Inference for High-Dimensional Linear Mixed-Effects Models: A Quasi-Likelihood Approach

Sai Li, T. Tony Cai & Hongzhe Li

View supplementary material

Published online: 20 Apr 2021.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

Inference for High-Dimensional Linear Mixed-Effects Models: A Quasi-Likelihood Approach

Sai Li[a], T. Tony Cai[b], and Hongzhe Li[a]

[a]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [b]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Linear mixed-effects models are widely used in analyzing clustered or repeated measures data. We propose a quasi-likelihood approach for estimation and inference of the unknown parameters in linear mixed-effects models with high-dimensional fixed effects. The proposed method is applicable to general settings where the dimension of the random effects and the cluster sizes are possibly large. Regarding the fixed effects, we provide rate optimal estimators and valid inference procedures that do not rely on the structural information of the variance components. We also study the estimation of variance components with high-dimensional fixed effects in general settings. The algorithms are easy to implement and computationally fast. The proposed methods are assessed in various simulation settings and are applied to a real study regarding the associations between body mass index and genetic polymorphic markers in a heterogeneous stock mice population.

1. Introduction

The results of scientific experiments are often subject to environmental effects as experimental units can be grouped and settled in diverse environments, where the observations within the same group can be dependent as a cluster. Clustered data commonly arise in many fields, such as biology, genetics, and economics. Linear mixed-effects models provide a flexible tool for analyzing such clustered data, which include repeated measures data, longitudinal data, and multilevel data (Pinheiro and Bates 2000; Goldstein 2011). The linear mixed-effects models incorporate both the fixed and random effects, where the random effects induce correlations among the observations within each cluster and accommodate the cluster structure. In many genomic and economic studies, the dimension of the covariates can be large and possibly much larger than the sample size. A variety of statistical models and approaches have been proposed and studied for analyzing high-dimensional data. However, most of them are restricted to dealing with independent observations, such as linear models and generalized linear models. Statistical inference for high-dimensional linear mixed-effects models remains to be a challenging problem. In this work, we consider estimation and inference of unknown parameters in high-dimensional mixed-effects models.

For ease of presentation, we use the setting for clustered data to present a linear mixed-effects model. For repeated measurement data, the repeated measures form a cluster. Let $i = 1, \ldots, n$ be the cluster indices. For the ith cluster, we have a response vector $y_i \in \mathbb{R}^{m_i}$, a design matrix for the fixed effects $X^i \in \mathbb{R}^{m_i \times p}$,

and a design matrix for the random effects $Z^i \in \mathbb{R}^{m_i \times q}$, where m_i is the size of the ith cluster. A linear mixed-effects model (Laird and Ware 1982) can be written as

$$y_i = X^i \beta^* + Z^i \gamma_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ is the vector of the fixed effects, $\gamma_i \in \mathbb{R}^q$ is the vector of the random effects of the ith cluster, and $\epsilon_i \in \mathbb{R}^{m_i}$ is the noise vector of the ith cluster. For $i = 1, \ldots, n$, we assume γ_i and ϵ_i are independently distributed with mean zero and variance $\Psi \in \mathbb{R}^{q \times q}$ and $\sigma_e^2 I_{m_i}$, respectively. Detailed assumptions are given in Sections 2 and 3.

Much existing literature on linear mixed-effects models assumes that the number of random effects q and cluster sizes m_i are fixed. Without special emphasis, we say a fixed-dimensional setting if p, q, and $\{m_i\}_{i=1}^n$ are all fixed numbers, and a high-dimensional setting if p is large and possibly much larger than N, where $N = \sum_{i=1}^n m_i$ is the total sample size. We refer to γ_i and ϵ_i as the random components.

1.1. Related Literature

In the fixed-dimensional setting, many methods have been proposed to jointly estimate the fixed effects and variance parameters. We refer to Gumedze and Dunne (2011) for a comprehensive review. Among them, the maximum likelihood estimators (MLEs) and restricted MLEs are most popular for estimation and inference in linear mixed-effects models. Restricted MLEs can produce unbiased estimators of the variance components in the low-dimensional

setting but it is not applicable in the high-dimensional setting. Furthermore, these likelihood-based estimators rely heavily on the normality assumptions of the random components. Computationally, maximizing the likelihood can generally lead to a nonconvex optimization problem that typically has multiple local maxima. Hence, the performance of likelihood-based methods lacks of guarantees in real applications.

As an alternative, Sun, Zhang, and Tong (2007) proposed moment estimators of the fixed effects and variance parameters for a random effect varying-coefficient model. Peng and Lu (2012) considered such moment estimators for fixed-dimensional linear mixed-effects models. Their proposed estimators have closed-form solutions and are computationally efficient. The consistency and asymptotic normality of these estimators are justified under certain conditions in the fixed-dimensional setting. Ahmn, Zhang, and Lu (2012) proposed another moment-based method for the estimation and selection of the variance components of the random effects in the fixed-dimensional setting. This method works especially well when the number of the random effects is as large as the cluster sizes, that is, $m_1 = \cdots = m_n = q$.

For inference of variance components in the fixed-dimensional setting, the likelihood ratio, score, and Wald tests (Stram and Lee 1994; Lin 1997; Verbeke and Molenberghs 2003; Demidenko 2004) are broadly used. However, when testing the existence of the random effects, the asymptotic distribution of the likelihood ratio is usually a mixture of chi-square distributions (Miller 1977; Self and Liang 1987). Since these methods are based on the MLEs or restricted MLEs as initial estimators, they also suffer from the drawbacks of likelihood-based methods discussed above.

In the high-dimensional setting, the problems are much more challenging. Assuming fixed cluster sizes, Schelldorfer, Bühlmann, and van De Geer (2011) analyzed the rate of convergence for the global maximizer of the ℓ_1-penalized likelihood with fixed designs. As mentioned before, the analysis for the global optimum may not apply to the realizations due to the existence of local maxima. Fan and Li (2012) studied the fixed effects and random effects selection in a high-dimensional linear mixed-effects model when the cluster sizes are balanced, that is, $(\max_i m_i)/(\min_i m_i) < \infty$. The selection consistency requires minimum signal strength conditions regarding the fixed effects and the random effects. Bradic, Claeskens, and Gueuning (2020) considered testing a single coefficient of the fixed effects in the high-dimensional linear mixed-effects models with fixed cluster sizes, fixed number of random effects, and sub-Gaussian designs. The theoretical analyses in all three aforementioned articles require the positive definiteness of the covariance matrix of the random effects. This condition takes prior knowledge on the existence of the random effects and can be hard to fulfill in applications. Moreover, the optimal convergence rate of parameter estimation remains unknown. In fact, estimators of fixed effects in Schelldorfer, Bühlmann, and van De Geer (2011) and Bradic, Claeskens, and Gueuning (2020) may not be rate optimal for estimation according to our analysis. Finally, estimation and inference of the variance components in the high-dimensional setting remain largely unknown.

The problems of estimation and inference of the fixed effects in linear mixed-effects models are related to high-dimensional linear models. Many penalized methods have been proposed for prediction, estimation, and variable selection in high-dimensional linear models (see, e.g., Tibshirani 1996; Fan and Li 2001; Zou 2006; Candes and Tao 2007; Meinshausen and Bühlmann 2010; Zhang 2010). Statistical inference on a low-dimensional component of a high-dimensional regression vector has been considered and studied in linear models and generalized linear models with "debiased" estimators (Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014), and the minimaxity and adaptivity of confidence intervals have been studied in Cai and Guo (2017) and Cai, Guo, and Ma (2020). The idea of debiasing has also been studied and extended to solve other statistical problems, such as statistical inference in Cox models (Fang, Ning, and Liu 2017), simultaneous inference (Dezeure, Bühlmann, and Zhang 2017; Zhang and Cheng 2017), and semisupervised inference (Cai and Guo 2020).

1.2. Our Contributions

In this article, we develop a simple but powerful method for inference of the unknown parameters in high-dimensional linear mixed-effects models. Our method is applicable to the settings where the number of random effects can possibly be large and the cluster sizes can be either fixed or growing, balanced or unbalanced. The proposed method is easy to implement and the optimization in each step is either analytic or convex.

Based on a proxy of the true covariance matrix, we develop a penalized quasi-likelihood approach for fixed effects estimation. The proposed estimator is minimax rate optimal under general conditions. We further develop a debiased estimator for hypothesis testing and construction of confidence intervals for the fixed effects. The proposed estimator does not require normality assumptions or the structural assumptions on the variance components. We further apply the idea of quasi-likelihood to estimate the variance components and prove its optimality under certain conditions.

Our analysis provides a novel insight for understanding and simplifying the linear mixed-effects models by approximating the true unknown covariance matrix of the random components with some simple proxy matrices. In this way, one separates the tasks of estimating the fixed effects and variance components and avoids the nuisance parameters in each optimization step. This improves the computational efficiency and simplifies the theoretical analysis.

1.3. Notation

Throughout the article, we use i to index the ith cluster and k to index the kth observation in each cluster. Let $\mathbf{y}$, $\boldsymbol{\gamma}$, $\boldsymbol{\epsilon}$, and $\mathbf{X}$ be obtained by stacking vectors y_i, γ_i, ϵ_i, and matrices X^i underneath each other, respectively. Let $\mathbf{Z} \in \mathbb{R}^{N \times (nq)}$ be a block diagonal matrix with the ith block being Z^i. Let $\Sigma_\theta^i = Z^i \Psi Z^i + \sigma_e^2 I_{m_i}$ and $\boldsymbol{\Sigma}_\theta \in \mathbb{R}^{N \times N}$ be a block diagonal matrix with the ith block being Σ_θ^i. Let $\Sigma_z^i = (Z^i)^\top Z^i / m_i$ and $\Sigma_{z,x}^i = (Z^i)^\top X^i / m_i$, $i = 1, \ldots, n$. For a random variable $u \in \mathbb{R}$, define its sub-Gaussian norm as $\|u\|_{\psi_2} = \sup_{l \geq 1} l^{-1/2} \mathbb{E}^{1/l}[|u|^l]$. We refer to $\|u\|_{\psi_2, \mathbf{Z}} = \sup_{l \geq 1} l^{-1/2} \mathbb{E}^{1/l}[|u|^l | \mathbf{Z}]$ as the conditional

sub-Gaussian norm of u. For a random vector $U \in \mathbb{R}^{n_0}$, define its sub-Gaussian norm as $\|U\|_{\psi_2} = \sup_{\|v\|_2=1, v \in \mathbb{R}^{n_0}} \|\langle U, v \rangle\|_{\psi_2}$. Define it conditional sub-Gaussian norm as $\|U\|_{\psi_2, \mathbf{Z}} = \sup_{\|v\|_2=1, v \in \mathbb{R}^{n_0}} \|\langle U, v \rangle\|_{\psi_2, \mathbf{Z}}$.

Let $A \in \mathbb{R}^{n_0 \times n_0}$ be a symmetric matrix. $A \succeq 0$ means that A is semipositive definite and $A \succ 0$ means that A is positive definite. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ denote the largest and smallest eigenvalues of A, respectively. Let $\|A\|_2$ denote $\Lambda_{\max}(A)$. Let $\|A\|_F^2 = \mathrm{Tr}(A^\top A)$, where $\mathrm{Tr}(A)$ is the trace of matrix A. Let $c, c_0, c_1, \ldots, C, C_0, C_1, \ldots$ denote some generic positive constants that can vary in different statements.

1.4. Organization of the Article

The rest of the article is organized as follows. Section 2 introduces the idea of quasi-likelihood and a procedure for the fixed effects inference. Section 3 provides a theoretical analysis for the inference procedures proposed in Section 2. Section 4 introduces estimators for the variance components and provides upper and lower bounds. Numerical performance of the proposed methods is investigated in Section 5 in various simulation settings. The proposed methods are applied in Section 6 to analyze a real study on the associations between the body mass index (BMI) and genetic variants in a stock mice population where the cage effect is modeled as a random effect. A discussion is given in Section 7. Proofs and more numerical results are provided in the supplementary materials.

2. Inference for Fixed Effects: The Method

In many applications of the linear mixed-effects models, inference of the fixed effects is of main interest. In this section, we present our method for fixed effects inference and describe its motivations. We assume that the vector of fixed effects β^* is sparse such that $\|\beta^*\|_0 \leq s$ with s unknown. We consider model (1) where p, s, and q can grow and p can be much larger than N. The cluster sizes $\{m_i\}_{i=1}^n$ can be either fixed or grow with n.

2.1. Motivations of the Proposed Method

For fixed effects estimation in model (1), the main challenges are posed by the high-dimensionality of the fixed effects and the clustered structure of the observations. Before developing a new method, it is helpful to understand the new challenges posed by the cluster structures in model (1) in terms of estimation and inference. For this purpose, we study the consequences of misspecifying a linear mixed-effects model as a standard linear model.

Applying Lasso (Tibshirani 1996) directly to the observations generated from (1), we analyze

$$\hat{\beta}^{(lm)} = \underset{b \in \mathbb{R}^p}{\arg \min} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{X}b\|_2^2 + \lambda^{(lm)} \|b\|_1 \right\} \qquad (2)$$

for some tuning parameter $\lambda^{(lm)} > 0$. In a typical analysis of the Lasso, the convergence rate of $\hat{\beta}^{(lm)}$ depends on the restricted isometries of the sample covariance matrix, $\mathbf{X}^\top \mathbf{X}/N$, the sparsity of the true coefficients, and the so-called "empirical process" part of the problem, $\|\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta^*)/N\|_\infty$. It is known

that for linear models with row-wise independent sub-Gaussian $(\mathbf{X}, \mathbf{y})$, the "empirical process" part is of order $\sqrt{\log p / N}$, which gives the optimal convergence rate in ℓ_2-norm. In the following proposition, we study the size of "empirical process" part when the true model is (1).

Proposition 2.1 (The rate of Lasso for linear mixed-effects models). Suppose that the responses y_i are generated with respect to model (1) and each row of $\mathbf{X}$ is independently generated with covariance matrix $\Sigma_{x|z}$ conditioning on $\mathbf{Z}$. Then for any fixed $j \in \{1, \ldots, p\}$,

$$\mathbb{E}\left[\left| \frac{1}{N} \mathbf{X}_{\cdot,j}^\top (\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \right|^2 \Big| \mathbf{Z} \right] = \frac{(\Sigma_{x|z})_{j,j} \sigma_e^2}{N} + \frac{(\Sigma_{x|z})_{j,j} \sum_{i=1}^n m_i \mathrm{Tr}(\Psi \Sigma_z^i)}{N^2}$$
$$+ \frac{\sum_{i=1}^n m_i^2 \|\Psi^{1/2} \mathbb{E}[\Sigma_{z,x}^i | \mathbf{Z}]\|_2^2}{N^2}. \qquad (3)$$

If Ψ is positive definite and $\{\Sigma_z^i\}_{i=1}^n$ have bounded diagonal elements, then the second term on the right-hand side of (3) is $\asymp q/N$. If it further holds that $\mathbb{E}[\Sigma_{z,x}^i | \mathbf{Z}] \neq 0$, that is, $\mathbf{X}$ and $\mathbf{Z}$ are correlated, then the third term can be $\gtrsim \min_{1 \leq i \leq n} m_i / N$. That is, the Lasso may not be rate optimal for clustered data if either q grows, or, $\{m_i\}_{i=1}^n$ grow and $\mathbf{X}$ and $\mathbf{Z}$ are correlated. On the other hand, if the q and m_i's are all constant, it is not hard to prove that the original Lasso is still rate optimal for model (1).

Therefore, proper methods need to be developed for high-dimensional linear mixed-effects models under general conditions on q and $\{m_i\}_{i=1}^n$. The main challenge comes from the correlation among observations induced by the random effects. For the ith block, the covariance of the random components is $\Sigma_{\theta^*}^i = Z^i \Psi (Z^i)^\top + \sigma_e^2 I_{m_i}$, which involves unknown parameters. We consider a proxy of $\Sigma_{\theta^*}^i$ as

$$\Sigma_a^i = a Z^i (Z^i)^\top + I_{m_i}$$

with some predetermined constant $a > 0$. The following proposition shows that this approximation is valid up to some scaling constant. Let $\mathbf{\Sigma}_a \in \mathbb{R}^{N \times N}$ be the block diagonal matrix with the ith block being Σ_a^i.

Proposition 2.2. If Ψ is positive definite, then for any constant $a > 0$,

$$\min \left\{ \frac{1}{\sigma_e^2}, \frac{a}{\Lambda_{\max}(\Psi)} \right\} \mathbf{\Sigma}_a^{-1} \preceq \mathbf{\Sigma}_{\theta^*}^{-1} \preceq \max \left\{ \frac{1}{\sigma_e^2}, \frac{a}{\Lambda_{\min}(\Psi)} \right\} \mathbf{\Sigma}_a^{-1}.$$

Therefore, if Ψ has positive and bounded eigenvalues, $\mathbf{\Sigma}_{\theta^*}^{-1}$ and $\mathbf{\Sigma}_a^{-1}$ are of the same rate and only differ by constants. This property of $\mathbf{\Sigma}_a^{-1}$ is crucial to achieve the general results in this work. A broader class of proxy matrices have been considered in Fan and Li (2012) for variable selection and in Bradic, Claeskens, and Gueuning (2020) for hypothesis testing, which include $\mathbf{\Sigma}_a^{-1}$ as a special case. As reviewed in Section 1.1, afore-mentioned two articles considered relatively restrictive scenarios in terms of group sizes and the dimension of the random effects. It is not clear whether the desired property proved in Proposition 2.2 holds for the general class of proxy matrices.

2.2. The Quasi-Likelihood Approach

We consider a quasi-likelihood approach which replaces $\Sigma_{\theta^*}^i$ with Σ_a^i for some constant $a > 0$ in the likelihood function

for Gaussian mixed-effects models. Specifically, let $\mathbf{X}_a$ and $\mathbf{y}_a$ denote the transformed observations such that $(\mathbf{X}_a, \mathbf{y}_a) = (\boldsymbol{\Sigma}_a^{-1/2}\mathbf{X}, \boldsymbol{\Sigma}_a^{-1/2}\mathbf{y})$.

First, we estimate the fixed effects via the Lasso based on the transformed data. For some fixed $a > 0$, define

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2\text{Tr}(\boldsymbol{\Sigma}_a^{-1})} \|\mathbf{y}_a - \mathbf{X}_a\beta\|_2^2 + \lambda\|\beta\|_1 \right\} \quad (4)$$

for some tuning parameter $\lambda > 0$. The quantity $\text{Tr}(\boldsymbol{\Sigma}_a^{-1})$ can be viewed as the effective sample size in the current problem and its magnitude is studied in Remark 3.1. The choice of a will be studied theoretically in Section 3 and numerically in Section 5.

Given the task of making inference for β_j^*, we propose the following debiased estimator. For $\hat{\beta}$ defined in (4),

$$\hat{\beta}_j^{(db)} = \hat{\beta}_j + \frac{\hat{w}_j^\top(\mathbf{y}_a - \mathbf{X}_a\hat{\beta})}{\hat{w}_j^\top(\mathbf{X}_a)_{.,j}}, \quad (5)$$

where $\hat{w}_j \in \mathbb{R}^N$ can be viewed as a correction score. It can be computed via another Lasso regression (van de Geer et al. 2014; Zhang and Zhang 2014) or via a quadratic optimization (Javanmard and Montanari 2014; Zhang and Zhang 2014). For computational convenience, we consider the Lasso approach based on the transformed data. Define the correction score $\hat{w}_j = (\mathbf{X}_a)_{.,j} - (\mathbf{X}_a)_{.,-j}\hat{\kappa}_j$, where

$$\hat{\kappa}_j = \arg\min_{\kappa_j \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2\text{Tr}(\boldsymbol{\Sigma}_a^{-1})} \left\| (\mathbf{X}_a)_{.,j} - (\mathbf{X}_a)_{.,-j}\kappa_j \right\|_2^2 + \lambda_j\|\kappa_j\|_1 \right\}, \quad (6)$$

for some tuning parameter $\lambda_j > 0$. A two-sided $100 \times (1-\alpha)\%$ confidence interval for β_j^* can be constructed as

$$\hat{\beta}_j^{(db)} \pm z_{\alpha/2}\sqrt{\widehat{V}_j}, \quad (7)$$

where z_τ is the τth quantile of a standard normal distribution and $\widehat{V}_j$ is an estimator of the variance of $\hat{\beta}_j^{(db)}$. We propose to use the following empirical variance estimate

$$\widehat{V}_j = \frac{\sum_{i=1}^n \left[(\hat{w}_j^i)^\top(y_a^i - X_a^i\hat{\beta}) \right]^2}{(\hat{w}_j^\top(\mathbf{X}_a)_{.,j})^2}, \quad (8)$$

where $\hat{\beta}$ is the initial Lasso estimator (4), $\hat{w}_j^i \in \mathbb{R}^{m_i}$ is the ith sub-vector of $\hat{w}_j$ such that $\hat{w}_j = ((\hat{w}_j^1)^\top, \dots, (\hat{w}_j^n)^\top)^\top$, and y_a^i is the ith sub-vector of $\mathbf{y}_a$. The idea of empirical variance estimator has been considered in Bühlmann and van de Geer (2015) to deal with the misspecified linear models. The format of (8) is however different from the one for linear models because it is an average over n groups rather than N observations. In this work, $\widehat{V}_j$ serves as a convenient alternative to the limiting distribution-based variance estimators. In fact, the limiting distribution of $\hat{\beta}_j^{(db)}$ involves nuisance parameters coming from the complicated variance components. By using the empirical residuals of the transformed data, we bypass the estimation of the nuisance parameters.

3. Inference for Fixed Effects: Theoretical Guarantees

In this section, we provide theoretical guarantees for the estimators described in Section 2.2. We first detail our assumptions.

Condition 3.1 (Sub-Gaussian random components). The random noises $\epsilon_{i,k}$, $i = 1, \dots, n$, $k = 1, \dots, m_i$, are independent with mean zero and variance $0 < \sigma_e^2 < K_0 < \infty$. The sub-Gaussian norms of $\epsilon_{i,k}$ are upper bounded by K_0. The random effects $\gamma_i \in \mathbb{R}^q$, $i = 1, \dots, n$, are independent with mean zero and covariance $\Psi \preceq K_1 I_q$ for some positive constant K_1. For $i = 1, \dots, n$, ϵ_i and γ_i are independent of each other and are independent of (X^i, Z^i). The sub-Gaussian norms of $\boldsymbol{\Sigma}_{\theta^*}^{-1/2}(\mathbf{Z}\gamma + \boldsymbol{\epsilon})$ are upper bounded by K_0.

Condition 3.1 assumes sub-Gaussian random components while that classical linear mixed-effects models always assume Gaussian random components (Pinheiro and Bates 2000). Hence, Condition 3.1 is less restrictive and is more robust to model misspecifications than the classical assumptions. In addition, we do not require Ψ to be strictly positive definite. A scenario of singular Ψ is that some components of the random effects do not exists such that some diagonal elements of Ψ are zero.

Regarding the conditions on the designs, the estimation and inference in the linear mixed-effects models are usually conditioning on $\mathbf{Z}$ to maintain the cluster structure. Schelldorfer, Bühlmann, and van De Geer (2011) and Fan and Li (2012) assume both $\mathbf{X}$ and $\mathbf{Z}$ are fixed. Jiang et al. (2016) considered estimation and inference in a misspecified linear model when both $\mathbf{X}$ and $\mathbf{Z}$ are random. Bradic, Claeskens, and Gueuning (2020) assumed $\mathbf{X}$ is sub-Gaussian with mean zero and $\mathbf{Z}$ is fixed, which implies that $\mathbf{X}$ and $\mathbf{Z}$ are independent. In the current work, we consider random designs satisfying the following condition.

Condition 3.2 (Sub-Gaussian $\mathbf{X}$ conditioning on $\mathbf{Z}$). Conditioning on $\mathbf{Z}$, each row of $\mathbf{X}$ is independent with mean zero and covariance matrix $\Sigma_{x|z}$ such that $0 < K_* \leq \Lambda_{\min}(\Sigma_{x|z}) \leq \Lambda_{\max}(\Sigma_{x|z}) \leq K^* < \infty$. Conditioning on $\mathbf{Z}$, the conditional sub-Gaussian norms of $X_{k,.}^i$ are upper bounded by K_0.

In Condition 3.2, we assume sub-Gaussian $\mathbf{X}$ and $\mathbf{Z}$ have mean independence, that is, $\mathbb{E}[\mathbf{X}|\mathbf{Z}] = 0$ for simplicity. This is slightly weaker than assuming $\mathbf{X}$ and $\mathbf{Z}$ are mutually independent and it holds when $\mathbf{Z}$ is deterministic including the random intercept model. In Section 3.4, we study the performance of our proposal when $\mathbb{E}[\mathbf{X}|\mathbf{Z}] \neq 0$.

3.1. Fixed Effects Estimation

In this subsection, we analyze the theoretical performance of (4) under Conditions 3.1 and 3.2. Define

$$\lambda_a^* = \frac{\sqrt{\text{Tr}(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_{\theta^*}\boldsymbol{\Sigma}_a^{-1})\log p}}{\text{Tr}(\boldsymbol{\Sigma}_a^{-1})}.$$

Lemma 3.1 (Fixed effects estimation with quasi-likelihood based Lasso). Assume that Conditions 3.1 and 3.2 hold true. There exists a constant c_1 such that for $\lambda \geq c_1\lambda_a^*$ and $\text{Tr}(\boldsymbol{\Sigma}_a^{-1}) \gg$

$s \log p$, we have with probability at least $1 - 2 \exp(-\log p)$,

$$\left\| \hat{\beta} - \beta^* \right\|_1 \leq C_1 s \lambda, \quad \left\| \hat{\beta} - \beta^* \right\|_2 \leq C_2 \sqrt{s} \lambda, \quad \text{and}$$

$$\frac{1}{\mathrm{Tr}(\Sigma_a^{-1})} \left\| \mathbf{X}_a(\hat{\beta} - \beta^*) \right\|_2^2 \leq C_3 s \lambda^2 \tag{9}$$

for some positive constants C_1, C_2, and C_3. Moreover, for any $a > 0$,

$$\lambda_a^* \leq \sqrt{\frac{(\Lambda_{\max}(\Psi)/a + \sigma_e^2) \log p}{\mathrm{Tr}(\Sigma_a^{-1})}}.$$

Remark 3.1. For any $a \geq 0$,

$$\sum_{i=1}^n \max\{m_i - q, 0\} \leq \mathrm{Tr}(\Sigma_a^{-1}) \leq N.$$

Lemma 3.1 provides upper bounds for the prediction error and the estimation errors in ℓ_1-norm and ℓ_2-norm. By setting a to be a positive constant, the ℓ_2-error of $\hat{\beta}$ is of order $\sqrt{s \log p / \mathrm{Tr}(\Sigma_a^{-1})}$. Remark 3.1 studies the magnitude of the effective sample size $\mathrm{Tr}(\Sigma_a^{-1})$. As pointed out by a reviewer, in the case of equal group sizes and $q/m \leq c_0 < 1$, $\mathrm{Tr}(\Sigma_a^{-1}) \asymp N$, that is, the convergence rates are the same as the rates in linear models. Revoking Proposition 2.1, it shows that $\hat{\beta}$ has a faster convergence rate than $\hat{\beta}^{(lm)}$ in the regime that q grows but remains relatively small to m.

The results of Lemma 3.1 hold for any positive constant a. Different choices of a can affect the constants in the upper bound and the empirical performance of the method. To understand the optimal choice of a, we prove the following remark.

Remark 3.2 (Effect of a). Suppose $\Psi = \eta^* I_q$. For any given $n, p, q, \{m_i\}_{i=1}^n$, $a = \eta^*/\sigma_e^2$ minimizes λ_a^* for $a \in (0, \infty)$. If it further holds that $\eta^* \neq 0$ and $q < \max_{i \leq n} m_i$, then $\lambda_0^* > \lambda_a^*$ for any $a \in (0, \lambda_a^*]$.

Remark 3.2 gives the optimal choice of a for $\Psi = \eta^* I_q$. In this case, setting $a = \eta^*/\sigma_e^2$ minimizes λ_a^* and hence minimizes the upper bound on the estimation errors when other parameters and constants are fixed. The optimal choice of a is intuitive as it mimics the MLE procedure. Furthermore, when the random effects exist and $q < \max_{i \leq n} m_i$, then setting $a = 0$ is strictly worse than the proposed quasi-likelihood approach with $a \in (0, \lambda_a^*]$. We mention that the condition $q < \max_{i \leq n} m_i$ is sufficient but not necessary. This remark sheds lights on the choice of a in general settings as any semipositive definite Ψ can be upper and lower bounded by diagonal matrices. From the optimization perspective, we treat a as a tuning parameter in the optimization (4). In Section 5, we carefully examine the effect of a on estimation accuracy in numerical experiments.

3.2. Rate Optimality of the Proposed Estimator

In this subsection, we study the minimax optimality of proposed estimator for the fixed effects. We consider

$$X_{k,\cdot}^i | \mathbf{Z} \sim_{\text{iid}} N(0, \Sigma_x), \quad \gamma_i \sim_{\text{iid}} N(0, \Psi) \text{ and } \epsilon_{i,k} \sim_{\text{iid}} N(0, \sigma_e^2), \tag{10}$$

$k = 1, \ldots, m_i$, $i = 1, \ldots, n$. Consider the following parameter space

$$\Xi(s, \mathbf{Z}) = \left\{ \mathbf{v} = (\beta^*, \Psi, \sigma_e^2, \Sigma_x, \mathbf{Z}) : \|\beta^*\|_0 \leq s, \right.$$
$$0 < \sigma_e^2 \leq K_0, \ 0 \preceq \Psi \preceq K_1,$$
$$\left. 1/K^* \leq \Lambda_{\min}(\Sigma_x) \leq \Lambda_{\max}(\Sigma_x) \leq K^* < \infty \right\}, \tag{11}$$

where $K^* \geq 1$. We see that (10) and (11) define a special case of Conditions 3.1 and 3.2. We prove the minimax optimal rate of convergence in ℓ_2-norm with respect to $\Xi(s)$.

Theorem 3.1 (Minimax lower bounds for estimating the fixed effects). Suppose that (1) and (10) are true. If $s \leq c \min\{\mathrm{Tr}(\Sigma_a^{-1})/\log p, p^\nu\}$ for $0 < \nu < 1/2$ and $c > 0$, then there exists some constant $c_1 > 0$ such that for any fixed $a > 0$,

$$\inf_{\hat{\beta}} \sup_{\mathbf{v} \in \Xi(s, \mathbf{Z})} \mathbb{P}_{\mathbf{v}} \left(\|\hat{\beta} - \beta^*\|_2^2 \geq c_1 \frac{s \log(p/s^2)}{\mathrm{Tr}(\Sigma_a^{-1})} | \mathbf{Z} \right) \geq \frac{1}{4}.$$

Together with (9), this shows that $\hat{\beta}$ is minimax rate optimal in ℓ_2-error in the parameter space $\Xi(s, \mathbf{Z})$. In the proof, we use the minimax optimality of ℓ_1-penalized MLE, which has $\Sigma_{\theta^*}^{-1}$ as the weighting matrix and use Proposition 2.2 to show the equivalence of MLE and proposed estimator in term of convergence rate.

3.3. Statistical Inference of the Fixed Effects

Debiased estimators can be used for statistical inference of linear combinations of regression coefficients in high-dimensional linear models (Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014). Under certain conditions, the debiased estimators are asymptotically normal and can be used to construct confidence interval with optimal lengths (Cai and Guo 2017). To make inference of β_j^*, we consider the debiased estimator proposed in (5). Let H_j be the support of $(\Sigma_{x|z}^{-1})_{\cdot j}$.

Theorem 3.2 (Asymptotic normality of the debiased estimator). Assume Conditions 3.1 and 3.2. Let $\lambda \wedge \lambda_j \geq c_1 \sqrt{\log p / \mathrm{Tr}(\Sigma_a^{-1})}$ with a large enough constant c_1. For $\hat{\beta}_j^{(db)}$ defined in (5), if $(s \log p)^2 \vee \log n \max_i m_i \ll \mathrm{Tr}(\Sigma_a^{-1}) \Lambda_{\min}(\Sigma_a^{-1/2} \Sigma_{\theta^*} \Sigma_a^{-1/2})$ and $|H_j| \log p \ll \mathrm{Tr}(\Sigma_a^{-1})$, then it holds that

$$V_j^{-1/2}(\hat{\beta}_j^{(db)} - \beta_j^*) = R_j + o_P(1),$$

where $R_j \xrightarrow{D} N(0, 1)$ for

$$V_j = \frac{\hat{w}_j^\top \Sigma_a^{-1/2} \Sigma_{\theta^*} \Sigma_a^{-1/2} \hat{w}_j}{\{\hat{w}_j^\top (\mathbf{X}_a)_{\cdot j}\}^2}.$$

The magnitude of V_j satisfies

$$V_j = \frac{(\Sigma_{x|z}^{-1})_{j,j} \mathrm{Tr}(\Sigma_a^{-1} \Sigma_{\theta^*} \Sigma_a^{-1})}{\mathrm{Tr}^2(\Sigma_a^{-1})} (1 + o_P(1)). \tag{12}$$

Theorem 3.2 shows the asymptotic normality of the proposed debiased estimator under the given conditions. The convergence rate of $\hat{\beta}_j^{(db)}$ is $V_j^{1/2}$ with magnitude provided in (12). Remark 3.2 helps to understand the effect of a on the inference results. As $\Sigma_{x|z}$ is positive definite, V_j is proportional to $(\lambda_a^*)^2$ for any given p. Hence, using the debiased Lasso for linear models, that is, setting $a = 0$, can lead to large V_j and low power in statistical inference. We verify these arguments numerically in Section 5. The sparsity of $(\Sigma_{x|z}^{-1})_{.j}$ guarantees that $\hat{\kappa}_j$ converges to its probabilistic limit so that the central limit theorem can be justified. When Ψ is positive definite, the sample size condition for asymptotic normality is $(s \log p)^2 \vee \log n \max_i m_i \vee |H_j| \log p \ll \mathrm{Tr}(\Sigma_a^{-1})$. When Ψ is singular, it is sufficient to require $(s \log p)^2 \max_i m_i \vee \log n \max_i m_i^2 \vee |H_j| \log p \ll \mathrm{Tr}(\Sigma_a^{-1})$.

Theorem 3.2 is related to the results in some other works. When there is no random effect components, that is, $\mathbf{Z} = 0$, the conditions and conclusions of Theorem 3.2 recover the conditions and conclusions for the debiased Lasso in linear models, say, in Theorem 2.4 of van de Geer et al. (2014). In comparison to BCG19 of Bradic, Claeskens, and Gueuning (2020), the limiting distribution of their test statistic under the null hypothesis does not require sparse regression coefficients but requires $|H_j| = o(\sqrt{n}/\log p/\log n)$, using our notations. The power analysis for their test statistic requires $\max\{s, |H_j|\} = o(\sqrt{n}/\log p/\log n)$. In comparison, the sample size condition in our work (in the fixed q and m scenario) is $s = o(\sqrt{n}/\log p)$ and $|H_j| = o(n/\log p)$, which is weaker. More importantly, the realization of $\hat{\beta}_j^{(db)}$ does not rely on the null hypothesis and hence can be directly used to construct confidence intervals. We examine the empirical performance of these two different approaches in Section 5.

If one has a consistent estimator of V_j, the confidence intervals of the fixed effects can be constructed based on the limiting distribution of $\hat{\beta}_j^{(db)}$. However, a plug-in estimate of V_j would require the knowledge of the structures of variance components and extra efforts on estimation. In the following, we show that an empirical estimator of V_j, $\widehat{V}_j$ defined in (8), is consistent under mild conditions. Let $c_n = \log n \max_i m_i / \Lambda_{\min}(\Sigma_a^{-1/2} \Sigma_{\theta^*} \Sigma_a^{-1/2})$.

Lemma 3.2 (Convergence rate of the variance estimator). Under the conditions of Theorem 3.2, for $\widehat{V}_j$ defined in (8),

$$|\widehat{V}_j/V_j - 1| = O_P\left(c_n^{1/2} \mathrm{Tr}^{-1/2}(\Sigma_a^{-1}) + c_n \frac{s \log p}{\mathrm{Tr}(\Sigma_a^{-1})} \right).$$

Lemma 3.2 implies that the proposed variance estimator is consistent if $c_n = o(\mathrm{Tr}^{1/2}(\Sigma_a^{-1}))$ and the conditions of Theorem 3.2 hold true. The quantity c_n is to account for the correlations within clusters. The proposed $\widehat{V}_j$ is robust in the sense that it does not rely on the specific structure of the variance components and is consistent under mild conditions. Based on Theorem 3.2 and Lemma 3.2, hypothesis testing and constructing confidence intervals are both achievable. The asymptotic validness of the proposed confidence interval (7) is guaranteed.

We conclude this section with a further comment on the benefits of using the quasi-likelihood. In fact, even if we have consistent estimators of the variance parameters, say $\hat{\theta}$, using proxy matrix Σ_a to compute the debiased estimator is still favorable over using $\Sigma_{\hat{\theta}}$. First, using $\hat{\theta}$ can bring the complex dependence structure to R_j as the correction score would also depend on the random components. This makes it difficult to justify the asymptotic normality of R_j. Second, $\Sigma_{\hat{\theta}}^{-1}$ may not approximate $\Sigma_{\theta^*}^{-1}$ well enough in the sense that the magnitude of the error in $\hat{\theta}$ can be larger than the magnitude of the bias of the debiased Lasso estimator. As a result, the sample size condition for the asymptotic normality may be impaired.

3.4. Results for Possibly Correlated X and Z

In this subsection, we consider a relaxed version of Condition 3.2 that allows for correlation between $\mathbf{X}$ and $\mathbf{Z}$.

Condition 3.3. Conditioning on $\mathbf{Z}$, each row of $\mathbf{X}$ is independently distributed with covariance matrix $\Sigma_{x|z}$ such that $0 < K_* \le \Lambda_{\min}(\Sigma_{x|z}) \le \Lambda_{\max}(\Sigma_{x|z}) \le K^* < \infty$. Conditioning on $\mathbf{Z}$, the conditional sub-Gaussian norms of $X_{k,.}^i$ are upper bounded by K_0. Moreover, $\max_{1 \le j \le p} \|\mathbb{E}[\mathbf{X}_{.,j}|\mathbf{Z}]\|_2^2 \le c_1 \mathrm{Tr}(\Sigma_a^{-1})$ for some large enough c_1.

Revoking that $\|\mathbf{X}_{.,j}\|_2^2 \le CN$ and Remark 3.1, a sufficient condition for the last statement to hold is $q/(\min_i m_i) \le c_0 < 1$.

Lemma 3.3 (Fixed effects estimation with Lasso). Assume that Conditions 3.1 and 3.3 hold true. There exist large enough constants c_1 and c_2 such that for $\lambda \ge c_1 \sqrt{(\sigma_e^2 + K_1/a) \log p / \mathrm{Tr}(\Sigma_a^{-1})}$ and $\mathrm{Tr}(\Sigma_a^{-1}) \gg s \log p$, we have with probability at least $1 - 2\exp(-c_2 \log p)$,

$$\left\| \hat{\beta} - \beta^* \right\|_1 \le C_1 s \lambda, \quad \left\| \hat{\beta} - \beta^* \right\|_2 \le C_2 \sqrt{s} \lambda, \quad \text{and}$$
$$\frac{1}{\mathrm{Tr}(\Sigma_a^{-1})} \left\| \mathbf{X}_a (\hat{\beta} - \beta^*) \right\|_2^2 \le C_3 s \lambda^2 \qquad (13)$$

for some large enough constants C_1, C_2, and C_3.

Under Condition 3.3, for any constant $a > 0$, the effective sample size for the proposed approach is still of order $\mathrm{Tr}(\Sigma_a^{-1})$. However, we may not have a clear understanding of the optimal choice of a under current conditions but a can still be chosen by cross-validation in applications.

For inference of the fixed effects, one issue caused by the correlation between $\mathbf{X}$ and $\mathbf{Z}$ is that the limit of $\hat{\kappa}_j$ in (6) can depend on a and its sparsity is hard to guarantee. If its limit is sparse indeed, then the central limit theory of Theorem 3.2 still holds. If its limit is not sparse, then one may consider the debiasing scheme for linear models with the initial estimator computed by the quasi-likelihood approach. We show in the supplementary materials (Theorem A) that our proposed debiased estimator with $a = 0$ in (5) and (6) is robust to the correlation between $\mathbf{X}$ and $\mathbf{Z}$. However, its asymptotic normality requires stronger sample size conditions when Ψ is positive definite and it can have significantly wider confidence intervals and hence lower statistical power. We examine this method in Section 5 numerically.

4. Variance Components Estimation

In this section, we consider estimating the unknown parameters of variance components. While fixed effects inference can be of main interest in many problems, estimation of variance components can provide insights into the structure of the data. As far as we know, this problem has not been studied in existence of high-dimensional fixed effects. We will demonstrate that the idea of quasi-likelihood approach can be applied to estimating the variance components in high-dimensional linear mixed-effects models.

We parameterize Ψ as follows. Let $\eta^* \in \mathbb{R}^d$ be the true parameter such that

$$\Psi = \Psi_{\eta^*} = \sum_{j=1}^d \eta_j^* G_j \in \mathbb{R}^{q \times q}, \tag{14}$$

where $G_1, \ldots, G_d$ are symmetric basis matrices that are linearly independent in the sense that

$$\sum_{j=1}^d c_j G_j = 0 \text{ iff } c_1 = \cdots = c_d = 0. \tag{15}$$

The dimension d is allowed to grow to infinity. The structure of Ψ_{η^*} (14) incorporates most commonly used models in applications, such as the random intercept model and the models used in twin or family studies (Wang, Guo, and He 2011). One should note that any symmetric Ψ can be represented via (14) with η^* being the vector of its upper diagonal elements. Without loss of generality, we assume the basis matrices have a constant scale, that is, $\max_{1 \le j \le d} \|G_j\|_2 \le C < \infty$.

4.1. Estimating the Variance Components

A widely used approach for estimating the variance components is the Gaussian maximum likelihood method. However, this approach is highly restricted to the Gaussian assumptions. We consider a different approach that deals with sub-Gaussian random components. We first split the data into two folds: Let $I_1 \cup I_2 = [n]$, $I_1 \cap I_2 = \emptyset$, and $|I_1| \approx |I_2| \approx n/2$. Let $\hat{\beta}^{(2)}$ be an initial estimate of β^* with the second half of the data $\{X^i, Z^i, y_i\}_{i \in I_2}$. We compute the residuals $\hat{r}_i = y_i - X_i \hat{\beta}^{(2)}$ for $i \in I_1$ and estimate σ_e^2 via

$$\hat{\sigma}_e^2 = \frac{1}{\sum_{i \in I_1} \text{Tr}(P_{Z^i}^\perp)} \sum_{i \in I_1} \|P_{Z^i}^\perp \hat{r}_i\|_2^2. \tag{16}$$

Next, we estimate η^* via

$$\hat{\eta} = \underset{\eta \in \mathbb{R}^d}{\arg\min} \sum_{i \in I_1} \left\| (\Sigma_a^i)^{-1/2} (\hat{r}_i \hat{r}_i^\top - Z^i \Psi_\eta (Z^i)^\top - \hat{\sigma}_e^2 I_{n_i}) (\Sigma_a^i)^{-1/2} \right\|_F^2, \tag{17}$$

where constant $K \ge K_2$ and $\hat{\sigma}_e^2$ is obtained via (16).

The rationale of (16) is that the observations $P_{Z^i}^\perp (y_i - X^i \beta^*)$ have covariance matrix $\sigma_e^2 P_{Z^i}^\perp$ which only involves the target parameter σ_e^2. Replacing β^* with its quasi-likelihood estimate gives (16). This estimator is meaningful only when $\sum_{i \in I_1} \text{Tr}(P_{Z^i}^\perp) > 0$, that is, $\sum_{i \in I_1} m_i \max\{0, 1 - q/m_i\} > 0$. The rationale of (17) comes from the MLE. One can check

that the derivative of the target function in (17) would be the score function with respect to η if we replace Σ_a with the MLE estimate of Σ_{θ^*}. Different from the MLE, we estimate σ_e^2 and η^* separately. This is because a joint estimation of η^* and σ_e^2 may have poor performance. The reason is that, loosely speaking, the observed data involves N independent observations of the random noise and n independent observations of random effects. When $N \gg n$, the convergence rate for estimating σ_e^2 and η^* can have different magnitudes and a joint estimation can lead to ill-positioned Hessian matrix and nonsharp convergence rate. The sample splitting is for technical reasons and it is for proving that the estimation error of $\hat{\eta}$ is independent of the error of the fixed-effects estimation.

Computationally, $\hat{\sigma}_e^2$ in (16) is a one-step estimator and (17) involves a convex optimization, which can be easily implemented. On the other hand, sample splitting can lead to suboptimal finite sample performance and it is worthwhile to perform a cross-fitting step. That is, one can run another round of (16) and (17) with samples in the two folds switched and report the average of two estimates as the final estimate.

4.2. Upper Bound Analysis

In this subsection, we analyze the proposed estimator of the variance components. Let $D_G \in \mathbb{R}^{d \times d}$ be such that

$$\{D_G\}_{j,k} = \text{Tr}\left(G_j G_k\right). \tag{18}$$

The matrix D_G only depends on the prespecified basis and $\Lambda_{\min}(D_G) > 0$ as $G_j, j = 1, \ldots, d$, are linearly independent.

Theorem 4.1 (Convergence rate of variance components estimates). Assume that Conditions 3.1 and 3.2 hold and $\sum_{i \in I_2} \text{Tr}((\Sigma_a^i)^{-1}) \gg s \log p$. Then

$$|\hat{\sigma}_e^2 - \sigma_e^2| = O_P\left(\left(\sum_{i \in I_1} \max m_i\{0, 1 - q/m_i\} \right)^{-1/2} + \frac{s \log p}{\sum_{i \in I_2} \text{Tr}((\Sigma_a^i)^{-1})} \right).$$

If further $n \ge c_1 \log d$ for some large enough c_1, $\min_{1 \le i \le n} \Lambda_{\min}(\Sigma_z^i) \ge c_0/m_i > 0$, and $0 < c_1 \le \Lambda_{\min}(D_G) \le \Lambda_{\max}(D_G) \le c_2 < \infty$, then

$$\|\hat{\eta} - \eta^*\|_2 = O_P\left(\sqrt{\frac{d \log d}{n}} \right).$$

The convergence rate of $\hat{\sigma}_e^2$ depends on the effective sample size in I_1 as well as the estimation error of $\hat{\beta}^{(2)}$. In comparison to the rate of variance estimation in linear models (Verzelen 2012), the current result replaces the total sample size with the effective sample size. On the other hand, $\hat{\eta}$ has the typical parametric rate when there are d unknown parameters and n independent observations of random effects. The estimation error of $\hat{\eta}$ is independent of the error of the fixed effects estimation.

In terms of conditions, D_G depends on prespecified basis matrices and it eigenvalues are positive and bounded in many cases. Consider the class of basis matrices where $G_{j,k} \in \mathbb{R}^{q \times q}$ such that $(G_{j,k})_{l,q} = (G_{j,k})_{q,l} = 1$ if $l = j, q = k$ and

$(G_{j,k})_{l,q} = 0$ otherwise. For any $1 \leq d \leq q(q+1)/2$, it is easy to check that in this case $\Lambda_{\min}(D_G) = \Lambda_{\max}(D_G) = 1$. For independent sub-Gaussian $Z^i_{k,\cdot}$, $\min_{1 \leq i \leq n} \Lambda_{\min}(\Sigma^i_z) \geq c_0$ with high probability if $q \log n \ll \min_{1 \leq i \leq m} m_i$. To summarize, the estimators proposed in Section 4.1 are mainly for the scenario where $q \leq c_0 \min_{1 \leq i \leq n} m_i$ for sufficiently small c_0.

4.3. Rate Optimality of Estimating Variance Components

Now we turn to study the minimax lower bound for estimating the variance parameters. We consider the random components satisfying (10) and parameter space $\Xi(s, \mathbf{Z})$ (11).

Theorem 4.2 (Minimax lower bounds for estimation of variance components). Suppose that (1) and (10) are true. If $s \leq c \min\{\mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1})/\log p, p^\nu\}$ for $0 < \nu < 1/2$ and $c > 0$ for some $c_0 > 0$, then there exists some constants $c_1 - c_3 > 0$ such that

$$\inf_{\hat{\sigma}_e^2} \sup_{\boldsymbol{v} \in \Xi(s, \mathbf{Z})} \mathbb{P}_{\boldsymbol{v}}\left(|\hat{\sigma}_e^2 - \sigma_e^2| \geq c_1 \mathrm{Tr}^{-1/2}(\boldsymbol{\Sigma}_a^{-1}) + c_2 \frac{s \log(p/s^2)}{\mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1})} \Big| \mathbf{Z}\right) \geq \frac{1}{4}.$$

If further $\Lambda_{\max}(D_G) \leq C < \infty$,

$$\inf_{\hat{\eta}} \sup_{\boldsymbol{v} \in \Xi(s, \mathbf{Z})} \mathbb{P}_{\boldsymbol{v}}\left(\|\hat{\eta} - \eta^*\|_2 \geq c_3 n^{-1/2} | \mathbf{Z}\right) \geq \frac{1}{4}.$$

Theorems 4.2 and 4.1 together imply that $\hat{\sigma}_e^2$ is rate optimal in $\Xi(s, \mathbf{Z})$ under the conditions of Theorem 4.2 if when $\mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1}) \asymp \sum_{i=1}^n m_i \max\{0, 1 - q/m_i\}$. As explained after Lemma 3.1, in the case where group sizes are equal and $q/m \leq c_0 < 1$, $\sum_{i=1}^n \max\{0, m_i - q\} \asymp \mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1}) \asymp N$. Moreover, $\hat{\eta}$ is rate optimal when d is finite. When d grows, regularized estimators of η^* can have smaller estimation error than $\hat{\eta}$, similar to the famous Stein's phenomenon.

5. Simulation Results

In this section, we present simulation results to evaluate the empirical performance of the proposed methods and compare it with some related methods. We examine the effect of a on estimation and inference of the fixed effects.

We generate data as follows. We set $N = 144$ and $p = 300$. Each row of $(\mathbf{X}, \mathbf{Z})$ is iid generated from a normal distribution with mean zero and covariance such that $\Sigma_x = I_p$, $\Sigma_z = I_q$, and $(\Sigma_{x,z})_{k,j} = \rho^j$ for $1 \leq j, k \leq q$ and $(\Sigma_{x,z})_{k,\cdot} = 0$ for $k >$

q. That is, the correlation between $\mathbf{X}_j$ and $\mathbf{Z}$ is nonzero if $j \leq p$ and is 0 if $j > q$. The random noises are iid generated via $\epsilon_i \sim N(0, 0.25 I_{m_i})$ and the random effects are iid generated via $\gamma_i \sim N(0, \Psi)$. We consider $q \in \{2, 8, 14\}$. The matrix Ψ will be specified later. The responses $\mathbf{y}$ are generated via model (1) with $s = 5$ and $\beta_{1:5} = (1, 0.5, 0.2, 0.1, 0.05)^\top$ and equal cluster sizes, that is, $m_1 = \cdots = m_n = m$. Each setting is replicated with 300 independent Monte Carlo simulations.

5.1. Statistical Inference for Fixed Effects

We first examine the empirical performance of the proposed confidence intervals (7) and hypothesis testing based on $\hat{\beta}_j^{(db)}$. We consider two covariance matrices of random effects, a "positive definite Ψ" where $\Psi_{j,k} = 0.56^{|j-k|}$ for $1 \leq j, k \leq q$, and a "singular Ψ" with a diagonal Ψ where $\Psi_{j,j} = 0.56$ for $1 \leq j \leq q/2$ and $\Psi_{j,j} = 0$ otherwise. For the proposed method, we first choose a by cross-validation using the error criteria $\|\mathbf{y} - \mathbf{X}\hat{\beta}(a)\|_2^2$, where $\hat{\beta}(a)$ the proposed estimate associated with a specific a. The tuning parameter λ is chosen as $\hat{\sigma}^{(\mathrm{init})}\sqrt{2 \log p/N}$, where $\hat{\sigma}^{(\mathrm{init})}$ is computed via the scaled-Lasso (Sun and Zhang 2012) with observations $\{\mathbf{X}, \mathbf{y}_a\}$. For computing $\hat{\beta}_j^{(db)}$, the tuning parameters λ_j are set to be $\hat{\sigma}_x \sqrt{2 \log p/N}$, where $\hat{\sigma}_x$ is computed via the scaled-Lasso with observations $\{(\mathbf{X}_a)_{\cdot,-j}, (\mathbf{X}_a)_{\cdot,j}\}$. The tuning parameters for BCG19 are chosen as in Section 5 of Bradic, Claeskens, and Gueuning (2020).

We see from Table 1 that the coverage probabilities of the proposed confidence intervals are close to the nominal level in most scenarios. It shows that the proposed method is robust to large m and q and singular Ψ. We see that the confidence intervals have shorter lengths when m increases. This is because when q is fixed and m grows, the effective sample size $\mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1})$ increases and the proposed estimators have smaller estimation errors. See Table 1 in the supplementary materials for details. When q grows and m is fixed, the effective sample size $\mathrm{Tr}(\boldsymbol{\Sigma}_a^{-1})$ is smaller and the proposed estimators have larger estimation errors. The results for $\rho = 0.2$ are reported in the Table 2 of supplementary materials.

In Table 2, we report the Type I error and power of our proposed method and those of BCG19. The computational time for our proposal is around 8 sec per experiment and that for BCG19 is around 20 sec per experiment. Ideally, the rejection rate for the true null should be close to 5% and the rejection

Table 1. 95%-confidence intervals given by the proposed approach with positive definite Ψ and singular Ψ when $\rho = 0$.

q	m	Positive definite Ψ				Singular Ψ			
		cov(0.5)	cov(0)	SD(0.5)	SD(0)	cov(0.5)	cov(0)	SD(0.5)	SD(0)
2	4	0.940	0.957	0.068	0.062	0.953	0.943	0.062	0.056
	8	0.943	0.967	0.063	0.053	0.938	0.981	0.058	0.049
	12	0.943	0.967	0.061	0.049	0.967	0.948	0.059	0.047
8	4	0.943	0.960	0.195	0.177	0.957	0.943	0.128	0.111
	8	0.960	0.940	0.148	0.123	0.933	0.919	0.105	0.088
	12	0.943	0.973	0.106	0.083	0.957	0.976	0.085	0.066
14	4	0.937	0.953	0.276	0.264	0.976	0.948	0.173	0.153
	8	0.937	0.947	0.243	0.217	0.924	0.929	0.158	0.132
	12	0.933	0.950	0.202	0.166	0.981	0.924	0.148	0.112

NOTE: "cov(0.5)" and "cov(0)" denote the coverage probabilities for $\beta_j = 0.5$ and $\beta_j = 0$, respectively. "SD(0.5)" and "SD(0)" denote the standard deviations for $\beta_j = 0.5$ and $\beta_j = 0$, respectively.

Table 2. The rejection rate for testing $H_0 : \beta_j^* = 0$ at 95% level for $\beta_j^* \in \{1, 0.5, 0.2, 0\}$ with positive definite Ψ (p.d.) and singular Ψ when $\rho = 0$.

Ψ	q	m	Proposed 1	0.5	0.2	0	BCG19 1	0.5	0.2	0
p.d.	2	4	1	1	0.793	0.043	1	1	0.627	0.043
		8	1	1	0.880	0.033	1	1	0.850	0.040
		12	1	1	0.940	0.033	1	1	0.936	0.040
	8	4	0.997	0.713	0.163	0.040	0.987	0.593	0.150	0.067
		8	1	0.923	0.243	0.060	0.987	0.747	0.173	0.060
		12	1	1	0.423	0.027	1	0.840	0.207	0.030
	14	4	0.943	0.437	0.117	0.047	0.867	0.327	0.123	0.037
		8	0.967	0.487	0.113	0.053	0.927	0.397	0.107	0.057
		12	0.993	0.610	0.157	0.050	0.930	0.477	0.150	0.060
Singular	2	4	1	1	0.895	0.057	1	1	0.900	0.029
		8	1	1	0.952	0.020	1	1	0.943	0.024
		12	1	1	0.914	0.052	1	1	0.927	0.057
	8	4	1	0.995	0.362	0.057	1	0.976	0.371	0.052
		8	1	0.986	0.438	0.071	1	0.986	0.438	0.062
		12	1	1	0.638	0.024	1	1	0.567	0.047
	14	4	1	1	0.638	0.024	1	1	0.567	0.048
		8	1	0.895	0.228	0.043	1	0.890	0.233	0.052
		12	1	0.895	0.229	0.067	1	0.890	0.233	0.052

rates for $\beta_j^* \in \{1, 0.5, 0.2\}$ should be larger than 5%. We see that both our proposal and BCG19 are effective at controlling the Type I error. However, BCG19 is less powerful than our proposal when q is large and Ψ is positive definite. When Ψ is singular, two methods have comparable performance in most scenarios.

Table 3 demonstrates the effect of a on estimation and inference of the fixed effects. The results for singular Ψ are reported in the supplementary materials. We see that choosing $a = 0$ can lead to large estimation errors and significantly wider confidence intervals. This implies that the Lasso for linear models is less accurate than our proposed methods with $a > 0$. In all the scenarios of (q, m), we see that the estimation error first decreases as a increases and then increases as a increases.

This phenomenon agrees with Remark 3.2. For the inference results, the proposed confidence interval has the desired coverage probabilities as long as a is not too large. We see that setting $a = 0$ has coverage probabilities close to the nominal level but the confidence intervals are significantly wider than setting $a > 0$. This implies that using the linear debiased Lasso can lead to low power in hypothesis testing for mixed-effects models.

5.2. Estimating Variance Components

In this subsection, we consider estimating variance components with the proposed method. The true fixed effects and data generation steps are the same as in Section 5.1. We use the whole data to estimate σ_e^2 and η^*. We set $\sigma_e^2 = 0.25$. We first consider diagonal Ψ with $d = 2$. The basis matrices are set to be

$$G_1 = \begin{pmatrix} I_{q/2} & 0 \\ 0 & 0 \end{pmatrix} \text{ and } G_2 = \begin{pmatrix} 0 & 0 \\ 0 & I_{q/2} \end{pmatrix}.$$

For diagonal Ψ, $\eta^* = (0.56, 0.56)^\top$. For singular Ψ, $\eta^* = (0.56, 0)^\top$. Table 4 shows the mean absolute errors of σ_e^2 (mae.σ_e^2), η_1^* (mae.η_1), and η_2^* (mae.η_2). A scenario with relatively large d is reported in the supplementary materials (Table 4).

Table 3. Effect of different a on sum of squared error (SSE) for estimating β^* and on the accuracy of confidence intervals with positive definite Ψ and $\rho = 0$.

(q, m)	a	SSE	$\mathrm{Tr}(\Sigma_a^{-1})$	cov(0.5)	cov(0)	SD(0.5)	SD(0)
	0	0.321	144.0	0.948	0.967	0.138	0.122
	2	0.134	87.3	0.962	0.962	0.074	0.065
	4	0.113	81.4	0.933	0.952	0.070	0.062
(2,4)	8	0.111	77.7	0.933	0.957	0.068	0.062
	16	0.103	75.2	0.900	0.967	0.065	0.060
	32	0.105	73.8	0.890	0.933	0.065	0.060
	0	0.753	144.0	0.943	0.943	0.261	0.235
	2	0.338	34.5	0.952	0.962	0.150	0.122
	4	0.342	26.2	0.948	0.933	0.148	0.121
(8,8)	8	0.349	19.6	0.943	0.967	0.144	0.119
	16	0.350	14.8	0.910	0.962	0.143	0.116
	32	0.402	10.9	0.895	0.967	0.142	0.119
	0	0.961	144.0	0.948	0.967	0.344	0.316
	2	0.531	19.9	0.981	0.948	0.211	0.167
	4	0.501	12.9	0.938	0.967	0.199	0.162
(14,12)	8	0.515	8.1	0.938	0.971	0.200	0.167
	16	0.504	5.0	0.948	0.957	0.200	0.161
	32	0.551	2.9	0.905	0.967	0.190	0.155

NOTE: "cov(0.5)" and "cov(0)" denote the coverage probabilities for $\beta_j = 0.5$ and $\beta_j = 0$, respectively. "SD(0.5)" and "SD(0)" denote the standard deviations for $\beta_j = 0.5$ and $\beta_j = 0$, respectively.

Table 4. Estimation of the variance components with the proposed method for positive definite and singular Ψ when $\rho = 0$.

		Positive definite Ψ			Singular Ψ		
m	q	mae.σ_e^2	mae.η_1	mae.η_2	mae.σ_e^2	mae.η_1	mae.η_2
4	2	0.115	0.206	0.207	0.076	0.150	0.050
8	2	0.091	0.212	0.249	0.070	0.171	0.020
	4	0.122	0.164	0.166	0.071	0.136	0.027
12	2	0.087	0.268	0.245	0.076	0.197	0.015
	6	0.126	0.163	0.160	0.078	0.116	0.019

6. Application to A Genome-Wide Association Study in a Mouse Population

We apply the proposed method to estimate the effects of genetic variants on the BMI in a heterogenous stock mice population generated by the Welcome Trust Centre for Human Genetics http://gscan.well.ox.ac.uk. The data is available in R package "BGLR" (Perez and de los Campos 2014). The dataset consists of 1814 mice, each genotyped over 10,346 polymorphic markers (SNPs) and has been used for genome-wide genetic association studies of multiple traits (Shifman, Bell, and Copley 2006; Valdar et al. 2006). This mice population consists of 8 liters and were housed in 523 different cages, each including a different number of mice. The distribution of cage density is in Figure 1. We are interested in identifying the genetic variants that are associated with the BMI phenotype. The measurements of BMI are transformed as described in Valdar et al. (2006) so that the data distribution is close to normal. In many mice experiments, cages often contribute significant environmental effects to the phenotypes such as BMI and mice in the same cage tend to be correlated in their phenotype measurements. It is therefore important to account for such cage effect in genetic association studies and the linear mixed-effects model can be employed.

In the current analysis, we incorporate the effect of cages as a single random effect and consider the following model

$$Y_{i,k} = \beta_0 + \sum_{j=1}^{10346} \beta_j X_{j,k}^i + \tau_1 age_{i,k} + \tau_2 gender_{i,k} + \gamma_i + \epsilon_{i,k},$$

where $Y_{i,k}$ is the BMI of the kth mouse in the ith cage, $X_{j,k}^i$ is the numerical genotype at the genetic variant j for the ith mouse in
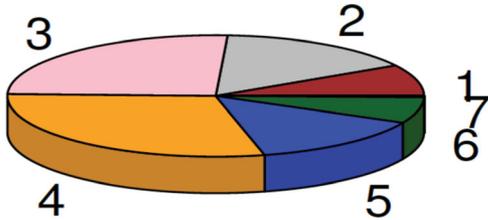
cage k, β_0 and β_j are the regression coefficients corresponding to the intercept and genetic variants, τ_1 and τ_2 are the regression coefficients for age and gender, γ_i is the cage-specific random effect for the ith cage. For cages with only one individual, we only fit the fixed effects.

The fixed effects are estimated via a weighted Lasso. To mitigate the relatively high correlation among the design, we first compute ridge regression estimates of the fixed effects, say $\hat{\beta}^{(rr)}$, with tuning parameter chosen by cross-validation and use normalized $\{1/|\hat{\beta}_j^{(rr)}|\}_{j=1}^p$ (sum up equal to p) as the weights for the Lasso estimates. The regression coefficient $\hat{\beta}$ is obtained by fitting (4) with tuning parameter $\lambda = 0.655 \times \sqrt{2 \log p/N}$, where 0.655 is the noise level estimated by the scaled Lasso. In terms of statistical inference, we compute the debiased Lasso estimates of the fixed effect via (5) and their variances according to (8). According to cross-validation, we set $a = 2$.

6.1. Identification of BMI Associated Genetic Variants

We control the false discovery rate (FDR) at 5% using the procedure proposed in Xia, Cai, and Cai (2018). Our method identifies 14 covariates with p-value threshold 6.7×10^{-5}. The QQ plot of the z-scores of all the covariates is given in the left panel of Figure 2. It shows some deviation from standard normal density at both tails, indicating that some variants can be associated with BMI (Table 5). Some of the genetic variants identified are in or near the genes known to be associated with body growth, body size, metabolism or obesity. For example, SNP rs13478535 is a variant in Auts2 gene, which has been shown to be related to with either low birth weight or small stature mice (Gao, Lee, and Stafford 2014). SNP rs13481413 is one of the genetic variants in gene Immp2l, which is associated with food intake and body weight (Han, Zhao, and Lu 2013). cAMP response element binding protein (Crebbp) has been postulated to play an important role downstream of the melanocortin-4 receptor and may affect other pathways that are implicated in the regulation of body weight (Chiappini et al. 2011).

We also consider applying the proposed procedure with $a = 0$. This is equivalent to applying the Lasso to fit the linear model to without considering the random cage effects. The tuning

Figure 1. Cage density of the stock mice population. The average density of cages with at least two individuals is 3.70.
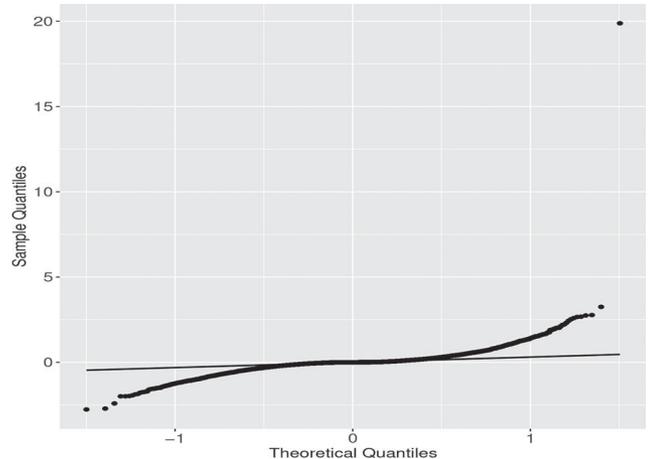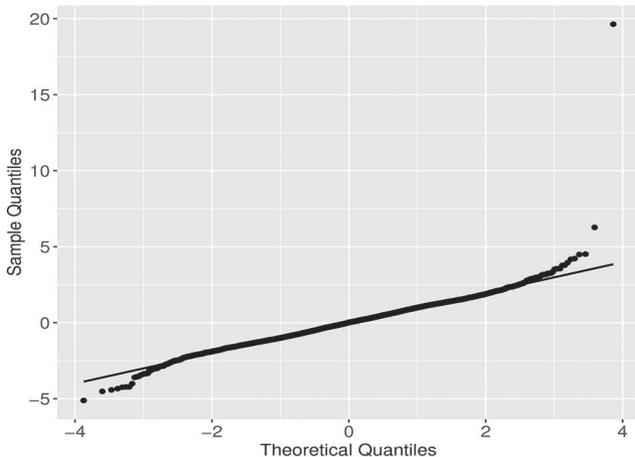
Figure 2. The normal QQ-plots of the z-scores of the debiased Lasso estimators with proposed approach (left) and the debiased Lasso estimates with $a = 0$ (right) for 10348 fixed effects. The straight reference line passes the first and third quantiles of the z-scores.

Table 5. Selected covariates at FDR level 0.05.

SNP	Gene	z-score	SNP	Gene	z-score
rs13476390		−5.11	rs13482464		−4.01
rs13478535	Auts2	−4.42	rs4152477	Crebbp	4.17
rs3023058	Srrm3	4.22	rs6251709	Csnka2ip	4.49
rs13480072		−4.52	rs6185805	Mtcl1	−4.22
rs13481413	Immp2l	−4.22	gnf18.028.738		−4.34
rs13481961		6.27	gnfX.070.167		−4.24
rs4139535		4.51	gender		19.63

NOTE: 13 SNPs and gender are selected at FDR ≤ 0.05 and their z-scores are reported. The genes where the SNPs are located are presented when they are available.

parameters are chosen in the same way as above. Only gender is selected as nonzero at FDR level 0.05. This is possibly due to the model misspecification and larger variances of the debiased Lasso estimators. The QQ-plot of the z-scores based on the debiased Lasso estimation of the linear model (right panel of Figure 2) shows that the z-scores clearly deviate from the standard normal distribution. These results indicate that the proposed estimation and inference methods for the linear mixed-effects model indeed provide an effective way of identifying important genetic variants associated with BMI in mice.

6.2. Evaluation of Cage Effect

For estimating the variance components, we only use the clusters with at least two observations. The estimated variance of the random effects is 0.202 and the estimated variance of the noise is 0.209. We compute the standard error of the estimated variance of the random effects assuming that the random components are normally distributed. The estimated standard deviation is 0.018, which indicates a strong cage effect.

7. Discussion

The present article considers estimation and inference of unknown parameters in a high-dimensional linear mixed-effects model. Optimal rate of convergence for estimation was established and rate-optimal estimators were developed. The proposed methods have general applicability in modeling repeated measures and longitudinal data, especially when the cluster sizes are large or heterogeneous. The desirable properties of the proposed estimators are mainly due to the proper approximations of the unknown oracle weighting matrix $\boldsymbol{\Sigma}_{\theta*}$. Our proposed estimation procedure is computationally efficient and does not require strong distributional assumptions on the random effects and error distributions.

The proposed methods have important applications in large-scale genetic association studies in humans, including both family-based studies where the kinship coefficients can be used to specify the random effects and population cohort studies where the random effects can be used to adjust for population stratification (Yang et al. 2014). Instead of considering one genetic variant at a time as in typical mixed-effects models in genetic association studies (Yang et al. 2014), our model considers all the variants jointly. We expect gain in power in detecting phenotype-associated genetic variants by allowing for flexible random effects and by considering all genetic variants jointly using high-dimensional mixed-effects models studied in this article.

Supplementary Materials

In the online supplementary materials, we provide proofs of all the theorems and lemmas and more numerical studies.

Funding

References

Ahmn, M., Zhang, H. H., and Lu, W. (2012), "Moment-Based Method for Random Effects Selection in Linear Mixed Models," *Statistica Sinica*, 100, 130–134. [2]

Bradic, J., Claeskens, G., and Gueuning, T. (2020), "Fixed Effects Testing in High-Dimensional Linear Mixed Models," *Journal of the American Statistical Association*, 115, 1835–1850. [2,3,4,6,8]

Bühlmann, P., and van de Geer, S. (2015), "High-Dimensional Inference in Misspecified Linear Models," *Electronic Journal of Statistics*, 9, 1449–1473. [4]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [2,5]

——— (2020), "Semi-Supervised Inference for Explained Variance in High-Dimensional Regression and Its Applications," *Journal of the Royal Statistical Society*, Series B, 82, 391–419. [2]

Cai, T. T., Guo, Z., and Ma, R. (2020), "Statistical Inference for High-Dimensional Generalized Linear Models With Binary Outcomes," Technical Report. [2]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When *p* Is Much Larger Than *n*," *The Annals of Statistics*, 35, 2313–2351. [2]

Chiappini, F., Cunha, L., Harris, J., and Hollenberg, A. (2011), "Lack of Camp-Response Element-Binding Protein 1 in the Hypothalamus Causes Obesity," *Journal of Biological Chemistry*, 286, 8094–8105. [10]

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, New York: Wiley. [2]

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017), "High-Dimensional Simultaneous Inference With the Bootstrap," *TEST*, 26, 685–719. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [2]

Fan, Y., and Li, R. (2012), "Variable Selection in Linear Mixed Effects Models," *The Annals of Statistics*, 40, 2043–2068. [2,3,4]

Fang, E. X., Ning, Y., and Liu, H. (2017), "Testing and Confidence Intervals for High Dimensional Proportional Hazards Models," *Journal of the Royal Statistical Society*, Series B, 79, 1415–1437. [2]

Gao, Z., Lee, P., and Stafford, J. M. (2014), "AUTS2 Confers Gene Activation to Polycomb Group Proteins in the CNS," *Nature*, 516, 349–354. [10]

Goldstein, H. (2011), *Multilevel Statistical Models* (Vol. 922), Chichester: Wiley. [1]

Gumedze, F. N., and Dunne, T. T. (2011), "Parameter Estimation and Inference in the Linear Mixed Model," *Linear Algebra and Its Applications*, 435, 1920–1944. [1]

Han, C., Zhao, Q., and Lu, B. (2013), "The Role of Nitric Oxide Signaling in Food Intake; Insights From the Inner Mitochondrial Membrane Peptidase 2 Mutant Mice," *Redox Biology*, 1, 498–507. [10]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [2,4,5]

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), "On High-Dimensional Misspecified Mixed Model Analysis in Genome-Wide Association Study," *The Annals of Statistics*, 44, 2127–2160. [4]

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974. [1]

Lin, X. (1997), "Variance Component Testing in Generalised Linear Models With Random Effects," *Biometrika*, 84, 309–326. [2]

Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection," *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [2]

Miller, J. J. (1977), "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance," *The Annals of Statistics*, 5, 746–762. [2]

Peng, H., and Lu, Y. (2012), "Model Selection in Linear Mixed Effect Models," *Journal of Multivariate Analysis*, 109, 109–129. [2]

Perez, P., and de los Campos, G. (2014), "Genome-Wide Regression and Prediction With the BGLR Statistical Package," *Genetics*, 198, 483–495. [10]

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer. [1,4]

Schelldorfer, J., Bühlmann, P., and van De Geer, S. (2011), "Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization," *Scandinavian Journal of Statistics*, 38, 197–214. [2,4]

Self, S. G., and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605. [2]

Shifman, S., Bell, J., and Copley, R. (2006), "A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome," *PLoS Biology*, 4, e395. [10]

Stram, D. O., and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, 1171–1177. [2]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [8]

Sun, Y., Zhang, W., and Tong, H. (2007), "Estimation of the Covariance Matrix of Random Effects in Longitudinal Studies," *The Annals of Statistics*, 35, 2795–2814. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [2,3]

Valdar, W., Solberg, L., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., Taylor, M. S., Rawlins, J. N. P., Mott, R., and Flint, J. (2006), "Genome-wide genetic association of complex traits in heterogeneous stock mice," *Nature Genetics*, 38, 879–887. [10]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2,4,5,6]

Verbeke, G., and Molenberghs, G. (2003), "The Use of Score Tests for Inference on Variance Components," *Biometrics*, 59, 254–262. [2]

Verzelen, N. (2012), "Minimax Risks for Sparse Regressions: Ultra-High Dimensional Phenomenons," *Electronic Journal of Statistics*, 6, 38–90. [7]

Wang, X., Guo, X., and He, M. (2011), "Statistical Inference in Mixed Models and Analysis of Twin and Family Data," *Biometrics*, 67, 987–995. [7]

Xia, Y., Cai, T., and Cai, T. T. (2018), "Two-Sample Tests for High-Dimensional Linear Regression With an Application to Detecting Interactions," *Statistica Sinica*, 28, 63–92. [10]

Yang, J., Zaitlen, N., Goddard, M., Visscher, P., and Price, A. (2014), "Advantages and Pitfalls in the Application of Mixed Model Association Methods," *Nature Genetics*, 46, 100–106. [11]

Zhang, C. H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [2]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [2,4,5]

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757–768. [2]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [2]