VARIABLE SELECTION IN LINEAR MIXED EFFECTS MODELS

BY YINGYING FAN[1] AND RUNZE LI[2]

University of Southern California and Pennsylvania State University

This paper is concerned with the selection and estimation of fixed and random effects in linear mixed effects models. We propose a class of nonconcave penalized profile likelihood methods for selecting and estimating important fixed effects. To overcome the difficulty of unknown covariance matrix of random effects, we propose to use a proxy matrix in the penalized profile likelihood. We establish conditions on the choice of the proxy matrix and show that the proposed procedure enjoys the model selection consistency where the number of fixed effects is allowed to grow exponentially with the sample size. We further propose a group variable selection strategy to simultaneously select and estimate important random effects, where the unknown covariance matrix of random effects is replaced with a proxy matrix. We prove that, with the proxy matrix appropriately chosen, the proposed procedure can identify all true random effects with asymptotic probability one, where the dimension of random effects vector is allowed to increase exponentially with the sample size. Monte Carlo simulation studies are conducted to examine the finite-sample performance of the proposed procedures. We further illustrate the proposed procedures via a real data example.

1. Introduction. During the last two decades, linear mixed effects models [Laird and Ware (1982), Longford (1993)] have been widely used to model longitudinal and repeated measurements data, and have received much attention in the fields of agriculture, biology, economics, medicine and sociology; see Verbeke and Molenberghs (2000) and references therein. With the advent of modern technology, many variables can be easily collected in a scientific study, and it is typical to include many of them in the full model at the initial stage of modeling to reduce model approximation error. Due to the complexity of the mixed effects models, the inferences and interpretation of estimated models become challenging as the dimension of fixed or random components increases. Thus the selection of impor-

tant fixed or random components becomes a fundamental problem in the analysis of longitudinal or repeated measurements data using mixed effects models.

Variable selection for mixed effects models has become an active research topic in the literature. Lin (1997) considers testing a hypothesis on the variance component. The testing procedures can be used to detect whether an individual random component is significant or not. Based on these testing procedures, a stepwise procedure can be constructed for selecting important random effects. Vaida and Blanchard (2005) propose the conditional AIC, an extension of the AIC [Akaike (1973)], for mixed effects models with detailed discussion on how to define degrees of freedom in the presence of random effects. The conditional AIC has further been discussed in Liang, Wu and Zou (2008). Chen and Dunson (2003) develop a Bayesian variable selection procedure for selecting important random effects in the linear mixed effects model using the Cholesky decomposition of the covariance matrix of random effects, and specify a prior distribution on the standard deviation of random effects with a positive mass at zero to achieve the sparsity of random components. Pu and Niu (2006) extend the generalized information criterion to select linear mixed effects models and study the asymptotic behavior of the proposed method for selecting fixed effects. Bondell, Krishna and Ghosh (2010) propose a joint variable selection method for fixed and random effects in the linear mixed effects model using a modified Cholesky decomposition in the setting of fixed dimensionality for both fixed effects and random effects. Ibrahim et al. (2011) propose to select fixed and random effects in a general class of mixed effects models with fixed dimensions of both fixed and random effects using maximum penalized likelihood method with the SCAD penalty and the adaptive least absolute shrinkage and selection operator penalty.

In this paper, we develop a class of variable selection procedures for both fixed effects and random effects in linear mixed effects models by incorporating the recent advances in variable selection. We propose to use the regularization methods to select and estimate fixed and random effects. As advocated by Fan and Li (2001), regularization methods can avoid the stochastic error of variable selection in stepwise procedures, and can significantly reduce computational cost compared with the best subset selection and Bayesian procedures. Our proposal differs from the existing ones in the literature mainly in two aspects. First, we consider the high-dimensional setting and allow dimension of fixed or random effects to grow exponentially with the sample size. Second, our proposed procedures can estimate the fixed effects vector without estimating the random effects vector and vice versa.

We first propose a class of variable selection methods for the fixed effects using penalized profile likelihood method. To overcome the difficulty of unknown covariance matrix of random effects, we propose to replace it with a suitably chosen proxy matrix. The penalized profile likelihood is equivalent to a penalized quadratic loss function of the fixed effects. Thus, the proposed approach can take advantage of the recent developments in the computation of the penalized least-squares methods [Efron et al. (2004), Zou and Li (2008)]. The optimization of the

penalized likelihood can be solved by the LARS algorithm without extra effort. We further systematically study the sampling properties of the resulting estimate of fixed effects. We establish conditions on the proxy matrix and show that the resulting estimate enjoys model selection oracle property under such conditions. In our theoretical investigation, the number of fixed effects is allowed to grow exponentially with the total sample size, provided that the covariance matrix of random effects is nonsingular. In the case of singular covariance matrix for random effects, one can use our proposed method in Section 3 to first select important random effects and then conduct variable selection for fixed effects. In this case, the number of fixed effects needs to be smaller than the total sample size.

Since the random effects vector is random, our main interest is in the selection of true random effects. Observe that if a random effect covariate is a noise variable, then the corresponding realizations of this random effect should all be zero, and thus the random effects vector is sparse. So we propose to first estimate the realization of random effects vector using a group regularization method and then identify the important ones based on the estimated random effects vector. More specifically, under the Bayesian framework, we show that the restricted posterior distribution of the random effects vector is independent of the fixed effects coefficient vector. Thus, we propose a random effect selection procedure via penalizing the restricted posterior mode. The proposed procedure reduces the impact of error caused by the fixed effects selection and estimation. The unknown covariance matrix is replaced with a suitably chosen proxy matrix. In the proposed procedure, random effects selection is carried out with group variable selection techniques [Yuan and Lin (2006)]. The optimization of the penalized restricted posterior mode is equivalent to the minimization of the penalized quadratic function of random effects. In particular, the form of the penalized quadratic function is similar to that in the adaptive elastic net [Zou and Hastie (2005), Zou and Zhang (2009)], which allows us to minimize the penalized quadratic function using existing algorithms. We further study the theoretical properties of the proposed procedure and establish conditions on the proxy matrix for ensuring the model selection consistency of the resulting estimate. We show that, with probability tending to one, the proposed procedure can select all true random effects. In our theoretical study, the dimensionality of random effects vector is allowed to grow exponentially with the sample size as long as the number of fixed effects is less than the total sample size.

The rest of this paper is organized as follows. Section 2 introduces the penalized profile likelihood method for the estimation of fixed effects and establishes its oracle property. We consider the estimation of random effects and prove the model selection consistency of the resulting estimator in Section 3. Section 4 provides two simulation studies and a real data example. Some discussion is given in Section 5. All proofs are presented in Section 6.

2. Penalized profile likelihood for fixed effects. Suppose that we have a sample of N subjects. For the ith subject, we collect the response variable

y_{ij}, the $d \times 1$ covariate vector $\mathbf{x}_{ij}$ and $q \times 1$ covariate vector $\mathbf{z}_{ij}$, for $j = 1, \ldots, n_i$, where n_i is the number of observations on the ith subject. Let $n = \sum_{i=1}^{N} n_i$, $m_n = \max_{1 \le i \le N} n_i$, and $\tilde{m}_n = \min_{1 \le i \le N} n_i$. We consider the case where $\limsup_n \frac{m_n}{\tilde{m}_n} < \infty$, that is, the sample sizes for N subjects are balanced. For succinct presentation, we use matrix notation and write $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i})^T$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{in_i})^T$. In linear mixed effects models, the vector of repeated measurements $\mathbf{y}_i$ on the ith subject is assumed to follow the linear regression model

$$(1) \qquad\qquad \mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\beta}$ is the $d \times 1$ population-specific fixed effects coefficient vector, $\boldsymbol{\gamma}_i$ represents the $q \times 1$ subject-specific random effects with $\boldsymbol{\gamma}_i \sim N(\mathbf{0}, G)$, $\boldsymbol{\varepsilon}_i$ is the random error vector with components independent and identically distributed as $N(0, \sigma^2)$, and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are independent. Here, G is the covariance matrix of random effects and may be different from the identity matrix. So the random effects can be correlated with each other.

Let vectors $\mathbf{y}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$, and matrix $\mathbf{X}$ be obtained by stacking vectors $\mathbf{y}_i$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\varepsilon}_i$, and matrices $\mathbf{X}_i$, respectively, underneath each other, and let $\mathbf{Z} = \text{diag}\{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\}$ and $\mathcal{G} = \text{diag}\{G, \ldots, G\}$ be block diagonal matrices. We further standardize the design matrix $\mathbf{X}$ such that each column has norm $\sqrt{n}$. The linear mixed effects model (1) can be rewritten as

$$(2) \qquad\qquad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

2.1. *Selection of important fixed effects.* In this subsection, we assume that there are no noise random effects, and $\mathcal{G}$ is positive definite. In the case where noise random effects exist, one can use the method in Section 3 to select the true ones. The joint density of $\mathbf{y}$ and $\boldsymbol{\gamma}$ is

$$
\begin{aligned}
f(\mathbf{y}, \boldsymbol{\gamma}) &= f(\mathbf{y}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}) \\
(3) \qquad &= (2\pi\sigma)^{-(n+qN)/2} |\mathcal{G}|^{-1/2} \\
&\quad \times \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}^T \mathcal{G}^{-1} \boldsymbol{\gamma} \right\}.
\end{aligned}
$$

Given $\boldsymbol{\beta}$, the maximum likelihood estimate (MLE) for $\boldsymbol{\gamma}$ is $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \mathbf{B}_z(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\mathbf{B}_z = (\mathbf{Z}^T\mathbf{Z} + \sigma^2\mathcal{G}^{-1})^{-1}\mathbf{Z}^T$. Plugging $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ into $f(\mathbf{y}, \boldsymbol{\gamma})$ and dropping the constant term yield the following profile likelihood function:

$$(4) \qquad\qquad L_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})) = \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{P}_z (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\},$$

where $\mathbf{P}_z = (\mathbf{I} - \mathbf{Z}\mathbf{B}_z)^T (\mathbf{I} - \mathbf{Z}\mathbf{B}_z) + \sigma^2\mathbf{B}_z^T \mathcal{G}^{-1} \mathbf{B}_z$ with $\mathbf{I}$ being the identity matrix. By Lemma 3 in Section 6, $\mathbf{P}_z$ can be rewritten as $\mathbf{P}_z = (\mathbf{I} + \sigma^{-2}\mathbf{Z}\mathcal{G}\mathbf{Z}^T)^{-1}$. To

select the important x-variables, we propose to maximize the following penalized profile log-likelihood function:

$$(5) \qquad \log\big(L_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}))\big) - n \sum_{j=1}^{d_n} p_{\lambda_n}(|\beta_j|),$$

where $p_{\lambda_n}(x)$ is a penalty function with regularization parameter $\lambda_n \geq 0$. Here, the number of fixed effects d_n may increase with sample size n.

Maximizing (5) is equivalent to minimizing

$$(6) \qquad Q_n(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{P}_z (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^{d_n} p_{\lambda_n}(|\beta_j|).$$

Since $\mathbf{P}_z$ depends on the unknown covariance matrix $\mathcal{G}$ and σ^2, we propose to use a proxy $\widetilde{\mathbf{P}}_z = (\mathbf{I} + \mathbf{Z}\mathcal{M}\mathbf{Z}^T)^{-1}$ to replace $\mathbf{P}_z$, where $\mathcal{M}$ is a pre-specified matrix. Denote by $\widetilde{Q}_n(\boldsymbol{\beta})$ the corresponding objective function when $\widetilde{\mathbf{P}}_z$ is used. We will discuss in the next section how to choose $\mathcal{M}$.

We note that (6) does not depend on the inverse of $\mathcal{G}$. So although we started this section with the nonsingularity assumption of $\mathcal{G}$, in practice our method can be directly applied even when noise random effects exist, as will be illustrated in simulation studies of Section 4.

Many authors have studied the selection of the penalty function to achieve the purpose of variable selection for the linear regression model. Tibshirani (1996) proposes the Lasso method by the use of L_1 penalty. Fan and Li (2001) advocate the use of nonconvex penalties. In particular, they suggest the use of the SCAD penalty. Zou (2006) proposes the adaptive Lasso by using adaptive L_1 penalty, Zhang (2010) proposes the minimax concave penalty (MCP), Liu and Wu (2007) propose to linearly combine L_0 and L_1 penalties and Lv and Fan (2009) introduce a unified approach to sparse recovery and model selection using general concave penalties. In this paper, we use concave penalty function for variable selection.

CONDITION 1. For each $\lambda > 0$, the penalty function $p_\lambda(t)$ with $t \in [0, \infty)$ is increasing and concave with $p_\lambda(0) = 0$, its second order derivative exists and is continuous and $p'_\lambda(0+) \in (0, \infty)$. Further, assume that $\sup_{t>0} p''_\lambda(t) \to 0$ as $\lambda \to 0$.

Condition 1 is commonly assumed in studying regularization methods with concave penalties. Similar conditions can be found in Fan and Li (2001), Fan and Peng (2004) and Lv and Fan (2009). Although it is assumed that $p''_\lambda(t)$ exists and is continuous, it can be relaxed to the case where only $p'_\lambda(t)$ exists and is continuous. All theoretical results presented in later sections can be generalized by imposing conditions on the local concavity of $p_\lambda(t)$, as in Lv and Fan (2009).

2.2. *Model selection consistency.* Although the proxy matrix $\widetilde{\mathbf{P}}_z$ may be different from the true one $\mathbf{P}_z$, solving the regularization problem (6) may still yield correct model selection results at the cost of some additional bias. We next establish conditions on $\widetilde{\mathbf{P}}_z$ to ensure the model selection oracle property of the proposed method.

Let $\boldsymbol{\beta}_0$ be the true coefficient vector. Suppose that $\boldsymbol{\beta}_0$ is sparse, and denote $s_{1n} = \|\boldsymbol{\beta}_0\|_0$, that is, the number of nonzero elements in $\boldsymbol{\beta}_0$. Write

$$\boldsymbol{\beta}_0 = (\beta_{1,0}, \ldots, \beta_{d_n,0})^T = (\boldsymbol{\beta}_{1,0}^T, \boldsymbol{\beta}_{2,0}^T)^T,$$

where $\boldsymbol{\beta}_{1,0}$ is an s_{1n}-vector and $\boldsymbol{\beta}_{2,0}$ is a $(d_n - s_{1n})$-vector. Without loss of generality, we assume that $\boldsymbol{\beta}_{2,0} = \mathbf{0}$, that is, the nonzero elements of $\boldsymbol{\beta}_0$ locate at the first s_{1n} coordinates. With a slight abuse of notation, we write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with $\mathbf{X}_1$ being a submatrix formed by the first s_{1n} columns of $\mathbf{X}$ and $\mathbf{X}_2$ being formed by the remaining columns. For a matrix $\mathbf{B}$, let $\Lambda_{\min}(\mathbf{B})$ and $\Lambda_{\max}(\mathbf{B})$ be its minimum and maximum eigenvalues, respectively. We will need the following assumptions.

CONDITION 2.   (A) Let $a_n = \min_{1 \leq j \leq s_{1n}} |\beta_{0,j}|$. It holds that

$$a_n n^\tau (\log n)^{-3/2} \to \infty$$

with $\tau \in (0, \frac{1}{2})$ being some positive constant, and $\sup_{t \geq a_n/2} p_{\lambda_n}''(t) = o(n^{-1+2\tau})$.

(B) There exists a constant $c_1 > 0$ such that $\Lambda_{\min}(c_1 \mathcal{M} - \sigma^{-2}\mathcal{G}) \geq 0$ and $\Lambda_{\min}(c_1 \sigma^{-2}(\log n)\mathcal{G} - \mathcal{M}) \geq 0$.

(C) The minimum and maximum eigenvalues of matrices $n^{-1}(\mathbf{X}_1^T \mathbf{X}_1)$ and $n^\theta (\mathbf{X}_1^T \mathbf{P}_z \mathbf{X}_1)^{-1}$ are both bounded from below and above by c_0 and c_0^{-1}, respectively, where $\theta \in (2\tau, 1]$ and $c_0 > 0$ is a constant. Further, it holds that

$$(7) \qquad \left\|\left(\frac{1}{n}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1\right)^{-1}\right\|_\infty \leq n^{-\tau}(\log n)^{3/4}/p_{\lambda_n}'(a_n/2),$$

$$(8) \qquad \|\mathbf{X}_2^T \widetilde{\mathbf{P}}_z \mathbf{X}_1 (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\|_\infty < p_{\lambda_n}'(0+)/p_{\lambda_n}'(a_n/2),$$

where $\|\cdot\|_\infty$ denotes the matrix infinity norm.

Condition 2(A) is on the minimum signal strength a_n. We allow the minimum signal strength to decay with sample size n. When concave penalties such as SCAD [Fan and Li (2001)] or SICA [Lv and Fan (2009)] are used, this condition can be easily satisfied with λ_n appropriately chosen. Conditions 2(B) and (C) put constraints on the proxy $\mathcal{M}$. Condition 2(C) is about the design matrices $\mathbf{X}$ and $\mathbf{Z}$. Inequality (8) requires noise variables and signal variables not highly correlated. The upper bound of (8) depends on the ratio $p_{\lambda_n}'(0+)/p_{\lambda_n}'(a_n/2)$. Thus, concave penalty functions relax this condition when compared to convex penalty functions. We will further discuss constraints (7) and (8) in Lemma 1.

If the above conditions on the proxy matrix are satisfied, then the bias caused by using $\widetilde{\mathbf{P}}_z$ is small enough, and the resulting estimate still enjoys the model selection oracle property described in the following theorem.

THEOREM 1. *Assume that $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$ and $\log d_n = o(n\lambda_n^2)$. Then under Conditions 1 and 2, with probability tending to 1 as $n \to \infty$, there exists a strict local minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ of $\widetilde{Q}_n(\boldsymbol{\beta})$ which satisfies*

$$(9) \qquad \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{0,1}\|_\infty < n^{-\tau}(\log n) \quad and \quad \widehat{\boldsymbol{\beta}}_2 = \mathbf{0}.$$

Theorem 1 presents the weak oracle property in the sense of Lv and Fan (2009) on the local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$. Due to the high dimensionality and the concavity of $p_\lambda(\cdot)$, the characterization of the global minimizer of $\widetilde{Q}(\boldsymbol{\beta})$ is a challenging open question. As will be shown in the simulation and real data analysis, the concave function $\widetilde{Q}(\boldsymbol{\beta})$ will be iteratively minimized by the local linear approximation method [Zou and Li (2008)]. Following the same idea as in Zou and Li (2008), it can be shown that the resulting estimate poesses the properties in Theorem 1 under some conditions.

2.3. *Choice of proxy matrix* $\mathcal{M}$. It is difficult to see from (7) and (8) on how restrictive the conditions on the proxy matrix $\mathcal{M}$ are. So we further discuss these conditions in the lemma below. We introduce the notation $\mathbf{T} = \sigma^2 \mathcal{G}^{-1} + \mathbf{Z}^T \mathbf{P}_x \mathbf{Z}$ and $\mathbf{E} = \sigma^2 \mathcal{G}^{-1} + \mathbf{Z}^T \mathbf{Z}$ with $\mathbf{P}_x = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1$. Correspondingly, when the proxy matrix $\mathcal{M}$ is used, define $\widetilde{\mathbf{T}} = \mathcal{M}^{-1} + \mathbf{Z}^T \mathbf{P}_x \mathbf{Z}$ and $\widetilde{\mathbf{E}} = \mathcal{M}^{-1} + \mathbf{Z}^T \mathbf{Z}$. We use $\|\cdot\|_2$ to denote the matrix 2-norm, that is, $\|\mathbf{B}\|_2 = \{\Lambda_{\max}(\mathbf{BB}^T)\}^{1/2}$ for a matrix $\mathbf{B}$.

LEMMA 1. *Assume that $\|(\frac{1}{n}\mathbf{X}_1^T \mathbf{P}_z \mathbf{X}_1)^{-1}\|_\infty < n^{-\tau}\sqrt{\log n}/p'_{\lambda_n}(a_n/2)$ and*

$$(10) \qquad \|\mathbf{T}^{-1/2}\widetilde{\mathbf{T}}\mathbf{T}^{-1/2} - \mathbf{I}\|_2 < \big(1 + n^\tau s_{1n}^{1/2} p'_{\lambda_n}(a_n/2)\|\mathbf{Z}\mathbf{T}^{-1}\mathbf{Z}^T\|_2\big)^{-1}.$$

Then (7) *holds.*

Similarly, assume that $\|\mathbf{X}_2^T \mathbf{P}_z \mathbf{X}_1(\mathbf{X}_1^T \mathbf{P}_z \mathbf{X}_1)^{-1}\|_\infty < p'_{\lambda_n}(0+)/p'_{\lambda_n}(a_n/2)$, and there exists a constant $c_2 > 0$ such that

$$
\begin{aligned}
&\|\mathbf{T}^{-1/2}\widetilde{\mathbf{T}}\mathbf{T}^{-1/2} - \mathbf{I}\|_2 \\
(11) \quad &< \big[1 + n^{-1}\|\mathbf{Z}\mathbf{T}^{-1}\mathbf{Z}^T\|_2 \\
&\qquad \times \max\{c_2 n^\theta, c_0^{-1}(\log n) s_{1n}^{1/2} \lambda_n^{-1} p'_{\lambda_n}(a_n/2)\|\mathbf{X}_2^T \mathbf{P}_z \mathbf{X}_1\|_2\}\big]^{-1},
\end{aligned}
$$

$$
\begin{aligned}
&\|\mathbf{E}^{-1/2}\widetilde{\mathbf{E}}\mathbf{E}^{-1/2} - \mathbf{I}\|_2 \\
(12) \quad &< \big[1 + \lambda_n^{-1}(\log n) s_{1n}^{1/2}(\log n) p'_{\lambda_n}(a_n/2) \\
&\qquad \times \|\mathbf{Z}\mathcal{G}\mathbf{Z}^T\|_2 \{\|(\mathbf{X}_1^T \mathbf{P}_z \mathbf{X}_1)^{-1}\|_2 \|\mathbf{X}_2^T \mathbf{P}_z \mathbf{X}_2\|_2\}^{1/2}\big]^{-1},
\end{aligned}
$$

then (8) *holds.*

Equations (10), (11) and (12) show conditions on the proxy matrix $\mathcal{M}$. Note that if penalty function used is flat outside of a neighborhood of zero, then $p'_{\lambda_n}(a_n/2) \approx 0$ with appropriately chosen regularization parameter λ_n, and conditions (10) and (12), respectively, reduce to

$$(13) \qquad \left\| \mathbf{T}^{-1/2}\widetilde{\mathbf{T}}\mathbf{T}^{-1/2} - \mathbf{I} \right\|_2 < 1, \qquad \left\| \mathbf{E}^{-1/2}\widetilde{\mathbf{E}}\mathbf{E}^{-1/2} - \mathbf{I} \right\|_2 < 1.$$

Furthermore, since $\mathbf{Z}$ is a block diagonal matrix, if the maximum eigenvalue of $\mathbf{Z}\mathbf{T}^{-1}\mathbf{Z}^T$ is of the order $o(n^{1-\theta})$, then condition (11) reduces to

$$(14) \qquad \left\| \mathbf{T}^{-1/2}\widetilde{\mathbf{T}}\mathbf{T}^{-1/2} - \mathbf{I} \right\|_2 < 1.$$

Conditions (13) and (14) are equivalent to assuming that $\mathbf{T}^{-1/2}\widetilde{\mathbf{T}}\mathbf{T}^{-1/2}$ and $\mathbf{E}^{-1/2}\widetilde{\mathbf{E}}\mathbf{E}^{-1/2}$ have eigenvalues bounded between 0 and 2. By linear algebra, they can further be reduced to $\|\mathbf{T}^{-1}\widetilde{\mathbf{T}}\|_2 < 2$ and $\|\mathbf{E}^{-1}\widetilde{\mathbf{E}}\|_2 < 2$. It is seen from the definitions of $\mathbf{T}$, $\widetilde{\mathbf{T}}$, $\mathbf{E}$ and $\widetilde{\mathbf{E}}$ that if eigenvalues of $\mathbf{Z}\mathbf{P}_x\mathbf{Z}^T$ and $\mathbf{Z}\mathbf{Z}^T$ dominate those of $\sigma^2\mathcal{G}^{-1}$ by a larger order of magnitude, then these conditions are not difficult to satisfy. In fact, note that both $\mathbf{Z}\mathbf{P}_x\mathbf{Z}^T$ and $\mathbf{Z}\mathbf{Z}^T$ have components with magnitudes increasing with n, while the components of $\sigma^2\mathcal{G}^{-1}$ are independent of n. Thus as long as both matrices $\mathbf{Z}\mathbf{P}_x\mathbf{Z}^T$ and $\mathbf{Z}\mathbf{Z}^T$ are nonsingular, these conditions will easily be satisfied with the choice $\mathcal{M} = (\log n)\mathbf{I}$ when n is large enough.

3. Identifying important random effects. In this section, we allow the number of random effects q to increase with sample size n and write it as q_n to emphasize its dependency on n. We focus on the case where the number of fixed effects d_n is smaller than the total sample size $n = \sum_{i=1}^{N} n_i$. We discuss the $d_n \geq n$ case in the discussion Section 5. The major goal of this section is to select important random effects.

3.1. *Regularized posterior mode estimate.* The estimation of random effects is different from the estimation of fixed effects, as the vector $\boldsymbol{\gamma}$ is random. The empirical Bayes method has been used to estimate the random effects vector $\boldsymbol{\gamma}$ in the literature. See, for example, Box and Tiao (1973), Gelman et al. (1995) and Verbeke and Molenberghs (2000). Although the empirical Bayes method is useful in estimating random effects in many situations, it cannot be used to select important random effects. Moreover, the performance of an empirical Bayes estimate largely depends on the accuracy of estimated fixed effects. These difficulties call for a new proposal for random effects selection.

Patterson and Thompson (1971) propose the error contrast method to obtain the restricted maximum likelihood of a linear model. Following their notation, define the $n \times (n - d)$ matrix $\mathbf{A}$ by the conditions $\mathbf{A}\mathbf{A}^T = \mathbf{P}_x$ and $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, where $\mathbf{P}_x = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Then the vector $\mathbf{A}^T\boldsymbol{\varepsilon}$ provides a particular set of $n - d$ linearly independent error contrasts. Let $\mathbf{w}_1 = \mathbf{A}^T\mathbf{y}$. The following proposition characterizes the conditional distribution of $\mathbf{w}_1$:

PROPOSITION 1. *Given* $\boldsymbol{\gamma}$, *the density function of* $\mathbf{w}_1$ *takes the form*

$$(15) \quad f_{\mathbf{w}_1}(\mathbf{A}^T \mathbf{y}|\boldsymbol{\gamma}) = (2\pi\sigma^2)^{-(n-d)/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})\right\}.$$

The above conditional probability is independent of the fixed effects vector $\boldsymbol{\beta}$ and the error contrast matrix $\mathbf{A}$, which allows us to obtain a posterior mode estimate of $\boldsymbol{\gamma}$ without estimating $\boldsymbol{\beta}$ and calculating $\mathbf{A}$.

Let $\mathfrak{M}_0 \subset \{1, 2, \ldots, q_n\}$ be the index set of the true random effects. Define

$$\overline{\mathfrak{M}}_0 = \{j : j = iq_n + k, \text{ for } i = 0, 1, 2, \ldots, N - 1 \text{ and } k \in \mathfrak{M}_0\}$$

and denote by $\overline{\mathfrak{M}}_0^c = \{1, 2, \ldots, Nq_n\} \setminus \overline{\mathfrak{M}}_0$. Then $\overline{\mathfrak{M}}_0$ is the index set of nonzero random effects coefficients in the vector $\boldsymbol{\gamma}$, and $\overline{\mathfrak{M}}_0^c$ is the index set of the zero ones. Let $s_{2n} = \|\mathfrak{M}_0\|_0$ be the number of true random effects. Then $\|\overline{\mathfrak{M}}_0\|_0 = Ns_{2n}$. We allow Ns_{2n} to diverge with sample size n, which covers both the case where the number of subjects N diverges with n alone and the case where N and s_{2n} diverge with n simultaneously.

For any $\mathcal{S} \subset \{1, \ldots, q_n N\}$, we use $\mathbf{Z}_{\mathcal{S}}$ to denote the $(q_n N) \times |\mathcal{S}|$ submatrix of $\mathbf{Z}$ formed by columns with indices in $\mathcal{S}$, and $\boldsymbol{\gamma}_{\mathcal{S}}$ to denote the subvector of $\boldsymbol{\gamma}$ formed by components with indices in $\mathcal{S}$. Then $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0} \sim N(\mathbf{0}, \mathcal{G}_{\overline{\mathfrak{M}}_0})$ with $\mathcal{G}_{\overline{\mathfrak{M}}_0}$ a submatrix formed by entries of $\mathcal{G}$ with row and column indices in $\overline{\mathfrak{M}}_0$. In view of (15), the restricted posterior density of $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}$ can be derived as

$$f_{\mathbf{w}_1}(\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}|\mathbf{A}^T \mathbf{y}) \propto f_{\mathbf{w}_1}(\mathbf{A}^T \mathbf{y}|\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}) f(\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0})$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Z}_{\overline{\mathfrak{M}}_0}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}_{\overline{\mathfrak{M}}_0}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}) - \frac{1}{2}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^T \mathcal{G}_{\overline{\mathfrak{M}}_0}^{-1} \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}\right\}.$$

Therefore, the restricted posterior mode estimate of $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}$ is the solution to the following minimization problem:

$$(16) \qquad \min_{\boldsymbol{\gamma}}\left\{(\mathbf{y} - \mathbf{Z}_{\overline{\mathfrak{M}}_0}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}_{\overline{\mathfrak{M}}_0}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}) + \sigma^2 \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^T \mathcal{G}_{\overline{\mathfrak{M}}_0}^{-1} \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}\right\}.$$

In practice, since the true random effects $\overline{\mathfrak{M}}_0$ are unknown, the formulation (16) does not help us estimate $\boldsymbol{\gamma}$. To overcome this difficulty, note that $\mathbf{Z}_{\overline{\mathfrak{M}}_0}\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0} = \mathbf{Z}\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^T \mathcal{G}_{\overline{\mathfrak{M}}_0}^{-1} \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0} = \boldsymbol{\gamma}^T \mathcal{G}^+ \boldsymbol{\gamma}$ with $\mathcal{G}^+$ the Moore–Penrose generalized inverse of $\mathcal{G}$. Thus the objective function in (16) is rewritten as

$$(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \sigma^2 \boldsymbol{\gamma}^T \mathcal{G}^+ \boldsymbol{\gamma},$$

which no longer depends on the unknown $\overline{\mathfrak{M}}_0$. Observe that if the kth random effect is a noise one, then the corresponding standard deviation is 0, and the coefficients γ_{ik} for all subjects $i = 1, \ldots, N$ should equal to 0. This leads us to consider group variable selection strategy to identify true random effects. Define

$\gamma_{\cdot k} = (\sum_{i=1}^{N} \gamma_{ik}^2)^{1/2}, k = 1, \ldots, q_n$, and consider the following regularization problem:

$$(17) \qquad \frac{1}{2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \frac{1}{2}\sigma^2 \boldsymbol{\gamma}^T \mathcal{G}^+ \boldsymbol{\gamma} + n \sum_{k=1}^{q_n} p_{\lambda_n}(\gamma_{\cdot k}),$$

where $p_{\lambda_n}(\cdot)$ is the penalty function with regularization parameter $\lambda_n \geq 0$. The penalty function here may be different from the one in Section 2. However, to ease the presentation, we use the same notation.

There are several advantages to estimating the random effects vector $\boldsymbol{\gamma}$ using the above proposed method (17). First, this method does not require knowing or estimating the fixed effects vector $\boldsymbol{\beta}$, so it is easy to implement, and the estimation error of $\boldsymbol{\beta}$ has no impact on the estimation of $\boldsymbol{\gamma}$. In addition, by using the group variable selection technique, the true random effects can be simultaneously selected and estimated.

In practice, the covariance matrix $\mathcal{G}$ and the variance σ^2 are both unknown. Thus, we replace $\sigma^{-2}\mathcal{G}$ with $\mathcal{M}$, where $\mathcal{M} = \text{diag}\{M, \ldots, M\}$ with M a proxy of G, yielding the following regularization problem:

$$(18) \qquad \widetilde{Q}_n^*(\boldsymbol{\gamma}) = \frac{1}{2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \frac{1}{2}\boldsymbol{\gamma}^T \mathcal{M}^{-1}\boldsymbol{\gamma} + n \sum_{k=1}^{q_n} p_{\lambda_n}(\gamma_{\cdot k}).$$

It is interesting to observe that the form of regularization in (18) includes the elastic net [Zou and Hastie (2005)] and the adaptive elastic net [Zou and Zhang (2009)] as special cases. Furthermore, the optimization algorithm for adaptive elastic net can be modified for minimizing (18).

3.2. *Asymptotic properties.* Minimizing (18) yields an estimate of $\boldsymbol{\gamma}$, denoted by $\widehat{\boldsymbol{\gamma}}$. In this subsection, we study the asymptotic property of $\widehat{\boldsymbol{\gamma}}$. Because $\boldsymbol{\gamma}$ is random rather than a deterministic parameter vector, the existing formulation for the asymptotic analysis of a regularization problem is inapplicable to our setting. Thus, asymptotic analysis of $\widehat{\boldsymbol{\gamma}}$ is challenging.

Let $\mathbf{T} = \mathbf{Z}^T \mathbf{P}_x \mathbf{Z} + \sigma^2 \mathcal{G}^+$ and $\widetilde{\mathbf{T}} = \mathbf{Z}^T \mathbf{P}_x \mathbf{Z} + \mathcal{M}^{-1}$. Denote by $\mathbf{T}_{11} = \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \times \mathbf{Z}_{\overline{\mathfrak{M}}_0} + \sigma^2 (\mathcal{G}_{\overline{\mathfrak{M}}_0})^{-1}$, $\mathbf{T}_{22} = \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}$ and $\mathbf{T}_{12} = \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}$. Similarly, we can define submatrices $\widetilde{\mathbf{T}}_{11}$, $\widetilde{\mathbf{T}}_{22}$ and $\widetilde{\mathbf{T}}_{12}$ by replacing $\sigma^{-2}\mathcal{G}$ with $\mathcal{M}$. Then it is easy to see that $\widetilde{\mathbf{T}}_{12} = \mathbf{T}_{12}$. Notice that if the oracle information of set $\overline{\mathfrak{M}}_0$ is available and $\mathcal{G}$, and σ^2 are known, then the Bayes estimate of the true random effects coefficient vector $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}$ has the form $\mathbf{T}_{11}^{-1}\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{y}$. Define $\boldsymbol{\gamma}^* = ((\boldsymbol{\gamma}_1^*)^T, \ldots, (\boldsymbol{\gamma}_N^*)^T)^T$ with $\boldsymbol{\gamma}_j^* = (\gamma_{j1}^*, \ldots, \gamma_{jq_n}^*)^T$ for $j = 1, \ldots, N$ as the oracle-assisted Bayes estimate of the random effects vector. Then $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0^c}^* = \mathbf{0}$ and $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^* = \mathbf{T}_{11}^{-1}\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{y}$. Correspondingly, define $\widetilde{\boldsymbol{\gamma}}^*$ as the oracle Bayses estimate with proxy matrix, that is, $\widetilde{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0^c}^* = \mathbf{0}$ and

$$(19) \qquad \widetilde{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^* = \widetilde{\mathbf{T}}_{11}^{-1}\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{y}.$$

For $k = 1, \ldots, q_n$, let $\gamma_{\cdot k}^* = \{\sum_{j=1}^N (\gamma_{jk}^*)^2\}^{1/2}$. Throughout we condition on the event

$$(20) \qquad \Omega^* = \left\{ \min_{k \in \mathfrak{M}_0} \gamma_{\cdot k}^* \geq \sqrt{N} b_0^* \right\}$$

with $b_0^* \in (0, \min_{\in \mathfrak{M}_0} \sigma_k)$ and $\sigma_k^2 = \text{var}(\gamma_{jk})$. The above event Ω^* is to ensure that the oracle-assisted estimator $\gamma_{\cdot k}^* / \sqrt{N}$ of σ_k is not too negatively biased.

CONDITION 3. (A) The maximum eigenvalues satisfy $\Lambda_{\max}(\mathbf{Z}_i G \mathbf{Z}_i^T) \leq c_3 s_{2n}$ for all $i = 1, \ldots, N$ and the minimum and maximum eigenvalues of $m_n^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0}$ and $G_{\mathfrak{M}_0}$ are bounded from below and above by c_3 and c_3^{-1}, respectively, with $m_n = \max_{1 \leq i \leq N} n_i$, where c_3 is a positive constant. Further, assume that for some $\delta \in (0, \frac{1}{2})$,

$$(21) \qquad \|\widetilde{\mathbf{T}}_{11}^{-1}\|_\infty \leq \frac{\sqrt{N} n^{-1-\delta}}{p_{\lambda_n}'(\sqrt{N} b_0^*/2)},$$

$$(22) \qquad \max_{j \in \mathfrak{M}_0^c} \|\widetilde{\mathbf{Z}}_j^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1}\|_2 < \frac{p_{\lambda_n}'(0+)}{p_{\lambda_n}'(\sqrt{N} b_0^*/2)},$$

where $\widetilde{\mathbf{Z}}_j$ is the submatrix formed by the N columns of $\mathbf{Z}$ corresponding to the jth random effect.

(B) It holds that $\sup_{\{t \geq \sqrt{N} b_0^*/2\}} p_{\lambda_n}''(t) = o(N^{-1})$.

(C) The proxy matrix satisfies $\Lambda_{\min}(\mathcal{M} - \sigma^{-2}\mathcal{G}) \geq 0$.

Condition 3(A) is about the design matrices $\mathbf{X}$, $\mathbf{Z}$ and covariance matrix $\mathcal{G}$. Since $\mathbf{Z}_{\overline{\mathfrak{M}}_0}$ is a block diagonal matrix and $\limsup \frac{\max_i n_i}{\min_i n_i} < \infty$, the components of $\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0}$ have magnitude of the order $m_n = O(n/N)$. Thus, it is not very restrictive to assume that the minimum and maximum eigenvalues of $\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0}$ are both of the order m_n. Condition (22) puts an upper bound on the correlation between noise covariates and true covariates. The upper bound of (22) depends on the penalty function. Note that for concave penalty we have $p_{\lambda_n}'(0+)/p_{\lambda_n}'(\sqrt{N} b_0^*/2) > 1$, whereas for L_1 penalty $p_{\lambda_n}'(0+)/p_{\lambda_n}'(\sqrt{N} b_0^*/2) = 1$. Thus, concave penalty relaxes (22) when compared with the L_1 penalty. Condition 3(B) is satisfied by many commonly used penalties with appropriately chosen λ_n, for example, L_1 penalty, SCAD penalty and SICA penalty with small a. Condition 3(C) is a restriction on the proxy matrix $\mathcal{M}$, which will be further discussed in the next subsection.

Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_N^T)^T$ with $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jq_n})^T$ being an arbitrary (Nq_n)-vector. Define $\gamma_{\cdot k} = (\sum_{j=1}^N \gamma_{jk}^2)^{1/2}$ for each $k = 1, \ldots, q_n$. Let

$$(23) \qquad \mathfrak{M}(\boldsymbol{\gamma}) = \{k \in \{1, \ldots, q_n\} : \gamma_{\cdot k} \neq 0\}.$$

Theorem 2 below shows that there exists a local minimizer of $\widetilde{Q}_n^*(\boldsymbol{\gamma})$ defined in (18) whose support is the same as the true one $\overline{\mathfrak{M}}_0$, and that this local minimizer is close to the oracle estimator $\widetilde{\boldsymbol{\gamma}}^*$.

THEOREM 2. *Assume that Conditions* 1 *and* 3 *hold,* $b_0^* n^\delta/\sqrt{N} \to \infty$, $\log(Nq_n) = o(n^2\lambda_n^2/(Ns_{2n}m_n))$, *and* $n^2\lambda_n^2/(Nm_ns_{2n}) \to \infty$ *as* $n \to \infty$. *Then, with probability tending to* 1, *there exists a strict local minimizer* $\widehat{\boldsymbol{\gamma}} \in \mathbf{R}^{Nq_n}$ *of* $\widetilde{Q}_n^*(\boldsymbol{\gamma})$ *such that*

$$\mathfrak{M}(\widehat{\boldsymbol{\gamma}}) = \mathfrak{M}_0 \quad and \quad \max_{k \in \mathfrak{M}_0} \left\{ \frac{1}{N} \sum_{j=1}^N (\widehat{\gamma}_{jk} - \widetilde{\gamma}_{jk}^*)^2 \right\}^{1/2} \le n^{-\delta},$$

where δ *is defined in* (21).

Using a similar argument to that for Theorem 1, we can obtain that the dimensionality Nq_n is also allowed to grow exponentially with sample size n under some growth conditions and with appropriately chosen λ_n. In fact, note that if the sample sizes $n_1 = \cdots = n_N \equiv m_n/N$, then the growth condition in Theorem 2 becomes $\log(Nq_n) = o(ns_{2n}^{-1}\lambda_n^2)$. Since the lowest signal level in this case is $\sqrt{N}b_0^*$, if b_0^* is a constant, a reasonable choice of tuning parameter would be of the order $\sqrt{N}n^{-\kappa}$ with some $\kappa \in (0, \frac{1}{2})$. For $s_{2n} = O(n^\nu)$ with $\nu \in [0, \frac{1}{2})$ and $Nn^{1-2\kappa-\nu} \to \infty$, we obtain that Nq_n can grow with rate $\exp(Nn^{1-2\kappa-\nu})$.

3.3. *Choice of proxy matrix* $\mathcal{M}$. Similarly as for the fixed effects selection and estimation, we discuss (21) and (22) in the following lemma.

LEMMA 2. *Assume that* $\|\mathbf{T}_{11}^{-1}\|_\infty < \frac{\sqrt{N}n^{-1-\delta}}{p'_{\lambda_n}(\sqrt{N}b_0^*/2)}[1 - \frac{1}{\sqrt{\log n}}]$ *and*

$$(24) \qquad \|\mathbf{T}_{11}^{-1}\widetilde{\mathbf{T}}_{11} - \mathbf{I}\|_2 \le [1 + \sqrt{s_{2n}\log n}\,n^{1+\delta}p'_{\lambda_n}(\sqrt{N}b_0^*/2)\|\mathbf{T}_{11}^{-1}\|_2]^{-1}.$$

Then (21) *holds.*

Assume that $\max_{j \in \mathfrak{M}_0^c} \|\widetilde{\mathbf{Z}}_j^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \mathbf{T}_{11}^{-1}\|_2 < \frac{p'_{\lambda_n}(0+)}{2p'_{\lambda_n}(\sqrt{N}b_0^*/2)}$ *with* $\widetilde{\mathbf{Z}}_j$ *defined in* (22) *and*

$$(25) \qquad\qquad\qquad \|\mathbf{T}_{11}\widetilde{\mathbf{T}}_{11}^{-1} - \mathbf{I}\|_2 \le 1.$$

Then (22) *holds.*

Conditions (24) and (25) put restrictions on the proxy matrix $\mathcal{M}$. Similarly to the discussions after Lemma 1, if $p'_{\lambda_n}(\sqrt{N}b_0^*/2) \approx 0$, then these conditions become $\|\mathbf{T}_{11}\widetilde{\mathbf{T}}_{11}^{-1} - \mathbf{I}\|_2 < 1$. If $\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0}$ dominates $\sigma^2\mathcal{G}_{\overline{\mathfrak{M}}_0}^{-1}$ by a larger magnitude, then conditions (24) and (25) are not restrictive, and choosing $\mathcal{M} = (\log n)\mathbf{I}$ should make these conditions as well as Condition 3(C) satisfied for large enough n.

We remark that using the proxy matrix $\mathcal{M} = (\log n)\mathbf{I}$ is equivalent to ignoring correlations among random effects. The idea of using diagonal matrix as a proxy of covariance matrix has been proposed in other settings of high-dimensional statistical inference. For instance, the naive Bayes rule (or independence rule), which replaces the full covariance matrix in Fisher's discriminant analysis with a diagonal matrix, has been demonstrated to be advantageous for high-dimensional classifications both theoretically [Bickel and Levina (2004), Fan and Fan (2008)] and empirically [Dudoit, Fridlyand and Speed (2002)]. The intuition is that although ignoring correlations gives only a biased estimate of covariance matrix, it avoids the errors caused by estimating a large amount of parameters in covariance matrix in high dimensions. Since the accumulated estimation error can be much larger than the bias, using diagonal proxy matrix indeed produces better results.

4. Simulation and application. In this section, we investigate the finite-sample performance of the proposed procedures by simulation studies and a real data analysis. Throughout, the SCAD penalty with $a = 3.7$ [Fan and Li (2001)] is used. For each simulation study, we randomly simulate 200 data sets. Tuning parameter selection plays an important role in regularization methods. For fixed effect selection, both AIC- and BIC-selectors [Zhang, Li and Tsai (2010)] are used to select the regularization parameter λ_n in (6). Our simulation results clearly indicate that the BIC-selector performs better than the AIC-selector for both the SCAD and the LASSO penalties. This is consistent with the theoretical analysis in Wang, Li and Tsai (2007). To save space, we report the results with the BIC-selector. Furthermore the BIC-selector is used for fixed effect selection throughout this section. For random effect selection, both AIC- and BIC-selectors are also used to select the regularization parameter λ_n in (18). Our simulation results imply that the BIC-selector outperforms the AIC-selector for the LASSO penalty, while the SCAD with AIC-selector performs better than the SCAD with BIC-selector. As a result, we use AIC-selector for the SCAD and BIC-selector for the LASSO for random effect selection throughout this section.

EXAMPLE 1. We compare our method with some existing ones in the literature under the same model setting as that in Bondell, Krishna and Ghosh (2010), where a joint variable selection method for fixed and random effects in linear mixed effects models is proposed. The underlying true model takes the following form with $q = 4$ random effects and $d = 9$ fixed effects:

$$(26) \quad y_{ij} = b_{i1} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{i2} z_{ij1} + b_{i3} z_{ij2} + \varepsilon_{ij}, \varepsilon_{ij} \sim_{\text{i.i.d.}} N(0, 1),$$

where the true parameter vector $\boldsymbol{\beta}_0 = (1, 1, 0, \ldots, 0)^T$, the true covariance matrix for random effects

$$\mathbf{G} = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}$$

and the covariates x_{ijk} for $k = 1, \ldots, 9$ and z_{ijl} for $l = 1, 2, 3$ are generated independently from a uniform distribution over the interval $[-2, 2]$. So there are three true random effects and two true fixed effects. Following Bondell, Krishna and Ghosh (2010), we consider two different sample sizes $N = 30$ subjects and $n_i = 5$ observations per subject, and $N = 60$ and $n_i = 10$. Under this model setting, Bondell, Krishna and Ghosh (2010) compared their method with various methods in the literature, and simulations therein demonstrate that their method outperforms the competing ones. So we will only compare our methods with the one in Bondell, Krishna and Ghosh (2010).

In implementation, the proxy matrix is chosen as $\mathcal{M} = (\log n)\mathbf{I}$. We then estimate the fixed effects vector $\boldsymbol{\beta}$ by minimizing $\widetilde{Q}_n(\boldsymbol{\beta})$, and the random effects vector $\boldsymbol{\gamma}$ by minimizing (18). To understand the effects of using proxy matrix $\mathcal{M}$ on the estimated random effects and fixed effects, we compare our estimates with the ones obtained by solving regularization problems (6) and (17) with the true value $\sigma^{-2}\mathcal{G}$.

Table 1 summarizes the results by using our method with the proxy matrix $\mathcal{M}$ and SCAD penalty (SCAD-P), our method with proxy matrix $\mathcal{M}$ and Lasso penalty (Lasso-P), our method with true $\sigma^{-2}\mathcal{G}$ and SCAD penalty (SCAD-T). When SCAD penalty is used, the local linear approximation (LLA) method proposed by Zou and Li (2008) is employed to solve these regularization problems. The rows "M-ALASSO" in Table 1 correspond to the joint estimation method by Bondell, Krishna and Ghosh (2010) using BIC to select the tuning parameter. As demonstrated in Bondell, Krishna and Ghosh (2010), the BIC-selector outperforms the AIC-selector for M-ALASSO. We compare these methods by calculating the percentage of times the correct fixed effects are selected (%CF), and the percentage of times the correct random effects are selected (%CR). Since these two measures were also used in Bondell, Krishna and Ghosh (2010), for simplicity and fairness of comparison, the results for M-ALASSO in Table 1 are copied from Bondell, Krishna and Ghosh (2010).

TABLE 1
*Fixed and random effects selection in Example 1 when $d = 9$
and $q = 4$*

Setting	Method	%CF	%CR
$N = 30$	Lasso-P	51	19.5
$n_i = 5$	SCAD-P	90	86
	SCAD-T	93.5	99
	M-ALASSO	73	79
$N = 60$	Lasso-P	52	50.5
$n_i = 10$	SCAD-P	100	100
	SCAD-T	100	100
	M-ALASSO	83	89

It is seen from Table 1 that SCAD-P greatly outperforms Lasso-P and M-ALASSO. We also see that when the true covariance matrix $\sigma^{-2}\mathcal{G}$ is used, SCAD-T has almost perfect variable selection results. Using the proxy matrix makes the results slightly inferior, but the difference vanishes for larger sample size $N = 60, n_i = 10$.

EXAMPLE 2. In this example, we consider the case where the design matrices for fixed and random effects overlap. The sample size is fixed at $n_i = 8$ and $N = 30$, and the numbers for fixed and random effects are chosen to be $d = 100$ and $q = 10$, respectively. To generate the fixed effects design matrix, we first independently generate $\widetilde{\mathbf{x}}_{ij}$ from $N_d(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{st})$ with $\sigma_{st} = \rho^{|s-t|}$ and $\rho \in (-1, 1)$. Then for the jth observation of the ith subject, we set $x_{ijk} = I(\widetilde{x}_{ijk} > 0)$ for covariates $k = 1$ and d, and set $x_{ijk} = \widetilde{x}_{ijk}$ for all other values of k. Thus 2 out of d covariates are discrete ones and the rest are continuous ones. Moreover, all covariates are correlated with each other. The covariates for random effects are the same as the corresponding ones for fixed effects, that is, for the jth observation of the ith subject, we set $z_{ijk} = x_{ijk}$ for $k = 1, \ldots, q = 10$. Then the random effect covariates form a subset of fixed effect covariates.

The first six elements of fixed effects vector $\boldsymbol{\beta}_0$ are $(2, 0, 1.5, 0, 0, 1)^T$, and the remaining elements are all zero. The random effects vector $\boldsymbol{\gamma}$ is generated in the same way as in Example 1. So the first covariate is discrete and has both nonzero fixed and random effect. We consider different values of correlation level ρ, as shown in Table 2. We choose $\mathcal{M} = (\log n)\mathbf{I}$.

Since the dimension of random effects vector $\boldsymbol{\gamma}$ is much larger than the total sample size, as suggested at the beginning of Section 2.1, we start with the ran-

TABLE 2
Fixed and random effects selection and estimation in Example 2 when $n_i = 8$, $N = 30$, $d = 100$,
$q = 10$ and design matrices for fixed and random effects overlap

| Setting | Method | Random effects | | | | Fixed effects | | | |
		FNR (%)	FPR (%)	MRL2	MRL1	FNR (%)	FPR (%)	MRL2	MRL1
$\rho = 0.3$	Lasso-P	11.83	9.50	0.532	0.619	62.67	0.41	0.841	0.758
	SCAD-P	0.50	1.07	0.298	0.348	0.83	0.03	0.142	0.109
	SCAD-T	3.83	0.00	0.522	0.141	0.33	0.02	0.102	0.082
$\rho = -0.3$	Lasso-P	23.67	7.64	0.524	0.580	59.17	0.41	0.802	0.745
	SCAD-P	1.83	0.71	0.308	0.352	0.67	0.05	0.141	0.109
	SCAD-T	3.17	0.00	0.546	0.141	0.17	0.02	0.095	0.078
$\rho = 0.5$	Lasso-P	9.83	10.07	0.548	0.631	60.33	0.48	0.844	0.751
	SCAD-P	1.67	0.50	0.303	0.346	0.17	0.05	0.138	0.110
	SCAD-T	5.00	0.00	0.532	0.149	0.50	0.02	0.113	0.091

dom effects selection by first choosing a relatively small tuning parameter λ and use our method in Section 3 to select important random effects. Then with the selected random effects, we apply our method in Section 2 to select fixed effects. To improve the selection results for random effects, we further use our method in Section 3 with the newly selected fixed effects to reselect random effects. This iterative procedure is applied to both Lasso-P and SCAD-P methods. For SCAD-T, since the true $\sigma^{-2}\mathcal{G}$ is used, it is unnecessary to use the iterative procedure, and we apply our methods only once for both fixed and random effects selection and estimation.

We evaluate each estimate by calculating the relative L_2 estimation loss

$$\mathrm{RL2}(\widehat{\boldsymbol{\beta}}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 / \|\boldsymbol{\beta}_0\|_2,$$

where $\widehat{\boldsymbol{\beta}}$ is an estimate of the fixed effects vector $\boldsymbol{\beta}_0$. Similarly, the relative L_1 estimation error of $\widehat{\boldsymbol{\beta}}$, denoted by $\mathrm{RL1}(\widehat{\boldsymbol{\beta}})$, can be calculated by replacing the L_2-norm with the L_1-norm. For the random effects estimation, we define $\mathrm{RL2}(\widehat{\boldsymbol{\gamma}})$ and $\mathrm{RL1}(\widehat{\boldsymbol{\gamma}})$ in a similar way by replacing $\boldsymbol{\beta}_0$ with the true $\boldsymbol{\gamma}$ in each simulation. We calculate the mean values of RL2 and RL1 in the simulations and denote them by MRL2 and MRL1 in Table 2. In addition to mean relative losses, we also calculate the percentages of missed true covaritates (FNR), as well as the percentages of falsely selected noise covariates (FPR), to evaluate the performance of proposed methods.

From Table 2 we see that SCAD-T has almost perfect variable selection results for fixed effects, while SCAD-P has highly comparable performance, for all three values of correlation level ρ. Both methods greatly outperform the Lasso-P method. For the random effects selection, both SCAD-P and SCAD-T perform very well with SCAD-T having slightly larger false negative rates. We remark that the superior performance of SCAD-P is partially because of the iterative procedure. In these high-dimensional settings, directly applying our random effects selection method in Section 3 produces slightly inferior results to the ones for SCAD-T in Table 2, but iterating once improves the results. We also see that as the correlation level increases, the performance of all methods become worse, but the SCAD-P is still comparable to SCAD-T, and both perform very well in all settings.

EXAMPLE 3.   We illustrate our new procedures through an empirical analysis of a subset of data collected in the Multi-center AIDs Cohort Study. Details of the study design, method and medical implications have been given by Kaslow et al. (1987). This data set comprises the human immunodeficiency virus (HIV) status of 284 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. All patients are scheduled to take measurements semiannually. However, due to the missing of scheduled visits and the random occurrence of HIV infections, there are an unequal number of measurements and different measurement times for each patients. The total number of observations is 1765.

TABLE 3
The estimated coefficients of fixed and random effects in Example 3

	Intercept	$b_1(t)$	$b_2(t)$	$b_3(t)$	$b_4(t)$	$b_5(t)$	x_1	x_2	x_3
Fixed	29.28	9.56	5.75	0	−8.32	0	4.95	0	0
Random	0	0	0	0	0	0	0	0	0

	$b_1(t)x_3$	$b_2(t)x_3$	$b_3(t)x_3$	$b_4(t)x_3$	$b_5(t)x_3$	x_1x_2	x_1x_3	x_2x_3
Fixed	0	0	0	0	0	0	0	0
Random	0.163	0.153	0.057	0.043	0.059	0	0.028	0.055

Of interest is to investigate the relation between the mean CD4 percentage after the infection (y) and predictors smoking status (x_1, 1 for smoker and 0 for non-smoker), age at infection (x_2), and pre-HIV infection CD4 percentage (Pre-CD4 for short, x_3). To account for the effect of time, we use a five-dimensional cubic spline $\mathbf{b}(t) = (b_1(t), b_2(t), \ldots, b_5(t))^T$. We take into account the two-way interactions $\mathbf{b}(t_{ij})x_{i3}$, $x_{i1}x_{i2}$, $x_{i1}x_{i3}$ and $x_{i2}x_{i3}$. These eight interactions together with variables $\mathbf{b}(t_{ij})$, x_{i1}, x_{i2} and x_{i3} give us 16 variables in total. We use these 16 variables together with an intercept to fit a mixed effects model with dimensions for fixed and random effects $d = q = 17$. The estimation results are listed in Table 3 with rows "Fixed" showing the estimated β_j's for fixed effects, and rows "Random" showing the estimates $\gamma_{\cdot k}/\sqrt{N}$. The standard error for the null model is 11.45, and it reduces to 3.76 for the selected model. From Table 3, it can be seen that the baseline has time-variant fixed effect and Pre-CD4 has time-variant random effect. Smoking has fixed effect while age and Pre-CD4 have no fixed effects. The interactions smoking × Pre-CD4 and age × Pre-CD4 have random effects with smallest standard deviations among selected random effects. The boxplots of the selected random effects are shown in Figure 1.

Our results have close connections with the ones in Huang, Wu and Zhou (2002) and Qu and Li (2006), where the former used bootstrap approach to test the significance of variables and the later proposed hypothesis test based on penalized spline and quadratic inference function approaches, for varying-coefficient models. Both papers revealed significant evidence for time-varying baseline, which is consistent with our discovery that basis functions $b_j(t)$'s have nonzero fixed effect coefficients. At 5% level, Huang, Wu and Zhou (2002) failed to reject the hypothesis of constant Pre-CD4 effect (*p*-value 0.059), while Qu and Li's (2006) test was weakly significant with *p*-value 0.045. Our results show that Pre-CD4 has constant fixed effect and time-varying random effect, which may provide an explanation on the small difference of *p*-values in Huang, Wu and Zhou (2002) and Qu and Li (2006).

To further access the significance of selected fixed effects, we refit the linear mixed effects model with selected fixed and random effects using the Matlab func-
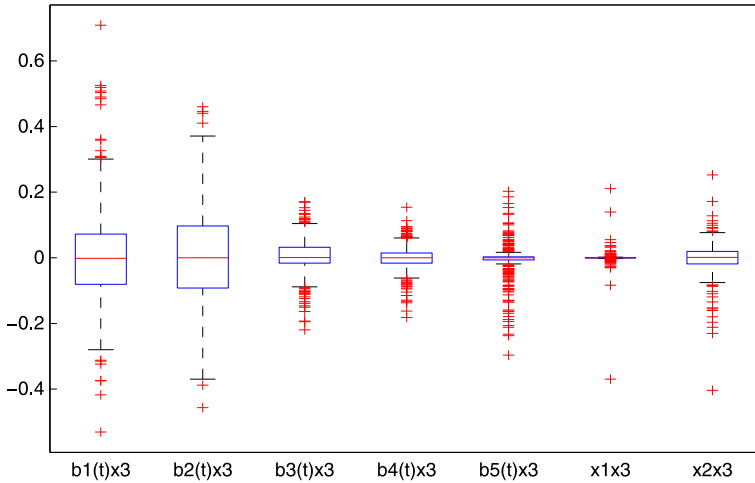
FIG. 1.   *Boxplots of selected random effects. From left to right*: $b_i(t)x_3$, $i = 1, 2, \ldots, 5$, $x_1 x_3$, $x_2 x_3$, *where* x_1 *is the smoking status*, x_2 *is the age at infection*, x_3 *is Pre-CD*4 *level and* $b_i(t)$*'s are cubic spline basis functions of time*.

tion "nlmefit." Based on the t-statistics from the refitted model, the intercept, the baseline functions $b_1(t)$ and $b_2(t)$ are all highly significant with t-statistics much larger than 7, while the t-statistics for $b_4(t)$ and x_1 (smoking) are -1.026 and 2.216, respectively. This indicates that $b_4(t)$ is insignificant, and smoking is only weakly significant at 5% significance level. This result is different from those in Huang, Wu and Zhou (2002) and Qu and Li (2006), where neither paper found significant evidence for smoking. A possible explanation is that by taking into account random effects and variable selection, our method has better discovery power.

5. Discussion.   We have discussed the selection and estimation of fixed effects in Section 2, providing that the random effects vector has nonsingular covariance matrix, while we have discussed the selection of random effects in Section 3, providing that the dimension of fixed effects vector is smaller than the sample size. We have also illustrated our methods with numerical studies. In practical implementation, the dimensions of the random effects vector and fixed effects vector can be both much larger than the total sample size. In such case, we suggest an iterative way to select and estimate the fixed and random effects. Specifically, we can first start with the fixed effects selection using the penalized least squares by ignoring all random effects to reduce the number of fixed effects to below sample size. Then in the second step, with the selected fixed effects, we can apply our new method in Section 3 to select important random effects. Third, with the selected random effects from the second step, we can use our method in Section 2 to further select important fixed effects. We can also iterate the second and third steps several times to improve the model selection and estimation results.

6. Proofs. Lemma 3 is proved in the supplemental article Fan and Li (2012).

LEMMA 3. *It holds that*

$$\mathbf{P}_z = (I - \mathbf{Z}\mathbf{B}_z)^T \mathcal{R}^{-1}(I - \mathbf{Z}\mathbf{B}_z) + \mathbf{B}_z^T \mathcal{G}^{-1} \mathbf{B}_z = (\mathcal{R} + \mathbf{Z}\mathcal{G}\mathbf{Z}^T)^{-1}.$$

6.1. *Proof of Theorem* 1. Let $\mathcal{N}_0 = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{0,1}\|_\infty \leq n^{-\tau}(\log n), \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbf{R}^{d_n - s_{1n}}\}$. We are going to show that under Conditions 1 and 2, there exists a strict local minimizer $\widehat{\boldsymbol{\beta}} \in \mathcal{N}_0$ of $\widetilde{Q}_n(\boldsymbol{\beta})$ with asymptotic probability one.

For a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, let $\bar{p}'_{\lambda_n}(\boldsymbol{\beta})$ be a vector of the same length whose jth component is $p'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(\beta_j)$, $j = 1, \ldots, d_n$. By Lv and Fan (2009), the sufficient conditions for $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T \in \mathbf{R}^{d_n}$ with $\widehat{\boldsymbol{\beta}}_1 \in \mathbf{R}^{s_{1n}}$ being a strict local minimizer of $\widetilde{Q}_n(\boldsymbol{\beta})$ are

$$(27) \qquad -\mathbf{X}_1^T \widetilde{\mathbf{P}}_z(\mathbf{y} - \mathbf{X}_1 \widehat{\boldsymbol{\beta}}_1) + n \bar{p}'_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1) = 0,$$

$$(28) \qquad \|\mathbf{v}_2\|_\infty < n p'_{\lambda_n}(0+),$$

$$(29) \qquad \Lambda_{\min}(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1) > -n p''_{\lambda_n}(|\widehat{\beta}_j|), \qquad j = 1, \ldots, s_{1n},$$

where $\mathbf{v}_2 = \mathbf{X}_2^T \widetilde{\mathbf{P}}_z(\mathbf{y} - \mathbf{X}_1 \widehat{\boldsymbol{\beta}}_1)$. So we only need to show that with probability tending to 1, there exists a $\widehat{\boldsymbol{\beta}} \in \mathcal{N}_0$ satisfying conditions (27)–(29).

We first consider (27). Since $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_{0,1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, equation (27) can be rewritten as

$$(30) \quad \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{0,1} = (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \mathbf{X}_1^T \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) - n(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \bar{p}'_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1).$$

Define a vector-valued continuous function

$$g(\boldsymbol{\beta}_1) = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_{0,1} - (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \mathbf{X}_1^T \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) + n(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \bar{p}'_{\lambda_n}(\boldsymbol{\beta}_1)$$

with $\boldsymbol{\beta}_1 \in \mathbf{R}^{s_{1n}}$. It suffices to show that with probability tending to 1, there exists $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T \in \mathcal{N}_0$ such that $g(\widehat{\boldsymbol{\beta}}_1) = 0$. To this end, first note that

$$(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \mathbf{X}_1^T \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) \sim N\big(0, (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{P}_z^{-1} \widetilde{\mathbf{P}}_z \mathbf{X}_1 (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\big).$$

By Condition 2(B), the matrix $c_1 \widetilde{\mathbf{P}}_z - \widetilde{\mathbf{P}}_z \mathbf{P}_z^{-1} \widetilde{\mathbf{P}}_z = \widetilde{\mathbf{P}}_z \mathbf{Z}(c_1 \mathcal{M} - \sigma^{-2}\mathcal{G})\mathbf{Z}^T \widetilde{\mathbf{P}}_z \geq 0$, where $A \geq 0$ means the matrix A is positive semi-definite. Therefore,

$$(31) \qquad \mathbf{V} \equiv (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{P}_z^{-1} \widetilde{\mathbf{P}}_z \mathbf{X}_1 (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \leq c_1 (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}.$$

Thus, the jth diagonal component of matrix $\mathbf{V}$ in (31) is bounded from above by the jth diagonal component of $c_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}$. Further note that by Condition 2(B), $\widetilde{\mathbf{P}}_z^{-1} - c_1(\log n)\mathbf{P}_z^{-1} \leq \mathbf{Z}(\mathcal{M} - c_1 \frac{(\log n)}{\sigma^2}\mathcal{G})\mathbf{Z}^T \leq 0$. Recall that by linear algebra, if two positive definite matrices A and B satisfy $A \geq B$, then it follows from the Woodbury formula that $A^{-1} \leq B^{-1}$. Thus, $(c_1 \log n)\widetilde{\mathbf{P}}_z \geq \mathbf{P}_z$

and $(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1} \leq (c_1 \log n)(\mathbf{X}_1^T \mathbf{P}_z \mathbf{X}_1)^{-1}$. So by Condition 2(C), the diagonal components of $\mathbf{V}$ in (31) are bounded from above by $O(n^{-\theta}(\log n))$. This indicates that the variance of each component of the normal random vector $(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})$ is bounded from above by $O(n^{-\theta}(\log n))$. Hence, by Condition 2(C),

$$
\begin{aligned}
\|(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1 \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\|_\infty &= O_p\big(n^{-\theta/2}\sqrt{(\log n)(\log s_{1n})}\big) \\
&= o_p\big(n^{-\tau}(\log n)\big).
\end{aligned}
\tag{32}
$$

Next, by Condition 2(A), for any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_n})^T \in \mathcal{N}_0$ and large enough n, we can obtain that

$$
|\beta_j| \geq |\beta_{0,j}| - |\beta_{0,j} - \beta_j| \geq a_n/2, \qquad j = 1, \ldots, s_{1n}.
\tag{33}
$$

Since $p'_{\lambda_n}(x)$ is a decreasing function in $(0, \infty)$, we have $\|\bar{p}'_{\lambda_n}(\boldsymbol{\beta}_1)\|_\infty \leq p'_{\lambda_n}(a_n/2)$. This together with Condition 2(C) ensures that

$$
\begin{aligned}
\|(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\bar{p}'_{\lambda_n}(\boldsymbol{\beta}_1)\|_\infty &\leq \|(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\|_\infty \|\bar{p}'_{\lambda_n}(\boldsymbol{\beta}_1)\|_\infty \\
&\leq o\big(n^{-\tau-1}(\log n)\big).
\end{aligned}
\tag{34}
$$

Combining (32) and (34) ensures that with probability tending to 1, if n is large enough,

$$
\|(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1 \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) + n(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\bar{p}'_\lambda(\boldsymbol{\beta}_1)\|_\infty < n^{-\tau}(\log n).
$$

Applying Miranda's existence theorem [Vrahatis (1989)] to the function $g(\boldsymbol{\beta}_1)$ ensures that there exists a vector $\widehat{\boldsymbol{\beta}}_1 \in \mathbf{R}^{s_{1n}}$ satisfying $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{0,1}\|_\infty < n^{-\tau}\log n$ such that $g(\widehat{\boldsymbol{\beta}}_1) = 0$.

Now we prove that the solution to (27) satisfies (28). Plugging $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_{0,1} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ into $\mathbf{v}$ in (28) and by (30), we obtain that

$$
\mathbf{v}_2 = \mathbf{X}_2^T \widetilde{\mathbf{P}}_z \mathbf{X}_1(\boldsymbol{\beta}_{0,1} - \widehat{\boldsymbol{\beta}}_1) + \mathbf{X}_2^T \widetilde{\mathbf{P}}_z(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) = \mathbf{v}_{2,1} + \mathbf{v}_{2,2},
$$

where $\mathbf{v}_{2,1} = [-\mathbf{X}_2^T \widetilde{\mathbf{P}}_z \mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z + \mathbf{X}_2^T \widetilde{\mathbf{P}}_z](\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})$ and $\mathbf{v}_{2,2} = \mathbf{X}_2^T \widetilde{\mathbf{P}}_z \times \mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\bar{p}_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1)$. Since $(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) \sim N(0, \mathbf{P}_z^{-1})$, it is easy to see that $\mathbf{v}_{2,1}$ has normal distribution with mean 0 and variance

$$
\mathbf{X}_2^T(\mathbf{I} - \widetilde{\mathbf{P}}_z \mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T)\widetilde{\mathbf{P}}_z \mathbf{P}_z^{-1}\widetilde{\mathbf{P}}_z(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z)\mathbf{X}_2.
$$

Since $\mathbf{P}_z^{-1} \leq c_1 \widetilde{\mathbf{P}}_z^{-1}$, $\mathbf{I} - \widetilde{\mathbf{P}}_z^{1/2}\mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z^{1/2}$ is a projection matrix, and $\widetilde{\mathbf{P}}_z$ has eigenvalues less than 1, it follows that for the unit vector $\mathbf{e}_k$,

$$
\begin{aligned}
\mathbf{e}_k^T \operatorname{var}(\mathbf{v}_{2,1})\mathbf{e}_k &\leq c_1 \mathbf{e}_k^T \mathbf{X}_2^T(\widetilde{\mathbf{P}}_z - \widetilde{\mathbf{P}}_z \mathbf{X}_1(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\mathbf{X}_1^T \widetilde{\mathbf{P}}_z)\mathbf{X}_2 \mathbf{e}_k \\
&\leq c_1 \mathbf{e}_k^T \mathbf{X}_2^T \widetilde{\mathbf{P}}_z \mathbf{X}_2 \mathbf{e}_k \leq \mathbf{e}_k^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{e}_k = c_1 n,
\end{aligned}
$$

where in the the last step, each column of $\mathbf{X}$ is standardized to have L_2-norm $\sqrt{n}$. Thus the diagonal elements of the covariance matrix of $\mathbf{v}_{1,2}$ are bounded from above by $c_1 n$. Therefore, for some large enough constant $C > 0$,

$$P(\|\mathbf{v}_{2,1}\|_\infty \geq \sqrt{2Cn \log d_n}) \leq (d_n - s_{1n}) P(|N(0, c_1 n)| \geq \sqrt{2Cn \log d_n})$$

$$= (d_n - s_{1n}) \exp(-c_1^{-1} C \log d_n) \to 0.$$

Thus, it follows from the assumption $\log d_n = o(n\lambda_n^2)$ that

$$\|\mathbf{v}_{2,1}\|_\infty = O_p(\sqrt{n \log d_n}) = o_p(np'_{\lambda_n}(0+)).$$

Moreover, by Conditions 2(B) and (C),

$$\|\mathbf{v}_{2,2}\|_\infty \leq n\|\mathbf{X}_2^T \widetilde{\mathbf{P}}_z \mathbf{X}_1 (\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1)^{-1}\|_\infty p'_{\lambda_n}(a_n/2) < np'_{\lambda_n}(0+).$$

Therefore inequality (28) holds with probability tending to 1 as $n \to \infty$.

Finally we prove that $\widehat{\boldsymbol{\beta}} \in \mathcal{N}_0$ satisfying (27) and (28) also makes (29) hold with probability tending to 1. By (33) and Condition 2(A),

$$0 \leq -np''_{\lambda_n}(|\widehat{\beta}_j|) \leq -n \sup_{t \geq a_n/2} p''_{\lambda_n}(t) = o(n^{2\tau}).$$

On the other hand, by Condition 2(C), $\Lambda_{\min}(\mathbf{X}_1^T \widetilde{\mathbf{P}}_z \mathbf{X}_1) \geq c_0 n^\theta$. Since $\theta > 2\tau$, inequality (34) holds with probability tending to 1 as $n \to \infty$.

Combing the above results, we have shown that with probability tending to 1 as $n \to \infty$, there exists $\widehat{\boldsymbol{\beta}} \in \mathcal{N}_0$ which is a strict local minimizer of $\widetilde{Q}_n(\boldsymbol{\beta})$. This completes the proof.

6.2. *Proof of Theorem* 2.   Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_N^T)^T \in \mathbf{R}^{q_n N}$ with $\boldsymbol{\gamma}_j^T = (\gamma_{j1}, \ldots, \gamma_{jq_n})$ be a $\mathbf{R}^{Nq_n}$-vector satisfying $\mathfrak{M}(\boldsymbol{\gamma}) = \mathfrak{M}_0$. Define $\mathbf{u}(\boldsymbol{\gamma}) = (\mathbf{u}_1^T, \ldots, \mathbf{u}_N^T)^T \in \mathbf{R}^{Nq_n}$ with $\mathbf{u}_j = (u_{j1}, \ldots, u_{jq_n})^T$, where for $j = 1, \ldots, N$,

$$(35) \qquad \lambda_n u_{jk} = p'_{\lambda_n}(\gamma_{\cdot k})\gamma_{jk}/\gamma_{\cdot k} \qquad \text{if } k \in \mathfrak{M}(\boldsymbol{\gamma})$$

and $\lambda_n u_{jk} = 0$ if $k \notin \mathfrak{M}(\boldsymbol{\gamma})$. Here, $\gamma_{\cdot k} = \{\sum_{j=1}^N \gamma_{jk}^2\}^{1/2}$. Let $\widetilde{\boldsymbol{\gamma}}^*$ be the oracle-assisted estimate defined in (19). By Lv and Fan (2009), the sufficient conditions for $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0^c} = \mathbf{0}$ being a strict local minimizer of (18) are

$$(36) \qquad \widetilde{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^* - \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0} = n\lambda_n \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{u}(\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}),$$

$$(37) \qquad \left(\sum_{j=1}^N w_{jk}^2\right)^{1/2} < np'_{\lambda_n}(0+), \qquad k \in \mathfrak{M}_0^c,$$

$$(38) \qquad \Lambda_{\min}(\widetilde{\mathbf{T}}_{11}) > n\Lambda_{\max}\left(-\frac{\partial^2}{\partial\boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^2}\left(\sum_{j=1}^{q_n} p_{\lambda_n}(\gamma_{\cdot k})\right)\right),$$

where $\mathbf{w}(\boldsymbol{\gamma}) = (\mathbf{w}_1^T, \ldots, \mathbf{w}_N^T)^T \in \mathbf{R}^{Nq_n}$ with $\mathbf{w}_j = (w_{j1}, \ldots, w_{jq_n})^T$, and

$$(39) \qquad \mathbf{w}(\boldsymbol{\gamma}) = \mathbf{Z}^T \mathbf{P}_x (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) - \mathcal{M}^{-1}\boldsymbol{\gamma}.$$

We will show that, under Conditions 1 and 3, conditions (36)–(38) above are satisfied with probability tending to 1 in a small neighborhood of $\widetilde{\boldsymbol{\gamma}}^*$.

In general, it is not always guaranteed that (36) has a solution. We first show that under Condition 3, there exists a vector $\widehat{\boldsymbol{\gamma}}^*$ with $\mathfrak{M}(\widehat{\boldsymbol{\gamma}}^*) = \mathfrak{M}_0$ such that $\widehat{\boldsymbol{\gamma}}^*_{\overline{\mathfrak{M}}_0}$ makes (36) hold. To this end, we constrain the objective function $\widetilde{Q}_n^*(\boldsymbol{\gamma})$ defined in (18) on the (Ns_{n2})-dimensional subspace $\mathcal{B} = \{\boldsymbol{\gamma} \in \mathbf{R}^{q_n N} : \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0^c} = 0\}$ of $\mathbf{R}^{q_n N}$. Next define

$$\mathcal{N}_1 = \left\{ \boldsymbol{\gamma} \in \mathcal{B} : \max_{k \in \mathfrak{M}_0} \left\{ \sum_{j=1}^N (\gamma_{jk} - \widetilde{\gamma}_{jk}^*)^2 \right\}^{1/2} \le \sqrt{N} n^{-\delta} \right\}.$$

For any $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\gamma}_{11}, \ldots, \widetilde{\gamma}_{1q_n}, \ldots, \gamma_{N1}, \ldots, \gamma_{Nq_n})^T \in \mathcal{N}_1$ and $k \in \mathfrak{M}_0$, we have

$$\|\widetilde{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}^*\|_\infty \le \max_{k \in \mathfrak{M}_0} \left\{ \sum_{j=1}^N (\gamma_{jk} - \widetilde{\gamma}_{jk}^*)^2 \right\}^{1/2} \le \sqrt{N} n^{-\delta} \quad \text{and}$$

$$(40) \qquad \widetilde{\gamma}_{\cdot k}^* = \left\{ \sum_{j=1}^N (\widetilde{\gamma}_{jk}^*)^2 \right\}^{1/2} \le \left\{ \sum_{j=1}^N (\widetilde{\gamma}_{jk}^* - \widetilde{\gamma}_{jk})^2 \right\}^{1/2} + \left\{ \sum_{j=1}^N (\widetilde{\gamma}_{jk})^2 \right\}^{1/2}$$

$$\le \sqrt{N} n^{-\delta} + \widetilde{\gamma}_{\cdot k}.$$

Note that by Condition 3(C), we have $\widetilde{\mathbf{T}}_{11}^{-1} \ge \mathbf{T}_{11}^{-1}$. Thus it can be derived using linear algebra and the definitions of $\widetilde{\gamma}_{\cdot k}^*$ and $\gamma_{\cdot k}^*$ that $\widetilde{\gamma}_{\cdot k}^* \ge \gamma_{\cdot k}^*$. Since we condition on the event Ω^* in (20), it is seen that for large enough n,

$$(41) \qquad \widetilde{\gamma}_{\cdot k} \ge \widetilde{\gamma}_{\cdot k}^* - \sqrt{N} n^{-\delta} \ge \gamma_{\cdot k}^* - \sqrt{N} n^{-\delta} > \sqrt{N} b_0^*/2$$

for $k \in \mathfrak{M}_0$ and $\widetilde{\boldsymbol{\gamma}} \in \mathcal{N}_1$. Thus, in view of the definition of $\mathbf{u}(\boldsymbol{\gamma})$ in (35), for $k \in \mathfrak{M}_0$, we have

$$\|\lambda_n \mathbf{u}(\widetilde{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0})\|_\infty \le \max_{k \in \mathfrak{M}_0} p'_{\lambda_n}(\widetilde{\gamma}_{\cdot k}) \le p'_{\lambda_n}(\sqrt{N} b_0^*/2),$$

where in the last step, $p'_{\lambda_n}(t)$ is decreasing in $t \in (0, \infty)$ due to the concavity of $p_{\lambda_n}(t)$. This together with (21) in Condition 3 ensures

$$(42) \qquad \|n\lambda_n \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{u}(\widetilde{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0})\|_\infty \le n\|\widetilde{\mathbf{T}}_{11}^{-1}\|_\infty p'_{\lambda_n}(\sqrt{N} b_0^*/2) \le \sqrt{N} n^{-\delta}.$$

Now define the vector-valued continuous function $\boldsymbol{\Psi}(\boldsymbol{\xi}) = \boldsymbol{\xi} - \widetilde{\boldsymbol{\gamma}}^*_{\overline{\mathfrak{M}}_0} - n\lambda_n \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{u}(\boldsymbol{\xi})$, with $\boldsymbol{\xi}$ a $\mathbf{R}^{Ns_{2n}}$-vector. Combining (40) and (42) and applying Miranda's existence theorem [Vrahatis (1989)] to the function $\boldsymbol{\Psi}(\boldsymbol{\xi})$, we conclude that there exists $\widehat{\boldsymbol{\gamma}}^* \in \mathcal{N}_1$ such that $\widehat{\boldsymbol{\gamma}}^*_{\overline{\mathfrak{M}}_0}$ is a solution to equation (36).

We next show that $\widehat{\boldsymbol{\gamma}}^*$ defined above indeed satisfies (38). Note that for any vector $\mathbf{x} \neq \mathbf{0}$,

$$(43) \qquad \frac{\partial^2}{\partial \mathbf{x}^2} p_{\lambda_n}(\|\mathbf{x}\|_2) = p_{\lambda_n}''(\|\mathbf{x}\|_2) \frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|_2} + p_{\lambda_n}'(\|\mathbf{x}\|_2)\left(\frac{1}{\|\mathbf{x}\|_2} - \frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|_2^3}\right).$$

Since $-p_{\lambda_n}'(t) \leq 0$ and $-p_{\lambda_n}''(t) \geq 0$ for $t \in (0, \infty)$, we have

$$\begin{aligned}
\Lambda_{\max}\left(-\frac{\partial^2}{\partial \mathbf{x}^2} p_{\lambda_n}(\|\mathbf{x}\|_2)\right) &\leq -p_{\lambda_n}''(\|\mathbf{x}\|_2) + \frac{p_{\lambda_n}'(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2} - \frac{p_{\lambda_n}'(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2} \\
&= -p_{\lambda_n}''(\|\mathbf{x}\|_2).
\end{aligned}$$

Since $\widehat{\boldsymbol{\gamma}}^*_{\overline{\mathfrak{M}}_0} \in \mathcal{N}_1$, by (41) we have $\widehat{\gamma}_{\cdot k}^* > \sqrt{N}b_0^*/2$ for $k \in \mathfrak{M}_0$. It follows from the above inequality and Condition 3(B) that with probability tending to 1, the maximum eigenvalue of the matrix $-\frac{\partial^2}{\partial \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^2}(\sum_{j=1}^{q_n} p_{\lambda_n}(\widehat{\gamma}_{\cdot k}^*))$ is less than

$$\max_{j \in \mathfrak{M}_0}\left(-p_{\lambda_n}''(\widehat{\gamma}_{\cdot j}^*)\right) = o(N^{-1}) = o(m_n/n).$$

Further, by Condition 3(A), $\frac{1}{n}\Lambda_{\min}(\widetilde{\mathbf{T}}_{11}) = \frac{1}{n}\Lambda_{\min}(\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0}) \geq c_3 \frac{m_n}{n}$. Thus the maximum eigenvalue of the matrix $-\frac{\partial^2}{\partial \boldsymbol{\gamma}_{\overline{\mathfrak{M}}_0}^2}(\sum_{j=1}^{q_n} p_{\lambda_n}(\widehat{\gamma}_{\cdot k}^*))$ is less than $n^{-1}\Lambda_{\min}(\widetilde{\mathbf{T}}_{11})$ with asymptotic probability 1, and (38) holds for $\widehat{\boldsymbol{\gamma}}^*$.

It remains to show that $\widehat{\boldsymbol{\gamma}}^*$ satisfies (37). Let $\widehat{\mathbf{v}} = \widehat{\boldsymbol{\gamma}}^* - \widetilde{\boldsymbol{\gamma}}^*$. Since $\widehat{\boldsymbol{\gamma}}^*$ is a solution to (36), we have $\widehat{\mathbf{v}} = n\lambda_n \widetilde{\mathbf{T}}_{11}^{-1}\mathbf{u}(\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0})$. In view of (39), we have

$$(44) \qquad \begin{aligned}
\mathbf{w}(\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0^c}^*) &= (\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T - \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T)\mathbf{P}_x\mathbf{y} + \widetilde{\mathbf{T}}_{12}^T \widehat{\mathbf{v}}_{\overline{\mathfrak{M}}_0} \\
&= (\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T - \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T)\mathbf{P}_x(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) + \widetilde{\mathbf{T}}_{12}^T \widehat{\mathbf{v}}_{\overline{\mathfrak{M}}_0} \\
&\equiv \widetilde{\mathbf{w}}_1 + \widetilde{\mathbf{w}}_2.
\end{aligned}$$

Since $\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \sim N(0, \mathbf{P}_z^{-1})$, we obtain that $\widetilde{\mathbf{w}}_1 \sim N(0, \mathbf{H})$ with

$$\mathbf{H} = (\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T - \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T)\mathbf{P}_x\mathbf{P}_z^{-1}\mathbf{P}_x(\mathbf{Z}_{\overline{\mathfrak{M}}_0^c} - \mathbf{Z}_{\overline{\mathfrak{M}}_0}\widetilde{\mathbf{T}}_{11}^{-1}\widetilde{\mathbf{T}}_{12}).$$

Note that $\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}$ is a block diagonal matrix, and the ith block matrix has size $n_i \times (q_n - s_{2n})$. By Condition 3(A), it is easy to see that $\Lambda_{\max}(\mathbf{Z}\mathcal{G}\mathbf{Z}^T) \leq \max_{1 \leq i \leq N} \Lambda_{\max}(\mathbf{Z}_i G\mathbf{Z}_i^T) \leq c_1 s_{2n}$. Thus, $\mathbf{P}_x\mathbf{P}_z^{-1}\mathbf{P}_x = \mathbf{P}_x(\sigma^2\mathbf{I} + \mathbf{Z}\mathcal{G}\mathbf{Z}^T)\mathbf{P}_x \leq (\sigma^2 + c_1 s_{2n})\mathbf{P}_x^2 = (\sigma^2 + c_1 s_{2n})\mathbf{P}_x$. Further, it follows from $\widetilde{\mathbf{T}}_{12} = \mathbf{T}_{12}$ and

$\mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \leq \widetilde{\mathbf{T}}_{11}$ that

$$\mathbf{H} \leq (\sigma^2 + c_1 s_{2n})(\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T - \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T) \mathbf{P}_x (\mathbf{Z}_{\overline{\mathfrak{M}}_0^c} - \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1} \widetilde{\mathbf{T}}_{12})$$

$$= (\sigma^2 + c_1 s_{2n})$$

$$\times (\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c} + \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{Z}_{\overline{\mathfrak{M}}_0}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1} \widetilde{\mathbf{T}}_{12} - 2 \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1} \widetilde{\mathbf{T}}_{12})$$

$$\leq (\sigma^2 + c_1 s_{2n})(\mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c} - \widetilde{\mathbf{T}}_{12}^T \widetilde{\mathbf{T}}_{11}^{-1} \widetilde{\mathbf{T}}_{12}) \leq (\sigma^2 + c_1 s_{2n}) \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}.$$

Thus, the ith diagonal element of $\mathbf{H}$ is bounded from above by the ith diagonal element of $(\sigma^2 + c_1 s_{2n}) \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0^c}$, and is thus bounded by $\widetilde{c}_1 s_{2n} m_n$ with $\widetilde{c}_1$ some positive constant. Therefore by the normality of $\widetilde{\mathbf{w}}_1$ we have

$$P(\|\widetilde{\mathbf{w}}_1\|_\infty \geq \{2\widetilde{c}_1 s_{2n} m_n \log(N(q_n - s_{2n}))\}^{1/2})$$

$$\leq N(q_n - s_{2n}) P(|N(0, \widetilde{c}_1 s_{2n} m_n)| \geq \{2\widetilde{c}_1 s_{2n} m_n \log(N(q_n - s_{2n}))\}^{1/2})$$

$$= O((\log(N(q_n - s_{2n})))^{-1/2}) = o(1).$$

Therefore, $\|\widetilde{\mathbf{w}}_1\|_\infty = o_p(\{s_{2n} m_n \log(N(q_n - s_{2n})\}^{1/2}) = o_p(nN^{-1/2}\lambda_n)$ and

$$(45) \qquad \max_{j > s_{2n}} \left\{ \sum_{k=1}^{N} \widetilde{w}_{1,jk}^2 \right\}^{1/2} \leq \sqrt{N} \|\widetilde{\mathbf{w}}_1\|_\infty = o_p(n\lambda_n) = o_p(1) n p'_{\lambda_n}(0+),$$

where $\widetilde{w}_{1,jk}$ is the $((j-1)q_n + k)$th element of Nq_n-vector $\widetilde{\mathbf{w}}_1$.

Now we consider $\widetilde{\mathbf{w}}_2$. Define $\widetilde{\mathbf{Z}}_j$ as the submatrix of $\mathbf{Z}$ formed by columns corresponding to the jth random effect. Then, for each $j = s_{2n} + 1, \ldots, q_n$, by Condition 3(A) we obtain that

$$\left\{ \sum_{k=1}^{N} \widetilde{w}_{2,jk}^2 \right\}^{1/2} = n\lambda_n \|\widetilde{\mathbf{Z}}_j^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1} \mathbf{u}(\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^*)\|_2$$

$$\leq n \|\lambda_n \mathbf{u}(\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^*)\|_2 \|\widetilde{\mathbf{Z}}_j^T \mathbf{P}_x \mathbf{Z}_{\overline{\mathfrak{M}}_0} \widetilde{\mathbf{T}}_{11}^{-1}\|_2,$$

where $\widetilde{w}_{2,jk}$ is the $((j-1)q_n + k)$th element of Nq_n-vector $\widetilde{\mathbf{w}}_2$. Since $\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^* \in \mathcal{N}_1$, by (35), (41) and the decreasing property of $p'_{\lambda_n}(\cdot)$ we have $\|\lambda_n \mathbf{u}(\widehat{\boldsymbol{\gamma}}_{\overline{\mathfrak{M}}_0}^*)\|_2 \leq p'_{\lambda_n}(\sqrt{N} b_0^*/2)$. By (22) in Condition 3(A),

$$\max_{j \geq s_{2n}+1} \left\{ \sum_{k=1}^{N} \widetilde{w}_{2,jk}^2 \right\}^{1/2} < n p'_{\lambda_n}(0+).$$

Combing the above result for $\widetilde{\mathbf{w}}_2$ with (44) and (45), we have shown that (37) holds with asymptotic probability one. This completes the proof.

SUPPLEMENTARY MATERIAL

Supplement to "Variable selection in linear mixed effects models" (DOI: 10.1214/12-AOS1028SUPP; .pdf). We included additional simulation examples and technical proofs omitted from the main text: simulation Examples A.1–A.3, and technical proofs of Lemmas 1–3 and Proposition 1.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csaki, eds.) 267–281. Akad. Kiadó, Budapest. MR0483125

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077. MR2758494

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA. MR0418321

CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. MR2025100

DUDOIT, S., FRIDLYAND, J. and SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97** 77–87. MR1963389

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–451.

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, Y. and LI, R. (2012). Supplement to "Variable selection in linear mixed effects models." DOI:10.1214/12-AOS1028SUPP.

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London. MR1385925

HUANG, J. Z., WU, C. O. and ZHOU, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503. MR2829018

KASLOW, R. A., OSTROW, D. G., DETELS, R., PHAIR, J. P., POLK, B. F. and RINALDO, C. R. (1987). The multicenter AIDS cohort study: Rationale, organization and selected characteristics of the participants. *American Journal Epidemiology* **126** 310–318.

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

LIANG, H., WU, H. and ZOU, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95** 773–778. MR2443190

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049

LIU, Y. and WU, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *J. Comput. Graph. Statist.* **16** 782–798. MR2412482

LONGFORD, N. T. (1993). *Random Coefficient Models. Oxford Statistical Science Series* **11**. Oxford Univ. Press, New York. MR1271143

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554. MR0319325

PU, W. and NIU, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *J. Multivariate Anal.* **97** 733–758. MR2236499

QU, A. and LI, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* **62** 379–391. MR2227487

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

VAIDA, F. and BLANCHARD, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370. MR2201364

VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer, New York. MR1880596

VRAHATIS, M. N. (1989). A short proof and a generalization of Miranda's existence theorem. *Proc. Amer. Math. Soc.* **107** 701–703. MR0993760

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, Y., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. MR2656055

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36** 1509–1566. MR2435443

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470

DEPARTMENT OF INFORMATION
   AND OPERATIONS MANAGEMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: fanyingy@marshall.usc.edu

DEPARTMENT OF STATISTICS
   AND THE METHODOLOGY CENTER
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: rli@stat.psu.edu