Ann Inst Stat Math (2017) 69:627–646
DOI 10.1007/s10463-016-0555-z

New variable selection for linear mixed-effects models

Ping Wu[1] · Xinchao Luo[1] · Peirong Xu[2] · Lixing Zhu[3,4]

Received: 5 June 2014 / Revised: 17 December 2015 / Published online: 25 February 2016

Abstract In this paper, we consider how to select both the fixed effects and the random effects in linear mixed models. To make variable selection more efficient for such models in which there are high correlations between covariates associated with fixed and random effects, a novel approach is proposed, which orthogonalizes fixed and random effects such that the two sets of effects can be separately selected with less influence on one another. Also, unlike most of existing methods with parametric assumptions, the new method only needs fourth order moments of involved random variables. The oracle property is proved. the performance of our method is examined by a simulation study.

Keywords Linear mixed-effects models · Fixed and random effects selection · Orthogonality

1 Introduction

Let the observations $\{y_i, x_i, z_i, l_i\}_{i=1}^n$ follow from the linear mixed-effects model

$$y_i = x_i\beta + z_i b_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

✉ Lixing Zhu
lzhu@hkbu.edu.hk

1 School of Statistics, East China Normal University, Shanghai 200241, China

2 Department of Mathematics, Southeast University, Nanjing 210096, China

3 School of Statistics, Beijing Normal University, Beijing 100875, China

4 Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

where $y_i = (y_{i1}, \ldots, y_{il_i})^\tau$ is the response for group i, l_i is the number of observations in the ith group, x_i and z_i are, respectively, the given between-individuals and within-individuals design matrices of dimensions $l_i \times p$ and $l_i \times q$ for the fixed effects vector β and the random effects vector b_i with zero mean and covariance matrix D, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{il_i})^\tau$ is the ith individual error with zero mean and covariance matrix $\sigma^2 I_{l_i}$. Here I_l means an identity matrix of $l \times l$. Assume that $b_1, \ldots, b_n$ are independent and identically distributed ($i.i.d.$), $\varepsilon_{i1}, \ldots, \varepsilon_{il_i}$ are also $i.i.d.$ for every group i, and are independent with $\{b_1, \ldots, b_n\}$. When there are many fixed and random effects, and actually some (maybe most) of them are unimportant for the response, variable selection is a natural and important topic.

To handle this problem, penalization-based approaches are useful. Under the normality assumption, examples include the following. Pu and Niu (2006) extended the generalized information criterion (Rao and Wu 1989) to select linear mixed-effects models. Bondell et al. (2010) suggested a joint log likelihood-based adaptive LASSO (Zou 2006). Ibrahim et al. (2011) developed another maximum penalized likelihood procedure for general mixed-effects models, but which was mainly used for linear mixed-effects models. Fan and Li (2012) applied a class of nonconcave penalized profile likelihood methods for selecting and estimating important fixed effects, and proposed a group variable selection strategy to simultaneously select and estimate important random effects. These likelihood-based approaches rely on the normality assumption that can at most be relaxed to parametric distribution, and cannot be applied to the cases without parametric assumptions. To relax parametric distribution assumption, Jiang and Rao (2003) proposed an alternative two-stage procedure. In their paper, the fixed effects are simply selected through marginal models that ignore the impact from the random effects. Jiang et al. (2008) proposed a 'fence' method to simultaneously select both the fixed and random effects for a general mixed model. As their simulations suggested, their method works well numerically for selecting the fixed effects. However, it is not clear whether it works well for selecting the random effects. Peng and Lu (2012) suggested an iterative estimating method with the SCAD penalty (Fan and Li 2001) to overcome the above issue. See Fan and Li (2012) for the recent developments on variable selection for linear mixed models.

There is another particularly important issue for variable selection in mixed-effects models. We know that the characteristics and roles that the fixed and random effects play are very different. The covariate, say, for an unimportant fixed effect may highly correlate with the covariate(s) for some important random effect(s) and vice versa. Therefore, when we have prior information on which are fixed effects and which are random effects, it should be a good strategy, via a separation selection approach, to avoid the impact from one set of effects when we select effects in the another set. However, how to achieve this goal is a challenge. All existing approaches of simultaneous selection are to transfer the selection for both the effects to a fixed effects selection in a conditional sense, particularly for the random effects, by using a sophisticated penalty such as the LASSO (Ibrahim et al. 2011). Note that for a successful selection, all existing penalized-based methods require strong constraint on the correlation between important and unimportant covariates. However, high correlation is just a case particularly for mixed-effects models.

Accounting for these issues, we develop an orthogonalization-based approach to separately select both the fixed and random effects. It is very easy to implement as all selection steps are based on the least squares, and requires no specific distribution assumption other than the existence of fourth order moments. The method is also two-stage: selecting the fixed effects first and the random effects next. However, unlike Jiang and Rao (2003), we do not simply use marginal models to select the fixed effects, which simply regards the random effects as error terms in this stage. Our procedure is as follows. First, a QR decomposition (see Gentle 1998) of the design matrices z_i is applied to constructing a simple homogeneous linear regression model [in (2) in the next section], which does not depend on the random effects. The resulting homogeneous linear model is regarded as a first-stage initial working model. The selection for the fixed effects is then based on this newly defined model. Second, after the important fixed effects are selected to form a second-stage working model, selecting the random effects from this working model is then implemented. Note that to remove random effects from a model, we need to eliminate the corresponding entire rows and columns of D to form the final working model. This is the main difficulty in this stage. To solve this problem, a corrected Kronecker tensor product of model residuals is defined. When the first-stage initial working model and the second-stage working model are defined, the fixed and random effects are separately selected by a sophisticated variable selection approach. In this paper, we use the SCAD penalty for models (2) and (12) below to select the two kinds of effects successively. Of course, one can also use other penalties such as the adaptive LASSO (Zou 2006) which has the oracle property too. With this method, the following results are acquired:

- Orthogonalization makes submodel (2) below be a simple homogeneous linear regression model without the random effects. Only the second order moment of the error is needed for us to define the SCAD penalized least squares estimate $\hat{\beta}$ of β with the oracle property. The estimation procedure is very simple to be implemented and then not affected by the random effects.
- The method is ready to handle nonlinear/semiparametric/non-parametric models in which the fixed and random effects are of an additive structure.

The rest of the article is organized as follows. The main selection and estimation procedures are described in Sect. 2. Section 3 provides the asymptotic properties of the resulting estimates. Section 4 reports the simulation results. Some discussion is given in Sect. 5. All technical details are relegated to Appendix.

2 Methodology development

2.1 Selection of fixed effects

It is obvious that if we want to select the fixed effects with less impact from the random effects, a direct way is to remove the random effects from the model such that the selection is based on a working model without them. To this end, we assume with no loss of generality that the design matrices z_i are of full column rank. Recalling the definition of the QR decomposition of a matrix, the design matrix z_i can be decomposed

as $z_i = Q_i \begin{pmatrix} R_i \\ 0 \end{pmatrix}$, where Q_i is an orthogonal matrix and R_{zi} an upper triangular one. Partition Q_i to be $Q_i = (Q_{i1}, Q_{i2})$, where Q_{i1} is a $l_i \times q$ matrix and Q_{i2} is a $l_i \times (l_i - q)$ matrix. It is easy to check that $z_i = Q_{i1} R_i$, $Q_{i2}^\tau Q_{i1} = 0$ and then $Q_{i2}^\tau z_i = 0$; see Gentle (1998) (Sect. 3.2.2, pp. 95–97) for more details. From model (1), we have that, for $i = 1, \ldots, n$,

$$Q_{i2}^\tau y_i = Q_{i2}^\tau x_i \beta + Q_{i2}^\tau \varepsilon_i. \tag{2}$$

Note that model (2) is constructed in the orthogonal column space of the design matrices of z_i, and thus does not depend on the random effects. See Wu and Zhu (2010) for details. Define $Q_2 = \text{diag}(Q_{12}, \ldots, Q_{n2})$, $X = (x_1^\tau, \ldots, x_n^\tau)^\tau$, $Y = (Y_1, \ldots, Y_n)$ and $\varepsilon = (\varepsilon_1^\tau, \ldots, \varepsilon_n^\tau)^\tau$. Then we can rewrite model (2) in matrix form as

$$Q_2^\tau Y = Q_2^\tau X \beta + Q_2^\tau \varepsilon. \tag{3}$$

Define $N = \sum_{i=1}^n l_i$. It follows from $E(Q_2^\tau \varepsilon \varepsilon^\tau Q_2^\tau) = \sigma^2 I_{\text{N-nq}}$ that model (3) is homogeneous. One can easily obtain the best linear unbiased estimation of β via minimizing the sum of squared residual errors. Furthermore, for selection, one can minimize this sum with a penalized function on β to decide whether to include or exclude some fixed effects.

In this paper, SCAD (Fan and Li 2001) is applied to selecting effects. Of course, other selection methods are also feasible. We use SCAD due to its merits resulting in an estimate with unbiasedness, sparsity an continuity. The SCAD penalized least squares is as follows:

$$S_1(\beta) = \frac{1}{2}(Y - X\beta)^\tau P_{z^\tau} (Y - X\beta)^\tau + N_1 \sum_{j=1}^{p} p_{\lambda_1}(|\beta_j|), \tag{4}$$

where $N_1 = N - nq$, $P_{z^\tau} = Q_2 Q_2^\tau = \text{diag}(P_{z_1^\tau}, \ldots, P_{z_n^\tau})$, and the function $p_\lambda(x)$ has the first derivative $p'_\lambda(x)$ at any value λ as follows:

$$p'_\lambda(x) = \lambda \left\{ I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda) \right\}. \tag{5}$$

Here λ and a in (4) are tuning parameters and need to be selected in practice. However, in Fan and Li (2001), the selection of a, by data-driven method, cannot improve the performance significantly, and $a = 3.7$ was suggested and then we use it throughout our study. In Sect. 3, we will suggest how to apply the GCV criterion to choose the tuning parameter.

Note that the penalized function in (5) is singular at the origin, and it does not have continuous second order derivative. Hence there not exist an exact solution by minimizing the objective function (4). The following approximation is used to solve this optimization problem, up to some constant terms:

$$2S_1(\beta) \propto (Y - X\beta)^\tau P_{z^\tau} (Y - X\beta)^\tau + N_1 \beta^\tau \Sigma_{\lambda_1}(\beta_0)\beta,$$

where

$$\Sigma(\lambda_1, \beta) = \text{diag}\left\{p'_{\lambda_1}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda_1}(|\beta_p|)/|\beta_p|\right\}. \tag{6}$$

See Fan and Li (2001) for details. Use the orthogonalization-based least squares estimate, as an initial estimate, $\hat{\beta}_{\text{obe}} = (X^\tau P_{z^\tau} X)^{-1} X^\tau P_{z^\tau} Y$, which is defined in Wu and Zhu (2010). Then the $(k+1)$th iterative quadratic minimization problem in (4) can be solved by computing the ridge estimation

$$\hat{\beta}^{(k+1)}_{\text{obpe}} = \left[X^\tau P_{z^\tau} X + N_1 \Sigma(\lambda_1, \hat{\beta}^{(k)}_{\text{obpe}})\right]^{-1} X^\tau P_{z^\tau} Y. \tag{7}$$

Now we study the asymptotic theory of $\hat{\beta}_{\text{obpe}}$. Rewrite $\beta = (\beta_1^\tau, \beta_2^\tau)^\tau$. Without loss of generality, we assume that $\beta_2 = 0$. Correspondingly, we have $x_i = (x_{i1}, x_{i2})$ and $X = (X_1, X_2)$. Let s denote the number of nonzero components of β_1. Denote $\beta_1 = (\beta_{11}, \ldots, \beta_{1s})^\tau$ and let

$$a_n(\lambda_1) = \max\left\{p'_{\lambda_1}(|\beta_{1j}|), j = 1, \ldots, s\right\}. \tag{8}$$

Then the following theorem states that $\hat{\beta}_{\text{obpe}}$ converges at the rate $O_p(n^{-1/2} + a_n(\lambda_1))$.

Theorem 1 *Assume that the moments up to fourth order of the errors exist and conditions* $(C1) - (C3)$ *in Appendix hold. When*

$$\max\left\{|p''_{\lambda_1}(|\beta_{1j}|)| : j = 1, \ldots, s\right\} \to 0, \tag{9}$$

there exists a local minimizer $\hat{\beta}_{\text{obpe}}$ *of* $S_1(\beta)$ *such that* $\| \hat{\beta}_{\text{obpe}} - \beta \| = O_p(n^{-1/2} + a_n(\lambda_1))$.

Let $\hat{\beta}_{\text{jobpe}}$ be the corresponding orthogonalization-based penalized estimates of β_j for $j = 1, 2$. We can prove the oracle property that $\hat{\beta}_{2\text{obpe}} = 0$ and $\hat{\beta}_{1\text{obpe}}$ is asymptotically normal. Define

$$\Gamma_1 = \text{diag}\left(p''_{\lambda_1}(|\beta_{11}|), \ldots, p''_{\lambda_1}(|\beta_{1s}|)\right),$$

and

$$\omega_1 = \left(p'_{\lambda_1}(|\beta_{11}|)\text{sgn}(\beta_{11}), \ldots, p'_{\lambda_1}(|\beta_{1s}|)\text{sgn}(\beta_{1s})\right)^\tau.$$

In the following we have the root n consistency and asymptotic normality of $\hat{\beta}_{\text{obpe}}$.

Theorem 2 *In addition to the conditions in Theorem 1, assume that*

$$\liminf_{n\to\infty} \liminf_{\beta\to 0+} p'_{\lambda_1}(\beta)/\lambda_1 > 0. \tag{10}$$

If $\lambda_1 \to 0$ *and* $\sqrt{n}\lambda_1 \to \infty$ *as* $n \to \infty$, *then with a probability tending to 1, the root-n consistent local minimizers* $\hat{\beta}_{\text{obpe}} = (\hat{\beta}^\tau_{1\text{obpe}}, \hat{\beta}^\tau_{2\text{obpe}})^\tau$ *in Theorem 1 satisfy:*

(a) Sparsity: $\hat{\beta}_{2\text{obpe}} = 0$. *Further, we have*
(b) Asymptotic normality: as $n \to \infty$,

$$\sqrt{n}(\Sigma_{11} + m_1\Gamma_1)(\hat{\beta}_{1\text{obpe}} - \beta_1 + (\Sigma_{11} + m_1\Gamma_1)^{-1}\omega_1) \xrightarrow{L} \mathcal{N}(0, \sigma^2\Sigma_{11}),$$

where $\Sigma_{11} = \lim_{n\to\infty} X_1^\tau P_{z^\tau} X_1/n$.

2.2 Selection of random effects

Now we turn to selecting the random effects. Write $D = (d_{ij})$. Since the random effects b_i have zero mean, the zero variance $d_{jj} = 0$ means that the jth component b_{ij} has the degenerate distribution with mass 1 at the origin and the corresponding row and column are also zero. Then it is sufficient to select the nonzero diagonal components of D.

From Theorem 2, we have $\hat{\beta}_{2\text{obpe}} = 0$ with a probability tending to one, and $\hat{\beta}_{1\text{obpe}}$ is asymptotically normal with the bias $(\Sigma_{11} + m_1\Gamma_1)^{-1}\omega_1$ tending to zero as $n \to \infty$. Let $\hat{\beta}_{1\text{obe}}$ be the corresponding orthogonalization-based least squares estimate of β_1. Define $V_i = E(y_i - x_i\beta)(y_i - x_i\beta)^\tau = z_i D z_i^\tau + \sigma^2 I_{l_i}$. Straightforwardly one can solve the following minimization problem to estimate D and σ^2:

$$W_2^*(D, \sigma^2) = \frac{1}{2}\sum_{i=1}^{n}\Big((y_i - x_i\hat{\beta}_{1\text{obe}}) \otimes (y_i - x_i\hat{\beta}_{1\text{obe}}) - \text{vec}(V_i)\Big)^\tau \times\Big((y_i - x_i\hat{\beta}_{1\text{obe}}) \otimes (y_i - x_i\hat{\beta}_{1\text{obe}}) - \text{vec}(V_i)\Big).$$

Here and throughout this paper a vector vec(A) is formed by stacking the complete columns of any matrix A, and $\otimes$ stands for the Kronecker tensor product. However, for fixed i,

$$\begin{aligned} &E(y_i - x_i\hat{\beta}_{1\text{obe}})(y_i - x_i\hat{\beta}_{1\text{obe}})^\tau \\ &\quad = V_i - \sigma^2 P_{z_i^\tau} x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau - \sigma^2 x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau P_{z_i^\tau} \\ &\qquad + \sigma^2 x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau. \end{aligned}$$

This means that the estimates are biased as the last three terms are not zero. To make a bias correction, we define

$$\begin{aligned} \tilde{y}_i = (y_i - x_{i1}\hat{\beta}_{\text{obe}1}) \otimes (y_i - x_{i1}\hat{\beta}_{\text{obe}1}) + \hat{\sigma}^2_{\text{obe}}\Big[& P_{z_i^\tau} x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau \\ & + x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau P_{z_i^\tau} - x_{i1}(X_1^\tau P_{z^\tau} X_1)^{-1} x_{i1}^\tau)\Big], \end{aligned} \quad (11)$$

where

$$\hat{\sigma}^2_{\text{obe}} = (Y - X_1\hat{\beta}_{\text{obe}1})^\tau P_{z^\tau}(Y - X_1\hat{\beta}_{\text{obe}1})/(N_1 - s)$$

can be proved to be an unbiased estimate of the variance σ^2. Let the vector vech(A) be formed by stacking the columns under the diagonal line of A. Define B_l be an $l^2 \times (l^2 + l)/2$ permutation matrix such that $\text{vec}(A) = B_l \text{vech}(A)$ for any $l \times l$ symmetric matrix A. Define $u_i = (\text{vec}(I_{l_i}), (z_i \otimes z_i)B_q)$ and $\theta = (\sigma^2, \text{vech}(D)^\tau)^\tau$. Then we can construct the following working model:

$$\tilde{y}_i = u_i\theta + \varepsilon_i^*, \tag{12}$$

where the $\varepsilon_i^* = (z_i b_i + \varepsilon_i) \otimes (z_i b_i + \varepsilon_i) - \text{vech}(V_i)$ are independent random vectors with zero mean. Define $\tilde{Y} = (\tilde{y}_1^\tau, \ldots, \tilde{y}_n^\tau)^\tau$ and $U = (u_1^\tau, \ldots, u_n^\tau)^\tau$. Again, one can estimate θ by the least squares:

$$\hat{\theta}_{\text{lse}} = (U^\tau U)^{-1} U^\tau \tilde{Y}.$$

It is easy to check that $\hat{\theta}_{\text{lse}}$ is an unbiased and consistent estimate of θ. Rewrite $V_i(\theta) = V_i$. Note that $V_i^{-1/2}(z_i b_i + \varepsilon_i)$ are independent with mean zero and covariance matrix I_{l_i}. Moreover, the covariance matrix of ε_i^* is unknown. By taking $W = \text{diag}(W_1, \ldots, W_n)$ with $W_i = V_i \otimes V_i$ as a weighted matrix, we suggest the following algorithm to estimate θ:

Step 1. Take $\hat{\theta}_{lse}$ as an initial estimate of θ

Step 2. Compute $\hat{V}_i = z_i \hat{D}_{\text{lse}} z_i^\tau + \hat{\sigma}^2_{\text{lse}} I_{l_i}$.

Step 3. Compute the orthogonalization-based weighted penalized least squares estimate $\hat{\theta}$ by minimizing the following objective function with the SCAD penalty:

$$S_2(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\tilde{Y} - u_i\theta)^\tau(\hat{V}_i \otimes \hat{V}_i)^{-1}(\tilde{Y} - u_i\theta) + N_2 \sum_{j=1}^{(q^2+q)/2+1} p_{\lambda_2}(|\theta_j|) \tag{13}$$

with $N_2 = \sum_{i=1}^n l_i^2$.

Step 4. Replace $\hat{D}_{\text{lse}}$ by $\hat{D}$ in Step 2. Repeat steps 2 and 3 until convergence.

Recall that the objective function (13) is solved by iteratively computing the ridge estimation of (7). Similarly, we can obtain the following $(k + 1)$th iterative ridge estimate of θ in Step 3

$$\hat{\theta}^{k+1} = \left[U^\tau \hat{W}^{-(k)} U + N_2 \Sigma_{\lambda_2}(\hat{\theta}^k)\right]^{-1} U^\tau \hat{W}^{-(k)} \tilde{Y}, \tag{14}$$

where $\hat{W}^{(k)}$ is the kth iterative plug-in estimate of W with $\hat{\theta}^k_{\text{obwpe}}$ replacing θ, and $\Sigma_{\lambda_2}(\theta) = \text{diag}(0, p'_{\lambda_2}(|\theta_1|)/|\theta_1|, \ldots, p'_{\lambda_2}(|\theta_{(q^2+q)/2}|)/|\theta_{(q^2+q)/2}|)$.

Without loss of generality, we assume that the first t random effects b_{i1} have nonzero variances and the last $q - t$ random effects b_{i2} are degenerate at 0. Correspondingly, rewrite $z_i = (z_{i1}, z_{i2})$ and $\theta = (\theta_1^\tau, \theta_2^\tau)^\tau$, where $\theta_2 = 0$. Moreover, denote $U_1 =$

$(u_{i1}^{\tau}, \ldots, u_{n1}^{\tau})^{\tau}$ with $u_{j1} = (z_{j1} \otimes z_{j1})B_t$. Rewrite $\theta_1 = (\theta_{11}, \ldots, \theta_{1r})^{\tau}$ with $r = t(2q - t + 1)/2 + 1$. Define

$$b_n(\lambda_2) = \max\left\{p'_{\lambda_2}(|\theta_{1j}|),\, j = 2, \ldots, r\right\}. \tag{15}$$

Then the penalized weighted least squares estimate $\hat{\theta}$ defined in the above procedure converges at the rate $O_p(n^{-1/2} + b_n(\lambda_2))$.

Theorem 3 *Assume that the moments up to fourth order of the random effects and errors exist and conditions* $(C1) - (C7)$ *are satisfied. If*

$$\max\left\{|p''_{\lambda_2}(|\theta_{1j}|)|,\, j = 1, \ldots, r\right\} \to 0 \tag{16}$$

holds, then there exists a local minimizer $\hat{\theta}$ *of the corresponding* $S_2(\theta)$ *such that* $\| \hat{\theta} - \theta \| = O_p(n^{-1/2} + b_n(\lambda_2))$.

Let $\hat{\theta}$ be the corresponding orthogonal-based penalized estimate θ_j for $j = 1, 2$. Define

$$\Gamma_2 = \mathrm{diag}(0,\, p''_{\lambda_2}(|\theta_{12}|), \ldots, p''_{\lambda_2}(|\theta_{1r}|)),$$

$$\omega_2 = (0,\, p'_{\lambda_2}(|\theta_{12}|)\mathrm{sgn}(\theta_{12}), \ldots, p'_{\lambda_2}(|\theta_{1r}|)\mathrm{sgn}(\theta_{1r}))^{\tau}.$$

The following theorem states the oracle property of $\hat{\theta}$.

Theorem 4 *In addition to the conditions in Theorem 3, assume that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_2}(\theta)/\lambda_1 > 0. \tag{17}$$

If $\lambda_2 \to 0$ *and* $\sqrt{n}\lambda_2 \to \infty$ *as* $n \to \infty$, *then with a probability tending to 1, the root-n consistent local minimizer* $\hat{\theta}$ *in Theorem 3 satisfies:*

(a) Sparsity: $\hat{\theta}_2 = 0$.
Further, we have
(b) Asymptotic normality: as $n \to \infty$,

$$\sqrt{n}(\Sigma_{21} + m_1\Gamma_2)(\hat{\theta}_1 - \theta_1 + (\Sigma_{21} + m_2\Gamma_2)^{-1}\omega_2) \xrightarrow{L} \mathcal{N}(0, \sigma^2\Psi_1),$$

where Σ_{21} *and* Ψ_1 *are defined similarly to* Σ_2 *(in condition* $(C2)$*) and* Ψ *(in condition* $(C5)$*) by using* z_{i1} *to replace* z_i.

3 Selection of tuning parameters

Note that properly choosing the tuning parameters λ_1 in (9) and λ_2 in (15) is crucial for implementation. In the literature, cross-validation (CV) and generalized cross-validation (GCV) are often used. In this paper, we adopt the idea of GCV. Refer to Fan and Li (2001) for details.

For the fixed effects, recall that we have used the homogeneous model (2) to get the orthogonalization-based penalized least squares estimate that is the minimizer of the quadratic function $S_1(\beta)$ of (4) with the SCAD penalty. The estimate of β relative to the fixed effects is obtained by computing the ridge estimation (7) iteratively. In the $(k+1)$th iterative step, the fitted value of $Q_2^\tau Y$ is $Q_2^\tau X\hat{\beta}_{\text{obe}}^{(k+1)}$. Let

$$P(\lambda_1, \beta) = \left(X^\tau P_{z^\tau} X + (N - nq)\Sigma(\lambda_1, \beta)\right]^{-1} X^\tau P_{z^\tau} X.$$

It follows that $tr(P(\lambda_1, \beta))$ is the number of effective parameters in the penalized least squares of (4). Therefore, the corresponding GCV score is defined as

$$\text{GCV}(\lambda_1) = \frac{1}{N_1} \frac{(Y - X\hat{\beta}_{\text{obe}}^{(k+1)})^\tau P_{z^\tau}(Y - X\hat{\beta}_{\text{obe}}^{(k+1)})}{|1 - \frac{1}{N-nq} tr(P(\lambda_1, \hat{\beta}_{\text{obe}}^{(k+1)}))|^2}.$$

For the random effects, by regressing $\tilde{Y}(\hat{\sigma}^2_{\text{obe}})$ on U, we have used the iterative generalized penalized estimation procedure which satisfies (13). Similarly, the corresponding GCV score in the $(k+1)$th iterative step can be defined as

$$\text{GCV}(\lambda_2) = \frac{1}{N_2} \frac{(\tilde{Y}(\hat{\sigma}^2) - U\hat{\theta}^{(k+1)})^\tau W^{-(k)} (\tilde{Y}(\hat{\sigma}^2) - U\hat{\theta}^{(k+1)})}{|1 - \frac{1}{N_2} \text{tr}(P(\lambda_2, \hat{\theta}^{(k+1)}))|^2},$$

where $P(\lambda_2, \theta) = \left[U^\tau W^{-1} U + \lambda_2 \Sigma_{\lambda_2}(\theta)\right]^{-1} U^\tau W^{-1} U$.

4 Simulation study

In this section, some simulation studies are carried out to assess the finite sample performance of the proposed orthogonal-based SCAD (O-SCAD) method. We also make a comparison with other methods: M-ALASSO, ALASSO, and B-SCAD suggested by Bondell et al. (2010), Ibrahim et al. (2011), and Peng and Lu (2012), respectively. The first two examples are in favor of M-ALASSO, ALASSO, and B-SCAD. Note that Peng and Lu (2012) showed some advantages of B-SCAD when compared with M-ALASSO and ALASSO in the settings they considered. We thus use their settings such that a comparison can be made to see what our method loses and gains. It is worth saying that we do not include a real data example in this section, because we mainly focus on the comparison between our method and existing ones to examine the performance when correlation between fixed and random effects is fairly high.

Example 1 Generate 100 datasets from the linear mixed-effects model

$$y_{ij} = \sum_{k=1}^{9} x_{ijk}\beta_k + b_{i0} + \sum_{k=1}^{3} z_{ijk}b_{ik} + \varepsilon_{ij},$$

where $i = 1, \ldots, n$, $j = 1, \ldots, l_i$, and $\varepsilon_{ij} \sim \mathrm{N}(0, 1)$. However, all the covariates x_{ijk} and z_{ijl} independently come from the uniform distribution $U(-2, 2)$, the true value of the nonzero fixed effects β_{10} is $(1, 1)^\tau$, and the first three efficient random effects are independently generated from the 3-dimensional multivariate normal distribution with zero-mean and covariance matrix

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

Then we consider two combinations:
$I = (n = 30, l_i \equiv 5)$; and
$II = (n = 60, l_i \equiv 10)$, for $i = 1, \ldots, n$.

Example 2 Consider two sample sizes $n = 10$ and 20, and $l_i \equiv 10$; the dimension of the fixed effects $p = 5$ with $\beta = (1, 0.5, 1, 0, 0)$. The model errors and covariates are all generated from the standard normal distribution. Generate 100 datasets from this linear mixed-effects model $y_{ij} = \sum_{k=1}^{5} x_{ijk}\beta_k + \sum_{k=1}^{4} z_{ijk}b_{ik} + \varepsilon_{ij}$, where the first two random effects are active with the covariance matrix $D = \begin{pmatrix} 0.5 & 0.345 \\ 0.345 & 1 \end{pmatrix}$, and the distributions of the corresponding ones are separately:

(2.1) $N(0, D)$;
(2.2) $\frac{\sqrt{3}}{2}t(8, D)$; and
(2.3) $C_t(8, \rho, \Gamma(0.5, 1) - 0.5, \Gamma(1, 1) - 1)$ that is a two-dimensional t copula distribution with degree 8 of freedom, correlation matrix $\rho = \begin{pmatrix} 1 & 0.5006 \\ 0.5006 & 1 \end{pmatrix}$, and the marginal distributions are two different Gamma ones.

Example 3 In this example, we generate 200 datasets from the following model

$$y_{ij} = \sum_{k=1}^{4} x_{ijk}\beta_k + \sum_{k=1}^{4} z_{ijk}b_{ik} + \varepsilon_{ij}, \quad i = 1, \ldots, 40, \ \ j = 1, \ldots, 10,$$

where $\beta = (1, -1, 0, 0)^\tau$, the first two efficient random effects follow a joint normal distribution with zero mean and covariance matrix D defined in Example 2, and the errors are generated from the standard normal distribution independently. Moreover, the covariates x_{ijk} and z_{ijk} are designed in the following two cases:

Table 1 For Example 1, the results of CM, CF, and CR

Case	Method	% CM	% CF	% CR
I	O-SCAD	84	95.5	87
	M-ALASSO	71	73	79
	ALASSO	62	63	68
	B-SCAD	19	43	20
II	O-SCAD	95	100	95
	M-ALASSO	83	83	89
	ALASSO	74	75	81
	B-SCAD	80	80	86

(3.1) For $k = 1, 2, 3, 4$, $(x_{ijk}, z_{ijk})^\tau \sim_{i.i.d} N(0, C_1)$ with $C_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho = 0, 0.85$;

(3.2) For $|k_1 - k_2| = 2$, $(x_{ijk_1}, z_{ijk_2})' \sim_{i.i.d} N(0, C_1)$.

The results for Example 1 with the following different measures are reported in Table 1: the percentage of the times of which the correct model is selected (CM); the percentage of times of which the correct fixed effects are selected (CF); the percentage of times of which the correct random effects are selected (CR). The comparison is made with M-ALASSO (Bondell et al. 2010), ALASSO (Ibrahim et al., 2011), and B-SCAD (Peng and Lu, 2012). For Example 2, the results are tabulated in Tables 2 and 3. In Table 2, the values are the numbers of correctly and incorrectly selected fixed and random effects by O-SCAD, M-ALASSO, ALASSO, and B-SCAD. In Table 3, we report the median of biases and of absolute deviation of the estimation of the nonzero fixed and random effects. Finally, Table 4 reports that the results for Example 3. R codes from Bondell et al. (2010) and Ibrahim et al. (2011) and Matlab code from Peng and Lu (2012) are used to implement their methods.

From Table 1, it is easy to see that our method performs efficiently. As the sample size grows, the number of correctly selecting the fixed and random effects reasonably grows. In the limited simulations, our method outperforms all the competitors.

Whether the normality assumption of the random effects is true or not, the results in Table 2 indicate that our method always performs best for selecting the fixed effects; ALASSO performs best to select the active random effects, but performs worst to remove the insignificant random effects; O-SCAD's performance is just opposite to that of ALASSO. Moreover, the performance of O-SCAD and A-LASSO are comparable in estimating the random effects.

As for the performance of parameter estimation, Table 3 shows that all of the methods perform satisfactorily. However, there are so many active random effects are estimated wrongly, the medians of B-SCAD estimates of D_{ij} are zero. For example, that of D_{11} is zero for $n = 10$ in cases (2.1) and (2.3).

Finally, Table 4 shows that our method uniformly outperforms M-ALASSO, ALASSO, and B-SCAD. A comparison between the results with $\rho = 0$ and $\rho = 0.8$ suggests that the performances of the competitors are significantly affected by the correlation between the inefficient fixed effects and efficient random effects. Our method still works well even when the correlated coefficient ρ increases to 0.8.

Table 2 For Example 2, in 100 replications, the times that β_4, β_5, D_{33}, and D_{44} are correctly selected to be zero, and the other parameters are incorrectly to be zero

Distribution	Method	β_1	β_2	β_3	β_4	β_5	D_{11}	D_{22}	D_{33}	D_{44}
		$n = 10$								
(2.1)	O-SCAD	0	0	0	100	100	8	7	92	93
	M-ALASSO	0	2	0	90	90	16	0	98	100
	ALASSO	0	0	0	55	47	2	0	38	35
	B-SCAD	0	0	0	91	93	57	25	97	100
	O-SCAD	0	0	0	100	100	2	0	97	98
	M-ALASSO	0	0	0	92	92	2	0	98	100
	ALASSO	0	0	0	65	67	0	0	44	58
	B-SCAD	0	0	0	93	93	35	7	99	99
		$n = 10$								
(2.2)	O-SCAD	0	0	0	100	100	4	8	93	93
	M-ALASSO	0	0	0	92	89	16	8	92	90
	ALASSO	0	0	0	63	49	1	1	31	36
	B-SCAD	0	0	0	89	90	37	23	100	100
		$n = 20$								
	O-SCAD	0	0	0	100	100	2	2	96	97
	M-ALASSO	0	0	0	94	96	18	11	97	99
	ALASSO	0	0	0	63	66	0	0	50	53
	B-SCAD	0	0	0	98	93	17	8	97	100
		$n = 10$								
(2.3)	O-SCAD	0	0	0	100	100	8	2	93	92
	M-ALASSO	0	2	0	99	92	31	3	95	92
	ALASSO	0	0	0	49	58	4	0	39	34
	B-SCAD	0	0	0	85	87	70	35	100	100
		$n = 20$								
	O-SCAD	0	0	0	100	100	5	1	94	93
	M-ALASSO	0	0	0	98	93	7	0	93	90
	ALASSO	0	0	0	75	72	1	0	54	59
	B-SCAD	0	0	0	95	95	49	18	99	99

5 Discussion

In this paper, we suggested an orthogonalization-based approach to separately select both the fixed and random effects. It is very easy to implement as all of the selection steps are based on the least squares, and requires no specific distribution assumption other than the existence of fourth order moments. However, the orthogonalization requires a strong condition on the separation of the fixed and random effects, otherwise, it may lead to loss of active efficient fixed effects. Theoretically, when the matrix Q_2 in Sect. 2.1 has lower rank, the loss of active fixed effects can occur. This is the case a

Table 3 Median of bias and median absolute deviation (MAD) of the significant fixed and random effects in 100 datasets (MAD is in parenthesis)

Case	Method	β_1	β_2	β_3	D_{11}	D_{22}	D_{12}
		$n = 10$					
(1.1)	O-SCAD	0.004 (0.083)	−0.008 (0.073)	0.013 (0.091)	−0.014 (0.166)	0.004 (0.406)	−0.354 (0.346)
	M-ALASSO	−0.051 (0.076)	−0.036 (0.074)	−0.031 (0.088)	−0.185 (0.244)	0.470 (0.470)	−0.721 (0.721)
	ALASSO	0.001 (0.074)	−0.026 (0.091)	−0.017 (0.077)	−0.108 (0.186)	0.420 (0.420)	−0.887 (0.887)
	B-SCAD	0.004 (0.091)	−0.015 (0.077)	−0.004 (0.072)	−0.500 (0.000)	−0.059 (0.436)	−0.354 (0.000)
		$n = 20$					
	O-SCAD	0.006 (0.066)	−0.007 (0.072)	0.007 (0.073)	−0.010 (0.147)	0.044 (0.196)	0.011 (0.154)
	M-ALASSO	−0.010 (0.058)	−0.028 (0.050)	−0.021 (0.059)	−0.071 (0.131)	0.495 (0.495)	−0.741 (0.741)
	ALASSO	−0.009 (0.064)	−0.017 (0.047)	−0.003 (0.045)	0.008 (0.128)	0.443 (0.443)	−0.853 (0.853)
	B-SCAD	0.000 (0.051)	−0.002 (0.065)	−0.004 (0.057)	−0.061 (0.266)	0.035 (0.278)	−0.068 (0.286)
		$n = 10$					
(2.2)	O-SCAD	0.010 (0.082)	−0.011 (0.078)	0.014 (0.104)	0.371 (0.357)	−0.126 (0.367)	−0.155 (0.199)
	M-ALASSO	−0.043 (0.089)	−0.067 (0.097)	−0.039 (0.090)	−0.372 (0.346)	0.404 (0.404)	−0.802 (0.802)
	ALASSO	0.044 (0.095)	0.012 (0.084)	−0.003 (0.056)	−0.084 (0.191)	0.465 (0.465)	−0.894 (0.894)
	B-SCAD	−0.008 (0.052)	−0.004 (0.060)	0.012 (0.052)	0.388 (0.303)	−0.157 (0.260)	−0.001 (0.292)
		$n = 20$					
	O-SCAD	0.001 (0.054)	0.016 (0.059)	0.005 (0.065)	−0.174 (0.211)	−0.399 (0.279)	−0.254 (0.010)
	M-ALASSO	−0.013 (0.062)	−0.041 (0.060)	−0.019 (0.072)	−0.238 (0.296)	0.444 (0.444)	−0.729 (0.729)
	ALASSO	−0.021(0.049)	−0.013 (0.063)	−0.008 (0.061)	−0.045 (0.177)	0.486 (0.486)	−0.868 (0.868)
	B-SCAD	0.007 (0.053)	−1.085 (0.103)	−0.003 (0.046)	0.352 (0.443)	−0.068 (0.382)	−0.089 (0.265)
		$n = 10$					
(2.3)	O-SCAD	0.017 (0.061)	−0.010 (0.063)	−0.003 (0.057)	−0.146 (0.213)	−0.086 (0.375)	−0.127 (0.227)
	M-ALASSO	−0.010 (0.077)	−0.046 (0.090)	−0.037 (0.067)	−0.342 (0.383)	0.199 (0.363)	−0.863 (0.868)
	ALASSO	0.001 (0.077)	−0.032 (0.085)	0.007 (0.074)	−0.176 (0.330)	0.325 (0.325)	−0.918 (0.918)

Table 3 continued

Case	Method	β_1	β_2	β_3	D_{11}	D_{22}	D_{12}
	B-SCAD	0.039 (0.067)	0.023 (0.084)	−0.062 (0.075)	−0.500 (0.000)	−0.273 (0.727)	−0.354 (0.000)
		$n = 20$					
	O-SCAD	0.014 (0.064)	−0.001 (0.066)	0.002 (0.050)	−0.148 (0.165)	−0.228 (0.289)	−0.198 (0.156)
	M-ALASSO	−0.022 (0.065)	−0.032 (0.067)	−0.024 (0.050)	−0.188 (0.281)	0.450 (0.450)	−0.776 (0.776)
	ALASSO	0.006 (0.062)	−0.005 (0.062)	−0.009 (0.052)	−0.163 (0.261)	0.476(0.476)	−0.886 (0.886)
	B-SCAD	−0.009 (0.060)	−0.001 (0.052)	−0.002 (0.059)	−0.273 (0.227)	−0.223 (0.394)	−0.354 (0.000)

Table 4 For Example 3, the times that the fixed effects and random effects are selected correctly, and the results of CM, CF, and CR

Case	ρ	Method	β	Diag (D)	% CM	% CF	% CR
(3.1)	0	O-SCAD	(200, 200, 200, 200)	(198, 200, 200, 200)	99	100	99
		M-ALASSO	(200, 200, 200, 200)	(199, 199, 200, 199)	91	97.5	98.5
		ALASSO	(200, 200, 192, 192)	(200, 200, 133, 133)	53	93	54
		B-SCAD	(200, 200, 197, 195)	(146, 186, 197, 198)	65	96	66
	0.8	O-SCAD	(200, 200, 200, 200)	(200, 200, 199, 199)	99	100	99
		M-ALASSO	(200, 200, 185, 192)	(200, 200, 198, 196)	88	88.5	97
		ALASSO	(200, 200, 185, 192)	(200, 200, 110, 111)	37	89.5	39.5
		B-SCAD	(200, 200, 196, 198)	(146, 180, 200, 198)	62.5	97	63.5
(3.2)	0	O-SCAD	(200, 200, 200, 200)	(198, 200, 199, 199)	98	100	98
		M-ALASSO	(200, 200, 193, 187)	(198, 199, 198, 198)	90	91	96
		ALASSO	(200, 200, 191, 185)	(200, 200, 124, 133)	54	89.5	55.5
		B-SCAD	(200, 200, 195, 192)	(162, 182, 196, 196)	70	94	72
	0.8	O-SCAD	(200, 200, 200, 200)	(200, 200, 198, 198)	98	100	98
		M-ALASSO	(200, 200, 189, 191)	(200, 200, 195, 197)	88	90	96
		ALASSO	(200, 200, 187, 180)	(200, 200, 138, 138)	52	84	57.5
		B-SCAD	(200, 200, 186, 195)	(146, 188, 199, 193)	67	91	67

variable is included as both a fixed effect and a random effect at the same time. Once the fixed effects and random effects are different from each other, the numerical studies suggest that O-SCAD is very efficient to select the fixed effects. But it is still in need to study how to improve the efficiency for selecting the insignificant random effect. Another important issue is how to apply our method when the dimension of fixed effects (random effects) tends to infinity as sample n goes to infinity. The research is ongoing.

Acknowledgements The first author was partially supported by the National Natural Science Foundation of China (Grant Nos. 11101157, 11371142), and the 111 project (B14019). The third author was supported by Natural Science Foundation of Jiangsu Province, China (No. BK20140617) and the NSFC Grant No. 11501099. The last author was supported by a grant from the University Grants Council of Hong Kong, China. The authors thank the editor, the associate editor and referees for their constructive suggestions that led to the improvement of an early manuscript. The authors are grateful to Dr. H. D. Bodell, H. Peng and H. Zhu for providing us the codes they used.

6 Appendix

We first assume the following conditions for the results.

(C1) Assume that $\lim_{n\to\infty} N/n = m_1 + q$, and $\lim_{n\to\infty} N_2/n = m_2$.

(C2) Assume that $\Sigma_1 = \lim_{n\to\infty} X^\tau P_{z^\tau} X/n$ and $\Sigma_2 = \lim_{n\to\infty} U^\tau W_0^{-1} U/n$

(C3) Assume that $\lim_{n\to\infty} \frac{\max_{1\le i\le n, 1\le j\le l_i} \|x_{ij}^\tau x_{ij}\|}{\sqrt{n}} = 0$.

(C4) Assume that $\lim_{n\to\infty} \frac{\max_{1\le i\le n, 1\le j\le q} \|z_{ij}^\tau z_{ij}\|}{\sqrt{n}} = 0$.

(C5) Assume that $\Psi = \lim_{n\to\infty} \frac{1}{n} Cov(I_1)$, where $Cov(I_1)$ are defined in (23).

Lemma 1 *Under the conditions in Theorem 2, we have, as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{1\text{obe}} - \beta) \xrightarrow{L} \mathcal{N}(0, \sigma^2 (X_1^\tau P_{z^\tau} X_1)^{-1}), \tag{18}$$

and $\sqrt{n}(\hat{\sigma}^2_{obe} - \sigma^2) = o_p(1)$.

Proof of Theorem 1 Let $\alpha_{n1} = n^{-1/2} + a_n(\lambda_1)$. For any given $\epsilon > 0$, if there exists a large constant C such that

$$P\left\{ \inf_{\|c\|=C} S_1(\beta_0 + \alpha_n c) > S_1(\beta_0) \right\} \ge 1 - \epsilon, \tag{19}$$

then there exists a local minimizer in the ball $\{\beta + \alpha_n c : \|c\| \le C\}$ with a probability at least $1 - \epsilon$. It follows that there exists a local minimizer such that $\|\hat{\beta}_{\text{obpe}} - \beta\| = O_p(\alpha_{n1})$. Hence it is sufficient to show that (19) is true.

Let $M_1(\beta) = \frac{1}{2}(Y - X\beta)^\tau P_{z^\tau}(Y - X\beta)$. Recalling $p_\lambda(0) = 0$, we have

$$
\begin{aligned}
&S_1(\beta + \alpha_{n1}c) - S_1(\beta) \qquad (20)\\
&\geq M_1(\beta + \alpha_{n1}c) - M_1(\beta) + (N - nq)\sum_{j=1}^{s}\{p_{\lambda_1}(|\beta_{1j} + \alpha_{n1}c_j|) - p_{\lambda_1}(|\beta_{1j}|)\}\\
&= -\alpha_{n1}c^\tau X^\tau P_{z^\tau}(Y - X\beta) + \frac{n\alpha_{n1}^2}{2}c^\tau \Sigma_1 c(1 + o_p(1))\\
&\quad + \sum_{j=1}^{s}(N - nq)\left[\alpha_{n1}p'_{\lambda_1}(|\beta_{1j}|)\mathrm{sgn}(\beta_{1j})c_j + \frac{\alpha_{n1}^2}{2}p''_{\lambda_1}(|\beta_{1j}|)c_j^2(1 + o(1))\right].
\end{aligned}
$$

By model (3), we have $X^\tau P_{z^\tau}(Y - X\beta_0) = X^\tau P_{z^\tau}\varepsilon$ which is a sum of zero mean independent random vectors. Under conditions $(C2)$ and $(C4)$, it is not difficult to verify that the Lindeberg's condition holds. By the Lindeberg–Feller central limit theorem, as $n \to \infty$

$$
\frac{1}{\sqrt{n}}X^\tau P_{z^\tau}(Y - X\beta) \xrightarrow{L} \mathcal{N}(0, \sigma^2(X^\tau P_{z^\tau}X)^{-1}).
$$

Thus the first term on the right-hand side of (20) is at a rate $O_p(n^{1/2}\alpha_n) = O_p(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|c\| = C$. Note that the third term in (20) is bounded by

$$
(N - nq)\left\{\sqrt{s}\alpha_n a_n\|c\| + \frac{1}{2}\alpha_n^2 \max\left\{|p''_{\lambda_n}(|\beta_{1j}|)| : \beta_{1j} \neq 0\right\}\|c\|^2\right\} = O_p(n\alpha_n^2).
$$

This is also dominated by the second term of (20). Hence, by choosing a sufficiently large C, (19) holds. This completes the proof of Theorem 1. □

Proof of Theorem 2 Consider part (a). Let $\beta_2 = (\beta_{21}, \ldots, \beta_{2(p-s)})^\tau$. Similar to Fan and Li (2001), it is sufficient to show that with a probability tending to 1 as $n \to \infty$, for any β_1 satisfying $\beta_1^* - \beta_1 = O_p(n^{-1/2})$ and for some $\epsilon_n = Cn^{-1/2}$ and $j = 1, \ldots, p - s$,

$$
\begin{aligned}
\frac{\partial S(\beta)}{\partial \beta_{2j}} &> 0 \quad \text{for} \quad 0 < \beta_{2j} < \epsilon_n \qquad (21)\\
&< 0 \quad \text{for} \quad -\epsilon_n < \beta_{2j} < 0. \qquad (22)
\end{aligned}
$$

By the Taylor expansion, we have

$$
\begin{aligned}
\frac{\partial S_1(\beta^*)}{\partial \beta_{ij}} &= \frac{\partial M_1(\beta^*)}{\partial \beta_{ij}} + (N - nq)p'_{\lambda_1}(|\beta_j^*|)\mathrm{sgn}(\beta_j^*)\\
&= \frac{\partial M_1(\beta)}{\partial \beta_{ij}} + \sum_{l=1}^{s}\frac{\partial^2 M_1(\beta)}{\partial \beta_{ij}\partial \beta_{1l}}(\beta_l^{**} - \beta_{1l}) + \sum_{l=1}^{p-s}\frac{\partial^2 M_1(\beta)}{\partial \beta_{ij}\partial \beta_{2l}}(\beta_l^{**} - \beta_{2l})
\end{aligned}
$$

$$+(N-nq)p'_{\lambda_1}(|\beta_j|)\mathrm{sgn}(\beta_j),$$

where β^{**} lies between β^* and β. By (21), $\frac{\partial M_1(\beta)}{\partial\beta} = O_p(n^{1/2})$. In view of condition (C2), $\frac{\partial^2 M_1(\beta)}{\partial\beta\partial\beta^\tau} = X^\tau P_{z^\tau} X = O(n)$ follows. If $\beta^* - \beta = O_p(n^{-1/2})$ and $N = m_1 O(n) + q$, we have

$$\frac{\partial S_1(\beta)}{\partial\beta_{ij}} = n\lambda_1\{\lambda_1^{-1} m_1 p'_{\lambda_1}(|\beta_{ij}|)\mathrm{sgn}(\beta_{ij}) + O_p(n^{-1/2}/\lambda_1)\}.$$

In view of

$$\liminf_{n\to\infty}\liminf_{\beta\to 0+} p'_{\lambda_1}(\beta)/\lambda_1 > 0$$

and $n^{-1/2}/\lambda_1 \to 0$, (21) and (22) follow. Thus $\hat\beta_2 = 0$ holds.

Now we prove part (b). It follows from Theorem 1 and the above proof of this theorem that there exists a root-n consistent local minimizer $\hat\beta_1$ such that

$$\left.\frac{\partial S_1(\beta)}{\partial\beta_{1j}}\right|_{\beta=(\hat\beta_{1\mathrm{obpe}}^\tau, 0^\tau)^\tau} = 0 \quad \text{for} \quad j = 1,\ldots,s.$$

Note that $\hat\beta_1$ is consistent, and

$$\begin{aligned}
&\left.\frac{\partial M_1(\beta)}{\partial\beta_{1j}}\right|_{\beta=(\hat\beta_{1\mathrm{obpe}}^\tau, 0^\tau)^\tau} + (N-nq)p'_{\lambda_1}(|\hat\beta_{1\mathrm{jobpe}}|)\mathrm{sgn}(\hat\beta_{1\mathrm{jobpe}})\\
&= \frac{\partial M_{11}(\beta_1)}{\partial\beta_{1j}} + \sum_{l=1}^{s}\frac{\partial^2 M_{11}(\beta_1)}{\partial\beta_{1j}\partial\beta_{1l}}(\hat\beta_{1\mathrm{lobpe}} - \beta_{1l})\\
&\quad + (N-nq)\Big[p'_{\lambda_1}(|\beta_{1j}|)\mathrm{sgn}(\beta_{1j}) + (p''_{\lambda_1}(|\beta_{1j}|) + o_p(1))(\hat\beta_{1\mathrm{jobpe}} - \beta_{1j})\Big],
\end{aligned}$$

where $M_{11}(\beta_1) = (Y - X_1\beta_1)^\tau P_{z^\tau}(Y - X_1\beta_1)$. Then $\frac{\partial M_{11}(\beta_1)}{\partial\beta_1} = -X_1^\tau P_{z^\tau}(Y - X_1\beta_1)$. Similar to (21), it is easy to verify

$$\frac{1}{\sqrt{n}}\frac{\partial M_{11}(\beta_1)}{\partial\beta} \xrightarrow{L} \mathcal{N}(0, \sigma^2\Sigma_{11}) \quad \text{as} \quad n\to\infty.$$

It follows from the Slutsky theorem that this theorem is proved. □

Proof of Theorem 3 Let $\alpha_{n2} = n^{-1/2} + b_n(\lambda_2)$. Similar to the proof of Theorem 1, it is sufficient to show that, for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|c\|=C} S_2(\theta + \alpha_{n2}c) > S_2(\theta)\right\} \geq 1 - \epsilon.$$

Define $M_2(\theta) = \frac{1}{2}(\tilde{Y}(\hat{\sigma}^2_{obe}) - U\theta)^\tau W^{-1}(\tilde{Y}(\hat{\sigma}^2_{obe}) - U\theta)$. In each iterative minimizing problem (13), the unknown parameter θ in W is replaced by the estimate obtained in the previous step in iteration. By computing the first derivative on both sides of the above equation, we have

$$-\frac{\partial M_2(\theta)}{\partial \theta} = \sum_{i=1}^{n} u_i^\tau W_i^{-1} \mathrm{vec}\Big((z_i b_i + \varepsilon_i)(z_i b_i + \varepsilon_i)^\tau - (z_i D z_i^\tau + \sigma^2 I_{l_i})\Big) + o_p(n^{-1/2}).$$

Define $\xi_i = V_{i0}^{-1/2}(z_i b_i + \varepsilon_i)$ which has mean zero, and covariance I_{l_i}. It follows that

$$nVar(I_1) = \frac{1}{n}\sum_{i=1}^{n} u_i^\tau W_i^{-1/2} \mathrm{vec}(\xi_i\xi_i^\tau - I_{l_i})\mathrm{vec}^\tau(\xi_i\xi_i^\tau - I_{l_i}) W_i^{-1/2} u_i \tag{23}$$

$$= \sum_{i=1}^{n} u_i^\tau W_i^{-1/2}\Big(E(\xi_i\xi_i^\tau \otimes \xi_i\xi_i^\tau) - \mathrm{vec}(I_{l_i})\mathrm{vec}^\tau(I_{l_i})\Big) W_i^{-1/2} u_i.$$

Under conditions $(C1) - (C6)$ and Lindeberg–Feller Central Limit theorem, we have

$$\frac{\partial M_{22}(\theta)}{\partial \theta} \xrightarrow{L} \mathcal{N}(0, \Psi) \quad \text{as} \quad n \to \infty.$$

Similar to the proof of Theorem 1, the proof is concluded. □

Proof of Theorem 4 Similar to the proof of Theorem 2, one can easily finish this proof, we then omit the details here.

References

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

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

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

Gentle, J. E. (1998). *Numerical linear algebra for applications in statistics*. Berlin: Springer.

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

Jiang, J. M., Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *The Indian Journal of Statistics*, *65*, 23–42.

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

Peng, H., Lu, Y. (2012). Models selection in linear mixed effect models. *Journal of Multivariate Analysis*, *109*, 109–129.

Pu, W. J., Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis*, *97*, 733–758.

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

Wu, P., Zhu, L. X. (2010). An orthogonality-based estimation of moments for linear mixed models. *Scandinavian Journal of Statistics*, *37*, 253–263.

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