluded from the model. The algorithm proposed in Section 3 can be applied to solve (18) with a slight modification.

APPENDIX A. TECHNICAL DETAILS

A.1 Proof of Proposition 3.1

The fact that the argument γ that maximizes the objective function (13) has the expression, $\tilde{\gamma}_k = \sqrt{\frac{\lambda_2}{2}\|\tilde{\mathbf{L}}_{(k)}\|_2}$, $k = 2, \ldots, q$, as described in (15), can be obtained by an application of the Cauchy-Schwarz inequality $a^2 + b^2 \geq 2ab$. Next, we prove $\hat{L}_{kj} = \tilde{L}_{kj}$.

Recall the definition of the objective functions, $Q_{1,\hat{\boldsymbol{\beta}}}(\mathbf{L})$ and $Q_{2,\hat{\boldsymbol{\beta}}}(\mathbf{L}, \boldsymbol{\gamma})$ in equations (11) and (12) in Section 3 and that $\hat{L}_{kj}$ and $(\tilde{\gamma}_k, \tilde{L}_{kj})$ maximize (11) and (12), respectively.

Direct derivations lead to that $Q_{1,\hat{\boldsymbol{\beta}}}(\tilde{\mathbf{L}}) = Q_{2,\hat{\boldsymbol{\beta}}}(\tilde{\mathbf{L}}, \tilde{\boldsymbol{\gamma}})$. Consequently, $Q_{1,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}}) \geq Q_{2,\hat{\boldsymbol{\beta}}}(\tilde{\mathbf{L}}, \tilde{\boldsymbol{\gamma}})$.

Letting $\hat{\gamma}_k = \sqrt{\frac{\lambda_2}{2}\|\hat{\mathbf{L}}_{(k)}\|_2}$, following some further derivations, we obtain that $Q_{2,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}}, \hat{\boldsymbol{\gamma}}) = Q_{1,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}})$, which leads to $Q_{2,\hat{\boldsymbol{\beta}}}(\tilde{\mathbf{L}}, \tilde{\boldsymbol{\gamma}}) \geq Q_{1,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}})$. That is, $Q_{1,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}}) = Q_{2,\hat{\boldsymbol{\beta}}}(\hat{\mathbf{L}}, \hat{\boldsymbol{\gamma}}) = Q_{2,\hat{\boldsymbol{\beta}}}(\tilde{\mathbf{L}}, \tilde{\boldsymbol{\gamma}})$.

Since the objective function Q is locally convex, the locally maximizer is unique. Thus, we have $\tilde{L}_{kj} = \hat{L}_{kj}$.

A.2 Proof of Theorem 4.1

The following technical regularity conditions are assumed throughout the proofs:

Assumption A.1. *Denote* $\mathbf{u}_i = \mathbf{Z}_i^T \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)$, $\tilde{\mathbf{u}}_i = \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)$, $\mathbf{A}_i = \mathbf{Z}_i^T \mathbf{V}_i^{*-1} \mathbf{Z}_i$, $\tilde{\mathbf{A}}_i = \mathbf{V}_i^{*-1}$, $W_{i,kl} = (\mathbf{u}_i)_k (\mathbf{u}_i)_l - (\mathbf{A}_i)_{kl}$ *and* $\tilde{W}_{i,kl} = (\tilde{\mathbf{u}}_i)_k (\tilde{\mathbf{u}}_i)_l - (\tilde{\mathbf{A}}_i)_{kl}$. *There are constants* τ_1 *and* τ_2 *such that* $0 < \tau_1 < \lambda_{\min}(\min_{i=1,\ldots,n} \mathbf{A}_i) \leq \lambda_{\max}(\max_{i=1,\ldots,n} \mathbf{A}_i) < \tau_2 < \infty$.

Assumption A.2. *For any* $\|\boldsymbol{\delta}\| \leq O_p\left((\log q_n / n)^{1/2}\right)$, *assume that*

$$\frac{1}{n}\sum_{i=1}^{n} Var\Big\{\boldsymbol{\delta}^T \mathbf{Z}_i^T \mathbf{V}_i^{*-1} \mathbf{Z}_i \otimes \mathbf{Z}_i^T \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)^T \mathbf{V}_i^{*-1} \mathbf{Z}_i \boldsymbol{\delta}\Big\} < \infty,$$

$$\frac{1}{n}\sum_{i=1}^{n} Var\Big\{\boldsymbol{\delta}^T \mathbf{Z}_i^T \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)^T \mathbf{V}_i^{*-1} \mathbf{Z}_i \otimes \mathbf{Z}_i^T \mathbf{V}_i^{*-1} \mathbf{Z}_i \boldsymbol{\delta}\Big\} < \infty,$$

and assume that

$$\frac{1}{n}\sum_{i=1}^{n} Var\left\{vec(\mathbf{I}_m)^T \mathbf{V}_i^{*-1} \otimes \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)^T \mathbf{V}_i^{*-1} vec(\mathbf{I}_m)\right\} < \infty,$$

$$\frac{1}{n}\sum_{i=1}^{n} Var\left\{vec(\mathbf{I}_m)^T \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i \mathbf{b}_i)^T \mathbf{V}_i^{*-1} \otimes \mathbf{V}_i^{*-1} vec(\mathbf{I}_m)\right\} < \infty.$$

Assumption A.3. *Restricted eigenvalue assumption* $RE(s, c_{01})$ *for* $\mathbf{X}$ *with* $1 \leq s \leq p_n$

$$\kappa_1^2 \equiv \min_{J_0 \subseteq \{1,\cdots,p_n\}:|J_0| \leq s} \min\nolimits_{\boldsymbol{\delta} \neq 0, \in \mathrm{R}^{p_n}: |\boldsymbol{\delta}_{J_0^c}|_1 \leq c_{01} |\boldsymbol{\delta}_{J_0}|_1} \frac{\sum_{i=1}^{n} \boldsymbol{\delta}^T \mathbf{X}_i^T \mathbf{V}_i^{*-1} \mathbf{X}_i \boldsymbol{\delta}}{n \boldsymbol{\delta}_{J_0}^T \boldsymbol{\delta}_{J_0}} > 0$$

holds for $c_{01} > 1$ *with probability one, and similarly restricted eigenvalue assumption* $RE(d, c_{02})$ *for* $\mathbf{Z}$ *with* $1 \leq d \leq q_n^2$

$$\kappa_2^2 \equiv \min_{J_0 \subseteq \{1,\cdots,q_n^2\}:|J_0| \leq d} \min_{\boldsymbol{\delta} \neq 0, \in \mathrm{R}^{q_n^2}: |\boldsymbol{\delta}_{J_0^c}|_1 \leq c_{02} |\boldsymbol{\delta}_{J_0}|_1} \frac{\sum_{i=1}^{n} \boldsymbol{\delta}^T (\mathbf{Z}_i \otimes \mathbf{Z}_i)^T (\mathbf{V}_i^* \otimes \mathbf{V}_i^*)^{-1} (\mathbf{Z}_i \otimes \mathbf{Z}_i) \boldsymbol{\delta}}{n \boldsymbol{\delta}_{J_0}^T \boldsymbol{\delta}_{J_0}} > 0$$

holds for $c_{02} > 0$ *with probability one.*

Assumption A.4. *The eigenvalues of* $n^{-1}\sum_{i=1}^{n} \mathbf{V}_i^{*-1} \otimes \mathbf{V}_i^{*-1}$ *are positive and bounded with probability one.*

Assumption A.5. *Denote* $|||\mathbf{X}_{[j]}|||_n^{(\nu)} = (n^{-1}\sum_{i=1}^{n} \mathbf{X}_{i[j]}^T \mathbf{V}_i^{*-\nu} \mathbf{X}_{i[j]})^{-1/2}$ *where* $\mathbf{X}_{i[j]}$ *is the* j*-th* *column of* $\mathbf{X}_i$. *Assume* $|||\mathbf{X}_{[j]}|||_n^{(1)} = O_p(1)$ *and* $|||\mathbf{X}_{[j]}|||_n^{(2)} = O_p(1)$. *Define* $\alpha = \max_j |||\mathbf{X}_{[j]}|||_n^{(1)}$

Proof of Theorem 4.1. The main idea of the proof follows Lam and Fan (2009)'s paper. Due to the high-dimensionality of p_n and q_n, some techniques in Bickel et al. (2009) will be applied. We divide the proof into three parts. In the first part we prove $Q_n(\boldsymbol{\beta}^*, \mathbf{L}^*, \sigma^{*2}) \geq Q_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^{*2})$ for $\|\mathbf{L} - \mathbf{L}^*\|_F^2 = O_p(\log q_n/n)$. In the second part we show that $Q_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^{*2}) \geq Q_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^2)$ for $|\sigma^2 - \sigma^{*2}|^2 = O_p(\log m/n)$. Finally $Q_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^2) \geq Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ for $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 = O_p(\log p_n/n)$ is shown in the third part of the proof.

We write $\tilde{\ell}_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ as the sum of two terms:

$$\frac{1}{2}\tilde{\ell}_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2) = \frac{1}{2n}\left(\sum_{i=1}^n \log|\mathbf{V}_i| + \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right).$$

Using Taylor expansion, we have $I_1 = \tilde{\ell}_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^{*2}) - \tilde{\ell}_n(\boldsymbol{\beta}^*, \mathbf{L}^*, \sigma^{*2}) = K_1 + K_2 + o_p(1)$, where

$$\begin{aligned} K_1 = \frac{1}{n}\sum_{i=1}^n \mathrm{Tr}\Big[&\mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i - \\ &\mathbf{Z}_i^T\mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}\mathbf{Z}_i)\Delta_{\mathbf{D}}\Big], \\ K_2 \;=\; &\mathrm{vec}(\Delta_{\mathbf{D}})^T \cdot G_n \cdot \mathrm{vec}(\Delta_{\mathbf{D}}), \end{aligned}$$

with $\Delta_{\mathbf{D}} = \Delta_{\mathbf{L}}\mathbf{L}^{*T} + \mathbf{L}^*\Delta_{\mathbf{L}}^T + \Delta_{\mathbf{L}}\Delta_{\mathbf{L}}^T$ with $\Delta_{\mathbf{L}} = \mathbf{L} - \mathbf{L}^*$, and $G_n = \frac{1}{n}\sum_{i=1}^n G_i$ and the i-th term is given by

$$\begin{aligned} G_i = \\ &\mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i \otimes \mathbf{Z}_i^T\mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}\mathbf{Z}_i \\ &+ \mathbf{Z}_i^T\mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}\mathbf{Z}_i \otimes \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i \\ &- \mathbf{Z}_i^T\mathbf{V}^{*-1}\mathbf{Z}_i \otimes \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i. \end{aligned}$$

Denote $S_{0n} = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i$ and

$$S_n = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^T\mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}\mathbf{Z}_i,$$

and we have

$$|K_1| \leq L_1 + L_2,$$

where

$$\begin{aligned} L_1 &= |\sum_{(i,j)\in\mathcal{S}} (S_{0n} - S_n)_{ij}(\Delta_{\mathbf{D}})_{ij}|, \\ L_2 &= |\sum_{(i,j)\in\mathcal{S}^c} (S_{0n} - S_n)_{ij}(\Delta_{\mathbf{D}})_{ij}|. \end{aligned}$$

If Assumption A.1 is satisfied, similar to Lemma A.2 and Lemma A.3 in Bickel and Levina (2008), we have

$$\max_{ij} |(S_{0n} - S_n)_{ij}| \leq O_p\left((\log q_n/n)^{1/2}\right).$$

Consequently, we have

$$\begin{aligned} L_1 &\leq \max_{ij} |(S_{0n} - S_n)_{ij}||\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1 \\ &\leq O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1, \end{aligned}$$

where $\mathcal{S}_v$ denotes the nonzero element set of $\mathrm{vec}(\mathbf{D}^*)$. Let $\mathcal{S}_1 = \{i : (i,j) \in \mathcal{S}\}$ and $\mathcal{S}_1^c = \{i : (i,j) \in \mathcal{S}^c\}$, and we write

$$\begin{aligned} I_2 &= \lambda_{2n}\sum_{i\in\mathcal{S}_1^c}(\|\mathbf{L}_{(i)}\| - \|\mathbf{L}_{(i)}^*\|), \\ I_3 &= \lambda_{2n}\sum_{i\in\mathcal{S}_1}(\|\mathbf{L}_{(i)}\| - \|\mathbf{L}_{(i)}^*\|). \end{aligned}$$

Then we have

$$\begin{aligned} &I_2 - L_2 \\ &\geq \sum_{i\in\mathcal{S}_1^c}\lambda_{2n}\|\mathbf{L}_{(i)}\| - \max_{ij}|(S_{0n} - S_n)_{ij}|\sum_{(i,j)\in\mathcal{S}^c}\|\mathbf{L}_{(i)}\|\|\mathbf{L}_{(j)}\| \\ &\geq \sum_{i\in\mathcal{S}_1^c}\lambda_{2n}\|\mathbf{L}_{(i)}\| - \max_{ij}|(S_{0n} - S_n)_{ij}|\sum_{i\in\mathcal{S}_1^c}\|\mathbf{L}_{(i)}\|\|\Delta_{\mathbf{L}}\|_F \\ &\geq \sum_{i\in\mathcal{S}_1^c}\left[\lambda_{2n} - O_p\left((\log q_n/n)^{1/2}\right)O_p\left((\log q_n/n)^{1/2}\right)\right]\|\mathbf{L}_{(i)}\| \\ &\geq 0 \end{aligned}$$

for $\lambda_{2n} = O_p\left((\frac{\log q_n}{n})^{1/2}\right)$. For the term I_3, we have

$$\begin{aligned} I_3 &= \lambda_{2n}\sum_{i\in\mathcal{S}_1}(\|\mathbf{L}_{(i)}\| - \|\mathbf{L}_{(i)}^*\|) \\ &\leq \lambda_{2n}\sum_{i\in\mathcal{S}_1}\|(\Delta_{\mathbf{L}})_{(i)}\| \\ &\leq \lambda_{2n}d^{1/2}\|(\Delta_{\mathbf{L}})_{\mathcal{S}}\|_F \\ &\leq 2\lambda_{2n}d^{1/2}\|\mathbf{L}^{*-1}\|_F\|(\Delta_{\mathbf{D}})_{\mathcal{S}}\|_F \\ &\leq O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1 \end{aligned}$$

Therefore, $|L_1 + I_3| \leq O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1$. From Assumption A.2 and Kolmogorov's one series Theorem, we have

$$\begin{aligned} K_2 = \mathrm{vec}(\Delta_{\mathbf{D}})^T\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i^T\mathbf{V}^{*-1}\mathbf{Z}_i \otimes \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i\mathrm{vec}(\Delta_{\mathbf{D}}) \\ + o_p(1) \\ = \mathrm{vec}(\Delta_{\mathbf{D}})^T\frac{1}{n}\sum_{i=1}^n(\mathbf{Z}_i \otimes \mathbf{Z}_i)^T(\mathbf{V}_i^* \otimes \mathbf{V}_i^*)^{-1}(\mathbf{Z}_i \otimes \mathbf{Z}_i) \\ \times \mathrm{vec}(\Delta_{\mathbf{D}}) + o_p(1). \end{aligned}$$

Consequently from Assumption A.3, if $|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v^c}|_1 \leq c_{02}|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1$, we have

$$K_2 \;=\; \mathrm{vec}(\Delta_{\mathbf{D}})^T G_n \mathrm{vec}(\Delta_{\mathbf{D}})$$

$$
\begin{aligned}
&\geq \quad \kappa_2^2 \|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}\|^2 \\
&\geq \quad \kappa_2^2 |\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1^2/d.
\end{aligned}
$$

From above, if $|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1 \geq O_p\left((\log q_n/n)^{1/2}\right)$, we have $|L_1 + I_3| \leq K_2$. Using the condition $|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v^c}|_1 \leq c_{02}|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1$, we have $|\mathrm{vec}(\Delta_{\mathbf{D}})|_1 \leq (c_{02}+1)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1$. This means $|L_1 + I_3| \leq K_2$ if $|\mathrm{vec}(\Delta_{\mathbf{D}})|_1 \geq O_p\left((\log q_n/n)^{1/2}\right)$.

If $|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v^c}|_1 > c_{02}|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1$, we can obtain

$$
\begin{aligned}
|L_1 + L_2 + I_3| &\leq O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v}|_1 \\
&\quad + O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v^c}|_1 \\
&\leq O_p\left((\log q_n/n)^{1/2}\right)|\mathrm{vec}(\Delta_{\mathbf{D}})_{\mathcal{S}_v^c}|_1 \\
&\leq O_p\left((\log q_n/n)^{1/2}\right)\sum_{(i,j)\in\mathcal{S}^c}\|\mathbf{L}_{(i)}\|\|\mathbf{L}_{(j)}\| \\
&\leq O_p\left((\log q_n/n)^{1/2}\right)\sum_{i\in\mathcal{S}_1^c}\|\mathbf{L}_{(i)}\|\|\Delta_{\mathbf{L}}\|_F \\
&\leq O_p\left((\log q_n/n)^{1/2}\right)O_p\left((\log q_n/n)^{1/2}\right)\sum_{i\in\mathcal{S}_1^c}\|\mathbf{L}_{(i)}\| \\
&\leq I_2,
\end{aligned}
$$

for $\lambda_{2n} = O_p\left((\log q_n/n)^{1/2}\right)$. Since $q_n^{-1}|\mathrm{vec}(\Delta_{\mathbf{D}})|_1 \leq \|\Delta_{\mathbf{D}}\|_F \leq |\mathrm{vec}(\Delta_{\mathbf{D}})|_1$, we complete the first part of the proof.

Now we start to show the second part of the proof. Let $\Delta_{\sigma^2} = \sigma^2 - \sigma^{*2}$ and $\tilde{I}_1 = \tilde{\ell}(\boldsymbol{\beta}^*, \mathbf{L}, \sigma) - \tilde{\ell}(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^*) = \tilde{K}_1 + \tilde{K}_2 + o_p(1)$, where

$$
\begin{aligned}
\tilde{K}_1 &= \frac{1}{n}\sum_{i=1}^n \mathrm{Tr}\Big[\tilde{\mathbf{V}}_i^{-1} - \\
&\qquad \tilde{\mathbf{V}}_i^{-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\tilde{\mathbf{V}}_i^{-1}\Big]\Delta_{\sigma^2}, \\
\tilde{K}_2 &= (\Delta_{\sigma^2})^2\mathrm{vec}(\mathbf{I}_m)^T\tilde{G}_n\mathrm{vec}(\mathbf{I}_m),
\end{aligned}
$$

with $\tilde{\mathbf{V}}_i = \mathbf{V}_i^* + \mathbf{Z}_i\Delta_{\mathbf{D}}\mathbf{Z}_i^T$, and

$$
\begin{aligned}
\tilde{G}_n &= \frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{V}}_i^{-1}\otimes\tilde{\mathbf{V}}_i^{-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\tilde{\mathbf{V}}_i^{-1} \\
&\quad + \tilde{\mathbf{V}}_i^{-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\tilde{\mathbf{V}}_i^{-1}\otimes\tilde{\mathbf{V}}_i^{-1} \\
&\quad - \tilde{\mathbf{V}}_i^{-1}\otimes\tilde{\mathbf{V}}_i^{-1}.
\end{aligned}
$$

Due to $\|\Delta_{\mathbf{D}}\|_F^2 = O_p(\log q_n/n) = o_p(1)$, we have

$$
\begin{aligned}
\tilde{K}_1 = \frac{\Delta_{\sigma^2}}{n}\sum_{i=1}^n \mathrm{Tr}\Big[\mathbf{V}_i^{*-1} - \\
\mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}\Big] + o_p(1).
\end{aligned}
$$

Denote $\tilde{S}_{0n} = \frac{1}{n}\sum_{i=1}^n \mathbf{V}_i^{*-1}$ and

$$
\tilde{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{V}_i^{*-1}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{*-1}.
$$

From Assumption A.1 and similar argument in the first part, we have

$$
\max_{ij}|(\tilde{S}_{0n} - \tilde{S}_n)_{ij}| \leq O_p\left((\log m/n)^{1/2}\right).
$$

Then we have

$$
|\tilde{K}_1| \leq m \cdot O_p(\log m/n),
$$

for $|\Delta_{\sigma^2}|^2 = O_p(\log m/n)$.

From Assumption A.2 and Kolmogorov's one series Theorem, we have

$$
\begin{aligned}
\tilde{K}_2 &= (\Delta_{\sigma^2})^2\mathrm{vec}(\mathbf{I}_m)^T\left(\frac{1}{n}\sum_{i=1}^n \mathbf{V}_i^{*-1}\otimes\mathbf{V}_i^{*-1}\right)\mathrm{vec}(\mathbf{I}_m) \\
&\quad + o_p(1) \\
&\geq \lambda_{min}\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{V}_i^{*-1}\otimes\mathbf{V}_i^{*-1}\Big)(m/2)|\Delta_{\sigma^2}|^2(1 + o_p(1)) \\
&= Cm\log m/n.
\end{aligned}
$$

The last inequality is due to Assumption A.4. Hence $\tilde{K}_1$ is dominated by $\tilde{K}_2$ and we complete the second part of the proof.

For the third part, we use similar notations as the first part and let $\Delta_{\boldsymbol{\beta}} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $I_1' = \tilde{\ell}_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2) - \tilde{\ell}_n(\boldsymbol{\beta}^*, \mathbf{L}, \sigma^2) = K_1' + K_2'$, where

$$
\begin{aligned}
K_2' &= \frac{2}{n}\sum_{i=1}^n \Delta_{\boldsymbol{\beta}}^T\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}, \\
\text{(19)} \qquad K_1' &= -\frac{2}{n}\sum_{i=1}^n(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}.
\end{aligned}
$$

where $\mathbf{V}_i = \mathbf{V}_i^* + \mathbf{Z}_i\Delta_{\mathbf{D}}\mathbf{Z}_i^T + \Delta_{\sigma^2}\mathbf{I}_m$. It can be seen

$$
|K_1'| \leq L_1' + L_2'
$$

where

$$
\begin{aligned}
L_1' &= \left|\frac{2}{n}\sum_{i=1}^n\sum_{j\in\mathcal{J}}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\Delta_{\boldsymbol{\beta}_j}\right|, \\
L_2' &= \left|\frac{2}{n}\sum_{i=1}^n\sum_{j\in\mathcal{J}^c}(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\Delta_{\boldsymbol{\beta}_j}\right|.
\end{aligned}
$$

Define the random variables $\eta_j = n^{-1}\sum_{i=1}^n(\boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\Delta_{\boldsymbol{\beta}_j}$, $1 \leq j \leq p$, and the event

$$
\mathcal{A} = \cap_{j=1}^p\{2|\eta_j| \leq c_1\lambda_{1n}\},
$$

where $c_1 < (c_{01}-1)/(c_{01}+1) < 1$ since $c_{01} > 1$. Denote $|||X_{[j]}|||_n^{'} = (n^{-1}\sum_{i=1}^n \mathbf{X}_{i[j]}^T\mathbf{V}_i^{-1}\mathbf{V}_i^*\mathbf{V}_i^{-1}\mathbf{X}_{i[j]})^{-1/2}$. We can have $|||\mathbf{X}_{[j]}|||_n^{'} = |||\mathbf{X}_{[j]}|||_n + o_p(1)$, using probability bound on the tails of standard Gaussian distribution, we have that the probability of $\mathcal{A}^c$ satisfies

$$\begin{aligned}P(\mathcal{A}^c) &\leq \sum_{j=1}^p P(\sqrt{n}|\eta_j| > \sqrt{n}\lambda_{1n}/2)\\ &\leq \sum_{j=1}^p P(|z| > \sqrt{n}\lambda_{1n}/(2|||\mathbf{X}_{[j]}|||_n^*))\\ &\leq \sum_{j=1}^p \exp\left(-\frac{nc_1\lambda_{1n}^2}{8|||\mathbf{X}_{[j]}|||_n^{*2}}\right)\\ &\leq p_n\exp\left(-\frac{nc_1^2\lambda_{1n}^2}{8\alpha^2}\right),\end{aligned}$$

which tends to 0 when $\lambda_{1n} = C_1\alpha(\log p_n/n)^{1/2}$ for $c_1C_1 > 2\sqrt{2}$. Denote

$$\begin{aligned}I_2' &= \lambda_{1n}\sum_{j\in\mathcal{J}^c}|\Delta_{\boldsymbol{\beta}_j}|,\\ I_3' &= \lambda_{1n}\sum_{j\in\mathcal{J}}|\beta_j^* + \Delta_{\boldsymbol{\beta}_j}| - |\beta_j^*|.\end{aligned}$$

By considering the event $\mathcal{A}$, we have $L_2' \leq I_2'$ for $c_1 < 1$ and $|L_1' + I_3'| \leq (c_1+1)\lambda_{1n}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1$.

Since $K_2' = \frac{2}{n}\sum_{i=1}^n \Delta_{\boldsymbol{\beta}}^T\mathbf{X}_i^T\mathbf{V}_i^{*-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}} + o_p(1)$, if $|\Delta_{\boldsymbol{\beta}_{\mathcal{J}^c}}|_1 \leq c_{01}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1$, by applying Assumption A.3, then $K_2' \geq \kappa_1^2\|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}\|^2 \geq \kappa_1^2|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1^2/s$. If $(c_1+1)\lambda_{1n}\kappa_1^{-2}s < |\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1$, we have $|L_1' + I_3'| < K_2'$. By using $|\Delta_\beta|_1 < (c_{01}+1)|\Delta_{\beta_{\mathcal{J}}}|_1$, if $|\Delta_{\boldsymbol{\beta}}|_1 > (c_1+1)(c_{01}+1)s\kappa_1^{-2}\lambda_{1n} = O_p((\log p_n/n)^{1/2})$, we have $|L_1' + I_3'| < K_2'$.

If $|\Delta_{\boldsymbol{\beta}_{\mathcal{J}^c}}|_1 > c_{01}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1$, $K_2' \geq 0$, and

$$\begin{aligned}|L_1' + L_2' + I_3'| &\leq (c_1+1)\lambda_{1n}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}}}|_1 + c_1\lambda_{1n}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}^c}}|_1\\ &\leq ((c_1+1)c_{01}^{-1} + c_1)\lambda_{1n}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}^c}}|_1\\ &\leq \lambda_{1n}|\Delta_{\boldsymbol{\beta}_{\mathcal{J}^c}}|_1 = I_2',\end{aligned}$$

as $c_1 < (c_{01}-1)/(c_{01}+1) < 1$ for $c_{01} > 1$. Since $p_n^{-1/2}|\Delta_{\boldsymbol{\beta}}|_1 \leq \|\Delta_{\boldsymbol{\beta}}\| \leq |\Delta_{\boldsymbol{\beta}}|_1$, we complete the proof. □

A.3 Proof of Theorem 4.2

Proof of Theorem 4.2. Suppose $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})$ is one set of minimizers of $Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 = O_p(\log p_n/n)$, $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 = O_p(\log q_n/n)$ and $|\hat{\sigma}^2 - \sigma^{*2}|^2 = O_p(\log m/n)$. Taking the derivative of $Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma^2)$ w.r.t β_j for $j \in \mathcal{J}^c$ at $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma}^2)$, we have

$$\begin{aligned}\frac{\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})}{\partial\beta_j} &= -\frac{2}{n}\sum_{i=1}^n(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})\mathbf{V}_i^{-1}\mathbf{X}_{i[j]} + \lambda_{1n}\text{sgn}(\beta_j)\\ &= |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^T\left[\frac{2}{n}\sum_{i=1}^n\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\right] - 2\eta_j\\ &\quad + \lambda_{1n}\text{sgn}(\beta_j)\\ &= O_p(\lambda_{1n}) + \lambda_{1n}\text{sgn}(\beta_j).\end{aligned}$$

The last step is due to

$$\begin{aligned}&|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^T\left[\frac{2}{n}\sum_{i=1}^n\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\right]\\ &\leq \|\Delta_\beta\|\|\frac{2}{n}\sum_{i=1}^n\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_{i[j]}\|\\ &\leq \|\Delta_\beta\|\left(\frac{4\phi_{\max}}{n}\sum_{i=1}^n\mathbf{X}_{i[j]}^T\mathbf{V}_i^{-2}\mathbf{X}_{i[j]}\right)^{1/2}\\ &\leq \|\Delta_\beta\|\left(4\phi_{\max}\|\mathbf{X}_{[j]}|||_n^{(2)} + o_p(1)\right)^{1/2}.\end{aligned}$$

Then the sign of $\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})/\partial\beta_j$ depends on $\text{sgn}(\beta_j)$ only with probability tending to one for $j \in \mathcal{J}^c$. and the sparsistency property for $\hat{\beta}_j$ in Theorem 4.2 is satisfied.

Next, taking the derivative of $Q_n(\boldsymbol{\beta}, \mathbf{L}, \sigma)$ w.r.t D_{kk} for $k \in \mathcal{S}_D^c$ at the minimizer $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})$, we have

$$\frac{\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})}{\partial D_{kk}} = (\tilde{S}_{0n} - \tilde{S}_n)_{kk} + \frac{\lambda_{2n}}{2\sqrt{D_{kk}}},$$

where $S_{0n}' = \frac{1}{n}\sum_{i=1}^n\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i$, and $S_n' = \frac{1}{n}\sum_{i=1}^n S_{ni}'$ with the i-th term S_{ni}' given by

$$\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{X}_i\Delta_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)(\mathbf{X}_i\Delta_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i)^T\mathbf{V}_i^{-1}\mathbf{Z}_i.$$

We decompose $(S_{0n}' - S_n')_{kk} = I_1 + I_2 + I_3$ where

$$I_1 = (S_{0n}' - S_{0n})_{kk},\ I_2 = (S_{0n} - S_n)_{kk},\ \text{and } I_3 = (S_n - S_n')_{kk}$$

Since $\mathbf{V}_i - \mathbf{V}_i^* = \mathbf{Z}_i\Delta_{\mathbf{D}}\mathbf{Z}_i^T + \Delta_{\sigma^2}\mathbf{I}_m$, we have $I_1 \leq \|S_{0n}' - S_{0n}\| = o_p(1)$. From the argument in the proof of Theorem 4.1, we have $I_2 = O_p\left((\log q_n/n)^{1/2}\right)$. Denote $\mathbf{e}_i = \boldsymbol{\varepsilon}_i + \mathbf{Z}_i\mathbf{b}_i$ and decompose $I_3 = I_{31} + I_{32} + I_{33}$, where

$$I_{31} = \frac{1}{n}\sum_{i=1}^n(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}^T\mathbf{V}_i^{-1}\mathbf{Z}_i)_{kk},$$

$$\begin{aligned}I_{32} &= \frac{1}{n}\sum_{i=1}^n\left(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}\mathbf{e}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i\right)_{kk}\\ &\quad + \frac{1}{n}\sum_{i=1}^n\left(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\Delta_{\boldsymbol{\beta}}\mathbf{e}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i\right)_{kk}^T,\end{aligned}$$

$$I_{33} = \frac{1}{n}\sum_{i=1}^n(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i)_{kk}.$$

From LLN, $I_{32} = o_p(1)$, and

$$I_{33} = E[(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{V}_i^*\mathbf{V}_i^{-1}\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i)_{kk}] + o_p(1)$$

$$\begin{aligned} &\leq E\|\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{V}_i^*\mathbf{V}_i^{-1}\mathbf{Z}_i - \mathbf{Z}_i^T\mathbf{V}_i^{*-1}\mathbf{Z}_i\| + o_p(1) \\ &= o_p(1). \end{aligned}$$

For I_{31} we have

$$I_{31} = \frac{1}{n}\sum_{i=1}^{n}\|(\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)_{(k)}\Delta_{\boldsymbol{\beta}}\|^2 \geq 0.$$

Hence, we have $(S'_{0n} - S'_n)_{kk} = I_{31} + O_p\left((\log q_n/n)^{1/2}\right)$. For $k \in \mathcal{S}_D^c$, since $D_{kk}^* = 0$, then we have $\hat{D}_{kk} \leq \|\Delta_{\mathbf{D}}\|_F = O_p\left((\log q_n/n)^{1/2}\right)$. Therefore

$$\frac{\lambda_{2n}}{2\sqrt{\hat{D}_{kk}}} \geq O_p\left((\log q_n/n)^{1/2}\right),$$

Then the sign of $\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\mathbf{L}}, \hat{\sigma})/\partial D_{kk}$ is always nonnegative with probability tending to one for $k \in \mathcal{S}_D^c$, and the sparsistency property for $\hat{D}_{kk}$ in Theorem 4.2 is satisfied. The proof of the theorem is completed. □

ACKNOWLEDGEMENTS

The authors are grateful to the editor and two anonymous reviewers for their constructive suggestions on the paper. They also thank Dr. Howard Bondell for sharing the R code of the LMM-EM method.

Received 12 November 2015

REFERENCES

Ahn, M., Zhang, H. H. and Lu. W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, **22**, 1539–1562. MR3027098

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds), Budapest: Akademiai Kiado. MR0483125

Albert, J. and Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, **92**, 916–925.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, **37**, 1705–1732. MR2533469

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics*, **66**, 1069–1077. MR2758494

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384. MR1365720

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25. MR1397972

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769. MR2025100

Cibrik, D. M.,Warner, R. L., Kommareddi, M., Song, P., Luan, F. L. and Johnson, K. J. (2013) Identification of a protein signature in allograft rejection. *Proteomics Clinical Applications*, **7**, 839–849.

Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B*, **59**, 157–171. MR1436561

Du, Y., Khalili, A., Nešlehová, J. and Steele, R. (2013). Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models. *Canadian Journal of Statistics*, **41**, 596–616. MR3146000

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499. MR2060166

Fan, Y. and Li. R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics* **40**, 2043–2068. MR3059076

Foster, S. D., Verbyla, A. P. and Pitchford, W.S. (2009). Estimation, prediction and inference for the LASSO random effects models. *Australian & New Zealand Journal of Statistics*, **51**, 43–61. MR2504102

Greven S. and Kneib T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models *Biometrika,* **97,** 773–789. MR2746151

Hall, D. B. and Praestgaard, J. T. (2001). Order-restricted score tests for homogeneity in generalized linear and nonlinear mixed models. *Biometrika*, **88**, 739–751. MR1859406

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385. MR0368279

Ibrahim, J. G., Zhu, H. T., Garcia, R. I. and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, **67**, 495–503. MR2829018

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**, 805–820. MR0872961

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, **36**, 1669–1692. MR2435452

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63,** 690–698. MR2395705

Lai, R. C. S., Huang, H.-C. and Lee, T. (2012). Fixed and random effects selection in nonparametric additive mixed models *Electronic Journal of Statistics*, **6**, 810–842. MR2988430

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, **37**, 4254–4278. MR2572459

Lan, L. (2006). *Variable selection in linear mixed model for longitudinal data.* Ph.D. dissertation, Department of Statistics, North Carolina State University. MR2709243

Lange, N. and Laird, N. M. (1989). The effect of covariance structures on variance estimation in balance-curve models with random parameters. *Journal of the American Statistical Association*, **84**, 241–247. MR0999684

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95,** 773–778. MR2443190

Lin, B., Pang, Z. and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, **22**, 341–355. MR3173718

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326. MR1467049

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Applied Statistics*, **34**, 2272–2297. MR2291500

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. MR0997577

Pan, J. and MacKenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244. MR1966564

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, **24**, 725–738. MR3229693

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677–690. MR1723786

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–435. MR1782488

Rotheram-Borus, M. J., Lee, M., Lin, Y. Y. and Lester, P. (2004). Six year intervention outcomes for adolescent children of parents with the human immunodeficiency virus. *Archives of Pediatrics and Adolescent Medicine*, **158**, 742–748.

Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, **38**, 197–214. MR2829596

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. MR0468014

Sinharay, S. and Stern, H. S. (2001). Bayes factors for variance component testing in generalized linear mixed models. *Bayesian Methods with Applications to Science, Policy and Official Statistics (ISBA 2000 Proceeding)*, 507–516. MR2702187

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*, Springer, New York. MR2377853

Stram, D. O. and Lee, J. W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288. MR1379242

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, **109**, 475–494. MR1835069

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

Van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike Information for mixed-effects models. *Biometrika*, **92**, 351–370. MR2201364

Verbeke, G. and Lesaffre, E. (1997). The Effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, **23**, 541–556. MR1437679

Wang, H., Li, R. and Tsai, C.-L. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568. MR2410008

Weiss, R. E. (2005). *Modeling Longitudinal Data*, Springer, New York. MR2208652

Yang, M. (2012) Bayesian variable selection for logistic mixed model with nonparametric random effects. *Computational Statistics and Data Analysis*, **56**, 2663–2674. MR2915153

Ye, H. and Pan, J. (2006) Modelling covariance structures in generalized estimating equations for longitudinal data. *Biometrika*, **93**, 927–941. MR2285080

Yuan, M. and Lin, Y. (2006) Model selection and Estimation in regression with grouped Variable. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67. MR2212574

Zhang, H. and Lu, W. (2007) Adaptive-Lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703. MR2410017

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

Yun Li
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: yrlee10@gmail.com

Sijian Wang
Department of Biostatistics and Medical Informatics
Department of Statistics
University of Wisconsin at Madison
Madison, WI 53705
USA
E-mail address: swang@biostat.wisc.edu

Peter X.-K. Song
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: pxsong@umich.edu

Naisyin Wang
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: nwangaa@umich.edu

Ling Zhou
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: zholing@umich.edu

Ji Zhu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: jizhu@umich.edu