Statistica Sinica **22** (2012), 1539-1562
doi:http://dx.doi.org/10.5705/ss.2011.054

MOMENT-BASED METHOD FOR RANDOM EFFECTS SELECTION IN LINEAR MIXED MODELS

Mihye Ahn, Hao Helen Zhang and Wenbin Lu

North Carolina State University

Abstract: The selection of random effects in linear mixed models is an important yet challenging problem in practice. We propose a robust and unified framework for automatically selecting random effects and estimating covariance components in linear mixed models. A moment-based loss function is first constructed for estimating the covariance matrix of random effects. Two types of shrinkage penalties, a hard thresholding operator and a new sandwich-type soft-thresholding penalty, are then imposed for sparse estimation and random effects selection. Compared with existing approaches, the new procedure does not require any distributional assumption on the random effects and error terms. We establish the asymptotic properties of the resulting estimator in terms of its consistency in both random effects selection and variance component estimation. Optimization strategies are suggested to tackle the computational challenges involved in estimating the sparse variance-covariance matrix. Furthermore, we extend the procedure to incorporate the selection of fixed effects as well. Numerical results show the promising performance of the new approach in selecting both random and fixed effects, and consequently, improving the efficiency of estimating model parameters. Finally, we apply the approach to a data set from the Amsterdam Growth and Health study.

Key words and phrases: Hard thresholding, linear mixed model, shrinkage estimation, variance component selection.

1. Introduction

In many applications, it is common practice to collect repeated measurements on a subject or take serial observations over time on the same unit, resulting in clustered data, longitudinal data, or spatial data. Linear mixed models (Laird and Ware (1982)) are a class of tools useful in the analysis of correlated data by introducing subject-specific random effects to account for the variation among subjects. The use of random effects models provides a convenient and effective way of describing the covariance structure of data. Thus, suppose there are m subjects under study and the number of measurements on subject i is n_i. Typically we assume that $m > n_i$. A general linear mixed model is written as

$$Y_{ij} = X_{ij}^T\beta + Z_{ij}^T\gamma_i + \varepsilon_{ij} \qquad (i = 1, \ldots, m; j = 1, \ldots, n_i), \tag{1.1}$$

where Y_{ij} is the response, $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ are the fixed-effect covariates for observation j for subject i, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the $p \times 1$ vector of fixed-effect coefficients, $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijq})^T$ are the random-effect covariates for observation j for subject i, $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iq})^T$ is the subject-specific $q \times 1$ vector of random-effects coefficients, and the ε_{ij}'s are the error terms. Furthermore, we assume that the γ_i have mean 0 and variance-covariance matrix $\Sigma = [\sigma_{jk}]$ with $1 \le j, k \le q$. The errors $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T$ are iid with mean 0 and variance-covariance matrix $\sigma_\varepsilon^2 I_{n_i}$, and the ε_i's are independent of the γ_i's. When X_i and Z_i are random variables, we also assume that $E(\gamma_i | X_i, Z_i) = 0$ and $\mathrm{var}(\gamma_i | X_i, Z_i) = \Sigma$, where the jth row vectors of X_i and Z_i are X_{ij}^T and Z_{ij}^T, respectively. These assumptions are made throughout the paper.

The selection of important random effects in (1.1) plays a crucial role in model estimation and inference. Here, important random effects refer to those whose coefficients actually vary among subjects. If important random effects are left out of the model, the covariance matrix of random effects would be underfitted, which would lead to bias in the estimated variance for fixed effects. On the other hand, if unnecessary random effects are included in the model, the covariance matrix of random effects could be nearly singular, which would cause numerical instability for model fitting. Furthermore, the correct selection of the important random effects helps to achieve the estimation efficiency for the fixed effects and the accuracy of future prediction.

Recently the problem of variable selection has received much attention and a large number of methods have been proposed, including the traditional forward selection and backward elimination methods (Breiman (1995)), and modern penalized regression (Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Zou (2006); Wang, Li, and Jiang (2007); Zhang and Lu (2007); and Bondell and Reich (2008)). Most of these methods are designed for selecting important fixed effects. In this paper, our main focus is on the selection of random effects. In the research literature, various model selection criteria have been proposed to compare candidate models, including the Akaike Information Criterion (AIC) (Akaike (1973)), Bayesian Information Criterion (BIC) (Schwarz (1978)), Generalized Information Criterion (GIC) (Rao and Wu (1989)), and conditional AIC (Vaida and Blanchard (2005)). However, the number of possible models is 2^{p+q}, which increases exponentially with the number of predictors and hence makes computation infeasible for large p or q. Some approaches have been proposed to reduce the number of possible models to $2^p + 2^q$, including the extended GIC (Niu and Pu (2006)) and the restricted information criterion (Wolfinger (1993); Diggle, Liang, and Zeger (1994)). Recently, Chen and Dunson (2003) and Kinney and Dunson (2007) proposed Bayesian approaches for variable selection in linear mixed models, while Bondell, Krishna, and Ghosh (2010) proposed a likelihood-based method for jointly selecting fixed and random effects. These approaches

all require the normality assumption for both the random effects and error terms; therefore, the validity of their inferences depends heavily on whether the distributional assumption is correct or not. This motivates us to develop a more robust and flexible approach for random effects selection in linear mixed models.

We construct a moment-based loss function for the variance components and then employ adaptive shrinkage and thresholding to achieve random effects selection and model estimation. The new estimator is robust against non-normality of the data, and the estimates are valid for any distribution of random effects and errors. We proved the selection consistency, the root-m consistency, and asymptotic normality for the new estimator. Furthermore, the proposed computational algorithm is fast and the procedure shows promising performance in numerical studies. The rest of the paper is organized as follows. In Sections 2 and 3 we propose a new class of procedures for random effects selection. We establish the theoretical properties of the proposed estimators and discuss computational issues. In Section 4 we extend the procedure to incorporate fixed effects selection. Section 5 contains numerical results and Section 6 gives concluding remarks. All the proofs are relegated to the Appendix.

2. Initial Moment-based Covariance Matrix Estimator

In a matrix form, model (1.1) can be written as

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

where Y_i is the $n_i \times 1$ response vector for the observations of subject i, X_i is the $n_i \times p$ design matrix for fixed effects, Z_i is the $n_i \times q$ design matrix for random effects, and ε_i is the $n_i \times 1$ vector of errors for the observations of subject i. Then, $\mathrm{var}(Y_i) = \sigma_\varepsilon^2 I_{n_i} + Z_i \Sigma Z_i^T$ for $i = 1, \ldots, m$, which naturally incorporates heterogeneity among the subjects. Typically the parameters in (2.1) are estimated by maximum likelihood (ML) and restricted maximum likelihood (REML) methods by assuming that the γ_i's and ε_i's are all normally distributed. See Laird and Ware (1982), Jennrich and Schluchter (1986), and Lindstrom and Bates (1988). In the following, we use a moment-based approach to estimate the model parameters, which does not require any specification of the distributions of random effects and errors.

Denote the total number of observations by $N = \sum_{i=1}^m n_i$. We further express (2.1) as

$$Y = X\beta + Z\gamma + \varepsilon,$$

where $Y = (Y_1^T, \ldots, Y_m^T)^T$ is a $N \times 1$ vector, $X = (X_1^T, \ldots, X_m^T)^T$ is a $N \times p$ matrix, $Z = \mathrm{diag}(Z_1, \ldots, Z_m)$ is a $N \times mq$ block diagonal matrix, $\gamma = (\gamma_1^T, \ldots, \gamma_m^T)^T$ is a $mq \times 1$ vector, and $\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_m^T)^T$ is a $N \times 1$ vector. The "diag" operator

is defined as: diag(A) is a vector of diagonal elements of A if A is a matrix, or a diagonal matrix with elements of A along the diagonal if A is a vector, or a block diagonal matrix with submatrices along the diagonal being $A_1, \ldots, A_a$ if A consists of submatrices $A_1, \ldots, A_a$.

Define

$$Y_{ijk} = (Y_{ij} - X_{ij}^T\beta)(Y_{ik} - X_{ik}^T\beta), \tag{2.2}$$

where Y_{ij} is the jth entry of Y_i and X_{ij}^T is the jth row of X_i ($i = 1, \ldots, m; j = 1, \ldots, n_i; k = j, \ldots, n_i$). It is easy to show that the expectation of Y_{ijk} is the second-order cross-moment of $Z_i\gamma_i + \varepsilon_i$:

$$\begin{aligned} E[Y_{ijk}|X_i, Z_i] &= E\big\{(Z_{ij}^T\gamma_i + \varepsilon_{ij})(Z_{ik}^T\gamma_i + \varepsilon_{ik})\big\} \\ &= \begin{cases} Z_{ij}^T\Sigma Z_{ik} + \sigma_\varepsilon^2 & \text{if } j = k, \\ Z_{ij}^T\Sigma Z_{ik} & \text{otherwise,} \end{cases} \end{aligned}$$

where Z_{ij}^T is the jth row of Z_i. If β is known, a moment estimator for Σ could be obtained by minimizing the quantity $\sum_{i=1}^m \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \big(Y_{ijk} - Z_{ij}^T\Sigma Z_{ik}\big)^2$. Since β is generally unknown, we propose to obtain an unbiased initial estimator first. A natural choice is the ordinary least squares (OLS) estimator, $\widetilde{\beta}$, obtained by fitting the simple linear model $Y = X\beta + \eta$ under the working independence assumption. Substituting $\widetilde{\beta}$ into (2.2), we get $\widetilde{Y}_{ijk} = (Y_{ij} - X_{ij}^T\widetilde{\beta})(Y_{ik} - X_{ik}^T\widetilde{\beta})$. We propose to obtain an initial estimator of Σ by minimizing

$$L_0(\Sigma) = \sum_{i=1}^m \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \Big(\widetilde{Y}_{ijk} - Z_{ij}^T\Sigma Z_{ik}\Big)^2. \tag{2.3}$$

Let $\widetilde{\Sigma} = [\widetilde{\sigma}_{jk}]$ be the solution to (2.3). By substituting $\widetilde{\Sigma}$ into $\sum_{i=1}^m \sum_{j=k} \big(\widetilde{Y}_{ijk} - Z_{ij}^T\Sigma Z_{ik} - \sigma_\varepsilon^2\big)^2$ and minimizing the quantity with respect to σ_ε^2, we obtain the estimator of the error variance $\widetilde{\sigma}_\varepsilon^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (\widetilde{Y}_{ijj} - Z_{ij}^T\widetilde{\Sigma}Z_{ij})/N$.

For convenient notation, we can reformat Σ into its vector form as $\kappa \equiv \text{vech}(\Sigma)$, a vector consisting of $q(q+1)/2$ elements that are on and above the diagonal of Σ. In this way, $L_0(\Sigma)$ can be written as a function of κ as

$$L_0(\kappa) = \sum_{i=1}^m \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \Big(\widetilde{Y}_{ijk} - Z_{ijk}^{*T}\kappa\Big)^2,$$

where Z_{ijk}^* is a $q(q+1)/2 \times 1$ vector such that $Z_{ij}^T\widetilde{\Sigma}Z_{ik} = Z_{ijk}^{*T}\kappa$. Correspondingly, let $\widetilde{\kappa} = \text{vech}(\widetilde{\Sigma})$. Throughout the paper, we use the subscript 'o' on parameters to denote the true values of corresponding parameters. In particular, β_o denotes the true fixed-effect coefficients, Σ_o denotes the true variance-covariance matrix

of the random effects, and $\kappa_o \equiv \text{vech}(\Sigma_o)$. Define $e_{ijk} = (Y_{ij} - X_{ij}^T\beta_o)(Y_{ik} - X_{ik}^T\beta_o) - Z_{ijk}^{*T}\kappa_o$ for $i = 1, \ldots, m; j = 1, \ldots, n_i - 1; k = j + 1, \cdots, n_i$. Let e_i be a column vector consisting of e_{ijk}'s, and let Z_i^* be the matrix whose column vectors are Z_{ijk}^*'s. In the following two lemmas, we establish the asymptotic normalities of $\widetilde{\beta}$ and $\widetilde{\kappa}$, respectively.

Lemma 1. *Let*

$$A_1 = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X_i^T X_i \quad \textit{and} \quad B_1 = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X_i^T (Z_i \Sigma Z_i^T + \sigma_\varepsilon^2 I_{n_i}) X_i.$$

Given that A_1 and B_1 are finite and nondegenerate, we have

$$m^{1/2}(\widetilde{\beta} - \beta_o) \to N(0, A_1^{-1} B_1 A_1^{-1})$$

in distribution, as $m \to \infty$.

Lemma 2. *Suppose that the assumptions of Lemma* 1 *hold. Let*

$$A_2 = \lim_{m\to\infty} \frac{1}{m} \sum_{i,j,k} Z_{ijk}^* Z_{ijk}^{*T} \quad \textit{and} \quad B_2 = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} Z_i^{*T} cov(e_i) Z_i^*.$$

Given that A_2 and B_2 are finite and nondegenerate, we have

$$m^{1/2}\big(\widetilde{\kappa} - \kappa_o\big) \to N(0, A_2^{-1} B_2 A_2^{-1})$$

in distribution, as $m \to \infty$.

The proofs of Lemmas 1 and 2 are straightforward and so are omitted here. For large samples, $\widetilde{\Sigma}$ is a non-negative definite matrix; however when the sample size is small, $\widetilde{\Sigma}$ is not guaranteed. In practice, we suggest minimizing (2.3) under the constraint $\Sigma \succeq 0$, *i.e.*,

$$\min_{\Sigma} L_0(\Sigma) = \sum_{i=1}^{m} \sum_{j=1}^{n_i - 1} \sum_{k=j+1}^{n_i} \Big(\widetilde{Y}_{ijk} - Z_{ij}^T \Sigma Z_{ik}\Big)^2 \quad \text{subject to} \quad \Sigma \succeq 0. \tag{2.4}$$

Similarly, to ensure that the estimated error variance is non-negative, we take

$$\widetilde{\sigma}_\varepsilon^2 = \max\Big(0, \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{(\widetilde{Y}_{ijj} - Z_{ij}^T \widetilde{\Sigma} Z_{ij})}{N}\Big)$$

in finite sample situations.

The optimization problem in (2.4) is a nonlinear semi-definite programming problem, We propose using the MATLAB toolbox YALMIP, which is a modeling

language for rapid optimization (Löfberg (2004)) that is publicly available for free. YALMIP provides a convenient interface to convert various optimization problems into a common format and solve them. Based on our comparison of different solvers in YALMIP for semi-definite programming, we find that the solver SeDuMi (Sturm (1999)) provides fast and robust performance overall. Therefore, we chose SeDuMi to solve (2.4).

3. Sparse Covariance Estimator by Thresholding and Shrinkage

The main goal of this paper is to identify the important random effects in model (1.1). We say the lth ($1 \le l \le q$) random effect is not important, if and only if $\text{var}(\gamma_{il}) = 0$ for all i, or equivalently, if all of the elements in the lth column and the lth row of Σ are zero. Though the moment-based estimator $\widetilde{\Sigma}$ is consistent, it does not have the desired sparse structure. In this section, we propose two effective procedures to achieve sparsity in the covariance estimation by imposing hard thresholding and shrinkage on the initial moment estimator. Generalized thresholding has been shown to be a promising technique for estimating the sample covariance matrix (Rothman, Levina, and Zhu (2009)). We will study the theoretical and computational properties of new estimators and discuss their advantages and disadvantages.

3.1. Hard thresholding method

For any $\nu > 0$, we define the hard thresholding estimator $\widehat{\Sigma}_\nu^H = [\widehat{\sigma}_{ij}^H]$ by

$$\widehat{\sigma}_{ij}^H = \widetilde{\sigma}_{ij} I(|\widetilde{\sigma}_{ij}| > \nu), \quad 1 \le i, j \le q, \tag{3.1}$$

where $\mathrm{I}(\cdot)$ is an indicator function and $\widetilde{\sigma}_{ij}$ is the moment-based estimate obtained from Section 2.1. Here $\nu \ge 0$ is the parameter that controls the thresholding criterion. Given ν and $\widetilde{\Sigma}$, it is easy to obtain $\widehat{\Sigma}_\nu^H$, and the computation cost is minimal. This is one of the main advantages of the hard thresholding estimator. In the first theorem, we show that $\widehat{\Sigma}_\nu^H$ is root-m consistent in both the operator and Frobenius norms. For any symmetric matrix A, we denote its largest eigenvalue in absolute value by $\zeta_{\max}$, its operator norm by $\|A\|_2 = \zeta_{\max}$, and the Frobenius norm by $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$.

Theorem 1 (Root-m Consistency)**.** *Under the assumptions of Lemma 2 hold. If $\nu \to 0$ as $m \to \infty$, then the hard thresholding estimator $\widehat{\Sigma}_\nu^H = [\widehat{\sigma}_{ij}^H]$ satisfies $\|\widehat{\Sigma}_\nu^H - \Sigma_o\|_2 = O_p(m^{-1/2})$ and $\|\widehat{\Sigma}_\nu^H - \Sigma_o\|_F = O_p(m^{-1/2})$.*

In the next theorem, we show that with probability approaching 1, the hard thresholding estimator can correctly identify true zero components. Furthermore,

under certain conditions $\mathrm{sgn}(\widehat{\sigma}_{ij}^H)$ matches the sign of the true nonzero σ_{ij} if the tuning parameter is chosen properly.

Theorem 2 (Sparsity)**.** *Under the assumptions of Lemma* 2 *suppose that* $\nu \to 0$ *and* $\sqrt{m}\nu \to \infty$ *as* $m \to \infty$.

(a) *With probability approaching* 1,

$$\widehat{\sigma}_{ij}^H = 0 \quad \textit{for all} \quad (i,j) \ \textit{such that} \quad \sigma_{ij,o} = 0.$$

(b) *With probability approaching* 1, *we have*

$$sgn(\widehat{\sigma}_{ij}^H \sigma_{ij,o}) = 1 \quad \textit{for all} \quad (i,j) \ \textit{such that} \quad \sigma_{ij,o} \neq 0.$$

The proofs of Theorems 1 and 2 are given in the Appendix.

One practical issue with $\widehat{\Sigma}^H$ is that the resulting matrix is not guaranteed to be positive semi-definite in finite sample situations. For example, it is possible to have nonzero $\widehat{\sigma}_{ij}^H$ while either $\widehat{\sigma}_{ii}^H$ or $\widehat{\sigma}_{jj}^H$ is zero. This is the main motivation for proposing a shrinkage estimator.

3.2. Sandwich estimator with shrinkage

We propose a new sandwich estimator for the covariance matrix. A shrinkage penalty is imposed to achieve a sparse structure. In particular, we propose minimizing the following objective function

$$Q_R(D) = \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \left(\widetilde{Y}_{ijk} - Z_{ij}^T D \widetilde{\Sigma} D Z_{ik} \right)^2 + \lambda \sum_{i=1}^{q} d_i \quad \text{subject to all } d_i \geq 0, \tag{3.2}$$

where $D = \mathrm{diag}(d_1, \ldots, d_q) \geq 0$, and $\lambda \geq 0$ is a tuning parameter which controls the amount of shrinkage. Let $\widehat{D} = \mathrm{diag}(\hat{d}_1, \ldots, \hat{d}_q)$ denote the minimizer of $Q_R(D)$. Once we obtain $\widehat{D}$, the final estimate of Σ is $\widehat{\Sigma} = \widehat{D}\widetilde{\Sigma}\widehat{D}$. The final estimate of the error variance is

$$\widehat{\sigma}_\varepsilon^2 = \max\left(0, \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{(\widetilde{Y}_{ijj} - Z_{ij}^T \widehat{\Sigma} Z_{ij})}{N}\right).$$

The sandwich structure of $D\widetilde{\Sigma}D$ assures the positive semi-definiteness of the estimator. In addition, if $\hat{d}_j = 0$, then the entire jth column and row in $\widehat{D}\widetilde{\Sigma}\widehat{D}$ are zero. That is, the variance of the jth random effect and its covariances with all the other random effects are zero.

3.2.1. Theoretical properties of sandwich estimators

For $j = 1, \ldots, q$, take $\psi_j = \sigma_{jj}^{1/2}$ and similarly, $\psi_{j,o} = \sigma_{jj,o}^{1/2}$, $\widetilde{\psi}_j = \widetilde{\sigma}_{jj}^{1/2}$, and $\widehat{\psi}_j = \widehat{\sigma}_{jj}^{1/2}$. An equivalent formulation of $Q_R(D)$ can be written as

$$Q_R(D) = \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} \Big(\widetilde{Y}_{ijk} - \sum_{r=1}^{q} \sum_{s=1}^{q} Z_{ijr} Z_{iks} \widetilde{\sigma}_{rs} d_r d_s \Big)^2 + \lambda \sum_{i=1}^{q} d_i,$$

where Z_{ijk} is the kth element of Z_{ij} and $\widetilde{\sigma}_{ij}$ is the (i,j)th entry of $\widetilde{\Sigma}$. From the relationship $\widehat{\sigma}_{ij} = \widetilde{\sigma}_{ij} \widehat{d}_i \widehat{d}_j$, we have $\widehat{d}_i = \widehat{\psi}_i / \widetilde{\psi}_i$, where $\widehat{\sigma}_{ij}$ is the (i,j)th entry of $\widehat{\Sigma}$. Therefore $\widehat{\sigma}_{ij}$ can be expressed as $\widetilde{\sigma}_{ij} (\widehat{\psi}_i \widehat{\psi}_j) / (\widetilde{\psi}_i \widetilde{\psi}_j)$. When $\widetilde{\sigma}_{ii}$ or $\widetilde{\sigma}_{jj} = 0$. We define $0/0$ as 0 throughout the paper. Accordingly, $Q_R(D)$ can be reparameterized as a function of $\psi = (\psi_1, \ldots, \psi_q)^T$,

$$Q_R(\psi) = \sum_{i=1}^{m} \sum_{j<k} \Big(\widetilde{Y}_{ijk} - \sum_{r=1}^{q} \sum_{s=1}^{q} Z_{ijr} Z_{iks} \frac{\widetilde{\sigma}_{rs}}{\widetilde{\psi}_r \widetilde{\psi}_s} \psi_r \psi_s \Big)^2 + \lambda \sum_{i=1}^{q} \frac{\psi_i}{\widetilde{\psi}_i}.$$

Note that $Q_R(\psi)$ involves $\widetilde{\kappa}$, and both $\widetilde{Y}_{ijk}$ and $\widetilde{\kappa}$ depend on $\widetilde{\beta}$. To emphasize this dependence, we write $\widetilde{\kappa} = \widetilde{\kappa}(\widetilde{\beta})$ and denote $Q_R(\psi)$ by $Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})$. Then the objective function $Q_R(D)$ in (3.2) is

$$Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) \equiv L_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) + \lambda \sum_{i=1}^{q} \frac{\psi_i}{\widetilde{\psi}_i}.$$

Let $\widehat{\psi}$ denote the minimizer of $Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})$. Using the asymptotic normality of the initial estimators $\widetilde{\beta}$ and $\widetilde{\Sigma}$, we can establish the theoretical properties of the final estimator. Without loss of generality, assume that the first b diagonal elements of Σ_o are nonzero and the remaining $q-b$ elements are zero. Accordingly, write $\psi_o = (\psi_{10}^T, 0^T)^T$ and $\widehat{\psi} = (\widehat{\psi}_1^T, \widehat{\psi}_2^T)^T$.

Theorem 3 (Root-m Consistency)**.** *Under the assumptions of Lemma* 2, $\lambda / \sqrt{m} \to 0$ *as* $m \to \infty$, *then the estimator* $\widehat{\Sigma}$ *satisfies* $\|\widehat{\kappa} - \kappa_o\| = O_p(m^{-1/2})$.

Theorem 4 (Selection Consistency and Asymptotic Normality)**.** *Under the assumptions of Lemma* 2, *suppose* $\lambda / \sqrt{m} \to 0$ *and* $\lambda \to \infty$ *as* $m \to \infty$. *Then, with probability approaching* 1, *the root-m consistent estimator* $(\widehat{\psi}_1^T, \widehat{\psi}_2^T)^T$ *satisfies the following:*

(a) *Sparsity :* $\widehat{\psi}_2 = 0$;

(b) *Asymptotic normality :* $m^{1/2}(\widehat{\psi}_1 - \psi_{10}) \to N(0, T)$ *in distribution as* m *goes to infinity, where* T *as defined at* (A.13) *in the Appendix.*

The proofs of Theorems 3 and 4 are given in the Appendix.

3.2.2. Computational algorithms

Assume λ is fixed. The estimation of $(\widehat{\Sigma}, \widehat{\sigma}_\varepsilon^2)$ can be implemented by the following algorithm.

Step 1 (obtain initial estimate of β): Fit a linear regression model $Y = X\beta + \eta$ by ordinary least squares. Denote the solution by $\widetilde{\beta}$.

Step 2 (obtain initial estimates of Σ and σ_ε^2): Compute $\widetilde{Y}_{ijk} = (Y_{ij} - X_{ij}^T\widetilde{\beta})(Y_{ik} - X_{ik}^T\widetilde{\beta})$ for all i, j, k. Then obtain $\widetilde{\Sigma}$ by minimizing $L_0(\Sigma)$ in (2.4), and compute $\widetilde{\sigma}_\varepsilon^2$.

Step 3 (obtain final estimates of Σ and σ_ε^2): Obtain $\widehat{D}$ by minimizing $Q_R(D)$ in (3.2), and compute $\widehat{\Sigma} = \widehat{D}\widetilde{\Sigma}\widehat{D}$ and $\widehat{\sigma}_\varepsilon^2$.

The minimization of $Q_R(D)$ in Step 3 is a nonlinear programming problem subject to a linear inequality constraint. We use a MATLAB toolbox TOMLAB for solving the optimization problem (Holmström (1999)). After numerous experiments in comparing the speed and accuracy of various solvers in TOMLAB, we found that the base module solver "clsSolve" is a good choice for our nonlinear least squares problems. Among seven optimization algorithms in "clsSolve", we selected the structured MBFGS method (Wang, Li, and Qi (2010)) because it is known to be the best theoretically and is expected to be the best in practice. In our numerical examples, these solvers proved to be stable and to demonstrate solid performance characteristics in various settings.

3.3. Tuning procedure

The choice of tuning parameters is crucial in practice, for example, of proper ν for the hard thresholding estimator in (3.1). We consider the value set $\{0, \widetilde{\sigma}_{11}, \widetilde{\sigma}_{12}, \ldots, \widetilde{\sigma}_{qq}\}$ and select the best value based on BIC. Since our estimator is based on the moment approach instead of the likelihood approach, we need to modify the traditional BIC for model selection. Denote the hard-thresholding estimator as $\widehat{\Sigma}_\nu^H$. A modified version of BIC is given as follows:

$$\mathrm{BIC}_R(\nu) = \frac{L_0(\widehat{\Sigma}_\nu^H)}{L_0(\widetilde{\Sigma})} + \frac{\log(N)}{N} \times df,$$

where df is the number of nonzeros on the diagonal of $\widehat{\Sigma}_\nu^H$. The first term in BIC_R is the ratio of the loss evaluated for the ν-selected model to the loss evaluated for the full model, which is an analog to the ratio of the residual sum of squares (RSS) defined in the standard BIC.

For the sandwich estimator, we also use the modified BIC method to select the optimal λ in (3.2) by simply replacing $L_0(\widehat{\Sigma}_\nu^H)$ with $L_0(\widehat{\Sigma}_\lambda)$. In Section 5, we examine the performance of this modified BIC.

Here we proposed some data-dependent way of choosing the tuning parameters. As pointed out by one referee, these tuning procedures are kind of ad hoc, and some theoretical results should be useful to justify their choices. A key question to answer is whether the selected tuning parameters satisfy the desired convergence rate required to assure the consistency and asymptotic normality of the estimators. To our best knowledge, this is still an open problem for model selection via shrinkage, even for simple linear models. This is an excellent topic for our future research. From a practical viewpoint, our simulation results in Section 5 suggest that the proposed criteria work quite well in various situations.

4. Fixed Effects Selection

We have focused on the estimation of $(\widehat{\Sigma}, \widehat{\sigma}_\varepsilon^2)$ and the selection of random effects for (1.1) so far. Next, we further extend our procedure to incorporate the selection of fixed effects and the estimation of nonzero regression coefficients.

Using a working variance-covariance matrix V, we can define a weighted estimator of β by

$$\min_\beta (Y - X\beta)^T V^{-1} (Y - X\beta).$$

If $V = I_N$, the solution to the above problem produces the ordinary least squares (OLS) estimator. If $V = Z\Sigma Z^T + \sigma_\varepsilon^2 I_N$, and Σ and σ_ε^2 were known, then we would obtain the generalized least squares (GLS) estimator. The GLS estimator is in theory the most efficient for the fixed effects β, but it is usually not practically available. In practice, since Σ and σ_ε^2 are generally unknown, we have to use their estimators. In this paper, we use our sparse estimators $(\widehat{\Sigma}, \widehat{\sigma}_\varepsilon^2)$, and correspondingly, $V = Z\widehat{\Sigma}Z^T + \widehat{\sigma}_\varepsilon^2 I_N$. We call the yielded estimator as the feasible generalized least squares (FGLS) estimator. The efficiency of these different estimators are compared via the simulations in Section 5.

The inverse matrix of $(Z\widehat{\Sigma}Z^T + \widehat{\sigma}_\varepsilon^2 I_N)$ can be decomposed into $Q^T Q$, where Q is an upper triangular matrix. Then, the weighted RSS can be written as

$$L_F(\beta \mid \widehat{\Sigma}, \widehat{\sigma}_\varepsilon^2) = (Y^* - X^*\beta)^T (Y^* - X^*\beta), \tag{4.1}$$

where $Y^* = QY$ and $X^* = QX$. By minimizing (4.1), we obtain the FGLS estimator $\widehat{\beta}_G$. To achieve sparsity in estimating β, we propose minimizing

$$Q_F(\beta) = L_F(\beta \mid \widehat{\Sigma}, \widehat{\sigma}_\varepsilon^2) + \tau \sum_{j=1}^{p} w_j |\beta_j|, \tag{4.2}$$

where $\tau \geq 0$ is a tuning parameter and w_j's are data-dependent weights. Here the subscript 'F' refers to the fixed effects selection. We propose to use $w_j = 1/|\widehat{\beta}_{G,j}|$, where $\widehat{\beta}_{G,j}$ is the jth component of $\widehat{\beta}_G$. Denote the solution to (4.2) by $\widehat{\beta}_\tau$. In summary, in order to select both random and fixed effects, we add the step of solving (4.2) into our algorithm proposed in Section 2.3:

Step 4 (final estimate of β): Obtain $\widehat{\beta}_\tau$ by minimizing $Q_F(\beta)$.

To select a proper τ for (4.2), we can use the selection criteria such as CV (cross-validation), GCV (generalized cross-validation) and BIC. Here we suggest modifying the BIC in a similar way to BIC_R in Section 3. Specifically, we use

$$\mathrm{BIC}_F(\tau) = \frac{L_F(\widehat{\beta}_\tau|\widehat{\Sigma}, \widehat{\sigma}^2_\varepsilon)}{L_F(\widehat{\beta}_G|\widehat{\Sigma}, \widehat{\sigma}^2_\varepsilon)} + \frac{\log(N)}{N} \times df,$$

where df is the number of nonzero $\widehat{\beta}_{\tau,i}$'s.

5. Numerical Studies

5.1. Simulation examples

In this section, we illustrate the performance of the proposed hard thresholding (HARD) and sandwich (SW) estimators under various scenarios and compare them with the MLE-based method of Bondell, Krishna, and Ghosh (2010) (denoted by BKG). We evaluate the performance of all the methods in three aspects: random effects selection, fixed effects selection, and the median of model errors (MME). Here the model error (ME) is given by $(\widehat{\beta} - \beta_o)^T E(XX^T)(\widehat{\beta} - \beta_o)$. In order to measure the variability of ME, we also present the median absolute deviation (MAD), that is, the median of the absolute deviations computed from the MEs. Four measures are used to assess the variable selection performance: the number of zero coefficients which are correctly estimated as zero (denoted by "CZ"), the number of nonzero coefficients which are incorrectly set to zero (denoted by "IZ"), the frequency of selecting the correct model (denoted by "C"), and the frequency of over-selecting variables (denoted by "O"). In each experimental setting, 100 data sets are simulated from the model, and we report the median performance over the 100 runs. As a baseline for comparison, we also present the results from the Oracle procedure (denoted by "Oracle"), assuming the true variance-covariance matrix for random effects is known.

We center the data before implementing each method, therefore it is not necessary to include a fixed intercept term in the model. For selecting fixed effects, we examined three different types of weights: $w_j = 1$, $1/|\widetilde{\beta}_j|$, and $1/|\widehat{\beta}_{G,j}|$. Recall that $\widetilde{\beta}_j$'s are the OLS estimates defined in Section 2, and $\widehat{\beta}_{G,j}$'s are the FGLS estimates defined in Section 4. Based on our numerous experiments, the

Table 1. Fixed and random effects selection and estimation results when $m = 100$ for Example 1.

Scenario	method	Random effect				Fixed effect				MME	MAD
		CZ	IZ	C	O	CZ	IZ	C	O		(ME)
1	HARD	2.80	0.00	82	18	5.38	0.00	72	28	0.011	0.008
	SW	2.76	0.00	83	17	5.89	0.00	89	11	0.009	0.005
	BKG	2.48	0.00	52	48	5.42	0.00	48	52	0.011	0.006
	Oracle	3	0	100	0	6	0	100	0	0.007	0.003
2	HARD	5.44	0.07	55	38	3.64	0.15	33	62	0.079	0.076
	SW	5.48	0.06	60	34	5.86	0.00	87	13	0.006	0.004
	BKG	5.05	0.00	28	72	5.59	0.00	64	36	0.008	0.005
	Oracle	6	0	100	0	6	0	100	0	0.004	0.002

weight $1/|\widehat{\beta}_{G,j}|$ consistently gives the best performance among the three. For parameter tuning in random effects selection, we use BIC_R to choose ν for the HARD estimator and λ for the SW estimator. For parameter tuning in fixed effects selection, we compared BIC_F, 5-fold CV, and various types of GCV and found that BIC_F yielded the best performance. Therefore, we report only the results obtained by using $w_j = 1/|\widehat{\beta}_{G,j}|$ and BIC_F for fixed effects selection in this paper. In addition, the BKG method applies the standard BIC as a tuning parameter selector.

Consider the following linear mixed model

$$Y = X\beta + \gamma_0 + Z\gamma_1 + \varepsilon, \tag{5.1}$$

where γ_0 is the random intercept, γ_1 is the random slope, β is fixed-effect coefficients, and $\gamma = (\gamma_0^T, \gamma_1^T)^T$ has mean 0 and variance-covariance matrix Σ. We consider two scenarios for sparse fixed and random effects. The second scenario is more complicated because it involves a larger number of covariates.

In all the examples, we set $\beta = (1, -0.9, 0.8, 0, 0, 0, 0, 0, 0)^T$. The covariates X and Z are generated from a normal distribution with mean 0 and an AR(1) covariance structure with $\rho = 0.5$; that is, the covariance between the ith and jth variables is $0.5^{|i-j|}$.

Scenario 1: $X \neq Z$, $n_i = 5$ for all i, Σ is a 5×5 matrix with $\sigma_{11} = 1$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 0.5$, and the remaining entries being 0.

Scenario 2: $X = Z$, $n_i = 10$ for all i, Σ is a 10×10 matrix with $\sigma_{11} = 1$, $\sigma_{55} = 1$, $\sigma_{15} = \sigma_{51} = -0.5$, $\sigma_{66} = 0.8$, $\sigma_{77} = 0.6$, $\sigma_{67} = \sigma_{76} = 0.2$, and the remaining entries being 0.

We designed three examples which correspond to three different types of error distributions: Gaussian, t_5, and centered exponential.

Table 2. Fixed and random effects selection and estimation results for Example 2.

Scenario	m	method	Random effect				Fixed effect				MME	MAD
			CZ	IZ	C	O	CZ	IZ	C	O		(ME)
1	100	HARD	2.82	0.00	84	16	5.55	0.02	78	21	0.019	0.011
		SW	2.72	0.00	80	20	5.87	0.00	89	11	0.013	0.008
		BKG	2.32	0.00	42	58	5.46	0.00	54	46	0.018	0.011
		Oracle	3	0	100	0	6	0	100	0	0.011	0.006
	200	HARD	2.79	0.00	81	19	5.88	0.00	91	9	0.006	0.004
		SW	2.81	0.00	84	16	5.94	0.00	95	5	0.007	0.004
		BKG	2.39	0.00	50	50	5.66	0.00	68	32	0.011	0.007
		Oracle	3	0	100	0	6	0	100	0	0.005	0.003
2	100	HARD	5.50	0.06	63	32	4.48	0.09	52	45	0.029	0.025
		SW	5.59	0.13	58	29	5.90	0.00	90	10	0.008	0.005
		BKG	4.91	0.00	29	71	5.62	0.00	73	27	0.013	0.006
		Oracle	6	0	100	0	6	0	100	0	0.005	0.003
	200	HARD	5.62	0.00	68	32	5.65	0.00	85	15	0.005	0.002
		SW	5.65	0.00	73	27	5.96	0.00	96	4	0.004	0.002
		BKG	5.16	0.00	39	61	5.45	0.00	57	43	0.007	0.005
		Oracle	6	0	100	0	6	0	100	0	0.003	0.002

Example 1. (Gaussian Example) Let ε have a standard Gaussian distribution. Table 1 summarizes the simulation results for the Gaussian distribution example in the case of $m = 100$. In both scenarios, the SW estimator gives the highest selection frequency in identifying the nonzero random effects and fixed effects. The BKG approach never misses any important random or fixed effects, but it tends to retain some unimportant variables in the model. With regard to the model estimation, the SW estimator and BKG method work similarly in both scenarios. The HARD estimator works quite well in Scenario 1, but gives a large MME in Scenario 2. Recall that the variance-covariance matrix given by the HARD estimator is not guaranteed to be non-negative definite which can lead to unreliable selection and estimation for the fixed effects.

Example 2. (t_5 Example) In this example, we let ε follow a t distribution with 5 degrees of freedom. Table 2 shows the results from the t_5 distribution when $m = 100$ and 200 for both scenarios. Again, the SW estimator gives the highest selection frequency in identifying the nonzero random effects and fixed effects. The BKG method tends to produce a larger-sized model but never misses any important term. In terms of model estimation, the SW estimator overall gives the best MME under all the settings, the BKG method is the second best when $m = 100$, and the HARD estimator is the second best when $m = 200$. It is observed that the variance matrix estimate given by the HARD estimator tends

Table 3. Fixed and random effects selection and estimation results for Example 3.

Scenario	m	method	Random effect				Fixed effect				MME	MAD
			CZ	IZ	C	O	CZ	IZ	C	O		(ME)
1	100	HARD	2.70	0.02	75	24	5.10	0.03	63	36	0.017	0.012
		SW	2.58	0.00	72	28	5.86	0.00	88	12	0.010	0.005
		BKG	2.33	0.00	42	58	5.49	0.00	58	42	0.013	0.007
		Oracle	3	0	100	0	6	0	100	0	0.006	0.004
	200	HARD	2.80	0.00	83	17	5.64	0.00	85	15	0.005	0.004
		SW	2.78	0.00	83	17	5.90	0.00	90	10	0.005	0.004
		BKG	2.41	0.00	45	55	5.72	0.00	76	24	0.004	0.003
		Oracle	3	0	100	0	6	0	100	0	0.003	0.002
2	100	HARD	5.53	0.05	61	34	3.55	0.18	34	60	0.144	0.142
		SW	5.57	0.05	64	31	5.84	0.00	85	15	0.006	0.004
		BKG	4.92	0.00	26	74	5.58	0.00	67	33	0.008	0.005
		Oracle	6	0	100	0	6	0	100	0	0.004	0.003
	200	HARD	5.59	0.01	67	32	4.27	0.00	55	45	0.007	0.006
		SW	5.55	0.00	65	35	5.92	0.00	93	7	0.003	0.002
		BKG	5.18	0.00	41	59	5.27	0.00	47	53	0.004	0.002
		Oracle	6	0	100	0	6	0	100	0	0.002	0.001

to suffer from being negative definite, but the results improve when the sample size increases.

Example 3. (Centered Exponential Example) In this example, we assume that ε has a centered exponential distribution with mean 0 and variance 1; that is, $\varepsilon \sim Exp(1)-1$. Table 3 summarizes the results from this example when $m = 100$ and 200. Overall, the performance of the three methods is consistent with that in the previous two examples. The SW estimator is the best for model selection, giving the highest selection frequency in identifying the nonzero random effects and fixed effects. The BKG method tends to produce a larger model. In terms of model estimation, the SW estimator overall gives the best MME under all the settings, the BKG method is the second best when $m = 100$, and the HARD estimator is the second best when $m = 200$.

In summary, the SW estimator is consistently the best in terms of both model selection and parameter estimation in all the three examples. The BKG method gives comparable performance with SW estimator in Example 1 where the error variable follows a Gaussian distribution, but when the error is not Gaussian, the BKG method misspecifies the likelihood function so its MME is worse than the SW estimator in Examples 2 and 3. The HARD estimator works well when the sample size is large. In addition, another advantage of the moment-based procedure is the computational cost. Based on our experience, the computation

Table 4. Median of mean squared errors (MMSE) of the estimated fixed-effect coefficients. MADs of the MSEs are given in parentheses.

Example#	method	Scenario 1			Scenario 2		
		m=100	m=200	m=300	m=100	m=200	m=300
1	OLS	0.079	0.039	0.030	0.082	0.041	0.027
		(0.024)	(0.015)	(0.012)	(0.028)	(0.014)	(0.009)
	FGLS	0.040	0.021	0.013	0.038	0.019	0.013
		(0.014)	(0.007)	(0.005)	(0.016)	(0.008)	(0.006)
	GLS	0.040	0.020	0.014	0.037	0.019	0.013
		(0.012)	(0.007)	(0.005)	(0.015)	(0.008)	(0.006)
2	OLS	0.093	0.048	0.036	0.084	0.047	0.029
		(0.029)	(0.013)	(0.013)	(0.037)	(0.017)	(0.011)
	FGLS	0.058	0.029	0.022	0.053	0.026	0.017
		(0.021)	(0.010)	(0.008)	(0.019)	(0.009)	(0.007)
	GLS	0.057	0.029	0.022	0.056	0.027	0.016
		(0.020)	(0.011)	(0.008)	(0.020)	(0.010)	(0.007)
3	OLS	0.077	0.042	0.025	0.084	0.043	0.030
		(0.033)	(0.014)	(0.010)	(0.036)	(0.013)	(0.010)
	FGLS	0.044	0.019	0.012	0.037	0.018	0.013
		(0.020)	(0.007)	(0.004)	(0.015)	(0.007)	(0.005)
	GLS	0.039	0.019	0.012	0.038	0.019	0.014
		(0.018)	(0.007)	(0.005)	(0.014)	(0.007)	(0.005)

speed of the HARD and SW estimators is much faster than that of the BKG method which employs the EM algorithm for computation.

One ultimate purpose of selecting random effects in linear mixed models is to improve the estimation efficiency of the fixed-effect β in the final model. In order to show that the efficiency is gained in fixed-effect estimation due to a proper random-effects selection, we consider the measure mean squared error (MSE)

$$MSE(\widehat{\beta}) = (\widehat{\beta} - \beta_o)^T(\widehat{\beta} - \beta_o),$$

where $\widehat{\beta}$ is any estimator and β_o is the true parameter, which can be used to compare different estimators in estimating fixed effects. In Table 4, we report the median of MSE (MMSE) and the corresponding MAD (median absolute deviation) for three weighted estimators under both scenarios with $m = 100$, 200, and 300. The OLS estimator is the standard ordinary least squares obtained by assuming working independence, which is essentially the model fit without random effects selection. The FGLS estimator is obtained by using the SW estimate $\widehat{\Sigma}$ and its corresponding $\widehat{\sigma}_\varepsilon^2$ to construct the weight. Finally, the GLS estimator is computed by using the true underlying variance-covariance matrix and true error variance. All estimators are computed without fixed effects selection and defined in Section 4. From Table 4, we observe that our FGLS estimator gives

Table 5. Estimates for the selected model for the Amsterdam Growth and Health Study data. Standard errors for fixed effects are given in parentheses.

Random effect Variable	REML estimate	Moment-based initial estimate	HARD estimate	SW estimate
$\text{var}(\gamma_{0i})$	0.522	0.405	0.405	0.347
$\text{var}(\gamma_{1i})$	0.014	0.026	0	0.006
$\text{var}(\gamma_{2i})$	0.042	0.005	0	0
$\text{var}(\gamma_{3i})$	0.073	0.149	0.149	0
$\text{var}(\gamma_{4i})$	0.961	0.668	0.668	0.624
$\text{var}(\gamma_{5i})$	0.037	0.016	0	0
$\text{var}(\varepsilon_i)=\sigma_\varepsilon^2$	0.197	0.220	0.268	0.253
Fixed effect Variable	**REML estimate**	**OLS estimate**	**HARD estimate**	**SW estimate**
X_1	-0.027 (0.050)	0.018 (0.030)	0	0
X_2	0.190 (0.035)	0.270 (0.029)	0.174	0.165
X_3	-0.106 (0.060)	-0.007 (0.067)	0	0
X_4	0.126 (0.072)	-0.017 (0.044)	0	0
X_5	0.169 (0.023)	0.149 (0.027)	0.156	0.167

great improvement in terms of MSE than the OLS, and its performance is very close to the oracle procedure GLS in most examples. These findings suggest that our random effect selection procedure is effective in producing efficiency gain for the fixed-effect estimation in linear mixed models.

5.2. Real example

For illustration, we consider a real data example: the Amsterdam Growth and Health Study (Kemper (1995)). The goal of this study is to investigate the relationship between lifestyle and health in adolescence and young adulthood. The response variable Y is the total serum cholesterol measured over six time points. There are five explanatory variables: X_1 is the fitness level at baseline measured as maximum oxygen uptake on a treadmill, X_2 is body fatness estimated by the sum of the thickness of four skinfolds, X_3 is smoking behavior (0=no,1=yes), X_4 is gender (0=female, 1=male), and X_5 is the measurement time coded as $(1, 2, \ldots, 6)$. The number of subjects is 147, and the total number of observations is 882. Twisk (2003) analyzed this data set using regression techniques for longitudinal data. Azari, Li, and Tsai (2006) conducted the fixed effects selection by including some quadratic and interaction terms as fixed effects. We fit the linear mixed model with all the five covariates for both fixed and random effects. We compare the new estimators with the REML estimator.

We centered the response variable Y and standardized all the inputs, so the fitted model does not include an intercept for the fixed effects, but a random

intercept is allowed. The BIC_R and BIC_F were used to choose the tuning parameters for random and fixed effects selection, respectively. In Step 4, we used the FGLS estimates $\widehat{\beta}_G$ for constructing the weights in (4.2). For comparison, we also fitted the full model by including all explanatory variables as fixed and random effects. We used the `lmer` function from `lme4` package in R, which used the REML estimation based on the normality assumption.

The real data analysis results are summarized in Table 5. Both HARD and SW estimators identify X_2 and X_5 as important fixed effects, which is consistent with the results from the REML estimation. In the analysis of REML, X_2 and X_5 have t-statistics of 5.50 and 7.26, respectively, which are the only two highly significant fixed effects. The fitted coefficients of the HARD and SW estimates are also similar to those obtained using OLS and REML methods. For random effects selection, the HARD estimator selects the random intercept, X_3 and X_4, and SW estimator selects the random intercept, X_1 and X_4. Due to a smaller number of random effects in the final model, the remaining unexplained variance is absorbed into the error variance, and therefore the error variances given by the HARD and SW estimators are slightly larger than that of the REML estimator.

Though the HARD estimator performs well in general for random effects selection, we note that the estimated covariance matrix is

$$\widehat{\Sigma}^H = \begin{pmatrix} 0.41 & 0.07 & -0.03 & -0.05 & -0.39 & 0 \\ 0.07 & 0 & 0 & 0.03 & -0.11 & 0 \\ -0.03 & 0 & 0 & 0 & 0 & 0 \\ -0.05 & 0.03 & 0 & 0.15 & -0.09 & -0.04 \\ -0.39 & -0.11 & 0 & -0.09 & 0.67 & 0.07 \\ 0 & 0 & 0 & -0.04 & 0.07 & 0 \end{pmatrix},$$

which is not an appropriate variance-covariance matrix. For example, even though $\widehat{\sigma}_{22}$ is zero, the entries $\widehat{\sigma}_{21}$ and $\widehat{\sigma}_{12}$ are not. Obviously, this matrix is not positive semi-definite. In practice, we recommend using the SW method for estimating the sparse variance-covariance matrix for random effects.

6. Discussion

We propose a new class of robust thresholding and shrinkage approaches for random and fixed effects selection in linear mixed models. Compared with existing methods, the new procedures do not rely on any distributional assumptions for random effects and errors, and are hence more robust for non-normal correlated data. The theoretical and numerical results suggest that the proposed methods provide a promising tool for the analysis of clustered data in practice.

The implementation of the proposed methods requires solving nonlinear optimization problems. We suggest some feasible strategies and solvers for computing the solutions, but the computation efficiency could be further improved. We also tried a quadratic approximation to the nonlinear objective function and iteratively solved the quadratic programming problem. Our preliminary analysis suggests that this approximation technique can greatly speed up the computation for our procedure and deserves further investigation in the future.

Acknowledgement

We thank Editor, Associate Editor, and both reviewers for their constructive comments to improve the paper. We acknowledge support from National Science Foundation grant DMS-0645293 and National Institutes of Health RO1 CA140632, P01 CA142538, R01 CA085848.

Appendix

Proof of Theorem 1. For a $q \times q$ symmetric matrix A, the following inequalities are satisfied:

$$\|A\|_2 \le \max_{1\le i\le q} \sum_{j=1}^{q} |a_{ij}| \qquad \text{and} \qquad \|A\|_F \le q \max_{1\le i,j\le q} |a_{ij}|.$$

Based on these inequalities, to prove the results in Theorem 1, it only needs to show the component-wise root-m consistency of $\widehat{\Sigma}_\nu^H$. Note that

$$\begin{aligned}
&\sqrt{m}|\widehat{\sigma}_{ij}^H - \sigma_{ij,o}| \\
&\quad = \sqrt{m}|\widehat{\sigma}_{ij}^H| I(\sigma_{ij,o} = 0) + \sqrt{m}|\widehat{\sigma}_{ij}^H - \sigma_{ij,o}| I(\sigma_{ij,o} \neq 0) \\
&\quad \le \sqrt{m}|\widetilde{\sigma}_{ij}| I(\sigma_{ij,o} = 0) + \sqrt{m}|\widetilde{\sigma}_{ij} - \sigma_{ij,o} - \widetilde{\sigma}_{ij} I(|\widetilde{\sigma}_{ij}| \le \nu)| I(\sigma_{ij,o} \neq 0) \\
&\quad \le O_p(1) + \sqrt{m}|\widetilde{\sigma}_{ij} - \sigma_{ij,o}| I(\sigma_{ij,o} \neq 0) + \sqrt{m}|\widetilde{\sigma}_{ij}| I(|\widetilde{\sigma}_{ij}| \le \nu) I(\sigma_{ij,o} \neq 0) \\
&\quad = O_p(1) + \sqrt{m}|\widetilde{\sigma}_{ij}| I(|\widetilde{\sigma}_{ij}| \le \nu) I(\sigma_{ij,o} \neq 0).
\end{aligned}$$

In addition, for $\sigma_{ij,o} \neq 0$, $P(\sqrt{m}|\widetilde{\sigma}_{ij}| I(|\widetilde{\sigma}_{ij}| \le \nu) = 0) = P(|\widetilde{\sigma}_{ij}| > \nu)$. Without loss of generality, assume $\sigma_{ij,o} > 0$. Then

$$\begin{aligned}
P(|\widetilde{\sigma}_{ij}| > \nu) = P(\sqrt{m}(\widetilde{\sigma}_{ij} - \sigma_{ij,o}) > \sqrt{m}(\nu - \sigma_{ij,o})) + P(\sqrt{m}(\widetilde{\sigma}_{ij} - \sigma_{ij,o}) \\
< \sqrt{m}(-\nu - \sigma_{ij,o})) \to 1.
\end{aligned}$$

Therefore, $\sqrt{m}|\widetilde{\sigma}_{ij}| I(|\widetilde{\sigma}_{ij}| \le \nu) I(\sigma_{ij,o} \neq 0) = o_p(1)$. The same result also holds for $\sigma_{ij,o} < 0$. Then it follows $\sqrt{m}|\widehat{\sigma}_{ij}^H - \sigma_{ij,o}| = O_p(1)$ for any i, j.

Proof of Theorem 2. We follow similar steps to prove Theorem 2 from Rothman, Levina, and Zhu (2009). If $\widehat{\sigma}_{ij}^H \neq 0$ and $\sigma_{ij,o} = 0$, then $|\widetilde{\sigma}_{ij} - \sigma_{ij,o}| > \nu$

holds. Using Chebyshev's inequality and Lemma 2, we have

$$P\Big\{\sum_{i=1}^{q}\sum_{j=1}^{q} I\big(\widehat{\sigma}_{ij}^{H}\neq 0,\sigma_{ij,o}=0\big)>0\Big\}$$
$$\leq P\Big\{\sum_{i,j} I\big(|\widetilde{\sigma}_{ij}-\sigma_{ij,o}|>\nu\big)>0\Big\}\leq P\Big\{\max_{1\leq i,j\leq q}|\widetilde{\sigma}_{ij}-\sigma_{ij,o}|>\nu\Big\}\leq\frac{\phi}{m\nu^2}\to 0,$$

where ϕ is the maximum diagonal element of $A_2^{-1}B_2A_2^{-1}$. Hence, with probability approaching 1, $\widehat{\sigma}_{ij}^{H}=0$ for all (i,j) such that $\sigma_{ij,o}=0$.

If $\widehat{\sigma}_{ij}^{H}$ and nonzero $\sigma_{ij,o}$ have different signs, then we have $|\widetilde{\sigma}_{ij}-\sigma_{ij,o}|>\delta-\nu$ for some $\delta>0$. Using Lemma 2, we have

$$P\Big\{\sum_{i,j} I\big(\widehat{\sigma}_{ij}^{H}\leq 0,\sigma_{ij,o}>0\quad or\quad \widehat{\sigma}_{ij}^{H}\geq 0,\sigma_{ij,o}<0\big)>0\Big\}$$
$$\leq P\Big\{\sum_{i,j} I\big(|\widetilde{\sigma}_{ij}-\sigma_{ij,o}|>\delta-\nu\big)>0\Big\}$$
$$\leq P\Big\{\max_{1\leq i,j\leq q}|\widetilde{\sigma}_{ij}-\sigma_{ij,o}|>\delta-\nu\Big\}\leq\frac{\phi}{m(\delta-\nu)^2}\to 0.$$

Hence, with probability approaching 1, nonzero $\sigma_{ij,o}$ and $\widehat{\sigma}_{ij}^{H}$ have the same sign.

Proof of Theorem 3. To prove Theorem 3, it is sufficient to show that for any given $\epsilon>0$, there exists a large constant C such that

$$\mathrm{pr}\Big[\inf_{\psi\in B_m(C)} Q_R(\psi;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})>Q_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big]\geq 1-\epsilon, \tag{A.1}$$

where the C-ball $B_m(C)=\{\psi:\psi=\psi_o+m^{-1/2}u,\|u\|\leq C\}$. The derivatives of Q_R and L_R with respect to ψ or $\widetilde{\kappa}(\widetilde{\beta})$ are the right derivatives because ψ is defined in the set of non-negative real q-vectors, denoted by $\mathbb{R}^{q+}$.

From the Taylor expansion of L_R around $\psi=\psi_o$, we have

$$Q_R(\psi_o+m^{-1/2}u;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})-Q_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})$$
$$=L_R(\psi_o+m^{-1/2}u;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})-L_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})+\lambda\sum_{i=1}^{q}\Big(\frac{\psi_{i,o}+m^{-1/2}u_i}{\widetilde{\psi}_i}-\frac{\psi_{i,o}}{\widetilde{\psi}_i}\Big)$$
$$=\Big(\frac{u}{\sqrt{m}}\Big)^T S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})+\frac{1}{2}\Big(\frac{u}{\sqrt{m}}\Big)^T\nabla S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big(\frac{u}{\sqrt{m}}\Big)+\lambda\sum_{i=1}^{q}\frac{u_i}{\widetilde{\psi}_i\sqrt{m}}, \tag{A.2}$$

where

$$S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})=\frac{\partial L_R(\psi;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})}{\partial\psi}\bigg|_{\psi_o},\nabla S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})=\frac{\partial^2 L_R(\psi;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})}{\partial\psi\partial\psi^T}\bigg|_{\psi_o}.$$

By the law of large numbers,

$$\frac{1}{m}\sum_{i,j,k} e_{ijk} \to 0. \tag{A.3}$$

Therefore, we can obtain

$$\frac{1}{m}\Big(\frac{\partial}{\partial\widetilde{\kappa}^T}S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big|_{\kappa_o,\beta_o}\Big) \to E \quad \text{and} \quad \frac{1}{m}\Big(\frac{\partial}{\partial\widetilde{\beta}^T}S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big|_{\kappa_o,\beta_o}\Big) \to 0, \tag{A.4}$$

where E is a $q \times q(q+1)/2$ matrix.

Using (A.3), (A.4), and Lemmas 1 and 2, the first-order Taylor expansion around $\widetilde{\kappa}(\widetilde{\beta}) = \kappa_o$ and $\widetilde{\beta} = \beta_o$ yields

$$\begin{aligned}
&S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\\
&= S_R(\psi_o;\kappa_o,\beta_o) + \Big(\frac{\partial}{\partial\widetilde{\kappa}^T}S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big|_{\kappa_o,\beta_o}\Big)(\widetilde{\kappa}(\beta_o)-\kappa_o)\\
&\quad+\Big(\frac{\partial}{\partial\widetilde{\kappa}^T}S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big|_{\kappa_o,\beta_o}\Big)\Big(\frac{\partial}{\partial\widetilde{\beta}^T}\widetilde{\kappa}(\widetilde{\beta})\Big|_{\beta_o}\Big)(\widetilde{\beta}-\beta_o)\\
&\quad+\Big(\frac{\partial}{\partial\widetilde{\beta}^T}S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})\Big|_{\kappa_o,\beta_o}\Big)(\widetilde{\beta}-\beta_o) + o_p(\|\widetilde{\kappa}(\widetilde{\beta})-\kappa_o\|) + o_p(\|\widetilde{\beta}-\beta_o\|)\\
&= S_R(\psi_o;\kappa_o,\beta_o) + o_p(m^{1/2})
\end{aligned} \tag{A.5}$$

and

$$\nabla S_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta}) = \nabla S_R(\psi_o;\kappa_o,\beta_o) + o_p(m^{1/2}). \tag{A.6}$$

The tth component of $S_R(\psi_o;\kappa_o,\beta_o)$ is

$$-2\sum_{i,j,k} e_{ijk}\sum_{l=1}^{q}(Z_{ijt}Z_{ikl}+Z_{ijl}Z_{ikt})\frac{\sigma_{tl,o}}{\psi_t},$$

where $\sigma_{tl,o}$ is the (t,l)th element of Σ_o. Hence, $S_R(\psi_o;\kappa_o,\beta_o)$ can be expressed as $\sum_{i=1}^m W_i^T e_i$, where e_i is a column vector consisting of e_{ijk}'s. By the central limit theorem, we have

$$\frac{1}{\sqrt{m}}S_R(\psi_o;\kappa_o,\beta_o) \to N(0,F) \quad \text{as} \quad m\to\infty, \tag{A.7}$$

where $F = \lim_{m\to\infty}\sum_{i=1}^m W_i^T\mathrm{cov}(e_i)W_i/m$ and is a $q\times q$ positive semi-definite matrix. This implies that $S_R(\psi_o;\kappa_o,\beta_o)/\sqrt{m} = O_p(1)$. In addition, by the law of large numbers, it follows that

$$\frac{1}{m}\nabla S_R(\psi_o;\kappa_o,\beta_o) = H + o_p(1), \tag{A.8}$$

where H is a $q \times q$ positive semi-definite matrix. Then, using (A.5) and (A.6), the first and second terms of (A.2) become

$$\frac{u^T}{\sqrt{m}}\Big(S_R(\psi_o;\kappa_o,\beta_o)+o_p(m^{1/2})\Big)+\frac{1}{2}\frac{u^T}{\sqrt{m}}\Big(\nabla S_R(\psi_o;\kappa_o,\beta_o)+o_p(m^{1/2})\Big)\frac{u}{\sqrt{m}}$$
$$=u^TO_p(1)+\frac{1}{2}u^T\big(H+o_p(1)\big)u. \tag{A.9}$$

We assume that the first b diagonal elements of ψ_o are nonzero. For $i=1,2,\ldots,b$, we have

$$\frac{1}{\widetilde{\psi}_i}=\frac{1}{\psi_{i,o}}-\frac{1}{\psi_{i,o}^2}(\widetilde{\psi}_i-\psi_{i,o})+o_p(|\widetilde{\psi}_i-\psi_{i,o}|)=\frac{1}{\psi_{i,o}}+O_p(m^{-1/2}).$$

Hence, if $\lambda/\sqrt{m}=o(1)$, then

$$\frac{\lambda}{\sqrt{m}}\sum_{i=1}^{q}\frac{u_i}{\widetilde{\psi}_i}\geq\frac{\lambda}{\sqrt{m}}\sum_{i=1}^{b}\frac{u_i}{\widetilde{\psi}_i}=\frac{\lambda}{\sqrt{m}}\sum_{i=1}^{b}u_i\Big(\frac{1}{\psi_{i,\,o}}+O_p(m^{-1/2})\Big)=\sum_{i=1}^{b}u_iO_p(1). \tag{A.10}$$

Combining (A.9) and (A.10), it follows that

$$Q_R(\psi_o+m^{-1/2}u;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})-Q_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})$$
$$\geq u^TO_p(1)+\frac{1}{2}u^T\big(H+o_p(1)\big)u+\sum_{i=1}^{b}u_iO_p(1).$$

If we choose a sufficiently large constant C, the second term dominates the other terms. Therefore, (A.1) holds. This means that $m^{1/2}(\widehat{\psi}_i-\psi_{i,\,o})=O_p(1)$. By the Delta method, it is easy to show that $m^{1/2}(\widehat{\psi}_i-\psi_{i,\,o})=O_p(1)$. Since $\widehat{\sigma}_{ij}=\widetilde{\sigma}_{ij}(\widehat{\psi}_i\widehat{\psi}_j/(\widetilde{\psi}_i\widetilde{\psi}_j)$, the root-$m$ consistency also holds for the off-diagonal elements of $\widehat{\Sigma}$. Hence, $m^{1/2}(\widehat{\sigma}_{ij}-\sigma_{ij,\,o})=O_p(1)$ for $i,j=1,\ldots,q$.

Proof of Theorem 4. We first prove that $\widehat{\psi}_2=0$ with probability approaching 1. It is enough to show that for any sequence ψ_1 satisfying $\|\psi_1-\psi_{10}\|=O_p(m^{-1/2})$ and for any constant C,

$$Q_R\big((\psi_1,0);\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta}\big)=\min_{\|\psi_2\|\leq Cm^{-1/2}}Q_R\big((\psi_1,\psi_2);\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta}\big).$$

From (A.5), (A.6), (A.8), and the second-order Taylor expansion around $\psi=\psi_o$, we obtain

$$L_R(\psi;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})=L_R(\psi_o;\widetilde{\kappa}(\widetilde{\beta}),\widetilde{\beta})+(\psi-\psi_o)^TS_R(\psi_o;\kappa_o,\beta_o)$$
$$+\frac{1}{2}(\psi-\psi_o)^T\nabla S_R(\psi_o;\kappa_o,\beta_o)(\psi-\psi_o)+o_p(1).$$

For $t = b+1, \ldots, q$,

$$\begin{aligned}\frac{\partial}{\partial \psi_t} Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) &= \frac{\partial}{\partial \psi_t} L_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) + \lambda \frac{1}{\widetilde{\psi}_t} \\ &= [S_R(\psi_o; \kappa_o, \beta_o)]_t + [\nabla S_R(\psi_o; \kappa_o, \beta_o)]_t \psi_t + \frac{m^{1/2}\lambda}{|O_p(1)|} \\ &= m^{1/2}\Big(O_p(1) + \frac{\lambda}{|O_p(1)|}\Big),\end{aligned}$$

where $[A]_t$ denotes the tth element if A is a vector, and the tth row vector if A is a matrix. Since $\lambda \to \infty$, $\partial Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})/\partial \psi_t > 0$. Hence, $Q_R((\psi_1, 0); \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) < Q_R((\psi_1, \psi_2); \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})$. This completes the proof.

Next we prove part (b) of Theorem 4, which deals with the asymptotic normality of $\widehat{\psi}_1$. Define $S_{1R}(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})$ be the first b elements of $S_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})$. From part (a), we can write $\widehat{\psi} = (\widehat{\psi}_1^T, 0^T)^T$. Hence, we have

$$0 = \frac{\partial}{\partial \psi_1} Q_R(\psi; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})\Big|_{\psi = (\widehat{\psi}_1^T, 0^T)^T} = S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) + \lambda\Big(\frac{1}{\widetilde{\psi}_1}, \ldots, \frac{1}{\widetilde{\psi}_b}\Big)^T.$$

From the first-order Taylor expansion around $(\widehat{\psi}, \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) = (\psi_o, \kappa_o, \beta_o)$, we have

$$\begin{aligned}&S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta}) \\ &\quad = S_{1R}(\psi_o; \kappa_o, \beta_o) + \Big(\frac{\partial}{\partial \widehat{\psi}_1^T} S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})\Big|_{\psi_o, \kappa_o, \beta_o}\Big)(\widehat{\psi}_1 - \psi_{10}) \\ &\qquad + \Big(\frac{\partial}{\partial \widetilde{\kappa}^T} S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})\Big|_{\psi_o, \kappa_o, \beta_o}\Big)\big(\widetilde{\kappa}(\beta_o) - \kappa_o\big) \\ &\qquad + \Big(\frac{\partial}{\partial \widetilde{\beta}^T} S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})\Big|_{\psi_o, \kappa_o, \beta_o}\Big)(\widetilde{\beta} - \beta_o) \\ &\qquad + \Big(\frac{\partial}{\partial \widetilde{\kappa}^T} S_{1R}(\widehat{\psi}; \widetilde{\kappa}(\widetilde{\beta}), \widetilde{\beta})\Big|_{\psi_o, \kappa_o, \beta_o}\Big)\Big(\frac{\partial}{\partial \widetilde{\beta}^T} \widetilde{\kappa}(\widetilde{\beta})\Big|_{\beta_o}\Big)(\widetilde{\beta} - \beta_o) + o_p(m^{1/2}). \quad \text{(A.11)}\end{aligned}$$

Since $\lambda/\sqrt{m} \to 0$ and $\widetilde{\psi}_i \xrightarrow{p} \psi_{i,o} \neq 0$ for $i = 1, \ldots, b$, we have

$$\frac{1}{\sqrt{m}} \lambda\Big(\frac{1}{\widetilde{\psi}_1}, \ldots, \frac{1}{\widetilde{\psi}_b}\Big)^T \to 0. \quad \text{(A.12)}$$

Putting (A.3), (A.4), (A.7), (A.8), (A.12), and Lemmas 1 and 2 into (A.11), we have, by Slutsky's theorem,

$$m^{1/2}(\widehat{\psi}_1 - \psi_{10}) \to N\big(0, T\big), \quad \text{(A.13)}$$

where $T = H_1^{-1}(F_1 + E_1 A_2^{-1} B_2 A_2^{-1} E_1^T) H_1^{-1}$ with E_1, F_1, and H_1 being the first $b \times q(q+1)/2$, $b \times b$, and $b \times b$ submatrices of E, F, and H, respectively.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255-265.

Azari, R., Li, L. and Tsai, C.-L. (2006). Longitudinal data model selection. *Comput. Statist. Data Anal.* **50**, 3053-3066.

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

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* **64**, 115-123.

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

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

Diggle, P., Liang, K. and Zeger, S. (1994). *Analysis of Longitudinal Data.* Oxford University Press, New York.

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

Holmström, K. (1999). The TOMLAB optimization environment in MATLAB. *Adv. Model. Optim.* **1**, 47-69.

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

Kemper, H. (1995). *The Amsterdam growth study: A longitudinal analysis of health, fitness and lifestyle.* In *HK Sport Science Monograph Series*, **6**, Human Kinetics Publishers, Champaign IL.

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

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

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *J. Amer. Statist. Assoc.* **83**, 1014-1022.

Löfberg, J. (2004). YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proc. of the CACSD Conf.*, Taipei, Taiwan.

Niu, F. and Pu, P. (2006). Selecting mixed-effects models based on generalized information criterion. *J. Multivariate Anal.* **97**, 733-758.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in regression problems. *Biometrika* **76**, 369-374.

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104**, 177-186.

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

Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optim. Meth. Softw.* **11-12**, 625-653.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Twisk, J. W. (2003). *Applied Longitudinal Data Analysis for Epidemiology-Practical Guide.* Cambridge University press, New York.

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

Wang, F., Li, D.-H. and Qi, L. (2010). Global convergence of Gauss-Newton-MBFGS method for solving the nonlinear least squares problem. *Adv. Model. Optim.* **12(1)**, 1-20.

Wang, F., Li, D.-H. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the lad-lasso. *J. Bus. Econ. Stat.* **20**, 347-355.

Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Comm. Statist. Simul. Comput.* **22**, 1079-1106.

Zhang, H. 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. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

E-mail: mahn@ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

E-mail: hzhang@stat.ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

E-mail: lu@stat.ncsu.edu

(Received February 2011; accepted September 2011)