Journal of Multivariate Analysis 109 (2012) 109–129

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Model selection in linear mixed effect models[☆]

Heng Peng[a], Ying Lu[b,*]

[a] *Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong*

[b] *New York University, Department of Humanities and Social Sciences, Steinhardt School of Culture, Education and Human Development, United States*

ARTICLE INFO

Article history:
Received 20 October 2010
Available online 1 March 2012

AMS subject classifications:
62H12
62J05
62J07

Keywords:
Model selection
SCAD function
Penalized least squares
Oracle property
Group selection

ABSTRACT

Mixed effect models are fundamental tools for the analysis of longitudinal data, panel data and cross-sectional data. They are widely used by various fields of social sciences, medical and biological sciences. However, the complex nature of these models has made variable selection and parameter estimation a challenging problem. In this paper, we propose a simple iterative procedure that estimates and selects fixed and random effects for linear mixed models. In particular, we propose to utilize the partial consistency property of the random effect coefficients and select groups of random effects simultaneously via a data-oriented penalty function (the smoothly clipped absolute deviation penalty function). We show that the proposed method is a consistent variable selection procedure and possesses some oracle properties. Simulation studies and a real data analysis are also conducted to empirically examine the performance of this procedure.

1. Introduction

Clustered data is a common phenomenon in modern data analysis. For example, in social surveys, individual respondents are often clustered under city blocks, neighborhoods or other geographical regions. Another example can be seen in longitudinal studies where repeated measurements on the same subject are taken over time. Mixed effect models are widely used statistical tools to deal with clustered data (see for examples, Goldstein [10], Bryk and Raudenbush [4]). In this paper, we aim to study the problem of variable selection and parameter estimation for linear mixed effect models.

In mixed effect models, it is assumed that the unobserved heterogeneity at cluster level causes intra-cluster correlation between the responses, and hence the mean level of the responses and/or the effects of the covariates can vary across clusters. Fixed effects and random effects are used to model such intra-cluster correlation. The key difference between fixed effect and random effect is that the former assumes that unobserved heterogeneity at cluster level is constant while the latter assumes that such quantity is random. Hence the estimation of the fixed effects concerns the actual sizes of the cluster-specific effects. When the number of clusters is large, the number of fixed effect coefficients increases rapidly. Conversely, researchers are more interested in the distribution of the random effects rather than the actual sizes of random effect coefficients. Random effects are often assumed to follow a zero-mean multivariate normal distribution, and its covariance matrix becomes our key interest since it summarizes the intra-cluster correlation. When the number of the random effect components is large, the estimation of random effects in a mixed effect model involves a high dimensional covariance matrix

[☆] Our thanks to Jianqing Fan, Wenyang Zhang, the attendants at 2009 Joint Statistical Meeting at Washington DC for their valuable suggestions and comments. This paper is partially supported by National Science Foundation of grant DMS-0631652, the CERG grant of Hong Kong Research Grant Council, HKBU 201707, HKBU 201610 and HKBU 201809 and FRG grants of Hong Kong Baptist University, FRG/08-09/II-33 and FRG/10-11/024 and a developmental grant from the NICHD-funded University of Colorado Population Center (grant R21 HD51146). This paper has also appeared as a Chapter in Ying Lu's Ph.D. thesis at the Department of Statistics at University of North Carolina at Chapel Hill.

[*] Corresponding author.
E-mail addresses: hpeng@math.hkbu.edu.hk (H. Peng), yl46@nyu.edu, ying.lu@nyu.edu (Y. Lu).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.02.005

that can greatly increase computational instability. Therefore, identifying the effective components of random effects is very important for applied researchers to build more interpretable models and to ease the computational burden.

Traditionally, variable selection for mixed effect models has relied on p value-based stepwise deletion, or more elaborately, the Akaike's information criterion [1], the FPE_λ method [19], and Mallow's C_p [14]. However, these procedures ignore stochastic errors inherited through the process of variable selection. The estimators based on these variable selection procedures suffer from lack of stability and it is hard to understand their theoretical properties [3]. Alternatively, the Bayesian information criterion [18] and Generalized information criterion [15,17] are used as consistent variable selection procedures for fixed effect parameters. But according to Pu and Niu [16], these procedures perform poorly in selecting random effect components. In addition, all of these variable selection procedures involve a combinatorial optimization problem which is *NP*-hard with computational time increasing exponentially with the number of parameters. (see comments in Fan and Li [6]). Hence it is not feasible to apply these procedures to the complete set of the candidate models when the number of parameters is large.

To address the weakness of traditional variable selection procedures, recent work has focused on selecting variables simultaneously with model estimation using data oriented penalty functions. For examples, the bridge regression [9], the least absolute shrinkage and selection operator (LASSO) [21], Tibshirani (1997), and the smoothly clipped absolute deviation penalty (SCAD) [7]. The SCAD penalty function has some oracle properties in that the estimators based on it converge to the true model while other alternative penalty functions are only shrinkage estimators. Fan and Peng [8] further established the asymptotic properties of the SCAD penalty function when the number of parameters increases with sample size.

When random effects are not subject to selection, the penalty method for the variable selection problem in linear mixed effect is straightforward. One can use a penalized likelihood estimation approach. When random effects are subject to selection, the problem becomes more complicated as the estimation of the covariance matrix involves a constrained optimization problem that is close to or on the boundary of the parameter space. In this situation, for most optimization procedures such as Newton–Raphson and the EM algorithm, the convergence can be slow and often fails. Only recently, Krishna [12] developed a restricted EM algorithm that uses the adaptive LASSO [25] to estimate and select linear mixed model under the penalized likelihood framework.

In this paper, we aim to develop an optimization-free variable selection procedure for linear mixed effect models. To ease the burden of computation, we adopt a simple iterative procedure that takes advantage of the partial consistency property of random effects. Then we extend this approach to select effective random effect components by penalizing random effect coefficients in groups. Antoniadis and Fan [2] pointed out that selecting variables based on the information of a group of variables will lead to better thresholding decision rules and faster convergence. As our simulation and theoretical results will show, this procedure selects both fixed effect and random effects consistently, and gives unbiased estimates. As sample size becomes large, the procedure has some oracle properties. Although our analysis is limited to linear mixed effect models, it provides important insights to generalized linear mixed effect models.

The rest of this paper is organized as follows: In Section 2 we present a simple iterative procedure that can effectively estimate the linear mixed effect model without burdensome optimization. In Section 3, we adapt this procedure to select random effect and fixed effect components simultaneously during estimation. Simulation results and an example of data analysis will be presented in Section 4. The paper ends with a discussion and future research directions.

2. A distribution-free procedure to estimate linear mixed effect model

To avoid constrained optimization problem, we hereby propose to select variables and estimate parameters for linear mixed effect models based on a simple iterative procedure. We first describe how we can use this procedure to estimate linear mixed effect models and the proposed estimators can achieve satisfactory sampling properties under mild conditions. Then we extend this procedure so that it also selects the effective components of fixed effects and random effects during model estimation.

Consider the linear mixed effect model (LMM) that was originally introduced by Laird and Ware [13]. For each cluster i,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad i = 1, \ldots, m, \tag{2.1}$$

where $\mathbf{Y}_i$ is a vector of dependent variables of length n_i, and the elements of Y_i are assumed to be independent across clusters, but correlated within the cluster. $\mathbf{X}_i$ is a given n_i by p matrix of covariates whose effects are assumed to fixed, $\boldsymbol{\beta}$ is a $p \times 1$ vector of corresponding fixed effect coefficients. To simplify the notations, we allow $\mathbf{X}$ to include both the traditional sensed covariates whose effects are constant across clusters and the cluster-specific fixed effects. $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{in_i})^T$ is a given n_i by q matrix of covariates whose effects are assumed to be random across clusters and $\mathbf{b}_i$ is a $q \times 1$ vector of random effect coefficients. $\boldsymbol{\epsilon}_i$ is a vector of random errors of length n_i that is independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{b}_i$. Typically, a linear mixed model makes the following distributional assumptions,

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2\mathbf{I}_{n_i}),$$
$$\mathbf{b}_i \sim \mathcal{N}(0, \sigma^2\mathbf{D}),$$
$$\mathbf{Y}_i \sim \mathcal{N}(X_i\boldsymbol{\beta}, \sigma^2\mathbf{V}_i),$$

where $\mathbf{D}$ is a $q \times q$ nonnegative definite matrix, and $\mathbf{V}_i = \mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$. However, it worth noting that the estimating procedure that we adopt to compute linear mixed effect model does not rely on the normality assumptions for both $\boldsymbol{\epsilon}_i$ and $\mathbf{b}_i$.

2.1. An iterative procedure to estimate LME

Inspired by Sun et al. [20], we consider the following two-step iterative procedure to estimate fixed and random effects. We start with initial values $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 = [\sum_i (\mathbf{X}_i^T \mathbf{X}_i)]^{-1} [\sum_i \mathbf{X}_i^T \mathbf{Y}_i]$.

Step 1: predict the residuals given $\hat{\boldsymbol{\beta}}$ for group i,

$$\mathbf{u}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}},$$

for $i = 1, \dots, m$ we can estimate

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{u}_i, \tag{2.2}$$

and residual $\boldsymbol{\gamma}_i = \mathbf{u}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i$. Based on $\boldsymbol{\gamma}_i$ and $\hat{\mathbf{b}}_i$, we propose an estimator of σ^2,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m \boldsymbol{\gamma}_i^T \boldsymbol{\gamma}_i}{(n - qm)}, \tag{2.3}$$

and an estimator of $\mathbf{D}$,

$$\hat{\mathbf{D}} = \frac{\sum_{i=1}^m \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T}{m \hat{\sigma}^2} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}}{m}. \tag{2.4}$$

The first term of (2.4) appears to be a naïve estimator of $\mathbf{D}$. But if we look closely,

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{u}_i = \mathbf{b}_i + (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \boldsymbol{\epsilon}_i.$$

This leads to

$$\sum_{i=1}^m \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T = \sum_{i=1}^m \mathbf{b}_i \mathbf{b}_i^T + \sum_{i=1}^m (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} + \sum_{i=1}^m (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \boldsymbol{\epsilon}_i \mathbf{b}_i^T + \sum_{i=1}^m \mathbf{b}_i \boldsymbol{\epsilon}_i^T \mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}.$$

As Sun et al. [20] pointed out, the last two terms are of order $Op(m^{1/2})$, hence

$$\begin{aligned} m^{-1} \sum_{i=1}^m \mathbf{b}_i \mathbf{b}_i^T &\approx m^{-1} \left\{ \sum_{i=1}^m \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T - \sum_{i=1}^m (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \mathbf{Z}_i (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \right\} \\ &\approx m^{-1} \left\{ \sum_{i=1}^m \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T - \sum_{i=1}^m \sigma^2 (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \right\}. \end{aligned}$$

Substituting σ^2 by $\hat{\sigma}^2$, we obtain the estimator of $\mathbf{D}$ in (2.4).

Step 2: given $\hat{\mathbf{D}}$, now we can estimate $\hat{\boldsymbol{\beta}}$ based on generalized least squares.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}, \tag{2.5}$$

where $\mathbf{W}$ is a block diagonal matrix with diagonal elements $(\mathbf{I}_{n_i} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1}$, $i = 1, \dots, m$, and $\mathbf{y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_m^T)^T$.

To achieve numerically stable estimates of $\hat{\sigma}^2$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{D}}$, we can iterate between steps 1 and 2 until convergence.

2.2. Asymptotic properties

The estimators we proposed in Section 2.1 have been mentioned in several papers in various contexts (for examples, Sun et al. [20], Demidenko [5].) In this section, we systematically show that the estimators of β, D are $\sqrt{n}$-consistent. To make the presentation clearer, we introduce the following notations and regular conditions,

Notations.

$$c_1 = \lim_{m \to \infty} \frac{n}{n - qm} \tag{2.6}$$

$$c_2 = \lim_{m \to \infty} \frac{n}{m} \tag{2.7}$$

$$\Gamma = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^m E\left[(\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \right], \tag{2.8}$$

$$\Delta_2 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} E\left[\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\otimes\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\right] \tag{2.9}$$

$$\Delta_3 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E\left[\mathrm{vec}\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\mathrm{vec}^T\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\right]$$

$$\Delta_4 = \mathrm{vec}\left\{\mathbf{D}+\frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\mathrm{vec}\left\{\mathbf{D}+\frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}^T \tag{2.10}$$

$$\gamma = \lim_{m\to\infty}(n-qm)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E\left[Z_{ij}^T(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}Z_{ij}\right]^2 - c_1 q/c_2 + 1 \tag{2.11}$$

and

$$\Delta_1 = \begin{pmatrix} \mathbf{D}\otimes\Gamma_{(1)} + \Gamma\otimes\mathbf{D}_{(1)} \\ \vdots \\ \mathbf{D}\otimes\Gamma_{(q)} + \Gamma\otimes\mathbf{D}_{(q)} \end{pmatrix}$$

where $\Gamma_{(r)}, \mathbf{D}_{(r)} (r = 1, \dots, q)$ denote the rth row of $\mathbf{D}$, Γ, respectively, and vec($\mathbf{M}$) denotes the vector by simply stacking the column vectors of the matrix $\mathbf{M}$ below one another. Obviously there exists a unique $q^2 \times q(q+1)/2$ matrix R_q such that $\mathrm{vec}(\mathbf{A}) = R_q\mathrm{vech}(\mathbf{A})$. vech($\mathbf{A}$) denotes the vector consisting of all elements on and below the diagonal of the matrix,

Regular conditions:

(A) The errors $\epsilon_{ij}, i = 1, \dots, m, j = 1, \dots, n_i$, are i.i.d. and $E(\epsilon_{11}^4) < \infty$. The random effects $\mathbf{b}_i, i = 1, \dots, m$ are i.i.d. and $E\|\mathbf{b}_1\|^4 < \infty$, where $\|\mathbf{b}_1\| = (\mathbf{b}_1^T\mathbf{b}_1)^{1/2}$. $Ex_{ilj}^{2k} < \infty$ and $Ez_{ilj}^{2k} < \infty$, where x_{ilj} denotes the (l, j) element of $\mathbf{X}_i$ and z_{ils} denotes the (l, s) element of $\mathbf{Z}_i$ for $k > 2, i = 1, \dots, m, l = 1, 2 \dots, n_i, j = 1, \dots, p$ and $s = 1, \dots, q$.
(B) The absolute value of elements of $\mathbf{Z}_i, i = 1, \dots, m$ are uniformly bounded by a positive constant.
(C) The minimum eigenvalue of $\mathbf{Z}_i^T\mathbf{Z}_i$ and $\mathbf{X}_i^T\mathbf{X}_i, i = 1, \dots, m$ are uniformly larger than a positive constant. Δ_2 is not a singular matrix. $\Sigma = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$ is a positive definite matrix.
(D) c_1 and γ are positive values, and Γ is a positive definite matrix. The size of each cluster, $n_i, i = 1, \dots, m$ is bounded by a positive constant which are larger than 1. So $1 < c_2 = \lim_{m\to\infty}\frac{n}{m} \le C$.

Comment: Condition (A) contains a set of mild conditions for the linear mixed model without any distributional assumption about normality. The bounded moments about fixed effects, random effects and error terms are sufficient for our proposed estimates to follow asymptotic normal distribution in general situations. Conditions (B) and (C) are used to guarantee the positive definiteness of the design matrix or the weighted design matrix so that the weighted least squares estimates for the linear mixed model have a positive definite asymptotic covariance matrix. Condition (D) is also reasonable in practice. Similar to conditions (B) and (C), it removes certain singular situations of the linear mixed model. Specially the conditions on c_1 and c_2 are used to guarantee that every cluster provides useful information to estimate the covariance matrix of the random effects. In fact, in the balance design, it is just required that the size of every cluster should be larger than the number of random effects.

Under these mild conditions above, we have the following results.

Lemma 2.1. *Under the regularity conditions* (A)–(D),

$$n^{1/2}\left\{\hat\sigma^2 - \sigma^2\right\} \xrightarrow{D} \mathcal{N}(0, 2\sigma^4(1+\gamma)c_1 + \mathrm{Var}(\epsilon_{11}^2)\gamma c_1).$$

Proposition 2.1. *Under the regularity conditions* (A)–(D), *given a $\sqrt{n}$-consistent estimator $\hat{\mathbf{D}}$ of* $\mathbf{D}$, *for the generalized least square estimator of $\boldsymbol\beta$,*

$$\hat{\boldsymbol\beta} = \underset{\beta}{\mathrm{argmin}}\sum_{i=1}^{m}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol\beta)^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol\beta),$$

we have

$$\sqrt{n}(\hat{\boldsymbol\beta} - \boldsymbol\beta) \xrightarrow{D} \mathcal{N}(0, \Sigma_\beta),$$

where Σ_β is a positive definite matrix and

$$\Sigma_\beta = \lim_{m\to\infty}\sigma^2\left(\frac{1}{n}\sum_{i=1}^{m}\mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1}.$$

Proposition 2.2. *Under the regularity conditions* (A)–(D), *given a* $\sqrt{n}$*-consistent estimate of* β, *if* ϵ_{ij}, $i = 1, \ldots, m, j = 1, \ldots, n_i$, *are i.i.d and follow* $\mathcal{N}(0, \sigma^2)$, *for the estimate of* $\mathbf{D}$ *by* (2.4), *we have*

$$\sqrt{n}\left\{\text{vec}(\hat{\mathbf{D}} - \mathbf{D})\right\} \xrightarrow{D} \mathcal{N}(0, (R_q^T R_q)^{-1} R_q^T \Delta R_q (R_q^T R_q)^{-1} c_2),$$

where

$$\Delta = \frac{1}{\sigma^4} E\left\{\mathbf{b}_1 \mathbf{b}_1^T \otimes \mathbf{b}_1 \mathbf{b}_1^T\right\} - \text{vec}(\mathbf{D})\text{vec}^T(\mathbf{D}) + \frac{1}{\sigma^2}\{\mathbf{D} \otimes \Gamma + \Gamma \otimes \mathbf{D} + \Delta_1\}$$
$$+2\left\{\Delta_2 - \Delta_3 + \frac{c_1}{c_2}(1+\gamma)\Delta_4\right\} + \frac{\text{var}(\epsilon_{11}^2)}{\sigma^4}\left\{\Delta_3 + \frac{c_1}{c_2}\gamma\Delta_4\right\}.$$

Comment: In Proposition 2.2, we add a normal assumption for error ϵ_{ij} in order to compute the close form of the asymptotic variance of the estimate $\hat{\mathbf{D}}$. Though without such assumption, we can still follow the same steps to show the asymptotic normality of the estimate $\hat{\mathbf{D}}$ without a close form of the variance of the estimate. Obviously a close form of the asymptotic variance of the estimate facilitates future statistical inference based on the estimates. By nature, as shown in the proofs of Propositions 2.1 and 2.2, this iterative procedure does not really rely on the normal assumption for the linear mixed model, and the proposed estimate can be regarded as a distribution-free estimate for the linear mixed model. It is also easy to compute since it does not involve complex optimization problems. Here some efficiency is traded for such advantages. In particular, if the cluster size n_i is small and close to q, the size of the random effects, $\mathbf{Z}_i^T \mathbf{Z}_i$ will tend to be singular and the constant c_1 will go to infinity. The standard errors of $\hat{\mathbf{b}}_i$ are going to be large and this will lead to inaccuracy in the estimation of $\mathbf{D}$.

3. Selecting effective fixed and random effects components

Since the procedure we proposed in Section 2.1 is an optimization-free one, it enjoys great computational stability even when the covariance matrix $\mathbf{D}$ is near singular. In this section, we consider selecting the effective components of fixed and random effects in linear mixed effect model via the penalty function SCAD.

Now suppose in model (2.1), some components of $\boldsymbol{\beta}$ are zero, and some random effects are zero such that the corresponding diagonal elements of $\mathbf{D}$ are zero. Without loss of generality, we write

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$$

where $\boldsymbol{\beta}_{20} = \mathbf{0}$, and $\text{diag}(\mathbf{D}_0) = (\mathbf{d}_{10}^T, \mathbf{d}_{20}^T)^T$, where $\mathbf{d}_{20} = \mathbf{0}$, and corresponding rows and columns of $\mathbf{D}_0$ are zero as well.

To simultaneously select the nonzero components of fixed effects and random effects during the estimation, we adjust the two-step estimating procedures in Section 2 such that the small fixed effect coefficients will be shrunk to zero and the effective dimension of $\mathbf{D}$ will be correctly identified.

3.1. An iterative procedure to select and estimate LME

Step 1: First observe that if the kth random effects component is effectively absent, then the (k, k) diagonal element of $\mathbf{D}$ is zero, so are the elements of the corresponding kth row and kth column since the correlation between the kth random effect and other components of random effects is zero. Hence we expect the estimate $\hat{\mathbf{D}}$ as given in (2.4) will be close to zero as well. Using this fact, we consider shrinking the corresponding random effect coefficients $\mathbf{b}_{ik}$, $i = 1, \ldots, m$ to zero if their variance is estimated to be sufficiently close to zero.

We propose to estimate $\mathbf{b}_i$, for each i, $i = 1, \ldots, m$, by minimizing the following penalized least squares

$$\frac{1}{2}(\mathbf{u}_i - \mathbf{b}_i \mathbf{Z}_i)^T (\mathbf{u}_i - \mathbf{b}_i \mathbf{Z}_i) + \sum_{k=1}^{q} n p_\xi(c_k), \tag{3.12}$$

where $c_k = \sqrt{|D_{kk}^*|}$, D_{kk}^* is the kth diagonal element of the estimate $\mathbf{D}^*$ of the covariance matrix $\mathbf{D}$,

$$\mathbf{D}^* = \frac{\sum_{i=1}^m \mathbf{b}_i \mathbf{b}_i^T}{m\hat{\sigma}^2} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}}{m}.$$

$p_\xi(\theta)$ is the smoothly clipped absolute deviation penalty function (SCAD) by Fan and Li [7], where

$$p'_\xi(\theta) = \xi\left\{I(\theta \le \xi) + \frac{(a\xi - \theta)_+}{(a-1)\xi} I(\theta > \xi)\right\}, \tag{3.13}$$

for some $a > 2$ and $\theta > 0$. ξ is the tuning parameter for penalty function $p(\theta)$. As Fan and Li pointed out, the SCAD function is singular at the origin and does not have a continuous secondary derivative. To solve this penalized least squares, one can locally approximate the penalty function by its quadratic function when $c_k \neq 0$ and given a initial value c_{k0}, $\mathbf{b}_{i0} = (b_{i10}, \dots, b_{iq0})^T$ and $\mathbf{D}_0^*$,

$$[p_\xi(c_k)]' \approx \left\{p'_\xi(c_{k0})/c_{k0}\right\} c_k, \quad \text{for } c_{k0} \approx c_k.$$

In other words,

$$p_\xi(c_k) \approx p_\xi(c_{k0}) + \frac{1}{2}\left\{p'_\xi(c_{k0})/c_{k0}\right\}(c_k^2 - c_{k0}^2).$$

Consequently, the solution to (3.12) can be updated based on the following ridge regression

$$\mathbf{b}_i^* = (\mathbf{Z}_i^T\mathbf{Z}_i + n\Sigma_\xi(\mathbf{c_0}))^{-1}\mathbf{Z}_i^T\mathbf{u}_i, \quad i = 1, \dots, m. \tag{3.14}$$

where $\mathbf{c}_0 = \text{diag}(c_{10}, \dots, c_{q0})$ and $\Sigma_\xi = \text{diag}\left(\frac{p'_\xi(c_{10})b_{i10}\text{sign}(D^*_{110})}{mc_{10}}, \dots, \frac{p'_\xi(c_{q0})b_{iq0}\text{sign}(D^*_{qq0})}{mc_{q0}}\right)$.

An estimator of $\mathbf{D}^*$ can be updated as,

$$\mathbf{D}^* = \frac{\sum_{i=1}^m \mathbf{b}_i^*\mathbf{b}_i^{*T}}{m\hat{\sigma}^2} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}}{m}. \tag{3.15}$$

When the variance of the kth random effect, c_k, is estimated to be small, we expect the solution b_{ik}^* in (3.14) will shrink to zero. This is true for all $i = 1, \dots, m$. The corresponding diagonal elements and the corresponding rows and columns of $\mathbf{D}^*$ in (3.15) will then be estimated to be zero.

Although the diagonal elements of $\mathbf{D}$ are always nonnegative, it is not always the case for its estimate $\mathbf{D}^*$. When some of the diagonal elements of $\mathbf{D}^*$ are negative, we choose to directly set those elements zero. For example, if $D_{qq}^* < 0$, then we set $D_{qi}^*, D_{iq}^*, i = 1, \dots, q$ as zero, and update c_q as 0 and shrink $b_{iq}^*, i = 1, \dots, m$ to zero. Intuited by Proposition 2 that $\sqrt{n}\{\text{vec}(\mathbf{D}^* - \mathbf{D})\}$ is asymptotic normally distributed, this is a reasonable treatment. When D_{qq} is greater than zero, the estimate D_{qq}^* should be positive and away from zero with large probability when sample size n is large enough. Only when $D_{qq} = 0$, there exists nonzero probability that D_{qq}^* would be negative.

Comment: Our approach of shrinking a group of random effect coefficients together is closely related to the block-wise penalized functions that were discussed in [2]. In particular, the block-wise penalized least squares problem takes the following form,

$$\|Z - \theta\|^2 + p_\lambda(\|\theta\|).$$

They argue that whenever applicable, to shrink the coefficients in groups will make the thresholding decision more accurate and improve the convergence rate since the information within a group is bigger. This idea is also seen in more recent work such as group LASSO [24].

Step 2: The selection of fixed effects given random effect variance estimates $\hat{D}$ is simply the solution to the penalized weighted least squares

$$\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{k=1}^p p_\lambda(|\beta_k|). \tag{3.16}$$

Similarly, we can use a local quadratic approximation of $p_\lambda(|\beta_k|)$ and update $\boldsymbol{\beta}^*$ based on the following ridge regression,

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{W}\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0))^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{3.17}$$

where λ is the tuning parameter for the penalty function, and

$$\Sigma_\lambda(\boldsymbol{\beta}_0) = \text{diag}\left[p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \dots, p'_\lambda(|\beta_{p0}|)/|\beta_{p0}|\right].$$

To achieve numerically stable estimates of $\boldsymbol{\beta}^*$ and $\mathbf{D}^*$, we can iterate between steps 1 and 2 until convergence.

3.2. Asymptotic properties

We can show that the estimators of $\mathbf{D}$ and $\boldsymbol{\beta}$ given in (3.15) and (3.17) are consistent and have some oracle properties.

Theorem 3.1. *Under the regularity conditions* (A)–(D), *given a $\sqrt{n}$-consistent estimate $\mathbf{D}^*$ of $\mathbf{D}$, if $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$ as $n \to \infty$, then there is a local minimizer $\boldsymbol{\beta}^*$ of* (3.16) *such that*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| = O_p(1/\sqrt{n}),$$

and this minimizer must satisfy

(a) *Sparsity:* $\boldsymbol{\beta}_2^* = 0$.

(b) *Asymptotic normality:*

$$\sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{01}) \xrightarrow{D} \mathcal{N}(0, \Sigma_{\beta_{01}})$$

where

$$\Sigma_{\beta_{01}} = \lim_{m\to\infty} \sigma^2 \left(\frac{1}{n} \sum_{i=1}^{m} \mathbf{X}_{i1}^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D}^* \mathbf{Z}_i^T)^{-1} \mathbf{X}_{i1} \right)^{-1}.$$

Theorem 3.2. *Under the regularity conditions* (A)–(D), *given a* $\sqrt{n}$*-consistent estimate of* β, *if* $\sqrt{n/\log(n)}\xi \to O(1)$ *as* $n \to \infty$, *then there is a local minimizer* $\mathbf{b}_i^*$, $i = 1, \ldots, m$ *of* (3.14) *such that for the estimate of* $\mathbf{D}^*$ *by* (3.15) *we have*

(a) *Sparsity:* $\mathbf{d}_2^* = 0$.

(b) *Asymptotic Normality: if* ϵ_{ij}, $i = 1, \ldots, m, j = 1, \ldots, n_i$, *are i.i.d and follow* $\mathcal{N}(0, \sigma^2)$, *then*

$$\sqrt{n}\left\{\text{vech}(\mathbf{D}_1^* - \mathbf{D}_{01})\right\} \xrightarrow{D} \mathcal{N}(0, (R_q^T R_q)^{-1} R_q^T \Delta R_q (R_q^T R_q)^{-1} c_2)$$

where $\mathbf{Z}_i, \mathbf{b}_i, q$ *are replaced by* $\mathbf{Z}_{i1}, \mathbf{b}_{i1}$ *and* q_1 *in* (2.6)–(2.11) *and the definition of* R_q, *and the definition of* Δ *is same as its definition in* Proposition 2.2.

3.3. *Tuning parameter selection and thresholding*

To implement the variable selection procedure in Section 3.1, we need to consider the choice of tuning parameter λ and ξ. Theoretically, we need $\lambda \to 0$, $\xi \to 0$ and $\sqrt{n}\lambda \to \infty$, $\sqrt{n}\xi \to \infty$, as $n \to \infty$ in order to consistently select fixed and random effects. In practice, the tuning parameter can be selected based on data oriented methods. Following Fan and Li [7], Wang et al. [23], we consider the following three criteria,

1. generalized cross-validation criterion

$$\text{GCV}_\lambda = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_{\mathbf{W}}^2}{n(1 - \text{Df}_\lambda/n)},$$

2. the AIC criterion

$$\text{AIC}_\lambda = \log \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_{\mathbf{W}}^2 + 2\text{Df}_\lambda/n,$$

3. the BIC criterion

$$\text{BIC}_\lambda = \log \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_{\mathbf{W}}^2 + \text{Df}_\lambda \log(n)/n$$

where $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_{\mathbf{W}}^2$ is the model error for linear mixed model, and $\mathbf{W}$ is a block diagonal matrix with diagonal elements $(\mathbf{I}_{n_i} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}$, $i = 1, \ldots, m$. Wang et al. [23] argued that the BIC criterion is an optimal and consistent tuning-parameter-selection procedure for linear regression, while GCV and AIC criteria tend to overfit the model. We expect this argument holds for our application as well.

The degree of freedom is difficult to determine in linear mixed effect model. Here we adopt Hodges & Sargent's formula [11] to calculate the degree of freedom. We write model (2.1) by adding a block of "pseudo data"

$$\mathbf{Y}^* = \mathbf{U}\boldsymbol{\delta} + \beta,$$

where

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ 0_{qm} \end{pmatrix}, \quad \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ 0 & -\Delta \end{pmatrix}, \quad \beta = \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{b} \end{pmatrix},$$

where $\Delta^T\Delta = G^{-1}$, and G is a $qm \times qm$ block diagonal matrix with diagonal element $\mathbf{D}$. Based on this, a quasi "Hat" matrix H_1 can be defined for linear mixed model,

$$H_1 = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \end{pmatrix} (\mathbf{U}^T\mathbf{U})^{-1} \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix}.$$

The effective degree of freedom is then trace(H_1).

4. Simulation studies and real data analysis

In this section, we conduct a set of simulation studies to assess the performance of the proposed variable selection and estimation procedure for linear mixed effect model. A real data analysis will also be conducted. We are particularly interested in model performance in the following aspects: whether the correct subsets of fixed effects and random effects can be

Table 1
Performance of fixed and random effect selection. "FPR%" is the average false positive rate which is defined as the percentage of the coefficients that are incorrectly estimated to be nonzero. "FNR%" is the average false negative rate that is the percentage of the coefficients that are incorrectly estimated to be zero. "Model size" reports the average size of nonzero fixed effect coefficients and nonzero random effect components.

Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
	Example 1			Example 2		
Fixed effects						
BIC	21.5	9.9	2.26	1.5	1.9	2.10
AIC	17	11.0	2.43	1.5	3.3	2.20
GCV	20.5	10.1	2.30	1.5	3	2.18
$\sqrt{\log(n)/n}$	21	15.6	2.67	1.5	4.1	2.26
Random effects						
BIC	27	6	2.25	0	0	3
AIC	25	12	2.37	0	0	3
GCV	26	6	2.28	0	0	3
$\sqrt{\log(n)/n}$	33	7	2.09	0	0	3

correctly selected; whether the parameter estimates are unbiased and efficient in small to medium sample sizes; and when the true models are ascertained, whether the iterative method has comparable sample properties to maximum likelihood method.

4.1. Simulation I

In the first set of simulations, we adopt the examples in [12]. There are two scenarios:

1. Example 1: Consider $m = 30$ subjects, $n_i = 5$ observations per subject. There are 9 fixed effects to be considered, and the true value of coefficients are $\boldsymbol{\beta} = [0, 1, 1, 0, 0, 0, 0, 0, 0]$. For random effects, we consider 4 dimensions, with the true covariance matrix
$$D = \begin{pmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$
The model variance $\sigma^2 = 1$. Furthermore, covariates $\mathbf{X}$ are generated from a uniform $(-2, 2)$ distribution, along with a vector of $\mathbf{1}$'s for the subject-specific intercept. The values of $\mathbf{Z}$ are taken to be the values of the first four columns of $\mathbf{X}$.
2. Example 2: The set up of the second example is the same as the first, except the number of the subjects increases to $m = 60$ and the number of the observations per subject increases to $n_i = 10$.

For each example, we randomly draw 200 samples and apply the proposed variable selection and estimation procedure to these data sets. In Table 1, we summarize the performance of the proposed iterative procedure under different tuning parameters. We notice that as sample size grows, the procedure selects the correct fixed and random effect components with increasing accuracy. Due to the benefit of group selection, the selected random effect components quickly converge to the true model. For different tuning parameter choices, we can see that in general the BIC criterion outperforms others choices. Compared to other criteria, the BIC-based false positive rate and the false negative rate are smaller and the average model size is closer to that of the true model as well.

We also compare the percentage of the models that are correctly identified by our procedure in comparison with Krishna's Table 3.1. In Table 2, We can see as the sample size increases, the performance of our method improves dramatically. With a sample size of 600, random effect selection has nearly 100% accuracy and the fixed effect selection (using BIC as tuning parameter criterion) outperforms all the other existing approaches. As a matter of fact, the overall model selection performance is partly impacted by the simulation setup. The first random effect is assumed a large variance of 9 which causes large uncertainty in estimating the random effect coefficients $\mathbf{b}_1$ and in turn affects the estimation and selection of the first fixed effect coefficient in our iterative procedure. A detailed look of the fixed effect selection reveals that our method successfully selects all but the first fixed effect in the 200 simulations we conducted.

On the other hand, our model performs relatively unsatisfactorily with low sample size, especially when the number of observations per group is low compared to the number of the random effect coefficients. As we commented earlier in Section 2, this is an inherent problem due to the two-step estimation procedure. With smaller cluster size, the large standard errors of $\mathbf{b}_i$ pose challenges for model selection. But our simulation also shows that this problem can be quickly remedied if the number of clusters increases.

4.2. Simulation II

In the second simulation study, we focus on the performance of parameter estimates of our proposed methods. We consider the following four different scenarios: the number of clusters is either 10 or 20, and the number of observations

Table 2
Comparing the model selection performance of the proposed iterative method with other existing methods. "%Correct" reports the percentage of times the correct true model was selected, "%CF" and "%RF" report the percentage of the times correct fixed effect components and random effect components are selected. The results for M-ALASSO, EGIC, RIC, Stepwise deletion and ALASSO are borrowed from Krishna [12].

Method	Tuning	%Correct	%CF	%CR	%Correct	%CF	%CR
		Example 1			Example 2		
Iterative method	BIC	19	49	35	86	86	100
Iterative method	AIC	21	46	35	77	77	100
Iterative method	GCV	20	49	37	79	79	100
Iterative method	$\sqrt{\log(n)/n}$	16	33	27	72	72	100
M-ALASSO	BIC	71	73	79	83	83	89
EGIC	BIC	47	56	52	48	59	53
RIC	AIC	19	21	62	31	34	74
RIC	BIC	59	59	68	77	79	81
Stepwise	AIC	17	21	62	26	28	74
Stepwise	BIC	51	53	68	68	69	81
ALASSO	AIC	21	24	62	39	41	74
ALASSO	BIC	62	63	68	74	75	81

Table 3
Numbers of fixed effects and random effects that are selected to be zero in 100 simulated data sets. For β_2, β_5, D_1 and D_3, the table reports the number of parameters that are correctly selected to be zero. For β_1, β_3, β_4, D_2 and D_4, the table reports the number that are incorrectly selected to be zero.

Sample size		Fixed effects					Random effects			
m	n_i	β_1	β_2	β_3	β_4	β_5	D_1	D_2	D_3	D_4
10	10	0	92	1	11	94	100	39	100	5
20	10	1	98	1	0	98	100	8	100	0
10	20	1	96	0	1	95	100	9	100	1
20	20	1	100	0	0	99	100	0	100	0

within each cluster is either 10 or 20. For each scenario, we assume the common model structure to be as follows: the dimension of fixed effects is $p = 5$ with true value $\boldsymbol{\beta} = [1, 0, 1.5, 1, 0]$. The dimension of the random effects is $q = 4$ with the covariance matrix of the random effect coefficients D,

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.354 \\ 0 & 0 & 0 & 0 \\ 0 & 0.354 & 0 & 1 \end{pmatrix}$$

so that only the second and the fourth random effect components are significant. Furthermore, the correlation between the second and the fourth random effects is 0.5. The model variance σ^2 is assumed to be 1. Without loss of generality, the components of $\mathbf{X}$ are generated from standard normal distributions, and $\mathbf{Z}$ assumes the same values as $\mathbf{X}_1, \ldots, \mathbf{X}_4$.

For each scenario, we simulate 100 data sets and run the iterative variable selection and estimation procedure for each data set. For the purpose of comparison, we estimate three different models. First we apply the proposed iterative penalized method to select and estimate both fixed effects and random effects simultaneously. Then we estimate the model using the iterative procedure and the maximum likelihood estimation assuming the true model is known.

The number of correctly and incorrectly selected fixed effects and random effects among the 100 simulated data sets are reported in Table 3. We can clearly see that as both the number of clusters and the number of within cluster observations increase, the fixed effects and random effects are selected with increasingly high accuracy.

Next we examine the performance of parameter estimation of our proposed models. For each simulation set up, we present bias and median absolute deviation of the nonzero fixed effect and random effect parameters in Table 4. These summary statistics demonstrate that the parameter estimators based on our proposed iterative procedure possess satisfying sampling properties. For both fixed effects and random effects, the estimators are unbiased and behave as if the true model is known when sample size is large. Moreover, we can see that when the true model is known, our proposed iterative procedure performs equally well as the maximum likelihood estimation.

4.3. *Real data analysis*

In this section, we apply the proposed method to the 2000 American National Election Study. The ANES is a series of surveys on voters' opinions before and after each election since 1948. The outcome variable we are interested in is the feeling thermometer reading for George W. Bush. Feeling thermometer is a widely accepted way of quantifying individuals' feeling toward public figures. It mimics a physical thermometer and ranges from 0° to 100°. In this example, the higher temperature an individual assigns indicates he/she feels more positive toward George W. Bush, and vice versa.

The 2004 ANES is a national representative sample of 1212 respondents from 29 states in the US. In this analysis, we will examine what factors affect individuals' feeling toward Bush. Since such effects tend to be mediated by social and cultural

Table 4
Bias and median absolute deviation (MAD) of the significant fixed effect and random effect parameter estimates. "iter" refers to iterative variable selection and estimation method, "iterO" refers to iterative estimation method under the true model, and "MLEO" refers to MLE under the true model.

m	n_i	Parameter	Bias			MAD		
			Iter	IterO	MLEO	Iter	IterO	MLEO
10	10	β_1	0.006	0.012	0.011	0.073	0.065	0.067
		β_3	−0.010	−0.003	0.000	0.064	0.065	0.064
		β_4	−0.112	0.022	0.019	0.292	0.260	0.251
		D_{22}	−0.124	−0.008	−0.002	0.275	0.135	0.130
		D_{44}	0.150	0.038	−0.040	0.437	0.321	0.313
		D_{24}	−0.120	−0.019	−0.023	0.115	0.185	0.189
20	10	β_1	−0.001	0.008	0.009	0.046	0.044	0.046
		β_3	−0.006	0.002	0.002	0.046	0.046	0.046
		β_4	−0.007	−0.010	−0.008	0.141	0.140	0.130
		D_{22}	−0.010	0.013	0.026	0.128	0.125	0.155
		D_{44}	−0.004	−0.031	−0.029	0.238	0.231	0.257
		D_{24}	−0.024	−0.005	0.008	0.117	0.109	0.112
10	20	β_1	−0.005	0.004	0.004	0.047	0.046	0.046
		β_3	−0.006	−0.008	−0.008	0.053	0.051	0.047
		β_4	0.052	0.063	0.063	0.207	0.186	0.189
		D_{22}	−0.062	−0.033	−0.033	0.183	0.160	0.141
		D_{44}	−0.006	−0.023	−0.090	0.253	0.269	0.241
		D_{24}	−0.033	−0.025	−0.027	0.188	0.185	0.164
20	20	β_1	−0.010	0.001	0.001	0.034	0.033	0.033
		β_3	0.007	0.004	0.004	0.037	0.037	0.036
		β_4	0.003	0.001	0.002	0.126	0.121	0.120
		D_{22}	0.013	0.013	0.020	0.113	0.112	0.111
		D_{44}	−0.021	−0.021	−0.040	0.201	0.201	0.210
		D_{24}	−0.003	−0.002	0.005	0.119	0.118	0.110

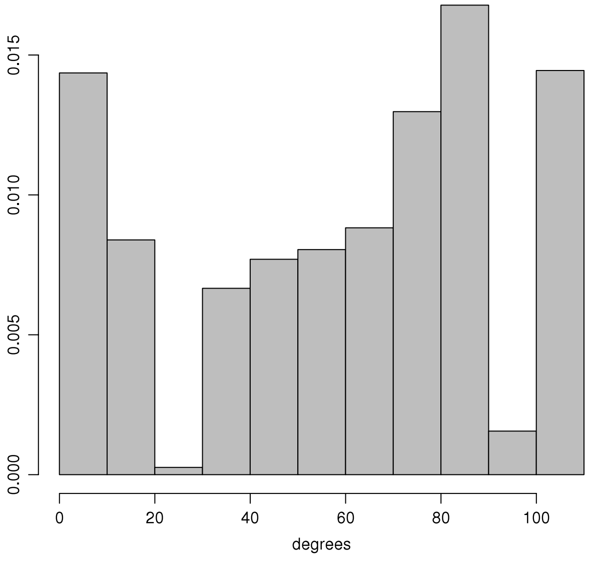

Fig. 1. Histogram of the outcome variable "Bush feeling thermometer readings".

contexts at the state level, we will further examine whether these effects vary across states. To do so, we fit a linear mixed effect model with individuals nested under states. After removing missing data and states with too few observations, the effective sample size consists of 1156 individuals from 24 states.

Fig. 1 shows the histogram of Bush feeling thermometer readings. We can see that there is considerable amount of variation and the distribution appears to be bimodal rather than normally distributed. Like many other social and behavioral studies, there exists a large amount of individual level heterogeneity that cannot be easily captured via systematic modeling. Our results reveal that the model variance σ^2 is rather large compared to the amount of variance systematically explained by fixed effects and random effects (the intra-cluster correlation is only 18%). Since there could be a wide arrays of factors influencing individuals' preference toward political figures, we start with a linear mixed effect model with a large number of fixed and random effects (see Table 5).

Table 5
The complete lists of the candidate fixed effect and random effect components. A detailed description of these variables is given in the Appendix.

Fixed effects	Random effects
Intercept, age, gender, education, income, Christian, black, other, gun control, liberal view, moderate view, defense issues, abortion right, death penalty, environment issues, social trust, church attendance, health insurance, Democrat, Independent, Iraq war	Intercept, gender, income, Christian, gun control, liberal view, moderate view, defense issues, abortion right, death penalty, health insurance, Democrat, Independent, Iraq war

Table 6
Parameter estimation of the fixed effect coefficients and the random effect covariance. The first three columns report the coefficients, standard errors and p-values estimated based on the iterative variable selection and estimation procedure. The last three columns report the corresponding estimates based on R package `nlme` under the model that is selected via iterative procedure.

Method	Iterative method			`nlme` package		
Fixed effects ($\boldsymbol{\beta}$)	Coefficient	s.e.	p value	Coefficient	s.e.	$p$value
(Intercept)	48.57	3.46	0.00	46.62	3.34	0.00
Age	0.06	0.04	0.12	0.09	0.04	0.02
Education	−4.96	1.36	0.00	−5.23	1.33	0.00
Christian	7.32	1.65	0.00	7.62	1.76	0.00
Black	−3.97	2.17	0.06	−2.52	2.01	0.21
Other	3.6	2.03	0.07	4.59	1.97	0.02
Liberal	−11.97	1.77	0.00	−11.68	1.75	0.00
Defense	2.39	1.35	0.07	1.95	1.31	0.14
Death penalty	4.17	1.44	0.00	4.62	1.43	0.00
Democrat	−24.48	2.08	0.00	−25.05	2.06	0.00
Independent	−14.21	1.73	0.00	−14.53	1.71	0.00
Iraq war	30.47	1.61	0.00	30.62	1.58	0.00
Random effects (D)						
Gender	52.32			55.38		
Christian	25.91			15.44		
Covariance	−19.65			−28.27		
Model variance (σ^2)						
	438.02			442.26		

In this data analysis, large model variance and a number of potentially nuisance random effect components can pose great challenge to maximum likelihood based approaches in estimating and selecting the correct submodel. In our attempts to fit the model using commercial software such as the `xtmixed` package in STATA and the `nlme` package in R, the initial full model and its many submodels fail to converge. To tackle this problem, we apply the proposed procedure in Section 3 to estimate the model while shrinking the insignificant fixed and random effects to zero simultaneously. We use general cross validation method to determine the values of the tuning parameters for selecting fixed effects and random effects via the SCAD function. The results are presented in Table 6. Among the 14 random effects listed in Table 5, only two are deemed to be effective random effect components. For comparison purposes, we also fit the same model using R package `nlme` that is based on RMLE. We can see that in general our model yields estimates that are close to the R package. However, as mentioned before, since the outcome variable Bush feeling thermometer appears to be bimodally distributed, our approach is expected to provide more robust results than the MLEs that are based on the assumption of normality.

5. Discussion

In this paper, we present a simple iterative penalized procedure that selects and estimates fixed effects and random effects simultaneously. The theoretical and simulation investigations of the proposed procedure have shown that it selects the correct submodel effectively and has some oracle properties. Although in mixed effect model it is well known that the random effect coefficients cannot be consistently estimated, we have demonstrated that the covariance of the random effect coefficients can be consistently estimated. We can further take advantage of the partial consistency property to select the effective component of random effects by penalizing the random effect coefficients in group. If the corresponding variance term is sufficiently small, the entire group of random effect coefficients will be shrunk to zero via penalized least squares.

Our method is based on the estimation of the random effect coefficients. The cost of relying on the estimation of the random effect coefficients is that we need sufficient number of observation within each cluster. When the cluster size is small relative to the dimension of random effects, our method does not perform as well as the likelihood based approaches that only concern the marginal distribution of the data. However in survey data analysis, the size of the clusters is typically large, so we expect this method offers a practical solution to many real data analysis problems. On the other hand, when the dimension of all candidate random effects is large compared to the cluster size, as long as the dimension of significant random effects is relatively small, the proposed variable selection method can still select and estimate the significant covariance matrix due to the oracle property of the proposed method and the ridge regression procedure. However, when the dimension of random effects is larger than the cluster size, our proposed penalized method will depend on the initial estimate of the covariance matrix, since the penalty function is non-concave and there can be more than one extreme value for the penalized least square function. In this case, we can consider using convex penalty function such as the L_1 penalty to estimate better initial values, then apply our method to achieve unbiased estimates.

In general, the proposed method enjoys many advantages over the classical likelihood based approaches. Compared to the classical likelihood approach, this procedure has greater computational stability since it avoids the complicated constrained optimization problem of estimating a high dimensional covariance matrix that is located at the boundary of the parameter space due to the inclusion of non-existing random effects. Since our method does not rely on multivariate normal distribution of the data, it is expected to be robust under model misspecification. In particular, we can further relax step 2 of the iterative procedure: instead of using (penalized) weighted least squared that takes the normal covariance structure of the error terms into account, we can simply use (penalized) ordinary least squares to calculate $\boldsymbol{\beta}$ based on $\mathbf{Y}_{0i} = \mathbf{Y}_i - \mathbf{Z}_i\mathbf{b}_i$.

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(|\boldsymbol{\beta}_0|))^{-1}\mathbf{X}^T\mathbf{Y}_0. \tag{5.18}$$

Based on simulation evidence (not shown in this paper), this distribution-free version of the iterative procedure can also select fixed effects and random effects satisfactorily. Although this procedure is less efficient when the errors are known to be normally distributed, it is more robust if the model is misspecified as it does not depend on particular information of the error structure.

Lastly, this method can be easily adapted to estimate multiple levels of hierarchical structure. To select and estimate fixed and random effects at multiple levels, we can simply condition on the random effect coefficients at lower level and partial consistency property will ensure the validity of this approach.

Appendix A. Proofs

In this section, we outline the detailed proofs of the asymptotic results in previous sections.

Proof of Lemma 2.1. It can be deduced directly from Theorem 2 of Sun et al. [20]. □

Proof of Proposition 2.1. Following the definition of $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{Y}_i,$$

we define

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}\mathbf{Y}_i.$$

Based on linear model theory, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \Sigma_\beta).$$

Hence to prove Proposition 2.1, by Slutsky's lemma (see Van der Vaart [22, p. 11]), we only need to prove that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) = o_p(1). \tag{A.1}$$

To show (A.1), we rewrite $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*$ as

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^* &= \left(\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{Y}_i \\ &\quad - \left(\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}\mathbf{Y}_i \\ &= I_1 + I_2\end{aligned}$$

where

$$I_1 = \left(\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\right)^{-1} \left\{\sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}(\mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i) - \sum_{i=1}^m \mathbf{X}_i^T(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)^{-1}(\mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i)\right\}$$

and

$$I_2 = \left\{ \left(\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \right)^{-1} - \left(\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \right)^{-1} \right\} \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i).$$

First notice that since $E\mathbf{b}_i\mathbf{b}_i^T = \sigma^2\mathbf{D}$ and $E\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \sigma^2\mathbf{I}_{n_i}$, by Central Limit Theorem and the regularity conditions (A) and (C), we can show that

$$\sum_{i=1}^m \mathbf{X}_i^T \mathbf{Z}_i \mathbf{Z}_i^T (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i) = O_p(\sqrt{n})$$

and

$$\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i) = O_p(\sqrt{n}).$$

Now let us consider I_1. Because $\hat{\mathbf{D}} - \mathbf{D} = O_p(1/\sqrt{n})$,

$$\begin{aligned} &\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i) - \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i) \\ &\quad = O_p(1/\sqrt{n}) \left| \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Z}_i \mathbf{Z}_i^T (\mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i) \right| \\ &\quad = O_p(1/\sqrt{n}) O_p(\sqrt{n}) = O_p(1). \end{aligned} \tag{A.2}$$

Next because the elements of $\mathbf{Z}_i$ are bounded and $\hat{\mathbf{D}} - \mathbf{D} = O_p(1/\sqrt{n})$, the eigenvalue of $\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T$, $i = 1, \dots, m$ should be bounded by 1 and a positive constant $1 + C$ where C is a positive constant determined by $\mathbf{Z}_i$, $i = 1, \dots, m$ and $\mathbf{D}$. Hence we have that

$$\sum_{i=1}^m \mathbf{X}_i^T \mathbf{X}_i \geq \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \geq (1 + C)^{-1} \sum_{i=1}^m \mathbf{X}_i^T \mathbf{X}_i.$$

Then by condition (C) that $\Sigma = 1/m \sum_{i=1}^m \mathbf{X}_i^T \mathbf{X}_i$ is a positive definite matrix, we just know that

$$\Sigma_\beta = \lim_{m \to \infty} \sigma^2 \left(\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \right)^{-1}$$

should be also a positive definite matrix. We also have

$$\begin{aligned} O_p(1/n) &= \left\{ \sum_{i=1}^m \mathbf{X}_i^T \mathbf{X}_i \right\}^{-1} \leq \left\{ \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \right\}^{-1} \\ &\leq (1 + C) \left\{ \sum_{i=1}^m \mathbf{X}_i^T \mathbf{X}_i \right\}^{-1} = O_p(1/n). \end{aligned} \tag{A.3}$$

By (A.2) and (A.3), we have that

$$I_1 = O_p(1/n) = o_p(1/\sqrt{n}). \tag{A.4}$$

Next consider I_2. First notice that

$$\begin{aligned} \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i - \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i &= O_p(1/\sqrt{n}) \sum_{i=1}^m \left| \mathbf{X}_i^T \mathbf{Z}_i \mathbf{Z}_i^T \mathbf{X}_i \right| \\ &= O_p(1/\sqrt{n}) O_p(n) \\ &= O_p(\sqrt{n}) \end{aligned}$$

and

$$\sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i = O_p(n), \qquad \sum_{i=1}^m \mathbf{X}_i^T (\mathbf{I} + \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i = O_p(n).$$

Hence

$$I_2 = \frac{O_p(\sqrt{n})}{O_p(n)\cdot O_p(n)} \cdot O_p(\sqrt{n}) = O_p(1/n) = o_p(1/\sqrt{n}). \tag{A.5}$$

Finally, by (A.4) and (A.5), (A.1) is obtained and the proof of Proposition 2.1 is complete. □

Proof of Proposition 2.2. If $\boldsymbol{\beta}$ is known, then we have

$$\mathbf{u}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} \quad \text{and} \quad \tilde{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i,$$

and an estimate of $\mathbf{D}$ is given

$$\tilde{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} \tag{A.6}$$

where $\hat{\sigma}^2$ is an estimate of σ^2 defined by the following (A.7).

Let $\hat{\boldsymbol{\beta}}$ is the $\sqrt{n}$ consistent estimate of $\boldsymbol{\beta}$, we can define

$$\hat{\mathbf{u}}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i,$$

and an estimate of σ^2 can be defined as

$$\hat{\sigma}^2 = (n-qm)^{-1}\sum_{i=1}^m \mathrm{RSS}_i, \quad \mathrm{RSS}_i = \hat{\mathbf{u}}_i^T(\mathbf{I}_{n_i} - \mathbf{P}_i)\hat{\mathbf{u}}_i \tag{A.7}$$

where $\mathbf{P}_i = \mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T$ is the project matrix of $\mathbf{Z}_i$.

By Lemma 2.1, it is known that $\hat{\sigma}^2$ is a $\sqrt{n}$-consistent estimate of σ^2. Then $\mathbf{D}$ can be estimated as

$$\hat{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}. \tag{A.8}$$

Next we first prove that

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = o_p(1/\sqrt{n}).$$

Then we only need study the asymptotic distribution of $\sqrt{n}\tilde{\mathbf{D}}$.

By (A.6) and (A.8), we have

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T. \tag{A.9}$$

Because

$$\begin{aligned}\hat{\mathbf{u}}_i &= \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}}\\ &= \mathbf{u}_i + \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{b}}_i &= (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i\\ &= (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\\ &\hat{=} \tilde{\mathbf{b}}_i + \mathbf{e}_i.\end{aligned}$$

Then

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = -\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \{\tilde{\mathbf{b}}_i\mathbf{e}_i^T + \mathbf{e}_i\tilde{\mathbf{b}}_i^T\} - \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \mathbf{e}_i\mathbf{e}_i^T.$$

Since $\mathbf{u}_i = \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$, it can be shown that

$$\tilde{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{u}_i = \mathbf{b}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\boldsymbol{\epsilon}_i.$$

Hence by the regularity conditions and the definition of $\mathbf{b}_i$, it can be shown that

$$\begin{aligned}\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \tilde{\mathbf{b}}_i\mathbf{e}_i^T &= \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m(\mathbf{b}_i + (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i)(\boldsymbol{\beta}^T - \hat{\boldsymbol{\beta}}^T)\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\\ &\le O_p(1/\sqrt{n})\left(\left|\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \mathbf{b}_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right| + \left|\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right|\right)\\ &= O_p(1/\sqrt{n})O_p(1/\sqrt{m}) = O_p(1/n). \end{aligned} \tag{A.10}$$

Similarly, we also have

$$\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \mathbf{e}_i\tilde{\mathbf{b}}_i^T = O_p(1/n). \tag{A.11}$$

On the other hand, because $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(1/\sqrt{n})$, we have

$$\begin{aligned}\frac{1}{m\sigma^2}\sum_{i=1}^m \mathbf{e}_i\mathbf{e}_i^T &= \frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\\ &= O_p(1/n)\left|\frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right|\\ &= O_p(1/n)O_p(1) = O_p(1/n). \end{aligned} \tag{A.12}$$

So by (A.10)–(A.12), it is obtained that

$$\tilde{\mathbf{D}} - \hat{\mathbf{D}} = O_p(1/n). \tag{A.13}$$

For $\tilde{\mathbf{D}}$, this can be written as

$$\tilde{\mathbf{D}} = \frac{1}{m\sigma^2}\sum_{i=1}^m \tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T - \frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \left\{\frac{1}{m\hat{\sigma}^2} - \frac{1}{m\sigma^2}\right\}\sum_{i=1}^m \tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T \,\hat{=}\, D_1 + D_2.$$

For D_1, we have

$$\begin{aligned} D_1 &= \frac{1}{m\sigma^2}\sum_{i=1}^m \mathbf{b}_i\mathbf{b}_i^T + \left\{\frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} - \frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\\ &\quad + \frac{1}{m\sigma^2}\sum_{i=1}^m\left\{(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\epsilon_i\mathbf{b}_i^T + \mathbf{b}_i\epsilon_i^T\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\\ &\hat{=}\ D_{11} + D_{12} + D_{13}. \end{aligned}$$

Because of the independence of $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, m$, D_{11}, D_{12} and D_{13} are linear independent and $ED_{12} = ED_{13} = 0$, and we have

$$ED_1 = ED_{11} \quad \text{and} \quad \mathrm{Var}(\mathrm{vec}(D_1)) = \mathrm{Var}(\mathrm{vec}(D_{11})) + \mathrm{Var}(\mathrm{vec}(D_{12})) + \mathrm{Var}(\mathrm{vec}(D_{13})).$$

By simple calculation, it can be shown that

$$ED_{11} = \mathbf{D} \quad \text{and} \quad \mathrm{Var}(\sqrt{m}\cdot\mathrm{vec}(D_{11})) = \frac{1}{\sigma^4}E\{\mathbf{b}_1\mathbf{b}_1^T \otimes \mathbf{b}_1\mathbf{b}_1^T\} - \mathbf{D}\otimes\mathbf{D}. \tag{A.14}$$

For D_{12}, it can be written as

$$\begin{aligned} &\frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} - \frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\\ &= \frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T(\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i - \sigma^2\mathbf{I}_{n_i})\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\\ &= \frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{A}\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \frac{1}{m\sigma^2}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{B}\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\\ &\hat{=} D_{121} + D_{122} \end{aligned}$$

where

$$\mathbf{A} = \mathrm{diag}(\epsilon_{i1}^2 - \sigma^2, \ldots, \epsilon_{in_i}^2 - \sigma^2)$$

and

$$B_{jl} = \epsilon_{ij}\epsilon_{il}, \quad j \neq l \text{ and } B_{jj} = 0, j, l = 1, \ldots, n_i.$$

It can be shown that the element of **A** and **B** are linear independent, and then because $\boldsymbol{\epsilon}_i$ and $\mathbf{Z}_i$, $i = 1, \ldots, m$ are independent, we have.

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{12})) = \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{121})) + \mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{122})).$$

Next some complex calculation, we obtain that

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{121})) = \mathrm{Var}(\epsilon_{11}^2)\Delta_3/\sigma^4, \tag{A.15}$$

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{122})) = 2(\Delta_2 - \Delta_3). \tag{A.16}$$

To D_{13}, by similar complicate calculation, we have

$$\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_{13})) = \frac{1}{\sigma^2}\{\mathbf{D} \otimes \Gamma + \Gamma \times \mathbf{D} + \Delta_1\}. \tag{A.17}$$

Next consider D_2. By Lemma 2.1, $\hat{\sigma}^2 - \sigma^2 = O_p(1/\sqrt{n})$, hence

$$D_2 = \left\{\frac{1}{m\hat{\sigma}^2} - \frac{1}{m\sigma^2}\right\}\sum_{i=1}^{m}\tilde{\mathbf{b}}_i\tilde{\mathbf{b}}_i^T = \frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\left(\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right) + o_p(1/\sqrt{n}), \tag{A.18}$$

and

$$\begin{aligned}
&\mathrm{Var}(\sqrt{m} \cdot \mathrm{vec}(D_2)) \\
&\quad = \left(2(1+\gamma)\frac{c_1}{c_2} + \frac{\mathrm{Var}(\epsilon_{11}^2)}{\sigma^4}\frac{\gamma c_1}{c_2}\right)\mathrm{vec}\left\{\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}\mathrm{vec}\left\{\mathbf{D} + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\right\}^T \\
&\quad = \left(2(1+\gamma)\frac{c_1}{c_2} + \frac{\mathrm{Var}(\epsilon_{11}^2)}{\sigma^4}\frac{\gamma c_1}{c_2}\right)\Delta_4.
\end{aligned} \tag{A.19}$$

Notice from the definition of $\hat{\sigma}^2$ and $\hat{\sigma}^2 - \sigma^2 \to 0$ as $n \to \infty$, and hence $ED_2 \to 0$ and

$$E(\tilde{\mathbf{D}}) \to E(D_1) \quad \text{as } n \to \infty. \tag{A.20}$$

According to the condition that $\boldsymbol{\epsilon}_i$ follows the normal distribution $\mathcal{N}(0, \sigma^2\mathbf{I}_{n_i})$, so $\mathbf{Z}_i^T\boldsymbol{\epsilon}_i$ and $(\mathbf{I}_{n_i} - \mathbf{P}_i)\boldsymbol{\epsilon}_i$ are independent, $i = 1, 2, \ldots, m$.

Notice (A.7), the structure of the estimate σ^2, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(1/\sqrt{n})$ and $\hat{\mathbf{u}} = \mathbf{u}_i + \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, by some calculation we have

$$\hat{\sigma}^2 = (n - qm)^{-1}\sum_{i=1}^{m}\mathrm{RSS}_i = (n - qm)^{-1}\sum_{i=1}^{m}\mathrm{RSS}_i^* + O_p\left(\frac{1}{n}\right)$$

where $\mathrm{RSS}_i^* = \boldsymbol{\epsilon}_i^T(\mathbf{I}_{n_i} - \mathbf{P}_i)\boldsymbol{\epsilon}$.

Then by the structure of D_{11}, D_{12}, D_{13} and combined with (A.18), it can be shown that D_1 and D_2 are asymptotic independent and Hence

$$\mathrm{Var}(\mathrm{vec}(\tilde{\mathbf{D}})) = \mathrm{Var}(\mathrm{vec}(D_1)) + \mathrm{Var}(\mathrm{vec}(D_2)) + o(1/n). \tag{A.21}$$

Finally, by (A.13)–(A.21), the transformation from vec to vec$h$ by R_q, and the central limited theory, the proof of Proposition 2.2 is complete. □

Proof of Theorem 3.1. Define

$$\begin{aligned}
Q(\boldsymbol{\beta}) &= \sum_{i=1}^{m}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T(\mathbf{I} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i')^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + n\sum_{j=1}^{p}p_{\lambda_n}(|\beta_j|) \\
&= L(\boldsymbol{\beta}) + n\sum_{j=1}^{p}p_{\lambda_n}(|\beta_j|)
\end{aligned}$$

and $\alpha_n = (1/\sqrt{n})$.

To prove the theorem, we first show that for any give $\varepsilon > 0$, there exist a large constant C such that

$$P\left\{\min_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} > 1-\varepsilon. \tag{A.22}$$

This implies with probability at least $1-\varepsilon$ that there exist local minimizer in the ball $\{\boldsymbol{\beta}_0+\alpha_n\mathbf{u}\colon \|\mathbf{u}\| \le C\}$. Hence there exists a local minimizer such that $\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0\| = O_p(\alpha_n)$.

Define $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ where the elements of $\boldsymbol{\beta}_{10}$ are nonzero components of $\boldsymbol{\beta}_0$, and the elements of $\boldsymbol{\beta}_{20}$ are zero components of $\boldsymbol{\beta}_0$. In the other means that $\boldsymbol{\beta}_{10}^T$ denotes the coefficients of significant predictors in the model and $\boldsymbol{\beta}_{10}^T$ denotes those coefficients of insignificant predictors in the model. Since $p_{\lambda_n}(0) = 0$. we have

$$D_n(\mathbf{u}) = Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{u}) - Q(\boldsymbol{\beta}_0) \ge L(\boldsymbol{\beta}_0+\alpha_n\mathbf{u}) - L(\boldsymbol{\beta}_0) + n\sum_{j=1}^{s}\left\{p_{\lambda_n}(|\beta_{j0}+\alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\right\}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. By simple calculation, we have

$$\begin{aligned} D_n(\mathbf{u}) \ge{}& \alpha_n^2\sum_{i=1}^{m}\mathbf{u}^T\mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\mathbf{u} - \alpha_n\sum_{i=1}^{m}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i)^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}\mathbf{X}_i\mathbf{u} \\ & - \alpha_n\sum_{i=1}^{m}\mathbf{u}^T\mathbf{X}_i^T(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1}(\mathbf{Z}_i\mathbf{b}_i+\epsilon_i) + n\sum_{j=1}^{s}\left\{p_{\lambda_n}(|\boldsymbol{\beta}_{j0}+\alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\right\} \\ \hat{=}{}& D_{n1}+D_{n2}+D_{n3}+D_{n4}. \end{aligned}$$

Under the regularity condition (C), we have

$$D_{n1} \ge O_p(n\alpha_n^2)\|\mathbf{u}\|^2, \tag{A.23}$$

and

$$D_{n2} = O_p(\alpha_n\sqrt{n})\|\mathbf{u}\| \quad \text{and} \quad D_{n3} = O_p(\alpha_n\sqrt{n})\|\mathbf{u}\|. \tag{A.24}$$

After taking the second order Taylor expansion of the first term around β_{j0} in D_{n4} we have

$$D_{n4} = n\sum_{j=1}^{s}p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\alpha_n u_j + n\sum_{j=1}^{s}p''_{\lambda_n}(|\beta_{j0}|)\alpha_n^2u_j^2\cdot(1+o(1)).$$

By the definition of the SCAD function, as $\lambda_n \to 0$, $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|)\colon \beta_{j0} \ne 0\} = 0$ and $\max(p''_{\lambda_n}(|\beta_{j0}|)) \to 0$. Hence when n is large enough, $D_{n4} = 0$. Moreover it is obvious that D_{n1} dominates D_{n2} and D_{n3}, therefore (A.22) holds. In other words, there is a local minimizer $\hat{\boldsymbol{\beta}}$ of (3.17) such that

$$\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\| = O_p(\alpha_n).$$

Next we show this minimizer $\hat{\boldsymbol{\beta}}$ has properties of (a) and (b). In fact if (a) is true, by the oracle properties of SCAD penalty, we know that the asymptotic normality of $\hat{\boldsymbol{\beta}}$ can be directly deduced from Proposition 2.1. Hence we only need to show that $\hat{\boldsymbol{\beta}}$ has the property (a).

For $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = O(1/\sqrt{n})$ and $\beta_{j0} = 0, j = s+1,\ldots,p$, we consider the derivative of $Q(\boldsymbol{\beta})$ with respect to β_j,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j} &= -\sum_{i=1}^{m}2\mathbf{X}_{ij}^T(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1} + np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= Q_1+Q_2. \end{aligned}$$

Based on the definition of SCAD penalty function, $p'_{\lambda_n}(|\beta_j|) = \lambda_n$ when $\beta_j = o(1/\sqrt{n})$ and $\sqrt{n}\lambda_n \to \infty$. Hence $Q_2 = n\lambda_n\mathrm{sgn}(\beta_j)$ when n is large enough.

Under the regularity condition (C), we know that

$$\sum_{i=1}^{m}2\mathbf{X}_{ij}^T(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{I}+\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)^{-1} = O_p(\sqrt{n}).$$

Since $\sqrt{n}\lambda_n \to \infty$ when $n \to \infty$, Q_1 is dominated by Q_2 and Q_2 determines the sign of the derivative above. This means that for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1,\ldots,p$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j} > 0 \quad \text{for } 0 < \beta_j < \varepsilon_n \tag{A.25}$$

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for } 0 > \beta_j > -\varepsilon_n. \tag{A.26}$$

Therefore, only when $\beta_j = 0, j = s+1, \ldots, p$, $Q(\boldsymbol{\beta})$ arrives its minimizer point.

Hereby we finish the proof of this theorem. □

Proof of Theorem 3.2. Similar as the proof of Proposition 2.2, we only need to show there exists a local minimizer $\mathbf{D}^*$ such that

$$\|\mathbf{D} - \mathbf{D}^*\| = O_p\left(\sqrt{\log n/n}\right), \quad \text{and} \quad \mathbf{d}_2^* = 0.$$

The asymptotic normality of $\mathbf{D}_1^*$ follows by the properties of the SCAD penalty function.

To show $\|\mathbf{D} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$, it suffices to show $\|\hat{\mathbf{D}} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$ since we showed in Proposition 2.2 that $\|\hat{\mathbf{D}} - \mathbf{D}\| = O_p(\sqrt{1/n})$. Moreover, since

$$\mathbf{D}^* = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} \mathbf{b}_i^* \mathbf{b}_i^{*T} - \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}$$

and

$$\hat{\mathbf{D}} = \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^{T} - \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}$$

we only need to show $\|\hat{\mathbf{B}} - \mathbf{B}^*\| = O_p(\sqrt{\log(n)/n})$, where $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_m^T)^T$.

First define $\hat{\mathbf{u}}_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$ and

$$Q(\mathbf{B}) = \sum_{i=1}^{m} (\hat{\mathbf{u}}_i - \mathbf{Z}_{\mathbf{i}} \mathbf{b}_i)^T (\hat{\mathbf{u}}_i - \mathbf{Z}_{\mathbf{i}} \mathbf{b}_i) + \sum_{i=1}^{q} n p_{\xi}(c_k),$$

where

$$c_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} b_{ik}^2 - \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)_{kk}^{-1} \right|^{\frac{1}{2}}.$$

To prove the theorem, we need to show that for any give $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \min_{\|\mathbf{v}\|=C} Q(\hat{\mathbf{B}} + \alpha_n \mathbf{v}) > Q(\hat{\mathbf{B}}) \right\} > 1 - \varepsilon \tag{A.27}$$

where $\alpha_n = \sqrt{\log n/n}$, $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1^T, \ldots, \hat{\mathbf{b}}_m^T)^T$ and $\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \hat{\mathbf{u}}_i$ is defined as in the proof of Proposition 2.2. This implies with probability at least $1 - \varepsilon$ that there exists a local minimizer such that $\|\mathbf{B}^* - \hat{\mathbf{B}}\| = O_p(\sqrt{\log n/n})$ and therefore $\|\hat{\mathbf{D}} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$. Then by Theorem 2.2, $\|\mathbf{D} - \mathbf{D}^*\| = O_p(\sqrt{\log n/n})$.

By the definition of $\hat{\mathbf{B}}$, we have

$$\begin{aligned} Q(\hat{\mathbf{B}} + \alpha_n \mathbf{v}) - Q(\hat{\mathbf{B}}) &= \sum_{i=1}^{m} \left((\hat{\mathbf{u}}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i)^T \mathbf{Z}_i \alpha_n v_i + \alpha v_i^T \mathbf{Z}_i^T (\hat{\mathbf{u}}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i) \mathbf{Z}_i + \alpha_n^2 v_i^T \mathbf{Z}_i^T \mathbf{Z}_i v_i \right) \\ &\quad + \sum_{i=1}^{q} n(p_{\xi}(\tilde{c}_k) - p_{\xi}(\hat{c}_k)) \\ &= Q_1 + Q_2 \end{aligned}$$

where $\hat{c}_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} \hat{b}_{ik}^2 - \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)_{kk}^{-1} \right|^{\frac{1}{2}}$ and $\tilde{c}_k = \left| \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^{m} (\hat{b}_{ik} + \alpha_n v_{ik})^2 - \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)_{kk}^{-1} \right|^{\frac{1}{2}}$.

First by the definitions of $\hat{\mathbf{b}}_i$, we know that the first two terms in Q_1 are equal to 0. By the regularity conditions, the third term of Q_1 is of order $O_p(n\alpha_n^2)C^2$.

After taking the Taylor expansion of $p_{\xi}(\tilde{c}_k)$ around $\hat{c}_k$,

$$Q_2 = \sum_{i=1}^{q} n p_{\xi}'(\hat{c}_k)(\tilde{c}_k - \hat{c}_k) + \sum_{i=1}^{q} n p_{\xi}''(\hat{c}_k)(\tilde{c}_k - \hat{c}_k)^2 (1 + o(1)).$$

For Q_2, because $\sqrt{n/\log n} \cdot \xi \to \infty$ and Proposition 2.2, when n is large enough, we have $p'_\xi(\hat{c}_k)$ is bounded by ξ and

$$\max\{p''_\xi(\hat{c}_k), k = 1, \ldots, q\} \to 0.$$

On the other hand, since $\frac{1}{m}\sum_{i=1}^m \hat{\mathbf{b}}_i = O_p(1/\sqrt{n})$, by regularity conditions we know that

$$\begin{aligned}
&\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m(\hat{\mathbf{b}}_i + \alpha_n v_i)(\hat{\mathbf{b}}_i + \alpha_n v_i)^T - \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T \\
&= \frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m\left(\alpha_n\hat{\mathbf{b}}_i^T v_i + \alpha_n v_i^T\hat{\mathbf{b}}_i + \alpha_n^2 v_i^T v_i\right) + \left(\frac{1}{m\hat{\sigma}^2} - \frac{1}{m\hat{\sigma}^2}\right)\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T \\
&= O_p(\alpha_n^2) \cdot C^2.
\end{aligned}$$

Hence $\tilde{c}_k^2 - \hat{c}_k^2 = O_p(\alpha_n^2) \cdot C^2$, and $\tilde{c}_k - \hat{c}_k \le O_p(\alpha_n) \cdot C$.

$$Q_2 = O_p(n\xi\alpha_n) \cdot C + o_p(n\alpha_n^2) \cdot C^2 = O_p(n\alpha_n^2) \cdot C + o_p(n\alpha_n^2) \cdot C^2.$$

It is obvious that Q_2 is dominated by Q_1 when C is large enough and $Q(\hat{\mathbf{B}} + \alpha_n\mathbf{u}) - Q(\hat{\mathbf{B}}) > 0$. Hence it is easy to see that (A.27) has been proved.

Next we want to show that $\mathbf{d}_2^* = 0$. To simplify the analysis, assume that $D_{qq} = 0$, and $\mathbf{D}^* - \mathbf{D}_0 = O_p(\sqrt{\log n/n})$. We want to show that $c_q^* = 0$ or $D_{qq}^* = 0$ where

$$c_q^* = \sqrt{\left|\frac{1}{\hat{\sigma}^2}\sum_{i=1}^n b_{iq}^{*2} - \frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_i^T\mathbf{Z}_i)_{qq}^{-1}\right|} = \sqrt{|D_{qq}^*|}$$

is the estimate of $D_{qq}^{\frac{1}{2}}$. According to our updating procedure (see Section 3.1), the estimate of D_{qq}^* cannot be negative, otherwise c_q^* or D_{qq}^* will be updated to zero.

Next we show by contradiction that $c_q^* = 0$. First we assume that $D_{qq}^* = O_p(\sqrt{\log n/n}) > 0$ or $c_q^* = O_p(\log^{\frac{1}{4}} n/n^{\frac{1}{4}}) > 0$ since $\mathbf{D}^* - \mathbf{D}_0 = O_p(\sqrt{\log n/n})$ and $D_{qq} = 0$.

For $i = 1, \ldots, m$, we have

$$\begin{aligned}
\frac{\partial Q(\mathbf{B}^*)}{\partial b_{iq}} &= (\hat{\mathbf{u}}_i - \mathbf{Z}_i\mathbf{b}_i^*)^T\mathbf{Z}_{iq} + np'_{\xi_n}(c_q^*) \cdot \frac{b_{iq}^*}{mc_q^*\hat{\sigma}^2} \\
&= (\mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\hat{\mathbf{u}}_i - \mathbf{Z}_i\mathbf{b}_i^*)^T\mathbf{Z}_{iq} + np'_{\xi_n}(c_q^*) \cdot \frac{b_{iq}^*}{mc_q^*\hat{\sigma}^2} \\
&= (\mathbf{Z}_i\hat{\mathbf{b}}_i - \mathbf{Z}_i\mathbf{b}_i^*)^T\mathbf{Z}_{iq} + np'_{\xi_n}(c_q^*) \cdot \frac{b_{iq}^*}{mc_q^*\hat{\sigma}^2}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sum_{i=1}^m \frac{\partial Q(\mathbf{B}^*)}{\partial b_{iq}} \cdot b_{iq}^* &= \sum_{i=1}^m(\mathbf{Z}_i\hat{\mathbf{b}}_i - \mathbf{Z}_i\mathbf{b}_i^*)^T\mathbf{Z}_{iq}b_{iq}^* + np'_{\xi_n}(c_q^*) \cdot \sum_{i=1}^m \frac{b_{iq}^{*2}}{mc_q^*\widehat{\sigma^2}} \\
&\hat{=} Q_{d1} + Q_{d2}.
\end{aligned}$$

For Q_{d1}, by Cauchy inequality, n_i and Z_{ijq} are bounded by constants, therefore

$$\begin{aligned}
Q_{d1}^2 &\le \left\{\sum_{i=1}^m \|\mathbf{b}_i^* - \hat{\mathbf{b}}_i\|^2\right\}\left\{\sum_{i=1}^m \|\mathbf{Z}_i^T\mathbf{Z}_{iq}\|^2 b_{iq}^{*2}\right\} = O_p(\log n) \cdot \sum_{i=1}^m b_{iq}^{*2} \\
&= O_p(n\log n) \cdot \left\{\frac{1}{m\hat{\sigma}^2}\sum_{i=1}^m b_{iq}^{*2} - \frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_{iq}^T\mathbf{Z}_{iq})_{qq}^{-1}\right\} + O_p(n\log n) \cdot \left\{\frac{1}{m}\sum_{i=1}^m(\mathbf{Z}_{iq}^T\mathbf{Z}_{iq})_{qq}^{-1}\right\} \\
&= O_p(n\log(n)) \cdot O_p\left(\sqrt{\log(n)/n}\right) + O_p(n\log n) \\
&= O_p(n\log n). \qquad \text{(A.28)}
\end{aligned}$$

For Q_{d2}, because $\mathbf{D}^* - \mathbf{D}_0 = O_p(\sqrt{\log n/n})$, we know that $c_k^* = O_p(\sqrt{\log n/n})$, and hence

$$\begin{aligned} Q_{d2} &= np'_{\xi_n}(c_q) \cdot \frac{\text{sign}(D^*_{qq})}{mc_q} \left\{ \sum_{i=1}^{m} \frac{b_{iq}^{*2}}{\hat{\sigma}^2} - \sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}_{qq} + \sum_{i=1}^{m} (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}_{qq} \right\} \\ &= np'_{\xi_n}(c_q) \cdot c_q + np'_{\xi_n}(c_q) \cdot \frac{1}{c_q} \cdot O(1) \\ &= O_p\left(\sqrt{n \log n} \cdot (\log n/n)^{\frac{1}{4}}\right) + O_p\left(\sqrt{n \log n} \cdot (n/\log n)^{\frac{1}{4}}\right) \\ &= O_p(n^{\frac{3}{4}} \cdot \log^{\frac{1}{4}} n) > O_p\left(\sqrt{n \log n}\right). \end{aligned} \tag{A.29}$$

Since $\mathbf{B}^* = (\mathbf{b}_1^{*T}, \ldots, \mathbf{b}_m^{*T})^T$ is the minimizer point of $Q(\mathbf{B})$ and c_q^* does not equal to zero, we should have that

$$\frac{\partial Q(\mathbf{B}^*)}{\partial b_{iq}} = 0, \quad i = 1, \ldots, m$$

and therefore

$$Q_{1d} + Q_{2d} = 0$$

However, it can be easily seen that Q_{d1} is dominated by Q_{d2}, and hence $Q_{1d} + Q_{2d}$ cannot equal to zero with probability tending to one. This contradicts the assumption that $c_q^* \neq 0$ or $\mathbf{D}^*_{qq} > 0$. Hence it is a necessary condition that $c_q^* = 0$ if c_q or D^*_{qq} is a local minimizer. Therefore for the local minimizer $\mathbf{b}_i^*$, $i = 1, \ldots, m$, the sparsity property must hold. The proof of this Theorem has been finished. □

Appendix B. Variable definitions

The definitions of the covariates in the data analysis section are as follows:

1. Age: respondent's age.
2. Gender: respondent's gender. 1 =“male”, 0 =“female”.
3. Education: respondent's education. 1 =“high school and higher”, 0 =“less than high school”.
4. Income: respondent's income.
5. Christian: respondent's religious denomination. 1 =“Christian/Catholic”, 0 =“other”.
6. Black: dummy variable for respondent's race where “white” is the reference group. 1 =“black”, 0 =“not”.
7. Other: dummy variable for respondent's race. 1 =“other race”, 0 =“not”.
8. Gun control: the importance of gun control issues to the respondent. 1 =“very important”, 0 =“not very important”.
9. Liberal view: dummy variable for respondent's ideology where “conservative” is the reference group. 1 =“liberal”, 0 =“not”.
10. Moderate view: dummy variable for respondent's ideology. 1 =“moderate”, 0 =“not”.
11. Defense issues: the importance of defense spending issues to the respondent. 1 =“very important”, 0 =“not very important”.
12. Abortion rights: the respondent's attitude toward abortion rights. 1 =“abortion should be permitted under special circumstances or should always be permitted”. 0 =“abortion should never be permitted”.
13. Death penalty: respondent's attitude toward death penalty. 1 =“favor”, 0 =“oppose”.
14. environment issues: the importance of environmental issues to the respondent. 1 =“very important”, 0 =“not very important”.
15. Social trust: the extend the respondent trust the $\cdots$ 1 =“high”, 0 =“low”.
16. Church attendance: how often the respondent attend religious services. 1 =“regularly”, 0 =“not”.
17. Health insurance: does the respondent have health insurance? 1 =“yes”, 0 =“no”.
18. Democrat: dummy variable for respondent's party identification where “Republican” is the reference group. 1 =“Democrat”, 0 =“not”.
19. Independent: dummy variable for respondent's party identification. 1 =“Independent”, 0 =“not”.
20. Iraq war: Does the respondent approve the way President Bush handles Iraq war? 1 =“approve”, 0 =“disapprove”.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki (Eds.), Second Internal Symposium on Information Theory, Akadémiai Kiado, Budapest, 1973, pp. 267–281.

[2] A. Antoniadis, J. Fan, Regularized wavelet approximations (with discussion), Journal of American Statistical Association 96 (2001) 939–967.

[3] L. Breiman, Heuristics of instability and stablilization in model selection, Annals of Statistics 24 (1996) 2350–2383.

[4] Bryk, Raudenbush, Hierarchical Linear Models: Applications and Data Analysis Methods, second ed., Sage Publication, 2001.

[5] E. Demidenko, Criteria for global minimum of sum of squares in nonlinear regression, Computational Statistics and Data Analysis 51 (3) (2006) 1739–1753.
[6] J. Fan, R. Li, New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, Journal of American Statistical Association 99 (2004) 710–723.
[7] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96 456 (2001) 1348–1360.
[8] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, The Annals of Statistics 32 (2004) 928–961.
[9] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–148.
[10] H. Goldstein, Multilevel Statistical Models, third ed., A Hodder Arnold Publication, 2002.
[11] J.S. Hodges, D.J. Sargent, Counting degrees of freedom in hierarchical and other richly parameterized models, Biometrika 88 (2001) 367–379.
[12] A. Krishna, 2008, Shrinkage-Based Variable Selection Methods for Linear Regression and Mixed-Effects Models, Dissertation, North Carolina State University.
[13] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics 38 (1982) 963–974.
[14] C.L. Mallow, Some comments on C_p, Technometric 15 (1973) 661–675.
[15] R. Nishii, Asymptotic properties of criteria for selection of variables in multiple regression, Annal of Statistics 12 (1984) 758–765.
[16] W. Pu, X. Niu, Selecting mixed-effects models based on a generalized information criterion, Journal of Multivariate Analysis 97 (2006) 733–758.
[17] C.R. Rao, Y. Wu, A strongly consistent procedure for model selection in a regression problem, Biometrika 76 (1989) 369–374.
[18] G. Schwartz, Estimating the dimensions of a model, Annals of Statistics 6 (1978) 461–464.
[19] R. Shibata, Approximate efficiency of a selection procedure for the number of regression variables, Biometrika 71 (1984) 43–49.
[20] Y. Sun, W. Zhang, H. Tong, Estimation of the covariance matrix of random effects in longitudinal studies, The Annals of Statistics 35 (2007) 2795–2814.
[21] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society B 58 (1996) 267–288.
[22] A.M. Van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.
[23] H. Wang, R. Li, C.-L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007) 553–568.
[24] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, B 68 (2006) 49–67.
[25] H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429.