

ABSTRACT

HU, LIUYI. MM Algorithms for Variance Components Models. (Under the direction of Wenbin Lu and Hua Zhou.)

The classical linear regression model assumes independence among the observations. However, in many real applications such as clustered data analysis and longitudinal data analysis, observations are correlated with each other and people are interested in estimating the variance from different sources. For example, in genetic analysis, people coming from the same family share common genetic information and researchers would like to know how much variation in the traits can be explained by the genetic effect. Despite the best efforts of generations of statisticians and numerical analysts, maximum likelihood estimation and restricted maximum likelihood estimation of variance component models remain numerically challenging. Building on the minorization-maximization (MM) principle, this thesis work presents novel iterative algorithms for variance components estimation as well as selection.

For the first part of the thesis, we develop MM algorithms in the linear mixed model framework where we assume the covariance structures are known. The proposed algorithm is trivial to implement and competitive on large data problems. The algorithm readily extends to more complicated problems such as linear mixed models, multivariate response models possibly with missing data, maximum a posteriori estimation and penalized estimation. We establish the global convergence of the MM algorithm to a KKT point and demonstrate, both numerically and theoretically, that it converges faster than the classical EM algorithm when the number of variance components is greater than two and all covariance matrices are positive definite.

In the second part, we extend algorithms to logistic linear mixed model, which is widely used in experimental designs and genetic analysis with binary traits. When the number of variance components is large, fitting the logistic linear mixed model is challenging. We develop two efficient and stable MM algorithms for the estimation of variance components based on the Laplace approximation of the logistic model since direct optimization of the likelihood function is intractable. One of them leads to a simple iterative soft-thresholding algorithm for variance component selection using maximum penalized approximated likelihood. We demonstrate the variance component estimation and selection performance of our algorithms by simulation studies and a real data analysis.

© Copyright 2018 by Liuyi Hu

All Rights Reserved

MM Algorithms for Variance Components Models

by
Liuyi Hu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2018

APPROVED BY:

Eric Chi

Luo Xiao

Wenbin Lu
Co-chair of Advisory Committee

Hua Zhou
Co-chair of Advisory Committee

DEDICATION

To my beloved family.

BIOGRAPHY

The author grew up in Wuhan, China. Liuyi got her bachelor's degree in Statistics from Wuhan University in 2013. After that, Liuyi was enrolled in the Ph.D program of Department of Statistics at North Carolina State University. Liuyi spent five years there and did her research under the direction Dr. Wenbin Lu and Dr. Hua Zhou. During her graduate study, she worked as Graduate Industrial Trainee at Bioventus in 2015 and worked as a summer intern at AT&T Big Data Team. Liuyi received her Ph.D degree in 2018.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Hua Zhou and Dr. Wenbin Lu, for their patient guidance, knowledge and inspiational ideas. I would also like to thank Dr. Eric Chi and Dr. Luo Xiao for serving on my committee. Dr. Hua Zhou's computing course inspired my interest in the statistical computing area and I am fortunate to have him as my co-advisor. Dr. Hua Zhou's dedication, enthusiam and passion into research has inspired me a lot. And his attitude towards work has deeply influenced my life. I am also fortunate to have Dr. Wenbin Lu as my co-advisor after Dr. Hua Zhou moved to UCLA. Dr. Wenbin Lu is very smart and knowledgeable and has always been able to provide insightful guidance during our weekly meeting. They have helped me a lot not only in research but also in career development. I am blessed to have such nice advisors.

I am also very grateful to the faculty and staff at North Carolina State Univeristy. Even though Dr. Howard Bondell is no longer my committee member, I still want to thank him for the time and valuable feedback for this thesis work. I appreciate the research assistantship position that Dr. Eric Chi offered me. I appreciate his generous guidance, patience and valuable mentorship. I would like to thank Terry Byron and Chris Waddell for their kind help on computer related problems. I also want to thank Alison McCoy for always willing to help me out.

Last but not least, I would like to thank my family for their support and unconditional love. I truly believe that I would not be here without them. My parents and my husband Yuan have always been standing by me for the ups and downs in my life. I would also like to thank my daughter Claire for all the happiness she has brought to me since she was born.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	viii
Chapter 1 Introduction	1
1.1 Variance Components Model	1
1.2 Optimization Techniques	3
Chapter 2 Preliminaries	6
2.1 The MM Principle	6
2.2 Convex Matrix Functions	7
2.3 Supporting Hyperplane Minorization	8
2.4 Quadratic Minorization	8
Chapter 3 MM Algorithms for Linear Mixed Model	10
3.1 Introduction	10
3.2 Univariate Response Model	11
3.3 Numerical Experiments	16
3.4 Global Convergence of the MM Algorithm	17
3.5 MM versus EM	26
3.6 Extensions	30
3.6.1 Multivariate Response Model	30
3.6.2 Multivariate Response Model with Missing Responses	36
3.6.3 Linear Mixed Model (LMM)	37
3.6.4 MAP Estimation	40
3.6.5 Variable Selection	41
3.7 A Numerical Example	42
3.8 Discussion	43
Chapter 4 MM Algorithms for Logistic Linear Mixed Model	45
4.1 Introduction	45
4.2 Algorithms for Estimation	47
4.2.1 Model Formulation 1	47
4.2.2 Model Formulation 2	52
4.2.3 MM Algorithm for Maximizing the Penalized Approximated Likelihood	55
4.2.4 Choice of Regularization Parameter	56
4.3 Simulation Studies	57
4.3.1 Random Effects ANOVA	57
4.3.2 Genetic Example	60
4.4 Real Data Analysis	62
4.5 Discussion	63
References	66

Appendix	75
Appendix A Supplementary Materials for Chapter 3	76
A.1 EM Algorithm for the Multivariate Response Model	76

LIST OF TABLES

Table 3.1	Average iterations until convergence for MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.	18
Table 3.2	Average run times ($\times 10^{-3}$ seconds) of MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.	19
Table 3.3	Rooted mean squared error (RMSE) of $\hat{\sigma}^2$ using MM, quasi-Newton, accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.	20
Table 3.4	Rooted mean squared error (RMSE) of fixed effects and variance components in the genetic model. Standard errors are given in parentheses.	21
Table 3.5	Average performance of MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a genetic model. Standard errors are given in parentheses.	22
Table 3.6	Top 10 genes selected by the lasso penalized variance component model (3.18) in an association study of 200 genes and the complex trait height	43
Table 4.1	Comparison of the MM algorithms with two different parameterizations (MMLA1 and MMLA2) and the <code>glmer()</code> function (with <code>nAGQ=1</code>) in the <code>lme4</code> package, <code>rstanarm</code> package, and <code>glmm</code> package. Standard errors are given in parentheses. Results for <code>rstanarm</code> and <code>glmm</code> with $c = 100, 200$ are not reported because the simulation takes more than 1 week.	59
Table 4.2	Estimation and selection results for Setting 1.	61
Table 4.3	Estimation and selection results for Setting 2.	61
Table 4.4	Estimation and selection results for Setting 3.	62
Table 4.5	Estimation and selection results for Setting 4.	62
Table 4.6	Top 5 genes selected by (1) the lasso penalized variance component model (4.17) with AIC criterion (PLVC-AIC) and (2) SKAT in an association study of 200 genes and the binary trait smoke	64
Table 4.7	5-fold cross validation performance on prediction accuracy with top 5 genes selected by PLVC-AIC and SKAT added to the model respectively in an association study of 200 genes and the complex trait smoke	64

LIST OF FIGURES

Figure 2.1	Log-likelihood surface of a 2-variance component model and the surrogate functions of EM and MM minorizing the objective function at point $(\sigma_1^{2(t)}, \sigma_2^{2(t)}) = (18.5, 0.7)$	9
Figure 3.1	Solution path of the lasso penalized variance component model (3.18) in an association study of 200 genes and the complex trait height	44
Figure 4.1	Log-likelihood evaluation with top 5 genes selected by PLVC-AIC and SKAT added to the model respectively in an association study of 200 genes and the complex trait smoke	65

Chapter 1

Introduction

1.1 Variance Components Model

In statistics, linear regression model is an approach for modeling the linear relationship between the mean of the response variable and a set of explanatory variables (or predictors). Let us denote the set of observed response as a $n \times 1$ column vector $\mathbf{y} = (y_1, \dots, y_n)^T$ and the set of explanatory variables as a $n \times p$ predictor matrix \mathbf{X} , where each row of \mathbf{X} refers an observation and each column refers to a predictor. The classical linear regression model assumes \mathbf{y} to be a realization of a random variable \mathbf{Y} , which is normally distributed with

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I},$$

where $\boldsymbol{\beta}$ is the unknown parameter. Thus the response variables are assumed to be independent of each other and to have equal variance. However, in many applications, responses are correlated in a certain way. For example, a genetic study might try to answer what are the factors affecting people's height. It is not reasonable to assume that people's heights are independent of each other since people from the same family share common genes which might lead to similar heights. Also longitudinal studies observes a response variable repeated for each subject at different time points. The observations coming from the same subject are usually correlated with each other. Analysis that ignores the correlation structure will lead to invalid standard errors. Sometimes the observations are correlated via different structures and researchers want to study the variance of different structures. In this scenario, variance components model comes into play.

The simplest form of variance components model assumes that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \sum_{i=1}^m \sigma_i^2 \mathbf{V}_i,$$

and the $\mathbf{V}_1, \dots, \mathbf{V}_m$ are m fixed positive semidefinite matrices. The parameters of interest are $\boldsymbol{\beta}$ and variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_m^2)^T$.

For example, a genetic study wants to analyze the traits of n related individuals. It is assumed that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_1^2\boldsymbol{\Phi} + \sigma_2^2\mathbf{I})$ where $\boldsymbol{\Phi}$ is the kinship coefficient matrix summarizing genetic similarity between individuals. The total variability, having two variance components, is $\sigma_1^2 + \sigma_2^2$. The heritability effect is the proportion of variation explained by genetic effects, that is $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$.

Another example is the multilevel models where the data structure is hierarchical with sampled units nested in clusters that are themselves nested in other clusters. Suppose researchers want to study the students' performance on GRE and they sample students from a sample of schools and a sample of departments under the school. Students in the same school and same department tend to be more alike than students in different schools or different departments. A multilevel model that takes into account the random effects of different levels has the form

$$y_{ijk} = \mathbf{x}_{ijk}^T\boldsymbol{\beta} + \alpha_i + \gamma_{ij} + \epsilon_{ijk},$$

where \mathbf{x}_{ijk} is the vector of characteristics of students, $\{\alpha_i\}$ are the random effects for schools and $\{\gamma_{ij}\}$ are the random effects for departments under each school. We assume that α_i , γ_{ij} and ϵ_{ijk} are independent with normal distributions $N(0, \sigma_\alpha^2)$, $N(0, \sigma_\gamma^2)$ and $N(0, \sigma^2)$. The school level random effects account for the variability among schools from unmeasured variables such as the quality of teachers. The department level random effects account for the department characteristics such as emphasis on math or verbals. The three variance components, σ_α^2 , σ_γ^2 and σ^2 can help us understand the intraclass correlation between scores of different students in the same department and same school.

An extension to the standard variance components models is to encompass non-normal response distributions. Just like the generalized linear models (GLM), the generalized variance components model assumes that $E(\mathbf{Y}) = \boldsymbol{\mu}$, $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ where $g(\cdot)$ is some link function and $\boldsymbol{\eta} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \sum_{i=1}^m \sigma_i^2 \mathbf{V}_i.$$

For example, when the response is binary, $g(\cdot)$ can take the logic link function $g(x) = \ln(x/(1-x))$. When the response variables have counts as their possible values, we can use the log link $g(x) = \ln(x)$.

In this thesis work, we mainly work with the standard variance components model and then extend to the logistic variance components model.

1.2 Optimization Techniques

In the above mentioned variance components model, the parameters of interest include coefficients β and the variance components $\sigma^2 = (\sigma_1^2, \dots, \sigma_m^2)^T$. In statistical models, maximum likelihood estimate (MLE) of parameters are widely used. However, in most applications, there are no explicit solutions to the original optimization problem. Therefore, we need iterative methods. Two commonly used iterative methods include Newton's method and expectation maximization (EM) algorithm. Here we give an overview of these optimization methods. Throughout we reserve Greek letters for parameters and indicate the current iteration number by a superscript t .

Newton's Method

The Newton's method is built on maximizing the quadratic approximation of the original objective function. Denote the objective function as $f(\theta)$ and take second-order Taylor expansion at the current iterate $\theta^{(t)}$, we have

$$f(\theta) \approx f(\theta^{(t)}) + \nabla f(\theta^{(t)})^T (\theta - \theta^{(t)}) + \frac{1}{2} (\theta - \theta^{(t)})^T \nabla^2 f(\theta^{(t)}) (\theta - \theta^{(t)})$$

where $\nabla f(\theta^{(t)})$ is the gradient of the $f(\theta)$ evaluated at current iterate $\theta^{(t)}$ and $\nabla^2 f(\theta^{(t)})$ is the Hessian matrix of $f(\theta)$ evaluated at current iterate $\theta^{(t)}$. To maximize the quadratic function, we set its gradient to zero, which yields the following update

$$\theta^{(t+1)} = \theta^{(t)} - \left\{ \nabla^2 f(\theta^{(t)}) \right\}^{-1} \nabla f(\theta^{(t)}).$$

The Newton's method is a second-order algorithm and can converge very fast. However, it has several drawbacks. The first is that in the vanilla version each iteration involves the evaluation and inversion of the Hessian matrix, which can be very computational expensive, especially when the parameter space is high dimensional. The second is that Newton's method can not guarantee ascent property when the Hessian matrix is not negative definite. The third is that when the Hessian matrix is close to singular matrix, the inverted Hessian can be numerically unstable and the solution may diverge. There are several remedies to address these issues. Most of them involve approximating $-\nabla^2 f(\theta^{(t)})$ by a positive definite matrix, which leads to variants of Newton's method.

The general form of Newton's method is

$$\theta^{(t+1)} = \theta^{(t)} + s \left\{ M(\theta^{(t)}) \right\}^{-1} \nabla f(\theta^{(t)}) \tag{1.1}$$

where s is the step size and $M(\boldsymbol{\theta}^{(t)})$ is some approximation of $-\nabla^2 f(\boldsymbol{\theta}^{(t)})$. When $M(\boldsymbol{\theta}^{(t)}) = \mathbf{I}$, (1.1) is the same as the gradient descent method. The gradient descent method does not use the second order information, so it converges at a linear rate instead of quadratic rate. When $M(\boldsymbol{\theta}^{(t)}) = \mathbb{E} \left\{ -\nabla^2 f(\boldsymbol{\theta}^{(t)}) \right\}$, (1.1) gives the Fisher's scoring method. The Fisher's information matrix is positive semi-definite under exchangeability of expectation and differentiation, however, it could be hard to derive and computationally expensive to evaluate.

EM Algorithm

The EM algorithm was first introduced by Dempster et al. (1977) and it is widely used in models that involve observed data \mathbf{X} , unobserved latent variables \mathbf{Z} and a vector of unknown parameters $\boldsymbol{\theta}$. The MLE of $\boldsymbol{\theta}$ is the one that maximizes the marginal likelihood of the observed data

$$f(\mathbf{X} | \boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) d\mathbf{Z}$$

where $p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$ is the density function of the complete data. However the marginal likelihood is often intractable especially when \mathbf{Z} involves high-dimensional data. The EM algorithm deals with the problem by working with the complete data model. It involves two steps:

- Expectation step (E step): Calculate the conditional expectation of the log-likelihood function of the complete data under the current estimate of the parameter $\boldsymbol{\theta}^{(t)}$

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{(t)}} \{ \ln L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) \}$$

where $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$.

- Maximization step (M step): Maximize $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}).$$

By the information inequality, we have

$$\begin{aligned} & Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) - \ln f(\mathbf{X} | \boldsymbol{\theta}) \\ &= \mathbb{E}_{\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{(t)}} \{ \ln L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) \} - \ln f(\mathbf{X} | \boldsymbol{\theta}) \\ &= \mathbb{E} \left\{ \ln \frac{L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})}{f(\mathbf{X} | \boldsymbol{\theta})} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)} \right\} \\ &\leq \mathbb{E} \left\{ \ln \frac{L(\boldsymbol{\theta}^{(t)}; \mathbf{X}, \mathbf{Z})}{f(\mathbf{X} | \boldsymbol{\theta}^{(t)})} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)} \right\} \\ &= Q(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t)}) - \ln f(\mathbf{X} | \boldsymbol{\theta}^{(t)}) \end{aligned}$$

which leads to the following inequality

$$\ln f(\mathbf{X} | \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t)}) + \ln f(\mathbf{X} | \boldsymbol{\theta}^{(t)})$$

for all $\boldsymbol{\theta}$ in the parameter space. Combining with the maximization step, we have

$$\begin{aligned} \ln f(\mathbf{X} | \boldsymbol{\theta}^{(t+1)}) &\geq Q(\boldsymbol{\theta}^{(t+1)} | \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t)}) + \ln f(\mathbf{X} | \boldsymbol{\theta}^{(t)}) \\ &\geq \ln f(\mathbf{X} | \boldsymbol{\theta}^{(t)}). \end{aligned}$$

Therefore we are guaranteed that the objective function moves towards the right direction at each iteration. Unlike Newton's method which can be applied to various optimization problems, EM algorithm depends on the construction of latent variables and complete data likelihood. Therefore, EM algorithm is more restricted and is often seen in maximizing likelihood functions.

In this thesis work, we develop a new set of algorithms based on MM principle, which will be described in Chapter 2, for maximizing the likelihood function of the variance components model and compare it to the Newton's method as well as the EM algorithm.

Chapter 2

Preliminaries

2.1 The MM Principle

The MM principle for maximizing an objective function $f(\boldsymbol{\theta})$ involves minorizing the objective function $f(\boldsymbol{\theta})$ by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ around the current iterate $\boldsymbol{\theta}^{(t)}$ of a search (Lange et al., 2000). Minorization is defined by the two conditions

$$\begin{aligned} f(\boldsymbol{\theta}^{(t)}) &= g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\ f(\boldsymbol{\theta}) &\geq g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}), \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}^{(t)}. \end{aligned} \tag{2.1}$$

In other words, the surface $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ lies below the surface $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ and is tangent to it at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Construction of the minorizing function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ constitutes the first M of the MM algorithm. The second M of the algorithm maximizes the surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than $f(\boldsymbol{\theta})$. The point $\boldsymbol{\theta}^{(t+1)}$ maximizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ satisfies the ascent property $f(\boldsymbol{\theta}^{(t+1)}) \geq f(\boldsymbol{\theta}^{(t)})$. This fact follows from the inequalities

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

reflecting the definition of $\boldsymbol{\theta}^{(t+1)}$ and the tangency and domination conditions (2.1). The ascent property makes the MM algorithm remarkably stable. The validity of the descent property depends only on increasing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, not on maximizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. With obvious changes, the MM algorithm also applies to minimization rather than to maximization. To minimize a function $f(\boldsymbol{\theta})$, we majorize it by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ and minimize $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ to produce the next iterate $\boldsymbol{\theta}^{(t+1)}$. The acronym should not be confused with the maximization-maximization algorithm in the variational Bayes context (Jeon, 2012).

The MM principle (De Leeuw, 1994; Heiser, 1995; Kiers, 2002; Lange et al., 2000; Hunter and Lange, 2004a; Wu and Lange, 2010) finds applications in multidimensional scaling (Borg

and Groenen, 2005), ranking of sports teams (Hunter, 2004), variable selection (Hunter and Li, 2005), optimal experiment design (Yu, 2010), multivariate statistics (Zhou and Lange, 2010), geometric programming (Lange and Zhou, 2014), and many other areas (Lange, 2016). The celebrated EM principle (Dempster et al., 1977) is a special case of the MM principle. The Q function produced in the E step of an EM algorithm minorizes the log-likelihood up to an irrelevant constant. Thus, both EM and MM share the same advantages: simplicity, stability, graceful adaptation to constraints, and the tendency to avoid large matrix inversion. The more general MM perspective frees algorithm derivation from the missing data straitjacket and invites wider applications (Wu and Lange, 2010). Figure 2.1 shows the minorization functions of EM and MM for a variance components model with $m = 2$ variance components.

EM and MM algorithms often exhibit slow convergence. Fortunately, this defect can be remedied by off-the-shelf acceleration techniques for fixed point iterations. The recently developed squared iterative method (SQUAREM) (Varadhan and Roland, 2008) and the quasi-Newton acceleration method (Zhou et al., 2011) are particularly attractive, given their simplicity and minimal memory and computational costs. Our numerical experiments feature the unadorned MM algorithm and the quasi-Newton accelerated MM (aMM) algorithm based on one secant pair. Using more secant pairs is likely to further improve performance.

2.2 Convex Matrix Functions

For symmetric matrices we write $\mathbf{A} \preceq \mathbf{B}$ when $\mathbf{B} - \mathbf{A}$ is positive semidefinite and $\mathbf{A} \prec \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive definite. A matrix-valued function f is said to be (matrix) convex if

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

for all \mathbf{A}, \mathbf{B} , and $\lambda \in [0, 1]$. Our derivation of the MM variance components algorithm hinges on the convexity of the two functions mentioned in the next lemma.

Lemma 1. (a) *The matrix fractional function $f(\mathbf{A}, \mathbf{B}) = \mathbf{A}^T \mathbf{B}^{-1} \mathbf{A}$ is jointly convex in the $m \times n$ matrix \mathbf{A} and the $m \times m$ positive definite matrix \mathbf{B} .* (b) *The log determinant function $f(\mathbf{B}) = \ln \det \mathbf{B}$ is concave on the set of positive definite matrices.*

Proof. The matrix fractional function is matrix convex because its epigraph

$$\{(\mathbf{A}, \mathbf{B}, \mathbf{C}) : \mathbf{B} \succ \mathbf{0}, \mathbf{A}^T \mathbf{B}^{-1} \mathbf{A} \preceq \mathbf{C}\} = \left\{ (\mathbf{A}, \mathbf{B}, \mathbf{C}) : \mathbf{B} \succ \mathbf{0}, \begin{pmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{C} \end{pmatrix} \succeq \mathbf{0} \right\}$$

is a convex set. Here \mathbf{C} varies over the set of $n \times n$ positive semidefinite matrices. The equivalence of these two epigraph representations is proved in Boyd and Vandenberghe (2004, A.5.5). For

the concavity of the log determinant, see Boyd and Vandenberghe (2004, p74). □

2.3 Supporting Hyperplane Minorization

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

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

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

Since the negative log determinant function $f(\mathbf{B}) = -\log \det \mathbf{B}$ is convex on the set of positive definite matrices (Boyd and Vandenberghe, 2004) and the supporting hyperplane of $f(\mathbf{B})$ is

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

the supporting hyperplane minorization described above yields the following inequality

$$-\log \det \mathbf{B} \geq -\log \det \mathbf{B}^{(t)} - \text{tr} \left\{ \left(\mathbf{B}^{(t)} \right)^{-1} \left(\mathbf{B} - \mathbf{B}^{(t)} \right) \right\}. \quad (2.4)$$

2.4 Quadratic Minorization

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

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

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

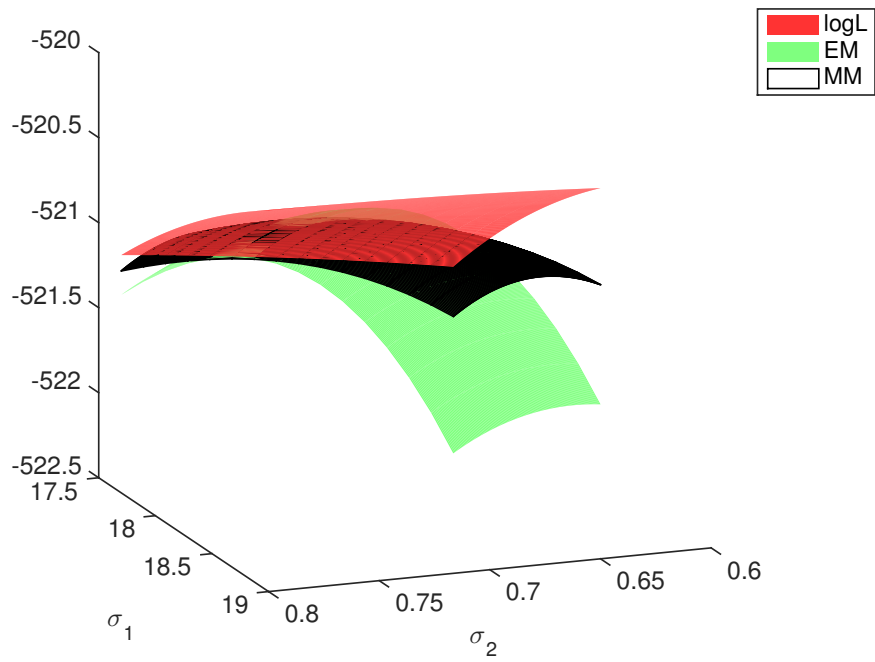


Figure 2.1: Log-likelihood surface of a 2-variance component model and the surrogate functions of EM and MM minorizing the objective function at point $(\sigma_1^{2(t)}, \sigma_2^{2(t)}) = (18.5, 0.7)$.

Chapter 3

MM Algorithms for Linear Mixed Model

3.1 Introduction

Variance components and linear mixed models are among the most potent tools in a statistician's toolbox. They are essential topics in graduate-level linear model courses and the subject of many current papers and research monographs (Rao and Kleffe, 1988; Searle et al., 1992; Rao, 1997; Khuri et al., 1998; Demidenko, 2013). Their applications in agriculture, biology, economics, genetics, epidemiology, and medicine are too numerous to cover here in detail. The recommended books (Verbeke and Molenberghs, 2000; Weiss, 2005; Fitzmaurice et al., 2011) stress longitudinal data analysis.

Given an observed $n \times 1$ response vector \mathbf{y} and $n \times p$ predictor matrix \mathbf{X} , the simplest variance components model postulates that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \sum_{i=1}^m \sigma_i^2 \mathbf{V}_i,$$

and the $\mathbf{V}_1, \dots, \mathbf{V}_m$ are m fixed positive semidefinite matrices. The parameters of the model can be divided into mean effects $(\beta_1, \dots, \beta_p)$ and variance components $(\sigma_1^2, \dots, \sigma_m^2)$, summarized by vectors $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Throughout we assume $\boldsymbol{\Omega}$ is positive definite. The extension to singular $\boldsymbol{\Omega}$ will not be pursued here. Estimation revolves around the log-likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (3.1)$$

Among the commonly used methods for estimating variance components, maximum likelihood estimation (MLE) (Hartley and Rao, 1967) and restricted (or residual) MLE (REML) (Harville,

1977) are the most popular. REML first projects \mathbf{y} to the null space of \mathbf{X} and then estimates variance components based on the projected responses. If the columns of the matrix \mathbf{B} span the null space of \mathbf{X}^T , then REML estimates the σ_i^2 by maximizing the log-likelihood of the redefined response vector $\mathbf{B}^T\mathbf{Y}$, which is normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{B}^T\boldsymbol{\Omega}\mathbf{B} = \sum_{i=1}^m \sigma_i^2 \mathbf{B}^T \mathbf{V}_i \mathbf{B}$.

There exists a large literature on iterative algorithms for finding MLE and REML (Laird and Ware, 1982; Lindstrom and Bates, 1988, 1990; Harville and Callanan, 1990; Callanan and Harville, 1991; Bates and Pinheiro, 1998; Schafer and Yucel, 2002). Fitting variance components models remains a challenge in models with a large sample size n or a large number of variance components m . Newton’s method (Lindstrom and Bates, 1988) converges quickly but is numerically unstable owing to the non-concavity of the log-likelihood. Fisher’s scoring algorithm replaces the observed information matrix in Newton’s method by the expected information matrix and yields an ascent algorithm when safeguarded by step halving. However the calculation and inversion of expected information matrices cost $O(mn^3) + O(m^3)$ flops for unstructured \mathbf{V}_i and quickly become impractical when either n or m is large. The expectation-maximization (EM) algorithm initiated by Dempster et al. is a third alternative (Dempster et al., 1977; Laird and Ware, 1982; Laird et al., 1987; Lindstrom and Bates, 1988; Bates and Pinheiro, 1998). Compared to Newton’s method, the EM algorithm is easy to implement and numerically stable, but painfully slow to converge. In practice, a strategy of priming Newton’s method by a few EM steps leverages the stability of EM and the faster convergence of second-order methods. Quasi-Newton methods dispense with explicit calculation of the observed information while achieving a superlinear rate of convergence.

In this chapter we derive a minorization-maximization (MM) algorithm for finding the MLE and REML estimates of variance components. We prove global convergence of the MM algorithm to a Karush-Kuhn-Tucker (KKT) point and explain why MM generally converges faster than EM for models with more than two variance components. We also sketch extensions of the MM algorithm to the multivariate response model with possibly missing responses, the linear mixed model (LMM), maximum a posteriori (MAP) estimation and penalized estimation. The numerical efficiency of the MM algorithm is illustrated through simulated data sets and a genomic example with more than 200 variance components.

3.2 Univariate Response Model

Our strategy for maximizing the log-likelihood (3.1) is to alternate updating the mean parameters $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\sigma}^2$. Updating $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ is a standard general least squares

problem with solution

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^T \boldsymbol{\Omega}^{-(t)} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-(t)} \mathbf{y}, \quad (3.2)$$

where $\boldsymbol{\Omega}^{-(t)}$ represents the inverse of $\boldsymbol{\Omega}^{(t)} = \sum_{i=1}^m \sigma_i^{2(t)} \mathbf{V}_i$. Updating $\boldsymbol{\sigma}^2$ given $\boldsymbol{\beta}^{(t)}$ depends on two minorizations. If we assume that all of the \mathbf{V}_i are positive definite, then the joint convexity of the map $(\mathbf{X}, \mathbf{Y}) \mapsto \mathbf{X}^T \mathbf{Y}^{-1} \mathbf{X}$ for positive definite \mathbf{Y} implies that

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} &= \left(\sum_{i=1}^m \sigma_i^{2(t)} \mathbf{V}_i \right) \left(\sum_{i=1}^m \sigma_i^2 \mathbf{V}_i \right)^{-1} \left(\sum_{i=1}^m \sigma_i^{2(t)} \mathbf{V}_i \right) \\ &= \left(\sum_{i=1}^m \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}} \frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \mathbf{V}_i \right) \cdot \left(\sum_{i=1}^m \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}} \frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^2 \mathbf{V}_i \right)^{-1} \\ &\quad \cdot \left(\sum_{i=1}^m \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}} \frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \mathbf{V}_i \right) \\ &\preceq \sum_{i=1}^m \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}} \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \mathbf{V}_i \right) \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^2 \mathbf{V}_i \right)^{-1} \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \mathbf{V}_i \right) \\ &= \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i \mathbf{V}_i^{-1} \mathbf{V}_i \\ &= \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i. \end{aligned}$$

When one or more of the \mathbf{V}_i are rank deficient, we replace each \mathbf{V}_i by $\mathbf{V}_{i,\epsilon} = \mathbf{V}_i + \epsilon \mathbf{I}$ for $\epsilon > 0$ small and let $\boldsymbol{\Omega}_\epsilon^{(t)} = \sum_i \sigma_i^{2(t)} \mathbf{V}_{i,\epsilon}$. Sending ϵ to 0 in the just proved majorization

$$\boldsymbol{\Omega}_\epsilon^{(t)} \boldsymbol{\Omega}_\epsilon^{-1} \boldsymbol{\Omega}_\epsilon^{(t)} \preceq \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_{i,\epsilon}$$

gives the desired majorization

$$\boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} \preceq \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i$$

in the general case. Negating both sides leads to the minorization

$$-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \succeq -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-(t)} \left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i \right) \boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.3)$$

that effectively separates the variance components $\sigma_1^2, \dots, \sigma_m^2$ in the quadratic term of the log-likelihood (3.1).

The convexity of the function $\mathbf{A} \mapsto -\log \det \mathbf{A}$ is equivalent to the supporting hyperplane minorization

$$-\ln \det \boldsymbol{\Omega} \geq -\ln \det \boldsymbol{\Omega}^{(t)} - \text{tr}\{\boldsymbol{\Omega}^{- (t)}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^{(t)})\} \quad (3.4)$$

that separates $\sigma_1^2, \dots, \sigma_m^2$ in the log determinant term of the log-likelihood (3.1). Combination of the minorizations (3.3) and (3.4) gives the overall minorization

$$\begin{aligned} & g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)}) \\ &= -\frac{1}{2} \text{tr}(\boldsymbol{\Omega}^{- (t)} \boldsymbol{\Omega}) - \frac{1}{2} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{- (t)} \left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i \right) \boldsymbol{\Omega}^{- (t)} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)}) + c^{(t)} \quad (3.5) \\ &= \sum_{i=1}^m \left\{ -\frac{\sigma_i^2}{2} \text{tr}(\boldsymbol{\Omega}^{- (t)} \mathbf{V}_i) - \frac{1}{2} \frac{\sigma_i^{4(t)}}{\sigma_i^2} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{- (t)} \mathbf{V}_i \boldsymbol{\Omega}^{- (t)} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)}) \right\} + c^{(t)}, \end{aligned}$$

where $c^{(t)}$ is an irrelevant constant. Maximization of $g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)})$ with respect to σ_i^2 yields the lovely multiplicative update

$$\sigma_i^{2(t+1)} = \sigma_i^{2(t)} \sqrt{\frac{(\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{- (t)} \mathbf{V}_i \boldsymbol{\Omega}^{- (t)} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(t)})}{\text{tr}(\boldsymbol{\Omega}^{- (t)} \mathbf{V}_i)}}, \quad i = 1, \dots, m. \quad (3.6)$$

To preserve the uniqueness and continuity of the algorithm map, we must take $\sigma_i^{2(t+1)} = 0$ whenever $\sigma_i^{2(t)} = 0$. As a sanity check on our derivation, consider the partial derivative

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{1}{2} \text{tr}(\boldsymbol{\Omega}^{-1} \mathbf{V}_i) + \frac{1}{2} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1} \mathbf{V}_i \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}). \quad (3.7)$$

Given $\sigma_i^{2(t)} > 0$, it is clear from the update formula (3.6) that $\sigma_i^{2(t+1)} < \sigma_i^{2(t)}$ when $\frac{\partial}{\partial \sigma_i^2} L < 0$. Conversely $\sigma_i^{2(t+1)} > \sigma_i^{2(t)}$ when $\frac{\partial}{\partial \sigma_i^2} L > 0$. Algorithm 1 summarizes the MM algorithm for MLE of the univariate response model (3.1).

The update formula (3.6) assumes that the numerator under the square root sign is nonnegative and the denominator is positive. The numerator requirement is a consequence of the positive semidefiniteness of \mathbf{V}_i . The denominator requirement can be verified through the Hadamard (elementwise) product representation $\text{tr}(\boldsymbol{\Omega}^{- (t)} \mathbf{V}_i) = \mathbf{1}^T (\boldsymbol{\Omega}^{- (t)} \odot \mathbf{V}_i) \mathbf{1}$. The following lemma of Schur (1911) is crucial. We give a self-contained probabilistic proof.

Lemma 2 (Schur). *The Hadamard product of a positive definite matrix with a positive semidefinite matrix with positive diagonal entries is positive definite.*

<p>Input : $\mathbf{y}, \mathbf{X}, \mathbf{V}_1, \dots, \mathbf{V}_m$ Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \dots, \hat{\sigma}_m^2$</p> <p>1 Initialize $\sigma_i^{(0)} > 0, i = 1, \dots, m$; 2 repeat 3 $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m \sigma_i^{2(t)} \mathbf{V}_i$; 4 $\boldsymbol{\beta}^{(t)} \leftarrow \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ $\sigma_i^{2(t+1)} \leftarrow \sigma_i^{2(t)} \sqrt{\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)} \mathbf{V}_i \boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})}{\text{tr}(\boldsymbol{\Omega}^{-(t)} \mathbf{V}_i)}}$, $i = 1, \dots, m$; 5 until objective value converges;</p>

Algorithm 1: MM algorithm for MLE of the variance components of model (3.1).

Proof. Let $\mathbf{X} = (X_1, \dots, X_n)^T$ be a random normal vector with mean $\mathbf{0}$ and positive definite covariance matrix \mathbf{A} . Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ be a random normal vector independent of \mathbf{X} with mean $\mathbf{0}$ and positive semidefinite covariance matrix \mathbf{B} having positive diagonal entries. Then $\mathbf{Z} = \mathbf{X} \odot \mathbf{Y}$ has covariances $E(Z_i Z_j) = E(X_i Y_i X_j Y_j) = E(X_i X_j) E(Y_i Y_j) = a_{ij} b_{ij}$. It follows that $\text{Cov}(\mathbf{Z}) = \mathbf{A} \odot \mathbf{B}$. To show $\mathbf{A} \odot \mathbf{B}$ is positive definite, suppose on the contrary that $\mathbf{v}^T (\mathbf{A} \odot \mathbf{B}) \mathbf{v} = \text{Var}(\mathbf{v}^T \mathbf{Z}) = 0$ for some $\mathbf{v} \neq \mathbf{0}$. Then

$$0 = \text{Var}(\mathbf{v}^T \mathbf{Z}) = E\left(\sum_i v_i X_i Y_i\right)^2 = E\left\{\left(\sum_i v_i X_i Y_i\right)^2 \mid \mathbf{Y}\right\} = E\{(\mathbf{v} \odot \mathbf{Y})^T \mathbf{A} (\mathbf{v} \odot \mathbf{Y})\}$$

implies $\mathbf{v} \odot \mathbf{Y} = \mathbf{0}$ with probability 1. Since $\mathbf{v} \neq \mathbf{0}$, $Y_i = 0$ with probability 1 for some i . This contradicts the assumption $b_{ii} = \text{Var}(Y_i) > 0$ for all i . \square

We can now obtain the following characterization of the MM iterates.

Proposition 1. *Assume \mathbf{V}_i has strictly positive diagonal entries. Then $\text{tr}(\boldsymbol{\Omega}^{-(t)} \mathbf{V}_i) > 0$ for all t . Furthermore if $\sigma_i^{2(0)} > 0$ and $\boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) \notin \text{null}(\mathbf{V}_i)$ for all t , then $\sigma_i^{2(t)} > 0$ for all t . When \mathbf{V}_i is positive definite, $\sigma_i^{2(t)} > 0$ holds if and only if $\mathbf{y} \neq \mathbf{X}\boldsymbol{\beta}^{(t)}$.*

Proof. The first claim follows easily from Schur's lemma. The second claim follows by induction. The third claim follows from the observation that $\text{null}(\mathbf{V}_i) = \{\mathbf{0}\}$. \square

In most applications, $\mathbf{V}_m = \mathbf{I}$. Proposition 1 guarantees that if $\sigma_m^{2(0)} > 0$ and the residual vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}$ is nonzero, then $\sigma_m^{2(t)}$ remains positive and thus $\boldsymbol{\Omega}^{(t)}$ remains positive definite throughout all iterations. This fact does not prevent any of the sequences $\sigma_i^{2(t)}$ from converging to 0. In this sense, the MM algorithm acts like an interior point method, approaching the optimum from inside the feasible region.

Univariate Response: Two Variance Components

<p>Input : $\mathbf{y}, \mathbf{X}, \mathbf{V}_1, \mathbf{V}_2$ Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$</p> <ol style="list-style-type: none"> 1 Simultaneous congruence decomposition: $(\mathbf{D}, \mathbf{U}) \leftarrow (\mathbf{V}_1, \mathbf{V}_2)$; 2 Transform data: $\tilde{\mathbf{y}} \leftarrow \mathbf{U}^T \mathbf{y}, \tilde{\mathbf{X}} \leftarrow \mathbf{U}^T \mathbf{X}$; 3 Initialize $\sigma_1^{(0)}, \sigma_2^{(0)} > 0$; 4 repeat 5 $w_i^{(t)} \leftarrow (\sigma_1^{2(t)} d_i + \sigma_2^{2(t)})^{-1}, \quad i = 1, \dots, n$; 6 $\boldsymbol{\beta}^{(t)} \leftarrow \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n w_i^{(t)} (\tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta})^2$ 7 $\sigma_1^{2(t+1)} \leftarrow \sigma_1^{2(t)} \sqrt{\frac{(\tilde{\mathbf{y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}^{(t)})^T (\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I})^{-1} \mathbf{D} (\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I})^{-1} (\tilde{\mathbf{y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}^{(t)})}{\text{tr}\{(\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I})^{-1} \mathbf{D}\}}}$; 7 $\sigma_2^{2(t+1)} \leftarrow \sigma_2^{2(t)} \sqrt{\frac{(\tilde{\mathbf{y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}^{(t)})^T (\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I})^{-2} (\tilde{\mathbf{y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}^{(t)})}{\text{tr}\{(\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I})^{-1}\}}}$; 8 until <i>objective value converges</i>;

Algorithm 2: Simplified MM algorithm for MLE of model (3.1) with $m = 2$ variance components and $\boldsymbol{\Omega} = \sigma_1^2 \mathbf{V}_1 + \sigma_2^2 \mathbf{V}_2$.

The major computational cost of Algorithm 1 is inversion of the covariance matrix $\boldsymbol{\Omega}^{(t)}$ at each iteration. The special case of $m = 2$ variance components deserves attention as repeated matrix inversion can be avoided by invoking the simultaneous congruence decomposition for two symmetric matrices, one of which is positive definite (Rao, 1973; Horn and Johnson, 1985). This decomposition is also called the generalized eigenvalue decomposition (Golub and Van Loan, 1996; Boyd and Vandenberghe, 2004). If one assumes $\boldsymbol{\Omega} = \sigma_1^2 \mathbf{V}_1 + \sigma_2^2 \mathbf{V}_2$ and lets $(\mathbf{V}_1, \mathbf{V}_2) \mapsto (\mathbf{D}, \mathbf{U})$ be the decomposition with \mathbf{U} nonsingular, $\mathbf{U}^T \mathbf{V}_1 \mathbf{U} = \mathbf{D}$ diagonal, and $\mathbf{U}^T \mathbf{V}_2 \mathbf{U} = \mathbf{I}$, then

$$\begin{aligned}
 \boldsymbol{\Omega}^{(t)} &= \mathbf{U}^{-T} (\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I}_n) \mathbf{U}^{-1} \\
 \boldsymbol{\Omega}^{-(t)} &= \mathbf{U} (\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I}_n)^{-1} \mathbf{U}^T \\
 \det(\boldsymbol{\Omega}^{(t)}) &= \det(\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I}_n) \det(\mathbf{U}^{-T} \mathbf{U}^{-1}) \\
 &= \det(\sigma_1^{2(t)} \mathbf{D} + \sigma_2^{2(t)} \mathbf{I}_n) \det(\mathbf{V}_2).
 \end{aligned} \tag{3.8}$$

With the revised responses $\tilde{\mathbf{y}} = \mathbf{U}^T \mathbf{y}$ and the revised predictor matrix $\tilde{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, the update (3.6) requires only vector operations and costs $O(n)$ flops. Updating the fixed effects is a weighted least squares problem with the transformed data $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ and observation weights

$w_i^{(t)} = (\sigma_1^{2(t)} d_i + \sigma_2^{2(t)})^{-1}$. Algorithm 2 summarizes the simplified MM algorithm for two variance components.

3.3 Numerical Experiments

This section compares the numerical performance of MM, quasi-Newton accelerated MM, EM, and Fisher scoring on simulated data from a two-way ANOVA random effects model and a genetic model. For ease of comparison, all algorithm runs start from $\boldsymbol{\sigma}^{2(0)} = \mathbf{1}$ and terminate when the relative change $(L^{(t+1)} - L^{(t)})/(|L^{(t)}| + 1)$ in the log-likelihood is less than 10^{-6} . In order to respect nonnegativity constraint, quasi-Newton acceleration is performed on positive square root of σ_i^2 .

Two-way ANOVA: We simulated data from a two-way ANOVA random effects model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, 1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq c,$$

where $\alpha_i \sim N(0, \sigma_1^2)$, $\beta_j \sim N(0, \sigma_2^2)$, $(\alpha\beta)_{ij} \sim N(0, \sigma_3^2)$, and $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ are jointly independent. Here i indexes levels in factor 1, j indexes levels in factor 2, and k indexes observations in the (i, j) -combination. This corresponds to $m = 4$ variance components. In the simulation, we set $\sigma_2^2 = \sigma_3^2 = \sigma_e^2$ and varied the ratio σ_1^2/σ_e^2 ; the numbers of levels a and b in factor 1 and factor 2, respectively; and the number of observations c in each combination of factor levels. For each simulation scenario, we simulated 50 replicates. The sample size was $n = abc$ for each replicate.

Tables 3.1 and 3.2 show the average number of iterations and the average runtimes when there are $a = b = 5$ levels of each factor. Based on these results and further results not shown for other combinations of a and b , we draw the following conclusions. Fisher scoring takes the fewest iterations. The MM algorithm always takes fewer iterations than the EM algorithm. Accelerated MM further improves the convergence rate of MM. The faster rate of convergence of Fisher scoring is outweighed by the extra cost of evaluating and inverting the covariance matrix. When the sample size $n = abc$ is large, Fisher scoring takes much longer than either EM or MM.

Table 3.3 summarizes the the rooted mean squared error (RMSE) of the variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_e^2)$. For each replicate, RMSE is calculated as

$$\sqrt{\frac{1}{4} \sum_{j=1}^4 (\hat{\sigma}_j^2 - \sigma_j^2)^2}. \quad (3.9)$$

We can see that the estimation performance using different algorithms are comparable.

Genetic model: We simulated a quantitative trait \mathbf{y} from a genetic model with two variance components and covariance matrix $\mathbf{\Omega} = \sigma_a^2 \widehat{\mathbf{\Phi}} + \sigma_e^2 \mathbf{I}$, where $\widehat{\mathbf{\Phi}}$ is a full-rank empirical kinship matrix estimated from the genome-wide measurements of 212 individuals using Option 29 of the Mendel software (Lange et al., 2013). Table 3.4 summarizes the RMSE of each parameter: the intercept β_0 , the slope for gender β_1 , and two variance components σ_a^2 and σ_e^2 . Table 3.5 summarizes the iteration number, runtime and objective value for different algorithms. In this example, Fisher scoring excels at smaller σ_a^2/σ_e^2 ratios, while accelerated MM is fastest at larger σ_a^2/σ_e^2 ratios.

In summary, the MM algorithm appears competitive even in small-scale examples. Modern applications often involve a large number of variance components. In this setting, the EM algorithm suffers from slow convergence and Fisher scoring from an extremely high cost per iteration. Our genomic example in Section 3.7 reinforces this point.

3.4 Global Convergence of the MM Algorithm

The KKT necessary conditions for a local maximum $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_m^2)$ of the log-likelihood (3.1) require each component of the score vector to satisfy

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^2) \in \begin{cases} \{0\} & \sigma_i^2 > 0 \\ (-\infty, 0] & \sigma_i^2 = 0. \end{cases}$$

In this section we establish the global convergence of Algorithm 1 to a KKT point. To reduce the notational burden, we assume that \mathbf{X} is null and omit estimation of fixed effects $\boldsymbol{\beta}$. The analysis easily extends to the MLE case. Our convergence analysis relies on characterizing the properties of the objective function $L(\boldsymbol{\sigma}^2)$ and the MM algorithmic mapping $\boldsymbol{\sigma}^2 \mapsto M(\boldsymbol{\sigma}^2)$ defined by equation (3.6). Special attention must be paid to the boundary values $\sigma_i^2 = 0$. We prove convergences for two cases, which cover most applications. The genetic model in Section 3.2 satisfies Assumption 1, while the two-way ANOVA model satisfies Assumption 2.

Assumption 1. *All \mathbf{V}_i are positive definite.*

Assumption 2. *\mathbf{V}_1 is positive definite, each \mathbf{V}_i is nontrivial, $\mathcal{H} = \text{span}\{\mathbf{V}_2, \dots, \mathbf{V}_m\}$ has dimension $q < n$, and $\mathbf{y} \notin \mathcal{H}$.*

The key condition $\mathbf{y} \notin \text{span}\{\mathbf{V}_2, \dots, \mathbf{V}_m\}$ in the second case is critical for the existence of an MLE or REML (Demidenko and Massam, 1999; Grzadziel and Michalski, 2014). We will derive a sequence of lemmas en route to the global convergence result declared in Theorem 1.

Lemma 3. *Under Assumption 1 or 2, the log-likelihood function (3.1) is coercive in the sense that the super-level set $S_c = \{\boldsymbol{\sigma}^2 \geq \mathbf{0} : L(\boldsymbol{\sigma}^2) \geq c\}$ is compact for every c .*

Table 3.1: Average iterations until convergence for MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

σ_1^2/σ_e^2	Method	$c = \#$ observations per combination			
		2	8	20	50
0	MM	34.52(15.79)	25.90(8.69)	18.62(7.22)	15.48(5.34)
	aMM	16.68(5.69)	13.48(3.47)	11.76(3.12)	10.88(2.43)
	EM	123.70(63.72)	61.58(31.36)	38.44(18.58)	25.66(10.31)
	FS	6.10(1.09)	6.74(0.99)	6.68(0.79)	6.36(0.72)
0.05	MM	27.78(13.05)	22.82(8.96)	19.82(6.55)	15.48(3.97)
	aMM	14.80(4.33)	12.32(3.27)	12.08(2.62)	11.20(2.52)
	EM	108.04(62.58)	58.42(33.67)	43.52(19.48)	27.62(12.47)
	FS	6.20(1.29)	6.72(1.25)	6.62(0.73)	6.60(1.07)
0.1	MM	31.26(14.90)	23.38(9.21)	16.84(6.72)	14.88(4.56)
	aMM	15.96(5.65)	12.72(3.59)	10.36(2.51)	10.80(2.46)
	EM	112.12(72.70)	62.26(28.87)	34.86(22.61)	24.10(11.96)
	FS	6.10(1.25)	6.90(0.79)	6.48(0.86)	6.52(0.86)
1	MM	29.72(15.85)	22.72(10.86)	17.78(8.18)	13.94(4.73)
	aMM	15.24(5.60)	12.40(4.12)	10.72(2.70)	10.24(2.01)
	EM	85.86(63.85)	41.50(30.46)	28.40(20.02)	21.36(13.86)
	FS	5.96(1.19)	6.90(0.91)	6.36(1.05)	6.44(0.93)
10	MM	16.46(9.74)	13.28(7.75)	12.80(6.41)	10.74(3.67)
	aMM	11.60(3.70)	9.36(2.78)	9.04(3.00)	8.68(2.54)
	EM	24.50(32.87)	16.18(23.06)	15.10(16.55)	12.36(11.13)
	FS	6.98(0.80)	6.96(0.70)	6.74(0.83)	6.76(0.52)
20	MM	17.34(10.70)	14.20(6.79)	11.58(4.46)	10.16(4.26)
	aMM	12.12(5.96)	9.92(2.68)	8.92(2.07)	8.48(2.00)
	EM	31.08(42.11)	20.50(24.55)	10.84(10.86)	8.98(8.94)
	FS	7.18(0.98)	7.02(0.82)	6.90(0.74)	6.78(0.79)

Table 3.2: Average run times ($\times 10^{-3}$ seconds) of MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

σ_1^2/σ_e^2	Method	$c = \#$ observations per combination			
		2	8	20	50
0	MM	30.50(81.58)	132.50(48.51)	739.80(272.66)	4004.10(1317.90)
	aMM	16.76(36.57)	92.06(34.42)	638.76(231.86)	3691.90(873.70)
	EM	70.76(43.58)	376.82(184.27)	1912.30(918.56)	8276.98(3269.26)
	FS	17.52(27.99)	241.06(51.79)	4039.73(6955.91)	43315.42(76563.64)
0.05	MM	16.79(11.73)	117.33(50.33)	867.41(296.24)	4083.60(1016.71)
	aMM	14.23(14.21)	80.93(31.83)	692.54(196.19)	3841.10(948.55)
	EM	66.73(44.87)	376.78(206.19)	2291.69(1035.80)	9054.02(3989.52)
	FS	13.33(18.33)	253.10(61.72)	3198.17(379.00)	34057.87(7132.36)
0.1	MM	17.01(8.97)	122.03(55.77)	733.37(329.08)	3992.52(1166.67)
	aMM	12.39(8.86)	88.34(33.87)	593.27(174.93)	3745.50(922.99)
	EM	76.65(53.37)	389.45(179.41)	1814.91(1152.64)	7951.40(3810.34)
	FS	10.24(8.57)	257.63(45.53)	3140.15(481.08)	33490.67(4533.51)
1	MM	16.50(13.08)	112.94(52.73)	736.32(322.26)	3746.87(1221.92)
	aMM	9.86(4.10)	80.98(39.49)	585.98(158.65)	3536.90(754.48)
	EM	56.93(48.16)	267.86(194.49)	1465.45(986.31)	7079.60(4430.10)
	FS	15.75(17.26)	262.49(44.28)	3003.68(481.92)	33215.82(4801.38)
10	MM	10.80(11.16)	70.94(47.94)	545.71(256.97)	2316.96(1022.51)
	aMM	8.64(4.17)	62.50(26.13)	483.63(183.47)	2317.43(1061.57)
	EM	21.51(31.61)	113.76(158.82)	803.36(816.05)	3256.65(2624.29)
	FS	12.32(9.81)	261.85(37.84)	3190.52(394.75)	26163.81(8451.96)
20	MM	8.83(5.05)	104.94(54.66)	552.13(190.42)	1706.71(680.84)
	aMM	9.57(9.80)	92.94(35.84)	524.70(137.22)	1750.99(489.96)
	EM	23.13(31.17)	175.12(198.18)	642.39(576.82)	2007.86(1901.66)
	FS	12.71(11.90)	340.81(48.29)	3543.18(464.36)	18796.59(2445.74)

Table 3.3: Rooted mean squared error (RMSE) of $\hat{\sigma}^2$ using MM, quasi-Newton, accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

σ_1^2/σ_e^2	Method	$c = \#$ observations per combination			
		2	8	20	50
0	MM	0.092(0.061)	0.041(0.027)	0.024(0.019)	0.016(0.012)
	aMM	0.092(0.061)	0.040(0.027)	0.024(0.019)	0.016(0.012)
	EM	0.092(0.060)	0.040(0.027)	0.024(0.019)	0.016(0.012)
	FS	0.093(0.062)	0.041(0.027)	0.024(0.019)	0.016(0.012)
0.05	MM	0.138(0.069)	0.061(0.027)	0.042(0.020)	0.031(0.011)
	aMM	0.138(0.069)	0.062(0.027)	0.042(0.020)	0.031(0.011)
	EM	0.137(0.069)	0.061(0.027)	0.042(0.020)	0.030(0.011)
	FS	0.141(0.072)	0.062(0.027)	0.043(0.021)	0.031(0.011)
0.1	MM	0.161(0.085)	0.083(0.026)	0.063(0.024)	0.063(0.030)
	aMM	0.162(0.084)	0.083(0.026)	0.064(0.024)	0.064(0.031)
	EM	0.159(0.085)	0.082(0.026)	0.063(0.024)	0.063(0.030)
	FS	0.168(0.088)	0.085(0.027)	0.064(0.024)	0.064(0.031)
1	MM	0.563(0.206)	0.525(0.273)	0.440(0.236)	0.504(0.230)
	aMM	0.566(0.206)	0.529(0.272)	0.446(0.238)	0.512(0.230)
	EM	0.561(0.205)	0.525(0.272)	0.440(0.236)	0.505(0.230)
	FS	0.567(0.207)	0.531(0.273)	0.447(0.238)	0.514(0.231)
10	MM	4.961(2.289)	4.419(1.986)	5.253(3.416)	4.819(2.208)
	aMM	4.974(2.297)	4.446(1.993)	5.304(3.456)	4.886(2.224)
	EM	4.962(2.290)	4.427(1.988)	5.258(3.454)	4.820(2.224)
	FS	4.976(2.298)	4.447(1.994)	5.311(3.457)	4.897(2.228)
20	MM	9.552(3.922)	10.595(4.606)	9.410(4.032)	8.731(3.391)
	aMM	9.575(3.923)	10.661(4.636)	9.507(4.071)	8.844(3.443)
	EM	9.550(3.917)	10.597(4.632)	9.424(4.057)	8.733(3.429)
	FS	9.579(3.924)	10.660(4.635)	9.512(4.073)	8.848(3.447)

Table 3.4: Rooted mean squared error (RMSE) of fixed effects and variance components in the genetic model. Standard errors are given in parentheses.

σ_1^2/σ_e^2	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$
0.00	MM	0.12(0.10)	0.18(0.14)	0.27(0.27)	0.20(0.16)
	aMM	0.12(0.10)	0.18(0.14)	0.24(0.26)	0.19(0.15)
	EM	0.12(0.10)	0.18(0.14)	0.35(0.24)	0.20(0.16)
	FS	0.12(0.10)	0.18(0.14)	0.23(0.26)	0.21(0.16)
0.05	MM	0.13(0.08)	0.16(0.11)	0.28(0.29)	0.21(0.17)
	aMM	0.13(0.08)	0.16(0.11)	0.26(0.27)	0.21(0.17)
	EM	0.13(0.08)	0.16(0.11)	0.29(0.29)	0.21(0.17)
	FS	0.13(0.08)	0.16(0.11)	0.25(0.27)	0.21(0.17)
0.10	MM	0.11(0.09)	0.15(0.13)	0.30(0.30)	0.22(0.18)
	aMM	0.11(0.09)	0.15(0.13)	0.29(0.31)	0.22(0.18)
	EM	0.11(0.09)	0.15(0.13)	0.31(0.32)	0.21(0.19)
	FS	0.11(0.09)	0.15(0.13)	0.29(0.30)	0.22(0.18)
1.00	MM	0.11(0.10)	0.17(0.13)	0.65(0.45)	0.25(0.19)
	aMM	0.11(0.10)	0.17(0.13)	0.69(0.45)	0.26(0.20)
	EM	0.11(0.10)	0.17(0.13)	0.66(0.42)	0.25(0.20)
	FS	0.11(0.10)	0.17(0.13)	0.69(0.46)	0.26(0.20)
10.00	MM	0.17(0.13)	0.29(0.23)	2.84(2.10)	0.80(0.56)
	aMM	0.17(0.14)	0.30(0.24)	5.04(3.45)	1.32(0.73)
	EM	0.17(0.13)	0.29(0.23)	2.71(1.92)	0.76(0.52)
	FS	0.17(0.13)	0.30(0.23)	3.66(2.77)	0.96(0.71)
20.00	MM	0.27(0.22)	0.45(0.38)	4.46(3.55)	0.89(0.70)
	aMM	0.27(0.22)	0.46(0.39)	7.94(3.94)	1.83(0.50)
	EM	0.27(0.21)	0.45(0.38)	4.39(3.53)	0.84(0.68)
	FS	0.27(0.22)	0.45(0.38)	5.75(3.85)	1.18(0.75)

Table 3.5: Average performance of MM, quasi-Newton accelerated MM (aMM), EM, and Fisher scoring (FS) for fitting a genetic model. Standard errors are given in parentheses.

σ_a^2/σ_e^2	Method	Iteration	Runtime (10^{-3} sec)	Objective
0	MM	88.10(29.01)	778.24(305.37)	-374.35(9.82)
	aMM	23.65(5.74)	293.16(146.23)	-374.34(9.82)
	EM	231.93(123.39)	3509.02(1851.11)	-374.41(9.83)
	FS	5.05(1.24)	137.76(65.74)	-374.36(9.83)
0.05	MM	84.97(31.18)	710.56(260.24)	-377.19(10.85)
	aMM	23.05(5.45)	272.04(67.01)	-377.18(10.85)
	EM	220.57(124.70)	3292.87(1865.91)	-377.25(10.85)
	FS	5.08(1.21)	136.47(33.18)	-377.21(10.83)
0.1	MM	82.45(34.39)	673.96(268.23)	-379.62(10.54)
	aMM	22.55(6.01)	269.55(86.69)	-379.61(10.54)
	EM	199.70(113.47)	2917.71(1607.33)	-379.68(10.54)
	FS	4.97(1.03)	129.71(40.51)	-379.62(10.54)
1	MM	31.00(15.59)	160.21(80.45)	-409.66(11.26)
	aMM	12.55(5.38)	90.21(43.54)	-409.66(11.26)
	EM	51.10(28.70)	550.55(321.89)	-409.67(11.26)
	FS	4.60(0.59)	80.28(25.56)	-409.66(11.26)
10	MM	72.67(39.23)	374.80(209.31)	-532.57(9.11)
	aMM	20.15(10.18)	146.25(81.06)	-531.24(9.28)
	EM	294.20(717.05)	3079.82(7520.30)	-532.71(9.11)
	FS	10.18(4.92)	168.63(80.34)	-532.08(9.21)
20	MM	78.35(34.32)	425.40(188.08)	-591.36(7.05)
	aMM	14.80(6.53)	117.14(71.14)	-589.13(7.15)
	EM	362.07(764.60)	4144.92(8862.65)	-591.62(6.82)
	FS	10.93(4.75)	181.48(83.96)	-590.68(7.08)

Proof. Let us first prove the assertion when all of the covariance matrices \mathbf{V}_i are positive definite. If we set $r = \|\boldsymbol{\sigma}^2\|_1$ and $\alpha_i = r^{-1}\sigma_i^2$ for each i , then the log-likelihood satisfies

$$L(\boldsymbol{\sigma}^2) = -\frac{n}{2} \ln r - \frac{1}{2} \ln \det \left(\sum_{i=1}^m \alpha_i \mathbf{V}_i \right) - \frac{1}{2r} \mathbf{y}^T \left(\sum_{i=1}^m \alpha_i \mathbf{V}_i \right)^{-1} \mathbf{y}.$$

The functions $\ln \det \left(\sum_{i=1}^m \alpha_i \mathbf{V}_i \right)$ and $\mathbf{y}^T \left(\sum_{i=1}^m \alpha_i \mathbf{V}_i \right)^{-1} \mathbf{y}$ of $\boldsymbol{\alpha}$ are defined and continuous on the unit simplex and hence bounded there. The dominant term $-\frac{n}{2} \ln r$ of the loglikelihood tends to $-\infty$ as r tends to ∞ .

To prove the assertion under Assumption 2, consider first the case $\mathbf{V}_1 = \mathbf{I}_n$. Setting $\alpha_i = \sigma_i^2/\sigma_1^2$ for $i = 2, \dots, m$ reduces the loglikelihood to

$$L(\sigma_1^2, \boldsymbol{\alpha}) = -\frac{n}{2} \ln \sigma_1^2 - \frac{1}{2} \ln \det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) - \frac{1}{2\sigma_1^2} \mathbf{y}^T \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right)^{-1} \mathbf{y}. \quad (3.10)$$

The middle term on the right satisfies

$$-\frac{1}{2} \ln \det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) \leq 0$$

because $\det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) \geq \det \mathbf{I}_n = 1$. Now let $\mathbf{U} = (\mathbf{U}_q, \mathbf{U}_{n-q})$ be an $n \times n$ orthogonal matrix whose left columns \mathbf{U}_q span \mathcal{H} and whose right columns \mathbf{U}_{n-q} span \mathcal{H}^\perp . The identity

$$\mathbf{U}^T \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) \mathbf{U} = \begin{pmatrix} \mathbf{I}_q + \sum_{i=2}^m \alpha_i \mathbf{U}_q^T \mathbf{V}_i \mathbf{U}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-q} \end{pmatrix}$$

follows from the orthogonality relations $\mathbf{U}_{n-q}^T \mathbf{V}_i = \mathbf{U}_{n-q}^T \mathbf{U}_q = \mathbf{0}_{(n-q) \times n}$. This in turn implies

$$\begin{aligned} \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right)^{-1} &= \mathbf{U} \begin{pmatrix} (\mathbf{I}_q + \sum_{i=2}^m \alpha_i \mathbf{U}_q^T \mathbf{V}_i \mathbf{U}_q)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-q} \end{pmatrix} \mathbf{U}^T \\ &\succeq \mathbf{U} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-q} \end{pmatrix} \mathbf{U}^T \\ &= \mathbf{U}_{n-q} \mathbf{U}_{n-q}^T. \end{aligned}$$

Therefore the quadratic term in equation (3.10) is bounded below by the positive constant

$$\mathbf{y}^T \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right)^{-1} \mathbf{y} \geq \mathbf{y}^T \mathbf{U}_{n-q} \mathbf{U}_{n-q}^T \mathbf{y} = \|\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y}\|^2 > 0.$$

Here the assumption $\mathbf{y} \notin \mathcal{H}$ guarantees the projection property $\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y} \neq \mathbf{0}$.

Next we show that the loglikelihood tends to $-\infty$ when σ_1^2 tends to 0 or ∞ or when $\|\boldsymbol{\alpha}\|_2$ tends to ∞ . The second of the two inequalities

$$\begin{aligned} L(\sigma_0^2, \boldsymbol{\alpha}) &\leq -\frac{n}{2} \ln \sigma_1^2 - \frac{1}{2} \ln \det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) - \frac{1}{2\sigma_1^2} \|\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y}\|^2 \\ &\leq -\frac{n}{2} \ln \sigma_1^2 - \frac{1}{2\sigma_1^2} \|\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y}\|^2 \end{aligned}$$

renders the claim about σ_1^2 obvious. To prove the claim about $\boldsymbol{\alpha}$, we make the worst case choice $\sigma_i^2 = \|\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y}\|^2$ in the first inequality. It follows that

$$L(\sigma_0^2, \boldsymbol{\alpha}) \leq -\frac{1}{2} \ln \det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) - \frac{n}{2} \ln \|\mathbb{P}_{\mathcal{H}^\perp} \mathbf{y}\|^2 - \frac{n}{2}.$$

If α_j tends to ∞ , then the inequality

$$-\frac{1}{2} \ln \det \left(\mathbf{I}_n + \sum_{i=2}^m \alpha_i \mathbf{V}_i \right) \leq -\frac{1}{2} \ln \det \left(\mathbf{I}_n + \alpha_j \mathbf{V}_j \right) = -\frac{1}{2} \sum_{k=1}^n \ln(1 + \alpha_j \lambda_{jk})$$

holds, where the λ_{jk} are the eigenvalues of \mathbf{V}_j . At least one of these eigenvalues is positive because \mathbf{V}_j is nontrivial. It follows that $L(\sigma_0^2, \boldsymbol{\alpha})$ tends to $-\infty$ in this case as well.

For the general case where \mathbf{V}_1 is non-singular but not necessarily \mathbf{I}_n , let $\mathbf{V}_1^{1/2}$ be the symmetric square root of \mathbf{V}_1 and write

$$\mathbf{V}_1 + \sum_{i=2}^m \sigma_i^2 \mathbf{V}_i = \mathbf{V}_1^{1/2} \left(\mathbf{I} + \sum_{i=2}^m \sigma_i^2 \mathbf{V}_1^{-1/2} \mathbf{V}_i \mathbf{V}_1^{-1/2} \right) \mathbf{V}_1^{1/2}.$$

The above arguments still apply since each $\mathbf{V}_1^{-1/2} \mathbf{V}_i \mathbf{V}_1^{-1/2}$ is nontrivial and \mathbf{y} belongs to the span $\{\mathbf{V}_2, \dots, \mathbf{V}_m\} = \mathbf{S}$ if and only if $\mathbf{V}_1^{-1/2} \mathbf{y}$ belongs to $\mathbf{V}_1^{-1/2} \mathbf{S} \mathbf{V}_1^{-1/2}$. \square

Lemma 4. *The iterates possess the ascent property $L(M(\boldsymbol{\sigma}^{2(t)})) \geq L(\boldsymbol{\sigma}^{2(t)})$. Furthermore, when $L(M(\boldsymbol{\sigma}_*^2)) = L(\boldsymbol{\sigma}_*^2)$, $\boldsymbol{\sigma}_*^2$ fulfills the fixed point condition $M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$, and each component satisfies either (i) $\sigma_{*i}^2 = 0$ or (ii) $\sigma_{*i}^2 > 0$ and $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}_*^2) = 0$.*

Proof. The ascent property is built into any MM algorithm. Suppose $L(M(\boldsymbol{\sigma}_*^2)) = L(\boldsymbol{\sigma}_*^2)$ at a point $\boldsymbol{\sigma}_*^2 \in \mathbb{R}_+^m$. Then equality must hold in the string of inequalities (2.2). It follows that

$$g(M(\boldsymbol{\sigma}_*^2) \mid \boldsymbol{\sigma}_*^2) = g(\boldsymbol{\sigma}_*^2 \mid \boldsymbol{\sigma}_*^2).$$

$g(\cdot \mid \boldsymbol{\sigma}_*^2)$ has a unique maximum since its Hessian is diagonal with strictly negative entries,

hence $M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$. If $\sigma_{*i}^2 > 0$, the stationarity condition

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}_*^2) = \frac{\partial}{\partial \sigma_i^2} g(\boldsymbol{\sigma}_*^2 | \boldsymbol{\sigma}_*^2) = 0$$

applies. The equivalence of the two displayed partial derivatives is a consequence of the fact that the difference $f(\boldsymbol{\sigma}^2) - g(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}_*^2)$ achieves its minimum of 0 at $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}_*^2$. \square

Lemma 5. *The distance between successive iterates $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2$ converges to 0.*

Proof. Suppose on the contrary that $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2$ does not converge to 0. Then one can extract a subsequence $\{t_k\}_{k \geq 1}$ such that

$$\|\boldsymbol{\sigma}^{2(t_k+1)} - \boldsymbol{\sigma}^{2(t_k)}\|_2 \geq \epsilon > 0 \quad (3.11)$$

for all k . Let C_0 be the compact super-level set $\{\boldsymbol{\sigma}^2 : L(\boldsymbol{\sigma}^2) \geq L(\boldsymbol{\sigma}^{2(0)})\}$. Since the sequence $\{\boldsymbol{\sigma}^{2(t_k)}\}_{k \geq 1}$ is confined to C_0 , one can pass to a subsequence if necessary and assume that $\boldsymbol{\sigma}^{2(t_k)}$ converges to a limit $\boldsymbol{\sigma}_*^2$ and that $\boldsymbol{\sigma}^{2(t_k+1)}$ converges to a limit $\boldsymbol{\sigma}_{**}^2$. Taking limits in the relation $\boldsymbol{\sigma}^{2(t_k+1)} = M(\boldsymbol{\sigma}^{2(t_k)})$ and invoking the continuity $M(\boldsymbol{\sigma}^2)$ imply that $\boldsymbol{\sigma}_{**}^2 = M(\boldsymbol{\sigma}_*^2)$. Because the sequence $L(\boldsymbol{\sigma}^{2(t_k)})$ is monotonically increasing in k and bounded above on C_0 , it converges to a limit L_* . Hence, the continuity of $L(\boldsymbol{\sigma}^2)$ implies

$$L(\boldsymbol{\sigma}_*^2) = \lim_k L(\boldsymbol{\sigma}^{2(t_k)}) = L_* = \lim_k L(\boldsymbol{\sigma}^{2(t_k+1)}) = L(\boldsymbol{\sigma}_{**}^2) = L(M(\boldsymbol{\sigma}_*^2)).$$

Lemma 4 therefore gives $\boldsymbol{\sigma}_{**}^2 = M(\boldsymbol{\sigma}_*^2) = \boldsymbol{\sigma}_*^2$, contradicting the bound $\|\boldsymbol{\sigma}_*^2 - \boldsymbol{\sigma}_{**}^2\|_2 \geq \epsilon$ entailed by inequality (3.11). \square

Theorem 1. *The MM sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t \geq 0}$ has at least one limit point. Every limit point is a fixed point of $M(\boldsymbol{\sigma}^2)$. If the set of fixed points is discrete, then the MM sequence converges to one of them. Finally, when the iterates converge, their limit is a KKT point.*

Proof. The sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t \geq 0}$ is contained in the super-level compact set C_0 defined in Lemma 5 and therefore admits a convergent subsequence $\boldsymbol{\sigma}^{2(t_k)}$ with limit $\boldsymbol{\sigma}^{2(\infty)}$. As argued in Lemma 5, $L(\boldsymbol{\sigma}^{2(\infty)}) = L(M(\boldsymbol{\sigma}^{2(\infty)}))$. Lemma 4 now implies that $\boldsymbol{\sigma}^{2(\infty)}$ is a fixed point of the algorithm map $M(\boldsymbol{\sigma}^2)$.

According to Ostrowski's theorem (Lange, 2010, Proposition 8.2.1), the set of limit points of a bounded sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t \geq 0}$ is connected and compact provided $\|\boldsymbol{\sigma}^{2(t+1)} - \boldsymbol{\sigma}^{2(t)}\|_2 \rightarrow 0$. If the set of fixed points is discrete, then the connected subset of limit points reduces to a single point. Hence, the bounded sequence $\boldsymbol{\sigma}^{2(t)}$ converges to this point. When the limit exists, one can check that $\boldsymbol{\sigma}^{2(\infty)}$ satisfies the KKT conditions by proving that each zero component of $\boldsymbol{\sigma}^{2(\infty)}$

has a non-positive partial derivative. Suppose on the contrary $\sigma_i^{2(\infty)} = 0$ and $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^{2(\infty)}) > 0$. By continuity $\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^{2(t)}) > 0$ for all large t . Therefore, $\sigma_i^{2(t+1)} > \sigma_i^{2(t)}$ for all large t by the observation made after equation (3.7). This behavior is inconsistent with the assumption that $\sigma_i^{2(t)} \rightarrow 0$. \square

3.5 MM versus EM

Examination of Tables 3.2 and 3.5 suggests that the MM algorithm usually converges faster than the EM algorithm. We now provide theoretical justification for this observation. Again for notational convenience, we consider the REML case where \mathbf{X} is null. Since the EM principle is just a special instance of the MM principle, we can compare their convergence properties in a unified framework. Consider an MM map $M(\boldsymbol{\theta})$ for maximizing the objective function $f(\boldsymbol{\theta})$ via the surrogate function $g(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$. Close to the optimal point $\boldsymbol{\theta}^\infty$,

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^\infty \approx dM(\boldsymbol{\theta}^\infty)(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^\infty),$$

where $dM(\boldsymbol{\theta}^\infty)$ is the differential of the mapping M at the optimal point $\boldsymbol{\theta}^\infty$ of $f(\boldsymbol{\theta})$. Hence, the local convergence rate of the sequence $\boldsymbol{\theta}^{(t+1)} = M(\boldsymbol{\theta}^{(t)})$ coincides with the spectral radius of $dM(\boldsymbol{\theta}^\infty)$. Familiar calculations (McLachlan and Krishnan, 2008; Lange, 2010) demonstrate that

$$dM(\boldsymbol{\theta}^\infty) = \mathbf{I} - \{d^2g(\boldsymbol{\theta}^\infty | \boldsymbol{\theta}^\infty)\}^{-1} d^2f(\boldsymbol{\theta}^\infty).$$

In other words, the local convergence rate is determined by how well the surrogate surface $g(\boldsymbol{\theta} | \boldsymbol{\theta}^\infty)$ approximates the objective surface $f(\boldsymbol{\theta})$ near the optimal point $\boldsymbol{\theta}^\infty$. In the EM literature, $dM(\boldsymbol{\theta}^\infty)$ is called the *rate matrix* (Meng and Rubin, 1991). Fast convergence occurs when the surrogate $g(\boldsymbol{\theta} | \boldsymbol{\theta}^\infty)$ hugs the objective $f(\boldsymbol{\theta})$ tightly around $\boldsymbol{\theta}^\infty$. Figure 2.1 shows a case where the MM surrogate locally dominates the EM surrogate. We demonstrate that this is no accident.

McLachlan and Krishnan (2008) derive the EM surrogate

$$g_{\text{EM}}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) = -\frac{1}{2} \sum_{i=1}^m \left\{ \text{rank}(\mathbf{V}_i) \ln \sigma_i^2 + \text{rank}(\mathbf{V}_i) \frac{\sigma_i^{2(t)}}{\sigma_i^2} - \frac{\sigma_i^{4(t)}}{\sigma_i^2} \text{tr}(\boldsymbol{\Omega}^{-(t)} \mathbf{V}_i) \right\} \\ - \frac{1}{2} \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{y}^T \boldsymbol{\Omega}^{-(t)} \mathbf{V}_i \boldsymbol{\Omega}^{-(t)} \mathbf{y}$$

minorizing the log-likelihood up to an irrelevant constant. Appendix A.1 gives a detailed derivation for the more general multivariate response case. The rank of the covariance ma-

trix \mathbf{V}_i appears because \mathbf{V}_i may not be invertible. Both of the surrogates $g_{EM}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(\infty)})$ and $g_{MM}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(\infty)})$ are parameter separated. This implies that both second differentials $d^2 g_{EM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ and $d^2 g_{MM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ are diagonal. A small diagonal entry of either matrix indicates fast convergence of the corresponding variance component. Our next result shows that, under Assumption 1, on average the diagonal entries of $d^2 g_{EM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ dominate those of $d^2 g_{MM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ when $m > 2$. Thus, the EM algorithm tends to converge more slowly than the MM algorithm, and the difference is more pronounced as the number of variance components m grows.

Theorem 2. *Let $\boldsymbol{\sigma}^{2(\infty)} \succ \mathbf{0}_m$ be a common limit point of the EM and MM algorithms. Then both second differentials $d^2 g_{MM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ and $d^2 g_{EM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})$ are diagonal with*

$$\begin{aligned} d^2 g_{EM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})_{ii} &= -\frac{\text{rank}(\mathbf{V}_i)}{2\sigma_i^{4(\infty)}} \\ d^2 g_{MM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})_{ii} &= -\frac{\mathbf{y}^T \boldsymbol{\Omega}^{-(\infty)} \mathbf{V}_i \boldsymbol{\Omega}^{-(\infty)} \mathbf{y}}{\sigma_i^{2(\infty)}} = -\frac{\text{tr}(\boldsymbol{\Omega}^{-(\infty)} \mathbf{V}_i)}{\sigma_i^{2(\infty)}}. \end{aligned}$$

Furthermore, the average ratio

$$\frac{1}{m} \sum_{i=1}^m \frac{d^2 g_{MM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})_{ii}}{d^2 g_{EM}(\boldsymbol{\sigma}^{2(\infty)} | \boldsymbol{\sigma}^{2(\infty)})_{ii}} = \frac{2}{mn} \sum_{i=1}^m \text{tr}(\boldsymbol{\Omega}^{-(\infty)} \sigma_i^{2(\infty)} \mathbf{V}_i) = \frac{2}{m} < 1$$

for $m > 2$ when all \mathbf{V}_i have full rank n .

Proof. **MM algorithm:** The minorizing function for the MM algorithm is

$$\begin{aligned} &g_{MM}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) \\ &= -\frac{1}{2} \text{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{\Omega}) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)} \left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \mathbf{V}_i \right) \boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) + c^{(t)} \\ &= \sum_{i=1}^m -\frac{\sigma_i^2}{2} \text{tr}(\boldsymbol{\Omega}^{-(t)} \mathbf{V}_i) - \frac{\sigma_i^{4(t)}}{2\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)} \mathbf{V}_i \boldsymbol{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) + c^{(t)}, \end{aligned}$$

where

$$c^{(t)} = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln \det \boldsymbol{\Omega}^{(t)} + \frac{n}{2}.$$

Taking derivatives, we have

$$\begin{aligned}\frac{\partial}{\partial \sigma_i^2} g_{\text{MM}}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) &= -\frac{1}{2} \text{tr}(\boldsymbol{\Omega}^{-t} \mathbf{V}_i) + \frac{\sigma_i^{4(t)}}{2\sigma_i^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-t} \mathbf{V}_i \boldsymbol{\Omega}^{-t} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}), \\ \frac{\partial^2}{\partial \sigma_i^2 \partial \sigma_j^2} g_{\text{MM}}(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) &= \begin{cases} -\frac{\sigma_i^{4(t)}}{\sigma_i^6} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-t} \mathbf{V}_i \boldsymbol{\Omega}^{-t} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) & i = j \\ 0 & i \neq j. \end{cases}\end{aligned}$$

EM algorithm: Assume $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^m \mathbf{Z}_i$, where $\mathbf{Z}_i \sim N(\mathbf{0}, \sigma_i^2 \mathbf{V}_i)$ are independent. Then the complete data is $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_m)$. From the information inequality, we have

$$L(\mathbf{y} | \boldsymbol{\sigma}^2) \geq Q(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) - Q(\boldsymbol{\sigma}^{2(t)} | \boldsymbol{\sigma}^{2(t)}) + L(\mathbf{y} | \boldsymbol{\sigma}^{2(t)}),$$

where

$$\begin{aligned}L(\mathbf{y} | \boldsymbol{\sigma}^2) &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\ Q(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{rank}(\mathbf{V}_i) \ln \sigma_i^2 + \frac{\sigma_i^{2(t)}}{\sigma_i^2} \text{rank}(\mathbf{V}_i) - \frac{\sigma_i^{4(t)}}{\sigma_i^2} \text{tr}(\boldsymbol{\Omega}^{-t} \mathbf{V}_i) \right\} \\ &\quad - \frac{1}{2} \sum_{i=1}^m \left\{ \frac{\sigma_i^{4(t)}}{\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-t} \mathbf{V}_i \boldsymbol{\Omega}^{-t} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) \right\}.\end{aligned}$$

The minorizing function

$$\begin{aligned}g_{\text{EM}}(\boldsymbol{\sigma} | \boldsymbol{\sigma}^{2(t)}) &= Q(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) - Q(\boldsymbol{\sigma}^{2(t)} | \boldsymbol{\sigma}^{2(t)}) + L(\mathbf{y} | \boldsymbol{\sigma}^{2(t)}) \\ &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{rank}(\mathbf{V}_i) \ln \sigma_i^2 + \frac{\sigma_i^{2(t)}}{\sigma_i^2} \text{rank}(\mathbf{V}_i) - \frac{\sigma_i^{4(t)}}{\sigma_i^2} \text{tr}(\boldsymbol{\Omega}^{-t} \mathbf{V}_i) \right\} \\ &\quad - \frac{1}{2} \sum_{i=1}^m \left\{ \frac{\sigma_i^{4(t)}}{\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-t} \mathbf{V}_i \boldsymbol{\Omega}^{-t} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) \right\} \\ &\quad - \frac{1}{2} \sum_{i=1}^m \left\{ -\text{rank}(\mathbf{V}_i) \ln \sigma_i^{2(t)} - \text{rank}(\mathbf{V}_i) \right\} - \frac{1}{2} \left\{ n \ln(2\pi) + \ln \det \boldsymbol{\Omega}^{(t)} + n \right\}\end{aligned}$$

of the EM algorithms depends on σ^2 only through $Q(\sigma^2|\sigma^{2(t)})$. Taking derivatives, we have

$$\begin{aligned}
& \frac{\partial}{\partial \sigma_i^2} g_{\text{EM}}(\sigma^2|\sigma^{2(t)}) \\
= & -\frac{\text{rank}(\mathbf{V}_i)}{2\sigma_i^2} \\
& + \frac{\text{rank}(\mathbf{V}_i)\sigma_i^{2(t)} - \sigma_i^{4(t)}\text{tr}(\boldsymbol{\Omega}^{-(t)}\mathbf{V}_i) + \sigma_i^{4(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)}\mathbf{V}_i\boldsymbol{\Omega}^{-(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})}{2\sigma_i^4}, \\
& \frac{\partial^2}{\partial \sigma_i^2 \partial \sigma_j^2} g_{\text{EM}}(\sigma^2|\sigma^{2(t)}) \\
= & \begin{cases} \frac{\text{rank}(\mathbf{V}_i)}{2\sigma_i^4} - \frac{\text{rank}(\mathbf{V}_i)\sigma_i^{2(t)} - \sigma_i^{4(t)}\text{tr}(\boldsymbol{\Omega}^{-(t)}\mathbf{V}_i) + \sigma_i^{4(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)}\mathbf{V}_i\boldsymbol{\Omega}^{-(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})}{\sigma_i^6} & i = j \\ 0 & i \neq j. \end{cases}
\end{aligned}$$

EM vs MM: Let $\sigma^{2(\infty)}$ be a common limit point of EM and MM. By Lemma 4, each component of $\sigma^{2(\infty)}$ is either 0 or has vanishing gradient. Therefore

$$\begin{aligned}
\frac{\partial^2}{(\partial \sigma_i^2)^2} g_{\text{EM}}(\sigma^2|\sigma^{2(\infty)}) \Big|_{\sigma^2=\sigma^{2(\infty)}} &= -\frac{\text{rank}(\mathbf{V}_i)}{2\sigma_i^{4(\infty)}}, \\
\frac{\partial^2}{(\partial \sigma_i^2)^2} g_{\text{MM}}(\sigma^2|\sigma^{2(\infty)}) \Big|_{\sigma^2=\sigma^{2(\infty)}} &= -\frac{\text{tr}(\boldsymbol{\Omega}^{-(\infty)}\mathbf{V}_i)}{\sigma_i^{2(\infty)}}
\end{aligned}$$

and, when all the \mathbf{V}_i all non-singular,

$$\begin{aligned}
& \frac{1}{m} \sum_{i=1}^m \frac{[d^2 g_{\text{MM}}(\sigma^{2(\infty)}|\sigma^{2(\infty)})]_{ii}}{[d^2 g_{\text{EM}}(\sigma^{2(\infty)}|\sigma^{2(\infty)})]_{ii}} \\
= & \frac{2}{m} \sum_{i=1}^m \frac{\sigma_i^{2(\infty)} \text{tr}(\boldsymbol{\Omega}^{-(\infty)}\mathbf{V}_i)}{\text{rank}(\mathbf{V}_i)} \\
= & \frac{2}{m} \leq 1.
\end{aligned}$$

□

Both the EM and MM algorithms must evaluate the traces $\text{tr}(\boldsymbol{\Omega}^{-(t)}\mathbf{V}_i)$ and quadratic forms $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-(t)}\mathbf{V}_i\boldsymbol{\Omega}^{-(t)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})$ at each iteration. Since these quantities are also the building blocks of the approximate rate matrices $d^2g(\sigma^{2(t)}|\sigma^{2(t)})$, one can rationally choose either the EM or MM updates based on which has smaller diagonal entries measured by the ℓ_1 , ℓ_2 , or ℓ_∞ norms. At negligible extra cost, this produces a hybrid algorithm that retains the ascent property and enjoys the better of the two convergence rates.

3.6 Extensions

Besides its competitive numerical performance, Algorithm 1 is attractive for its simplicity and ease of generalization. In this section, we outline MM algorithms for multivariate response models possibly with missing data, linear mixed models, MAP estimation, and penalized estimation.

3.6.1 Multivariate Response Model

Consider the multivariate response model with $n \times d$ response matrix \mathbf{Y} , mean $E\mathbf{Y} = \mathbf{X}\mathbf{B}$, and covariance

$$\mathbf{\Omega} = \text{Cov}(\text{vec}\mathbf{Y}) = \sum_{i=1}^m \mathbf{\Gamma}_i \otimes \mathbf{V}_i.$$

The $p \times d$ coefficient matrix \mathbf{B} collects the fixed effects, the $\mathbf{\Gamma}_i$ are unknown $d \times d$ covariance matrices, and the \mathbf{V}_i are known $n \times n$ covariance matrices. If the vector $\text{vec}\mathbf{Y}$ is normally distributed, then \mathbf{Y} equals a sum of independent matrix normal distributions (Gupta and Nagar, 1999). We now make this assumption and pursue estimation of \mathbf{B} and the $\mathbf{\Gamma}_i$, which we collectively denote as $\mathbf{\Gamma}$. Under the normality assumption, the Kronecker product identity $\text{vec}(\mathbf{CDE}) = (\mathbf{E}^T \otimes \mathbf{C})\text{vec}(\mathbf{D})$ yields the log-likelihood

$$\begin{aligned} & L(\mathbf{B}, \mathbf{\Gamma}) \\ &= -\frac{1}{2} \ln \det \mathbf{\Omega} - \frac{1}{2} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T \mathbf{\Omega}^{-1} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \\ &= -\frac{1}{2} \ln \det \mathbf{\Omega} - \frac{1}{2} \{ \text{vec}\mathbf{Y} - (\mathbf{I}_d \otimes \mathbf{X})\text{vec}\mathbf{B} \}^T \mathbf{\Omega}^{-1} \{ \text{vec}\mathbf{Y} - (\mathbf{I}_d \otimes \mathbf{X})\text{vec}\mathbf{B} \}. \end{aligned} \quad (3.12)$$

Updating \mathbf{B} given $\mathbf{\Gamma}^{(t)}$ is accomplished by solving the general least squares problem met earlier in the univariate case. Maximization of the log-likelihood (3.12) is difficult due to the requirement that each $\mathbf{\Gamma}_i$ be positive semidefinite. Typical solutions involve reparameterization of the covariance matrix (Pinheiro and Bates, 1996). The MM algorithm derived in this section gracefully accommodates the covariance constraints.

Updating $\mathbf{\Gamma}$ given $\mathbf{B}^{(t)}$ requires generalizing the minorization (3.3). In view of Lemma 1 and

the identities $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ and $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, we have

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} &= m \left\{ \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Gamma}_i^{(t)} \otimes \mathbf{V}_i \right\} \left\{ \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Gamma}_i \otimes \mathbf{V}_i \right\}^{-1} \left\{ \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Gamma}_i^{(t)} \otimes \mathbf{V}_i \right\} \\ &\preceq m \frac{1}{m} \sum_{i=1}^m (\boldsymbol{\Gamma}_i^{(t)} \otimes \mathbf{V}_i) (\boldsymbol{\Gamma}_i \otimes \mathbf{V}_i)^{-1} (\boldsymbol{\Gamma}_i^{(t)} \otimes \mathbf{V}_i) \\ &= \sum_{i=1}^m (\boldsymbol{\Gamma}_i^{(t)} \boldsymbol{\Gamma}_i^{-1} \boldsymbol{\Gamma}_i^{(t)}) \otimes \mathbf{V}_i, \end{aligned}$$

or equivalently

$$\boldsymbol{\Omega}^{-1} \preceq \boldsymbol{\Omega}^{-(t)} \left\{ \sum_{i=1}^m (\boldsymbol{\Gamma}_i^{(t)} \boldsymbol{\Gamma}_i^{-1} \boldsymbol{\Gamma}_i^{(t)}) \otimes \mathbf{V}_i \right\} \boldsymbol{\Omega}^{-(t)}. \quad (3.13)$$

This derivation relies on the invertibility of the matrices \mathbf{V}_i . One can relax this assumption by substituting $\mathbf{V}_{\epsilon,i} = \mathbf{V}_i + \epsilon \mathbf{I}_n$ for \mathbf{V}_i and sending ϵ to 0.

The majorization (3.13) and the minorization (3.4) jointly yield the surrogate

$$g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}^{(t)}) = -\frac{1}{2} \sum_{i=1}^m \left[\text{tr} \left\{ \boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Gamma}_i \otimes \mathbf{V}_i) \right\} + (\text{vec } \mathbf{R}^{(t)})^T \left\{ (\boldsymbol{\Gamma}_i^{(t)} \boldsymbol{\Gamma}_i^{-1} \boldsymbol{\Gamma}_i^{(t)}) \otimes \mathbf{V}_i \right\} (\text{vec } \mathbf{R}^{(t)}) \right] + c^{(t)},$$

where $\mathbf{R}^{(t)}$ is the $n \times d$ matrix satisfying $\text{vec } \mathbf{R}^{(t)} = \boldsymbol{\Omega}^{-(t)} \text{vec}(\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)})$ and $c^{(t)}$ is an irrelevant constant. Based on the Kronecker identities $(\text{vec } \mathbf{A})^T \text{vec } \mathbf{B} = \text{tr}(\mathbf{A}^T \mathbf{B})$ and $\text{vec}(\mathbf{CDE}) = (\mathbf{E}^T \otimes \mathbf{C}) \text{vec}(\mathbf{D})$, the surrogate can be rewritten as

$$\begin{aligned} g(\boldsymbol{\Gamma} \mid \boldsymbol{\Gamma}^{(t)}) &= -\frac{1}{2} \sum_{i=1}^m \left[\text{tr} \left\{ \boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Gamma}_i \otimes \mathbf{V}_i) \right\} + \text{tr}(\mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \boldsymbol{\Gamma}_i^{(t)} \boldsymbol{\Gamma}_i^{-1} \boldsymbol{\Gamma}_i^{(t)}) \right] + c^{(t)} \\ &= -\frac{1}{2} \sum_{i=1}^m \left[\text{tr} \left\{ \boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Gamma}_i \otimes \mathbf{V}_i) \right\} + \text{tr}(\boldsymbol{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \boldsymbol{\Gamma}_i^{(t)} \boldsymbol{\Gamma}_i^{-1}) \right] + c^{(t)}. \end{aligned}$$

The first trace is linear in $\boldsymbol{\Gamma}_i$ with the coefficient of entry $(\boldsymbol{\Gamma}_i)_{jk}$ equal to

$$\text{tr}(\boldsymbol{\Omega}_{jk}^{-(t)} \mathbf{V}_i) = \mathbf{1}_n^T (\mathbf{V}_i \odot \boldsymbol{\Omega}_{jk}^{-(t)}) \mathbf{1}_n,$$

where $\boldsymbol{\Omega}_{jk}^{-(t)}$ is the (j, k) -th $n \times n$ block of $\boldsymbol{\Omega}^{-(t)}$. The matrix \mathbf{M}_i of these coefficients can be

<p>input : Y, X, V_1, \dots, V_m output: MLE $\hat{B}, \hat{\Gamma}_1, \dots, \hat{\Gamma}_m$</p> <p>1 Initialize $\Gamma_i^{(0)}$ positive definite, $i = 1, \dots, m$; 2 repeat 3 $\Omega^{(t)} \leftarrow \sum_{i=1}^m \Gamma_i^{(t)} \otimes V_i$; 4 $B^{(t)} \leftarrow \arg \min_B \{ \text{vec} Y - (I_d \otimes X) \text{vec} B \}^T \Omega^{-(t)} \{ \text{vec} Y - (I_d \otimes X) \text{vec} B \}$; 5 $R^{(t)} \leftarrow \text{reshape}(\Omega^{-(t)} \text{vec}(Y - XB^{(t)}), n, d)$; 6 for $i = 1, \dots, m$ do 7 Cholesky $L_i^{(t)} L_i^{(t)T} \leftarrow (I_d \otimes \mathbf{1}_n)^T \{ (\mathbf{1}_d \mathbf{1}_d^T \otimes V_i) \odot \Omega^{-(t)} \} (I_d \otimes \mathbf{1}_n)$; 8 $\Gamma_i^{(t+1)} \leftarrow L_i^{-(t)T} \{ L_i^{(t)T} (\Gamma_i^{(t)} R^{(t)T} V_i R^{(t)} \Gamma_i^{(t)}) L_i^{(t)} \}^{1/2} L_i^{-t}$ 9 end 10 until <i>objective value converges</i>;</p>
--

Algorithm 3: The MM algorithm for MLE of the multivariate response model (3.12).

written as

$$\begin{aligned}
M_i &= \begin{pmatrix} \mathbf{1}^T & \mathbf{0}^T & \dots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{1}^T & \dots & \mathbf{0}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \dots & \mathbf{1}^T \end{pmatrix} \left\{ \begin{pmatrix} V_i & \dots & V_i \\ \vdots & \ddots & \vdots \\ V_i & \dots & V_i \end{pmatrix} \odot \Omega^{-(t)} \right\} \begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{pmatrix} \\
&= (I_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes V_i) \odot \Omega^{-(t)}] (I_d \otimes \mathbf{1}_n).
\end{aligned}$$

The directional derivative of $g(\Gamma | \Gamma^{(t)})$ with respect to Γ_i in the direction Δ_i is

$$\begin{aligned}
&-\frac{1}{2} \text{tr}(M_i \Delta_i) + \frac{1}{2} \text{tr}(\Gamma_i^{(t)} R^{(t)T} V_i R^{(t)} \Gamma_i^{(t)} \Gamma_i^{-1} \Delta_i \Gamma_i^{-1}) \\
&= -\frac{1}{2} \text{tr}(M_i \Delta_i) + \frac{1}{2} \text{tr}(\Gamma_i^{-1} \Gamma_i^{(t)} R^{(t)T} V_i R^{(t)} \Gamma_i^{(t)} \Gamma_i^{-1} \Delta_i).
\end{aligned}$$

Because all directional derivatives of $g(\Gamma | \Gamma^{(t)})$ vanish at a stationarity point, the matrix equation

$$M_i = \Gamma_i^{-1} \Gamma_i^{(t)} R^{(t)T} V_i R^{(t)} \Gamma_i^{(t)} \Gamma_i^{-1} \quad (3.14)$$

holds. Fortunately, this equation admits an explicit solution. For positive scalars a and b , the solution to the equation $b = x^{-1} a x^{-1}$ is $x = \pm \sqrt{a/b}$. The matrix analogue of this equation is the Riccati equation $B = X^{-1} A X^{-1}$, whose solution is summarized in the next lemma.

Lemma 6. *Assume A and B are positive definite and L is the Cholesky factor of B . Then*

$\mathbf{Y} = \mathbf{L}^{-T}(\mathbf{L}^T \mathbf{A} \mathbf{L})^{1/2} \mathbf{L}^{-1}$ is the unique positive definite solution to the matrix equation $\mathbf{B} = \mathbf{X}^{-1} \mathbf{A} \mathbf{X}^{-1}$.

Proof. Direct substitution shows that \mathbf{Y} solves the equivalent equation $\mathbf{X} \mathbf{B} \mathbf{X} = \mathbf{A}$. To show uniqueness, suppose $\mathbf{Y}^{-1} \mathbf{A} \mathbf{Y}^{-1} = \mathbf{B}$ and $\mathbf{Z}^{-1} \mathbf{A} \mathbf{Z}^{-1} = \mathbf{B}$. The equations

$$\begin{aligned} (\mathbf{B}^{1/2} \mathbf{Y} \mathbf{B}^{1/2})^2 &= \mathbf{B}^{1/2} \mathbf{Y} \mathbf{B} \mathbf{Y} \mathbf{B}^{1/2} = \mathbf{B}^{1/2} \mathbf{A} \mathbf{B}^{1/2} \\ (\mathbf{B}^{1/2} \mathbf{Z} \mathbf{B}^{1/2})^2 &= \mathbf{B}^{1/2} \mathbf{Z} \mathbf{B} \mathbf{Z} \mathbf{B}^{1/2} = \mathbf{B}^{1/2} \mathbf{A} \mathbf{B}^{1/2} \end{aligned}$$

imply $\mathbf{B}^{1/2} \mathbf{Y} \mathbf{B}^{1/2} = \mathbf{B}^{1/2} \mathbf{Z} \mathbf{B}^{1/2}$ by virtue of the uniqueness of symmetric square root. Since $\mathbf{B}^{-1/2}$ is positive definite, $\mathbf{Y} = \mathbf{Z}$. \square

The Cholesky factor \mathbf{L} in Lemma 6 can be replaced by the symmetric square root of \mathbf{B} . The solution, which is unique, remains the same. The Cholesky decomposition is preferred for its cheaper computational cost and better numerical stability.

Algorithm 3 summarizes the MM algorithm for fitting the multi-response model (3). Each iteration invokes m Cholesky decompositions and symmetric square roots of $d \times d$ positive definite matrices. Fortunately in most applications, d is a small number. The following result guarantees the non-singularity of the Cholesky factor throughout iterations.

Proposition 2. *Assume \mathbf{V}_i has strictly positive diagonal entries. Then the symmetric matrix $\mathbf{M}_i = (\mathbf{I}_d \otimes \mathbf{1}_n)^T \left\{ (\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \boldsymbol{\Omega}^{-t} \right\} (\mathbf{I}_d \otimes \mathbf{1}_n)$ is positive definite for all t . Furthermore if $\boldsymbol{\Gamma}_i^{(0)} \succ \mathbf{0}$ and no column of $\mathbf{R}^{(t)}$ lies in the null space of \mathbf{V}_i for all t , then $\boldsymbol{\Gamma}_i^{(t)} \succ \mathbf{0}$ for all t .*

Proof. If \mathbf{V}_i has strictly positive diagonal entries, then so does $\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i$, and the Hadamard product $(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \boldsymbol{\Omega}^{-t}$ is positive definite by Schur's lemma. Since the matrix $\mathbf{I}_d \otimes \mathbf{1}_n$ has full column rank d , the matrix \mathbf{M}_i is also positive definite. Finally, if no column of $\mathbf{R}^{(t)}$ lies in the null space of \mathbf{V}_i , and $\boldsymbol{\Gamma}^{(t)}$ is positive definite, then $\boldsymbol{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \boldsymbol{\Gamma}_i^{(t)}$ is positive definite. The second claim follows by induction and Lemma 6. \square

Multivariate Response, Two Variance Components

When there are $m = 2$ variance components $\boldsymbol{\Omega} = \boldsymbol{\Gamma}_1 \otimes \mathbf{V}_1 + \boldsymbol{\Gamma}_2 \otimes \mathbf{V}_2$, repeated inversion of the $nd \times nd$ covariance matrix $\boldsymbol{\Omega}$ reduces to a single $nd \times nd$ simultaneous congruence decomposition and, per iteration, two $d \times d$ Cholesky decompositions and one $d \times d$ simultaneous congruence decomposition. The simultaneous congruence decomposition of the matrix pair $(\mathbf{V}_1, \mathbf{V}_2)$ involves generalized eigenvalues $\mathbf{d} = (d_1, \dots, d_n)$ and a nonsingular matrix \mathbf{U} such that $\mathbf{U}^T \mathbf{V}_1 \mathbf{U} = \mathbf{D} = \text{diag}(\mathbf{d})$ and $\mathbf{U}^T \mathbf{V}_2 \mathbf{U} = \mathbf{I}$. If the simultaneous congruence decomposition of $(\boldsymbol{\Gamma}_1^{(t)}, \boldsymbol{\Gamma}_2^{(t)})$ is

$(\mathbf{\Lambda}^{(t)}, \mathbf{\Phi}^{(t)})$ with $\mathbf{\Phi}^{(t)T} \mathbf{\Gamma}_1^{(t)} \mathbf{\Phi}^{(t)} = \mathbf{\Lambda}^{(t)} = \text{diag}(\boldsymbol{\lambda}^{(t)})$ and $\mathbf{\Phi}^{(t)T} \mathbf{\Gamma}_2^{(t)} \mathbf{\Phi}^{(t)} = \mathbf{I}_d$, then

$$\begin{aligned}
\boldsymbol{\Omega}^{(t)} &= (\mathbf{\Phi}^{-(t)} \otimes \mathbf{U}^{-1})^T (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n) (\mathbf{\Phi}^{-(t)} \otimes \mathbf{U}^{-1}) \\
\boldsymbol{\Omega}^{- (t)} &= (\mathbf{\Phi}^{(t)} \otimes \mathbf{U}) (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} (\mathbf{\Phi}^{(t)} \otimes \mathbf{U})^T \\
\det \boldsymbol{\Omega}^{(t)} &= \det(\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n) \det \left\{ (\mathbf{\Phi}^{-(t)} \otimes \mathbf{U}^{-1})^T (\mathbf{\Phi}^{-(t)} \otimes \mathbf{U}^{-1}) \right\} \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n) \det(\mathbf{\Gamma}_2^{(t)} \otimes \mathbf{V}_2) \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n) \det(\mathbf{\Gamma}_2^{(t)})^n \det(\mathbf{V}_2)^d.
\end{aligned}$$

To update the fixed effects \mathbf{B} given $\mathbf{\Gamma}_1^{(t)}$ and $\mathbf{\Gamma}_2^{(t)}$, the general least squares criterion is

$$\begin{aligned}
&\frac{1}{2} \{ \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \}^T \boldsymbol{\Omega}^{- (t)} \{ \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \} \\
&= \frac{1}{2} \{ \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \}^T (\mathbf{\Phi}^{(t)} \otimes \mathbf{U}) (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} (\mathbf{\Phi}^{(t)} \otimes \mathbf{U})^T \{ \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \} \\
&= \frac{1}{2} \text{vec} \left\{ \mathbf{U}^T (\mathbf{Y} - \mathbf{X}\mathbf{B}) \mathbf{\Phi}^{(t)} \right\}^T (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} \text{vec} \left\{ \mathbf{U}^T (\mathbf{Y} - \mathbf{X}\mathbf{B}) \mathbf{\Phi}^{(t)} \right\} \\
&= \frac{1}{2} \left\{ \text{vec}(\mathbf{U}^T \mathbf{Y} \mathbf{\Phi}^{(t)}) - (\mathbf{\Phi}^{(t)T} \otimes \mathbf{U}^T \mathbf{X}) \text{vec} \mathbf{B} \right\}^T (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} \\
&\quad \cdot \left\{ \text{vec}(\mathbf{U}^T \mathbf{Y} \mathbf{\Phi}^{(t)}) - (\mathbf{\Phi}^{(t)T} \otimes \mathbf{U}^T \mathbf{X}) \text{vec} \mathbf{B} \right\}.
\end{aligned}$$

Minimization of this criterion reduces to a weighted least squares problem for the transformed responses $\tilde{\mathbf{Y}} = \mathbf{U}^T \mathbf{Y}$, transformed predictor matrix $\tilde{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, and observation weights $(\lambda_k^{(t)} d_i + 1)^{-1}$. To update $\mathbf{\Gamma}_1^{(t)}$ and $\mathbf{\Gamma}_2^{(t)}$, we need to evaluate the matrices \mathbf{M}_i and $\mathbf{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \mathbf{\Gamma}_i^{(t)}$ that appear in the stationarity condition (3.14).

Evaluation of \mathbf{M}_i : Note the (j, k) -th entry of \mathbf{M}_i is $\text{tr}(\boldsymbol{\Omega}_{jk}^{- (t)} \mathbf{V}_i)$, where $\boldsymbol{\Omega}_{jk}^{- (t)}$ is the (j, k) -th block of

$$\boldsymbol{\Omega}^{- (t)} = (\mathbf{\Phi}^{(t)} \otimes \mathbf{U}) (\boldsymbol{\Lambda}^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} (\mathbf{\Phi}^{(t)} \otimes \mathbf{U})^T,$$

which can be expressed as

$$\boldsymbol{\Omega}_{jk}^{- (t)} = \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \mathbf{U} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \mathbf{U}^T.$$

Therefore \mathbf{M}_1 has entries

$$\begin{aligned}
(\mathbf{M}_1)_{jk} &= \text{tr}(\mathbf{V}_1 \boldsymbol{\Omega}_{ij}^{-\langle t \rangle}) \\
&= \text{tr} \left\{ \mathbf{U}^{-T} \mathbf{D} \mathbf{U}^{-1} \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \mathbf{U} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \mathbf{U}^T \right\} \\
&= \text{tr} \left\{ \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \mathbf{D} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \right\} \\
&= \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \text{tr} \left\{ \mathbf{D} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \right\},
\end{aligned}$$

and \mathbf{M}_2 has entries

$$\begin{aligned}
(\mathbf{M}_2)_{jk} &= \text{tr}(\mathbf{V}_2 \boldsymbol{\Omega}_{ij}^{-\langle t \rangle}) \\
&= \text{tr} \left\{ \mathbf{U}^{-T} \mathbf{U}^{-1} \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \mathbf{U} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \mathbf{U}^T \right\} \\
&= \text{tr} \left\{ \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \right\} \\
&= \sum_{l=1}^d \phi_{jl}^{(t)} \phi_{lk}^{(t)} \text{tr}(\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1}.
\end{aligned}$$

Collectively we have

$$\begin{aligned}
\mathbf{M}_1 &= \boldsymbol{\Phi}^{(t)} \text{diag} \left[\text{tr} \left\{ \mathbf{D} (\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \right\} \right] \boldsymbol{\Phi}^{(t)T} \\
\mathbf{M}_2 &= \boldsymbol{\Phi}^{(t)} \text{diag} \left\{ \text{tr}(\lambda_l^{(t)} \mathbf{D} + \mathbf{I}_n)^{-1} \right\} \boldsymbol{\Phi}^{(t)T}.
\end{aligned}$$

Evaluation of $\boldsymbol{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \boldsymbol{\Gamma}_i^{(t)}$: Write

$$\begin{aligned}
\boldsymbol{\Gamma}_1^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_1 \mathbf{R}^{(t)} \boldsymbol{\Gamma}_1^{(t)} &= \mathbf{N}_1^T \mathbf{N}_1 \\
\boldsymbol{\Gamma}_2^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_2 \mathbf{R}^{(t)} \boldsymbol{\Gamma}_2^{(t)} &= \mathbf{N}_2^T \mathbf{N}_2,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{N}_1 &= \mathbf{D}^{1/2} \mathbf{U}^{-1} \mathbf{R}^{(t)} \boldsymbol{\Phi}^{-\langle t \rangle T} \boldsymbol{\Lambda}^{(t)} \boldsymbol{\Phi}^{-\langle t \rangle} \\
\mathbf{N}_2 &= \mathbf{U}^{-1} \mathbf{R}^{(t)} \boldsymbol{\Phi}^{-\langle t \rangle T} \boldsymbol{\Phi}^{-\langle t \rangle}.
\end{aligned}$$

To further simplify, note

$$\begin{aligned}
& \text{vec } \mathbf{N}_1 \\
&= (\Phi^{-(t)T} \Lambda^{(t)} \Phi^{-(t)} \otimes \mathbf{D}^{1/2} \mathbf{U}^{-1}) \text{vec} \mathbf{R}^{(t)} \\
&= (\Phi^{-(t)T} \Lambda^{(t)} \Phi^{-(t)} \otimes \mathbf{D}^{1/2} \mathbf{U}^{-1}) \Omega^{-(t)} \text{vec}(\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)}) \\
&= (\Phi^{-(t)T} \Lambda^{(t)} \Phi^{-(t)} \otimes \mathbf{D}^{1/2} \mathbf{U}^{-1}) (\Phi^{(t)} \otimes \mathbf{U}) (\Lambda^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} (\Phi^{(t)} \otimes \mathbf{U})^T \\
&\quad \text{vec}(\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)}) \\
&= (\Phi^{-(t)T} \Lambda^{(t)} \otimes \mathbf{D}^{1/2}) (\Lambda^{(t)} \otimes \mathbf{D} + \mathbf{I}_d \otimes \mathbf{I}_n)^{-1} \text{vec}(\mathbf{U}^T (\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)}) \Phi^{(t)}) \\
&= (\Phi^{-(t)T} \Lambda^{(t)} \otimes \mathbf{D}^{1/2}) \text{vec} \left\{ \mathbf{U}^T (\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)}) \Phi^{(t)} \oslash (\mathbf{d} \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right\} \\
&= \text{vec} \left[\mathbf{D}^{1/2} \left\{ (\mathbf{U}^T (\mathbf{Y} - \mathbf{X} \mathbf{B}^{(t)}) \Phi^{(t)}) \oslash (\mathbf{d} \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right\} \Lambda^{(t)} \Phi^{-(t)} \right],
\end{aligned}$$

where \oslash denotes a Hadamard quotient. Thus,

$$\mathbf{N}_1 = \mathbf{D}^{1/2} \left[\left\{ (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \mathbf{B}^{(t)}) \Phi^{(t)} \right\} \oslash (\mathbf{d} \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right] \Lambda^{(t)} \Phi^{-(t)},$$

and similarly

$$\mathbf{N}_2 = \left[\left\{ (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \mathbf{B}^{(t)}) \Phi^{(t)} \right\} \oslash (\mathbf{d} \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right] \Phi^{-(t)}.$$

Algorithm 4 summarizes the simplified MM algorithm.

3.6.2 Multivariate Response Model with Missing Responses

In many applications the multivariate response model (3.12) involves missing responses. For instance, in testing multiple longitudinal traits in genetics, some trait values y_{ij} may be missing due to dropped patient visits, while their genetic covariates are complete. Missing data destroys the symmetry of the log-likelihood (3.12) and complicates finding the MLE. Fortunately, MM algorithm 3 easily adapts to this challenge.

The familiar EM argument (McLachlan and Krishnan, 2008, Section 2.2) shows that

$$-\frac{n}{2} \ln \det \Omega^{(t)} - \frac{1}{2} \text{tr} \left[\Omega^{-(t)} \left\{ \text{vec}(\mathbf{Z}^{(t)} - \mathbf{X} \mathbf{B}^{(t)}) \text{vec}(\mathbf{Z}^{(t)} - \mathbf{X} \mathbf{B}^{(t)})^T + \mathbf{C}^{(t)} \right\} \right] \quad (3.15)$$

minorizes the observed log-likelihood at the current iterate $(\mathbf{B}^{(t)}, \Gamma_1^{(t)}, \dots, \Gamma_m^{(t)})$. Here $\mathbf{Z}^{(t)}$ is the completed response matrix given the observed responses $\mathbf{Y}_{\text{obs}}^{(t)}$ and the current parameter values. The complete data \mathbf{Y} is assumed to be normally distributed $N(\text{vec}(\mathbf{X} \mathbf{B}^{(t)}), \Omega^{(t)})$. The block matrix $\mathbf{C}^{(t)}$ is 0 except for a lower-right block consisting of a Schur complement.

To maximize the surrogate (3.15), we invoke the familiar minorization (3.4) and majorization

<p>Input : Y, X, V_1, V_2 Output: MLE $\hat{B}, \hat{\Gamma}_1, \hat{\Gamma}_2$</p> <ol style="list-style-type: none"> 1 Simultaneous congruence decomposition: $(D, U) \leftarrow (V_1, V_2)$; 2 Transform data: $\tilde{Y} \leftarrow U^T Y, \tilde{X} \leftarrow U^T X$; 3 Initialize $\Gamma_1^{(0)}, \Gamma_2^{(0)}$ positive definite ; 4 repeat 5 Simultaneous congruence decomposition $(\Lambda^{(t)}, \Phi^{(t)}) \leftarrow (\Gamma_1^{(t)}, \Gamma_2^{(t)})$; 6 $B^{(t)} \leftarrow \arg \min_B \left\{ \text{vec}(\tilde{Y} \Phi^{(t)}) - (\Phi^{(t)T} \otimes \tilde{X}) \text{vec} B \right\}^T (\Lambda^{(t)} \otimes D + I_d \otimes I_n)^{-1} \left\{ \text{vec}(\tilde{Y} \Phi^{(t)}) - (\Phi^{(t)T} \otimes \tilde{X}) \text{vec} B \right\}$; 7 Cholesky $L_1^{(t)} L_1^{(t)T} \leftarrow \Phi^{(t)} \text{diag} \left(\text{tr} \left(D(\lambda_k^{(t)} D + I_n)^{-1} \right), k = 1, \dots, d \right) \Phi^{(t)T}$; 8 Cholesky $L_2^{(t)} L_2^{(t)T} \leftarrow \Phi^{(t)} \text{diag} \left(\text{tr} \left((\lambda_k^{(t)} D + I_n)^{-1} \right), k = 1, \dots, d \right) \Phi^{(t)T}$; 9 $N_1^{(t)} \leftarrow D^{1/2} \left\{ (\tilde{Y} - \tilde{X} B^{(t)}) \Phi^{(t)} \oslash (d \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right\} \Lambda^{(t)} \Phi^{-t}$; 10 $N_2^{(t)} \leftarrow \left\{ (\tilde{Y} - \tilde{X} B^{(t)}) \Phi^{(t)} \oslash (d \lambda^{(t)T} + \mathbf{1}_n \mathbf{1}_d^T) \right\} \Phi^{-t}$; 11 $\Gamma_i^{(t+1)} \leftarrow L_i^{-t)T} (L_i^{(t)T} M_i^{(t)T} M_i^{(t)} L_i^{(t)})^{1/2} L_i^{-t}$, $i = 1, 2$; 12 until <i>objective value converges</i>;
--

Algorithm 4: MM algorithm for multivariate response model $\Omega = \Gamma_1 \otimes V_1 + \Gamma_2 \otimes V_2$ with two variance components matrices. Note that \oslash denotes a Hadamard quotient.

(3.13) to separate the variance components Γ_i . At each iteration we impute missing entries by their conditional means, compute their conditional variances and covariances to supply the Schur complement, and then update the fixed effects and variance components by the explicit updates of Algorithm 3. The required conditional means and conditional variances can be conveniently obtained in the process of inverting $\Omega^{(t)}$ by the sweep operator of computational statistics (Lange, 2010, Section 7.3).

3.6.3 Linear Mixed Model (LMM)

The linear mixed model plays a central role in longitudinal data analysis. For the sake of simplicity, consider the single-level LMM (Laird and Ware, 1982; Bates and Pinheiro, 1998) for n independent data clusters (y_i, X_i, Z_i) with

$$Y_i = X_i \beta + Z_i \gamma_i + \epsilon_i, \quad i = 1, \dots, n,$$

where β is a vector of fixed effects, the $\gamma_i \sim N(\mathbf{0}, R_i(\theta))$ are independent random effects, and $\epsilon_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$ captures random noise independent of γ_i . We assume the matrices Z_i have full column rank. The within-cluster covariance matrices $R_i(\theta)$ depend on a parameter

vector $\boldsymbol{\theta}$; typical choices for $\mathbf{R}_i(\boldsymbol{\theta})$ impose autocorrelation, compound symmetry, or unstructured correlation. It is clear that \mathbf{Y}_i is normal with mean $\mathbf{X}_i\boldsymbol{\beta}$, covariance $\boldsymbol{\Omega}_i = \mathbf{Z}_i\mathbf{R}_i(\boldsymbol{\theta})\mathbf{Z}_i^T + \sigma^2\mathbf{I}_{n_i}$, and log-likelihood

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = -\frac{1}{2} \ln \det \boldsymbol{\Omega}_i - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

The next three facts about pseudo-inverses are used in deriving the MM algorithm for LMM.

Lemma 7. *If \mathbf{A} has full column rank and \mathbf{B} has full row rank, then $(\mathbf{AB})^+ = \mathbf{B}^+\mathbf{A}^+$.*

Proof. Under the hypotheses, the representations $\mathbf{A}^+ = (\mathbf{A}^T\mathbf{A})^+\mathbf{A}^T = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ and $\mathbf{B}^+ = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}$ are well known. The choice $\mathbf{B}^+\mathbf{A}^+ = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ satisfies the four equations characterizing the pseudo-inverse of \mathbf{AB} . \square

Lemma 8. *If \mathbf{A} and \mathbf{B} are positive semidefinite matrices with the same range, then*

$$\lim_{\epsilon \downarrow 0} (\mathbf{B} + \epsilon\mathbf{I})(\mathbf{A} + \epsilon\mathbf{I})^{-1}(\mathbf{B} + \epsilon\mathbf{I}) = \mathbf{B}\mathbf{A}^+\mathbf{B}.$$

Proof. Suppose \mathbf{A} has spectral decomposition $\sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T$. The matrix $\mathbf{P} = \sum_{\lambda_i > 0} \mathbf{u}_i \mathbf{u}_i^T$ projects onto the range of \mathbf{A} and therefore also projects onto the range of \mathbf{B} . It follows that $\mathbf{PB} = \mathbf{B}$ and by symmetry that $\mathbf{BP} = \mathbf{B}$. This allows us to write

$$\begin{aligned} & (\mathbf{B} + \epsilon\mathbf{I})(\mathbf{A} + \epsilon\mathbf{I})^{-1}(\mathbf{B} + \epsilon\mathbf{I}) \\ &= \mathbf{BP}(\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{PB} + \epsilon\mathbf{BP}(\mathbf{A} + \epsilon\mathbf{I})^{-1} + \epsilon(\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{PB} + \epsilon^2(\mathbf{A} + \epsilon\mathbf{I})^{-1}. \end{aligned}$$

The last three of these terms vanish as $\epsilon \downarrow 0$; the first term tends to the claimed limit. These assertions follow from the expressions

$$\mathbf{P}(\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{P} = \mathbf{P}(\mathbf{A} + \epsilon\mathbf{I})^{-1} = (\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{P} = \sum_{\lambda_i > 0} \frac{1}{\lambda_i + \epsilon} \mathbf{u}_i \mathbf{u}_i^T$$

$$\text{and } \epsilon^2(\mathbf{A} + \epsilon\mathbf{I})^{-1} = \sum_i \frac{\epsilon^2}{\lambda_i + \epsilon} \mathbf{u}_i \mathbf{u}_i^T. \quad \square$$

Lemma 9. *If \mathbf{R} and \mathbf{S} are positive definite matrices, and the conformable matrix \mathbf{Z} has full column rank, then the matrices $\mathbf{Z}\mathbf{R}\mathbf{Z}^T$ and $\mathbf{Z}\mathbf{S}\mathbf{Z}^T$ share a common range.*

Proof. In fact, both matrices have range equal to the range of \mathbf{Z} . The matrices \mathbf{Z} and $\mathbf{Z}\mathbf{R}^{1/2}$ clearly have the same range. Furthermore, the matrices $\mathbf{Z}\mathbf{R}^{1/2}$ and $\mathbf{Z}\mathbf{R}^{1/2}\mathbf{R}^{1/2}\mathbf{Z}^T$ also have the same range. \square

The convexity of the map $(\mathbf{X}, \mathbf{Y}) \mapsto \mathbf{X}^T \mathbf{Y}^{-1} \mathbf{X}$ and Lemmas 7, 8, and 9 now yield via the obvious limiting argument the majorization

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} &= (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T + \sigma^{2(t)} \mathbf{I}_{n_i}) (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}) \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_{n_i})^{-1} (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T + \sigma^{2(t)} \mathbf{I}_{n_i}) \\ &\succeq (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T) (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}) \mathbf{Z}_i^T)^+ (\mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T) + \frac{\sigma^{4(t)}}{\sigma^2} \mathbf{I}_{n_i} \\ &= \left\{ \mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T \mathbf{Z}_i^{T+} \right\} \mathbf{R}_i^{-1}(\boldsymbol{\theta}) \left\{ \mathbf{Z}_i^+ \mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T \right\} + \frac{\sigma^{4(t)}}{\sigma^2} \mathbf{I}_{n_i} \end{aligned}$$

In combination with the minorization (3.4), this gives the surrogate

$$\begin{aligned} g_i(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{\theta}^{(t)}, \sigma^{2(t)}) &= -\frac{1}{2} \text{tr} \left(\mathbf{Z}_i^T \boldsymbol{\Omega}_i^{-2(t)} \mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}) \right) - \frac{1}{2} \mathbf{r}_i^{(t)T} \mathbf{R}_i^{-1}(\boldsymbol{\theta}) \mathbf{r}_i^{(t)} \\ &\quad - \frac{\sigma^2}{2} \text{tr}(\boldsymbol{\Omega}_i^{-2(t)}) - \frac{\sigma^{4(t)}}{2\sigma^2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}_i^{-2(t)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)}) + c^{(t)}, \end{aligned}$$

for the log-likelihood $L_i(\boldsymbol{\theta}, \sigma^2)$, where

$$\mathbf{r}_i^{(t)} = \left(\mathbf{Z}_i^+ \mathbf{Z}_i \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T \right) \boldsymbol{\Omega}_i^{-2(t)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)}) = \mathbf{R}_i(\boldsymbol{\theta}^{(t)}) \mathbf{Z}_i^T \boldsymbol{\Omega}_i^{-2(t)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)}).$$

The parameters $\boldsymbol{\theta}$ and σ^2 are nicely separated. To maximize the overall minorization function $\sum_i g_i(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{\theta}^{(t)}, \sigma^{2(t)})$, we update σ^2 via

$$\sigma^{2(t+1)} = \sigma^{2(t)} \sqrt{\frac{\sum_i (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}_i^{-2(t)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(t)})}{\sum_i \text{tr}(\boldsymbol{\Omega}_i^{-2(t)})}}.$$

For structured models such as autocorrelation and compound symmetry, updating $\boldsymbol{\theta}$ is a low-dimensional optimization problem that can be approached through the stationarity condition

$$\sum_i \text{vec} \left(\mathbf{Z}_i^T \boldsymbol{\Omega}_i^{(t)} \mathbf{Z}_i - \mathbf{R}_i^{-1}(\boldsymbol{\theta}) \mathbf{r}_i^{(t)} \mathbf{r}_i^{(t)T} \mathbf{R}_i^{-1}(\boldsymbol{\theta}) \right)^T \frac{\partial}{\partial \theta_j} \text{vec} \mathbf{R}_i(\boldsymbol{\theta}) = 0$$

for each component θ_j . For the unstructured model with $\mathbf{R}_i(\boldsymbol{\theta}) = \mathbf{R}$ for all i , the stationarity condition reads

$$\sum_i \mathbf{Z}_i^T \boldsymbol{\Omega}_i^{(t)} \mathbf{Z}_i = \mathbf{R}^{-1} \left(\sum_i \mathbf{r}_i^{(t)} \mathbf{r}_i^{(t)T} \right) \mathbf{R}^{-1}$$

and admits an explicit solution based on Lemma 6.

Similar tactics apply to a multilevel LMM (Bates and Pinheiro, 1998) with responses

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_{i1} \boldsymbol{\gamma}_{i1} + \cdots + \mathbf{Z}_{im} \boldsymbol{\gamma}_{im} + \boldsymbol{\epsilon}_i.$$

Minorization separates parameters for each level (variance component). Depending on the complexity of the covariance matrices, maximization of the surrogate can be accomplished analytically. For the sake of brevity, details are omitted.

3.6.4 MAP Estimation

Suppose $\boldsymbol{\beta}$ follows an improper flat prior, the variance components σ_i^2 follow inverse gamma priors with shapes $\alpha_i > 0$ and scales $\gamma_i > 0$, and these priors are independent. The log-posterior density then reduces to

$$\begin{aligned} & \ln f(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \mathbf{y}, \mathbf{X}) \\ = & -\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \sum_{i=1}^m (\alpha_i + 1) \ln \sigma_i^2 - \sum_{i=1}^m \frac{\gamma_i}{\sigma_i^2} + c, \end{aligned} \quad (3.16)$$

where c is an irrelevant constant. The MAP estimator of $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ is the mode of the posterior distribution. The update (3.2) of $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ remains the same. To update $\boldsymbol{\sigma}^2$ given $\boldsymbol{\beta}$, apply the same minorizations (3.3) and (3.4) to the first two terms of equation (3.16). This separates parameters and yields a convex surrogate for each σ_i^2 . The minimum of the σ_i^2 surrogate is defined by the stationarity condition

$$0 = -\frac{1}{2} \text{tr}(\boldsymbol{\Omega}^{-t} \mathbf{V}_i) + \frac{\sigma_i^{4(t)}}{2\sigma_i^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \boldsymbol{\Omega}^{-t} \mathbf{V}_i \boldsymbol{\Omega}^{-t} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}) - \frac{\alpha_i + 1}{\sigma_i^2} + \frac{\gamma_i}{\sigma_i^4}.$$

Multiplying this by σ_i^4 gives a quadratic equation in σ_i^2 . The positive root should be taken to meet the nonnegativity constraint on σ_i^2 .

For the multivariate response model (3.12), we assume the variance components $\boldsymbol{\Gamma}_i$ follow independent inverse Wishart distributions with degrees of freedom $\nu_i > d - 1$ and scale matrix $\boldsymbol{\Psi}_i \succ \mathbf{0}$. The log density of the posterior distribution is

$$\begin{aligned} L(\mathbf{B}, \boldsymbol{\Gamma} | \mathbf{X}, \mathbf{Y}) &= -\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B})^T \boldsymbol{\Omega}^{-1} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \\ &\quad - \frac{1}{2} \sum_{i=1}^m (\nu_i + d + 1) \ln \det \boldsymbol{\Gamma}_i - \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Psi}_i \boldsymbol{\Gamma}_i^{-1}) + c, \end{aligned} \quad (3.17)$$

where c is an irrelevant constant. Invoking the minorizations (3.4) and (3.13) for the first two terms and the supporting hyperplane minorization

$$-\ln \det \boldsymbol{\Gamma}_i \geq -\ln \det \boldsymbol{\Gamma}_i^{(t)} - \text{tr} \left\{ \boldsymbol{\Gamma}_i^{-t} (\boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_i^{(t)}) \right\}$$

for $-\ln \det \mathbf{\Gamma}_i$ gives the surrogate function

$$\begin{aligned} g(\mathbf{\Gamma}|\mathbf{\Gamma}^{(t)}) &= -\frac{1}{2} \sum_{i=1}^m \text{tr} \left\{ \mathbf{\Omega}^{-(t)} (\mathbf{\Gamma}_i \otimes \mathbf{V}_i) \right\} - \frac{1}{2} \sum_{i=1}^m \text{tr} \left(\mathbf{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \mathbf{\Gamma}_i^{(t)} \mathbf{\Gamma}_i^{-1} \right) \\ &\quad - \frac{1}{2} \sum_{i=1}^m (\nu_i + d + 1) \text{tr} (\mathbf{\Gamma}_i^{-t} \mathbf{\Gamma}_i) - \frac{1}{2} \sum_{i=1}^m \text{tr} (\mathbf{\Psi}_i \mathbf{\Gamma}_i^{-1}) + c^{(t)}. \end{aligned}$$

The optimal $\mathbf{\Gamma}_i$ satisfies the stationarity condition

$$\begin{aligned} &(\mathbf{I}_d \otimes \mathbf{1}_n)^T \left\{ (\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i) \odot \mathbf{\Omega}^{-(t)} \right\} (\mathbf{I}_d \otimes \mathbf{1}_n) + (\nu_i + d + 1) \mathbf{\Gamma}_i^{-t} \\ &= \mathbf{\Gamma}_i^{-1} (\mathbf{\Gamma}_i^{(t)} \mathbf{R}^{(t)T} \mathbf{V}_i \mathbf{R}^{(t)} \mathbf{\Gamma}_i^{(t)} + \mathbf{\Psi}_i) \mathbf{\Gamma}_i^{-1} \end{aligned}$$

and can be found using Lemma 6.

3.6.5 Variable Selection

In the statistical analysis of high-dimensional data, the imposition of sparsity leads to better interpretation and more stable parameter estimation. MM algorithms mesh well with penalized estimation. The simple variance components model (3.1) illustrates this fact. For the selection of fixed effects, minimizing the lasso-penalized log-likelihood

$$-L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_j |\beta_j|$$

is often recommended (Schelldorfer et al., 2011). The only change to the MM Algorithm 1 is that in estimating $\boldsymbol{\beta}$, one solves a lasso penalized general least squares problem rather than an ordinary general least squares problem. The updates of the variance components σ_i^2 remain the same. For selection among a large number of variance components, one can minimize the ridge-penalized log-likelihood

$$-L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^m \sigma_i^2$$

subject to the nonnegativity constraints $\sigma_i^2 \geq 0$. Here the standard deviations σ_i are the underlying parameters. The variance update (3.6) becomes

$$\sigma_i^{2(t+1)} = \sigma_i^{2(t)} \sqrt{\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})^T \mathbf{\Omega}^{-(t)} \mathbf{V}_i \mathbf{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)})}{\text{tr}(\mathbf{\Omega}^{-(t)} \mathbf{V}_i) + 2\lambda}}, \quad i = 1, \dots, m,$$

which clearly exhibits shrinkage but no thresholding. The lasso penalized log-likelihood

$$-L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) + \lambda \sum_{i=1}^m \sigma_i \quad (3.18)$$

subject to nonnegativity constraint $\sigma_i \geq 0$ achieves both ends. The update of σ_i is chosen among the positive roots of a quartic equation and the boundary 0, whichever yields a lower objective value.

3.7 A Numerical Example

Quantitative trait loci (QTL) mapping aims to identify genes associated with a quantitative trait. Current sequencing technology measures millions of genetic markers in study subjects. Traditional single-marker tests suffer from low power due to the low frequency of many markers and the corrections needed for multiple hypothesis testing. Region-based association tests are a powerful alternative for analyzing next generation sequencing data with abundant rare variants.

Suppose \mathbf{y} is a $n \times 1$ vector of quantitative trait measurements on n people, \mathbf{X} is an $n \times p$ predictor matrix (incorporating predictors such as sex, smoking history, and principal components for ethnic admixture), and \mathbf{G} is an $n \times m$ genotype matrix of m genetic variants in a pre-defined region. The linear mixed model assumes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{I}), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n),$$

where $\boldsymbol{\beta}$ are fixed effects, $\boldsymbol{\gamma}$ are random genetic effects, and σ_g^2 and σ_e^2 are variance components for the genetic and environmental effects, respectively. Thus, the phenotype vector \mathbf{Y} has covariance $\sigma_g^2 \mathbf{G}\mathbf{G}^T + \sigma_e^2 \mathbf{I}_n$, where $\mathbf{G}\mathbf{G}^T$ is the kernel matrix capturing the overall effect of the m variants. Current approaches test the null hypothesis $\sigma_g^2 = 0$ for each region separately and then adjust for multiple testing (Lee et al., 2014). Instead of this marginal testing strategy, we consider the joint model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + s_1^{-1/2} \mathbf{G}_1 \boldsymbol{\gamma}_1 + \cdots + s_m^{-1/2} \mathbf{G}_m \boldsymbol{\gamma}_m + \boldsymbol{\epsilon}, \\ \boldsymbol{\gamma}_i &\sim N(\mathbf{0}, \sigma_i^2 \mathbf{I}), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n) \end{aligned}$$

and select the variance components σ_i^2 via the penalization (3.18). 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.

We illustrate this approach using the COPDGene exome sequencing study (<http://www.copdgene.org/>) (Regan et al., 2010). After quality control, 399 individuals and 646,125 genetic variants remain for analysis. Genetic variants are grouped into 16,619 genes to expose those

Table 3.6: Top 10 genes selected by the lasso penalized variance component model (3.18) in an association study of 200 genes and the complex trait **height**.

Lasso Rank	Gene	Marginal P-value	# Variants
1	DOLPP1	2.35×10^{-6}	2
2	C9orf21	3.70×10^{-5}	4
3	PLS1	2.29×10^{-3}	5
4	ATP5D	6.80×10^{-7}	3
5	ADCY4	1.01×10^{-3}	11
6	SLC22A25	3.95×10^{-3}	14
7	RCSD1	9.04×10^{-4}	4
8	PCDH7	1.20×10^{-4}	7
9	AVIL	8.34×10^{-4}	11
10	AHR	1.14×10^{-3}	7

genes associated with the complex trait **height**. We include **age**, **sex**, and the top 3 principal components in the mean effects. Because the number of genes vastly exceeds the sample size $n = 399$, we first pare the 16,619 genes down to 200 genes according to their marginal likelihood ratio test p-values and then carry out penalized estimation of the 200 variance components in the joint model (3.18). This is similar to the sure independence screening strategy for selecting mean effects (Fan and Lv, 2008). Genes are ranked according to the order they appear in the lasso solution path. Table 3.6 lists the top 10 genes together with their marginal LRT p-values. Figure 3.1 displays the corresponding segment of the lasso solution path. It is noteworthy that the ranking of genes by penalized estimation differs from the ranking according to marginal p-values. The same phenomenon occurs in selection of highly correlated mean predictors. This penalization approach for selecting variance components warrants further theoretical study. It is reassuring that the simple MM algorithm scales to high-dimensional problems.

3.8 Discussion

The current chapter leverages the MM principle to design powerful and versatile algorithms for variance components estimation. The MM algorithms derived are notable for their simplicity, generality, numerical efficiency, and theoretical guarantees. Both ordinary MLE and REML are apt to benefit. Other extensions are possible. In nonlinear models (Bates and Watts, 1988; Lindstrom and Bates, 1990), the mean response is a nonlinear function in the fixed effects β . One can easily modify the MM algorithms to update β by a few rounds of Gauss-Newton iteration. The variance components updates remain unchanged.

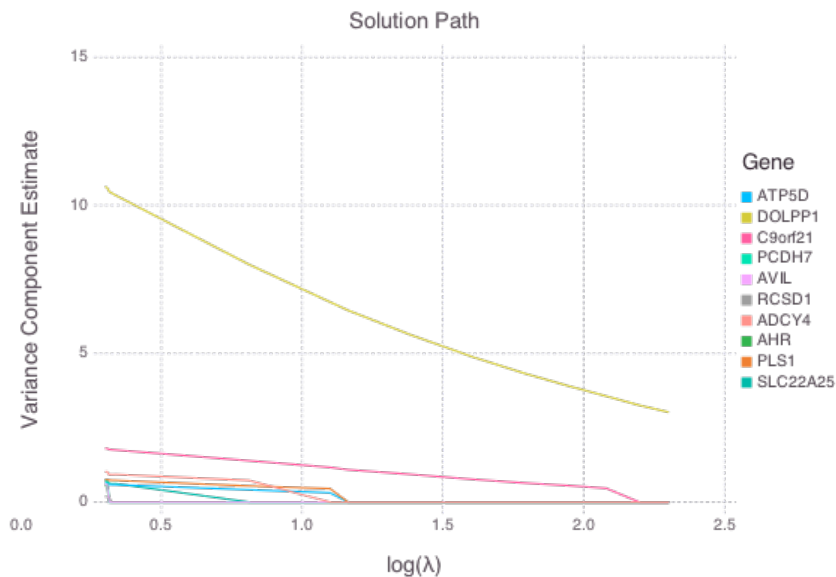


Figure 3.1: Solution path of the lasso penalized variance component model (3.18) in an association study of 200 genes and the complex trait **height**.

One can also extend our MM algorithms to elliptically symmetric densities

$$f(\mathbf{y}) = \frac{e^{-\frac{1}{2}\kappa(\delta^2)}}{(2\pi)^{\frac{n}{2}}(\det \mathbf{\Omega})^{\frac{1}{2}}}$$

defined for $\mathbf{y} \in \mathbb{R}^n$, where $\delta^2 = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ denotes the Mahalanobis distance between \mathbf{y} and $\boldsymbol{\mu}$. Here we assume that the function $\kappa(s)$ is strictly increasing and strictly concave. Examples of elliptically symmetric densities include the multivariate t , slash, contaminated normal, power exponential, and stable families. Previous work (Huber and Ronchetti, 2009; Lange and Sinsheimer, 1993) has focused on using the MM principle to convert parameter estimation for these robust families into parameter estimation under the multivariate normal. One can chain the relevant majorization $\kappa(s) \leq \kappa(s^{(t)}) + \kappa'(s^{(t)})(s - s^{(t)})$ with our previous minorizations and simultaneously split variance components and pass to the more benign setting of the multivariate normal.

Chapter 4

MM Algorithms for Logistic Linear Mixed Model

4.1 Introduction

Generalized linear mixed model (GLMM) is an extension of generalized linear model to incorporate random effects accounting for heterogeneity among responses (McCulloch and Neuhaus, 2001; Stroup, 2012). It is widely used in clustered, longitudinal, and panel data analysis (Zeger and Karim, 1991; Breslow and Clayton, 1993). Logistic linear mixed model is one of the GLMMs for binary responses and assumes

$$\begin{aligned} y_j | \eta_j &\sim \text{Bernoulli}(\mu_j) \\ \mu_j &= 1 / \{1 + \exp(-\eta_j)\} \end{aligned} \tag{4.1}$$

for $j = 1, \dots, n$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ takes the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \dots + \mathbf{Z}_m\mathbf{u}_m,$$

where \mathbf{X} and $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_m)$ are known predictor matrices, $\boldsymbol{\beta}$ is the coefficient vector for fixed effects, and $\mathbf{u}_i \sim N(\mathbf{0}_{q_i}, \sigma_i^2 \mathbf{I}_{q_i})$ are independent random effects. Because

$$\boldsymbol{\eta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_1^2 \mathbf{Z}_1 \mathbf{Z}_1^T + \dots + \sigma_m^2 \mathbf{Z}_m \mathbf{Z}_m^T),$$

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

Logistic linear mixed model finds applications in agriculture, econometrics, biology and genetics. Two motivating examples are the analysis of variance (ANOVA) for dichotomous responses (Anderson and Aitkin, 1985; Quené and Van den Bergh, 2008) and the quantitative

trait loci (QTL) mapping for binary traits (Yi and Xu, 1999; Che and Xu, 2012). In ANOVA, \mathbf{Z}_i corresponds to 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, \mathbf{Z}_i corresponds to a gene region. The number of genes m is at order of $10^2 \sim 10^3$ in a typical genetic study. In Section 4.3 and 4.4, we will discuss further about these two applications as well as 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. The existing methods involve various forms of approximations. The first class of methods use numerical integration such as Gaussian quadrature (Davidian and Gallant, 1992) and adaptive Gaussian quadrature (Pinheiro and Bates, 1995). These methods are applicable only to low dimensional integrals and thus limited to problems where data form very small independent clusters. The second type of methods invoke the 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 et al., 2009). The third class of methods resort to Monte Carlo methods to approximate either the original integral (Sung and Geyer, 2007) or the E step of EM algorithm (Booth and Hobert, 1999). Pinheiro and Bates (1995) compare and discuss penalized quasi-likelihood (PQL), Laplace approximation, importance sampling, Gaussian quadrature, and adaptive Gaussian quadrature (AGQ). They conclude that Laplace approximation and adaptive Gaussian quadrature give the “best mix of efficiency and accuracy”. In this chapter, 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 are considering.

Our primary interest is in the estimation and selection of variance components. Researchers have worked on selecting fixed effects in GLMMs (Groll and Tutz, 2014; Schelldorfer et al., 2014). For random effects selection, however, most procedures are developed in the framework of linear mixed models (Bondell et al., 2010; Ahn et al., 2012) for quantitative responses. In contrast only few references discuss random effects selection in GLMM. Ibrahim et al. (2011) develop a simultaneous fixed and random effects selection procedure based on the 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 the Bayesian framework using a stochastic search MCMC algorithm. Pan and Huang (2014) propose a backfitting algorithm to select effective random effects based on penalized quasi-likelihood (PQL) function. However all the above mentioned papers study the clustered data with repeated measurements on the subjects. They assume n independent subjects with observations $(\mathbf{y}_1, \mathbf{X}_1, \mathbf{Z}_1), \dots, (\mathbf{y}_n, \mathbf{X}_n, \mathbf{Z}_n)$ and

$$E(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) = g(\boldsymbol{\eta}_i) = g(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i), \quad (4.2)$$

where $g(\cdot)$ is some known link function, \mathbf{X}_i and \mathbf{Z}_i are known matrices and $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$ is the random effect. Here, \mathbf{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 \mathbf{D} , denoted as $\mathbf{\Gamma}$, such that each row of $\mathbf{\Gamma}$ either are all not zero or all zero and Pan and Huang (2014) penalize on positive elements proportional to the standard deviation of the random effects \mathbf{b}_i . In this chapter, we propose an algorithm for selection of random effects by shrinking the variances of ineffective random effects towards zero based on penalized likelihood defined in Section 4.2.3. There are two key differences between our variance components selection and previous work. First, model (4.2) is not the same as model (4.1) we want to address in this chapter even though they both can deal with clustered data as well as non-clustered data. Model (4.1) assumes that the random effects $\mathbf{u}_i \sim N(\mathbf{0}_{q_i}, \sigma_i^2 \mathbf{I}_{q_i})$ are independent. If we write model (4.1) in the framework of model (4.2), then the covariance in model (4.1) is diagonal with some equality constraints on the random effect variances while the covarariance in model (4.2) can be any covariance matrix. Second, the random effects selection on model (4.2) is selecting individual random effect while for model (4.1) we are selecting groups of random effects, i.e. the random effects in each u_i are either all selected or not. To the best of our knowledge, there exist no literature about variance components selection for model (4.1).

In this chapter, based on the minorization-maximization (MM) principle (Lange et al., 2000), we propose two novel algorithms for variance component estimation under two different parameterizations of logistic linear mixed model and then extend to variance component selection by incorporating penalization. The first parameterization is efficient for estimating parameters without penalty, while 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.

4.2 Algorithms for Estimation

4.2.1 Model Formulation 1

The likelihood for model (4.1) is

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \int \exp\{h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2)\} d\mathbf{u}, \quad (4.3)$$

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

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

Direct optimization of the likelihood defined in (4.3) is computationally challenging because of the integral. The Laplace approximation (LA) to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\sigma})$ is obtained by replacing $h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ by its second-order Taylor expansion at the conditional maximum. Given current iterate $(\boldsymbol{\beta}, \boldsymbol{\sigma})$, let \mathbf{u}^* be the maximizer of h and $\boldsymbol{\eta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^*$ where $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_m)$. Then the approximated log-likelihood is

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

where

$$h(\mathbf{u}^* \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \sum_{i=1}^m q_i \ln \sigma_i^2 - \frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2}.$$

The gradient and Hessian of $h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma})$ at $\mathbf{u} = \mathbf{u}^*$ are

$$\begin{aligned} \nabla_{\mathbf{u}} h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2)_{\mid \mathbf{u}=\mathbf{u}^*} &= \mathbf{Z}^T (\mathbf{y} - \mathbf{p}^*) - \begin{pmatrix} \sigma_1^{-2} \mathbf{u}_1^* \\ \vdots \\ \sigma_m^{-2} \mathbf{u}_m^* \end{pmatrix}, \\ \nabla_{\mathbf{u}}^2 h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2)_{\mid \mathbf{u}=\mathbf{u}^*} &= -\{ \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} + \text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m}) \}, \end{aligned}$$

where $\mathbf{p}^* = (p_1^*, \dots, p_n^*)^T$ with $p_j^* = e^{\eta_j^*} / (1 + e^{\eta_j^*})$ and $\mathbf{W}^* = \text{diag}(\mathbf{w}^*)$ is a diagonal matrix with entries

$$w_j^* = p_j^* (1 - p_j^*) = \frac{e^{\eta_j^*}}{(1 + e^{\eta_j^*})^2}.$$

Therefore,

$$L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \sum_{i=1}^m q_i \ln \sigma_i^2 - \frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} - \frac{1}{2} \ln \det \{ \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} + \text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m}) \}. \quad (4.4)$$

Using the matrix determinant lemma, we have

$$\begin{aligned} & \ln \det \{ \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} + \text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m}) \} \\ = & \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) + \ln \det (\text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m})) + \ln \det \mathbf{W}^* \\ = & \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) - \sum_{i=1}^m q_i \ln \sigma_i^2 + \ln \det \mathbf{W}^*. \end{aligned} \quad (4.5)$$

Substitute (4.5) to (4.4) gives

$$L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} - \frac{1}{2} \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) - \frac{1}{2} \ln \det \mathbf{W}^* + \text{constant term}, \quad (4.6)$$

where the constant term equals $-\frac{n}{2} \ln 2\pi$.

The MM algorithm cycles through following updates of \mathbf{u} , $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$.

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

$$\begin{aligned} \nabla_{\mathbf{u}} h &= \mathbf{Z}^T (\mathbf{y} - \mathbf{p}) - \begin{pmatrix} \sigma_1^{-2} \mathbf{u}_1 \\ \vdots \\ \sigma_m^{-2} \mathbf{u}_m \end{pmatrix} \\ \nabla_{\mathbf{u}}^2 h &= - \{ \mathbf{Z}^T \mathbf{W} \mathbf{Z} + \text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m}) \}, \end{aligned}$$

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

$$\nabla_{\mathbf{u}}^2 h \succeq - \{ 0.25 \mathbf{Z}^T \mathbf{Z} + \text{blkdiag}(\sigma_1^{-2} \mathbf{I}_{q_1}, \dots, \sigma_m^{-2} \mathbf{I}_{q_m}) \}.$$

Thus we can construct a quadratic minorization function at $\mathbf{u}^{(l)}$ using (2.5) and maxi-

mizing the quadratic surrogate gives the MM update

$$\mathbf{u}^{(l+1)} = \mathbf{u}^{(l)} + \{0.25\mathbf{Z}^T\mathbf{Z} + \text{blkdiag}(\sigma_1^{-2}\mathbf{I}_{q_1}, \dots, \sigma_m^{-2}\mathbf{I}_{q_m})\}^{-1} \nabla_{\mathbf{u}} h(\mathbf{u}^{(l)}). \quad (4.7)$$

To find the maximizer \mathbf{u}^* given $\boldsymbol{\beta}, \boldsymbol{\sigma}^2$, we iterate the MM update (4.7) until convergence. Note that the indicated matrix inverse in (4.7) only needs to be done once and remains constant through the iterations.

2. Updating $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ and \mathbf{u}^* is a regular logistic regression with offset $\mathbf{Z}\mathbf{u}^*$. We invoke a similar MM update as above

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (0.25\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T(\mathbf{y} - \mathbf{p}^*). \quad (4.8)$$

Again the matrix inverse $(0.25\mathbf{X}^T\mathbf{X})^{-1}$ only needs to be done once.

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

$$g(\boldsymbol{\sigma}^2 | \boldsymbol{\sigma}^{2(t)}) = -\frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} - \frac{1}{2} \sum_{i=1}^m \sigma_i^2 \text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{z}_i \mathbf{z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{z}_i \mathbf{z}_i^T \right\} + c^{(t)}, \quad (4.9)$$

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

$$\sigma_i^{2(t+1)} = \left[\frac{\|\mathbf{u}_i^*\|_2^2}{\text{tr} \left\{ \mathbf{z}_i^T (\sum_i \sigma_i^{2(t)} \mathbf{z}_i \mathbf{z}_i^T + \mathbf{W}^{*-1})^{-1} \mathbf{z}_i \right\}} \right]^{\frac{1}{2}}.$$

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

$$\begin{aligned} & \left(\sum_i \sigma_i^{(t)2} \mathbf{z}_i \mathbf{z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \\ &= \mathbf{W}^* - \mathbf{W}^* \mathbf{Z}(\boldsymbol{\sigma}) \{ \mathbf{I}_q + \mathbf{Z}(\boldsymbol{\sigma})^T \mathbf{W}^* \mathbf{Z}(\boldsymbol{\sigma}) \}^{-1} \mathbf{Z}(\boldsymbol{\sigma})^T \mathbf{W}^*, \end{aligned}$$

where $\mathbf{Z}(\boldsymbol{\sigma}) = (\sigma_1 \mathbf{Z}_1, \dots, \sigma_m \mathbf{Z}_m)$. Since the iterate is derived based on MM principle, it possesses the ascent property

$$L_{LA}(\boldsymbol{\sigma}^{(t+1)} | \boldsymbol{\beta}, \mathbf{u}^*) \geq L_{LA}(\boldsymbol{\sigma}^{(t)} | \boldsymbol{\beta}, \mathbf{u}^*). \quad (4.10)$$

Proof. From (4.6), the approximated log-likelihood is

$$\begin{aligned} L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) &= \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} \\ &\quad - \frac{1}{2} \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) - \frac{1}{2} \ln \det \mathbf{W}^* \\ &\quad + \text{terms without } \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \end{aligned}$$

where \mathbf{u}^* is the maximizer of $h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma})$, $\boldsymbol{\eta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^*$ and $\mathbf{W}^* = \text{diag}(\mathbf{w}^*)$ is a diagonal matrix with entries

$$w_j^* = p_j^*(1 - p_j^*) = \frac{e^{\eta_j^*}}{(1 + e^{\eta_j^*})^2}.$$

Thus

$$L_{\text{LA}}(\boldsymbol{\sigma} \mid \boldsymbol{\beta}, \mathbf{u}^*) = -\frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} - \frac{1}{2} \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) + c,$$

where $c = \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \ln \det \mathbf{W}^* - \frac{n}{2} \ln 2\pi$ is a constant not involving $\boldsymbol{\sigma}$.

The minorization (2.4) leads to the surrogate function of $L_{\text{LA}}(\boldsymbol{\sigma} \mid \boldsymbol{\beta}, \mathbf{u}^*)$

$$\begin{aligned} g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)}) &= -\frac{1}{2} \sum_{i=1}^m \frac{\|\mathbf{u}_i^*\|_2^2}{\sigma_i^2} - \frac{1}{2} \sum_{i=1}^m \sigma_i^2 \text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \right\} \\ &\quad + c^{(t)}, \end{aligned}$$

where $c^{(t)}$ is a constant irrelevant to optimization. Since $\boldsymbol{\sigma}^{2(t+1)} = (\sigma_1^{2(t+1)}, \dots, \sigma_m^{2(t+1)})$ with

$$\sigma_i^{2(t+1)} = \left[\frac{\|\mathbf{u}_i^*\|_2^2}{\text{tr} \left\{ \mathbf{Z}_i^T (\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1})^{-1} \mathbf{Z}_i \right\}} \right]^{\frac{1}{2}}$$

maximizes the surrogate function $g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)})$, we have the following inequality satisfied

$$L_{\text{LA}}(\boldsymbol{\sigma}^{(t+1)} \mid \boldsymbol{\beta}, \mathbf{u}^*) \geq g(\boldsymbol{\sigma}^{2(t+1)} \mid \boldsymbol{\sigma}^{2(t)}) \geq g(\boldsymbol{\sigma}^{2(t)} \mid \boldsymbol{\sigma}^{2(t)}) = L_{\text{LA}}(\boldsymbol{\sigma}^{(t)} \mid \boldsymbol{\beta}, \mathbf{u}^*).$$

Therefore, the iterates possess the ascent property. \square

Like the penalized iteratively reweighted least squares (PIRLS) algorithm described in Bates et al. (2015), parameter estimates are determined for a fixed weights matrix \mathbf{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 5 summarizes the MM algorithm for parameter estimation of the logistic linear mixed model (4.1). Each iteration involves one-step update of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Several more steps of updating $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ give similar results in practice.

Input : $\mathbf{y}, \mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_m$
Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \dots, \hat{\sigma}_m^2$

- 1 Initialize $\boldsymbol{\beta}^{(0)}, \sigma_i^{(0)} > 0, i = 1, \dots, m$;
- 2 **repeat**
- 3 $\mathbf{u}^* \leftarrow \arg \max_{\mathbf{u}} h(\mathbf{u} \mid \boldsymbol{\sigma}^{2(t)}, \boldsymbol{\beta}^{(t)})$;
- 4 $\mathbf{p}^* \leftarrow 1 / \left\{ 1 + \exp \left(-\mathbf{X}\boldsymbol{\beta}^{(t)} - \mathbf{Z}\mathbf{u}^* \right) \right\}$;
- 5 $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} + \left(0.25\mathbf{X}^T\mathbf{X} \right)^{-1} \mathbf{X}^T(\mathbf{y} - \mathbf{p}^*)$;
- 6 $\mathbf{p}^* \leftarrow 1 / \left\{ 1 + \exp \left(-\mathbf{X}\boldsymbol{\beta}^{(t+1)} - \mathbf{Z}\mathbf{u}^* \right) \right\}$;
- 7 $\mathbf{W}^* \leftarrow \text{diag} \{ \mathbf{p}^*(1 - \mathbf{p}^*) \}$;
- 8 $\sigma_i^{2(t+1)} \leftarrow \left[\frac{\|\mathbf{u}_i^*\|_2^2}{\text{tr} \{ \mathbf{Z}_i^T (\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1})^{-1} \mathbf{Z}_i \}} \right]^{\frac{1}{2}}, \quad i = 1, \dots, m$;
- 9 **until** *objective value converges*;

Algorithm 5: MMLA1 - a MM algorithm to maximize the Laplace approximation of likelihood for model (4.1).

4.2.2 Model Formulation 2

In Laplace approximated log-likelihood (4.6), we have σ_i in the denominator, thus it cannot be combined with penalized estimation which will shrink some of σ_i s to zero. Therefore we consider another reparameterization of model (4.1) by assuming that $\boldsymbol{\eta}$ takes the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \sigma_1\mathbf{Z}_1\mathbf{u}_1 + \dots + \sigma_m\mathbf{Z}_m\mathbf{u}_m, \quad (4.11)$$

where $\mathbf{u}_i \sim N(\mathbf{0}_{q_i}, \mathbf{I}_{q_i})$ are independent. Let $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_m^T)^T \in \mathcal{R}^q$ be the concatenated random effects and $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_m) \in \mathcal{R}^{n \times q}$, $q = \sum_{i=1}^m q_i$. Then $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{D}\mathbf{u}$, where $\mathbf{D} = \text{blkdiag}(\sigma_1\mathbf{I}_{q_1}, \dots, \sigma_m\mathbf{I}_{q_m})$ and the complete log-likelihood is

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

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

$$L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) = h(\mathbf{u}^* | \boldsymbol{\beta}, \boldsymbol{\sigma}) - \frac{1}{2} \ln \det \nabla^2 \{-h(\mathbf{u}^* | \boldsymbol{\beta}, \boldsymbol{\sigma}^2)\},$$

where

$$h(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_j \{y_j \eta_j - \ln(1 + e^{\eta_j})\} - \frac{1}{2} \|\mathbf{u}\|_2^2 - \frac{n}{2} \ln 2\pi.$$

The gradient and Hessian $h(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma})$ at $\mathbf{u} = \mathbf{u}^*$ are

$$\begin{aligned} \nabla_{\mathbf{u}} h(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma})|_{\mathbf{u}=\mathbf{u}^*} &= \mathbf{D}^T \mathbf{Z}^T (\mathbf{y} - \mathbf{p}^*) - \mathbf{u}^*, \\ \nabla_{\mathbf{u}}^2 h(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma})|_{\mathbf{u}=\mathbf{u}^*} &= -(\mathbf{D}^T \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} \mathbf{D} + \mathbf{I}_q), \end{aligned}$$

where $\mathbf{p}^* = (p_1^*, \dots, p_n^*)^T$ with $p_j^* = e^{\eta_j^*} / (1 + e^{\eta_j^*})$ and $\mathbf{W}^* = \text{diag}(\mathbf{w}^*)$ is a diagonal matrix with entries

$$w_j^* = p_j^*(1 - p_j^*) = \frac{e^{\eta_j^*}}{(1 + e^{\eta_j^*})^2}.$$

Using the matrix determinant lemma, we have

$$\ln \det (\mathbf{D}^T \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} \mathbf{D} + \mathbf{I}_q) = \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) + \ln \det \mathbf{W}^*.$$

Therefore,

$$\begin{aligned} &L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) \\ &= \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \|\mathbf{u}^*\|_2^2 - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln \det (\mathbf{D}^T \mathbf{Z}^T \mathbf{W}^* \mathbf{Z} \mathbf{D} + \mathbf{I}_q) \\ &= \sum_j \left\{ y_j \eta_j^* - \ln(1 + e^{\eta_j^*}) \right\} - \frac{1}{2} \|\mathbf{u}^*\|_2^2 - \frac{1}{2} \ln \det \left(\mathbf{W}^{*-1} + \sum_i \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T \right) \\ &\quad - \frac{1}{2} \ln \det \mathbf{W}^* + \text{constant term}, \end{aligned} \tag{4.12}$$

where the constant term equals $-\frac{n}{2} \ln 2\pi$.

Maximizing $h(\mathbf{u} | \boldsymbol{\beta}, \boldsymbol{\sigma})$ follows similar MM updates as in (4.7). Given $\boldsymbol{\sigma}^2$ and $\boldsymbol{\beta}$, \mathbf{u}^* can be

found through MM iterates

$$\mathbf{u}^{(l+1)} = \mathbf{u}^{(l)} + \{0.25(\mathbf{ZD})^T \mathbf{ZD} + \mathbf{I}_q\}^{-1} \nabla_{\mathbf{u}} h(\mathbf{u}^{(l)} | \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$$

until convergence, where $\nabla_{\mathbf{u}} h(\mathbf{u}^{(l)} | \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \mathbf{D}^T \mathbf{Z}^T (\mathbf{y} - \mathbf{p}) - \mathbf{u}^{(l)}$. Updating $\boldsymbol{\beta}$ given \mathbf{u}^* and $\boldsymbol{\sigma}^2$ is the same as update in (4.8).

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

$$\begin{aligned} -1^T \ln(1 + e^{\boldsymbol{\eta}^*}) &\geq -\mathbf{p}^{(t)T} (\boldsymbol{\eta}^* - \boldsymbol{\eta}^{*(t)}) - \frac{1}{8} \|\boldsymbol{\eta}^* - \boldsymbol{\eta}^{*(t)}\|_2^2 + c^{(t)} \\ &= -\mathbf{p}^{(t)T} \mathbf{ZD}\mathbf{u}^* - \frac{1}{8} \|\mathbf{Z}(\mathbf{D} - \mathbf{D}^{(t)})\mathbf{u}^*\|_2^2 + c^{(t)}, \end{aligned} \quad (4.13)$$

where $c^{(t)}$ is an irrelevant constant, $\mathbf{p}^{(t)}$ is a vector with the j th element equal to

$$p_j^{(t)} = e^{\eta_j^{*(t)}} / (1 + e^{\eta_j^{*(t)}})$$

and $\eta_j^{*(t)}$ is the j th element of $\boldsymbol{\eta}^{*(t)} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZD}^{(t)}\mathbf{u}^*$. The Cauchy inequality implies that

$$\begin{aligned} -\|\mathbf{Z}(\mathbf{D} - \mathbf{D}^{(t)})\mathbf{u}^*\|_2^2 &= -\left\| \sum_{i=1}^m \mathbf{Z}_i \mathbf{u}_i^* (\sigma_i - \sigma_i^{(t)}) \right\|_2^2 \\ &\geq -\left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \sum_{i=1}^m (\sigma_i - \sigma_i^{(t)})^2, \end{aligned} \quad (4.14)$$

where $(\mathbf{Z}_i \mathbf{u}_i^*)_j$ is the j th element of vector $\mathbf{Z}_i \mathbf{u}_i^*$. Combining (4.13), (4.14) and (2.4) gives the overall minorization function

$$\begin{aligned} g(\boldsymbol{\sigma} | \boldsymbol{\sigma}^{(t)}) &= \sum_{i=1}^m \sigma_i (\mathbf{y} - \mathbf{p}^{(t)})^T \mathbf{Z}_i \mathbf{u}_i^* - \frac{1}{8} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \sum_{i=1}^m (\sigma_i - \sigma_i^{(t)})^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^m \sigma_i^2 \text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \right\} + c^{(t)}, \end{aligned} \quad (4.15)$$

where σ_i are nicely separated and only involve quadratic terms. Maximization of $g(\boldsymbol{\sigma} | \boldsymbol{\sigma}^{(t)})$

results the following update

$$\sigma_i^{(t+1)} = \frac{(\mathbf{y} - \mathbf{p}^{(t)})^T \mathbf{Z}_i \mathbf{u}_i^* + \frac{1}{4} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \sigma_i^{(t)}}{\text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \right\} + \frac{1}{4} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\}}. \quad (4.16)$$

To account for the non-negative constraint of $\boldsymbol{\sigma}$, at each iteration we set $\sigma_i^{(t+1)} = \max(0, \sigma_i^{(t+1)})$. Algorithm 6 summarizes the MM algorithm for model formulation 2 defined in (4.11).

Input : $\mathbf{y}, \mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_m$
Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \dots, \hat{\sigma}_m^2$

- 1 Initialize $\boldsymbol{\beta}^{(0)}, \sigma_i^{(0)} > 0, i = 1, \dots, m$;
- 2 **repeat**
- 3 $\mathbf{D}^{(t)} = \text{diag} \left(\sigma_1^{(t)} \mathbf{1}_{q_1}, \dots, \sigma_m^{(t)} \mathbf{1}_{q_m} \right)$;
- 4 $\mathbf{u}^* \leftarrow \arg \max_{\mathbf{u}} h(\mathbf{u} \mid \boldsymbol{\sigma}^{2(t)}, \boldsymbol{\beta}^{(t)})$;
- 5 $pbf^{(t)} \leftarrow 1 / \left\{ 1 + \exp \left(-\mathbf{X} \boldsymbol{\beta}^{(t)} - \mathbf{Z} \mathbf{D}^{(t)} \mathbf{u}^* \right) \right\}$;
- 6 $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} + (0.25 \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{p}^{(t)})$;
- 7 $\mathbf{p}^{(t)} \leftarrow 1 / \left\{ 1 + \exp \left(-\mathbf{X} \boldsymbol{\beta}^{(t+1)} - \mathbf{Z} \mathbf{D}^{(t)} \mathbf{u}^* \right) \right\}$;
- 8 $\mathbf{W}^* \leftarrow \text{diag} \left\{ \mathbf{p}^{(t)} (1 - \mathbf{p}^{(t)}) \right\}$;
- 9 $\sigma_i^{2(t+1)} \leftarrow \max \left[0, \frac{(\mathbf{y} - \mathbf{p}^{(t)})^T \mathbf{Z}_i \mathbf{u}_i^* + \frac{1}{4} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \sigma_i^{(t)}}{\text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \right\} + \frac{1}{4} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\}} \right]$, $i =$
 $1, \dots, m$;
- 10 **until** objective value converges;

Algorithm 6: MMLA2 - a MM algorithm to maximize the Laplace approximation of likelihood for model (4.11).

4.2.3 MM Algorithm for Maximizing the Penalized Approximated Likelihood

For variance component selection, we consider the penalization approach using lasso penalty. Since the minorization function of $\boldsymbol{\sigma}$ derived in second model formulation is a quadratic function of $\boldsymbol{\sigma}$, it meshes well with penalized estimation. Other penalties such as the adaptive lasso (Zou, 2006) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) lead to similar algorithms.

The lasso penalized approximated log-likelihood is

$$-L_{\text{LA}}(\boldsymbol{\beta}, \boldsymbol{\sigma}) + \lambda \sum_{i=1}^m |\sigma_i|. \quad (4.17)$$

Finding \mathbf{u}^* to maximize $h(\mathbf{u} \mid \boldsymbol{\beta}, \boldsymbol{\sigma})$ and updating $\boldsymbol{\beta}$ are the same as described in algorithm 6. The only difference lies in the update of $\boldsymbol{\sigma}$ given \mathbf{u}^* and $\boldsymbol{\beta}$ in (4.16), which now becomes

$$\begin{aligned} \sigma_i^{(t+1)} &= \arg \min_{\sigma_i} \sigma_i^2 \left[\frac{1}{2} \text{tr} \left\{ \left(\sum_i \sigma_i^{2(t)} \mathbf{Z}_i \mathbf{Z}_i^T + \mathbf{W}^{*-1} \right)^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \right\} + \frac{1}{8} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \right] \\ &\quad - \sigma_i \left[\left(\mathbf{y} - \mathbf{p}^{(t)} \right)^T \mathbf{Z}_i \mathbf{u}_i^* + \frac{1}{4} \left\{ \sum_{j=1}^n \sum_{i=1}^m (\mathbf{Z}_i \mathbf{u}_i^*)_j^2 \right\} \sigma_i^{(t)} \right] + \lambda |\sigma_i| \\ &= ST(z_i, \gamma_i), \end{aligned} \quad (4.18)$$

where

$$ST(z, \gamma) = \arg \min_x \frac{1}{2}(x - z)^2 + \gamma|x| = \text{sng}(z) (|z| - \gamma)_+ \quad (4.19)$$

is the soft-thresholding operator and

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

4.2.4 Choice of Regularization Parameter

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

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

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

4.3 Simulation Studies

4.3.1 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) `glmm()` function in the `glmm` package in R (Knudson, 2016) and (3) `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 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 variance components. `rstanarm` package is an R interface to the Stan C++ library for Bayesian estimation. `stan_glmer()` adds independent prior distributions on the regression coefficients as well as priors on the covariance matrices of the group-specific parameters and perform Bayesian inference via MCMC.

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

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

where $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\gamma_j \sim N(0, \sigma_\gamma^2)$ and $(\alpha\gamma)_{ij} \sim N(0, \sigma_{\alpha\gamma}^2)$ are jointly independent. Here i indexes levels in factor 1, j indexes levels in factor 2, and k indexes observations in the (i, j) -combination. This corresponds to $m = 3$ variance components. Table 4.1 displays the results when there are $a = b = 5$ levels of each factor, the number of observations c in each combination of factor levels varies from 2 to 200, and the true parameter values are $(\beta_1, \beta_2, \beta_3, \sigma_\alpha^2, \sigma_\gamma^2, \sigma_{\alpha\gamma}^2) = (0.6, 1.0, -1.0, 0.5, 0.9, 0.3)$. For each scenario, we simulated 50 replicates. The sample size was $n = abc$ for each replicate. Therefore the largest model in Table 4.1 involves covariance matrix of size 5000×5000 . For $c = 100$ and 200, we omit the results of `glmm` and `rstanarm` since they take too much time when sample size gets larger and the whole simulation takes more than a week to complete.

We made the following observations. Two MM algorithms (MMLA1 and MMLA2) have very close results, but MMLA2 takes longer time to converge than MMLA1, especially when the number of groups c is large. This is what we expected since the surrogate function derived in MMLA2 involves two more 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 MM algorithms. For other values of c , `glmer()` delivered estimates comparable

to MM algorithm but was 3 ~ 4 fold slower than MMLA1. `glmm()` and `stan_glmer()` are much slower since they involve sampling and their estimation performance are 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 structure. Although it is hard to draw conclusions based on implementations in different languages, this example clearly demonstrates the efficiency and scalability of the MM algorithms for GLMM estimation.

Table 4.1: 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. Results for `rstanarm` and `glmm` with $c = 100, 200$ are not reported because the simulation takes more than 1 week.

c	Method	runtime	$\beta_1(0.6)$	$\beta_2(1.0)$	$\beta_3(-1.0)$	$\sigma_\alpha^2(0.5)$	$\sigma_\gamma^2(0.9)$	$\sigma_{\alpha\gamma}^2(0.3)$
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(2049)
	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	3613.26(2272.85)	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	—	—	—	—	—	—	—

4.3.2 Genetic Example

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

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma},$$

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

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + s_1^{-1/2} \mathbf{G}_1 \boldsymbol{\gamma}_1 + \cdots + s_m^{-1/2} \mathbf{G}_m \boldsymbol{\gamma}_m, \quad (4.20)$$

where $\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \sigma_i^2 \mathbf{I})$ and select the variance components σ_i^2 via the penalization (4.17). 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 COPDGene exome sequencing study (Regan et al., 2010), which has 399 subjects and genotype information of 16,610 genes. The covariate matrix \mathbf{X} contains `intercept`, `age`, `sex`, and the top 3 principal components in the mean effects. We consider four experimental settings for sparse random effects. In all the examples, we set $\boldsymbol{\beta} = (0.1, -1.0, 0.8, -0.3, -1.2, 1.5)$ and randomly select m genes \mathbf{G}_i , $i = 1, \dots, m$, from the COPD data.

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

We use mean squared error (MSE) = $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ to evaluate the performance of 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 AIC and BIC. Table 4.2, 4.3, 4.4 and 4.5 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 AIC and BIC can identify the truly significant random effects 97% \sim 99% of the time with AIC more prone to over-selection than BIC. Setting 3 and Setting 4 are more challenging since they involve a larger number of random effects. But our method can still identify the non-zero random effect 96% of the time under $m = 10$ when using AIC.

Table 4.2: Estimation and selection results for Setting 1.

m	Criteria	MSE (β)	Variance components selection			
			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 4.3: Estimation and selection results for Setting 2.

m	Criteria	MSE (β)	Variance components selection			
			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 4.4: Estimation and selection results for Setting 3.

m	Criteria	MSE (β)	Variance components selection			
			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 4.5: Estimation and selection results for Setting 4.

m	Criteria	MSE (β)	Variance components selection			
			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%

4.4 Real Data Analysis

In this real data analysis, we still use the data from COPDGene exome sequencing study described in the above simulated genetic example. The binary trait is smoke or not (denoted as `smoke`). There are 399 individuals with 646,125 genetic variants in 16,610 genes. The covariates include `age`, `sex`, and the top 3 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) and then carry out penalized estimation of the 200 variance components in the joint model (4.20). This is similar to the sure independence screening strategy for selecting mean effects (Fan and Lv, 2008). AIC selects 16 genes, while BIC criteria selects only one gene “AFAP1L2”. Table 4.6 lists the top 5 genes selected using AIC criteria (PLVC-AIC) and SKAT. We can see that the top 3 genes selected using both methods are the same but with different order. To compare the selection performance between SKAT and PLVC-AIC, we evaluate the log-likelihood of model (4.20) with the top 5 genes listed in Table 4.6 entering the model one by one. 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 4.1, we can see that the log-likelihood with genes selected by PLVC-AIC is above that of SKAT, which in some sense indicates that genes selected by PLVC-AIC explain more variability in the model.

Besides, we also compare the prediction performance between the top 5 genes selected by PLVC-AIC and SKAT. We evaluate the prediction performance using model (4.20) by including the genotype matrix \mathbf{G}_i of the corresponding selected genes similar to what is done in Wu et al. (2011). For example, if the genotype matrix of the top k genes selected are $\mathbf{G}_{h_1}, \mathbf{G}_{h_2}, \dots, \mathbf{G}_{h_k}$, then the predictive model becomes

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + s_{h_1}^{-1/2}\mathbf{G}_{h_1}\boldsymbol{\gamma}_1 + \dots + s_{h_k}^{-1/2}\mathbf{G}_{h_k}\boldsymbol{\gamma}_k = \mathbf{X}^*\boldsymbol{\beta}^*,$$

where $\mathbf{X}^* = (\mathbf{X}, s_{h_1}^{-1/2}\mathbf{G}_{h_1}, \dots, s_{h_k}^{-1/2}\mathbf{G}_{h_k})$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_k^T)$. This is the ordinary logistic regression model that can be used for prediction. Table 4.7 summarizes the prediction performance using 5-fold cross validation as the top 5 genes selected by both methods entering the model (4.20) one by one. We can see 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.

4.5 Discussion

This chapter discusses two MM algorithms for variance component estimation and selection in 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. This chapter only considers the binary response. The extension of the algorithm MMLA1 to the Poisson count data is straightforward with almost identical derivation. There is work on selecting fixed effects in GLMMs in literature. Here we only focus on random effects selection. Our algorithms can be easily extend to selecting fixed and random effects simultaneously. We leave a thorough study

Table 4.6: Top 5 genes selected by (1) the lasso penalized variance component model (4.17) with AIC criterion (PLVC-AIC) and (2) SKAT in an association study of 200 genes and the binary trait **smoke**.

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

Table 4.7: 5-fold cross validation performance on prediction accuracy with top 5 genes selected by PLVC-AIC and SKAT added to the model respectively in an association study of 200 genes and the complex trait **smoke**.

No. of genes entered into model	Prediction accuracy	
	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%)

of these topics to future research.

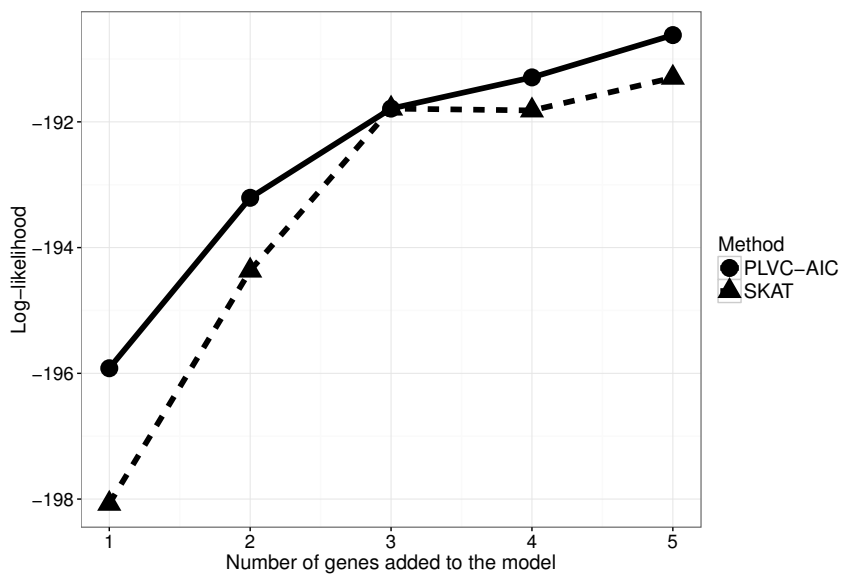


Figure 4.1: Log-likelihood evaluation with top 5 genes selected by PLVC-AIC and SKAT added to the model respectively in an association study of 200 genes and the complex trait `smoke`.

REFERENCES

- Ahn, M., Zhang, H. H., and Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, 22(4):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)*, pages 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):1–48.
- Bates, D. and Pinheiro, J. (1998). Computational methods for multilevel models. Technical Report Technical Memorandum BL0112140-980226-01TM, Bell Labs, Lucent Technologies, Murray Hill, NJ.
- Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York.
- 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(4):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(1):265–285.
- Borg, I. and Groenen, P. J. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.
- Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62(2):446–457.
- Callanan, T. P. and Harville, D. A. (1991). Some new algorithms for computing restricted maximum likelihood estimates of variance components. *J. Statist. Comput. Simulation*, 38(1-4):239–259.
- Che, X. and Xu, S. (2012). Generalized linear mixed models for mapping multiple quantitative trait loci. *Heredity*, 109(1):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(5):529–556.
- De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In *Information Systems and Data Analysis*, pages 308–324. Springer.
- Demidenko, E. (2013). *Mixed Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition. Theory and applications with R.
- Demidenko, E. and Massam, H. (1999). On the existence of the maximum likelihood estimate in variance components models. *Sankhyā Ser. A*, 61(3):431–443.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B.*, 39(1-38).
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):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(5):849–911.

- Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.
- Geyer, C. J. (1990). Likelihood and exponential families.
- Glanz, H. and Carvalho, L. (2013). An expectation-maximization algorithm for the matrix normal distribution. *arXiv*, 1309.6609.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition.
- Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by l₁-penalized estimation. *Statistics and Computing*, 24(2):137–154.
- Grzadziel, M. and Michalski, A. (2014). A note on the existence of the maximum likelihood estimate in variance components models. *Discuss. Math. Probab. Stat.*, 34(1-2):159–167.
- Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.
- Hartley, H. O. and Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93–108.
- Harville, D. and Callanan, T. (1990). Computational aspects of likelihood-based inference for variance components. In Gianola, D. and Hammond, K., editors, *Advances in Statistical Methods for Genetic Improvement of Livestock*, volume 18 of *Advanced Series in Agricultural Sciences*, pages 136–176. Springer Berlin Heidelberg.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72(358):320–340. With a comment by J. N. K. Rao and a reply by the author.
- Heiser, W. J. (1995). Convergent computation by iterative majorization: theory and applications

- in multidimensional data analysis. *Recent Advances in Descriptive Multivariate Analysis*, pages 157–189.
- Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press, Cambridge.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.
- Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32(1):384–406.
- Hunter, D. R. and Lange, K. (2004a). A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37.
- Hunter, D. R. and Lange, K. (2004b). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.
- Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.*, 33(4):1617–1642.
- Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.
- Jeon, M. (2012). *Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth*. PhD thesis, University of California, Berkeley.
- Khuri, A. I., Mathew, T., and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.
- Kiers, H. A. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational Statistics & Data Analysis*, 41(1):157–170.

- Knudson, C. (2016). *glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation*. R package version 1.1.1.
- Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Amer. Statist. Assoc.*, 82(397):97–105.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.
- Lange, K. (2010). *Numerical Analysis for Statisticians*. Statistics and Computing. Springer, New York, second edition.
- Lange, K. (2016). *MM optimization algorithms*, volume 147. SIAM.
- Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.*, 9(1):1–59. With discussion, and a rejoinder by Hunter and Lange.
- Lange, K., Papp, J., Sinsheimer, J., Sripracha, R., Zhou, H., and Sobel, E. (2013). Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics*, 29:1568–1570.
- Lange, K. and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2:175–198.
- Lange, K. and Zhou, H. (2014). MM algorithms for geometric and signomial programming. *Mathematical Programming Series A*, 143:339–356.
- Lee, S., Abecasis, G., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5 – 23.
- Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.*, 83(404):1014–1022.

- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687.
- McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized linear mixed models*. Wiley Online Library.
- McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.
- Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909.
- Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24(5):725–738.
- Pinheiro, J. and Bates, D. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.
- 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(1):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(4):413–425.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. John Wiley & Sons.
- Rao, C. R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, volume 3 of *North-Holland Series in Statistics and Probability*. North-Holland Publishing Co., Amsterdam.

- Rao, P. S. R. S. (1997). *Variance Components Estimation*, volume 78 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. Mixed models, methodologies and applications.
- 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. (2010). Genetic epidemiology of COPD (COPDGene) study designs. *COPD*, 7:32–43.
- Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.*, 79(386):406–414.
- 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(2):319–392.
- Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.*, 11(2):437–457.
- Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1 -penalization. *Scand. J. Stat.*, 38(2):197–214.
- Schelldorfer, J., Meier, L., and Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using 1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.
- Schur, J. (1911). Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *J. Reine Angew. Math.*, (140):1–28.
- Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.
- 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*, pages 990–1011.
- Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Statist.*, 35(2):335–353.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer-Verlag, New York.
- Weiss, R. E. (2005). *Modeling Longitudinal Data*. Springer Texts in Statistics. Springer, New York.
- Wolfinger, R. (1993). Laplace’s approximation for nonlinear mixed models. *Biometrika*, pages 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(1):82–93.
- Wu, T. T. and Lange, K. (2010). The MM alternative to EM. *Statistical Science*, 25:492–505.
- Yi, N. and Xu, S. (1999). Mapping quantitative trait loci for complex binary traits in outbred populations. *Heredity*, 82(6):668–676.
- Yu, Y. (2010). Monotonic convergence of a general algorithm for computing optimal designs. *Ann. Statist.*, 38(3):1593–1606.
- 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(413):79–86.

- Zhou, H., Alexander, D., and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing*, 21:261–273.
- Zhou, H. and Lange, K. (2010). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19:645–665.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

APPENDIX

Appendix A

Supplementary Materials for Chapter 3

A.1 EM Algorithm for the Multivariate Response Model

In this section we review the derivation of the EM algorithm for the multivariate response model (Glanz and Carvalho, 2013; Reinsel, 1984). If the response matrix \mathbf{Y} can be written as the sum $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{Z}_1 + \cdots + \mathbf{Z}_m$ of independent random matrices with $\text{vec } \mathbf{Z}_i \sim N(\mathbf{0}, \mathbf{\Omega}_i)$, then $\text{vec } \mathbf{Y} \sim N(\text{vec}(\mathbf{X}\mathbf{B}), \mathbf{\Omega})$, where $\mathbf{\Omega} = \sum_{i=1}^m \mathbf{\Omega}_i$. Under the matrix normal assumption, $\mathbf{\Omega}_i = \mathbf{\Gamma}_i \otimes \mathbf{V}_i$. As in the text, the $p \times d$ coefficient matrix \mathbf{B} collects the fixed effects, the $\mathbf{\Gamma}_i$ are unknown $d \times d$ covariance matrices, and the \mathbf{V}_i are known $n \times n$ covariance matrices. The complete data log-likelihood for the unobserved \mathbf{Z}_i is

$$-\frac{1}{2} \sum_{i=1}^m \ln \det^+ \mathbf{\Omega}_i - \frac{1}{2} \sum_{i=1}^m \text{vec}(\mathbf{Z}_i)^T \mathbf{\Omega}_i^+ \text{vec}(\mathbf{Z}_i),$$

where $\det^+ \mathbf{\Omega}_i$ denotes the pseudo-determinant of $\mathbf{\Omega}_i$ and $\mathbf{\Omega}_i^+$ the pseudo-inverse of $\mathbf{\Omega}_i$. To compute the surrogate function for the EM algorithm, one needs the conditional expectations

$$E(\text{vec } \mathbf{Z}_i \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)}) = \mathbf{\Omega}_i^{(t)} \mathbf{\Omega}^{-(t)} \text{vec}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)}) = \mathbf{E}_i^{(t)}$$

and the conditional covariances

$$\text{Cov}(\text{vec } \mathbf{Z}_i \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)}) = \mathbf{\Omega}_i^{(t)} - \mathbf{\Omega}_i^{(t)} \mathbf{\Omega}^{-(t)} \mathbf{\Omega}_i^{(t)} = \mathbf{F}_i^{(t)},$$

where $\boldsymbol{\theta}$ is the parameter vector. These are employed to compute the conditional second moments

$$\mathbb{E}(\text{vec } \mathbf{Z}_i \text{ vec } \mathbf{Z}_i^T \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)}) = \mathbf{F}_i^{(t)} + \mathbf{E}_i^{(t)}(\mathbf{E}_i^{(t)})^T = \mathbf{G}_i^{(t)}.$$

Here the random vector \mathbf{Z}_i should be replaced by $\mathbf{Z}_m - \mathbf{X}\mathbf{B}^{(t)}$ when $i = m$.

One can readily check that $\boldsymbol{\Omega}_i^+ = \boldsymbol{\Gamma}_i^+ \otimes \mathbf{V}_i^+ = \boldsymbol{\Gamma}_i^{-1} \otimes \mathbf{V}_i^+$ for $\boldsymbol{\Gamma}_i$ invertible. Since the pseudo-determinant of a positive semidefinite matrix equals the product of its positive eigenvalues, the formulas

$$\begin{aligned} \det^+ \boldsymbol{\Omega}_i &= (\det \boldsymbol{\Gamma}_i)^{r_i} (\det^+ \mathbf{V}_i)^{s_i} \\ \ln \det^+ \boldsymbol{\Omega}_i &= r_i \ln \det \boldsymbol{\Gamma}_i + s_i \ln \det^+ \mathbf{V}_i \end{aligned}$$

apply, where $r_i = \text{rank}(\mathbf{V}_i^+)$ and $s_i = \text{rank}(\boldsymbol{\Gamma}_i^+)$. In the M step of the EM algorithm, one maximizes the surrogate

$$-\frac{1}{2} \sum_{i=1}^m r_i \ln \det \boldsymbol{\Gamma}_i - \frac{1}{2} \sum_{i=1}^m \text{tr}[(\boldsymbol{\Gamma}_i^{-1} \otimes \mathbf{V}_i^+) \mathbf{G}_i^{(t)}]. \quad (\text{A.1})$$

For $\boldsymbol{\Gamma}_i$ unstructured, we substitute $\boldsymbol{\Lambda}_i = \boldsymbol{\Gamma}_i^{-1}$ and maximize with respect to $\boldsymbol{\Lambda}_i$. Fortunately, the next lemma can be invoked.

Lemma 10. *If the matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} are $d \times d$, $n \times n$, and $dn \times dn$ respectively, then*

$$\text{tr}[(\mathbf{A} \otimes \mathbf{B})\mathbf{C}^T] = \text{tr}\{(\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{B}) \odot \mathbf{C}] (\mathbf{I}_d \otimes \mathbf{1}_n) \mathbf{A}^T\}.$$

Proof. This trace identity is essentially proved in the text. □

Lemma 10 yields

$$\text{tr}[(\boldsymbol{\Lambda}_i \otimes \mathbf{V}_i^+) \mathbf{G}_i^{(t)}] = \text{tr}\{(\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i^+) \odot \mathbf{G}_i^{(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n) \boldsymbol{\Lambda}_i\}.$$

The stationarity condition

$$\mathbf{0} = \frac{1}{2} r_i \boldsymbol{\Lambda}_i^{-1} - \frac{1}{2} (\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i^+) \odot \mathbf{G}_i^{(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n).$$

now entails the update

$$\boldsymbol{\Gamma}_i^{(t+1)} = \frac{1}{r_i} (\mathbf{I}_d \otimes \mathbf{1}_n)^T [(\mathbf{1}_d \mathbf{1}_d^T \otimes \mathbf{V}_i^+) \odot \mathbf{G}_i^{(t)}] (\mathbf{I}_d \otimes \mathbf{1}_n).$$

In the case $m = 1$, the single update reduces to

$$\mathbf{\Gamma}^{(t+1)} = \frac{1}{r}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)})^T \mathbf{V}^+ (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)}),$$

which matches the earlier result of Glanz and Carvalho (2013). When $\mathbf{\Gamma}_i$ is the scalar σ_i^2 ,

$$\begin{aligned} \mathbf{E}_i^{(t)} &= \sigma_i^{2(t)} \mathbf{V}_i \mathbf{\Omega}^{-(t)} (\mathbf{y} - \mathbf{X}\mathbf{\beta}^{(t)}) \\ \mathbf{F}_i^{(t)} &= \sigma_i^{2(t)} \mathbf{V}_i - \sigma_i^{2(t)} \mathbf{V}_i \mathbf{\Omega}^{-(t)} \sigma_i^{2(t)} \mathbf{V}_i. \end{aligned}$$

One recovers the representation (3.12) by substituting these quantities in equation (A.1) and invoking the identities $\mathbf{V}_i \mathbf{V}_i^+ \mathbf{V}_i = \mathbf{V}_i$ and $\text{tr}(\mathbf{V}_i \mathbf{V}_i^+) = \text{rank}(\mathbf{V}_i)$ and the cyclic permutation property of the trace.