University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2010 *Paper* 89

Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models

Sijian Wang[*] Peter Xuewin Song[†]

Ji Zhu[‡]

[*]University of Michigan, sijwang@umich.edu

[†]University of Michigan - Ann Arbor, pxsong@umich.edu

[‡]University of Michigan, jizhu@umich.edu

Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models

Sijian Wang, Peter Xuewin Song, and Ji Zhu

Abstract

The linear mixed effects model (LMM) is widely used in the analysis of clustered or longitudinal data. In the practice of LMM, the inference on the structure of the random effects component is of great importance, not only to yield proper interpretation of subject-specific effects but also to draw valid statistical conclusions. This task of inference becomes significantly challenging when a large number of fixed effects and random effects are involved in the analysis. The difficulty of variable selection arises from the need of simultaneously regularizing both mean model and covariance structures, with possible parameter constraints between the two. In this paper, we propose a novel method of doubly regularized restricted maximum likelihood to select fixed and random effects simultaneously in the LMM. The Cholesky decomposition is invoked to ensure the positive-definiteness of the selected covariance matrix of random effects, and selected random effects are invariant with respect to the ordering of predictors appearing in the Cholesky decomposition. We then develop a new algorithm that solves the related optimization problem effectively, in which the computational cost is comparable with that of the Newton-Raphson algorithm for MLE or REML in the LMM. We also investigate large sample properties for the proposed method, including the oracle property. Both simulation studies and data analysis are included for illustration.

Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models [*]

Sijian Wang, Peter X.-K. Song and Ji Zhu

Abstract

The linear mixed effects model (LMM) is widely used in the analysis of clustered or longitudinal data. In the practice of LMM, the inference on the structure of the random effects component is of great importance, not only to yield proper interpretation of subject-specific effects but also to draw valid statistical conclusions. This task of inference becomes significantly challenging when a large number of fixed effects and random effects are involved in the analysis. The difficulty of variable selection arises from the need of simultaneously regularizing both mean model and covariance structures, with possible parameter constraints between the two. In this paper, we propose a novel method of doubly regularized restricted maximum likelihood to select fixed and random effects simultaneously in the LMM. The Cholesky decomposition is invoked to ensure the positive-definiteness of the selected covariance matrix of random effects, and selected random effects are invariant with respect to the ordering of predictors appearing in the Cholesky decomposition. We then develop a new algorithm that solves the related optimization problem effectively, in which the computational cost is comparable with that of the Newton-Raphson algorithm for MLE or REML in the LMM. We also investigate large sample properties for the proposed method, including the oracle property. Both simulation studies and data analysis are included for illustration.

Key Words: Oracle property; REML; regularization; variable selection.

[*] S. Wang is Assistant Professor, Department of Biostatistics and Medical Informatics and Department of Statistics, University of Wisconsin at Madison, Madison, WI 53705. (Email: *swang@biostat.wisc.edu*). P. Song is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, (Email: *pxsong@umich.edu*). Ji Zhu is Associate Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (Email: *jizhu@umich.edu*).

1 Introduction

High-dimensional clustered or longitudinal 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 a large number of covariates. 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.

In the current literature, predictors in a variable selection problem often refer to, in the LMM context, covariates of fixed effects. 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. The selection of fixed effects covariates may be done through the subset selection method using AIC (Akaike, 1973), or BIC (Schwarz, 1978) or conditional AIC (Vaida and Blanchard, 2005). However, this selection procedure is known to be inefficient or even infeasible when the number of predictors is large.

In addition to selecting fixed effects, another important challenge in LMM is to determine the structure of the random effects component. The selection of the random effects is equally important to the selection of fixed effects. This is because the random effects component not only determines the marginal covariance structure of the correlated data, but also pertains to 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 (e.g. Lange and Laird, 1989). Lange and Laird (1989) showed that an under-specified random-effects component would lead to biased estimation for the variance of the fixed effects. On the other hand, when the random-effects component is over-specified, the covariance structure becomes over-parameterized, which may lead to loss of estimation efficiency. Therefore, an appropriate composition of the random-effects component is critical for valid statistical inference.

At the same time, we would also like to note that determining the configuration of the random effects component is important for researchers to understand and interpret mechanisms of population heterogeneity in longitudinal studies. In complex correlated data, subject-specific characterizations may arise from multiple sources involving many predictors. Thus, learning which covariates exhibit subject-specific effects, and consequently their random effects be included in the model, is of practical importance. In the current literature, it is suggested that one may run a preliminary analysis based on individual cluster regression models to examine which covariates exhibit potential subject-specific effects. Clearly, this approach is rather limited and may become unreliable when the cluster size is small. Consequently, the resulting random-effects component specification can be subjective.

In this paper, we develop a data-driven procedure that enables us to select both fixed and random effects simultaneously, in the case where the number of candidate fixed and random effects can be large. There are several other work that have also contributed to the selection of fixed effects or the random effects component for LMM. Stram and Lee (1994) discussed the asymptotic behavior of a likelihood ratio test for nonzero random effect 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. Jiang et al. (2008) developed a "fence" method for variable selection in a general mixed effects model. In a PhD thesis, Lan (2006) developed a penalized likelihood-based approach to select the fixed effects component with a given structure of random effects, but they did not consider the random effects component selection. Foster et al. (2009) proposed a LASSO random effects models with no fixed effects, where random effects are 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 selection of random effects using Bayesian approaches. Like Foster et al. (2009), these papers did not consider the fixed effects component selection.

In this paper, we propose a method based on doubly regularized restricted maximum likelihood (REML) in that regularization takes place simultaneously at estimation of both fixed and random effects. In the context of the LMM, the REML method has been shown to have advantages over many of its competitors such as the conditional likelihood method and the EM algorithm based method, in terms of both small-sample properties and numerical performance (e.g. Harville, 1977; Lindstrom and Bates, 1988). We use the Cholesky decomposition to ensure the positive-definiteness of the selected covariance matrix of random effects. The resulting random effects selection is invariant with respect to the ordering of predictors appearing in the Cholesky decomposition. We develop an effective algorithm to deal with the involved optimization, where the computational cost is similar to that of the Newton-Raphson algorithm for MLE or REML in the LMM. Furthermore, we investigate large-sample properties of the proposed method, and show that when tuning parameters are appropriately chosen, the proposed estimation enjoys the oracle property (Fan and Li, 2001); that is, it performs as well as if the correct underlying model were given in advance.

The rest of the paper is organized as follows. In Section 2, we introduce our new method: the doubly regularized REML. In Section 3, we discuss a new algorithm to carry out the related optimization. In Section 4, we study the asymptotic behavior of the doubly regularized REML and propose an improvement for the method. In Sections 5 and 6, we demonstrate our method via simulations and a real data analysis, respectively. We conclude the paper with Section 7. All technical proofs are given in the Supplemental Material.

2 Method

2.1 Linear Mixed Model

Suppose there are n subjects under study, and there are m_i observations for subject i, $i = 1, \dots, n$. There are p fixed effects covariates: $X_1, \dots, X_p$, and q random effects covariates: $Z_1, \dots, Z_q$. Usually, the q random effects covariates are a subset of the p fixed effects covariates. For subject i at observation j, let Y_{ij} denote the response variable, $\boldsymbol{x}_{ij}$ be the vector of p predictors in the fixed effects component, and $\boldsymbol{z}_{ij}$ be the vector of q predictors in the random effects component. The linear mixed effects model is then written as

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

where ϵ_{ij}'s are assumed i.i.d. $N(0, \sigma^2)$, and the random effects, $\boldsymbol{b}_i = (b_{i1}, \dots, b_{iq})^T$, are i.i.d. according to a multivariate normal distribution $\mathrm{MVN}_q(0, \sigma^2 \boldsymbol{D})$. The set of parameters to be estimated is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)$. 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.

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

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

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

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

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

2.2 MLE and REML

Our variable selection method is built upon standard methods of estimation in the LMM, specifically, maximum likelihood (ML) and restricted maximum likelihood (REML) methods (e.g., Laird and Ware, 1982; Jennrich and Schluchter, 1986; Lindstrom and Bates, 1988).

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

$$\boldsymbol{Y}_i \sim \mathrm{MVN}_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_i), \tag{3}$$

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

$$\ell_F(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\log\Big|\sigma^2\boldsymbol{V_i}\Big| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{Y_i} - \boldsymbol{X_i\beta})^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i} - \boldsymbol{X_i\beta}), \tag{4}$$

and the ML estimates of parameters $\boldsymbol{\beta}, \boldsymbol{D}$ and σ^2 can be obtained by maximizing the log-likelihood function (4). Note that when $\boldsymbol{D}$ is known, the MLE for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}(\boldsymbol{D}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n}\Big(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\Big)^T\boldsymbol{V}_i^{-1}\Big(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\Big). \tag{5}$$

One well-known criticism on the ML estimation is that for the variance components (i.e. $\boldsymbol{D}$), 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 defining 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 log-likelihood, according to

Harville (1974), is

$$\begin{aligned}\ell_R(\boldsymbol{D},\sigma^2) &= -\frac{1}{2}\sum_{i=1}^{n}\log\left|\sigma^{-2}\boldsymbol{V}_i\right| - \frac{1}{2}\log\left|\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big\{\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{D})\Big\}^T\boldsymbol{V}_i^{-1}\Big\{\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{D})\Big\}, \qquad (6)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{D})$ is given by (5).

One way to obtain the estimate of $(\boldsymbol{\beta},\boldsymbol{D},\sigma^2)$ is to solve (5) and (6) iteratively until convergence. When convergence is achieved, one can estimate the random effects using BLUP (e.g. Song, 2007, Chapter 9):

$$\hat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}).$$

Joining the estimator (5) and the REML (6), we may write a modified log-likelihood as

$$\begin{aligned}\ell_n(\boldsymbol{\beta},\boldsymbol{D},\sigma^2) &= -\frac{1}{2}\sum_{i=1}^{n}\log\left|\sigma^{-2}\boldsymbol{V}_i\right| - \frac{1}{2}\log\left|\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\Big)^T\boldsymbol{V}_i^{-1}\Big(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\Big), \qquad (7)\end{aligned}$$

Clearly, the MLE of $\boldsymbol{\beta}$ and the REML of $\boldsymbol{D}$ can be obtained by jointly maximizing (7).

2.3 Doubly Regularized REML Estimation

The selection of fixed effects and random effects components can be realized through the selection of nonzero elements in $\boldsymbol{\beta}$ and $\boldsymbol{D}$. 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 $\boldsymbol{D}$ simultaneously, i.e.,

$$\max Q_n(\boldsymbol{\beta},\boldsymbol{D},\sigma^2) = \ell_n(\boldsymbol{\beta},\boldsymbol{D},\sigma^2) - \lambda_1 J_1(\boldsymbol{\beta}) - \lambda_2 J_2(\boldsymbol{D}), \qquad (8)$$

where 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 controls the selection of fixed effects. The second penalty function $J_2(\boldsymbol{D})$ controls the sparsity of the final estimation of $\boldsymbol{D}$, and hence controls the selection of random effects.

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

$$J_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|. \tag{9}$$

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 $\boldsymbol{D}$, we use the Cholesky decomposition, i.e., $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. This decomposition converts a constrained optimization into an unconstrained problem, and the resulting computation is more stable and faster. Consequently, the selection procedure will target $\boldsymbol{L}$, rather than $\boldsymbol{D}$. The relation between the sparsity of $\boldsymbol{D}$ and the sparsity of $\boldsymbol{L}$ is given by the following Lemma.

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

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

The proof is straightforward, and we omit it in this paper. Lemma 1 indicates that if the vector $\boldsymbol{L}_{(k)} = \boldsymbol{0}$, then the diagonal element D_{kk}, known as the variance of the random effect b_k, is zero. Furthermore, for any $j \neq k$, off-diagonal elements $D_{kj} = 0$, which implies that the covariances between b_k and all 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 $\boldsymbol{L}_{(k)}$ towards a zero vector. Therefore, we adopt the L_2-norm penalty (Yuan and Lin, 2005) for $J_2(\boldsymbol{D})$, i.e.,

$$J_2(\boldsymbol{L}) = \sum_{k=2}^{q} \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}. \tag{10}$$

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

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

3 Algorithm

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

$$Q_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) = \ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\boldsymbol{L}_{(k)}\|_2, \tag{11}$$

where $\|\boldsymbol{L}_{(k)}\|_2 = \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}$.

To simplify the computation, following Lindstrom and Bates (1988), we estimate σ^2 by

$$\hat{\sigma}^2(\boldsymbol{\beta}, \boldsymbol{L}) = \frac{1}{N-p} \sum_{i=1}^{n} (\boldsymbol{Y_i} - \boldsymbol{X_i}\boldsymbol{\beta})^T \boldsymbol{V_i}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i}\boldsymbol{\beta}). \tag{12}$$

We substitute this expression into $\ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2)$ to obtain the doubly regularized profile log-REML, which is

$$Q_R(\boldsymbol{\beta}, \boldsymbol{L}) = P_R(\boldsymbol{\beta}, \boldsymbol{L}) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\boldsymbol{L}_{(k)}\|_2, \tag{13}$$

where

$$\begin{aligned} P_R(\boldsymbol{\beta}, \boldsymbol{L}) \;\; = \;\; & -\frac{1}{2} \sum_{i=1}^{n} \log\Big|\boldsymbol{V}_i\Big| - \frac{1}{2} \log\Big|\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\Big| \\ & - \frac{N-p}{2} \log\left\{ \sum_{i=1}^{n} \Big(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\Big)^T \boldsymbol{V_i}^{-1} \Big(\boldsymbol{Y_i} - \boldsymbol{X_i}\boldsymbol{\beta}\Big) \right\}. \end{aligned} \tag{14}$$

The estimation of $\boldsymbol{\beta}$ and $\boldsymbol{L}$ can be obtained through an iterative algorithm: we first fix $\boldsymbol{L}$ and estimate $\boldsymbol{\beta}$, then we fix $\boldsymbol{\beta}$ and estimate $\boldsymbol{L}$; we iterate between these two steps until the algorithm converges. Since the value of the objective function (13) decreases over iterations, convergence is guaranteed.

When $\boldsymbol{L}$ is fixed, maximizing (13) 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 (13) with respect to $\boldsymbol{L}$ is challenging. Following the same spirit as Lin and Zhang (2006), we transform the optimization to an equivalent problem that is easier to solve.

Proposition 1 *For any given $\hat{\boldsymbol{\beta}}$ and λ_2, consider the following two optimization problems:*

$$\max_{L_{kj}} Q_1(\hat{\boldsymbol{\beta}}, \boldsymbol{L}) = P_R(\hat{\boldsymbol{\beta}}, \boldsymbol{L}) - \lambda_2 \sum_{k=2}^{q} \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}, \tag{15}$$

$$\max_{L_{kj},\gamma_k} Q_2(\hat{\boldsymbol{\beta}}, \boldsymbol{L}) = P_R(\hat{\boldsymbol{\beta}}, \boldsymbol{L}) - \sum_{k=2}^{q} \gamma_k^2 - \frac{\lambda_2^2}{4} \sum_{k=2}^{q} \frac{1}{\gamma_k^2} \Big(\sum_{j=1}^{q} L_{kj}^2 \Big). \tag{16}$$

Let $\hat{L}_{kj}$ be the maximizer of (15), and (γ_k^*, L_{kj}^*) be the maximizer of (16), $k = 2, \ldots, q, j = 1, \ldots, q$. Then we have

$$\hat{L}_{kj} = L_{kj}^*, \; k = 2, \ldots, q, j = 1, \ldots, q; \tag{17}$$

$$\gamma_k^* = \sqrt{\frac{\lambda_2}{2} \|\boldsymbol{L}_{(k)}^*\|_2}, \; k = 2, \ldots, q. \tag{18}$$

The proof of Proposition 1 is given in the Supplemental Material. This proposition suggests that, instead of maximizing (15) with respect to $\boldsymbol{L}$ directly, one can maximize (16) iteratively between γ_k and L_{kj}. Note that when γ_k is fixed, the objective function (16) 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 (18). 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.

2. Update L_{kj}: For iteration r, let

$$L_{kj}^{(r)} = \arg\max_{L_{kj}} P_R(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{D}) - \frac{\lambda_2^2}{4} \sum_{k=1}^{q} \frac{1}{\Big(\gamma_k^{(r-1)}\Big)^2} \Big(\sum_{j=1}^{k} L_{kj}^2 \Big). \tag{19}$$

3. Update γ_k:

$$\gamma_k^{(r)} = \sqrt{\frac{\lambda_2}{2} \|\boldsymbol{L}_{(k)}^{(r)}\|_2}. \tag{20}$$

4. Update $\boldsymbol{\beta}$ by LASSO:

$$\boldsymbol{\beta}^{(r)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} \Big(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta} \Big)^T \boldsymbol{V}_i^{(r)^{-1}} \Big(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta} \Big) + \lambda_1 \sum_{j=1}^{p} |\beta_j|. \tag{21}$$

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 some large-sample properties for the proposed method. Proofs are given in the Supplemental Material.

4.1 Main Results

Our main results are established on general penalty functions, including the $J_1(\boldsymbol{\beta})$ given in (9) and $J_2(\boldsymbol{L})$ given in (10) as special cases. Denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, vec(\boldsymbol{L})^T, \sigma^2)^T$. Consider the doubly regularized log-REML function of the following form:

$$Q_n(\boldsymbol{\theta}) \ = \frac{1}{n}\ell_n(\boldsymbol{\theta}) - \sum_{j=1}^{p} f_{\lambda_{1n}}(|\beta_j|) - \sum_{k=2}^{q} g_{\lambda_{2n}}(|L_{k1}|, \ldots, |L_{kq}|), \tag{22}$$

where $\ell_n(\boldsymbol{\theta})$ is given by (7), and both penalty functions $f_{\lambda_{1n}}(|\beta_j|)$ and $g_{\lambda_{2n}}(|L_{k1}|, \ldots, |L_{kq}|)$ are specified in general forms with tuning parameters λ_{1n} and λ_{2n} being allowed to change with sample size n. Moreover, the penalty functions are assumed to satisfy the following conditions of convexity and monotonicity:

1. $f_{\lambda_{1n}}(|a|) \geq 0$, for $\forall a \in \mathbb{R}$, with $f_{\lambda_{1n}}(0) = 0$, and $f_{\lambda_{1n}}(|a_1|) \leq f_{\lambda_{1n}}(|a_2|)$, if $|a_1| \leq |a_2|$.

2. $g_{\lambda_{2n}}(|a_1|, \ldots, |a_q|) \geq 0$, for $\forall (a_1, \ldots, a_q)^T \in \mathbb{R}^q$, with $g_{\lambda_{2n}}(\mathbf{0}) = 0$, and $g_{\lambda_{2n}}(|a_1|, \ldots, |a_q|) \leq g_{\lambda_{2n}}(|b_1|, \ldots, |b_q|)$, if $|a_l| \leq |b_l|$, $\forall\ l = 1, \ldots, q$.

We use $\boldsymbol{\theta}^*$ to denote the true parameter vector $\boldsymbol{\theta}^* = ({\boldsymbol{\beta}^*_{\mathcal{A}}}^T, {\boldsymbol{\beta}^*_{\mathcal{B}}}^T, vec(\boldsymbol{L}^*_{\mathcal{C}})^T, vec(\boldsymbol{L}^*_{\mathcal{D}})^T, \sigma^2_*)^T$, where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ are index sets, defined as

$$\begin{aligned} \mathcal{A} &= \{j : \beta^*_j \neq 0\}, \\ \mathcal{B} &= \{j : \beta^*_j = 0\}, \\ \mathcal{C} &= \{(k, j) : {L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2 \neq 0\}, \\ \mathcal{D} &= \{(k, j) : {L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2 = 0\}. \end{aligned}$$

Note that $\mathcal{A}$ contains the indices of coefficients for fixed effects which are truly non-zero, $\mathcal{B}$ contains the indices of coefficients for fixed effects which are truly zero, $\mathcal{C}$ contains the indices of elements in $\boldsymbol{L}$ whose row (and the variance of the corresponding random effect component) is truly non-zero, and $\mathcal{D}$ contains the indices of elements in $\boldsymbol{L}$ whose row (and the variance of the corresponding random effect component) is truly zero.

In addition to the common regularity conditions for MLE (Lehmann and Casella, 1998) and REML (Jiang, 2007), we also assume the following regularity conditions:

$$A1: \lim_{n\to\infty}\frac{1}{n}\boldsymbol{X_i}^T(\sigma_*^2\boldsymbol{V_i}(\boldsymbol{L}^*))^{-1}\boldsymbol{X_i} = \boldsymbol{I}(\boldsymbol{\beta}^*),\ \text{where } \boldsymbol{I}(\boldsymbol{\beta}^*) \text{ is a positive definite matrix.}$$

$$A2: \frac{1}{n}\frac{\partial^3 \ell_n(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k} = O_p(1),\ \text{for all } \boldsymbol{\theta} \text{ in a small neighborhood of } \boldsymbol{\theta}^*.$$

$$A3: \text{For any two bounded column vectors } \boldsymbol{a} \text{ and } \boldsymbol{b},\ \text{and for any two rows of } \boldsymbol{X_i}: \boldsymbol{X_{i(j)}} \text{ and } \boldsymbol{X_{i(j')}},$$
$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X_{i}}_{(j)}^{T}\boldsymbol{a})(\boldsymbol{X_{i}}_{(j')}^{T}\boldsymbol{b})^T = O_p(1).$$

Condition $A1$ guarantees that the Fisher information matrix for $\boldsymbol{\beta}$ exists and is positive definite. Condition $A2$ implies that the third and higher order expansions of ℓ_n are ignorable. Condition $A3$ indicates that the product of two bounded linear combinations of two rows of the design matrix $\boldsymbol{X}$ is bounded. Note that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are chosen to be column vectors with only 1 element being nonzero, then condition $A3$ is equivalent to the common condition $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i = O_p(1)$ for a linear model.

We also define

$$\begin{aligned}
a_n &= \max\Big\{\frac{\partial f_{\lambda_{1n}}(|\beta_j^*|)}{\partial|\beta_j|} : \beta_j^* \neq 0\Big\},\\
b_n &= \max\Big\{\frac{\partial^2 f_{\lambda_{1n}}(|\beta_j^*|)}{\partial|\beta_j|^2} : \beta_j^* \neq 0\Big\},\\
c_n &= \max\Big\{\frac{\partial g_{\lambda_{2n}}(|L_{k1}^*|,\ldots,|L_{kq}^*|)}{\partial|L_{kj}|} : {L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 \neq 0\Big\},\\
d_n &= \max\Big\{\frac{\partial^2 g_{\lambda_{2n}}(|L_{k1}^*|,\ldots,|L_{kq}^*|)}{\partial|L_{kj}|\partial|L_{kj'}|} : {L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 \neq 0\Big\}.
\end{aligned}$$

Then we have the theorem that contains the result on estimation consistency.

Theorem 2 *Suppose conditions A1-A3 hold. If both a_n and c_n are of order $O(n^{-1/2})$, and both b_n and d_n are of order $o(1)$, then there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of $Q_n(\boldsymbol{\theta})$ in (22) such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_p(n^{-1/2})$.*

Theorem 2 implies that by choosing proper penalty functions $f_{\lambda_{1n}}$ and $g_{\lambda_{2n}}$ as well as proper tuning parameters λ_{1n} and λ_{2n}, the doubly regularized REML estimator is root-n consistent. It immediately leads to the following corollary for the doubly regularized REML proposed in Section 2.3.

Corollary 1 *Consider the following penalty functions:*

$$\begin{aligned} f_{\lambda_{1n}}(|\beta_j|) &= \lambda_{1n}|\beta_j|, \\ g_{\lambda_{2n}}(|L_{k1}|,\ldots,|L_{kq}|) &= \lambda_{2n}\sqrt{L_{k1}^2+\cdots+L_{kq}^2},\ k \geq 2. \end{aligned}$$

If $\lambda_{1n} = O(n^{-1/2})$ and $\lambda_{2n} = O(n^{-1/2})$, then there exists a $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)$ of $Q_n(\boldsymbol{\theta})$ given in (11).

Below we establish the sparsity property and the asymptotic normality.

Theorem 3 *Suppose $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, vec(\hat{\boldsymbol{L}})^T, \hat{\sigma}^2)^T$ is a $\sqrt{n}$-consistent local maximizer of $Q_n(\boldsymbol{\theta})$ in (22). Under conditions A1-A3, we have the following results:*

(a) For all $j \in \mathcal{B}$ (i.e. $\beta_j^* = 0$), if $\sqrt{n}\frac{\partial f_{\lambda_{1n}}(|\hat{\beta}_j|)}{\partial|\beta_j|} \to \infty$, then $Pr(\hat{\beta}_j = 0) \to 1$ as $n \to \infty$.

(b) For all $(k,j) \in \mathcal{D}$ (i.e. ${L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 = 0$), if $\sqrt{n}\frac{\partial g_{\lambda_{2n}}(|\hat{L}_{k1}|,\ldots,|\hat{L}_{kq}|)}{\partial|L_{kj}|} \to \infty$, then $Pr(\hat{L}_{kj} = 0) \to 1$ as $n \to \infty$.

(c) If for all $j \in \mathcal{A}$ (i.e. $\beta_j^* \neq 0$), $\sqrt{n}\frac{\partial f_{\lambda_{1n}}(|\beta_j^*|)}{\partial|\beta_j|} \to 0$, and $b_n = o(1)$, then under part (a), we have $\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\Big) \xrightarrow{d} MVN\big(\mathbf{0}, \boldsymbol{I}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}^*_{\mathcal{A}})\big)$, $n \to \infty$, where $\boldsymbol{I}_{\mathcal{A}}$ is the part of the (full) Fisher Information matrix corresponding to the parameter subvector $\boldsymbol{\beta}_{\mathcal{A}}$.

(d) If for all $(k,j) \in \mathcal{C}$ (i.e. ${L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 \neq 0$), $\sqrt{n}\frac{\partial g_{\lambda_{2n}}(|L_{k1}^*|,\ldots,|L_{kq}^*|)}{\partial|L_{kj}|} \to 0$, and $d_n = o(1)$, then under part (b), we have $\sqrt{n}\Big(\hat{\boldsymbol{L}}_{\mathcal{C}} - \boldsymbol{L}^*_{\mathcal{C}}\Big) \xrightarrow{d} MVN\big(\mathbf{0}, \boldsymbol{I}_{\mathcal{C}}^{-1}(\boldsymbol{L}^*_{\mathcal{C}})\big)$, $n \to \infty$, where $\boldsymbol{I}_{\mathcal{C}}$ is the part of the (full) Fisher Information matrix corresponding to the parameter subvector $vec(\boldsymbol{L}_{\mathcal{C}})$.

Theorem 3 implies that by choosing proper penalty functions $f_{\lambda_{1n}}$ and $g_{\lambda_{2n}}$, as well as proper tuning parameters λ_{1n} and λ_{2n}, the doubly regularized REML estimators hold the sparse property for the zero parameters indexed by $\mathcal{B}$ and $\mathcal{D}$; that is, with probability tending to 1, $\hat{\boldsymbol{\beta}}_{\mathcal{B}} = \mathbf{0}$ and $\hat{\boldsymbol{L}}_{\mathcal{D}} = \mathbf{0}$. Moreover, the doubly regularized REML estimators for the nonzero parameters, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ and $\hat{\boldsymbol{L}}_{\mathcal{C}}$, follow the same asymptotic distributions as they would follow if the zero parameters were known in advance. Therefore, we can declare that asymptotically, the doubly regularized REML estimators perform as well as if the true underlying model were provided in advance; in other words, the proposed doubly regularized REML estimation method possesses the oracle property of Fan and Li (2001).

4.2 Improving Regularized REML Regression

Though Corollary 1 indicates that, when λ_{1n} and λ_{2n} are properly selected, there exists a root-n consistent estimate for the doubly regularized REML regression (11), the sparse property, however, may not hold for (11), i.e., there is no guarantee that $\hat{\boldsymbol{\beta}}_{\mathcal{B}} = \mathbf{0}$ or $\hat{\boldsymbol{L}}_{\mathcal{D}} = \mathbf{0}$ with probability approaching 1. To overcome this limitation, we employ the idea of adaptive regularization that has been used in the literature, for example, Breiman (1995), Wang et al. (2007), Zhang and Lu (2007), Zou (2006), among others. Essentially, the adaptive idea allocates different penalty weights on different parameters. Specifically, we propose a modified version of the doubly regularized REML function given as follows:

$$Q_n^W(\boldsymbol{\theta}) = \frac{1}{n}\ell_n(\boldsymbol{\theta}) - \lambda_{1n}\sum_{j=1}^{p} w_{nj}^{\beta}|\beta_j| - \lambda_{2n}\sum_{k=2}^{q} w_{nk}^{L}\sqrt{L_{k1}^2 + \cdots + L_{kq}^2}, \tag{23}$$

where $w_{nj}^{\beta} \geq 0, j = 1, \ldots, p$ and $w_{nk}^{L} \geq 0,\ k = 2, \ldots, q$ are pre-specified non-negative weights. The intuition behind this modification is that if a fixed effect or a random effect appears strong, its associated regularization weight should be small, so that the corresponding regression coefficient or variance component will be lightly penalized. On the other hand, if a fixed effect or a random effect appears weak, its associated regularization weight should be large, hence the corresponding regression coefficient or variance component is heavily penalized. With a proper choice for the adaptive weights, the weighted doubly regularized REML regression possesses the oracle property as in Theorem 3. The details are stated in the following theorem.

Theorem 4 *Define*

$$w_{n,max}^{\beta} = \max\{w_{nj}^{\beta} : \beta_j^* \neq 0\},\ w_{n,max}^{L} = \max\{w_{nk}^{L} : {L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 \neq 0\},$$
$$w_{n,min}^{\beta} = \min\{w_{nj}^{\beta} : \beta_j^* = 0\},\ \ w_{n,min}^{L} = \min\{w_{nk}^{L} : {L_{k1}^*}^2 + \cdots + {L_{kq}^*}^2 = 0\}.$$

Under conditions A1-A3, if $\sqrt{n}\lambda_{1n}w_{n,max}^{\beta} = O_p(1)$*,* $\sqrt{n}\lambda_{1n}w_{n,min}^{\beta} \to \infty$*,* $\sqrt{n}\lambda_{2n}w_{n,max}^{L} = O_p(1)$*, and* $\sqrt{n}\lambda_{2n}w_{n,min}^{L} \to \infty$*, then there exists a* $\sqrt{n}$*-consistent local maximizer* $\hat{\boldsymbol{\theta}}$ *of (23) such that*

$Pr(\hat{\boldsymbol{\beta}}_{\mathcal{B}} = \mathbf{0}) \to 1$ and $Pr(\hat{\boldsymbol{L}}_{\mathcal{D}} = \mathbf{0}) \to 1$ as $n \to \infty$. Furthermore, if $\sqrt{n}\lambda_{1n} w^{\beta}_{n,max} = o_p(1)$ and $\sqrt{n}\lambda_{2n} w^{L}_{n,max} = o_p(1)$, then we have $\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\Big) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{I}^{-1}_{\mathcal{A}}(\boldsymbol{\beta}^*_{\mathcal{A}}))$ and $\sqrt{n}\Big(vec(\hat{\boldsymbol{L}}_{\mathcal{C}}) - vec(\boldsymbol{L}^*_{\mathcal{C}})\Big) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{I}^{-1}_{\mathcal{C}}(\boldsymbol{L}^*_{\mathcal{C}}))$ as $n \to \infty$.

The following corollary provides one set of choices for proper tuning parameters λ_{1n} and λ_{2n} as well as proper weights w^{β}_{nj} and w^{L}_{nk}, which satisfy the conditions required in Theorem 4.

Corollary 2 *Let $\tilde{\beta}_j$ and $\tilde{L}_{kj}$ be n^{τ}-consistent estimators with $0 < \tau \leq 0.5$. If $\lambda_{1n} = \lambda_{2n} = 1/\{\sqrt{n}\log(n)\}$, $w^{\beta}_{nj} = 1/|\tilde{\beta}_j|^{r_1}$, $j = 1, \ldots, p$, and $w^{L}_{nk} = 1/(\tilde{L}^2_{k1} + \cdots + \tilde{L}^2_{kq})^{r_2}$, $k = 2, \ldots, q$, with $r_1 > 0, r_2 > 0$, then there exists a $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\theta}}$ of (23) such that as $n \to \infty$,*

$$Pr(\hat{\boldsymbol{\beta}}_{\mathcal{B}} = \mathbf{0}) \to 1, \ \sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\Big) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{I}^{-1}_{\mathcal{A}}(\boldsymbol{\beta}^*_{\mathcal{A}}))$$
$$Pr(\hat{\boldsymbol{L}}_{\mathcal{D}} = \mathbf{0}) \to 1, \ \sqrt{n}\Big(vec(\hat{\boldsymbol{L}}_{\mathcal{C}}) - vec(\boldsymbol{L}^*_{\mathcal{C}})\Big) \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{I}_{\mathcal{C}}(\boldsymbol{L}^*_{\mathcal{C}})^{-1}).$$

In practice, we may choose $\tilde{\beta}_j$ and $\tilde{L}_{kj}$ as the consistent estimates from the unpenalized LMM when $p < n$, and ridge LMM regression when $p > n$.

5 Simulation Studies

In this section, we report simulation results concerning the performance of the doubly regularized REML estimation. We considered four examples. In each example, there are n=200 clusters, with m_i=5 repeated observations in each cluster. The LMM used to generate data is detailed as follows.

- Example 1: There are p=6 predictors. Data are generated from the following LMM:

$$\begin{aligned} Y_{ij} &= 1 + 2X_{ij,1} + 2X_{ij,2} + 2X_{ij,3} + 0X_{ij,4} + 0X_{ij,5} + 0X_{ij,6} \\ &\quad + b_{i0} + b_{i1}X_{ij,1} + b_{i3}X_{ij,3} + \epsilon_{ij}, \ i = 1, \ldots, 200, \ j = 1, \ldots, 5, \end{aligned}$$

  where $X_{ij,1} \sim N(0, 2^2)$, $X_{ij,2} = X_{i2} \sim$ Bernoulli(0.5), $X_{ij,3} = j$, and the other three predictors $X_{ij,4}, X_{ij,5}, X_{ij,6}$ are independent $N(0,1)$ variables; random effects b_{i0}, b_{i1}, b_{i3} are independently generated from $N(0, 0.5^2)$; and errors ϵ_{ij} are i.i.d. $N(0,1)$.

- Example 2: The LMM is the same as that in Example 1, except for nonzero correlations among the random effects, which are given as: $b_{i0}, b_{i1}, b_{i3} \sim N(0, 0.5^2)$, corr$(b_{i0}, b_{i1}) = 0.5$, corr$(b_{i0}, b_{i3}) = 0.2$, and corr$(b_{i1}, b_{i3}) = 0.3$.

- Example 3: There are p=8 predictors that are serially correlated. The LMM that generates the data is as follows:

$$\begin{aligned} Y_{ij} &= 1 + 3X_{ij,1} + 1.5X_{ij,2} + 0X_{ij,3} + 0X_{ij,4} + 2X_{ij,5} + 0X_{ij,6} + 0X_{ij,7} + 0X_{ij,8} \\ &\quad + b_{i0} + b_{i1}X_{ij,1} + b_{i5}X_{ij,5} + \epsilon_{ij},\ i = 1, \ldots, 200,\ j = 1, \ldots, 5, \end{aligned}$$

  where $X_{ij,k} \sim N(0,1)$, $k = 1, \ldots, 8$, with $\text{corr}(X_{ij,k}, X_{ij,k'}) = 0.5^{|k-k'|}$; random effects b_{i0}, b_{i1}, b_{i5} are independent according to $N(0, 0.8^2)$, and errors ϵ_{ij} are i.i.d. $N(0, 1)$.

- Example 4: The model is the same as that in Example 3, except for correlated random effects: $b_{i0}, b_{i1}, b_{i5} \sim N(0, 0.8^2)$, $\text{corr}(b_{i0}, b_{i1}) = 0.5$, $\text{corr}(b_{i0}, b_{i5}) = 0.2$, and $\text{corr}(b_{i1}, b_{i5}) = 0.3$.

When fitting the model, we included all the predictors in both the fixed effects component ($p = 8$) and the random effects component ($q = 8$) plus the random intercept. We applied both the non-adaptive doubly regularized REML regression and its adaptive version to select important effects. Following Wang et al. (2007), we selected tuning parameters λ_1 and λ_2 by minimizing the BIC criterion:

$$\text{BIC} = -2P_R(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{L}}) + d\log(n), \tag{24}$$

where d is the total number of nonzero estimates in $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{L}})$.

For each example, we repeated for 200 times. The results are summarized in Tables 1-4. In particular, we recorded the selection frequency of fixed effects and random effects, and calculated average estimates of regression coefficients and variance components. Empirical standard errors of the average estimates are also reported. Since the random intercept was always included in the model, we omit the corresponding selection frequency.

As we can see from the tables, in all four simulation settings, the non-adaptive doubly regularized REML was very effective at identifying important fixed and random effects and reasonably effective in removing unimportant ones. We can also see that the non-adaptive doubly regularized REML method had very little bias in estimation for the fixed effects, but there was noticeable downward bias in the estimation of the variance components (upper parts of Tables 1-4). This bias, however, was reduced significantly by the adaptive doubly regularized REML method (lower parts of Tables 1-4). The adaptive version of the doubly regularized REML was also more effective at removing unimportant fixed and random effects than the non-adaptive method. In conclusion, both versions of doubly regularized REML are effective to identify positive signals and useful in the building of prediction models. However, for the purpose of discovery, the adaptive version is recommended, since it appears to have a better control of false discoveries than the non-adaptive vesion.

6 Data Analysis

In this section, we apply the proposed method in a real world data analysis. The data were collected from a longitudinal randomized controlled intervention trial on 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 6 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 race (1 if the subject is Hispanic and 0 otherwise), time of visit (logarithm of year), and season of visit. Seasonality was coded according to three different periods: Winter refers to November through February, Spring corresponds to March through June, and Summer represents July through October. 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 time, gender, or Hispanic. Thus, the LMM for the data analysis takes the following form:

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

where these 10 predictors were included in both $\boldsymbol{X}_i$ for fixed effects and $\boldsymbol{Z}_i$ for random effects, that is, $p = 10$ and $q = 11$ (one for the random intercept). The treatment effect was evaluated by the interaction between Time and Treatment in terms of whether there is a difference in the trend of changes of BSI in control and treatment groups. We fitted the model using the non-adaptive doubly regularized REML regression and selected tuning parameters using the BIC in (24). The results are summarized in Table 5 (left part). As we can see, our 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.

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 6.

We can see that Time, Summer, Winter, Time*Treatment and Gender*Treatment had high selection frequencies while Hispanic had low selection frequency among the fixed effects; regarding the random effects, Time, Summer, Winter and Time*Treatment had high selection frequencies.

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 5 (right part). As we can see, 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 6, 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 some evidence that the treatment program was effective and the program worked better for boys than 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 fitted coefficients for Winter and Summer also indicate 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.

7 Discussion

We have proposed a doubly regularized REML to select important fixed effects and random effects simultaneously. We have shown that an adaptive version of the doubly regularized REML enjoys the oracle property; that is, it performs as well as if the true model structure were given in advance. Numerical results indicate that our methods work well for the selection of both fixed and random effects. There is a downward bias in the estimation of the variance components, however, it can be reduced by the adaptive method.

We would like to note that our choice of REML is rooted in the fact that the REML method is more popular and superior over many other methods for estimation and inference in the LMM. For example, EM may be an alternative for estimation in LMM; however, since the regularization shrinks the number of random effects, the dimension of the posterior distribution of the random effects may vary from iteration to iteration. In such situations, it is not clear whether the EM algorithm would still converge. The slow convergence rate of the EM also limits its capability for handling a large number of random effects, the scenario where the proposed method intends to be effective. Furthermore, the doubly regularized REML can be naturally extended to the generalized linear mixed-effects models (GLMM) via the Laplace approximation (e.g. Breslow and Clayton, 1993), which is currently being investigated by the authors.

Another direction for future work arises from the possible hierarchy between fixed effects and random effects. That is, 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 doubly regularized REML can be easily generalized to handle this constraint.

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

$$\boldsymbol{D} = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_q \end{pmatrix} \boldsymbol{L}\boldsymbol{L}^T \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_q \end{pmatrix}, \tag{25}$$

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

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

$$(\hat{\boldsymbol{\beta}}, \hat{L}_{ij}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{L}} P_R - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \|\boldsymbol{L}_{(k)}\|_2. \tag{26}$$

As pointed out above, if $\hat{\beta}_j = 0$, from (25) the penalty on $\boldsymbol{L}$ will guarantee that $\hat{\boldsymbol{L}}_{(j)}$ 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 (26) with a slight modification.

Acknowledgements

The research of Song and Zhu are supported in part by their respective NSF grants.

References

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.

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.

Breiman, L. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**, 373-384.

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

Chen, Z. and Dunson, D.B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, **59**, 762-769.

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.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407-499.

Fan, J., and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

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.

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.

Harville, D.A. (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, **61**, 383-385.

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320-338.

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

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics *The Annals of Statistics*, **24**, 255-286.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

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

Laird, N.M. and Ware, J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974.

Lan, L. (2006). Variable Selection in Linear Mixed Model for Longitudinal Data. Ph.D. dissertation, Department of Statistics, North Carolina State University.

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.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

Lin, X. (1997). Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika*, **84**, 309-326.

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

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.

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.

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

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

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.

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

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

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

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

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.

Zhang, H. and Lu, W. (2007) Adaptive-Lasso for Cox's Proportional Hazards Model. *Biometrika*, **94**, 691-703.

Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

Table 1: Simulation results for Example 1. The upper part is for the non-adaptive method, and the lower part is for the adaptive method. “Sel. Freq.” represents the selection frequency over 200 repetitions. Averaged estimates over 200 repetitions and the corresponding standard errors (numbers in the parentheses) are also reported.

	Intercept	X_1	X_2	X_3	X_4	X_5	X_6
Non-adaptive DRLMM							
Fixed Effects							
Sel. Freq. (%)	—	100	100	100	21	25	21
$\hat{\beta}_j$	1.14	1.96	1.82	1.98	0.002	-0.001	0.001
	(0.13)	(0.05)	(0.20)	(0.04)	(0.03)	(0.03)	(0.03)
Random Effects							
Sel. Freq. (%)	—	100	5.5	100	17	17.5	18.5
$\sqrt{\hat{D}_{kk}}$	—	0.37	0.004	0.37	0.005	0.004	0.005
	(—)	(0.04)	(0.02)	(0.04)	(0.01)	(0.01)	(0.01)
Adaptive DRLMM							
Fixed Effects							
Sel. Freq. (%)	—	100	100	100	4	3.5	3.5
$\hat{\beta}_j$	1.05	2.00	1.92	1.99	0.001	-0.001	0.001
	(0.12)	(0.04)	(0.17)	(0.05)	(0.01)	(0.01)	(0.02)
Random Effects							
Sel. Freq. (%)	—	100	10	100	0.5	0	0
$\sqrt{\hat{D}_{kk}}$	—	0.46	0.03	0.45	0.0001	0	0
	—	(0.046)	(0.058)	(0.050)	(0.001)	(0)	(0)

Table 2: Simulation results for Example 2. Descriptions are referred to the caption of Table 1.

	Intercept	X_1	X_2	X_3	X_4	X_5	X_6
Non-adaptive DRLMM							
Fixed Effects							
Sel. Freq. (%)	—	100	100	100	21	24.5	23
$\hat{\beta}_j$	1.14	1.96	1.80	1.98	0.0001	0.0004	0.0023
	(0.16)	(0.05)	(0.21)	(0.05)	(0.03)	(0.02)	(0.03)
Random Effects							
Sel. Freq. (%)	—	100	4.4	100	20	16.7	16.7
$\sqrt{\hat{D}_{kk}}$	—	0.38	0.003	0.37	0.005	0.005	0.004
	—	(0.04)	(0.02)	(0.04)	(0.01)	(0.01)	(0.01)
Adaptive DRLMM							
Fixed Effects							
Sel. Freq. (%)	—	100	100	100	7.5	6.5	6.0
$\hat{\beta}_j$	1.04	1.99	1.93	2.00	0.0004	0.0002	-0.001
	(0.12)	(0.04)	(0.19)	(0.05)	(0.02)	(0.01)	(0.01)
Random Effects							
Sel. Freq. (%)	—	100	5.5	100	0	1	0
$\sqrt{\hat{D}_{kk}}$	—	0.46	0.02	0.45	0	0.001	0
	—	(0.046)	(0.038)	(0.051)	(0)	(0.008)	(0)

Table 3: Simulation results for Example 3. Descriptions are referred to the caption of Table 1.

	Intercept	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Non-adaptive DRLMM									
Fixed Effects									
Sel. Freq. (%)	—	100	100	17	13.5	100	13.5	9	7.5
$\hat{\beta}_j$	1.00	2.91	1.46	0.009	0.009	1.89	0.007	0.004	0.001
	(0.08)	(0.09)	(0.05)	(0.02)	(0.02)	(0.09)	(0.02)	(0.02)	(0.02)
Random Effects									
Sel. Freq. (%)	—	98	22	13.5	16.5	91	18.5	11.5	12
$\sqrt{\hat{D}_{kk}}$	—	0.54	0.009	0.006	0.007	0.50	0.009	0.004	0.006
	(—)	(0.12)	(0.02)	(0.02)	(0.02)	(0.19)	(0.02)	(0.01)	(0.02)
Adaptive DRLMM									
Fixed Effects									
Sel. Freq. (%)	—	100	100	6.5	2.0	100	3.0	5.0	2.5
$\hat{\beta}_j$	1.00	3.00	1.49	0.0002	-0.0001	1.98	0.0003	0.0005	0.0001
	(0.077)	(0.077)	(0.047)	(0.012)	(0.008)	(0.073)	(0.009)	(0.008)	(0.005)
Random Effects									
Sel. Freq. (%)	—	100	2.5	4.0	4.0	100	6.5	4.0	2.0
$\sqrt{\hat{D}_{kk}}$	—	0.75	0.002	0.003	0.003	0.75	0.005	0.004	0.002
	(—)	(0.08)	(0.01)	(0.01)	(0.01)	(0.08)	(0.02)	(0.01)	(0.01)

Table 4: Simulation results for Example 4. Descriptions are referred to the caption of Table 1.

	Intercept	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Non-adaptive DRLMM									
Fixed Effects									
Sel. Freq. (%)	—	100	100	18	17.5	100	19.5	15	12
$\hat{\beta}_j$	1.00	2.91	1.47	0.007	0.007	1.89	0.008	0.001	-0.0004
	(0.07)	(0.08)	(0.05)	(0.03)	(0.02)	(0.10)	(0.03)	(0.03)	(0.02)
Random Effects									
Sel. Freq. (%)	—	100	21	7.5	19.5	99	12	10	7.5
$\sqrt{\hat{D}_{kk}}$	—	0.50	0.007	0.003	0.006	0.49	0.004	0.004	0.003
	(—)	(0.12)	(0.02)	(0.01)	(0.02)	(0.16)	(0.02)	(0.02)	(0.01)
Adaptive DRLMM									
Fixed Effects									
Sel. Freq. (%)	—	100	100	5.5	4.0	100	7.5	5.5	4.0
$\hat{\beta}_j$	1.00	2.98	1.49	0.0001	0.0002	1.98	0.0006	0.0021	0.0002
	(0.067)	(0.071)	(0.048)	(0.014)	(0.009)	(0.075)	(0.020)	(0.020)	(0.006)
Random Effects									
Sel. Freq. (%)	—	100	3.0	4.5	3.5	100	4.0	7.0	1.0
$\sqrt{\hat{D}_{kk}}$	—	0.76	0.002	0.004	0.004	0.76	0.003	0.005	0.002
	(—)	(0.08)	(0.01)	(0.02)	(0.01)	(0.08)	(0.01)	(0.01)	(0.01)

Table 5: Results for the psychiatric symptom data analysis. The numbers are estimated fixed effects $\hat{\beta}_j$'s and the estimated variance components of random effects $\sqrt{\hat{D}_{kk}}$'s.

	Non-Adaptive		Adaptive	
	Fixed Effect	Variance Component	Fixed Effect	Variance Component
Age at baseline	0	0	0	0
Gender	0	0	0	0
Hispanic	0.010	0	0	0
Time	−0.060	0.142	-0.069	0.182
Summer	−0.045	0.014	-0.033	0.033
Winter	−0.041	0.010	-0.029	0.029
Treatment	0	0	0	0
Time*Trt	−0.027	0.002	-0.006	0.005
Gender*Trt	0.075	0	0.065	0
Hispanic*Trt	0	0	0	0

Table 6: 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)

Supplemental Material

Proof for Proposition 1 First, (18) can be obtained by using the inequality $a^2+b^2 \geq 2ab$. Next, we prove $\hat{L}_{kj} = L^*_{kj}$.

Denote P and Q be the objective functions corresponding to the two optimization problems:

$$\begin{aligned} P &= -P_R + \lambda_2 \sum_{k=2}^{q} \|L_{(k)}\|_2 \\ Q &= -P_R + \sum_{k=2}^{q} \gamma_k^2 + \frac{\lambda_2^2}{4} \sum_{k=2}^{q} \frac{1}{\gamma_k^2}\Big(\sum_{j=1}^{k} L_{kj}^2\Big) \end{aligned}$$

After some algebra, we can see that $P(L^*_{kj}) = Q(\gamma^*_k, L^*_{kj})$, so $P(\hat{L}_{kj}) \leq Q(\gamma^*_k, L^*_{kj})$. Then let $\hat{\gamma}_k = \sqrt{\frac{\lambda_2}{2}\|\hat{L}_{(k)}\|_2}$. After some algebra, we can see that $Q(\hat{\gamma}_k, \hat{L}_{kj}) = P(\hat{L}_{kj})$, so $Q(\gamma^*_k, L^*_{kj}) \leq P(\hat{L}_{kj})$. Therefore, $P(\hat{L}_{kj}) = Q(\gamma^*_k, L^*_{kj}) = Q(\hat{\gamma}_i, \hat{L}_{kj})$. Since the objective function Q is convex, so the minimizer is unique, then we have $L^*_{kj} = \hat{L}_{kj}$.

In order to prove Theorem 2 and Theorem 3, we need the log REML function $\ell_n(\boldsymbol{\theta})$ has several properties, which are stated in the following lemma.

Lemma 2 *Denote*

$$\begin{aligned} \ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) &= -\frac{1}{2}\sum_{i=1}^{n} \log\Big|\sigma^{-2}\boldsymbol{V_i}\Big| - \frac{1}{2}\log\Big|\sigma^{-2}\sum_{i=1}^{n} \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\Big| \\ & \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big\{\boldsymbol{Y_i} - \boldsymbol{X_i\beta}\Big\}^T \boldsymbol{V_i}^{-1}\Big\{\boldsymbol{Y_i} - \boldsymbol{X_i\beta}\Big\}, \\ \ell_R(\tilde{\boldsymbol{\beta}}(\boldsymbol{L}), \boldsymbol{L}, \sigma^2) &= -\frac{1}{2}\sum_{i=1}^{n} \log\Big|\sigma^{-2}\boldsymbol{V_i}\Big| - \frac{1}{2}\log\Big|\sigma^{-2}\sum_{i=1}^{n} \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\Big| \\ & \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big\{\boldsymbol{Y_i} - \boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{L})\Big\}^T \boldsymbol{V_i}^{-1}\Big\{\boldsymbol{Y_i} - \boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{L})\Big\}, \end{aligned}$$

where

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{L}) = \arg\min_{\beta} \sum_{i=1}^{n} (\boldsymbol{Y_i} - \boldsymbol{X_i\beta})^T \boldsymbol{V_i}^{-1}(\boldsymbol{Y_i} - \boldsymbol{X_i\beta}). \tag{27}$$

Denote $\boldsymbol{\tau} = (vec(\boldsymbol{L})^T, \sigma^2)^T$, and $\boldsymbol{I_\beta}$, $\boldsymbol{I_\tau}$ and $\boldsymbol{I_\theta}$ be three positive definite matrices given by

$$\boldsymbol{I_\beta} \quad = \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X_i}^T (\sigma^2 \boldsymbol{V_i})^{-1} \boldsymbol{X_i} \tag{28}$$

$$\boldsymbol{I_\tau} \quad = \quad \begin{pmatrix} \boldsymbol{I_L} & \boldsymbol{a} \\ \boldsymbol{a}^T & I_\sigma \end{pmatrix} = \lim_{n\to\infty} -\frac{1}{n} \frac{\partial^2 \ell_R}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T} \tag{29}$$

$$\boldsymbol{I_\theta} \quad = \quad \begin{pmatrix} \boldsymbol{I_\beta} & \\ & \boldsymbol{I_\tau} \end{pmatrix} \tag{30}$$

Under assumptions $A1 - A3$, we claim that

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = O_p(1);\ \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} \xrightarrow{d} MVN(0, \boldsymbol{I_\beta}(\boldsymbol{\beta}^*));\ \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial vec(\boldsymbol{L})} \xrightarrow{d} MVN(0, \boldsymbol{I_L}(\boldsymbol{L}^*)) \tag{31}$$

$$-\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \xrightarrow{p} \boldsymbol{I_\theta}(\boldsymbol{\theta}^*) \tag{32}$$

Proof : First, under the true parameter $(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)$, from the estimating equation theory, we know that $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator, and we also have

$$\sqrt{n}\Big(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\Big) \xrightarrow{d} \mathrm{MVN}(0, \boldsymbol{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}^*)) \tag{33}$$

Second, the REML function ℓ_R has the following properties (Jiang, 1996):

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_R(\tilde{\boldsymbol{\beta}}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\tau}} \quad \xrightarrow{d} \quad \mathrm{MVN}(0, \boldsymbol{I_\tau}(\boldsymbol{\tau}^*)), \tag{34}$$

$$-\frac{1}{n} \frac{\partial^2 \ell_R(\tilde{\boldsymbol{\beta}}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T} \quad \xrightarrow{p} \quad \boldsymbol{I_\tau}(\boldsymbol{\tau}^*) \tag{35}$$

In order to show the properties of ℓ_n, we decompose ℓ_n to be the summation of ℓ_R and two items, which is

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma) = \ell_R(\tilde{\boldsymbol{\beta}}, \boldsymbol{L}, \sigma) + E_1 + E_2, \tag{36}$$

where

$$E_1 \quad = \quad -\frac{1}{2\sigma^2} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \Big(\sum_{i=1}^{n} \boldsymbol{X_i}^T \boldsymbol{V_i}^{-1} \boldsymbol{X_i} \Big) (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{37}$$

$$E_2 \quad = \quad -\frac{1}{2\sigma^2} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \sum_{i=1}^{n} \boldsymbol{X_i}^T \boldsymbol{V_i}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i} \tilde{\boldsymbol{\beta}}) \tag{38}$$

Consider the first derivatives of E_1 and E_2. After some algebra, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\beta}} &= \Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X_i}^T(\sigma_*^2\boldsymbol{V_i})^{-1}\boldsymbol{X_i}\Big)\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})} &= \frac{1}{2\sigma_*^2}\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \sigma_*^2} &= \frac{1}{2\sigma_*^4}\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\beta}} &= \frac{1}{2\sigma_*^2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}) \\
\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})} &= \frac{1}{2\sigma_*^2}\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\frac{1}{n}\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) \\
&\quad -\frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})} \\
\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \sigma^2} &= \frac{1}{2\sigma_*^4}\frac{1}{\sqrt{n}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}})
\end{aligned}$$

With the assumption $A1-A3$ and equation (33), by using Slutsky theorem, we have

$$\frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\beta}} \rightarrow_d N(0, \boldsymbol{I_\beta});\ \frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})} = o_p(1);\ \frac{1}{\sqrt{n}}\frac{\partial E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \sigma_*^2} = o_p(1);$$

From equation (27), we can see that $\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}) = 0$. Therefore,

$$\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \boldsymbol{\beta}} = 0;\ \frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial \sigma^2} = 0$$

For $\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})}$, consider $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta})$.

Denote $S_i = \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta})$, then we have

$$\begin{aligned}
E(S_i) &= 0 && (39) \\
Var(S_i) &= \boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial \boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i} && (40)
\end{aligned}$$

With assumption $A1-A3$, we have $\frac{1}{n^2}\sum_{i=1}^n Var(S_i) \rightarrow 0$. Then by Chebyshev's LLN, we have $\frac{1}{n}\sum_i S_i \rightarrow_p 0$. Therefore,

$$\frac{1}{\sqrt{n}}\frac{\partial E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})} = o_p(1).$$

Combining the properties of ℓ_R and properties of E_1 and E_2, we have proved:

$$\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = O_p(1);\ \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} \rightarrow_d N(0, \boldsymbol{I_\beta});\ \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial vec(\boldsymbol{L})} \rightarrow_d N(0, \boldsymbol{I_L}).$$

Next, we consider the second derivatives of E_1. After some algebra, we have

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T(\sigma_*^2\boldsymbol{V_i})^{-1}\boldsymbol{X_i} \\
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial vec(\boldsymbol{L})^T} &= -\frac{1}{\sigma_*^2}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\Big)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial\sigma^2} &= -\frac{1}{\sigma_*^4}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\Big)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial\sigma^2} &= -\frac{1}{2\sigma_*^4}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial(\sigma^2)^2} &= -\frac{1}{4\sigma_*^6}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}) \\
\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial vec(\boldsymbol{L})^T} &= \frac{1}{2\sigma_*^2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(-2G_1+G_2)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}),
\end{aligned}$$

where

$$G_1 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i} \tag{41}$$

$$G_2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial^2\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})\partial vec(\boldsymbol{L})^T}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i} \tag{42}$$

With assumptions $A1-A3$ and the consistency of $\tilde{\boldsymbol{\beta}}$, we can see that

$$-\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \boldsymbol{I_{\beta}} + o_p(1),$$

and all other second derivatives of E_1 at true parameters $(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)$ are $o_p(1)$.

Then, we consider the second derivatives of E_2. After some algebra, we have

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= 0 \\
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial vec(\boldsymbol{L})^T} &= -\frac{1}{2\sigma_*^2}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) \\
&\quad -\frac{1}{2\sigma_*^2}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{X_i}\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
&\quad -\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial vec(\boldsymbol{L})^T} \\
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial\boldsymbol{\beta}\partial\sigma^2} &= -\frac{1}{n\sigma_*^4}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}) = 0 \text{ (from the definition of } \tilde{\boldsymbol{\beta}}) \\
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial\sigma^2} &= -\frac{1}{2\sigma_0^4}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) \\
&\quad -\frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial\sigma^2} \\
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial(\sigma^2)^2} &= -\frac{1}{4\sigma_*^6}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}) = 0 \text{ (from the definition of } \tilde{\boldsymbol{\beta}}) \\
\frac{1}{n}\frac{\partial^2 E_2(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial vec(\boldsymbol{L})^T} &= \frac{1}{2\sigma_*^2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(-2G_3+G_4) - \frac{1}{n}\frac{\partial^2 E_1(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)}{\partial vec(\boldsymbol{L})\partial vec(\boldsymbol{L})^T},
\end{aligned}$$

where

$$\begin{aligned}
G_3 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) \qquad (43) \\
G_4 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})\partial vec(\boldsymbol{L})^T}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) \qquad (44)
\end{aligned}$$

With assumptions $A1-A3$, by using Chebyshev'LLN, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X_i}^T\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}\frac{\partial\boldsymbol{L}\boldsymbol{L}^T}{\partial vec(\boldsymbol{L})}\boldsymbol{Z_i}^T\boldsymbol{V_i}^{-1}(\boldsymbol{Y_i}-\boldsymbol{X_i}\boldsymbol{\beta}) = o_p(1) \\
&G_3 = o_p(1);\ G_4 = o_p(1)
\end{aligned}$$

Then it is straightforward to prove all of second derivatives of E_2 at true parameters $(\boldsymbol{\beta}^*, \boldsymbol{L}^*, \sigma_*^2)$ are $o_p(1)$. Then we have proved

$$-\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \rightarrow_p \boldsymbol{I_\theta}$$

This finishes the proof for the lemma.

Proof of Theorem 2 : It is sufficient to show that for any given $\epsilon > 0$, there exists a large constant M_ϵ such that

$$P\left\{\sup_{\|\boldsymbol{u}\|_2=M_\epsilon} Q_n(\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}) < Q_n(\boldsymbol{\theta}^*)\right\} \geq 1-\epsilon, \tag{45}$$

where

$$\begin{aligned} \boldsymbol{\theta}^* &= (\boldsymbol{\beta}^{*T}, vec(\boldsymbol{L}^*)^T, \sigma_*^2)^T && (46)\\ \boldsymbol{u} &= (\boldsymbol{u_1}^T, \boldsymbol{u_2}^T, u_3)^T && (47) \end{aligned}$$

This implies with probability at least $1-\epsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq M_\epsilon\}$. Therefore, there exists a local maximizer such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_p(n^{-1/2})$.

Consider

$$\begin{aligned} D_n(\boldsymbol{u}) &= Q_n(\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}) - Q_n(\boldsymbol{\theta}^*)\\ &= \frac{1}{n}\Big(\ell_n(\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}) - \ell_n(\boldsymbol{\theta}^*)\Big)\\ &\quad - \sum_{j=1}^{p}\Big(f_{\lambda_{1n}}(|\beta_j^* + n^{-1/2}u_{1,j}|) - f_{\lambda_{1n}}(|\beta_j^*|)\Big)\\ &\quad - \sum_{k=2}^{q}\Big(g_{\lambda_{2n}}(|L_{k1}^* + n^{-1/2}u_{2,k1}|, \ldots, |L_{kq}^* + n^{-1/2}u_{2,kq}|) - g_{\lambda_{2n}}(|L_{k1}^*|, \ldots, |L_{kq}^*|)\Big) \end{aligned}$$

Without loss of generality, we assume the first p_1 fixed effects are important, i.e., $\beta_1^*, \ldots, \beta_{p_1}^* \neq 0, \beta_{p_1+1}^* = \ldots \beta_p^* = 0$, and the first q_1 random effects are important, i.e., $\boldsymbol{L}_{(1)}^*, \ldots, \boldsymbol{L}_{(q_1)}^* \neq \boldsymbol{0}, \boldsymbol{L}_{(q_1+1)}^* = \cdots = \boldsymbol{L}_{(q)}^* = \boldsymbol{0}$.

Using the fact that $f_{\lambda_{1n}}(0) = 0$ and $g_{\lambda_{2n}}(0, \ldots, 0) = 0$, we have

$$\begin{aligned} D_n(\boldsymbol{u}) &\leq \frac{1}{n}\Big(\ell_n(\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}) - \ell_n(\boldsymbol{\theta}^*)\Big)\\ &\quad - \sum_{j=1}^{p_1}\Big(f_{\lambda_{1n}}(|\beta_j^* + n^{-1/2}u_{1,j}|) - f_{\lambda_{1n}}(|\beta_j^*|)\Big)\\ &\quad - \sum_{k=2}^{q_1}\Big(g_{\lambda_{2n}}(|L_{k1}^* + n^{-1/2}u_{2,k1}|, \ldots, |L_{kq}^* + n^{-1/2}u_{2,kq}|) - g_{\lambda_{2n}}(|L_{k1}^*|, \ldots, |L_{kq}^*|)\Big)\\ &\widehat{=} A_n - B_n - C_n. \end{aligned}$$

First, by applying Taylor expansion around $\boldsymbol{\theta}^*$ to the log-REML function, we have

$$A_n \;=\; n^{-1}\Big\{n^{-1/2}\frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\Big\}^T\boldsymbol{u} - \frac{1}{2}n^{-1}\boldsymbol{u'}\Big(-n^{-1}\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big)\boldsymbol{u} + n^{-1}o_p(n^{-1}\|\boldsymbol{u}\|_2^2)$$

From Lemma 2, we have $n^{-1/2}\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = O_p(1)$ and $-n^{-1}\frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \boldsymbol{I}_{\boldsymbol{\theta}} + o_p(1)$. Then we have

$$\begin{aligned}
A_n &\leq n^{-1}O_p(1)\|\boldsymbol{u}\|_1 - \frac{1}{2}n^{-1}\boldsymbol{u}^T\Big\{\boldsymbol{I}_\theta + o_p(1)\Big\}\boldsymbol{u} + n^{-1}o_p(n^{-1}\|\boldsymbol{u}\|_2^2)\\
&\leq \sqrt{p+q+1}n^{-1}\|\boldsymbol{u}\|_2 O_p(1) - \frac{1}{2}n^{-1}\boldsymbol{u}^T\boldsymbol{I}_\theta\boldsymbol{u} + o_p(n^{-1}\|\boldsymbol{u}\|_2^2)\\
&= A_{1n} + A_{2n} + A_{3n},
\end{aligned}$$

where $\|\boldsymbol{u}\|_1$ is the L_1-norm of $\boldsymbol{u}$, i.e., $\|\boldsymbol{u}\|_1 = |u_1| + \cdots + |u_t|$ with t be the length of $\boldsymbol{u}$, and it can be easily checked that $\|\boldsymbol{u}\| \leq t\|\boldsymbol{u}\|_2$.

Second, by applying Taylor expansion to the penalty function, we have

$$\begin{aligned}
B_n &= \sum_{j=1}^{p_1}\Big(\frac{\partial f_{\lambda_{1n}}(|\beta_j^*|)}{\partial|\beta_j|}\mathrm{sgn}(\beta_j^*)n^{-1/2}u_{1,j} + \frac{1}{2}\frac{\partial^2 f_{\lambda_{1n}}(|\beta_j^*|)}{\partial|\beta_j|^2}n^{-1}u_{1,j}^2 + o_p(n^{-1}u_{1,j}^2)\Big)\\
&\leq \sqrt{p}n^{-1/2}a_n\|\boldsymbol{u_1}\|_2 + \frac{1}{2n}b_n\|\boldsymbol{u_1}\|_2^2 + o_p(n^{-1}\|\boldsymbol{u_1}\|_2^2)\\
&= \sqrt{p}\|\boldsymbol{u_1}\|_2 O_p(n^{-1}) + o_p(n^{-1}\|\boldsymbol{u_1}\|_2^2) \;\; (\text{using } a_n = O_p(n^{-1/2}),\ b_n = o_p(1))\\
&= B_{1n} + B_{2n}
\end{aligned}$$

$$\begin{aligned}
C_n &= \sum_{k=2}^{q_1}\Big(\sum_{l=1}^{q}\frac{\partial g_{\lambda_{2n}}(|L_{k1}^*|,\ldots,|L_{kq}^*|)}{\partial|L_{kl}|}\mathrm{sgn}(L_{kl}^*)n^{-1/2}u_{2,kl}\\
&\quad + \frac{1}{2}\sum_{l_1=1}^{q}\sum_{l_2=1}^{q}\frac{\partial^2 g_{\lambda_{2n}}(|L_{k1}^*|,\ldots,|L_{kq}^*|)}{\partial|L_{kl_1}|\partial|L_{kl_2}|}\mathrm{sgn}(L_{kl_1}^*)\mathrm{sgn}(L_{kl_2}^*)n^{-1}u_{2,kl_1}u_{2,kl_2} + o_p(n^{-1}(u_{2,k1}^2 + \cdots + u_{2,kq}^2)\Big)\\
&\leq \sqrt{q}n^{-1/2}c_n\|\boldsymbol{u_2}\|_2 + \frac{1}{2n}d_n\|\boldsymbol{u_2}\|_2^2 + o_p(n^{-1}\|\boldsymbol{u_2}\|_2^2)\\
&= \sqrt{q}\|\boldsymbol{u_2}\|_2 O_p(n^{-1}) + o_p(n^{-1}\|\boldsymbol{u_2}\|_2^2) \;\; (\text{using } c_n = O_p(n^{-1/2}),\ d_n = o_p(1))\\
&= C_{1n} + C_{2n}
\end{aligned}$$

We can see that, by choosing a sufficiently large M_ϵ, A_{2n} dominates $A_{1n}, A_{3n}, B_{1n}, B_{2n}, C_{1n}, C_{2n}$ uniformly in $\|\boldsymbol{u}\|_2 = M_\epsilon$. This completes the proof.

Proof of Theorem 3 :

To prove (a), it is sufficient to show that, for any constant M, with probability tending to 1 as $n \to \infty$,

$$Q_n(\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2) = \max_{\|\boldsymbol{\beta}_{\mathcal{B}}\|_2 \le M n^{-1/2}} Q_n(\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{B}}, \hat{\boldsymbol{L}}, \hat{\sigma}^2) \tag{48}$$

By applying Taylor's expansion around $\boldsymbol{\theta}_0$ to $\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j}$, the first derivative of Q_n, we have

$$\begin{aligned} \frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j} &= n^{-1}\frac{\partial \ell_n(\boldsymbol{\theta}^*)}{\partial \beta_j} + \frac{1}{2}\sum_{l=1}^{p} n^{-1}\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_l^*) + \frac{1}{2n}\sum_{l=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 \ell_n(\bar{\boldsymbol{\theta}})}{\partial \beta_j \partial \beta_l \partial \beta_k}(\hat{\beta}_l - \beta_l^*)(\hat{\beta}_k - \beta_k^*) \\ &\quad - \frac{f_{\lambda_{1n}}(|\hat{\beta}_j|)}{\partial |\beta_j|}\operatorname{sgn}(\hat{\beta}_j), \end{aligned}$$

where $\bar{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. From Lemma 2 and assumption, we have

$$n^{-1}\frac{\partial \ell_n(\boldsymbol{\theta}^{\mathbf{0}})}{\partial \beta_j} = n^{-1/2}\Big(n^{-1/2}\frac{\partial \ell_n(\boldsymbol{\theta}^{\mathbf{0}})}{\partial \beta_j}\Big) = O_p(n^{-1/2}),\ \frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\theta}^{\mathbf{0}})}{\partial \beta_j \partial \beta_l} = O_p(1),\ \frac{1}{n}\frac{\partial^3 \ell_n(\bar{\boldsymbol{\theta}})}{\partial \beta_j \partial \beta_l \partial \beta_k} = O_p(1).$$

Then since $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}\|_2 = O_p(n^{-1/2})$, we have

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j} = n^{-1/2}\left\{O_p(1) - n^{1/2}\frac{\partial f_{\lambda_{1n}}(|\hat{\beta}_j|)}{\partial |\beta_j|}\operatorname{sgn}(\hat{\beta}_j)\right\} \tag{49}$$

If for any j with $\beta_j^* = 0$, $\sqrt{n}\frac{\partial f_{\lambda_{1n}}(|\beta_j|)}{\partial |\beta_j|} \to \infty$ with probability tending to 1 as $n \to \infty$, then when n is large we have

$$\begin{aligned} \frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j} &< 0,\ 0 < \hat{\beta}_j < M n^{-1/2}, \\ \frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j} &> 0,\ \ -M n^{-1/2} < \hat{\beta}_j < 0, \end{aligned}$$

which indicates $Q_n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\mathcal{A}}}, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2) = \max_{\|\boldsymbol{\beta}_{\mathcal{B}}\| \le M n^{-1/2}} Q_n(\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{B}}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)$. This completes the proof for (a).

For (b), similarly we can have

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial L_{kj}} = n^{-1/2}\left\{O_p(1) - n^{1/2}\frac{\partial g_{\lambda_{2n}}(|\hat{L}_{k1}|, \ldots, |\hat{L}_{kq}|)}{\partial |L_{kj}|}\operatorname{sgn}(\hat{L}_{kj})\right\} \tag{50}$$

If for any $(k, j) \in \mathcal{D}$, $\sqrt{n}\frac{\partial g_{\lambda_{2n}}(|\hat{L}_{k1}|, \ldots, |\hat{L}_{kq}|)}{\partial |L_{kj}|} \to \infty$ with probability tending to 1 as $n \to \infty$, then for any constant $M > 0$, when n is large we have

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial L_{kj}} < 0,\ 0 < \hat{L}_{kj} < M n^{-1/2}, \tag{51}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial L_{kj}} > 0,\ \ -M n^{-1/2} < \hat{L}_{kj} < 0, \tag{52}$$

With the similar argument in the proof for (a), we can prove (b).

For (c), following Theorem 2, 3(a) and 3(b), there exists a $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^T, \mathbf{0}^T, vec(\hat{\boldsymbol{L}})^T, \hat{\sigma}^2)^T$ that satisfy the equation

$$\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)}{\partial \boldsymbol{\beta}_{\mathcal{A}}} = 0 \tag{53}$$

By applying Taylor expansion around $\boldsymbol{\beta}_{\mathcal{A}}^*$ to $\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)}{\partial \boldsymbol{\beta}_{\mathcal{A}}}$, for any $j \in \mathcal{A}$, we have

$$\begin{aligned}
\sqrt{n} \cdot 0 &= \sqrt{n}\left(\frac{1}{n}\frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial \beta_j} - \frac{\partial f_{\lambda_{1n}}(|\hat{\beta}_j|)}{\partial |\beta_j|}\text{sgn}(|\hat{\beta}_j|)\right) \\
&= \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)}{\partial \beta_j} + \frac{1}{n}\sum_{k=1}^{p_1}\left\{\frac{\partial^2 \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \hat{\boldsymbol{L}}, \hat{\sigma}^2)}{\partial \beta_j \partial \beta_k}\sqrt{n}(\hat{\beta}_k - \beta_k^*) + o_p\Big(\sqrt{n}(\hat{\beta}_k - \beta_k^*)\Big)\right\} \\
&\quad -\sqrt{n}\frac{\partial f_{\lambda_{1n}}(|\beta_j^*|)}{\partial |\beta_j|}\text{sgn}(\beta_j^*) - \frac{\partial^2 f_{\lambda_{1n}}(|\beta_j^*|)}{\partial |\beta_j|^2}\sqrt{n}(\hat{\beta}_j - \beta_j^*) + o_p\Big(\sqrt{n}(\hat{\beta}_j - \beta_j^*)\Big)
\end{aligned}$$

Under assumptions $A1 - A3$, if $\sqrt{n}\frac{\partial f_{\lambda_{1n}}(|\beta_j^*|)}{\partial |\beta_j|} = o_p(1)$ and $b_n = o_p(1)$, then by the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}, \hat{\boldsymbol{L}}, \hat{\sigma}^2$, we have

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \beta_j} + \left\{\frac{1}{n}\sum_{k=1}^{p_1}\frac{\partial^2 \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \beta_j \partial \beta_k}\right\}\sqrt{n}(\hat{\beta}_k - \beta_k^*) + o_p(1) \\
\Rightarrow \quad &\left\{-\frac{1}{n}\sum_{k=1}^{p_1}\frac{\partial^2 \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \beta_j \partial \beta_k}\right\}\sqrt{n}(\hat{\beta}_k - \beta_k^*) = \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\beta}_{\mathcal{A}}^*, \mathbf{0}, \boldsymbol{L}^*, \sigma_*^2)}{\partial \beta_j} + o_p(1)
\end{aligned}$$

Then by Slutsky's theorem, we have

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*\Big) \to \text{MVN}(\mathbf{0}, \boldsymbol{I}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}_{\mathcal{A}}^*)), \tag{54}$$

where $\boldsymbol{I}_{\mathcal{A}}$ is the corresponding part for $\boldsymbol{\beta}_{\mathcal{A}}$ in Fisher's information matrix.

For (d), Similarly to (c), following Theorem 2, 3(c) and 3(d), there exists a $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, vec(\hat{\boldsymbol{L}}_{\mathcal{C}})^T, \mathbf{0}^T, \hat{\sigma}^2)^T$ that satisfy the equation

$$\frac{\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{L}}_{\mathcal{C}}, \mathbf{0}, \hat{\sigma}^2)}{\partial vec(\boldsymbol{L}_{\mathcal{C}})} = 0 \tag{55}$$

By applying Taylor expansion around $vec(\boldsymbol{L}^*_{\mathcal{C}})$ to $\frac{\partial Q_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{L}}_{\mathcal{C}}, \mathbf{0}, \hat{\sigma}^2)}{\partial vec(\boldsymbol{L}_{\mathcal{C}})}$, for any $(k,j) \in \mathcal{C}$, we have

$$\begin{aligned}
\sqrt{n} \cdot 0 &= \sqrt{n}\left(\frac{1}{n}\frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial L_{kj}} - \frac{\partial g_{\lambda_{2n}}(|\hat{L}_{k1}|, \ldots, |\hat{L}_{kq}|)}{\partial |L_{kj}|}\mathrm{sgn}(|\hat{L}_{kj}|)\right) \\
&= \frac{1}{\sqrt{n}}\frac{\partial \ell_n(\hat{\boldsymbol{\beta}}, \boldsymbol{L}^*_{\mathcal{C}}, \mathbf{0}, \hat{\sigma}^2)}{\partial L_{kj}} + \frac{1}{n}\sum_{l=1}^{q_1}\sum_{m=1}^{q_1}\left\{\frac{\partial^2 \ell_n(\hat{\boldsymbol{\beta}}, \boldsymbol{L}^*_{\mathcal{C}}, \mathbf{0}, \hat{\sigma}^2)}{\partial L_{kj}\partial L_{lm}}\sqrt{n}(\hat{L}_{lm} - L^*_{lm}) + o_p\Big(\sqrt{n}(\hat{L}_{lm} - L^*_{lm})\Big)\right\} \\
&\quad -\sqrt{n}\frac{\partial g_{\lambda_{2n}}(|L^*_{k1}|, \ldots, |L^*_{kq}|)}{\partial |L_{kj}|}\mathrm{sgn}(L^*_{kj}) \\
&\quad -\sum_{l=1}^{q}\left\{\frac{\partial^2 g_{\lambda_{2n}}(|L^*_{k1}|, \ldots, |L^*_{kq}|)}{\partial |L_{kj}|\partial |L_{kl}|}\sqrt{n}(\hat{L}_{kl} - L^*_{kl}) + o_p\Big(\sqrt{n}(\hat{L}_{kl} - L^*_{kl})\Big)\right\}
\end{aligned}$$

Then with the similar argument in (c), we can prove (d).

Proof of Corollary 1 : We only need to check the corresponding $a_n, c_n = O_p(n^{-1/2})$ and $b_n, d_n = o_p(1)$. Since both numbers of fixed effects and random effects are fixed, it is straightforward to check the two conditions are satisfied when $\lambda_{1n} = O_p(n^{-1/2})$ and $\lambda_{2n} = O_p(n^{-1/2})$.

Proof of Theorem 4 :

For consistency, by Theorem 2, we only need to check $a_n, c_n = O_p(n^{-1/2})$ and $b_n, d_n = o_p(1)$.

For $j : \beta^*_j \neq 0$,

$$\begin{aligned}
\frac{\partial f_{\lambda_{1n}}(|\beta^0_j|)}{\partial |\beta_j|} &= \lambda_{1n} w^{\beta}_{nj} \leq \lambda_{1n} w^{\beta}_{n,max}, &(56)\\
\frac{\partial^2 f_{\lambda_{1n}}(|\beta^0_j|)}{\partial |\beta_j|^2} &= 0. &(57)
\end{aligned}$$

Obviously $b_n = 0$. If $\sqrt{n}\lambda_{1n} w^{\beta}_{n,max} = O_p(1)$, then $a_n = O_p(n^{-1/2})$.

For $k : \sqrt{{L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2} > 0$,

$$\begin{aligned}
\frac{\partial g_{\lambda_{2n}}(|L^*_{k1}|, \ldots, |L^*_{kq}|)}{\partial |L_{kj}|} &= \lambda_{2n} w^L_{nj} \frac{|L^*_{kj}|}{\sqrt{{L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2}} \leq \lambda_{2n} w^L_{n,max}, &(58)\\
\frac{\partial^2 g_{\lambda_{2n}}(|L^*_{k1}|, \ldots, |L^*_{kq}|)}{\partial |L_{kj}|^2} &= \lambda_{2n} w^L_{nj} \frac{{L^*_{k1}}^2 + \cdots + {L^*_{kj-1}}^2 + {L^*_{kj+1}}^2 + \cdots + {L^*_{kq}}^2}{({L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2)^{3/2}} \leq \lambda_{2n} w^L_{n,max}, &(59)\\
\frac{\partial^2 g_{\lambda_{2n}}(|L^*_{k1}|, \ldots, |L^*_{kq}|)}{\partial |L_{kj}|\partial |L_{kj'}|} &= \lambda_{2n} w^L_{nj} \frac{-|L^*_{kj}||L^*_{kj'}|}{({L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2)^{3/2}} \leq \lambda_{2n} w^L_{n,max} M_1, &(60)
\end{aligned}$$

where

$$M_1 = \max_{(k,j)} \Big\{ \frac{-|L^*_{kj}||L^*_{kj'}|}{({L^0_{k1}}^2 + \cdots + {L^0_{kq}}^2)^{3/2}} \Big\}.$$

If $\sqrt{n}\lambda_{2n} w^L_{n,max} = O_p(1)$ (hence $\lambda_{2n} w^L_{n,max} = o_p(1)$), then $c_n = O_p(n^{-1/2})$ and $d_n = o_p(1)$.

Now we prove the sparsity.

For $j : \beta^*_j = 0$, if $\hat{\beta}_j$ is a $\sqrt{n}$-consistent estimator, then

$$\sqrt{n}\frac{f_{\lambda_{1n}}(|\hat{\beta}_j|)}{|\beta_j|} = \sqrt{n}\lambda_{1n} w^{\beta}_{nj} \geq \sqrt{n}\lambda_{1n} w^{\beta}_{n,min} \tag{61}$$

For $k : \sqrt{{L^*_{k1}}^2 + \cdots + {L^*_{kq}}^2} = 0$, if $\hat{L}_{k1}, \ldots, \hat{L}_{kq}$ are $\sqrt{n}$-consistent estimators, then

$$\sqrt{n}\frac{g_{\lambda_{2n}}(|\hat{L}_{k1}|, \ldots, |\hat{L}_{kq}|)}{|L_{kj}|} = \frac{\sqrt{n}\lambda_{2n} w^L_{nj} \hat{L}_{kj}}{\sqrt{\hat{L}^2_{k1} + \cdots + \hat{L}^2_{kq}}}$$

Since $\hat{L}_{kj}$'s are $\sqrt{n}$-consistent, we have $\hat{L}_{kj} \Big/ \sqrt{\hat{L}^2_{k1} + \cdots + \hat{L}^2_{kq}} \xrightarrow{p} C > 0$. Therefore, for any $\epsilon > 0$, there is a constant M_ϵ, such that when n is large, $P\left(\hat{L}_{kj} \Big/ \sqrt{\hat{L}^2_{k1} + \cdots + \hat{L}^2_{kq}} > M_\epsilon\right) \geq 1-\epsilon$. Then

$$P\left(\sqrt{n}\frac{g_{\lambda_{2n}}(|\hat{L}_{k1}|, \ldots, |\hat{L}_{kq}|)}{|L_{kj}|} = \frac{\sqrt{n}\lambda_{2n} w^L_{nj} \hat{L}_{kj}}{\sqrt{\hat{L}^2_{k1} + \cdots + \hat{L}^2_{kq}}} \geq \sqrt{n}\lambda_{2n} w^L_{n,min} M_\epsilon\right) \geq 1 - \epsilon \tag{62}$$

If $\sqrt{n}\lambda_{1n} w^{\beta}_{n,min} \xrightarrow{p} \infty$ and $\sqrt{n}\lambda_{2n} w^L_{n,min} \xrightarrow{p} \infty$, then (61) and (61) tend to infinity with probability tending to 1 when n tends to infinity. By Theorem 3, we have

$$Pr(\hat{\boldsymbol{\beta}}_{\mathcal{B}} = \mathbf{0}) \rightarrow 1, \ Pr(\hat{\boldsymbol{L}}_{\mathcal{D}} = \mathbf{0}) \rightarrow 1.$$

For asymptotic normality, using (56) and (58), by Theorem 3, if $\sqrt{n}\lambda_{1n} w^{\beta}_{n,max}$, $\sqrt{n}\lambda_{2n} w^L_{n,max} = o_p(1)$, we have

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}\Big) \xrightarrow{d} \mathrm{MVN}(\mathbf{0}, \boldsymbol{I}_{\mathcal{A}}(\boldsymbol{\beta}^*_{\mathcal{A}})), \ \sqrt{n}\Big(\hat{\boldsymbol{L}}_{\mathcal{C}} - \boldsymbol{L}^*_{\mathcal{C}}\Big) \xrightarrow{d} \mathrm{MVN}(\mathbf{0}, \boldsymbol{I}_{\mathcal{C}}(\boldsymbol{L}^*_{\mathcal{C}})).$$

Proof of Corollary 2 : It is straightforward to check that the conditions in Theorem 4 are satisfied.