Statistica Sinica **28** (2018), 1585-1605
doi:https://doi.org/10.5705/ss.202017.0220

# MM ALGORITHMS FOR VARIANCE COMPONENT ESTIMATION AND SELECTION IN LOGISTIC LINEAR MIXED MODEL

Liuyi Hu[1], Wenbin Lu[1], Jin Zhou[2] and Hua Zhou[3]

[1]*North Carolina State University,* [2]*University of Arizona and* [3]*University of California*

*Abstract:* Logistic linear mixed models are widely used in experimental designs and genetic analyses of binary traits. Motivated by modern applications, we consider the case of many groups of random effects, where each group corresponds to a variance component. When the number of variance components is large, fitting a logistic linear mixed model is challenging. Thus, we develop two efficient and stable minorization–maximization (MM) algorithms for estimating variance components based on a Laplace approximation of the logistic model. One of these leads to a simple iterative soft-thresholding algorithm for variance component selection using the maximum penalized approximated likelihood. We demonstrate the variance component estimation and selection performance of our algorithms by means of simulation studies and an analysis of real data.

*Key words and phrases:* Generalized linear mixed model (GLMM), Laplace approximation, MM algorithm, variance components selection.

## 1. Introduction

The generalized linear mixed model (GLMM) is an extension of the generalized linear model, incorporating random effects that account for heterogeneity among responses (McCulloch and Neuhaus (2001); Stroup (2012)). The GLMM is widely used in clustered, longitudinal, and panel data analyses (Zeger and Karim (1991); Breslow and Clayton (1993)). The logistic linear mixed model is a GLMM used for binary responses, and assumes

$$\begin{aligned} y_j \mid \eta_j &\sim \text{Bernoulli}(\mu_j) \\ \mu_j &= \frac{1}{\{1+\exp(-\eta_j)\}}, \end{aligned} \tag{1.1}$$

for $j = 1, \ldots, n$. Here, $\eta = (\eta_1, \ldots, \eta_n)^T$ takes the form

$$\eta = X\beta + Z_1 u_1 + \cdots + Z_m u_m,$$

where $X$ and $Z = (Z_1, \ldots, Z_m)$ are known predictor matrices, $\beta$ is the coefficient vector for fixed effects, and $u_i \sim N(0_{q_i}, \sigma_i^2 I_{q_i})$ are independent random effects. Because

$$\eta \sim N(X\beta, \sigma_1^2 Z_1 Z_1^T + \cdots + \sigma_m^2 Z_m Z_m^T),$$

we call $\sigma_1^2, \ldots, \sigma_m^2$ variance components.

The logistic linear mixed model has been applied in agriculture, econometrics, biology, and genetics. Here, examples include using an analysis of variance (ANOVA) for dichotomous responses (Anderson and Aitkin (1985); Quené and Van den Bergh (2008)) and quantitative trait loci (QTL) mappings of binary traits (Yi and Xu (1999); Che and Xu (2012)). In the ANOVA, $Z_i$ denotes each factor or their interactions. In modern applications, the number of factors can be large, and the number of interaction terms increases quadratically with the number of factors. In QTL mapping, $Z_i$ corresponds to a gene region. The number of genes $m$ is of order $10^2 \sim 10^3$ in a typical genetic study. In Section 4 and 5, we discuss these two applications in further detail, as well as presenting an associated analysis using our proposed algorithms.

In general, direct maximization of the GLMM likelihood function is computationally intractable because it involves potentially high-dimensional integrals. Thus, existing methods involve various forms of approximations. The first class of methods uses numerical integration, such as Gaussian quadrature (Davidian and Gallant (1992)) and adaptive Gaussian quadrature (AGQ) (Pinheiro and Bates (1995)). These methods apply only to low-dimensional integrals, and thus are limited to problems in which data form very small independent clusters. The second type of method employs Laplace approximation (Wolfinger (1993); Shun and McCullagh (1995)) or its variants, such as the penalized quasi-likelihood (Breslow and Clayton (1993)) and the integrated nested Laplace approximation (Rue, Martino and Chopin (2009)). The third class of methods uses Monte Carlo methods to approximate either the original integral (Sung and Geyer (2007)) or the E step of the EM algorithm (Booth and Hobert (1999)). Pinheiro and Bates (1995) compare and discuss the penalized quasi-likelihood, the Laplace approximation, importance sampling, Gaussian quadrature, and AGQ. They conclude that the Laplace approximation and AGQ give the "best mix of efficiency and accuracy." Thus, we propose algorithms based on the Laplace approximation of the log-likelihood function because AGQ is numerically infeasible for the ANOVA and genetic applications we consider here.

Our primary interest is the estimation and selection of variance components.

Several studies have proposed ways of selecting fixed effects in GLMMs (Groll and Tutz (2014); Schelldorfer, Meier and Bühlmann (2014)). However, for the selection of random effects, most procedures are developed in the framework of linear mixed models (Bondell, Krishna and Ghosh (2010); Ahn, Zhang and Lu (2012)) for quantitative responses. In contrast, few works discuss random effects selection in GLMMs. Ibrahim et al. (2011) develop a simultaneous fixed and random effects selection procedure based on the smoothly clipped absolute deviation (SCAD) and adaptive LASSO penalties using a Monte Carlo EM for general mixed models. Cai and Dunson (2006) propose a method for random effect selection in GLMMs within a Bayesian framework using a stochastic search MCMC algorithm. Pan and Huang (2014) propose using a backfitting algorithm to select effective random effects based on a penalized quasi-likelihood function. However, the above-mentioned studies all examine clustered data with repeated measurements on the subjects. They assume $n$ independent subjects with observations $(y_1, X_1, Z_1), \ldots, (y_n, X_n, Z_n)$ and

$$\mathrm{E}\left(y_i \mid X_i, Z_i, b_i\right) = g\left(\eta_i\right) = g\left(X_i\beta + Z_i b_i\right), \tag{1.2}$$

where $g(\cdot)$ is some known link function, $X_i$ and $Z_i$ are known matrices, and $b_i \sim N_q\left(0, D\right)$ is the random effect. Here, $D$ is the unknown covariance matrix shared by the subjects that is to be estimated by maximizing some penalized likelihood. For example, Ibrahim et al. (2011) perform the penalization on the Cholesky decomposition of $D$, denoted as $\Gamma$, such that each row of $\Gamma$ is either all not zero or all zero, and Pan and Huang (2014) penalize on positive elements proportional to the standard deviation of the random effects $b_i$. We propose an algorithm for selecting random effects in which we shrink the variances of ineffective random effects toward zero based on the penalized likelihood defined in Section 3.3. There are two key differences between our variance component selection and those of previous works. First, model (1.2) is not the same as model (1.1) that we address in this study, even though they can both deal with clustered data and non-clustered data. Model (1.1) assumes that the random effects $u_i \sim N(0_{q_i}, \sigma_i^2 I_{q_i})$ are independent. If we write model (1.1) in the framework of model (1.2), then the covariance in model (1.1) is diagonal with some equality constraints on the random effect variances, whereas the covarariance in model (1.2) can be any covariance matrix. Second, model (1.2) selects individual random effects, whereas model (1.1) is used to select groups of random effects, that is, the random effects in each $u_i$ are either all selected or none are selected. To the best of our knowledge, no studies investigate variance component selection

using model (1.1).

Based on the minorization–maximization (MM) principle (Lange, Hunter and Yang (2000)), we propose two novel algorithms for variance component estimation under two different parameterizations of logistic linear mixed models. Then, we extend the algorithms to variance component selection by incorporating penalization. The first parameterization is efficient for estimating parameters without penalty, whereas the second easily generalizes to penalized estimation. Both algorithms are simple to implement and numerically stable. Our simulation studies and real-data analysis demonstrate that the proposed algorithms outperform the commonly used tools and are scalable to high-dimensional problems.

## 2. Preliminaries

Throughout, we reserve Greek letters for parameters and indicate the current iteration number by a superscript $t$.

### 2.1. The MM principle

The MM principle (Lange, Hunter and Yang (2000); Hunter and Lange (2004)) for maximizing an objective function $f(\theta)$ involves two M-steps. The first M-step minorizes the objective function $f(\theta)$ by a surrogate function $g(\theta \mid \theta^{(t)})$ at the current iterate $\theta^{(t)}$. Minorization is a combination of a tangent condition $f(\theta^{(t)}) = g(\theta^{(t)} \mid \theta^{(t)})$ and a domination condition $f(\theta) \geq g(\theta \mid \theta^{(t)})$, for $\theta \neq \theta^{(t)}$. The second M-step is defined by the iterates:

$$\theta^{(t+1)} = \arg\max_{\theta} g(\theta \mid \theta^{(t)}). \tag{2.1}$$

Because

$$f(\theta^{(t+1)}) \geq g(\theta^{(t+1)} \mid \theta^{(t)}) \geq g(\theta^{(t)} \mid \theta^{(t)}) = f(\theta^{(t)}), \tag{2.2}$$

the MM iterates satisfy the ascent property, which drives the objective function uphill and makes the MM algorithm remarkably stable.

Our derivation of the MM algorithms for variance component estimation and selection hinges on two minorizations.

### 2.2. Supporting hyperplane minorization

If $f(\theta)$ is convex and differentiable, then the supporting hyperplane

$$g(\theta) = f(\theta^{(t)}) + \nabla f(\theta^{(t)})^T(\theta - \theta^{(t)}) \tag{2.3}$$

is a minorization function of $f(\theta)$ at $\theta^{(t)}$ (Hunter and Lange (2004)).

For symmetric matrices, we write $A \preceq B$ when $B - A$ is positive semidefinite.

A matrix-valued function $f$ is said to be (matrix) convex if

$$f\{\lambda A + (1-\lambda)B\} \preceq \lambda f(A) + (1-\lambda)f(B),$$

for all $A$, $B$, and $\lambda \in [0,1]$. Because the negative log determinant function $f(B) = -\log\det B$ is convex on the set of positive definite matrices (Boyd and Vandenberghe (2004)) and the supporting hyperplane of $f(B)$ is

$$\begin{aligned} g(B) &= f(B^{(t)}) + \nabla f(B^{(t)})^T (B - B^{(t)}) \\ &= -\log\det B^{(t)} - \operatorname{tr}\left\{ \left(B^{(t)}\right)^{-1} \left(B - B^{(t)}\right) \right\}, \end{aligned}$$

the supporting hyperplane minorization described above yields the following inequality:

$$-\log\det B \geq -\log\det B^{(t)} - \operatorname{tr}\left\{ \left(B^{(t)}\right)^{-1} \left(B - B^{(t)}\right) \right\}. \tag{2.4}$$

### 2.3. Quadratic minorization

If a convex function $f(\theta)$ is twice differentiable and there exists a matrix $M$ such that $M \preceq \nabla^2 f(\theta)$ for all $\theta$, then

$$g(\theta) = f(\theta^{(t)}) + \nabla f(\theta^{(t)})^T (\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T M (\theta - \theta^{(t)}) \tag{2.5}$$

is a minorization function of $f(\theta)$ at $\theta^{(t)}$ Hunter and Lange (2004).

## 3. Algorithms for Estimation

### 3.1. Model formulation 1

The likelihood for model (1.1) is

$$L(\beta, \sigma) = \int \exp\{h(u \mid \beta, \sigma^2)\}\, du, \tag{3.1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)^T$, with $\sigma_i \geq 0$ for $i = 1, \ldots, m$, $\sigma^2 = \left(\sigma_1^2, \ldots, \sigma_m^2\right)^T$, and the complete log-likelihood is

$$\begin{aligned} h(u \mid \beta, \sigma^2) &= \sum_j \{y_j \eta_j - \ln(1 + e^{\eta_j})\} - \frac{1}{2}\sum_{i=1}^m \left( q_i \ln \sigma_i^2 + \frac{\|u_i\|_2^2}{\sigma_i^2} \right) \\ &= \sum_j \{y_j \eta_j - \ln(1 + e^{\eta_j})\} - \frac{1}{2}\sum_{i=1}^m \frac{\|u_i\|_2^2}{\sigma_i^2} + \text{ terms without } u_i. \end{aligned}$$

Direct optimization of the likelihood defined in (3.1) is computationally challenging because of the integral. The Laplace approximation to the likelihood $L(\beta, \sigma)$ is obtained by replacing $h(u \mid \beta, \sigma^2)$ by its second-order Taylor expan-

sion at the conditional maximum. Given the current iterate $(\beta, \sigma)$, let $u^*$ be the maximizer of $h$ and $\eta^* = X\beta + Zu^*$, where $Z = (Z_1, Z_2, \ldots, Z_m)$. Then, the approximated log-likelihood is

$$\begin{aligned} L_{\mathrm{LA}}(\beta, \sigma) &= h(u^* \mid \beta, \sigma^2) - \frac{1}{2} \ln \det \nabla^2 \left\{ -h(u^* \mid \beta, \sigma^2) \right\} \\ &= \sum_j \left\{ y_j \eta_j^* - \ln \left( 1 + e^{\eta_j^*} \right) \right\} - \frac{1}{2} \sum_{i=1}^m q_i \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^m \frac{\|u_i^*\|_2^2}{\sigma_i^2} \\ &\quad - \frac{1}{2} \ln \det \left\{ Z^T W^* Z + \mathrm{blkdiag}(\sigma_1^{-2} I_{q_1}, \ldots, \sigma_m^{-2} I_{q_m}) \right\} \\ &= \sum_j \left\{ y_j \eta_j^* - \ln \left( 1 + e^{\eta_j^*} \right) \right\} - \frac{1}{2} \sum_{i=1}^m \frac{\|u_i^*\|_2^2}{\sigma_i^2} \\ &\quad - \frac{1}{2} \ln \det \left( W^{*-1} + \sum_i \sigma_i^2 Z_i Z_i^T \right) - \frac{1}{2} \ln \det W^* \\ &\quad + \text{terms without } \beta, \sigma^2, \end{aligned} \tag{3.2}$$

where $W^* = \mathrm{diag}(w^*)$ is a diagonal matrix with entries

$$w_j^* = p_j^*(1 - p_j^*) = \frac{e^{\eta_j^*}}{\left(1 + e^{\eta_j^*}\right)^2} \quad \text{and} \quad p_j^* = \frac{e^{\eta_j^*}}{\left(1 + e^{\eta_j^*}\right)}.$$

Detailed derivations of the approximated log-likelihood (3.2) are provided in the Supplementary Material S1. The MM algorithm cycles through updates of $u$, $\beta$ and $\sigma^2$, as follows:

1. To maximize $h(u \mid \beta, \sigma^2)$, the gradient and Hessian are

$$\begin{aligned} \nabla_u h &= Z^T (y - p) - \begin{pmatrix} \sigma_1^{-2} u_1 \\ \vdots \\ \sigma_m^{-2} u_m \end{pmatrix} \\ \nabla_u^2 h &= - \left\{ Z^T W Z + \mathrm{blkdiag}(\sigma_1^{-2} I_{q_1}, \ldots, \sigma_m^{-2} I_{q_m}) \right\}, \end{aligned}$$

   respectively, where $p = (p_1, \ldots, p_n)^T$ with $p_j = e^{\eta_j}/(1 + e^{\eta_j})$, and $W = \mathrm{diag}(w_1, \ldots, w_n)$ with $w_j = p_j(1 - p_j)$. Because each $w_j$ is upper-bounded by 0.25, it follows that

$$\nabla_u^2 h \succeq - \left\{ 0.25 Z^T Z + \mathrm{blkdiag}(\sigma_1^{-2} I_{q_1}, \ldots, \sigma_m^{-2} I_{q_m}) \right\}.$$

   Thus, we can construct a quadratic minorization function at $u^{(l)}$ using (2.5), and maximizing the quadratic surrogate gives the MM update

$$u^{(l+1)} = \left\{ 0.25 Z^T Z + \mathrm{blkdiag}(\sigma_1^{-2} I_{q_1}, \ldots, \sigma_m^{-2} I_{q_m}) \right\}^{-1} \nabla_u h(u^{(l)}) + u^{(l)}. \tag{3.3}$$

To find the maximizer $u^*$ given $\beta$ and $\sigma^2$, we iterate the MM update (3.3) until convergence. Note that the indicated matrix inverse in (3.3) only needs to be performed once and remains constant through the iterations.

2. Updating $\beta$ given $\sigma^2$ and $u^*$ is a regular logistic regression with offset $Zu^*$. We invoke a similar MM update to that described above

$$\beta^{(t+1)} = \beta^{(t)} + \left(0.25X^TX\right)^{-1} X^T(y - p^*). \tag{3.4}$$

Again, the matrix inverse $\left(0.25X^TX\right)^{-1}$ only needs to be performed once.

3. To update $\sigma^2$ given $\beta$ and $u^*$, the minorization (2.4) leads to the surrogate function

$$\begin{aligned} g(\sigma^2 \mid \sigma^{2(t)}) = \; & c^{(t)} - \frac{1}{2}\sum_{i=1}^{m} \frac{\|u_i^*\|_2^2}{\sigma_i^2} \\ & -\frac{1}{2}\sum_{i=1}^{m} \sigma_i^2 \mathrm{tr}\left\{ \left( \sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1} \right)^{-1} Z_i Z_i^T \right\}, \end{aligned} \tag{3.5}$$

where $c^{(t)}$ is a constant irrelevant to the optimization. The maximization of $g(\sigma^2 \mid \sigma^{2(t)})$ with respect to $\sigma^2$ yields the explicit MM update

$$\sigma_i^{2(t+1)} = \left[ \frac{\|u_i^*\|_2^2}{\mathrm{tr}\left\{ Z_i^T (\sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1})^{-1} Z_i \right\}} \right]^{1/2}.$$

When $q \ll n$, the Woodbury formula facilitates the inversion

$$\begin{aligned} & \left( \sum_i \sigma_i^{(t)2} Z_i Z_i^T + W^{*-1} \right)^{-1} \\ & = W^* - W^* Z(\sigma)\{I_q + Z(\sigma)^T W^* Z(\sigma)\}^{-1} Z(\sigma)^T W^*, \end{aligned}$$

where $Z(\sigma) = (\sigma_1 Z_1, \ldots, \sigma_m Z_m)$. Because the iterate is derived based on the MM principle, it possesses the ascent property

$$L_{LA}(\sigma^{(t+1)} \mid \beta, u^*) \geq L_{LA}(\sigma^{(t)} \mid \beta, u^*). \tag{3.6}$$

A detailed proof is presented in the Supplementary Material S3.

As in the penalized iteratively reweighted least squares (PIRLS) algorithm described in Bates et al. (2015), parameter estimates are determined for a fixed-weights matrix $W^*$, and then the weights are updated to the current estimates and the process is repeated. The resulting algorithm is extremely simple to implement. Algorithm 1 summarizes the MM algorithm for the parameter estimation

**Input :** $y$, $X$, $Z_1, \ldots, Z_m$
**Output:** MLE $\hat{\beta}$, $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2$
Initialize $\beta^{(0)}, \sigma_i^{(0)} > 0$, $i = 1, \ldots, m$ ;
**repeat**

$$u^* \leftarrow \arg\max_u \; h(u \mid \sigma^{2(t)}, \beta^{(t)}) \; ;$$

$$p^* \leftarrow 1/\left\{1 + \exp\left(-X\beta^{(t)} - Zu^*\right)\right\} \; ;$$

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \left(0.25X^TX\right)^{-1} X^T(y - p^*) \; ;$$

$$p^* \leftarrow 1/\left\{1 + \exp\left(-X\beta^{(t+1)} - Zu^*\right)\right\} \; ;$$

$$W^* \leftarrow \operatorname{diag}\{p^*(1 - p^*)\} \; ;$$

$$\sigma_i^{2(t+1)} \leftarrow \left[\frac{\|u_i^*\|_2^2}{\operatorname{tr}\left\{Z_i^T(\sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1})^{-1} Z_i\right\}}\right]^{1/2}, \quad i = 1, \ldots, m \; ;$$

**until** *objective value converges*;

**Algorithm 1:** MMLA1 – an MM algorithm to maximize the Laplace approximation of the likelihood for model (1.1).

of the logistic linear mixed model (1.1). Each iteration involves a one-step update of $\beta$ and $\sigma^2$. Several additional steps updating $\beta$ and $\sigma^2$ give similar results, in practice.

### 3.2. Model formulation 2

In the Laplace-approximated log-likelihood (3.2), we have $\sigma_i$ in the denominator. Thus, it cannot be combined with the penalized estimation, which will shrink some of the $\sigma_i$s to zero. Therefore, we consider another reparameterization of model (1.1) by assuming that $\eta$ takes the form

$$\eta = X\beta + \sigma_1 Z_1 u_1 + \cdots + \sigma_m Z_m u_m, \tag{3.7}$$

where $u_i \sim N(0_{q_i}, I_{q_i})$ are independent. Let $u = (u_1^T, \ldots, u_m^T)^T \in \mathcal{R}^q$ be the concatenated random effects and $Z = (Z_1, \ldots, Z_m) \in \mathcal{R}^{n \times q}$, $q = \sum_{i=1}^m q_i$. Then, $\eta = X\beta + ZDu$, where $D = \operatorname{blkdiag}(\sigma_1 I_{q_1}, \ldots, \sigma_m I_{q_m})$ and the complete log-likelihood is

$$h(u \mid \beta, \sigma) = \sum_j \{y_j \eta_j - \ln(1 + e^{\eta_j})\} - \frac{1}{2}\|u\|_2^2 + \text{ terms without } u.$$

Given the current iterate $(\beta, \sigma)$, let $u^*$ be the maximizer of $h$ and $\eta^* = X\beta + ZDu^*$. Then the approximated log-likelihood is

$$L_{\mathrm{LA}}(\beta, \sigma)$$
$$= h(u^* \mid \beta, \sigma) - \frac{1}{2} \ln \det \nabla^2\{-h(u^* \mid \beta, \sigma)\}$$

$$= \sum_j \left\{ y_j \eta_j^* - \ln\left(1 + e^{\eta_j^*}\right) \right\} - \frac{1}{2}\|u^*\|_2^2 - \frac{1}{2} \ln\det\left(D^T Z^T W^* Z D + I_q\right)$$

$$= \sum_j \left\{ y_j \eta_j^* - \ln\left(1 + e^{\eta_j^*}\right) \right\} - \frac{1}{2}\|u^*\|_2^2 - \frac{1}{2} \ln\det\left(W^{*-1} + \sum_i \sigma_i^2 Z_i Z_i^T\right)$$

$$-\frac{1}{2}\ln\det W^* + \text{terms without } \beta, \sigma^2. \tag{3.8}$$

Detailed derivations of the above approximated log-likelihood can be found in the Supplementary Material S2. Maximizing $h(u \mid \beta, \sigma)$ follows similar MM updates to those in (3.3). Given $\sigma^2$ and $\beta$, $u^*$ can be found through the MM iterates

$$u^{(l+1)} = u^{(l)} + \left\{0.25(ZD)^T ZD + I_q\right\}^{-1} \nabla_u h(u^{(l)} \mid \beta, \sigma^2)$$

until convergence, where $\nabla_u h(u^{(l)} \mid \beta, \sigma^2) = D^T Z^T (y - p) - u^{(l)}$. Updating $\beta$ given $u^*$ and $\sigma^2$ is the same as the update in (3.4).

Updating $\sigma^2$ given $\beta$ and $u^*$ depends on three minorizations, which differ from the first reparameterization. Quadratic minorization implies that

$$\begin{aligned} -1^T \ln\left(1 + e^{\eta^*}\right) &\geq -p^{(t)T}\left(\eta^* - \eta^{*(t)}\right) - \frac{1}{8}\|\eta^* - \eta^{*(t)}\|_2^2 + c^{(t)} \\ &= -p^{(t)T} Z D u^* - \frac{1}{8}\|Z(D - D^{(t)})u^*\|_2^2 + c^{(t)}, \end{aligned} \tag{3.9}$$

where $c^{(t)}$ is an irrelevant constant, $p^{(t)}$ is a vector with the $j$th element equal to $e^{\eta_j^{*(t)}}/(1 + e^{\eta_j^{*(t)}})$, and $\eta_j^{*(t)}$ is the $j$th element of $\eta^{*(t)} = X\beta + ZD^{(t)}u^*$. The Cauchy inequality implies that

$$\begin{aligned} -\|Z(D - D^{(t)})u^*\|_2^2 &= -\left\| \sum_{i=1}^m Z_i u_i^* (\sigma_i - \sigma_i^{(t)}) \right\|_2^2 \\ &\geq -\left\{ \sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2 \right\} \sum_{i=1}^m (\sigma_i - \sigma_i^{(t)})^2, \end{aligned} \tag{3.10}$$

where $(Z_i u_i^*)_j$ is the $j$th element of vector $Z_i \mu_i^*$. Combining (3.9), (3.10), and (2.4) gives the overall minorization function

$$\begin{aligned} g(\sigma \mid \sigma^{(t)}) = &\sum_{i=1}^m \sigma_i \left(y - p^{(t)}\right)^T Z_i u_i^* - \frac{1}{8}\left\{ \sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2 \right\} \sum_{i=1}^m (\sigma_i - \sigma_i^{(t)})^2 \\ &-\frac{1}{2}\sum_{i=1}^m \sigma_i^2 \mathrm{tr}\left\{ \left( \sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1} \right)^{-1} Z_i Z_i^T \right\} + c^{(t)}, \end{aligned} \tag{3.11}$$

where $\sigma_i$ are nicely separated and only involve quadratic terms. The maximiza-

**Input :** $y$, $X$, $Z_1, \ldots, Z_m$
**Output:** MLE $\hat{\beta}$, $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2$
Initialize $\beta^{(0)}, \sigma_i^{(0)} > 0$, $i = 1, \ldots, m$ ;
**repeat**

$$D^{(t)} = \operatorname{diag}\left(\sigma_1^{(t)} 1_{q_1}, \ldots, \sigma_m^{(t)} 1_{q_m}\right) ;$$

$$u^* \leftarrow \arg\max_u \; h(u \mid \sigma^{2(t)}, \beta^{(t)}) ;$$

$$p^{(t)} \leftarrow 1/\left\{1 + \exp\left(-X\beta^{(t)} - ZD^{(t)}u^*\right)\right\} ;$$

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \left(0.25X^TX\right)^{-1} X^T(y - p^{(t)}) ;$$

$$p^{(t)} \leftarrow 1/\left\{1 + \exp\left(-X\beta^{(t+1)} - ZD^{(t)}u^*\right)\right\} ;$$

$$W^* \leftarrow \operatorname{diag}\left\{p^{(t)}(1 - p^{(t)})\right\};$$

$$\sigma_i^{2(t+1)} \leftarrow \max\left[0, \frac{\left(y - p^{(t)}\right)^T Z_i u_i^* + 1/4\left\{\sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2\right\}\sigma_i^{(t)}}{\operatorname{tr}\left\{\left(\sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1}\right)^{-1} Z_i Z_i^T\right\} + 1/4\left\{\sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2\right\}}\right],$$

$i = 1, \ldots, m$ ;
**until** *objective value converges*;

**Algorithm 2:** MMLA2 – an MM algorithm to maximize the Laplace approximation of the likelihood for model (3.7).

tion of $g(\sigma \mid \sigma^{(t)})$ results in the following update:

$$\sigma_i^{(t+1)} = \frac{\left(y - p^{(t)}\right)^T Z_i u_i^* + 1/4\left\{\sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2\right\}\sigma_i^{(t)}}{\operatorname{tr}\left\{\left(\sum_i \sigma_i^{2(t)} Z_i Z_i^T + W^{*-1}\right)^{-1} Z_i Z_i^T\right\} + 1/4\left\{\sum_{j=1}^n \sum_{i=1}^m (Z_i u_i^*)_j^2\right\}}. \tag{3.12}$$

To account for the non-negative constraint of $\sigma$, at each iteration, we set $\sigma_i^{(t+1)} = \max(0, \sigma_i^{(t+1)})$. Algorithm 2 summarizes the MM algorithm for model formulation 2 defined in (3.7).

### 3.3. MM algorithm for maximizing the penalized approximated like lihood

For the variance component selection, we consider the penalization approach using a lasso penalty.

Because the minorization function of $\sigma$ derived in the second model formulation is a quadratic function of $\sigma$, it meshes well with the penalized estimation. Other penalties such as the adaptive lasso (Zou (2006)) and SCAD (Fan and Li (2001)) lead to similar algorithms.

The lasso-penalized approximated log-likelihood is

$$-L_{\mathrm{LA}}(\beta,\sigma)+\lambda\sum_{i=1}^{m}|\sigma_i|. \tag{3.13}$$

Finding $u^*$ to maximize $h(u \mid \beta,\sigma)$ and updating $\beta$ follow the same steps described in algorithm 2. The only difference lies in the update of $\sigma$ given $u^*$ and $\beta$ in (3.12), which now becomes

$$\begin{aligned}\sigma_i^{(t+1)} &= \arg\min_{\sigma_i}\sigma_i^2\left[\frac{1}{2}\mathrm{tr}\left\{\left(\sum_i \sigma_i^{2(t)}Z_iZ_i^T+W^{*-1}\right)^{-1}Z_iZ_i^T\right\}+\frac{1}{8}\left\{\sum_{j=1}^{n}\sum_{i=1}^{m}(Z_iu_i^*)_j^2\right\}\right]\\ &\quad-\sigma_i\left[\left(y-p^{(t)}\right)^T Z_iu_i^*+\frac{1}{4}\left\{\sum_{j=1}^{n}\sum_{i=1}^{m}(Z_iu_i^*)_j^2\right\}\sigma_i^{(t)}\right]+\lambda|\sigma_i| \qquad (3.14)\\ &= ST\left(z_i,\gamma_i\right),\end{aligned}$$

where

$$ST(z,\gamma)=\arg\min_x\frac{1}{2}(x-z)^2+\gamma|x|=\mathrm{sng}(z)\left(|z|-\gamma\right)_+ \tag{3.15}$$

is the soft-thresholding operator, and

$$\begin{aligned}z_i &= \frac{\left(y-p^{(t)}\right)^T Z_iu_i^*+1/4\left\{\sum_{j=1}^{n}\sum_{i=1}^{m}(Z_iu_i^*)_j^2\right\}\sigma_i^{(t)}}{\mathrm{tr}\left\{(\sum_i\sigma_i^{2(t)}Z_iZ_i^T+W^{*-1})^{-1}Z_iZ_i^T\right\}+1/4\left\{\sum_{j=1}^{n}\sum_{i=1}^{m}(Z_iu_i^*)_j^2\right\}},\\ \gamma_i &= \frac{\lambda}{\mathrm{tr}\left\{(\sum_i\sigma_i^{2(t)}Z_iZ_i^T+W^{*-1})^{-1}Z_iZ_i^T\right\}+1/4\left\{\sum_{j=1}^{n}\sum_{i=1}^{m}(Z_iu_i^*)_j^2\right\}}.\end{aligned}$$

### 3.4. Choice of regularization parameter

The best $\lambda$ can be selected over a grid using the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or cross-validation. Here, we consider the AIC and BIC. Because it is difficult to evaluate the log-likelihood function, we replace it by its Laplace approximation. Specifically, we use

$$\begin{aligned}\mathrm{BIC}(\lambda) &= -2L_{\mathrm{LA}}(\hat\beta,\hat\sigma^2)+\log(n)\times\mathrm{df}(\lambda)\\ \mathrm{AIC}(\lambda) &= -2L_{\mathrm{LA}}(\hat\beta,\hat\sigma^2)+2\times\mathrm{df}(\lambda),\end{aligned}$$

where $\mathrm{df}(\lambda)$ is the number of non-zeros in $\hat\sigma^2(\lambda)$. In the following simulation studies, we compare the AIC and BIC on variance component selection.

## 4. Simulation Studies

### 4.1. Random effects ANOVA

In this section, we compare the estimation error and runtime of the MM algorithms (MMLA1 and MMLA2) to three different implementations: (1) the `glmer()` function in the popular `lme4` package in R (Bates et al. (2015)); (2) the `glmm()` function in the `glmm` package in R (Knudson (2016)); and (3) the `stan_glmer()` function in the `rstanarm` package in R (Stan Development Team (2016)). `glmer()` fits a generalized linear mixed-effects model and the default (nAGQ=1) uses a Laplace approximation to approximate the original log-likelihood. `glmm()` calculates and maximizes the Monte Carlo likelihood approximation (MCLA) (Geyer (1990)) to find Monte Carlo maximum likelihood estimates (MCMLEs) (Sung and Geyer (2007)) for the fixed effects and the variance components. The `rstanarm` package is an R interface to the Stan C++ library for Bayesian estimations. `stan_glmer()` adds independent prior distributions to the regression coefficients, as well as priors on the covariance matrices of the group-specific parameters. Then, it performs a Bayesian inference via MCMC.

We simulate data from the following two-way ANOVA model with crossed random effects:

$$\begin{aligned} P(y_{ijk} = 1) &= \frac{1}{(\exp(-\eta_{ijk}))} \\ \eta_{ijk} &= x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \alpha_i + \gamma_j + (\alpha\gamma)_{ij}, \\ i &= 1, \ldots, 5, j = 1, \ldots, 5, k = 1, \ldots c, \end{aligned}$$

where $\alpha_i \sim N(0, \sigma^2_\alpha)$, $\gamma_j \sim N(0, \sigma^2_\gamma)$ and $(\alpha\gamma)_{ij} \sim N(0, \sigma^2_{\alpha\gamma})$ are jointly independent. Here $i$ indexes the levels in factor 1, $j$ indexes the levels in factor 2, and $k$ indexes observations in the $(i, j)$-combination. This corresponds to $m = 3$ variance components. Table 1 displays the results when there are $a = b = 5$ levels for each factor, the number of observations $c$ in each combination of factor levels varies from 2 to 200, and the true parameter values are $(\beta_1, \beta_2, \beta_3, \sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_{\alpha\gamma}) = (0.6, 1.0, -1.0, 0.5, 0.9, 0.3)$. For each scenario, we simulate 50 replicates. The sample size is $n = abc$ for each replicate. Therefore, the largest model in Table 1 involves a covariance matrix of size 5,000 × 5,000. For $c = 100$ and 200, we omit the results of `glmm` and `rstanarm` because they take too much time when the sample size increases (the simulation takes more than a week to complete).

We observe the following. The results of the two MM algorithms (MMLA1 and MMLA2) are very similar, but MMLA2 takes longer to converge than MMLA1

Table 1. Comparison of the MM algorithms with two different parameterizations (MMLA1 and MMLA2) and the `glmer()` function (with `nAGQ=1`) in the `lme4` package, `rstanarm` package, and `glmm` package. Standard errors are given in parentheses. The results for `rstanarm` and `glmm` with $c = 100, 200$ are not reported because the simulation takes more than a week to complete.

| c | Method | runtime | $\beta_1(0.6)$ | $\beta_2(1.0)$ | $\beta_3(-1.0)$ | $\sigma^2_\alpha(0.5)$ | $\sigma^2_\gamma(0.9)$ | $\sigma^2_{\alpha\gamma}(0.3)$ |
|---|---|---|---|---|---|---|---|---|
| 2 | MMLA1 | 0.19 (0.55) | 0.68(0.51) | 1.08(0.43) | −0.92(0.51) | 0.52(0.91) | 1.03 (1.55) | 0.22 (0.37) |
| | MMLA2 | 0.14 (0.12) | 0.68(0.51) | 1.08(0.43) | −0.92(0.51) | 0.52(0.91) | 1.04 (1.56) | 0.22 (0.37) |
| | lme4 | 0.46 (0.37) | 2.83(7.22) | 3.52(7.39) | −2.42(4.04) | 187(753) | 108 (580) | 558 (2,049) |
| | rstanarm | 8.15 (0.49) | 0.91(0.69) | 1.42(0.45) | −1.20(0.58) | 1.38(1.32) | 2.14 (2.23) | 2.60 (1.86) |
| | glmm | 23.95 (45.66) | 0.64(0.53) | 0.91(0.55) | −0.76(0.59) | 1.54(3.13) | 0.03 (0.07) | 0.06 (0.14) |
| 8 | MMLA1 | 0.10 (0.03) | 0.55(0.21) | 0.96(0.24) | −0.98(0.20) | 0.36(0.33) | 0.96 (0.94) | 0.34 (0.34) |
| | MMLA2 | 0.17 (0.08) | 0.55(0.21) | 0.96(0.24) | −0.98(0.20) | 0.36(0.33) | 0.96 (0.94) | 0.34 (0.34) |
| | lme4 | 0.37 (0.10) | 0.60(0.23) | 1.04(0.27) | −1.07(0.22) | 0.42(0.38) | 1.15 (1.13) | 0.47 (0.48) |
| | rstanarm | 21.85 (1.15) | 0.61(0.24) | 1.05(0.27) | −1.09(0.22) | 0.68(0.44) | 1.48 (1.20) | 0.72 (0.53) |
| | glmm | 224.53 (492.52) | 0.46(0.17) | 0.82(0.24) | −0.85(0.17) | 0.78(1.50) | 0.02 (0.03) | 0.04 (0.08) |
| 50 | MMLA1 | 0.19 (0.10) | 0.58(0.07) | 1.01(0.08) | −1.00(0.08) | 0.52(0.43) | 0.96 (0.81) | 0.31 (0.16) |
| | MMLA2 | 1.65 (0.52) | 0.58(0.07) | 1.01(0.08) | −1.00(0.08) | 0.52(0.43) | 0.94 (0.72) | 0.31 (0.16) |
| | lme4 | 0.92 (0.12) | 0.59(0.07) | 1.03(0.08) | −1.02(0.09) | 0.54(0.45) | 1.01 (0.86) | 0.32 (0.17) |
| | rstanarm | 198.38 (26.88) | 0.59(0.07) | 1.04(0.08) | −1.02(0.09) | 0.82(0.58) | 1.37 (0.92) | 0.42 (0.21) |
| | glmm | 3,613.26 (2,272.85) | 0.48(0.09) | 0.86(0.12) | −0.84(0.12) | 0.88(1.39) | 0.04 (0.06) | 0.04 (0.07) |
| 100 | MMLA1 | 0.58 (0.18) | 0.61(0.06) | 1.01(0.06) | −1.00(0.06) | 0.65(0.46) | 0.94 (0.61) | 0.30 (0.11) |
| | MMLA2 | 4.28 (0.78) | 0.61(0.06) | 1.01(0.06) | −1.00(0.06) | 0.67(0.44) | 0.91 (0.54) | 0.30 (0.11) |
| | lme4 | 1.49 (0.18) | 0.62(0.06) | 1.02(0.06) | −1.01(0.06) | 0.67(0.47) | 0.97 (0.63) | 0.31 (0.12) |
| | rstanarm | — | — | — | — | — | — | — |
| | glmm | — | — | — | — | — | — | — |
| 200 | MMLA1 | 0.98 (0.16) | 0.60(0.04) | 0.99(0.04) | −0.99(0.04) | 0.45(0.33) | 0.92 (0.62) | 0.29 (0.12) |
| | MMLA2 | 13.49 (3.42) | 0.60(0.04) | 0.99(0.04) | −0.99(0.04) | 0.50(0.33) | 0.91 (0.51) | 0.29 (0.12) |
| | lme4 | 2.76 (0.33) | 0.60(0.04) | 1.00(0.04) | −1.00(0.04) | 0.46(0.33) | 0.94 (0.63) | 0.30 (0.13) |
| | rstanarm | — | — | — | — | — | — | — |
| | glmm | — | — | — | — | — | — | — |

does, especially when the number of groups $c$ is large. This is expected because the surrogate function derived in MMLA2 involves two additional layers of minorizations, which result in slower convergence. The `glmer()` function failed to converge in many replicates when $c = 2$ and produced much worse estimates than those of the MM algorithms. For other values of $c$, `glmer()` delivered estimates comparable to those of the MM algorithm, but was three to four times slower than MMLA1. `glmm()` and `stan_glmer()` are much slower because they involve sampling and their estimation performance is not good. The core algorithm in `glmer()` is coded in C and extensively utilizes sparse linear algebra. Our MM algorithms are implemented in the high-level Julia language and ignore sparsity structures. Although it is difficult to draw conclusions based on implementations in different languages, this example clearly demonstrates the efficiency and scalability of the MM algorithms for GLMM estimation.

### 4.2. Genetic example

In this section, we use a genetic example to demonstrate the performance of the variable selection using our algorithm derived in Section 3.3. Consider the QTL mapping example introduced in Section 1:

$$g(\mu) = X\beta + G\gamma,$$

where $G$ is an $n \times k$ genotype matrix for $k$ variants of interest, $g(\mu) = \text{logit}\,(\mu)$, $\beta$ are fixed effects, and $\gamma$ are random genetic effects with $\gamma \sim \text{Normal}\left(0, \sigma^2 I_k\right)$. The response $y$ is an $n \times 1$ vector of binary trait measurements with mean $\mu$. One way to identify important genes is to test the null hypothesis $\sigma^2 = 0$ for each region separately, and then to adjust for multiple testing (Lee et al. (2014)). Here, we consider the joint model for all regions rather than using marginal tests:

$$g(\mu) = X\beta + s_1^{-1/2} G_1\gamma_1 + \cdots + s_m^{-1/2} G_m\gamma_m, \tag{4.1}$$

where $\gamma_i \sim N(0, \sigma_i^2 I)$ and we select the variance components $\sigma_i^2$ via the penalization (3.13). Here, $s_i$ is the number of variants in region $i$, and the weights $s_i^{-1/2}$ put all variance components on the same scale.

In this simulation study, we use the genetic data from the COPDGene exome sequencing study (Regan et al. (2011)), which has 399 subjects and genotype information of 16,610 genes. The covariate matrix $X$ contains `intercept`, `age`, `sex`, and the top three principal components in the mean effects. We consider four experimental settings for sparse random effects. In all of the examples, we set $\beta = (0.1, -1.0, 0.8, -0.3, -1.2, 1.5)$ and randomly select $m$ genes $G_i$, $i = 1, \ldots, m$, from the COPD data.

- Setting 1: $\sigma^2 = \left(5.0, 7.5, 10.0, 0_{m-3}^T\right)^T$, with $m$ varying from 5, 10, 20, 100
- Setting 2: $\sigma^2 = \left(10, 15, 20, 0_{m-3}^T\right)^T$, with $m$ varying from 5, 10, 20, 100
- Setting 3: $\sigma^2 = \left(5, 6, 7, 8, 9, 10, 0_{m-6}^T\right)^T$, with $m$ varying from 10, 20, 40, 100
- Setting 4: $\sigma^2 = \left(10, 12, 14, 16, 18, 20, 0_{m-6}^T\right)^T$, with $m$ varying from 10, 20, 40, 100

We use the mean squared error (MSE) $= \|\hat{\beta}-\beta\|^2$ to evaluate the performance of the fixed effect estimation. Four measures are used to assess the variable selection performance: the number of truly non-zero variance components that are selected as non-zero variance components (denoted as "True Positive"), the number of truly zero variance components that are selected as non-zero variance components (denoted as "False Positive"), the frequency of exactly selecting the correct variance components (denoted by "Exact"), and the frequency of over-selecting variance components (denoted by "Over"). In each experimental setting, 100 data sets are simulated from the model, and we report the average performance over the 100 runs for both the AIC and the BIC. Tables 2, 3, 4, and 5 summarize the results for the above four settings. We can see that our proposed method for variable selection does a good job in identifying the significant random effects. For example, under Setting 1 and Setting 2 for different $m$, our method based on both the AIC and the BIC can identify the truly significant random effects 97% - 99% of the time, with the AIC more prone to over-selection than is the BIC. Setting 3 and Setting 4 are more challenging because they involve a larger number of random effects. However, our method can still identify the non-zero random effect 96% of the time under $m = 10$ when using the AIC.

## 5. Real-Data Analysis

In this real-data analysis, we again use the data from the COPDGene exome sequencing study described in the previous simulated genetic example. The binary trait indicates whether or not an individual smokes (denoted as `smoke`). There are 399 individuals with 646,125 genetic variants in 16,610 genes. The covariates include `age`, `sex`, and the top three principal components. Because the number of genes is too large, we first screen the 16,610 genes down to 200 genes according to their marginal p-values from the Sequence Kernel Association Test (SKAT). Then, we perform a penalized estimation of the 200 variance components in the joint model (4.1). This is similar to the sure independence screening

Table 2. Estimation and selection results for Setting 1.

| | | | Variance components selection | | | |
|---|---|---|---|---|---|---|
| m | Criteria | MSE ($\beta$) | True Positive (3) | False Positive (0) | Exact | Over |
| 5 | AIC | 0.31(0.20) | 2.98 | 0.33 | 66% | 32% |
| | BIC | 0.31(0.20) | 2.98 | 0.15 | 84% | 14% |
| 10 | AIC | 0.27(0.17) | 2.96 | 1.14 | 26% | 70% |
| | BIC | 0.29(0.18) | 2.93 | 0.61 | 50% | 44% |
| 20 | AIC | 0.26(0.16) | 2.96 | 2.01 | 11% | 86% |
| | BIC | 0.29(0.17) | 2.87 | 1.25 | 17% | 72% |
| 100 | AIC | 0.30(0.18) | 2.74 | 2.95 | 4% | 71% |
| | BIC | 0.38(0.21) | 2.50 | 0.57 | 27% | 24% |

Table 3. Estimation and selection results for Setting 2.

| | | | Variance components selection | | | |
|---|---|---|---|---|---|---|
| m | Criteria | MSE ($\beta$) | True Positive (3) | False Positive (0) | Exact | Over |
| 5 | AIC | 0.37(0.22) | 2.99 | 0.40 | 63% | 36% |
| | BIC | 0.38(0.22) | 2.99 | 0.22 | 79% | 20% |
| 10 | AIC | 0.33(0.20) | 2.98 | 1.17 | 28% | 70% |
| | BIC | 0.36(0.21) | 2.98 | 0.68 | 44% | 54% |
| 20 | AIC | 0.34(0.22) | 2.98 | 1.60 | 25% | 74% |
| | BIC | 0.38(0.24) | 2.95 | 0.85 | 39% | 58% |
| 100 | AIC | 0.37(0.19) | 2.83 | 3.31 | 3% | 80% |
| | BIC | 0.48(0.22) | 2.68 | 0.61 | 38% | 30% |

Table 4. Estimation and selection results for Setting 3.

| | | | Variance components selection | | | |
|---|---|---|---|---|---|---|
| m | Criteria | MSE ($\beta$) | True Positive (6) | False Positive (0) | Exact | Over |
| 10 | AIC | 0.78(0.30) | 5.96 | 0.84 | 34% | 62% |
| | BIC | 0.83(0.32) | 5.66 | 0.33 | 54% | 25% |
| 20 | AIC | 0.73(0.27) | 5.88 | 1.49 | 15% | 73% |
| | BIC | 0.82(0.32) | 5.56 | 0.48 | 41% | 32% |
| 40 | AIC | 1.04(0.33) | 5.68 | 1.96 | 15% | 57% |
| | BIC | 1.17(0.37) | 4.96 | 0.74 | 29% | 27% |
| 100 | AIC | 0.85(0.34) | 5.40 | 2.54 | 2% | 48% |
| | BIC | 0.98(0.38) | 4.82 | 0.63 | 12% | 14% |

Table 5. Estimation and selection results for Setting 4.

| | | | Variance components selection | | | |
|---|---|---|---|---|---|---|
| m | Criteria | MSE ($\beta$) | True Positive (6) | False Positive (0) | Exact | Over |
| 10 | AIC | 1.06(0.32) | 5.97 | 0.85 | 32% | 65% |
| | BIC | 1.09(0.32) | 5.91 | 0.56 | 45% | 47% |
| 20 | AIC | 1.02(0.34) | 5.96 | 1.36 | 15% | 81% |
| | BIC | 1.07(0.34) | 5.92 | 0.70 | 38% | 54% |
| 40 | AIC | 1.44(0.39) | 5.74 | 1.82 | 13% | 62% |
| | BIC | 1.51(0.40) | 5.54 | 0.85 | 29% | 39% |
| 100 | AIC | 1.18(0.42) | 5.72 | 2.10 | 6% | 68% |
| | BIC | 1.29(0.43) | 5.29 | 0.71 | 21% | 22% |

Table 6. Top five genes selected by (1) the lasso penalized variance component model (3.13) with the AIC (PLVC-AIC) and (2) SKAT in an association study of 200 genes and the binary trait `smoke`.

| | PLVC-AIC | | | SKAT | | |
|---|---|---|---|---|---|---|
| No. | Gene | Marginal p-value | # Variants | Gene | Marginal p-value | # Variants |
| 1 | AFAP1L2 | $6.0 \times 10^{-4}$ | 18 | KIAA1377 | $5.7 \times 10^{-4}$ | 14 |
| 2 | RREB1 | $6.0 \times 10^{-4}$ | 18 | RREB1 | $6.0 \times 10^{-4}$ | 18 |
| 3 | KIAA1377 | $5.7 \times 10^{-4}$ | 14 | AFAP1L2 | $6.0 \times 10^{-4}$ | 18 |
| 4 | PSG5 | $3.7 \times 10^{-3}$ | 11 | KARS | $6.1 \times 10^{-4}$ | 15 |
| 5 | TDRD1 | $1.2 \times 10^{-3}$ | 14 | PZP | $1.0 \times 10^{-3}$ | 21 |

Table 7. Five-fold cross-validation performance on prediction accuracy with the top five genes selected by PLVC-AIC and SKAT added to the model in an association study of 200 genes and the complex trait `smoke`.

| | Prediction accuracy | |
|---|---|---|
| No. of genes entered into model | PLVC-AIC | SKAT |
| 1 | 79.4%(6.2%) | 78.2%(4.6%) |
| 2 | 79.9%(6.0%) | 77.9%(2.9%) |
| 3 | 80.7%(4.1%) | 80.7%(4.1%) |
| 4 | 81.7%(2.3%) | 80.7%(5.4%) |
| 5 | 81.4%(3.4%) | 78.7%(5.8%) |

strategy for selecting mean effects (Fan and Lv (2008)). The AIC selects 16 genes, whereas the BIC selects only one gene "AFAP1L2". Table 6 lists the top five genes selected using the AIC (PLVC-AIC) and SKAT. We find that the top three genes selected using the two methods are the same, but in a different order. To compare the selection performance between SKAT and PLVC-AIC, we eval-

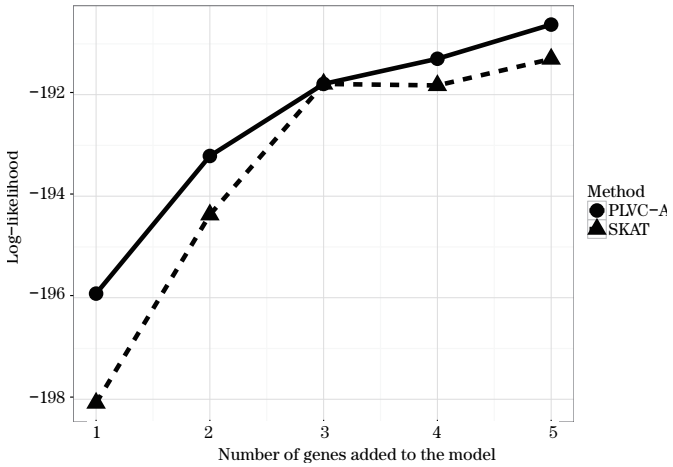


Figure 1. Log-likelihood evaluation with the top five genes selected by PLVC-AIC and SKAT added to the model in an association study of 200 genes and the complex trait `smoke`.

uate the log-likelihood of model (4.1) using the top five genes listed in Table 6, inputted to the model individually. To evaluate the log-likelihood, we use the R package `bernor`, which implements the Monte Carlo approximation method described in Sung and Geyer (2007). From Figure 1, we find that the log-likelihood with genes selected by PLVC-AIC is above that of SKAT, indicating that genes selected by PLVC-AIC explain more variability in the model.

In addition, we compare the prediction performance between the top five genes selected by PLVC-AIC and SKAT. We evaluate the prediction performance using model (4.1) by including the genotype matrix $G_i$ of the corresponding selected genes, similarly to the approach adopted in Wu et al. (2011). For example, if the genotype matrix of the top $k$ genes selected is $G_{h_1}, G_{h_2}, \ldots, G_{h_k}$, then the predictive model becomes

$$g(\mu) = X\beta + s_{h_1}^{-1/2} G_{h_1}\gamma_1 + \cdots + s_{h_k}^{-1/2} G_{h_k}\gamma_k = X^*\beta^*,$$

where $X^* = (X, s_{h_1}^{-1/2} G_{h_1}, \ldots, s_{h_k}^{-1/2} G_{h_k})$ and $\beta^* = \left(\beta^T, \gamma_1^T, \ldots, \gamma_k^T\right)$. This is the ordinary logistic regression model that can be used for predictions. Table 7 summarizes the prediction performance using five-fold cross validation as the top five genes selected by both methods are inputted to model (4.1), one by one. We find that, on average, the model with genes selected by PLVC-AIC performs slightly better than SKAT in terms of prediction. The penalization approach for selecting variance components warrants further theoretical study. This real-

data analysis demonstrates that the proposed simple MM algorithm scales to high-dimensional problems.

## 6. Discussion

This paper discusses two MM algorithms for variance component estimation and selection using the logistic linear mixed model. The algorithms are simple to implement and scale to models with a large number of variance components. Other extensions are possible. Here, we only consider binary responses. Extending the algorithm MMLA1 to Poisson count data is straightforward, with an almost identical derivation. Several studies have examined the selection of fixed effects using GLMMs. Here, we focus only on the selection of random effects. Our algorithms can be extended easily to select fixed and random effects simultaneously. We leave these topics to future research.

## Supplementary Materials

The supplementary material includes detailed derivations of (3.2), and (3.8), and a technical proof of the ascent property in (3.6).

## Acknowledgment

We sincerely thank the AE and two reviewers for their valuable comments and suggestions on this manuscript. HZ is partially supported by National Science Foundation (NSF) grant DMS-1310319 and National Institutes of Health (NIH) grants HG006139, GM105785, and GM53275. JZ is supported by NIH grant K01DK106116 and partially by DHS-14-GPD-044-000-98 and U01AI122275. The COPDGene study is supported by NIH R01 HL089856 and R01 HL089897. The exome sequencing was supported by the NHLBI Exome Sequencing Project.

# References

Ahn, M., Zhang, H. H. and Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica* **22**, 1539.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological)*, 203–210.

Bates, D., Mchler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48.

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.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 265–285.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9–25.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62**, 446–457.

Che, X. and Xu, S. (2012). Generalized linear mixed models for mapping multiple quantitative trait loci. *Heredity* **109**, 41–49.

Davidian, M. and Gallant, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Pharmacodynamics* **20**, 529–556.

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Geyer, C. (1990). Likelihood and Exponential Families. PhD thesis, University of Washington.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing* **24**, 137–154.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician* **58**, 30–37.

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.

Knudson, C. (2016). glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation. R package version 1.1.1.

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9**, 1–59. With discussion, and a rejoinder by Hunter and Lange.

Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, **95** 5–23.

McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized Linear mixed Models*. Wiley Online Library.

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing* **24**, 725–738.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Quené, H. and Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* **59**, 413–425.

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K. and Crapo, J. D. (2011). Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease* **7**, 32–43.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian

models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)* **71**, 319–392.

Schelldorfer, J., Meier, L. and Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using L1-penalization. *Journal of Computational and Graphical Statistics* **23**, 460–477.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 749–760.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1.

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC press.

Sung, Y. J. and Geyer, C. J. (2007). Monte carlo likelihood inference for missing data models. *The Annals of Statistics*, 990–1011.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, 791–795.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93.

Yi, N. and Xu, S. (1999). Mapping quantitative trait loci for complex binary traits in outbred populations. *Heredity* **82**, 668–676.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.

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

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

E-mail: lhu@ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

E-mail: lu@stat.ncsu.edu

Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ 85721-0066, U.S.A.

E-mail: jzhou@email.arizona.edu

Department of Biostatistics, University of California, Los Angeles, California 90095-1772, U.S.A.

E-mail: huazhou@ucla.edu

(Received ; accepted )