Published in final edited form as:

J Comput Graph Stat. 2013 ; 22(2): 341–355. doi:10.1080/10618600.2012.681219.

Fixed and Random Effects Selection by REML and Pathwise Coordinate Optimization

Bingqing Lin,

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371

Zhen Pang, and

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371

Jiming Jiang

Department of Statistics, University of California, Davis, CA 95616

Zhen Pang: ZPang@ntu.edu.sg

Abstract

We propose a two-stage model selection procedure for the linear mixed-effects models. The procedure consists of two steps: First, penalized restricted log-likelihood is used to select the random effects, and this is done by adopting a Newton-type algorithm. Next, the penalized log-likelihood is used to select the fixed effects via pathwise coordinate optimization to improve the computation efficiency. We prove that our procedure has the oracle properties. Both simulation studies and a real data example are carried out to examine finite sample performance of the proposed fixed and random effects selection procedure. Supplementary materials including R code used in this article and proofs for the theorems are available online.

Keywords

BIC; LASSO; Mixed-effects models

1. INTRODUCTION

Mixed-effects models are well-suited to the analysis of longitudinal, clustered, and panel data and have been used in nearly all areas of statistical applications such as biology, medicine, finance, and surveys (Demidenko 2004; McCulloch, Searle, and Neuhaus 2008). With the development of data collection technology, oftentimes there are a large number of variables that could be collected easily in the study, but usually only a subset of them are significantly related to the response. At the initial stage of statistical modeling, more variables are usually included in the model to reduce the approximation bias. Besides the

SUPPLEMENTARY MATERIALS

R code: The zip file (code.zip) contains R code that performs the method proposed in the manuscript and a readme file.

Proofs: Proofs for the theorems in Section 4.

selection of the variables, determining the structure of the random effects is another challenge in linear mixed-effects modeling. If the random effects are overspecified, the covariance matrix will be (nearly) singular. This imposes difficulties for the fitting algorithm; as a result, the results may not be reliable. On the other hand, Lange and Laird (1989) showed theoretically that an underspecified random effects covariance structure may lead to biased variance estimates of the fixed effects. In practice, inference and interpretation of parameter estimation are closely related to which variables exhibit subject-specific effects in the model. For example, in a study on students' attitude toward sciences, survey results are collected from 20 classes in 12 private schools and 46 classes in 29 public schools in and around Canberra, Australia. Maindonald and Braun (2010) showed that the between-class random effect is significant while the between-school random effect is not. However, if one mistakenly ignored the between-class random effect, the between-school random effect is found to be substantial, but this may be misleading because the main difference is between classes, not between schools.

While there is an extensive literature on the estimation and prediction problems in linear and generalized linear mixed-effects models, only few references discuss about model selection problems regarding the fixed and random effects in mixed-effects models. Jiang et al. (2008) introduced a method, known as the fence, for mixed-effects model selection that builds a statistical fence to separate a subset of what are called correct models (of which the optimal model is a member). Under some conditions, the authors showed that the fence method is consistent. Pu and Niu (2006) proposed a method for the variable selection in linear mixed-effects models. Their method divides the model selection process into two steps, the first step selecting the fixed effects by fixing the random effects and the second step selecting the random effects after choosing the fixed effects. However, when the number of variables is large, these methods become infeasible for model selection due to the expensive computational cost and lack of stability for the discrete property as analyzed, for instance, by Breiman (1996).

On the other hand, since the proposal of the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and the smoothly clipped absolute deviation (SCAD) method (Fan and Li 2001), there have been extensive developments on model selection based on penalized log-likelihood and the computational issues of solving these problems in linear models. See, for example, Fan and Li (2001), Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), Zou, Hastie, and Tibshirani (2007), Wu and Lange (2008), Zou and Zhang (2009), and Breheny and Huang (2011). However, there are only some papers discussing penalized variable selection approach for models with random effects. Ni, Zhang, and Zhang (2010) considered variable selection for semiparametric mixed-models in longitudinal studies, while Schelldorfer, Bühlmann, and van de Geer (2011) considered selection of the fixed effects variables for linear mixed-effects models. However, these authors did not consider selection of the random effects. Using the fact that the LASSO penalty may be thought of as a Laplacian prior for β (Park and Casella 2008), Foster, Verbyla, and Pitchford (2007) incorporated the LASSO into the mixed-effects models. However, the computation aspects of the Bayesian approach still need development. There are also some papers that have studied the modeling of the covariance matrix either

nonparametrically (Huang et al. 2006) or in Bayesian way (Daniels and Zhao 2003). Bondell, Krishna, and Ghosh (2010) proposed simultaneous selection of the fixed and random effects in linear mixed-effects models using an adaptive LASSO-type method. The authors applied an EM algorithm to solve the penalized estimation problem. More specifically, in the maximization step, a quadratic programming technique is used. Almost at the same time, Ibrahim et al. (2011) proposed a penalized likelihood method to simultaneously select the fixed and random effects by penalizing the fixed and random effects parameters with different tuning parameters λ 's. The latter authors maximized the penalized log-likelihood by an EM algorithm that involves Markov chain Monte Carlo (MCMC) to approximate the intractable integrals. In particular, in the maximization step, the authors used a modification of the local linear approximation algorithm (LLA) (Zou and Li 2008). The use of several different λ 's as the tuning parameters for the fixed and random effects would add much more computational burden to search among a grid of λ 's. If the same λ is used to penalize both the fixed and random effects, the difference in the scale of these two kinds of parameters would compromise the ability of the method in identifying whether a parameter is truly zero.

Several methods have been used to estimate the variance component parameters in linear mixed-effects models. The two most popular methods are maximum likelihood (ML) and restricted maximum likelihood (REML). Between the two methods, REML is preferred for several reasons. First, the REML estimators correct the bias in the ML estimators by accounting for the loss of degrees of freedoms attributable to the estimation of the fixed effects (Jiang 2007). Second, REML estimators seem more robust to outliers in data than ML estimators (Verbyla 1993). Third, fixed effects are not involved in REML, so that the estimators of unknown parameters in covariance matrix are invariant to the value of fixed effects (McCulloch, Searle, and Neuhaus 2008). The last point is especially important for computational consideration since the dimension involved in REML is generally much lower. However, both Bondell, Krishna, and Ghosh (2010) and Ibrahim et al. (2011) used the ML instead of REML in their approaches to model selection.

To make full use of REML's strength and increase the computational efficiency, we propose a two-stage model selection procedure for linear mixed-effects models. We first use the penalized restricted log-likelihood to select the random effects, in which a Newton-type algorithm is adopted. We then use the penalized log-likelihood to select the fixed effects, in which a pathwise coordinate optimization is developed to improve the computational efficiency. The dimension in each of the two steps is lower than the combined dimension of both fixed and random effects. This makes the computation more efficient and stable. We prove that our procedure has the oracle properties (Fan and Li 2001). Simulation results show that our procedure selects the random effects more accurately than the fixed effects. Our procedure is also more computationally efficient than that by Ibrahim et al. (2011) since we only need to search two sets of λ 's while Ibrahim et al. (2011) requires searching in a two-dimensional grid. The simulation results also show that our procedure is more efficient than Bondell, Krishna, and Ghosh (2010), which uses one common λ for the fixed and random effects.

The rest of the article is organized as follows. In Section 2, we propose the penalized restricted log-likelihood for the random effects selection. In Section 3, we consider the penalized log-likelihood function for the fixed effects selection and apply the pathwise coordinate optimization to the penalized log-likelihood. Theoretical results are given in Section 4. Simulation studies and a real data example are presented in Section 5. Some remarks are given in Section 6.

2. PENALIZED RESTRICTED LOG-LIKELIHOOD FOR RANDOM EFFECTS

2.1 Restricted Log-Likelihood

Consider the linear mixed-effects model

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i=1, \dots, n, \quad (1)$$

where $\mathbf{y}_i$ is an $n_i \times 1$ vector, $\mathbf{X}_i$ is an $n_i \times p$ matrix, $\mathbf{Z}_i$ is an $n_i \times q$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector for the fixed effects, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of random errors. We further assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$. Denote $\boldsymbol{\theta} = (\boldsymbol{\phi}^T, \sigma^2)^T$, where $\boldsymbol{\phi}$ is the vector of $k = \frac{q(q+1)}{2}$ unknown covariance parameters contained in $\mathbf{D}$.

The log-likelihood function, excluding constant terms, is given by

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{1}{2} \sum_{i=1}^n \log |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}),$$

where $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{n_i}$. And the restricted log-likelihood (REML) function (Lindstrom and Bates 1988), excluding constant terms, is given by

$$\ell^R(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}) - \frac{1}{2} \log \left| \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right|, \quad (2)$$

where $\hat{\boldsymbol{\beta}}$ is obtained by the generalized least squares

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i \right), \quad (3)$$

if $\mathbf{V}_i$ is assumed to be known for each i . From Jennrich and Schluchter (1986), the elements of the score vector $\mathbf{s} = \partial \ell / \partial \boldsymbol{\theta}$ and Hessian matrix $\mathbf{H} = \partial^2 \ell / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ are

$$[\mathbf{s}_\theta]_r = \frac{1}{2} \sum_{i=1}^n \text{tr} [\mathbf{V}_i^{-1} (\mathbf{e}_i \mathbf{e}_i^T - \mathbf{V}_i) \mathbf{V}_i^{-1} \dot{\mathbf{V}}_{ir}], \quad r=1, \dots, k+1;$$

$$[\mathbf{H}_{\theta\theta}]_{rs} = -\frac{1}{2} \sum_{i=1}^n \text{tr} [\mathbf{V}_i^{-1} \dot{\mathbf{V}}_{ir} \mathbf{V}_i^{-1} (2\mathbf{e}_i \mathbf{e}_i^T - \mathbf{V}_i) \mathbf{V}_i^{-1} \dot{\mathbf{V}}_{is}] + \frac{1}{2} \sum_{i=1}^n \text{tr} [\mathbf{V}_i^{-1} (\mathbf{e}_i \mathbf{e}_i^T - \mathbf{V}_i) \mathbf{V}_i^{-1} \ddot{\mathbf{V}}_{i,rs}], \quad r, s=1, \dots, k+1,$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}$, $\dot{\mathbf{V}}_{ir} = \partial \mathbf{V}_i / \partial \theta_r$, and $\ddot{\mathbf{V}}_{i,rs} = \partial^2 \mathbf{V}_i / \partial \theta_r \partial \theta_s$. The elements of $\dot{\mathbf{V}}_{ir}$ and $\ddot{\mathbf{V}}_{i,rs}$, respectively, are the first and second derivatives of the elements of $\mathbf{V}_i$ with respect to $\theta_1, \dots, \theta_{k+1}$.

The first and second derivatives of $-\log|\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i|/2$ can be found in Lindstrom and Bates (1988) and we list here for completeness. We have

$$\frac{\partial \log|\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i|}{\partial \theta_j} = -\sum_{i=1}^n \text{tr} [\mathbf{H}_0^{-1} \mathbf{X}_i^T \mathbf{A}_{ij} \mathbf{X}_i]$$

and

$$\frac{\partial^2 \log|\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i|}{\partial \theta_l \partial \theta_j} = -\text{tr} \left[\mathbf{H}_0^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{A}_{il} \mathbf{X}_i \times \mathbf{H}_0^{-1} \sum_{i=1}^n \mathbf{X}_i^T \mathbf{A}_{ij} \mathbf{X}_i \right] - \text{tr} \left[\mathbf{H}_0^{-1} \sum_{i=1}^n \left(\mathbf{X}_i^T \frac{\partial \mathbf{A}_{ij}}{\partial \theta_l} \mathbf{X}_i \right) \right],$$

where

$$\mathbf{A}_{ij} = \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_j} \mathbf{V}_i^{-1},$$

$$\frac{\partial \mathbf{A}_{ij}}{\partial \theta_l} = -\mathbf{V}_i^{-1} \left(\frac{\partial \mathbf{V}_i}{\partial \theta_l} \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_j} - \frac{\partial^2 \mathbf{V}_i}{\partial \theta_l \partial \theta_j} + \frac{\partial \mathbf{V}_i}{\partial \theta_j} \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_l} \right) \mathbf{V}_i^{-1}.$$

and

$$\mathbf{H}_0 = \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i.$$

2.2 Penalized Restricted Log-Likelihood

Zou (2006) showed that the adaptive LASSO has the oracle properties if the weights are cleverly chosen and it is essentially an ℓ_1 -penalization method for a convex optimization problem. The maximization algorithms for the REML are usually based on two basic approaches: the Expectation–Maximization (EM) and Newton–Raphson Method (NR). We factorize $\boldsymbol{\theta}$ as $(\mathbf{d}, \boldsymbol{\gamma}, \sigma^2)$, where $\mathbf{d}$ is the vector of the diagonal elements of the covariance matrix and $\boldsymbol{\gamma}$ is the vector of the unknown parameters of the covariance matrix besides the diagonal part.

The penalized restricted log-likelihood function is given by

$$Q^R(\boldsymbol{\theta}) = \ell^R(\boldsymbol{\theta}) - \sum_{j=1}^q \lambda_j w_j |d_j|, \quad (4)$$

where d_j is the j th element of the diagonal of $\mathbf{D}$ and w_j is the known weight. It is reasonable to define the weight vector $\hat{\mathbf{w}} = 1/|\hat{\mathbf{d}}|$, where $\hat{\mathbf{d}}$ is a $\sqrt{n}$ -consistent estimator of $\mathbf{d}$. In the theoretical analysis, each λ_j may be different. However, for simplicity, we assume that all the λ_j 's are the same and are denoted as λ_{1n} .

To maximize $Q^R(\boldsymbol{\theta})$, we apply the Newton–Raphson algorithm whose rate of convergence is quadratic with a good initial value $\boldsymbol{\theta}_0$ (Jennrich and Schluchter 1986). The Newton–Raphson algorithm is an iterative procedure which has the form

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s - \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \mathbf{s}_{\boldsymbol{\theta}}, \quad s=0, 1, \dots,$$

where $\boldsymbol{\theta}_s$ is the estimate of Equation (4) at the s th iteration, $\boldsymbol{\theta}_{s+1}$ is the updated estimator, $\mathbf{s}_{\boldsymbol{\theta}}$ is a $(k+1) \times 1$ vector of the first derivative of the objective function, and $\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is a $(k+1) \times (k+1)$ matrix of the second derivatives of the objective function.

In practice, it is difficult to implement the Newton–Raphson algorithm directly, since the objective function (4) is nondifferentiable at the origin. Motivated by the idea of Fan and Li (2001) and Li and Liang (2008), we locally approximate the penalty function by a quadratic function as follows. Given an initial value $\mathbf{d}^0$ that is close to the maximizer of (4), if $d_j^{(0)}$ is very close to 0, the entire j th column and j th row of $\mathbf{D}$ are set to 0. Otherwise, the ℓ_1 -penalty function can be locally approximated by a quadratic function as

$$|d_j| \approx \frac{1}{2} |d_j^{(0)}| + \frac{1}{2} \frac{d_j^2}{|d_j^{(0)}|}.$$

At step s of the iteration, the Newton–Raphson algorithm can be used to solve

$$\boldsymbol{\theta}^{s+1} = \arg \max \left\{ \ell^R(\boldsymbol{\theta}) - \lambda_{1n} \sum_{j=1}^q w_j \frac{d_j^2}{2|d_j^{(s)}|} \right\}.$$

Stop the iteration if the sequence of $\boldsymbol{\theta}$ converges.

3. PATHWISE COORDINATE OPTIMIZATION FOR FIXED EFFECTS

The penalized log-likelihood function when the covariance matrix of random effects is known, excluding constant terms, is given by

$$Q^f(\beta) = -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \beta)^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta) - \lambda_{2n} \sum_{j=1}^p w_j |\beta_j|, \quad (5)$$

where $\mathbf{w} = 1/\bar{\beta}$ is a $p \times 1$ vector and $\bar{\beta}$ is the estimator of the fixed effects using (3) after the covariance matrix is estimated.

Denote $\ell^f(\beta) = -\sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \beta)^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta) / 2$. The first derivative of $\ell^f(\beta)$ is given by

$$\begin{aligned} \frac{\partial \ell^f(\beta)}{\partial \beta} &= \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta) \\ &= \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i - \sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \beta \end{aligned}$$

and its j th component, $\partial \ell^f(\beta) / \partial \beta_j$, is denoted by $T_j(\beta, \mathbf{X}, \mathbf{y})$, where $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T$ and $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_n^T]^T$. We have

$$\frac{\partial Q^f(\beta)}{\partial \beta_j} = T_j(\beta, \mathbf{X}, \mathbf{y}) - \lambda_{2n} \text{sgn}(\beta_j) / |\bar{\beta}_j|.$$

It is easy to see that $T_j(\beta, \mathbf{X}, \mathbf{y})$ is a linear function of β_j given the β_i 's other than β_j in β . Motivated by the idea of the shooting procedure by Fu (1998) or pathwise coordinate optimization algorithm by Friedman et al. (2007), we maximize (5) as follows. Denote

$\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i$ and $\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i$ as $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively. For the j th element of β , let $T_j(0, \beta^{(-j)}, \mathbf{X}, \mathbf{y})$ be $\partial \ell^f(\beta) / \partial \beta_j$ evaluated at $\beta_j = 0$ and $\beta^{(-j)} = \hat{\beta}^{(-j)}$, where $\hat{\beta}^{(-j)}$ is a $(p-1) \times 1$ vector of estimator of β other than β_j . It is easy to verify that

$$T_j(0, \hat{\beta}^{(-j)}, \mathbf{X}, \mathbf{y}) = \mathbf{M}_1[j] - \mathbf{M}_2[j, -j] \hat{\beta}^{(-j)}.$$

So, the algorithm could be described as follows:

1. Start with $\hat{\beta}_0 = \hat{\beta}_{GLS}$,
2. At step m , for each $j = 1, \dots, p$, let $T_0 = T_j(0, \hat{\beta}^{(-j)}, \mathbf{X}, \mathbf{y})$ and set

$$\hat{\beta}_j = \begin{cases} \frac{T_0 - \lambda / |\bar{\beta}_j|}{\mathbf{M}_2[j, j]}, & \text{if } T_0 > 0 \quad \text{and} \quad \lambda / |\bar{\beta}_j| < |T_0|, \\ \frac{T_0 + \lambda / |\bar{\beta}_j|}{\mathbf{M}_2[j, j]}, & \text{if } T_0 < 0 \quad \text{and} \quad \lambda / |\bar{\beta}_j| < |T_0|, \\ 0, & \text{if } \lambda / |\bar{\beta}_j| \geq |T_0|, \end{cases}$$

3. Repeat Step 2 until the sequence of $\hat{\beta}$ converges.

4. THEORETICAL RESULTS

Denote the true value of θ as $\theta_0 = (\theta_{10}^T, \theta_{20}^T, \sigma_0^2)^T$, where $\theta_{10} = (\mathbf{d}_{10}^T, \gamma_{10}^T)^T$, $\mathbf{d}_{10}$ is a $q_1 \times 1$ vector whose components are nonzero, and $\theta_{20} = (\mathbf{d}_{20}^T, \gamma_{20}^T)^T$, all of the elements in θ_{20} are zero. Let the true value of β be $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$, where β_{10} is a $p_1 \times 1$ vector whose components are nonzero and β_{20} is a $(p - p_1) \times 1$ vector whose components are zero. The objective functions (4) and (5) could be written as follows:

$$Q^R(\theta) = \ell^R(\theta) - \lambda_{1n} \sum_{j=1}^q \frac{|d_j|}{|\bar{d}_j|} \quad (6)$$

and

$$Q^f(\beta) = \ell^f(\beta) - \lambda_{2n} \sum_{j=1}^p \frac{|\beta_j|}{|\bar{\beta}_j|}, \quad (7)$$

where

$$\ell^R(\theta) = -\frac{1}{2} \log |\mathbf{X}^T \mathbf{V} \mathbf{X}| - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X} \hat{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta})$$

and

$$\ell^f(\beta) = -\frac{1}{2} (\mathbf{y} - \mathbf{X} \beta)^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X} \beta),$$

where $\mathbf{Z}$ is the block diagonal matrix of all $\mathbf{Z}_i$, $\mathcal{D}$ is the block diagonal matrix of $\mathbf{D}$, Λ is the block diagonal matrix of $\sigma^2 \mathbf{I}_{n_i}$'s, and $\mathbf{V} = \mathbf{Z} \mathcal{D} \mathbf{Z}^T + \Lambda$.

Theorem 1—Under Assumptions 1–4 given in the Appendix, we have

- 1a** If $\lambda_{1n}/\sqrt{n} \rightarrow 0$, then there exists a local maximizer $\hat{\theta}_{\lambda_{1n}}$ of $Q^R(\theta)$ such that $\hat{\theta}_{\lambda_{1n}}$ is $\sqrt{n}$ -consistent for θ_0 ;
- 1b** If $\lambda_{1n} \rightarrow \infty$, then there exists some constant $M > 0$, for any given $\hat{\theta}$ satisfying $\|\hat{\theta}_1 - \theta_{10}\| \leq M n^{-1/2}$, $P(\hat{\theta}_2 = 0) \rightarrow 1$;
- 1c** As $\lambda_{1n} \rightarrow \infty$ and $\lambda_{1n}/\sqrt{n} \rightarrow 0$, we have

$$\sqrt{n} \mathbf{I}(\theta_{10}) \left\{ \hat{\theta}_1 - \theta_{10} + \frac{\lambda_{1n}}{n} \mathbf{I}(\theta_{10})^{-1} \mathbf{b}(\theta_{1n}) \right\} \rightarrow_d N\{0, \mathbf{I}(\theta_{10})\},$$

where $\mathbf{b}(\theta_{10}) = (\text{sgn}(\theta_{10})/\theta_{11}^-, \dots, \text{sgn}(\theta_{q10})/\theta_{q1}^-)^T$ and $\mathbf{I}(\theta_{10}) = \mathbf{I}(\theta_{10}, \theta_{20} = 0)$.

Theorem 1 presents the asymptotic property of the adaptive LASSO as a simultaneous random effects selection and parameter estimation tool. From Theorem 1, we see that $\hat{\theta}_{\lambda_{1n}}$ is a $\sqrt{n}$ -consistent estimator of θ and the true model could be identified with probability tending to 1 and $\hat{\theta}_1$ is asymptotically normal.

Theorem 2—Under Assumptions 1–4 given in the Appendix. We have

- 2a** If $\lambda_{2n}/\sqrt{n} \rightarrow 0$, then there exists a local maximizer $\hat{\beta}_{\lambda_{2n}}$ of $Q^f(\beta)$ such that $\hat{\beta}_{\lambda_{2n}} - \beta_0 = O_p(n^{-1/2})$;
- 2b** If $\lambda_{2n} \rightarrow \infty$, then for any given $\hat{\beta}$ satisfying $\|\hat{\beta}_1 - \beta_{10}\| \leq Mn^{-1/2}$ and some constant $M > 0$, $P(\hat{\beta}_2 = 0) \rightarrow 1$;
- 2c** As $\lambda_{2n} \rightarrow \infty$ and $\lambda_{2n}/\sqrt{n} \rightarrow 0$, we have

$$\sqrt{n}\mathbf{I}(\beta_{10}) \left\{ \hat{\beta}_1 - \beta_{10} + \frac{\lambda_{2n}}{n} \mathbf{I}(\beta_{10})^{-1} \mathbf{b}(\beta_{10}) \right\} \rightarrow_d N\{0, \mathbf{I}(\beta_{10})\},$$

where $\mathbf{b}(\beta_{10}) = (\text{sgn}(\beta_{10})/\beta_1^-, \dots, \text{sgn}(\beta_{p10})/\beta_{p10}^-)^T$ and $\mathbf{I}(\beta_{10}) = \mathbf{I}(\beta_{10}, \beta_{20} = 0)$.

Theorem 2 says that given the estimation of θ in the first step, $\hat{\beta}$ is $\sqrt{n}$ -consistent, asymptotically normal, and possesses the sparsity property when $\lambda_{2n}/\sqrt{n} \rightarrow 0$ and $\lambda_{2n} \rightarrow \infty$, that is, estimator of β enjoys the oracle properties as it performs as well as knowing $\beta_2 = 0$ before the estimation.

5. NUMERICAL EXAMPLES

5.1 Selection of Tuning Parameters

It is important to select the proper tuning parameters in the penalized likelihood method. For linear models, there have been several commonly used tuning parameter selection methods, such as the Akaike information criterion (AIC) (Akaike 1974), the Bayesian information criterion (BIC) (Schwarz 1978), cross-validation (CV) (Shao 1993), the generalized cross-validation (GCV) (Craven and Wahba 1979), and bootstrap (Shao 1996; Pang, Lin, and Jiang 2012). Zhang, Li, and Tsai (2010) showed that the BIC-type selector can identify the true model consistently while the AIC-type cannot. Denote $N = \sum_{i=1}^n n_i$. We apply the BIC-type criteria given by

$$\text{BIC}_{\lambda_{1n}} = -2\ell^R(\hat{\theta}) + \log(N)df_{\lambda_{1n}}$$

and

$$\text{BIC}_{\lambda_{2n}} = -2\ell^f(\hat{\beta}) + \log(N)df_{\lambda_{2n}}$$

for the random and fixed effects parts, respectively. Note that $\ell^R(\hat{\boldsymbol{\theta}})$ is the restricted log-likelihood function $\ell^R(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$ and $\ell^f(\hat{\boldsymbol{\beta}})$ is $\ell^f(\boldsymbol{\beta})$ evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. $df_{\lambda_{1n}}$ and $df_{\lambda_{2n}}$ are taken as the dimensions of the nonzero parts of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, respectively.

5.2 Simulation Studies

In this section, we compare the finite sample performance of the proposed fixed and random effects selection method with the existing approach of Bondell, Krishna, and Ghosh (2010). In each simulation example, the data are generated from the linear mixed-effects model

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n.$$

Example 1—In this example, we have the true $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0, 0, 0)^T$; $p = 9$, $q = 4$; $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ with $\sigma^2 = 1$; each element of $\mathbf{X}_i$ and $\mathbf{Z}_i$ except the first column of $\mathbf{Z}_i$ are generated from a uniform $(-2, 2)$ distribution, the first column of $\mathbf{Z}_i$ is $\mathbf{1}$'s for the subject-specific intercept and $\mathbf{b}_i = (\mathbf{b}_{i0}, \mathbf{0})^T$ with $\mathbf{b}_{i0} \sim N(\mathbf{0}, \mathbf{D})$, where

$$\mathbf{D} = \begin{bmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{bmatrix}.$$

In our simulation, we consider $n = 30$, $n_i = 5$ and $n = 60$, $n_i = 10$, respectively. For each case, 100 datasets were generated.

Example 2—In this example, we have the true $\boldsymbol{\beta} = (1, 0, 1, 0, 0, 0, 0, 0)^T$ and $\mathbf{b} = (\mathbf{b}_0, 0, 0, 0, 0, 0, 0, 0)^T$. $\boldsymbol{\varepsilon}_i$, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are generated using the same method as in Example 1 except that $p = 9$ and $q = 10$ for this example. We set $n = 60$ subjects and $n_i = 5$ per subject and repeat the simulation 100 times.

Example 3—In this example, the true $\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)^T$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$, the correlation between x_l and x_j in the 12×8 matrix $\mathbf{X}_i$ is $\rho^{|l-j|}$ with $\rho = 0.5$; $\mathbf{Z}_i = \mathbf{X}_i$ and $\mathbf{b} = (\mathbf{b}_0, 0, 0, 0, 0, 0, 0, 0)^T$ with $\mathbf{b}_0 \sim N(\mathbf{0}, \mathbf{D})$ where

$$\mathbf{D} = \begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{bmatrix}.$$

In this example, we consider two factors: $n = 50, 100, 200$ and $\sigma^2 = 1, 3$.

We adopt two model accuracy indexes to compare with different methods. First, the Kullback–Leibler discrepancy (KLD) for linear mixed-effects models is

$$\text{KLD} = E \left\{ \log f(\mathbf{Y}, \mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\beta}) - \log f(\mathbf{Y}, \mathbf{X}, \mathbf{Z} | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) \right\}, \quad (8)$$

where

$$\begin{aligned} E \{ \log f(\mathbf{Y}, \mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\beta}) \} &= E \left\{ -\frac{1}{2} \sum_{i=1}^n \log |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}) \right\} \\ &= E \left\{ -\frac{1}{2} \sum_{i=1}^n \log |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^n \text{tr} \left\{ (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{V}_i^{-1} \right\} \right\} \\ &= -\frac{1}{2} \sum_{i=1}^n \log |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^n n_i, \end{aligned}$$

and

$$E \{ \log f(\mathbf{Y}, \mathbf{X}, \mathbf{Z} | \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) \} = E \left\{ -\frac{1}{2} \sum_{i=1}^n \log |\hat{\mathbf{V}}_i| - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})^T \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \right\},$$

which is approximated by its empirical value given the estimated $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$

Let KLD_{fit} be the KLD of the selected model and KLD_{full} be the KLD of the full model. We consider the ratios $\text{KLD}_{\text{fit}}/\text{KLD}_{\text{full}}$ and its standard error as our first index and let MKLD be the median of the ratios. We expect the Kullback–Leibler discrepancy for the selected model by our procedure to be smaller than the one of full model. So, the smaller the MKLD the better.

Similarly, we consider the ratios of model error of the selected model based on our penalized estimation (ME_{pen}) to that of the unpenalized estimation (ME_{upen}): $\text{ME}_{\text{pen}}/\text{ME}_{\text{upen}}$. The second index is the median of the ratios (MRME). This is the mean square error $\text{ME}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for the fixed effects and the quadratic loss error $\text{ME}(\hat{\boldsymbol{\theta}}) = [\text{tr}(\hat{\mathbf{D}} - \mathbf{D})]^{1/2}$ for the random effects. Let %CF, %CR, and %C denote the percentages of times that the fixed effects, random effects, and their combinations are correctly selected, respectively. In all the tables, except for the MKLD where we report the standard error in the brackets, the numbers inside the brackets are for the random effects and those outside the brackets are for the fixed effects.

From Table 1, we can see that our two-stage procedure could identify nonzero fixed and random effects with a high probability, especially for random effects. For example, in Example 1, even when the sample size is small, that is, $n = 30$ and $n_i = 5$, the proportion of correctly identified random effects is 96% which is quite close to 1. Our method also significantly reduces the model error in the sense of both MKLD and MRME. When the sample size is bigger, that is, $n = 60$ and $n_i = 10$, the MKLD is 0.462 which is very close to that of the oracle (0.424). In this case, our procedure can correctly select the model with frequency 95%. Thus, our simulations show that our method works better when the sample size n increases which confirms the asymptotic results, that is, the performance of the method could be expected to be as good as if we know the correct model first with big

enough sample size n . In Example 2, when the dimension involved in the random effects is big ($q = 10$), our method can still identify the random effects with high frequency (98%).

From Table 2, we can see that in Example 3, our method only becomes a little worse when the variance of the error increases and the performance increases when the sample size n increases.

5.3 Real Data Examples

We apply the proposed method to a subsample from the 1982 High School and Beyond (HSB) Survey (Singer 1998; Raudenbush and Bryk 2002). The dataset includes information on 7185 students nested within 160 schools. There are 14 students in the smallest school and 67 students in the largest school in the survey. The response, y_{ij} , is a standardized measure of math achievement for $i = 1, \dots, 160, j = 1, \dots, n_i$. The predictors selected in our analysis are described as follows. There are three predictors at student-level: $Cses_{ij}$ = centered student socioeconomic status (SES), which is a composite of parental education, parental occupation, and parental income; $Minority_{ij}$ = a factor with levels 0 and 1 indicating if the student is a member of a minority racial group; $Gender_{ij}$ = gender of the student which has two levels male and female coded as 1 and 0; there are also two predictors at school-level: $Sector_i$ = an indicator variable taking on a value of 0 for public schools and 1 for Catholic schools; $Meanses_i$ = mean SES of the school.

We build a hierarchical model to the data following Raudenbush and Bryk (2002). First, the student-level equation for student j in school i is $y_{ij} = a_{0i} + a_{1i} Cses_{ij} + a_{2i} Minority_{ij} + a_{3i} Gender_{ij} + \varepsilon_{ij}$. Second, the intercepts and slopes may depend upon two school-level predictors: $Sector$ and $Meanses$. Finally, the model is rewritten in the convention of linear mixed-effects models as follows:

$$\begin{aligned} y_{ij} = & \beta_1 + \beta_2 Cses_{ij} + \beta_3 Minority_{ij} + \beta_4 Gender_{ij} + \beta_5 Meanses_i + \beta_6 Sector_i \\ & + \beta_7 Meanses_i Cses_{ij} + \beta_8 Sector_i Cses_{ij} + \beta_9 Meanses_i Minority_{ij} \\ & + \beta_{10} Sector_i Minority_{ij} + \beta_{11} Meanses_i Gender_{ij} + \beta_{12} Sector_i Gender_{ij} \\ & + b_{i1} + b_{i2} Cses_{ij} + b_{i3} Minority_{ij} + b_{i4} Gender_{ij} + \varepsilon_{ij} \end{aligned}$$

where β 's are fixed effects and b 's are random effects.

The results of the analysis are reported in Table 3. Furthermore, we compare the full model and the model selected using hypothesis test. The p -value is 0.93. This suggests that the selected model fits the data quite well while excluding unimportant variables. It is interesting to find that centered student socioeconomic status is a fixed effect, but not a random effect. This can be seen from the relative small value (0.04) as compared with other variance components under REML and our method correctly penalized it to zero. The average slope relating math achievement to students' SES is larger in public schools than in Catholic schools ($SectorCses$ equals -1.30). Minority students performed better on average in Catholic schools than in public schools. In Catholic schools, minority students scored 1.93 points behind their classmates (the difference between $Minority$ and $SectorMinority$), while the value is 3.81 in public schools. Boys had better performance in the school ($Gender$ equals 1.27) no matter the schools' sector ($SectorGender$ equals -0.02) and average SES

score (MeasuresGender equals 0.002). Whether a student is a minority or not also does not affect the average SES score (MeasuresMinority equals -0.77). The t -values for SectorGender, MeasuresGender, and MeasuresMinority are quite small and these three variables are penalized to zero under our method. The students' SES is related to the school average SES score (MeasuresCses equals 0.99 with a big t -value) and this variable is correctly kept under our penalized method.

6. REMARKS

We propose a two-stage model selection procedure for linear mixed-effects models and show through simulation studies that our method outperforms the existing approach of Bondell, Krishna, and Ghosh (2010) across the cases considered. There are several reasons for the outperformance. First, we believe that it is more efficient to use the two-stage selection procedure, first for the random effects and then for the fixed effects, than simultaneous selecting the fixed effects and variance components. This is because the natures of these parameters are quite different; it would seem unnatural to select them altogether. Second, we have adopted the REML for selecting the random effects. REML is known to be optimal in estimating the variance components (e.g., Jiang 2007), so it is not surprising that it outperforms the ML-based method in this regard. Also, the number of parameters involved in the REML part is much smaller (because it excludes all the fixed effects), which makes the procedure more efficient and stable. The second stage of our procedure for selecting the fixed effects only involves the regression coefficients. The adaptive LASSO is well known to be useful.

Acknowledgments

The authors thank the editor, the associate editor, and two referees for their valuable comments and helpful suggestions which have led to substantial improvement in the presentation of the manuscript. Zhen Pang's research is partly supported by a grant from the Ministry of Education, Singapore.

References

- Akaike H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.
- Bondell HD, Krishna A, Ghosh SK. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*. 2010; 66:1069–1077. [PubMed: 20163404]
- Breheny P, Huang J. Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection. *The Annals of Applied Statistics*. 2011; 5:232–253. [PubMed: 22081779]
- Breiman L. Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*. 1996; 24:2350–2383.
- Craven P, Wahba G. Smoothing Noisy Data With Spline Functions. *Numerische Mathematik*. 1979; 31:377–403.
- Daniels M, Zhao Y. Modelling the Random Effects Covariance Matrix in Longitudinal Data. *Statistics in Medicine*. 2003; 22:1631–1647. [PubMed: 12720301]
- Demidenko, E. *Mixed Models Theory and Applications*. New York: Wiley; 2004.
- Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Oracle Properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.

- Foster S, Verbyla A, Pitchford W. Incorporating Lasso Effects Into a Mixed Model for Quantitative Trait Loci Detection. *Journal of Agricultural, Biological, and Environmental Statistics*. 2007; 12:300–314.
- Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*. 2007; 1:302–332.
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010; 33:1–22. [PubMed: 20808728]
- Fu WJ. Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*. 1998; 7:397–416.
- Huang J, Liu N, Pourahmadi M, Liu L. Covariance Matrix Selection and Estimation via Penalized Normal Likelihood. *Biometrika*. 2006; 93:85–98.
- Ibrahim JG, Zhu H, Garcia RI, Guo R. Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*. 2011; 67:495–503. [PubMed: 20662831]
- Jennrich RI, Schluchter MD. Unbalanced Repeated-Measures Models With Structured Covariance Matrices. *Biometrics*. 1986; 42:805–820. [PubMed: 3814725]
- Jiang, J. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer; 2007.
- Jiang J, Rao JS, Gu Z, Hguyen T. Fence Methods for Mixed Model Selection. *The Annals of Statistics*. 2008; 36:1669–1692.
- Lange N, Laird N. The Effect of Covariance Structures on Variance Estimation in Balance-Curve Models With Random Parameters. *Journal of the American Statistical Association*. 1989; 84:241–247.
- Li R, Liang H. Variable Selection in Semiparametric Regression Modeling. *The Annals of Statistics*. 2008; 36:261–286.
- Lindstrom MJ, Bates DM. Newton–Raphson and em Algorithms for Linear Mixed-Effects Models for Repeated-Measure Data. *Journal of the American Statistical Association*. 1988; 83:1014–1022.
- Maindonald, J.; Braun, J. *Data Analysis and Graphics Using R: an Example-Based Approach*. New York: Cambridge University Press; 2010.
- McCulloch, CE.; Searle, SR.; Neuhaus, JM. *Generalized, Linear, and Mixed Models*. 2. New York: Wiley; 2008.
- Ni X, Zhang D, Zhang H. Variable Selection for Semiparametric Mixed Models in Longitudinal Models. *Biometrics*. 2010; 66:79–88. [PubMed: 19397585]
- Pang Z, Lin B, Jiang J. Regularization Parameter Selections via Bootstrapping. 2012 Submitted.
- Park T, Casella G. The Bayesian Lasso. *Journal of the American Statistical Association*. 2008; 103:681–686.
- Pu W, Niu X. Selecting Mixed-Effects Models Based on Generalized Information Criterion. *Journal of Multivariate Analysis*. 2006; 97:733–758.
- Raudenbush, SW.; Bryk, AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage; 2002.
- Schelldorfer J, Bühlmann P, van de Geer S. Estimation for High-Dimensional Linear Mixed-Effects Models Using l1-Penalization. *Scandinavian Journal of Statistics*. 2011; 38:197–214.
- Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978; 6:461–464.
- Shao J. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*. 1993; 88:486–494.
- Shao J. Bootstrap Model Selection. *Journal of the American Statistical Association*. 1996; 91:655–665.
- Singer JD. Using SAS Proc Mixed to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*. 1998; 23:323–355.
- Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:268–288.
- Verbyla A. Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics. *Journal of the Royal Statistical Society, Series B*. 1993; 55:493–508.
- Wu T, Lange K. Coordinate Descent Algorithms for Lasso Penalized Regression. *The Annals of Applied Statistics*. 2008; 2:224–244.

- Zhang Y, Li R, Tsai CL. Regularization Parameter Selections via Generalized Information Criterion. *Journal of the American Statistical Association*. 2010; 105:312–323. [PubMed: 20676354]
- Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T, Tibshirani R. On the ‘Degrees of Freedom’ of the Lasso. *The Annals of Statistics*. 2007; 35:2173–2192.
- Zou H, Li R. One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics*. 2008; 36:1509–1533.
- Zou H, Zhang HH. On the Adaptive Elastic-Net With a Diverging Number of Parameters. *The Annals of Statistics*. 2009; 37:1733–1751.

Table 1

Simulation results for Examples 1 and 2

Method	MKLD	MRME	%C	%CF	%CR
Example 1 ($p = 9, q = 4$)					
$(n = 30, n_i = 5)$					
ALASSO	0.526 (0.059)	0.189 (0.960)	79	81	96
M-ALASSO			71	73	79
oracle	0.384 (0.027)	0.111 (0.928)	100	100	100
$(n = 60, n_i = 10)$					
ALASSO	0.462 (0.021)	0.189 (0.990)	95	96	99
M-ALASSO			83	83	89
oracle	0.424 (0.020)	0.176 (0.975)	100	100	100
Example 2 ($p = 9, q = 10$)					
$(n = 60, n_i = 5)$					
ALASSO	0.233 (0.022)	0.268 (1.167)	85	86	98
M-ALASSO			61	63	84
oracle	0.102 (0.018)	0.165 (0.613)	100	100	100

NOTE: ALASSO is our method and M-ALASSO is the method by Bondell, Krishna, and Ghosh (2010). The results of M-ALASSO are copied from Table 1 in Bondell, Krishna, and Ghosh (2010).

Table 2

Simulation results for Example 3

Example	MKLD	MRME	%C	%CF	%CR	Zero ₁	Zero ₂
$n = 50, \sigma^2 = 1$	0.407 (0.046)	0.924 (0.697)	90	92	98	0.11 (0.02)	0.00 (0.00)
oracle	0.271 (0.016)	0.854 (0.604)	100	100	100	0.00 (0.00)	0.00 (0.00)
$n = 50, \sigma^2 = 3$	0.554 (0.052)	0.743 (0.656)	85	89	95	0.11 (0.09)	0.00 (0.00)
oracle	0.288 (0.022)	0.710 (0.467)	100	100	100	0.00 (0.00)	0.00 (0.00)
$n = 100, \sigma^2 = 1$	0.323 (0.016)	0.903 (0.666)	95	96	99	0.04 (0.01)	0.00 (0.00)
oracle	0.289 (0.017)	0.870 (0.640)	100	100	100	0.00 (0.00)	0.00 (0.00)
$n = 100, \sigma^2 = 3$	0.384 (0.021)	0.770 (0.535)	96	97	99	0.03 (0.01)	0.00 (0.00)
oracle	0.322 (0.023)	0.759 (0.499)	100	100	100	0.00 (0.00)	0.00 (0.00)
$n = 200, \sigma^2 = 1$	0.289 (0.012)	0.870 (0.658)	96	96	100	0.04 (0.00)	0.00 (0.00)
oracle	0.274 (0.010)	0.830 (0.656)	100	100	100	0.00 (0.00)	0.00 (0.00)
$n = 200, \sigma^2 = 3$	0.289 (0.013)	0.715 (0.515)	96	96	100	0.04 (0.00)	0.00 (0.00)
oracle	0.278 (0.012)	0.663 (0.503)	100	100	100	0.00 (0.00)	0.00 (0.00)

Table 3

Results for math achieve data

Variable	REML for full model		AdLASSO	
	Fixed (<i>t</i> value)	Random	Fixed	Random
Intercept	12.2 (62.9)	1.11	12.3	1.02
Cses	2.48 (16.1)	0.04	2.43	0.00
Minority	-3.81 (-11.2)	1.05	-3.53	0.64
Gender	1.27 (5.38)	0.95	1.21	0.67
Meanses	4.28 (10.6)	(-)	4.03	(-)
Sector	1.18 (3.84)	(-)	1.19	(-)
MeansesCses	0.99 (3.41)	(-)	0.84	(-)
SectorCses	-1.30 (-5.57)	(-)	-1.20	(-)
MeansesMinority	-0.77 (-1.39)	(-)	0.00	(-)
SectorMinority	1.88 (3.99)	(-)	1.46	(-)
MeansesGender	0.002 (0.00)	(-)	0.00	(-)
SectorGender	-0.02 (-0.05)	(-)	0.00	(-)