MULTILEVEL MODEL SELECTION: A REGULARIZATION APPROACH INCORPORATING HEREDITY CONSTRAINTS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Elizabeth A. Stone
Diploma Date August, 2013

Examining Committee Members:

Dr. Alan J. Izenman, Advisory Chair, Department of Statistics

Dr. Richard M. Heiberger, Department of Statistics
Dr. Zhigen Zhao, Department of Statistics
Dr. Jeremy Mennis, External Member,
Department of Geography & Urban Studies, Temple University

ABSTRACT

MULTILEVEL MODEL SELECTION: A REGULARIZATION APPROACH INCORPORATING HEREDITY CONSTRAINTS

Elizabeth A. Stone

DOCTOR OF PHILOSOPHY

Temple University, August, 2013

Dr. Alan J. Izenman, Advisory Chair, Department of Statistics

This dissertation focuses on estimation and selection methods for a simple linear model with two levels of variation. This model provides a foundation for extensions to more levels. We propose new regularization criteria for model selection, subset selection, and variable selection in this context. Regularization is a penalized-estimation approach that shrinks the estimate and selects variables for structured data. This dissertation introduces a procedure (HM-ALASSO) that extends regularized multilevel-model estimation and selection to enforce principles of fixed heredity (e.g., including main effects when their interactions are included) and random heredity (e.g., including fixed effects when their random terms are included). The goals in developing this method were to create a procedure that provided reasonable estimates of all parameters, adhered to fixed and random heredity principles, resulted in a parsimonious model, was theoretically justifiable, and was able to be implemented and used in available software. The HM-ALASSO incorporates heredity-constrained selection directly into the estimation process. HM-ALASSO is shown to enjoy the properties of consistency, sparsity, and asymptotic normality. The ability of HM-ALASSO to produce quality estimates of the underlying parameters while adhering to heredity principles is demonstrated using simulated data. The performance of HM-ALASSO is illustrated using a subset of the High School and Beyond (HS&B) data set that includes math-achievement outcomes modeled via student- and school-level predictors. The HM-ALASSO framework is flexible enough that it can be adapted for various rule sets and parameterizations.

ACKNOWLEDGMENTS

Given the many years that it took for me to make this journey, I could never list all the people who helped me to complete it. First, I could not have done this without my advisor, Dr. Alan J. Izenman. I asked Dr. Izenman to work with me because of the passion he displayed for innovative statistical techniques in the advanced topics classes in which I was his student. Dr. Izenman has been a thoughtful, motivating, and generous collaborator throughout the years during which I worked on this research. His belief in me and in my work has been directly responsible for my successfully completing and defending this dissertation. My committee, Professors Richard Heiberger, Zhigen Zhao, and Jeremy Mennis were supportive and encouraging throughout, and I cannot thank them enough. Dr. Heiberger, in particular, went above and beyond to provide me with copious, constructive input that made this dissertation stronger. Other Temple professors were instrumental in educating me as a statistician: Burt Holland, Francis Hsuan, Damaraju Raghavarao, Sanat Sarkar, and William Wei. I am thankful to have been given the opportunity to pursue this degree under their guidance. I also thank Jagbir Singh and Louise Northington-Jones of the Statistics Department and Lisa Fitch of the Graduate Office. For their assistance, and for the assistance of others who worked behind the scenes, I am grateful. I could not have made it this far without my graduate school colleagues, past and present. I cherish those friendships deeply, and I think and hope that we will always have those ties.

I also learned a great deal from professors that I had at previous institutions. I would specifically like to thank Professors Carlos Alves, Andrew Clifford, and Guy Stroh, each of whom helped to prepare me for this work and made me strive beyond what I thought was possible. For the past 15 years, I have worked at Educational Testing Service. I have been fortunate to work alongside some of the most brilliant minds in the field of psychometrics, and I am even more fortunate to call many of these colleagues my friends. I truly thank them for their incredible support and accommodation along the way. In particular, Cara, Fred, Linda, and Mary have been there for me in the toughest times.

I must thank my family above all. They have made many sacrifices to get me through this process. My mother Marian, daughter Carolyn, husband Matt, and brother Patrick have been given me unconditional love and strength. I hope that this work also honors the memory of my father Patrick, cousin Gary, and grandmother Anna. I also have a large, loving, extended family, and I truly appreciate the kindness and understanding that they have shown me. This family also includes the friends with whom I have shared enough good times to get me through the difficult ones (especially E3 and the PHS crew). I cannot leave out my favorite furry trio, Paris, Nicolas, and Murphy, who have provided comfort and laughs along with the many distractions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The discussion of data sets with complex structure typically focuses on data sets with large numbers of predictors relative to the number of observations. Greater numbers of high-dimensional data sets have become available due to increased computational and storage capability. However, other complex-structured datasets have appeared due to advances in sampling and modeling. The particular data structures that motivated this work are models that arise from multilevel sampling. The type of multilevel data set on which this dissertation focuses is one in which, for example, students are nested within schools.

Data sets with cluster-based groupings have observations that may be dependent. For example, students within a school are expected to be more alike with respect to some predictor variables than are students in different schools.

This structure requires advanced modeling and analysis techniques. Reducing dimensionality through variable selection is, therefore, a key area of interest.

In this dissertation, the term *multilevel model* is used as an extension of the classical *single-level regression model*. A multilevel model is taken to be a model describing a structured data set with more than one level of variation. The terms *hierarchical linear model*, *random coefficients model*, and *linear mixed-effects model* are often used synonymously. The more general terminology will be used in this dissertation. This allows for extensions to cross-classified and multiple-membership models (Rasbash & Browne, 2007). The focus is on a Gaussian linear model with two levels of variation (e.g., students and schools).

Selection criteria developed for single-level models may require modification for multilevel regression. Multilevel data structures have dependencies between and within levels. Random elements make selection a more complicated issue.

Multilevel modeling is a thriving area of research (Browne & Steele, 2009). There is evidence of this popularity in its application in many fields. These fields include management (Gilbert & Shultz, 1998; Short *et al.*, 2006); law (Hoffman *et al.*, 2007); political analysis (Shor *et al.*, 2007); economics (Goodman & Thibodeau, 1998); health and medicine (Willems *et al.*, 2001; Merlo *et al.*, 2006; Goldstein & Kounali, 2009; Leyland & Næss, 2009); and other sciences (Hernández-Lloreda *et al.*, 2003; Wagner *et al.*, 2006). Most of these applications employ two-level models. Short *et al.* (2006) provides an example of a three-level model examining individuals within firms within industries, and Hoffman *et al.* (2007) includes dockets within cases within jurisdictions. Multilevel models have been used in applications from sports statistics (Albert, 2009) to genomics (Wolfinger *et al.*, 2001).

Multilevel modeling techniques are especially important in education research (Trautwein *et al.*, 2002; Leckie, 2009). Education data sets often have a natural hierarchical data structure. To motivate this discussion, consider the *High School and Beyond* (HS&B) data set (Zahs *et al.*, 1995). The data set includes information from a national education survey of high-school sophomores and seniors in the United States. The original data set contained more than 58,000 students in 1,015 schools. However, it is useful to focus on a popular subset of the data that is used frequently for illustrating multilevel modeling. The multilevel structure includes 7,185 students (level one) in 160 schools (level two) (Raudenbush & Bryk, 2002). The student-level outcome variable, mathematics achievement, is modeled using predictors at the student and school levels. Student-level predictors include measures of student socioeconomic status, minority status, and gender. School-level predictors include school enrollment, school type, disciplinary climate, and aggregate school-level measures of minority representation, socioeconomic status, and academic climate. The level-one predictors can also be considered to be random, allowing variation at the school level in addition to variation at the individual level. It is important to find a model that is optimal in some sense for relating math achievement to these predictors. This evaluation can depend on different perspectives using various criteria. The HS&B data set is used throughout the dissertation to illustrate various aspects of estimation and selection.

The popularity and complexity of multilevel regression underscore the importance of considering the following caution (George, 2000):

> With the availability of so many variable selection procedures and so many different justifications, it has becomes [*sic*] increasingly easy to be misled and to mislead. Faced with too many choices and too little guidance, practitioners continue to turn to the old standards such as stepwise selection based on AIC or minimum C_p, followed by a report of the conventional estimates and inferences.

The complexity introduced by including predictors associated with both levels of the model and, thus, interactions within or between levels, has been acknowledged by practitioners (Dedrick *et al.*, 2009). However, this complexity makes the concern expressed by George (2000) well-founded. As Kinney & Dunson (2008) have also noted with respect to variance components,

> For fixed effects models, there is a rich literature on methods for subset selection and inferences from both frequentist and Bayesian perspectives; however, subset selection for the random effects component has received limited attention.

This notion applies generally to selection for multilevel models.

We introduce an estimator (HM-ALASSO) that responds to the concerns of George (2000) and Kinney & Dunson (2008). The HM-ALASSO extends regularized multilevel-model estimation and selection to enforce principles of heredity. It does this with a regularization penalty that capitalizes on the hierarchical model structure. The principles of interest are strong versions of *fixed heredity* (e.g., including main effects when their interactions are included) and *random heredity* (e.g., including fixed effects when their random terms are included). This estimator can be applied by modifying the code of a regularization function that is currently available in `R`.

This work is organized as follows. Chapter 2 provides an overview of the multilevel model structure and estimation. The literature review includes frequentist and Bayesian selection guidelines and criteria. Commonly used selection criteria are illustrated using the HS&B data set. We provide a synthesis of regularization methods and selection approaches that have been applied for models with fixed and random effects. Chapter 3 details a new procedure, HM-ALASSO, that provides regularized estimation and selection of both fixed and random effects in accordance with fixed and random heredity rules. The Chapter describes the asymptotic properties of the estimator and illustrates the method in a simulation study and an application to the HS&B data. The goals in developing this method were to create a procedure that provided reasonable estimates of all parameters, adhered to fixed and random heredity principles, resulted in a parsimonious model, was theoretically justifiable, and was able to be implemented and used in available software (e.g., in `R`).

This last criterion has been cited as important for realizing the benefits of new procedures (Hesterberg *et al.*, 2008). The HM-ALASSO incorporates heredity-constrained selection directly into the estimation process. Chapter 4 summarizes directions for future research in this area. We discuss the need for optimization of the current algorithm and list several approaches to meet this need. We claim that heredity-focused selection criteria must be developed in order to select models with the desired structure.

CHAPTER 2

Literature Review

In this Chapter, we discuss the multilevel model structure. We present background on estimation and guidelines for model building. Commonly used frequentist and Bayesian selection criteria are described and are applied to the HS&B data. The Chapter closes with an overview of regularization literature and a specific focus on regularization procedures for models with multilevel data structures.

2.1 Multilevel Model Structure

The classical single-level regression model can be extended conceptually to two levels. This is accomplished by considering one or more components of the vector of regression parameters to have its own distribution (i.e., to be random rather than fixed). Multilevel models are defined in this dissertation to refer to models for data that have structures with at least two levels of variation. The numerous ways to display a multilevel model each have advantages in particular applications. It is useful to begin the discussion of the model structure using the descriptive two-level format (see, e.g., Raudenbush & Bryk, 2002, p. 35).

In the multilevel context with two levels of variation, we distinguish between the number of individuals, N, and the number of groups, J. These numbers represent, respectively, the sample sizes of the level-one (micro) units and the level-two (macro) units (see, e.g., Snijders & Bosker, 1999, p. 142). We also distinguish between the numbers of fixed effects (p) and random coefficients (q). We refer to the ith individual in the jth group, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. We reserve the symbol n to represent the sizes of the groups in a balanced data set (i.e., $n_j = n$ for all $j = 1, 2, \ldots, J$, so that $N = nJ$). Due to the extensive notation used here and in the relevant literature, we present the model within the following framework.

To fix ideas, consider a simple model with one level-one predictor X_1 (with values X_{1ij}). Let X_1 be assumed to have both an associated random intercept and an associated random slope. We add a group-level covariate X_2 (with values X_{2ij}) at level two. A group descriptor, X_{2ij}, will be common to all individuals in the jth group. Let Y be the dependent variable, with Y_{ij} the outcome for the ith individual in the jth group. For the jth group, this model can be represented as

$$Y_{ij} = A_{1j} + A_{2j}X_{1ij} + R_{ij}, \tag{2.1}$$

where

$$A_{1j} = \beta_1 + \beta_2 X_{2ij} + B_{1j} \text{ and } A_{2j} = \beta_3 + \beta_4 X_{2ij} + B_{2j}, \tag{2.2}$$

and $R_{ij} \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. The random-effects vector $\mathbf{B}_j = (B_{1j}, B_{2j})'$, contains the level-two errors, and $\mathbf{B}_j \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$. The random-effects covariance matrix $\mathbf{\Sigma}$ is assumed to be the same for all J groups.

In this model, A_{1j} and A_{2j} are the random-intercept coefficient and the random-slope coefficient for the regression of the random variable Y on X_1, respectively. The vector of fixed effects is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$. We can combine the levels into one model equation that displays the main effects and cross-level interactions,

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{1ij} + \beta_4 X_{2ij} X_{1ij} + B_{1j} + B_{2j} X_{1ij} + R_{ij}, \tag{2.3}$$

where we note that there are predictors associated with fixed effects ($\boldsymbol{\beta}$) and random effects (B). The random-effects predictors are a subset of the fixed-effects predictors. To clarify the distinction between the fixed and random parts of the model, we rename the random predictors from X to Z. We do this in the example model by replacing X_{1ij} with Z_{1ij} in the random term.

In general, for the ith individual in the jth group, let Y be the random variable representing the outcome. Let $X_1, X_2, \ldots, X_p$ be the regressors for the fixed effects. Let $Z_1, Z_2, \ldots, Z_q$ be the regressors for the jth group's level-one predictors that are considered random. The general form of the model (2.3) can be written as

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots + \beta_p X_{pij} + B_{1j} Z_{1ij} + B_{2j} Z_{2ij} + \ldots + B_{qj} Z_{qij} + R_{ij}, \quad (2.4)$$

$i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. Note that X_{1ij} is often taken to be the constant 1 to represent the intercept. If the intercept is random, then Z_{1ij} will also be 1. Fixed-effects terms are reindexed in this compressed representation. Cross-level interactions will no longer be displayed as products of other terms. It will be convenient to express the various models using matrix representation. For the jth group, let $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, \ldots, Y_{n_j j})'$ be the response vector. Let $\mathbf{X}_j$ be the $(n_j \times p)$ design matrix of fixed-effects covariates. Let $\mathbf{Z}_j$ be the $(n_j \times q)$ design matrix for the random-effects covariates. The model for the jth group can be represented as

$$\overset{n_j\times 1}{\mathbf{Y}_j} = \overset{n_j\times p}{\mathbf{X}_j}\ \overset{p\times 1}{\boldsymbol{\beta}} + \overset{n_j\times q}{\mathbf{Z}_j}\ \overset{q\times 1}{\mathbf{B}_j} + \overset{n_j\times 1}{\boldsymbol{R}_j}, \quad (2.5)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-vector of unknown fixed effects (common to all individuals in all groups). The q-vector of random effects is denoted by $\mathbf{B}_j = (B_{1j}, \ldots, B_{qj})'$ (common to all individuals in the jth group), with $\mathbf{B}_j \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$. An n_j-vector of level-one errors is given by $\mathbf{R}_j$ where we assume that $\mathbf{R}_j \sim \mathcal{N}_{n_j}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_j})$. The covariance structures in this model can be generalized. In moving from (2.4) to (2.5), predictors with β coefficients are in the $\mathbf{X}_j$ matrix. Predictors with B coefficients are in the $\mathbf{Z}_j$ matrix.

The positive-definite covariance matrix, $\boldsymbol{\Sigma}$, is often represented by $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. This representation shows that the covariance matrix is dependent on an underlying parameter vector $\boldsymbol{\theta}$. The underlying parameter vector varies according to the parameterization of the model used for estimation. For example, the matrix $\boldsymbol{\Sigma}$ could be factored using a Cholesky decomposition to make estimation of the underlying parameters more efficient. In that case, we could take $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}'_{\boldsymbol{\theta}}$. In that decomposition, σ^2 is the level-one error variance. The matrix $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ is a lower-triangular, relative-precision factor dependent upon the underlying parameter vector $\boldsymbol{\theta}$ (see, e.g., Bates & Pinheiro, 1998). The diagonal of $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ must contain only non-negative entries. The particular set of parameters to be estimated would be different than if a modified-Cholesky decomposition were performed. One particular modified-Cholesky decomposition yields $\boldsymbol{\Sigma} = \sigma^2 \mathbf{D} \boldsymbol{\Gamma} \boldsymbol{\Gamma}' \mathbf{D}$. In this decomposition, $\mathbf{D}$ is a non-negative diagonal matrix, and $\boldsymbol{\Gamma}$ is a lower-triangular matrix (see, e.g., Chen & Dunson, 2003). The modified-Cholesky decomposition separates out dependencies and is useful for random-effects selection.

It is assumed that $\mathbf{B}$ is independent of $\mathbf{R}$ within each group and that each is independently distributed across groups. The model parameters are the fixed-effects parameters ($\boldsymbol{\beta}$), the random-effects covariance matrix $\boldsymbol{\Sigma}$ parameters ($\boldsymbol{\theta}$), and the level-one error variance (σ^2). Random effects are unobserved random variables with values referred to as "conditional modes" rather than as estimates. The random effects are level-two residuals that are analogous to the level-one residuals. They are computed based on the values of the parameter estimates and are not directly estimated.

We can combine the model (2.5) across groups. Let $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_J')'$ be an N-vector of responses. Let $\mathbf{X}$ be an $(N \times p)$ fixed-effects design matrix, and let $\mathbf{Z}$ be an $(N \times qJ)$ random-effects design matrix. We take $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ to be known. In this structure, $\mathbf{Z}$ is a block-diagonal matrix with the group-level $\mathbf{Z}_j$ matrices on the diagonal, $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_J')'$, and $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_J']'$, and $\tilde{\boldsymbol{\Sigma}}$ is a block-diagonal matrix with the $\boldsymbol{\Sigma}$ matrix repeated on the diagonal.

Then, the multilevel model is given by

$$\overset{N\times 1}{\mathbf{Y}} = \overset{N\times p}{\mathbf{X}}\,\overset{p\times 1}{\boldsymbol{\beta}} + \overset{N\times qJ}{\mathbf{Z}}\,\overset{qJ\times 1}{\mathbf{B}} + \overset{N\times 1}{\boldsymbol{R}}, \tag{2.6}$$

where $\boldsymbol{\beta}$ is a p-vector of unknown fixed-effects parameters, $\mathbf{B} \sim \mathcal{N}_{qJ}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$,

Table 2.1: Predictors Included in the HS&B Data Set.

Level Number Level Name	Predictor	Description
1 Individual	`minority`	student minority status (1=minority, 0=other)
	`ses`	student socioeconomic status
	`female`	gender (1=female, 0=male)
2 School	`size`	school enrollment
	`sector`	school type (1=Catholic, 0=public)
	`disclim`	school disciplinary climate
	`himinty`	students with minority status (1= more than 40% minority, 0=less than 40% minority)
	`meanses`	mean socioeconomic status for school
	`pracad`	proportion of students on the academic track

and $\mathbf{R} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\mathbf{I}_N)$.

Illustrative example. Consider again the HS&B data set. This major effort by the National Center for Education Statistics was designed to track achievement, choices, and demographic characteristics throughout a student's school years. A subset of the much larger data set is used extensively to illustrate multilevel modeling. The individual- and school-level data sets are packaged with the `nlme` package in `R` as the `MathAchieve` and `MathAchSchool` data, respectively. The student-level outcome variable, mathematics achievement (`mathach`), can be modeled using the predictor variables in Table 2.1.

The predictors `ses`, `sector`, and `meanses` are often of interest in developing a model for `mathach`. We can use the group-level predictors to model the random coefficients for the intercept and `ses`. The full model involving these three predictors can be written as three regression equations involving the

ith student in the jth school. The model includes one level-one equation for the outcome and two level-two equations that model the random regression coefficients. This two-level model can be written as

$$Y_{ij} = A_{1j} + A_{2j}(\texttt{ses}_{ij}) + R_{ij}, \tag{2.7}$$

where

$$A_{1j} = \beta_1 + \beta_2(\texttt{sector}_{ij}) + \beta_3(\texttt{meanses}_{ij}) + B_{1j}, \tag{2.8}$$

$$A_{2j} = \beta_4 + \beta_5(\texttt{sector}_{ij}) + \beta_6(\texttt{meanses}_{ij}) + B_{2j}, \tag{2.9}$$

and $R_{ij} \sim \mathcal{N}(0, \sigma^2)$, with $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, 160$, $N = 7,185$. This particular model has six fixed effects: parameters for the intercept, `ses`, `sector`, and `meanses`, and the interactions of `ses` with `sector` and `meanses`. The model includes four variance components: two random-effects variances, the covariance between the random effects, and the level-one error variance.

Figure 2.1 shows the grouping structure of the HS&B data set. For a random sample of 10 Catholic schools (group names starting with C) and 10 public schools (with group names starting with P), it shows the relationship of a student's math achievement to their grand-mean-centered `ses`. Plots are ordered from lower left to upper right by maximum `ses` in each school. Note that there does not appear to be a strong effect of `ses` by school or by school type.

The model (2.6) can be extended to more than two levels of variation or can be formulated for longitudinal data. This work focuses on the basic two-level cross-sectional model. The motivation for using a multilevel model over a single-level model for structured data is due to the information available for estimation. Observations within a group may be dependent. This dependence violates an assumption required for ordinary-least-squares (OLS) regression. Consequently, alternative parameter-estimation methods must be employed. Multilevel-model estimation methods capitalize on the information available both within and between groups.

The *intraclass correlation* (ICC) measures how much variation in the outcome variable is explained by clustering. There are many definitions of the ICC (Shrout & Fleiss, 1979). The formula that is usually applied in multilevel regression to estimate the ICC is the ratio of the estimate of the between-group variance to an estimate of the total variance. We use the popular definition,

$$ICC = \frac{\widehat{\Sigma}_{11}}{\widehat{\Sigma}_{11} + \widehat{\sigma}^2} \in [0,1], \tag{2.10}$$

where $\widehat{\Sigma}_{11}$ and $\widehat{\sigma}^2$ are estimates of the between-group and within-group variances, respectively. Larger ICC values indicate a greater need to take into account the multilevel structure of the data. What constitutes "large" ICC

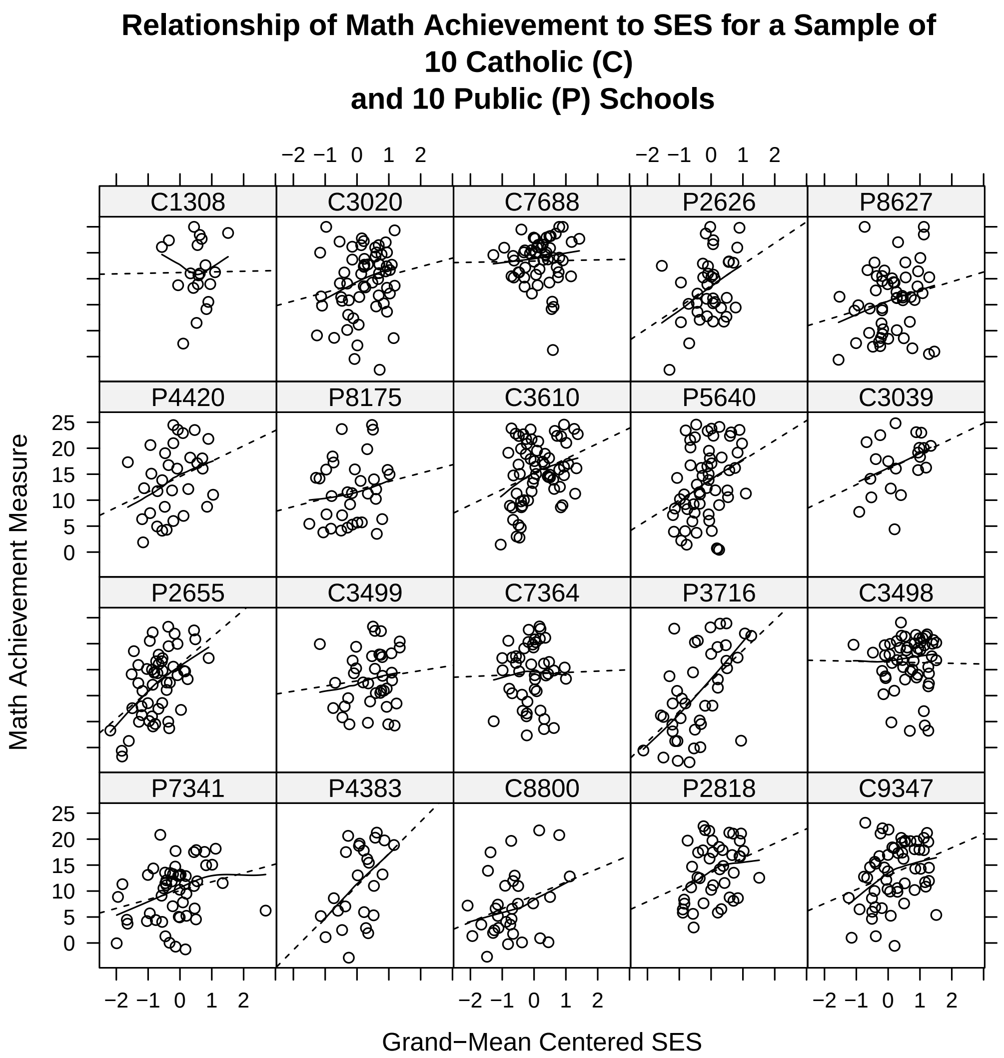

Figure 2.1: Math Achievement versus SES for 10 Catholic Schools (C) and 10 Public Schools (P) in the HS&B Data Set

values varies by context. For social-science data, ICC typically ranges from 0.00 to 0.40 (Snijders & Bosker, 1999, p. 151). ICCs for other types of data (e.g., health care) can exceed 0.50 (Rowe *et al.*, 2002). For the two-level model, the ICC is typically calculated using the variance estimates from a variance-components model. The variance-components model is alternatively known as a one-way random-effects analysis of variance (ANOVA) model. The ANOVA model contains no predictors at either level and can be represented as

$$Y_{ij} = A_{1j} + R_{ij}, \quad \text{where} \quad A_{1j} = \beta_1 + B_{1j}, \tag{2.11}$$

$i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. For balanced groups of size n, the effective sample size can be calculated as

$$N_{\text{eff}} = \frac{N}{1 + (n-1)ICC}. \tag{2.12}$$

This reduced sample size represents how many independent observations the structured sample would correspond to in terms of information available for statistical inference (see, e.g., Snijders & Bosker, 1999, p. 23). The clustering effect is important, because sample size is often used in selection criteria in a way analogous to that of a single-level model. Therefore, the question of which sample size to use in multilevel selection criteria is an important issue.

Applying single-level-regression selection approaches could yield inaccurate results. Standard errors are typically underestimated, leading to inflated Type I error and possible inclusion of falsely-significant terms into the model (Barcikowski, 1981; Hox, 1998).

2.2 Multilevel Model Estimation

For models with a multilevel-data structure, fixed-effect and random-effect covariance-matrix estimates do not have closed-form expressions. Estimation must be achieved by using an iterative procedure. Full maximum likelihood (FML; Fisher, 1922) and restricted (or residual) maximum likelihood (REML; Patterson & Thompson, 1971) are commonly used in multilevel-model estimation. Note that we can express the probability model for the jth group as

$$\mathbf{Y}_j \sim N_{n_j}(\mathbf{X}_j\boldsymbol{\beta}, \mathbf{V}_j(\boldsymbol{\theta}, \sigma^2)), \tag{2.13}$$

with

$$\mathbf{V}_j(\boldsymbol{\theta}, \sigma^2) = \mathbf{Z}_j\boldsymbol{\Sigma}_j\mathbf{Z}_j' + \sigma^2\mathbf{I}_{n_j}. \tag{2.14}$$

Letting $\mathbf{V} = \{\mathbf{V}_j\} \otimes \mathbf{I}_J$ (with $\mathbf{A} \otimes \mathbf{B}$ denoting the direct product of $\mathbf{A}$ and $\mathbf{B}$), the FML log-likelihood can be expressed as

$$l = -\frac{1}{2}\left\{\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{2.15}$$

where $\mathbf{y}$ is the observed value of $\mathbf{Y}$, and $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$ (see, e.g., Longford, 1993, p. 106). FML estimates all parameters (fixed effects and variance components) simultaneously. However, the estimated covariance-matrix parameters may be biased downward. The denominator of the FML estimator is not adjusted for the degrees of freedom (df) lost due to estimation of the fixed effects. This difference in the resulting estimates may be minor. However, the disparity increases with the number of predictors p in the model. Large bias would have an influence on hypothesis testing.

REML adjusts for the estimation of the fixed effects. Conceptually, REML does not maximize the likelihood based on the outcome vector directly. Instead, the method is based on a type of linear combination that is referred to as an *error contrast*. The term was used to describe estimation of the log-likelihood in two parts for an incomplete block design (Patterson & Thompson, 1971). The method involved selection of a matrix $\mathbf{S}$ of rank $N - p$ such that $\mathbf{SX} = 0$ and $E(\mathbf{SY}) = 0$, with $E(\cdot)$ denoting the expected value over the outcome vector. The matrix $\mathbf{S} = \mathbf{I}_N - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is an example of a contrast matrix with these characteristics. Based on the work of Patterson and Thompson, Harville (1974) developed a useful expression of the REML log-likelihood,

$$l = C - \frac{1}{2}\left\{\log|\mathbf{V}| + \log\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \qquad (2.16)$$

with the constant term C based on the particular set of contrasts that are used. Estimation of the variance components requires maximization of this function only over the error contrasts, which do not involve the regression parameters. The regression parameters can be estimated separately in a generalized-least-squares (GLS) step using the estimated covariance matrix. REML produces an unbiased estimator of the level-one error variance and reduced bias in the covariance parameters. Therefore, REML is more often used than FML to estimate the final covariance parameters. However, REML covariance-parameter estimates are not necessarily unbiased, particularly in the case of an unbalanced design (Rabe-Hesketh & Skrondal, 2008, pp. 185).

The FML and REML estimators yield nearly identical estimates of the fixed-effects parameters. The FML and REML estimates of the variance components are asymptotically equivalent. The asymptotic equivalence occurs in the sense that their distributions converge as sample sizes tend to infinity when considering a fixed number of parameters. As the number of parameters grows, REML is preferred due to the reduced bias in the variance-component estimates. However, a large sample can offset the differences between the procedures. These estimation methods can be used together in the model-building process to focus on different components of the model (see, e.g., Wolfinger, 1993). Both FML and REML are implemented through iterative procedures.

See Hartley & Rao (1967), Harville (1977), and (Raudenbush & Bryk, 2002, pp. 42–45) for details of the distributional properties of FML estimators for these models. The terms GLS and FML or ML are often used interchangeably in multilevel literature, particularly under the multivariate normality assumptions used for the model of interest. Goldstein (1986) showed that the iterative GLS procedure (IGLS) converges to produce estimates equivalent to those from FML. An analogous procedure, restricted unbiased iterative GLS (RIGLS), produces estimates that are equivalent to REML estimates (Goldstein, 1989).

The frequentist and Bayesian approaches to estimation follow different philosophies. The focus of this dissertation is on frequentist estimation. However, we briefly review Bayesian estimation because of its extensive use in multilevel-model estimation and inference. In frequentist estimation based on the log-likelihood (2.15), one attempts to determine parameter estimates that maximize the likelihood of observing those particular data points. From a Bayesian standpoint, all parameters and observed quantities are considered jointly. Prior distributions $p(\theta)$ are placed on the parameters to be estimated. Bayesian analysis is developed through the use of Bayes's rule,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})},$$

in which $p(\theta)$ is the prior probability distribution of the parameter of interest θ,

$p(\mathbf{y}|\theta)$ is the likelihood of observing the response vector $\mathbf{y}$ given the parameter, and $p(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$. The result, $p(\theta|\mathbf{y})$, is the posterior distribution of the parameter θ given the data $\mathbf{y}$. The relation between prior and posterior is often expressed as $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$. This representation conveys that the posterior distribution of a parameter is an updated version of the prior once the data have been observed.

Bayesian parameter estimation involves finding the posterior distribution of a parameter, from which various estimates can be derived. This evaluation is often too difficult (or too time- or computationally-intensive) to carry out analytically or numerically. Markov chain Monte Carlo (MCMC) simulations can be used to sample from a distribution that approximates the marginal posterior distribution of the parameter of interest. When more than one parameter is being estimated, MCMC is used to approximate the joint posterior distribution of the parameters. The mean of a parameter's posterior distribution is then found by computing the sample mean of these draws. Other summary measures can be found analogously. The sampling mechanism is constructed such that a Markov chain of updated parameter estimates converges to the joint posterior distribution.

MCMC estimation algorithms start with initial values (e.g., sample estimates) and iteratively update parameters until convergence. Chains of simulated parameters use different starting points for robustness. Because of the initial instability in the updating, a number of early *burn-in* steps are discarded. Chain *thinning*, in which not all iterations are recorded, can be performed to reduce the storage required and the autocorrelation among samples.

MCMC estimation is a popular class of Bayesian estimation algorithms. Excellent references on MCMC estimation include Gelman *et al.* (2003, Chapter 11), Gilks *et al.* (1996), and Gelman and Hill (2007, Chapter 18). Gelman & Hill (2007) provides examples of Bayesian estimation for many types of multilevel models. Browne (1998, 2005) are useful guides with respect to the consideration of priors. With large numbers of models under consideration, MCMC estimation may be impractical due to computational and storage requirements. However, for small samples, Bayesian modeling provides the ability to allow for uncertainty (Gelman & Hill, 2007, p. 262).

2.3 General Guidelines for Model Building

A multilevel model typically includes many parameters. For example, even though the subset of predictors represented by Table 2.1 is small, it would yield a model with 39 parameters. The parameters would include 28 parameters for the fixed-effects, 10 random-effects variance and covariance parameters, and one level-one variance parameter. Clearly, it would be worthwhile to reduce the number of parameters by judicious selection of predictors.

General guidelines commonly associated with single-level regression often still hold when there are multiple levels in the regression model. In multilevel models, these guidelines must be altered or expanded to take the more complex structure into account.

We have to be concerned with the possibility of overfit when sample sizes cannot support estimating so many parameters. Incorporation of more predictors decreases the bias but increases the variances of the estimates. This situation is often referred to as the *curse of dimensionality*.

As a basic starting point for multilevel-model building, Gelman (2008), in discussing Bartels (2008), listed several guidelines. Intercepts and slopes should be allowed to vary, and group averages of level-one predictors should be included at level two. Bartels also recommended using group-mean centering for level-one predictors. However, Gelman noted that this may not make sense when, for example, the level-one predictor is an indicator variable (e.g., of race).

The principle of *heredity* for single-level models states that main effects should be included whenever their interactions are included. Similar heredity rules have been suggested for multilevel models (see, e.g., Snijders & Bosker, 1999, p. 92; Yuan *et al.* 2009). In describing the relationship between terms, lower-order terms in the hierarchy (controlling) are referred to as "parents." Higher order terms (controlled) are referred to as "descendants" or "children."

The principle of including interactions only in the presence of their main effects is an example of a *fixed heredity* rule. Another fixed-heredity rule states that the random-intercept model should include all predictors used to model the random slope.

Strong heredity, in which all parent terms must be in the model, is related to the principle of *marginality.* The core marginality argument does not allow the inclusion of higher-order terms without their lower-order counterparts. If the hierarchical structure is not respected, the outcome variable is no longer invariant to scale translation. Inclusion of an interaction without both parents also implies that the point at which the parent terms are 0 is a special scale point. This may be appropriate in some contexts but not in general (see, e.g., McCullagh & Nelder, 1989, pp. 69–70). A weak version of heredity requires only that at least one parent term (e.g., one main effect that contributes to an interaction) be in the model.

When a level-one predictor has a random slope, it is suggested that the predictor's fixed effect be included. This is an example of a *random heredity* rule. Including level-one and level-two covariates produces cross-level interactions for which the main effect must be present under this principle. It has also been suggested that a model should not have random slopes if it does not have a random intercept. However, Gelman and Hill (2007, p. 283) provide a counterexample: a series of independent experiments, with the control constant among the experiments and the treatment varying. Clearly, it is important to know the underlying theory, data structure, and research design when applying these rules.

2.4 Hypothesis Testing Guidelines

Hypothesis testing is a popular way to evaluate the importance of fixed and random effects. It often plays an important role in variable and model selection.

The use of p-values for significance testing in multilevel models has been criticized due to problematic df approximation for unbalanced data sets. The quantities under consideration also do not have exact t or F distributions (see, e.g., Bates, 2006). One suggested approach is to obtain and evaluate p-values based on MCMC sampling from the posterior distribution of the parameter vector (Baayen *et al.*, 2008). Statistical power for hypothesis testing depends on the amount of information at the level of the model term being tested. It is generally better from an estimation standpoint to have more groups, with fewer individuals per group, if total sample size is held constant (Snijders, 2005).

Single fixed effects can be evaluated with a Wald test (Wald, 1943) using a t-ratio of the estimate to its estimated standard error. For multilevel models, the df used for the test depend upon which level the predictor belongs to in the model (Snijders & Bosker, 1999, p. 86). Fixed effects associated with levels of a categorical variable (represented by dummy variables) should be tested simultaneously. See Raudenbush and Bryk (pp. 58–59) for an example of a test of the importance of the predictor `sector` in the HS&B data set. Likelihood-ratio tests (LRTs) can be used to evaluate the regression parameters. Likelihood-ratio-test (LRT) statistics are calculated as $2[l_2 - l_1]$, where l_2 is the log-likelihood of the more general model and l_1 is the log-likelihood of the more specific model nested within that model. This statistic is asymptotically distributed as $\chi^2_{k_2-k_1}$ $(k_2 > k_1)$, where k_2 and k_1 are the respective df of the models (Pinheiro & Bates, 2000, p. 83).

Single-parameter tests and LRTs are also used to evaluate the random structure of the model. The statistic $z = \widehat{\Sigma}_{11}/[\mathrm{Var}(\widehat{\Sigma}_{11})]^{1/2}$, where the estimated standard error is $\mathrm{Var}(\widehat{\Sigma}_{11})$ from the inverse information matrix, can test $H_0 : \Sigma_{11} = 0$ with an asymptotically Gaussian distribution. Approximate normality may not be achieved when the true variance component is close to zero (Raudenbush & Bryk, 2002, p. 64). Therefore, comparisons of the fits of nested models are recommended to ascertain whether an additional term makes a significant contribution to the model.

Variances must be non-negative. Therefore, a test that assumes that a random component's variance is 0 places the null hypothesis on the boundary of the parameter space. LRTs for variance parameters should use a halved

p-value due to the one-sided alternative hypothesis that results from this constraint (Snijders & Bosker, 1999, p. 90). The random intercept can also be tested, controlling for the predictors, using the F-test from an ANCOVA that includes the predictors as covariates (Snijders & Bosker, 1999, p. 91).

A random-slope variance that is estimated using an iterative FML or REML procedure can be tested by estimating the parameters of a model without the random slope and, then, using those parameter estimates as the initial estimates in an updated model that includes the random slope. The significance of the random slope can be tested by comparing the ratio of the estimated random-slope variance to its estimated standard error to a standard Gaussian distribution (Snijders & Bosker, 1999, p. 123). See Berkhof & Snijders (2001) for a simulation study comparing several single- and multiple-parameter approaches to evaluating variance components. See Crainiceanu (2008) for a discussion of using LRTs with variance components.

2.5 Model Selection Guidelines

Models estimated with REML can only be compared, using likelihood-based criteria, in terms of their variance components (Singer & Willett, 2003, p. 118). In other words, models estimated with REML can only be compared if they have the same fixed effects. This warning is repeated throughout the literature. Inappropriate model comparisons arise when attention is not paid to the estimation method being used (see, e.g., Pinheiro & Bates, 2000, p. 18; Snijders & Bosker, 1999, p. 89).

Adding parameters tends to increase goodness of fit. Penalized-likelihood criteria incorporate model dimensionality by minimizing

$$-2l + \lambda d, \tag{2.17}$$

where l is the model log-likelihood, λ is a penalty parameter, and d represents model dimension. For multilevel models, the dimension d usually refers to the number of estimated parameters for the model, including fixed effects and variance components. The penalty coefficient, λ, determines which criterion is being applied. Akaike's criterion AIC uses a λ of approximately 2. The BIC of Schwarz includes the sample size in the form of $\lambda = \log N$. These criteria are sometimes described in terms of model deviance (McCullagh & Nelder, 1989, p. 33). The model deviance formulation reduces to the representation (2.17) when comparing models. The full model from which the given model's deviance is computed will be common to the models being compared.

A criterion used in Bayesian multilevel modeling is the *deviance information criterion* (DIC) (Spiegelhalter *et al.*, 2002). The DIC is often referred to as the "Bayesian analogue" of AIC. The DIC also combines model fit and complexity into one measure,

$$\textsc{dic} = \text{mean deviance} + 2p_D, \tag{2.18}$$

where p_D is the effective number of parameters (Gelman & Hill, 2007, p. 525). Challenges that arise when using DIC are similar to those found in the frequentist context. These include the question of how to count parameters for the dimensionality penalty. Rather than compute the number of parameters directly, p_D is often computed from simulation output in two different ways. One method counts the difference between the posterior mean deviance and the deviance of posterior means ($p_D = \overline{D(\theta)} - D(\bar{\theta})$, where $\overline{D(\theta)}$ is the posterior mean deviance and $D(\bar{\theta})$ is the deviance of the posterior means). The other method counts df as half the variance of the posterior deviance (offered as an alternative parameter-counting method in Gelman *et al.*, 2004, p. 182). The latter's use has been debated (Gelman, 2006b). The former, analogous to the hat-matrix parameter count used for single-level models, appears to be the current gold standard. This popularity has arisen despite the method's unstable estimation of p_D (Gelman & Hill, 2007, p. 526).

These guidelines represent some of the goals of building a "good" multilevel model. It is important to perform model checking and diagnostics before selecting the final model. Estimation is fairly robust to minor violations of the underlying assumptions of homogeneity of variance and normality. However, inaccurate standard errors may result for the random-effect covariance parameters (Maas & Hox, 2004). Violations can lead to bias in the estimates or underestimates of their standard errors. More general multilevel models will be more appropriate in those cases (Raudenbush & Bryk, 2002, p. 253).

It is important to specify the model correctly and to take into account the computational time and storage required to carry out simulations on large data sets. Detailed descriptions of these types of models and ways to check for convergence and improve computational efficiency are given in Gelman and Hill (2007, Chapters 17, 19). These general models are illustrated throughout that reference using the *radon* data set, which can be used to model the radon level for an individual home based on house- and county-level information. See Gelman (2006a) for a concise discussion and application with these data.

For additional guidance on multilevel model building and checking, see Longford (1993, Section 4.8), Hox (2002, Chapter 3), Raudenbush and Bryk (2002, Chapter 9), Snijders & Berkhof (2007), Gelman and Hill (2007, Chapter 24), and Ntzoufras (2009, Section 10.3).

See Stone & Izenman (2012) for a more comprehensive review of model-building guidelines and selection criteria that have been offered for multilevel models.

2.5.1 Example: High School and Beyond Data

We return to the HS&B data described previously. Predictors in this data set are listed in Table 2.1. The ICC for this example is 0.18 (Raudenbush & Bryk, 2002, p. 71). This ICC value is mid-range for social-science data in general. However, the ICC falls into the high end of the range for clustered standardized test score data (Schochet, 2008).

We follow the convention of centering `ses` around its school mean `meanses` (i.e., group-mean centering) (Gelman, 2008). We focus only on the predictors `ses`, `meanses`, and `sector` to keep the illustration of model comparison clear. The model involving these three predictors is

$$\begin{aligned} Y_{ij} =& \beta_1 + \beta_2(\texttt{sector}_{ij}) + \beta_3(\texttt{meanses}_{ij}) + \beta_4(\texttt{ses}_{ij}) \\ &+ \beta_5(\texttt{sector}_{ij})(\texttt{ses}_{ij}) + \beta_6(\texttt{meanses}_{ij})(\texttt{ses}_{ij}) \\ &+ B_{1j} + B_{2j}(\texttt{ses}_{ij}) + R_{ij}, \end{aligned} \tag{2.19}$$

with all elements as previously defined.

Several versions of this model were run using frequentist and Bayesian estimation and selection criteria. The frequentist models used FML so that model criteria could be compared directly.

In Table 2.2, we include the AIC, BIC, and DIC for the models derived from (2.19). Smaller relative values are desirable. The magnitude of the criterion

Table 2.2: Criterion Values of Models for the HS&B Data Set.[a]

Predictors (Level)	Random	AIC[b]	BIC[b]	DIC[c]
SES(1) [d]	N/A	47880	47801	47786
N/A[e]	Int[f]	47122	47142	46909
SES(1)	Int	46728	46756	46508
Sec[g](2)	Int	47087	47115	46914
MSES[h](2)	Int	46967	46995	46909
Sec(2), MSES(2)	Int	46954	46988	46918
SES(1), Sec(2), SES×Sec	Int	46694	46728	46507
SES(1), MSES(2), SES×MSES	Int	46574	46608	46512
SES(1), Sec(2), MSES(2), SES×Sec, SES×MSES	Int	46560	46602	46512
SES(1)	Int, SES	46723	46764	**46471**
SES(1), Sec(2), SES×Sec	Int, SES	46650	46705	**46471**
SES(1), MSES(2), SES×MSES	Int, SES	46568	46623	46643
SES(1), Sec(2), MSES(2), SES×Sec, SES×MSES	Int, SES	**46516**	**46585**	46476

[a] Boldface values indicate the model chosen using each criterion.
[b] Values are from the `lme4` package in R using FML.
[c] Values are from the `arm` package in R.
[d] OLS model: note that there is no random term in this model (N/A= Not Applicable)
[e] Variance components model: note that there are no predictors in this model (N/A= Not Applicable)
[f] Int refers to the intercept.
[g] The `Sector` variable is abbreviated as `Sec`.
[h] The `MeanSES` variable is abbreviated as `MSES`.

is important only as it compares to criterion values of other models. Boldface values in Table 2.2 indicate the model chosen using each criterion.

Simply examining these criteria will not necessarily lead to the best model. The model chosen by AIC and BIC was explored to determine whether to include the random slope for `ses`. Raudenbush and Bryk (2002, pp. 84–85) used an LRT to determine whether the additional variance and covariance components induced by the larger model added significant explanatory power of the model and found that they did not. This may indicate that the smaller model is adequate. Simulation results show that these criteria, in particular, do not necessarily lead to the selection of optimal multilevel models in a random-effects selection context (McMurray, 2010).

Examining differences in predictive ability between models provides evidence for which components may be important to the model (Breiman, 2001a). For example, we could be interested in predicting math achievement for a new student in one of the 160 schools. A comparison of OLS, prior only, and multilevel prediction rules was undertaken (Afshartous & de Leeuw, 2005a,b). The analysis found that multilevel rules had the lowest predictive mean-squared error (PMSE) estimates for multilevel data sets.

The multilevel cross-validation method in Table 2.3 was used to obtain a leave-one-out cross-validation (CV) estimate of the PMSE (Afshartous & de Leeuw, 2004). The full data set was divided into training and test data sets within school. One student from each school was set aside as a test case for each iteration. The data from the remaining students were used to train the model. This process was repeated for each model 100 times. The bootstrap (BS) PMSE estimate (see Table 2.4) was also calculated (Afshartous & de Leeuw, 2004). In that method, bootstrap samples the same size as the original data set are drawn with replacement (students sampled within school). These bootstrap samples are used for training a model that is then applied to the original data set.

The bootstrap method may lead to estimation problems if there are groups containing few individuals. If sampling leads to the same individual being

Table 2.3: Leave-One-Out Cross-validation PMSE (CV1 PMSE) Method.

1.	Randomly select one individual in each of the J groups, forming a test group of size J.
2.	Use the remaining N-J individuals (the training group) to estimate the model.
3.	Calculate predicted outcomes, and find the average of the sum of squared errors based on these predictions, for the J test group individuals.
4.	Repeat steps 1-3 until the desired number of iterations is complete.
5.	Calculate the mean of the average error sum of squares over those iterations. This average is the CV PMSE estimate.

Table 2.4: Bootstrap PMSE (BS PMSE) Method.

1.	Keep a copy of the entire original data set aside to be used as test cases.
2.	In each of the J groups, sample n_j individuals with replacement. This new set of observations, of size N, will be used as the training cases.
3.	Estimate the model using the training cases (bootstrap sample).
4.	Apply the estimated model coefficients to the testing cases (original data) to form predicted outcomes, and find the average sum of squared errors based on these predictions.
5.	Repeat steps 2-4 until the desired number of iterations is completed.
6.	Calculate the mean of the average error sum of squares over those iterations. This average is the BS PMSE estimate.

chosen n_j times, the level-one design matrix will be singular. A more general problem with this approach is that the bootstrap samples and the original sample have individuals in common. This leads to underestimation of the prediction error (Izenman, 2008, pp. 123–125). One way of improving the estimate is to calculate the PMSE on the out-of-bootstrap sample only (i.e., the members of the original data set who were not sampled for the bootstrap sample in each iteration).

The overly-optimistic in-sample (IS) training PMSE estimate is included here for reference. The IS PMSE is obtained by training the model on and applying prediction to the original data set. This approach does not involve sampling and is calculated only once. An example of prediction is provided using the models from Table 2.2. Under the full model (2.19), using results obtained from REML, prediction for a new student in a school already in the sample could be accomplished using the following prediction rule, with i_{new} referring to the new observation(s) in each school, $j = 1, 2, \ldots, 160$,

$$\begin{aligned}\widehat{y}_{i_{\text{new}}j} =& \widehat{\beta}_1 + \widehat{\beta}_2(\texttt{sector}_{ij}) + \widehat{\beta}_3(\texttt{meanses}_{ij}) \\ & + \widehat{\beta}_4(\texttt{ses}_{i_{\text{new}}j}) + \widehat{\beta}_5(\texttt{sector}_{ij})(\texttt{ses}_{i_{\text{new}}j}) + \widehat{\beta}_6(\texttt{meanses}_{ij})(\texttt{ses}_{i_{\text{new}}j}) \\ & + \tilde{b}_{1j} + \tilde{b}_{2j}(\texttt{ses}_{i_{\text{new}}j}).\end{aligned} \tag{2.20}$$

Note that random effects have tildes rather than hats to indicate that they are not estimates. As noted previously, the random effects are conditional modes given the parameter estimates. The prediction rules for the other models in Table 2.2 follow analogously.

The BS procedure involves making predictions for all the students in each school, for a total of 7,185 predictions. For the CV approach, only one new prediction is made for each school, for a total of 160 predictions. REML estimation was used to reduce bias in the variance-component estimates used in prediction. Table 2.5 contains the PMSE estimate values for the models in Table 2.2. Boldface values in Table 2.5 indicate the smallest values of each estimate of PMSE. From Table 2.5, it appears that the BS and CV PMSE estimates favor the model incorporating `ses`, `sector`, `meanses`, and the cross-level interactions, with a random intercept. These results do not agree with the models chosen by the information-type criteria, which include a random intercept and slope. Such a model was favored by the IS PMSE estimate.

However, obtaining predictions from the same sample on which a model is built provides an unrealistically optimistic estimate and is not encouraged. The model chosen using AIC and BIC was a more general form of the model chosen using the BS and CV PMSE estimates. The more general model included a random slope for `ses` as well as for the intercept. As stated previously, Raudenbush and Bryk explored this more general model and found that the random slope for `ses` was not significant. The smaller prediction error for the

Table 2.5: PMSE of models for the HS&B data set.[a]

Predictors (Level)	Random	IS[b]	BS[b]	CV[b]
SES[c]	N/A	45.21	45.22	45.88
N/A[d]	Int[e]	38.36	39.04	40.38
SES(1)	Int	36.26	36.91	38.02
Sec[f](2)	Int	38.39	39.02	40.35
MSES[g](2)	Int	38.51	38.96	40.27
Sec(2), MSES(2)	Int	38.53	38.95	40.24
SES(1), Sec(2), SES*Sec	Int	36.08	36.70	37.84
SES(1), MSES(2), SES*MSES	Int	36.38	36.84	37.93
SES(1), Sec(2), MSES(2), SES*Sec, SES*MSES	Int	36.15	**36.57**	**37.70**
SES(1)	Int, SES	**35.74**	50.06	37.98
SES(1), Sector(2), SES*Sec	Int, SES	35.93	44.16	37.81
SES(1), MSES(2), SES*MSES	Int, SES	35.88	42.81	37.89
SES(1), Sector(2), MSES(2), SES*Sec, SES*MSES	Int, SES	36.07	40.40	37.73

[a] Boldface values indicate the model chosen using each criterion.
[b] Calculations are based on REML estimates from the `lme4` package in R.
[c] OLS model: note that there is no random term in this model (N/A= Not Applicable)
[d] Variance components model: note that there are no predictors in this model (N/A= Not Applicable)
[e] Int refers to the intercept.
[f] The `Sector` variable is abbreviated as `Sec`.
[g] The `MeanSES` variable is abbreviated as `MSES`.

more specific model lends additional evidence that this random component is not useful. Choice of the final model would depend on what values one places on parsimony, interpretation, and other desired characteristics. Prediction for students in schools not in the original data set would be more complicated and is outside the scope of this dissertation.

2.6 Regularization Approaches

It is critical that practitioners heed the message of George (2000) not to rely on familiar methods when they may not be most appropriate, especially given the development of novel methods. One familiar paraphrase regarding the problematic one-tool-fits-all idea has been attributed variously. The phrase comfortably applies to the reluctance of data investigators to explore unfamiliar techniques, even if they may be more appropriate (Breiman, 2001b): "If all a man has is a hammer, then every problem looks like a nail." Even the best-crafted hammer may not be the most appropriate tool for every job.

Selection of fixed and random effects in the multilevel context is a key issue that has received attention from frequentist and Bayesian perspectives. Careful selection in multilevel models is key (Kreft & de Leeuw, 1998, p. 103):

> Data reduction is a very necessary strategy before starting any multilevel analyses. Multilevel analysis is not a tool for exploring large numbers of variables. Multilevel analysis is a nice tool for exploring small numbers of variables and making small changes in the models one at a time.

Note that this advice was offered in 1998. The field of statistics has advanced considerably since that time, particularly in the area of multilevel regression. While the general idea of avoiding overfit is reasonable, this should not be understood to mean that structured data sets involving large number of predictors cannot be analyzed. The same is true for low-dimensional data sets with complex structures.

Examples of multilevel analysis of larger data sets can be found in various fields. In the field of education, standardized tests are administered annually to almost all students, with performance at the school level being one of the measurement goals. Additionally, epidemiology data contain disease or other health-related information across many centers (see, e.g., Armagan & Dunson, 2011).

While it is tempting to focus on only a portion of the variables available to researchers, much information is lost in that reduction. It is still true that care must be taken because of the potential multicollinearity, instability, and computational difficulties that could result from haphazard kitchen-sink approaches to model building.

Recent work to address the use of multilevel modeling with large numbers of predictors has focused on shrinkage of estimates, but these methods are still under development (see, e.g. Gelman, 2011). One selection approach that has been a fertile area of research for single-level regression is that of regularized regression. In contrast with the unconstrained procedures (e.g., IGLS) described in Section 2.2, regularized estimates are found by constraining the solution region.

The regularization framework has recently been explored for models with multiple levels of variation. Regularization involves the use of one or more penalty parameters to shrink or select coefficients. The penalty term is a function of the parameters that are to be estimated. The addition of the penalty function has the effect of reducing, or shrinking, the sizes of the resulting parameter estimates. Some penalty functions allow estimates to be shrunken to 0. Such functions can be used for selection of model components, reducing model overfit. In single-level regression, regularization is often applied for high dimensional data when the true non-zero coefficients are assumed to be sparse in the set of all components.

Multilevel modeling raises additional dimensionality issues that must be taken into account, even when there are are relatively few predictors overall. No longer are there predictors at just one level. With covariates at level two, we must deal with potential cross-level interactions. Regularization is a useful technique for multilevel models because such models often have a complicated structure with parameters to estimate in both the fixed and random parts of the model. There is a tendency toward instability when there are more than a few random coefficients under consideration. The more complex the mean and covariance structures, the more of a challenge it is to obtain precise estimates and to be able to detect important features in the data.

Regularization approaches may add bias to the estimated coefficients, but the standard errors of the estimates are decreased. Therefore, regularization reduces the instability in parameter estimates. Constraints are placed on functions of the sizes of the regression coefficients (e.g., on the sum of the absolute values of the coefficients) to keep them as small as possible. The shape of the penalty function that is applied will determine whether shrinkage of coefficients and/or selection of predictors results. Hybrid procedures that combine shrinkage and selection have gained popularity in recent years. However, as most multilevel regularization research extends the work from the single-level context, it focuses more on the fixed effects than the random components.

Clustering in data sets further reduces the amount of information available for estimation, and this increases with the strength of the clustering effect. While problems with the inclusion of too many (or too few) fixed effects are familiar from single-level regression, selection of the proper covariance structure is also key. If the structure is underfit, there may be bias in the estimated fixed-effect variances. Alternatively, overfitting could lead (as it may also in the fixed-effects structure) to singularity of the associated covariance matrix (Bondell *et al.*, 2010). An additional motivation is that selection of important fixed and random effects from among many models via traditional subset selection methods is time-consuming and introduces stochastic error (Fan & Li, 2001). There are also computational problems leading to a lack of convergence in some algorithms when backward-deletion methods are employed (Schelldorfer *et al.*, 2011).

This Section provides an overview of regularized regression and procedures that perform regularization of components in hierarchical models.

2.6.1 Overview of Regularization

Let $\mathcal{L}(\boldsymbol{\phi})$ be a function of the parameters $\boldsymbol{\phi}$ that is to be minimized (e.g., a negative log-likelihood). Let λ be a tuning parameter that determines how much shrinkage will be applied. Let $p(\cdot)$ be the penalty function that is applied to the parameters to be estimated. Penalized, or regularized, regression estimates have the following general formula,

$$\widehat{\boldsymbol{\phi}}_{\lambda} = \underset{\boldsymbol{\phi}}{\arg\min} \left(\mathcal{L}\left(\boldsymbol{\phi}\right) + \lambda p\left(\boldsymbol{\phi}\right)\right) \tag{2.21}$$

Larger values of λ lead to larger penalties, and, thus, to more shrinkage or selection (i.e., fewer predictors with non-zero coefficients in the resulting model). In this format, which is referred to as a Lagrangian function format, λ is a Lagrangian multiplier (see, e.g., Nocedal & Wright, 2006, p. 310). The Lagrangian multipliers are referred to in this context as regularization parameters. An alternative, equivalent format is one in which the function to be minimized is constrained, or bounded, by requiring that the penalized term is smaller than a threshold. In this format, (2.21) would be written as

$$\widehat{\boldsymbol{\phi}}_{t} = \underset{\boldsymbol{\phi}}{\arg\min} \left(\mathcal{L}\left(\boldsymbol{\phi}\right) \ni p\left(\boldsymbol{\phi}\right) \leq t\right), \tag{2.22}$$

for a bound t.

Choice of model is made by determining which vector of parameter estimates minimizes the criterion that is being applied. Criteria typically include CV or other prediction-error estimates or penalized-likelihood functions with the general form (2.17).

For single-level regression models,

$$\boldsymbol{\beta}_{\phi} = \arg\min_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{N} (y_i - \sum_{r=1}^{p} x_{ir}\beta_r)^2 + \lambda \sum_{r=1}^{p} \phi(\beta_r) \Big\}. \tag{2.23}$$

Some of the most common regularization procedures introduced for single-level regression include ridge regression ($p(\beta_r) = \beta_r^2$) and the Lasso ($p(\beta_r) = |\beta_r|$). Ridge regression shrinks coefficient estimates but does not function as a selection procedure. The Lasso (Tibshirani, 1996) allows for selection, dropping any variable whose regularization path crosses zero. In general, the regularization penalty $p(\beta_r) = |\beta_r|^{\gamma}$ is referred to as bridge regression (Frank & Friedman, 1993). Shapes of the constraint regions of various regularization penalties are explored in Chapter 3 of Hastie *et al.* (2009) and Chapter 5 of Izenman (2008).

Convexity of the loss and penalty functions and piecewise linearity of the solution paths are desirable characteristics of a regularization procedure in terms of computation. The least-angle regression algorithm (LAR) (Efron *et al.*, 2004) can be modified to compute the Lasso solutions more efficiently with the general algorithm known as LARS.

The non-negative garrote shrinks and selects by scaling the OLS estimates (Breiman, 1995). As noted previously, the penalized function can be thought of as being bounded by a decreasing bound t:

$$\boldsymbol{\beta}_t = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N}(y_i - \sum_{r=1}^{p} x_{ir}\beta_r)^2 \ni \sum_{r=1}^{p} \phi(\beta_r) \leq t. \tag{2.24}$$

Fan & Li (2001) proposed the smoothly-clipped absolute deviation (SCAD) penalty, a quadratic spline penalty that is continuously differentiable (except at 0), symmetric, and non-concave. SCAD solutions are unbiased, sparse, and continuous, and they possess the oracle property: they lead to selection of the "true" data-generating model when it is one of the candidate models. The SCAD-penalized log-likelihood is not convex in $\boldsymbol{\beta}$, making optimization challenging. The SCAD penalty function has the following parametric form for a spline with knots at a and $a\lambda$:

$$p_\lambda(\omega) = \begin{cases} \lambda\omega & \text{if } 0 < \omega \leq \lambda \\ -\frac{\omega^2 - 2a\lambda\omega - \lambda^2}{2(a-1)} & \text{if } \lambda < \omega \leq a\lambda \\ \frac{1}{2}(a+1)\lambda^2 & \text{if } a\lambda < \omega \end{cases}$$

Selection of λ is accomplished using a grid search.

Extensions of the Lasso have been developed for more complicated scenarios. The fused Lasso (Tibshirani *et al.*, 2005) was developed to handle situations such as those involving microarray data, in which the predictors may greatly outnumber the observations and may also be ordered. In this case it is useful to constrain the sizes of adjacent predictors because they are related. The penalized-least-squares criterion applied for the fused Lasso,

$$\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\beta)^2 + \lambda_1\sum_{r=1}^{p}|\beta_r| + \lambda_2\sum_{r=2}^{p}|\beta_r - \beta_{r-1}|, \tag{2.25}$$

contains two tuning parameters that shrink independently based on the sum of all coefficients (λ_1) and the sum of differences of adjacent coefficients (λ_2), respectively, for observations $i = 1, 2, \ldots, N$ and predictors $r = 1, 2, \ldots, p$.

The group Lasso (Yuan & Lin, 2006) allows selection of groups of predictors that form factors, rather than just the individual predictors. The estimate is given by the group Lasso criterion,

$$\frac{1}{2}\|Y - \sum_{r=1}^{p} X_r\beta_r\|^2 + \lambda\sum_{r=1}^{p}\|\beta_r\|_{K_r}, \tag{2.26}$$

in which K_r are covariance matrices that are often taken to be the identity matrix. Sparseness of solutions is focused at the factor level. Other developments include a group non-negative garrote and a group LARS, which are extended in a similar way from their original versions.

Zhou & Zhu (2010) proposed a method to drive grouped selection of predictors in single-level models by separating out an 'intrinsic' parameter and a group-control parameter from each β in a single-level model. They conceptualized the β_r for the rth fixed effect as being a product of the variable's intrinsic parameter α_r and the control parameter d_r. While this approach is appealing, the iterative nature of estimation for multilevel models can lead to difficulties due to a non-positive-definite quadratic matrix in this type of algorithm. Another variation on grouped estimation is the group bridge (Huang *et al.*, 2009). This method also allows both group and individual parameter selection, but it requires multiple tuning parameters that are not simple to choose.

The composite absolute penalties (CAP) family (Zhao *et al.*, 2009) builds on the ideas of group Lasso and bridge regression. CAP takes into account a priori predictor groupings and hierarchies. The regularization process includes an overall l_1 penalty and a within-group l_∞ penalty on maximum coefficient size. This invokes penalization across and within overlapping or non-overlapping groups. Predictors can be selected in a particular order or in accordance with the correlation structure. The term *hierarchical gap* was coined therein as a measure of the distance between the candidate model structure and a minimally suitable hierarchical structure. For instance, adhering to the principle of fixed heredity, interactions should enter a model only after their corresponding main effects. Prior to modifications, this had been an issue with LAR, because interactions tend to be attractive to the algorithm (c.f. Turlach in the discussion of Efron *et al.* (2004)). The hierarchical gap is the number of predictors that one would have to add to the current model so that the model adheres to the principle of heredity. Closure of the hierarchical gap can also take place by removing terms that violate the hierarchical structure.

It would be useful in hierarchical models with heredity structures to regulate the order of entry of predictors. This can be done with CAP by creating overlapping groups, with the secondary predictors included in all groups that include their parents and also in their own groups. This allows additional penalization to take place for the descendents to control their entry into the active set after the entries of their parents. Simulation studies showed that CAP outperformed Lasso when interaction strength was low or moderate and was comparable for strong interactions, in terms of model error. The CAP also selected smaller models than Lasso and resulted in smaller hierarchical gap.

Yuan *et al.* (2009) incorporated heredity structure in estimation and selection through the use of linear inequality constraints with non-negative garrote regularization. Denote the set of coefficients of parents of the rth predictor by P_r. Rather than sacrificing convexity by requiring that all coefficients in P_r be non-zero when $\widehat{\beta}_r$ is non-zero, Yuan *et al.* used indicators of inclusion (non-negative garrote scale coefficients) such that $s_r \leq s_t$ for β_t in P_r. The estimation procedure involves two steps. First, the solution path is computed. Then, tuning to minimize the CV prediction error is applied to select the final estimate. Taking the simple example of a model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + R, \tag{2.27}$$

in which β_{12} is the coefficient for the interaction of the terms X_1 and X_2, strong heredity would require that $s_{12} \leq s_1$ and $s_{12} \leq s_2$. The relevant constraint matrix could be represented as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_{12} \end{bmatrix} \succcurlyeq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \implies \begin{bmatrix} s_1 \geq 0 \\ s_2 \geq 0 \\ s_{12} \geq 0 \\ s_1 - s_{12} \geq 0 \\ s_2 - s_{12} \geq 0 \end{bmatrix}. \tag{2.28}$$

Because the sum of the scaling factors is bounded, the bound can be chosen so that one or more of the scaling factors is exactly zero to invoke selection. Although the approach performed well on simulated and real data sets, it suffered in comparison to unconstrained methods when the data-generating model did not follow strong heredity. A weak heredity version was also offered, and extensions of both types to generalized linear models were described.

2.6.2 Regularization for Hierarchical Models

Selection of Fixed Effects Only

Lan (2006) proposed a penalized method for selecting fixed effects in longitudinal data using the SCAD penalty. Covariance structures were specified a priori in each model that was analyzed. The dissertation focused on both estimation and selection of fixed effects, with estimation of random-effect covariance parameters using REML. The BIC criterion was found to be optimal, of the three investigated, for choosing the tuning parameter λ. First, the penalized model that maximized the REML log-likelihood was selected. Then, BIC values were compared for the models' log-likelihoods evaluated at the SCAD estimates.

Schelldorfer *et al.* (2011) focused on high-dimensional linear mixed models. Such models have many more potential predictors than observations (i.e., $p \gg N$). However, true non-zero coefficients are assumed to be sparse in the vector of all coefficients. Schelldorfer *et al.* developed an `R` package, `lmmLasso`, which invokes an l_1 penalization of the fixed effects. The l_1-penalized objective function to minimize was

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{r=2}^{p} |\beta_k| \, , \tag{2.29}$$

with $\mathbf{V}$ a block-diagonal matrix with group-level matrices $\mathbf{V}_j$ on the diagonal, $\lambda \geq 0$ the penalty parameter ($\lambda = 0$ producing estimates comparable to that of `lmer`), and $r = 2, 3, \ldots, p$ indexing the fixed effects without the intercept. Schelldorfer *et al.* did not penalize fixed effects that had associated random terms (e.g., the intercept in this case). The set of random coefficients was specified a priori, because the high-dimensional context makes joint or simultaneous selection of random effects complicated. An adaptive version is available that is based on the adaptive Lasso (ALasso). The adaptive version uses weights on the covariates. Weights can be specified as the estimates from the non-adaptive version. The regularization parameter λ is chosen using BIC. The model dimensionality penalty is based on the expected df of the Lasso (Zou *et al.*, 2007): $\log N \times \widehat{df}_\lambda$, with $\widehat{df}_\lambda$ the number of non-zero regression coefficients in the vector $\widehat{\boldsymbol{\beta}}$. A block coordinate gradient descent (BCGD) algorithm (Tseng & Yun, 2009a,b) was employed to compute the solution.

Selection Focused on Random Effects

Ahn (2010) proposed a moment-based loss function to select variance components. The goal was to produce a method that would be more robust to violations of the distributional assumptions for these components. The BIC was used for model selection. A selection mechanism for selecting both fixed and random components was also proposed, in which an ALasso was used to select fixed effects. Ahn found that the random-effects structure specification was key to achieving efficiency and model accuracy.

Selection of Fixed and Random Components

Ibrahim *et al.* (2010) examined joint selection of fixed and random effects in mixed-effects models using maximum penalized likelihood (MPL). They applied SCAD and ALasso. The procedure optimizes over mixed effects and the penalty parameters using the E-M algorithm. Ibrahim *et al.* proposed two model selection criteria to be used with the objective function $Q(\widehat{\boldsymbol{\theta}})$. The first, $IC_Q(\boldsymbol{\lambda})$, minimizes

$$IC_Q(\boldsymbol{\lambda}) = -2Q(\widehat{\boldsymbol{\theta}}_\lambda|\widehat{\boldsymbol{\theta}}_0) + c_J(\widehat{\boldsymbol{\theta}}_\lambda), \tag{2.30}$$

in which $\widehat{\boldsymbol{\theta}}_0$ is the unpenalized MLE and $c_J(\boldsymbol{\theta})$ is a function that invokes an AIC- or BIC-type criterion. The second, a random-effects penalty estimator, applies marginal-maximized-likelihood estimation. In that process, $\boldsymbol{\lambda}$ is treated as a hyperparameter of the distribution of the random-effects vector.

Wang *et al.* (2010) proposed a doubly-regularized REML process with a Cholesky decomposition of the random-effects covariance matrix to produce a positive semidefinite final covariance matrix. Wang *et al.* noted that it is clear from the mean and variance of $\mathbf{y}_j$ in the multilevel model structure that the mean and covariance structures are influenced separately by the fixed- and random-effects components. However, the goal was to select on both sets of effects. The method of selection proposed by Wang *et al.* entails applying penalizations to both the fixed and random structures. Some fixed effects are set to zero via an l_1 penalty. Variance parameters (and associated covariances) can be zeroed out using the structure of the Cholesky decomposition. Setting all of the elements in the kth row of the relative precision factor $\mathbf{\Lambda}_{(k)}$ to zero for a given k through the use of a modified l_2 penalty will remove the random effect b_k from the model. In an adaptive version, different penalty weights are imposed depending on the strength of each term's effect, penalizing important components more lightly than components that are less important.

Bondell *et al.* (2010) proposed a method for joint fixed- and random-effects selection. This method was used as the basis for developing the heredity-constrained estimator introduced in the next Chapter. Therefore, a more extensive overview is discussed here.

Bondell *et al.* introduced the M-ALASSO, an adaptive Lasso suited for multilevel data. Estimation was performed using a constrained E-M algorithm to select fixed and random effects at the same time. Joint selection was desirable over separate selection performed iteratively, because the selection of one structure affects the selection of the other. Removal of a random effect entails deleting the row and column associated with the effect in the decomposed scaled covariance matrix, $\boldsymbol{\psi} = \mathbf{D}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}$. The reparameterized model can be expressed as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{D}\boldsymbol{\Gamma}\mathbf{b}_j + \mathbf{r}_j. \tag{2.31}$$

As explained previously, in this type of modified-Cholesky decomposition, $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_q)$ is a $q \times q$ diagonal matrix. The $q \times q$ lower triangular matrix $\boldsymbol{\Gamma} = \{\gamma_{lr}\}$ has a diagonal consisting of 1's. The free elements of $\boldsymbol{\Gamma}$ are $\boldsymbol{\gamma} = (\gamma_{lr} : l = 1, \ldots, q; r = l+1, \ldots, q)'$. Under the reparameterization, the random-effects covariance matrix is a function of $\mathbf{d}$ and $\boldsymbol{\gamma}$.

The data $(\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j)$ within the jth group, $j = 1, 2, \ldots, J$, are independently and identically distributed according to

$$\mathbf{y}_j \sim \mathcal{N}_{n_j}(\mathbf{X}_j\boldsymbol{\beta}, \mathbf{V}_j), \tag{2.32}$$

where

$$\mathbf{V}_j = \sigma^2(\mathbf{Z}_j\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}\mathbf{Z}_j' + \mathbf{I}_{n_j}). \tag{2.33}$$

For $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \mathbf{d}, \boldsymbol{\gamma})'$, the log-likelihood function for the jth group is

$$l_j(\boldsymbol{\Phi}) = -\frac{1}{2}\log|\mathbf{V}_j| - \frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})'(\mathbf{V}_j)^{-1}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}). \tag{2.34}$$

The log-likelihood can be expressed across all groups as

$$l(\boldsymbol{\Phi}) = \sum_{j=1}^{J} l_j(\boldsymbol{\Phi}) = -\frac{1}{2}\log\left|\tilde{\mathbf{V}}\right| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.35}$$

in which $\tilde{\mathbf{V}}$ is a block-diagonal matrix with the group-level $\mathbf{V}_j$ matrices on the diagonal, $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_J')'$, and $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_J']'$.

Building on the ALasso, Bondell *et al.* proposed a penalized criterion,

$$\mathbf{Q}_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b}) = \left\| \mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\boldsymbol{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda_j \left(\sum_{r=1}^{p} \frac{|\beta_r|}{|\overline{\beta}_r|} + \sum_{r=1}^{q} \frac{|d_r|}{|\overline{d}_r|} \right), \tag{2.36}$$

in which they take $\overline{\beta}_r$ to be the GLS estimate of β_r and $\overline{d}_r$ from the modified-Cholesky decomposition of the estimated covariance matrix.

The resulting estimator can be implemented in a linearly constrained quadratic programming framework. In order to use quadratic programming in the package `quadprog`, the quadratic matrix was written in a quadratic form in $(\boldsymbol{\beta}', \mathbf{d}')'$. The transformed objective function is

$$\begin{aligned} \mathbf{Q}_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b}) = & \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathrm{diag}\left(\tilde{\boldsymbol{\Gamma}}\mathbf{b}\right)\left(\mathbf{1}_q \otimes \mathbf{I}_j\right)\mathbf{d} \right\|^2 \\ & + \lambda_j \left(\sum_{r=1}^{p} \frac{|\beta_r|}{|\overline{\beta}_r|} + \sum_{r=1}^{q} \frac{|d_r|}{|\overline{d}_r|} \right). \end{aligned} \tag{2.37}$$

The constrained E-M procedure treats the random effects as unobserved data. In the E-step, the conditional expectation of (2.37) is computed. Then, updated values for the parameter estimates are obtained by minimization with respect to $(\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$, iterating until convergence is achieved. The minimization step, or M-step, alternates between $(\boldsymbol{\beta}', \mathbf{d}')'$ and $\boldsymbol{\gamma}$.

The method used in `quadprog` is an algorithm for linearly constrained convex quadratic programming (Goldfarb & Idnani, 1983). Let $\mathbf{d}$ and $\mathbf{D}$ be the vector and matrix, respectively, in the quadratic function that is to be minimized with respect to parameters $\mathbf{b}$. Let $\mathbf{A}$ be a matrix of linear equality and/or inequality constraints under which the quadratic function is being minimized. These constraints must be linearly independent. The constraints determine the feasible solution space. Let $\mathbf{b}_0$ be the vector of bounding values for those constraints. This optimization scheme solves problems of the form

$$\min_{b}(\mathbf{d}'\mathbf{b} + \frac{1}{2}\mathbf{b}'\mathbf{D}\mathbf{b}) \ni \mathbf{A}'\mathbf{b} \geq \mathbf{b}_0. \tag{2.38}$$

Strict convexity is required. The quadratic matrix must be positive definite. This is required to ensure that the initial unconstrained minimum can be found. This solution acts as the starting point. The linear constraint matrix is evaluated at that starting point to find a constraint which is not met at that point (i.e., is violated). That constraint is then added to the *active set* of constraints. This set contains equality constraints and inequality constraints that are equalities at the solution point. In each iteration, a new solution is found (if possible) that satisfies the active set constraints, and another violated constraint is added to the active set. Constraints may also become inactive during the course of the optimization process. If an ultimate feasible solution is obtained, it meets Kuhn-Tucker conditions for optimality. See Goldfarb & Idnani

(1983) for more technical details about this standard quadratic programming algorithm.

The `Pen.LME` program contributed by Bondell *et al.* incorporates only inequality constraints. The constraint matrix forces the penalty term to be bounded by a sequence of decreasing fractions of the total number of terms. This imposes an adaptive Lasso penalty on both fixed and random terms. The GLS estimates used as weights are obtained by using the `lmer` function.

A grid search is used to find the fractional threshold that produces the model for which BIC is minimized. The df count used is the number of nonzero fixed effects and random parameters.

CHAPTER 3

Heredity-Constrained Multilevel Regularization

This Chapter introduces a procedure (HM-ALASSO) that produces regularized multilevel-model estimates that adhere to principles of strong heredity in both their fixed and random parts. The Chapter includes both proofs of the theoretical qualities of HM-ALASSO and results of an empirical test on a subset of the HS&B data.

In this Chapter, we will further expand the model and make its notation more explicit. This model structure is used in the simulation studies in this Chapter. Consider a model that has an intercept term ($X1_1$) associated with a random intercept and two level-one predictors ($X1_2$,$X1_3$) associated with random slopes. Add two group-level covariates $X2_1$, $X2_2$ at level two. For the

jth group, this model can be represented as

$$Y_{ij} = A_{1j}X1_{1ij} + A_{2j}X1_{2ij} + A_{3j}X1_{3ij} + R_{ij}, \tag{3.1}$$

where

$$\begin{aligned} A_{1j} &= \beta_1 + \beta_2 X2_{1ij} + \beta_3 X2_{2ij} + B_{1j}, \\ A_{2j} &= \beta_4 + \beta_5 X2_{1ij} + \beta_6 X2_{2ij} + B_{2j}, \\ A_{3j} &= \beta_7 + \beta_8 X2_{1ij} + \beta_9 X2_{2ij} + B_{3j}. \end{aligned} \tag{3.2}$$

All elements are as previously defined. We make appropriate substitutions, reordering the equation and relabeling random terms, as described in Chapter 2. We rename interactions using explicit subscripting (e.g., $X1_{2ij}X2_{2ij}$ becomes $X12_{2\_2ij}$). The model equation is written as

$$\begin{aligned} Y_{ij} = {} & \beta_1 X1_{1ij} + \beta_2 X1_{2ij} + \beta_3 X1_{3ij} + \beta_4 X2_{1ij} + \beta_5 X2_{2ij} \\ & + \beta_6 X12_{2\_1ij} + \beta_7 X12_{2\_2ij} + \beta_8 X12_{3\_1ij} + \beta_9 X12_{3\_2ij} \\ & + B_{1j}Z_{1ij} + B_{2j}Z_{2ij} + B_{3j}Z_{3ij} + R_{ij}. \end{aligned} \tag{3.3}$$

In this Chapter, the focus is on both fixed and random heredity. Fixed and random heredity were described in Chapter 2. It is important to consider whether these types of heredity control make sense for a particular model, as in some cases it is theoretically sound to allow violations. In this dissertation, it was of interest to invoke the strong versions of both fixed and random heredity. Meeting the goals of both adhering to heredity and optimally fitting the model to the data is a challenging task. We make the reasonable assumption that the modeler knows the structure of the model (i.e., to which levels of the model potential predictors belong) a priori.

3.1 Heredity-Constrained M-ALASSO

The procedure developed in the course of this dissertation research builds on the joint selection algorithm devised by Bondell *et al.* (2010). The new approach incorporates additional linear constraints in the quadratic programming solution to control selection via heredity-based inequalities (Yuan *et al.*, 2009). This approach avoids the estimation of additional model parameters.

The HM-ALASSO estimation algorithm takes advantage of a versatile linearly constrained quadratic programming framework. A dynamic system of weights is applied in the linear constraint matrix to control estimation and selection. An alternative approach to heredity control that adds linear constraints that provide bounds on various component values may appear to be appealing. For example, a constraint could be constructed to limit the sizes of all interactions associated with a parent to be less than the size of the parent term. This would force descendants' parameter estimates to be 0 if the parent estimate is 0. While this extension addresses very simply the adherence to heredity, it invokes too harsh a bound. For example, when controlling random heredity, it is unrealistic to require that the magnitude of a fixed effect be larger than its associated random-variance component. The addition of linear inequality constraints also increases the difficulty involved in computing the constrained solution of the quadratic programming problem.

One of the main points of the Bondell *et al.* (2010) paper was that the fixed and random structures are not independent. Indeed, including informative fixed effects can reduce residuals and, therefore, random-effects covariance parameters. Therefore, it was of interest in this disseration to determine how one might connect the predictors and their associated random terms more effectively during estimation to produce a more optimal regularized solution.

The M-ALASSO estimator developed by Bondell *et al.* can be extended to incorporate heredity by changing the penalty constraint so that it capitalizes on the hierarchical structure of the data. The estimator can be implemented through a simple modification to the constraint matrix of the `Pen.LME` function. As with M-ALASSO as implemented in `Pen.LME`, the fractional thresholding approach is implemented, constraining the total penalty term by a fraction of the full estimates.

3.2 Description of the HM-ALASSO

The goals in developing this method were to create a procedure that provided reasonable estimates of all parameters, adhered to fixed and random heredity principles, resulted in a parsimonious model, was theoretically justifiable, and was able to be implemented and used in available software (e.g., in R). This last criterion has been cited as important for realizing the benefits of new procedures (Hesterberg *et al.*, 2008). The resulting estimator, dubbed HM-ALASSO for the heredity-constrained M-ALASSO, incorporates heredity-constrained selection directly into the estimation process.

It is clear from the model structure (3.2) that the level-one components (the intercept term $X1_1$, with $X1_2$ and $X1_3$) are components that control both the fixed and random heredity. The intercept is associated with the fixed-effect component for level-two effects. The other random level-one predictors are components of the cross-level interactions (with the level-two predictors). All three terms are components of their respective random portions of the model. The corresponding heredity rules are those in Table 3.1. The HM-ALASSO capitalizes on the model structure to implement heredity controls.

We emphasize that the method was developed to address this strict set of rules under a specific model structure. However, it can be adapted for use with other rule sets and models that can be similarly formulated. It should be noted that this set of logical rules implies that the fixed intercept

Table 3.1: Strong Rules of Fixed and Random Heredity.

The fixed intercept must be in the model for the level-two effects to be in the model.
Level-one fixed effects must be in the model if descendant cross-level interactions are in the model.
The fixed intercept must be in the model if the random intercept is in the model.
Level-one fixed effects must be in the model if corresponding random slopes are in the model.

must be in the model for the model to keep its multilevel structure. The seeming importance of this term compared to other components may not seem desirable. However, the modeler may choose to center or standardize following reasonable guidelines that make sense for the particular model. In that case, if the model can be formulated and parameterized in a hierarchical structure similar to the explicit one used in this dissertation, the penalization framework can be adapted appropriately and applied.

To ensure fixed and random heredity, it is required that the principles in Table 3.1 are followed. At the same time, it is of interest to apply shrinkage and selection to the estimates. The question of how to translate these constraints into the estimation process then arises. A natural constraint would be an indicator function constraint that says that if the controlling term is 0, then the controlleded term must also be 0. Such a function is not readily included as a constraint in (2.38) as implemented in available programming packages,

Table 3.2: Examples of Heredity Weights.

Group	Members	Parents	Weight
$\mathcal{G}_1$	$\{\beta_1, \beta_2, \beta_3\}$	N/A	$\zeta_1 = 1$
$\mathcal{G}_2$	$\{\beta_4, \beta_5, d_1\}$	$X1_1$	$\zeta_2 = \left(\frac{\lvert\widehat{\beta_1}\rvert}{\lvert\overline{\beta_1}\rvert}\right)$
$\mathcal{G}_3$	$\{\beta_6\}$	$X1_2$, $X2_1$	$\zeta_3 = \left(\frac{\lvert\widehat{\beta_2}\rvert}{\lvert\overline{\beta_2}\rvert}\right)\left(\frac{\lvert\widehat{\beta_4}\rvert}{\lvert\overline{\beta_4}\rvert}\right)$
$\mathcal{G}_4$	$\{\beta_7\}$	$X1_2$, $X2_2$	$\zeta_4 = \left(\frac{\lvert\widehat{\beta_2}\rvert}{\lvert\overline{\beta_2}\rvert}\right)\left(\frac{\lvert\widehat{\beta_5}\rvert}{\lvert\overline{\beta_5}\rvert}\right)$
$\mathcal{G}_5$	$\{\beta_8\}$	$X1_3$, $X2_1$	$\zeta_5 = \left(\frac{\lvert\widehat{\beta_3}\rvert}{\lvert\overline{\beta_3}\rvert}\right)\left(\frac{\lvert\widehat{\beta_4}\rvert}{\lvert\overline{\beta_4}\rvert}\right)$
$\mathcal{G}_6$	$\{\beta_9\}$	$X1_3$ $X2_2$	$\zeta_6 = \left(\frac{\lvert\widehat{\beta_3}\rvert}{\lvert\overline{\beta_3}\rvert}\right)\left(\frac{\lvert\widehat{\beta_5}\rvert}{\lvert\overline{\beta_5}\rvert}\right)$
$\mathcal{G}_7$	$\{d_2\}$	$X1_2$	$\zeta_7 = \left(\frac{\lvert\widehat{\beta_2}\rvert}{\lvert\overline{\beta_2}\rvert}\right)$
$\mathcal{G}_8$	$\{d_3\}$	$X1_3$	$\zeta_8 = \left(\frac{\lvert\widehat{\beta_3}\rvert}{\lvert\overline{\beta_3}\rvert}\right)$

which take numeric matrices representing linear constraints.

The HM-ALASSO estimator uses a Lasso-type penalization with dynamic implicit tuning parameters that defines the penalty on each group of predictors based on the amount of shrinkage of parent terms.

Let $\mathcal{G}_k$ denote the kth group of model terms, $k = 1, 2, \ldots, K$. Each group consists of one or more terms in $\boldsymbol{\phi} = (\boldsymbol{\beta}, \mathbf{d})$, and each group is constrained by a heredity weight ζ_k, which is assigned to enforce heredity control. The magnitude of the penalty for each group is varied based on the estimated importance, the importance ratios, of its parent components. In particular, define heredity weights ζ_k for component groups $1, 2, \ldots, K$ in the example model, where $\overline{\phi}$ denotes the full GLS estimate and $\widehat{\phi}$ the previous HM-ALASSO estimate, according to Table 3.2.

The more important a parent component is (i.e., the less shrinkage applied to it), the larger the possible estimate(s) of its descendant(s). If the parent component is shrunken all the way to 0 and selected out, its descendants must also be selected out. The objective function is given in (3.4).

$$\begin{aligned}\mathbf{Q}_c(\boldsymbol{\beta}, \mathbf{d}|\mathbf{y}, \mathbf{b}) = & \left\|\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\mathbf{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\right\|^2 \\ & + \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi \in \mathcal{G}_k} \frac{|\phi|}{|\overline{\phi}|} \right)\end{aligned} \tag{3.4}$$

Note that the heredity weight applies in addition to the adaptive Lasso weight set by Bondell *et al.*. We can also denote the heredity weight with respect to a particular component (e.g., ϕ_r), which would be common to the weighted group $\mathcal{G}$ to which the component belongs, as $\zeta(P(\phi_r))$. For example, $P(\beta_6) = \{\beta_2, \beta_4\}$ in the example model being used in this Chapter.

In the estimation algorithm, if the ζ based on the previous iteration's estimate is 0, the term is removed from the penalty (inequality) constraint. A new equality constraint (to 0) is added for the constrained term for the current iteration. The HM-ALASSO estimation algorithm is outlined in Table 3.3.

3.3 Theoretical Results

HM-ALASSO is an estimator. Therefore, it is necessary to evaluate the asymptotic properties of HM-ALASSO. As is standard for regularized estimators of this form, we prove the consistency, sparsity, and asymptotic normality

Table 3.3: Heredity-Constrained Joint Regularization and Selection Algorithm (HM-ALASSO).

	Compute initial GLS estimates for the full model. Set initial $\boldsymbol{\beta}$ to be the fixed-effects estimates. From the modified Cholesky decomposition of the scaled random-effects covariance matrix, take $\mathbf{d}$ to be the diagonal elements and $\boldsymbol{\gamma}$ to be the lower triangle of elements (divided by $\mathbf{d}$). Iterate Steps 1 and 2 until convergence before moving to the next fractional threshold.
0.	Begin estimation of $\boldsymbol{\beta}$ and $\mathbf{d}$. Initialize $\beta_r = \overline{\beta_r}$ for $r = 1, 2, \ldots, p$ and $d_s = \overline{d_s}$ for $s = 1, 2, \ldots, q$ from GLS estimates.
1.	Update the constraint matrix to incorporate current heredity weights. Minimize the objective function (3.4) using linearly constrained quadratic programming (2.38) to find new estimates for $\boldsymbol{\beta}$ and $\mathbf{d}$.
2	Calculate $\widehat{\gamma}$ using the final $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{d}}$ from 1.
3.	Choose the fractional threshold(s) at which the final model produces the best value of the chosen selection criterion.

of the HM-ALASSO.

For each Theorem, it is assumed that the data $(\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j)$ within the jth group, $j = 1, 2, \ldots, J$, are independently and identically distributed within group according to (2.32) and (2.33). The log-likelihood function is (2.34) for the jth group and (2.35) across all groups.

Let $\mathbf{\Phi}^* = (\boldsymbol{\beta}^{*'}, \mathbf{d}^{*'}, \boldsymbol{\gamma}^{*'})'$ denote the true values of $(\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$, with $\overline{\mathbf{\Phi}} = (\overline{\boldsymbol{\beta}}', \overline{\mathbf{d}}', \overline{\gamma}')'$ the GLS estimates and $\tilde{\mathbf{\Phi}} = (\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{d}}', \tilde{\gamma}')$ the estimates produced by the HM-ALASSO estimator. Let $\mathbf{\Phi}_{0^c}$ represent the nonzero terms and $\mathbf{\Phi}_0$ represent the zero terms of the parameter set, with analogous notation for true values and estimates. Split the true parameter vector into its true zero and nonzero parts (i.e., $(\boldsymbol{\beta}'^*, \mathbf{d}'^*, \boldsymbol{\gamma}'^*)' = ((\boldsymbol{\beta}_0'^*, \mathbf{d}_0'^*, \boldsymbol{\gamma}_0'^*)', (\boldsymbol{\beta}_{0^c}'^*, \mathbf{d}_{0^c}'^*, \boldsymbol{\gamma}_{0^c}'^*)'))$. Estimates are defined analogously to their true counterparts.

Let $p^* = p_1^* + p_2^* + p_{12}^*$ and q^* be the true number of predictors (by levels 1, 2, and crosslevel) and random coefficients, respectively. Let $p = p1 + p2 + p12$ and $q = p1$ represent numbers of possible fixed-effect terms (by levels 1, 2, and crosslevel) and random coefficients, respectively. We let $q = q_{0^c}$ and $q = q_0$ denote the numbers of nonzero and zero random terms, respectively. Similarly, we partition each level of the fixed effects (i.e., $p1 = p1_{0^c} + p1_0$). Estimates and true parameters are represented analogously. Without loss of generality, we can index the nonzero terms first for the random coefficients. We can do the same within each level of the fixed coefficent set (e.g., for level-one fixed effects, taking $r = 1, 2, \ldots, p1_{0^c}, p1_{0^c}+1, \ldots, p1_{0^c}+p1_0$.) Let $T = p+q(q+1)/2$ be the total number of fixed effects and random-effects covariance-matrix parameters.

We impose typical regularity conditions.

(1) First and second derivatives of the log-likelihood function exist and and are bounded in probability. The first and second derivatives of the log-likelihood with respect to particular terms in $\mathbf{\Phi}$ are as follows (see, e.g., Longford, 1993, pp. 41–42). The first partial derivative with respect to β_r is

$$\frac{\partial}{\partial \beta_r} l\left(\mathbf{\Phi}\right) = \mathbf{X}_r' \tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{3.5}$$

with $\mathbf{X}_r$ the rth column of the stacked $\mathbf{X}$ matrix. For $\phi_r \in (\mathbf{d}', \boldsymbol{\gamma}')'$,

$$\frac{\partial}{\partial \phi_r} l\left(\mathbf{\Phi}\right) = -\frac{1}{2}\left[\mathrm{tr}\left(\tilde{\mathbf{V}}^{-1}\frac{\partial}{\partial \phi_r}\right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}\frac{\partial}{\partial \phi_r}\tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right].$$

Note that $\tilde{\mathbf{V}}^{-1}$ is the inverse of a block diagonal matrix and is itself block diagonal. For each diagonal block of the form 2.33 for the jth group, denote

$$\mathbf{S}_j^{(r)} = \frac{\partial}{\partial d_r}\mathbf{V}_j = \mathbf{Z}_j\left[\frac{\partial}{\partial d_r}\left(\mathbf{D}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}\right)\right]\mathbf{Z}_j',$$

and

$$\mathbf{T}_j^{(r)} = \frac{\partial}{\partial \gamma_r}\mathbf{V}_j = \mathbf{Z}_j\mathbf{D}\left[\frac{\partial}{\partial d_r}\left(\mathbf{\Gamma}\mathbf{\Gamma}'\right)\right]\mathbf{D}\mathbf{Z}_j'.$$

Both $\mathbf{S}$ and $\mathbf{T}$ matrices consist of a function of: ϕ_r in cell $[r, r]$; 0, 1 or a function of the remaining covariance parameters in the cells in the rth row and rth column; and 0 elsewhere. Further, let $\tilde{\mathbf{S}}^{(r)}$ and $\tilde{\mathbf{T}}^{(r)}$ denote the block diagonal matrices formed from these group level matrices, respectively. Then,

$$\frac{\partial}{\partial d_r} l\left(\boldsymbol{\Phi}\right) = -\frac{1}{2}\left[\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\tilde{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right], \tag{3.6}$$

and

$$\frac{\partial}{\partial \gamma_r} l\left(\boldsymbol{\Phi}\right) = -\frac{1}{2}\left[\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{T}}^{(r)}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{T}}^{(r)}\tilde{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right].$$

The second partial derivatives with respect to terms ϕ_r and ϕ_s are

$$\frac{\partial^2}{\partial \beta_s\,\partial \beta_s} l\left(\boldsymbol{\Phi}\right) = -\mathbf{X}_r\tilde{\mathbf{V}}^{-1}\mathbf{X}_s \tag{3.7}$$

$$\frac{\partial^2}{\partial \beta_r\,\partial d_s} l\left(\boldsymbol{\Phi}\right) = -\mathbf{X}_r'\tilde{\mathbf{S}}^{(s)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.8}$$

$$\frac{\partial^2}{\partial \beta_r\,\partial \gamma_s} l\left(\boldsymbol{\Phi}\right) = -\mathbf{X}_r'\tilde{\mathbf{T}}^{(s)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.9}$$

$$\frac{\partial^2}{\partial d_r\,\partial d_s} l\left(\boldsymbol{\Phi}\right) = \frac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(s)}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(s)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.10}$$

$$\frac{\partial^2}{\partial d_r\,\partial \gamma_s} l\left(\boldsymbol{\Phi}\right) = \frac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{T}}^{(s)}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{S}}^{(r)}\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{T}}^{(s)}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.11}$$

When evaluated at the true parameters, the first partial derivatives have expectation 0 and variance the expected information matrix.

(2) The Fisher (expected) information matrix $\mathbf{I}(\mathbf{\Phi})$ is finite and positive definite when evaluated at the true parameters $\mathbf{\Phi}^*$.

(3) An open subset of the parameter space of $\mathbf{\Phi}$ exists that contains the true parameter set $\mathbf{\Phi}^*$ such that all third derivatives of the parameter set exist in the open subset and are bounded by functions with finite expectations. Bounds are also placed on the third partial derivative with respect to the random terms. For any three parameters $\phi_{r1}, \phi_{r2}, \phi_{r3}$ in $(\boldsymbol{\beta}, \mathbf{d}, \boldsymbol{\gamma})'$, there exists a function $M_{r1,r2,r3}$ with finite expectation, where the function

$$M_{r1,r2,r3}(\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j) > \left| \frac{\partial^3}{\partial \phi_{r1}\, \partial \phi_{r2}\, \partial \phi_{r3}} l_j \left(\mathbf{\Phi}\right) \right|.$$

With these regularity conditions in place, the following proofs can be presented. In Theorem 1, we prove that there exists a local maximizer of our objective function in the neighborhood of the true parameter set. The local maximizer is $\sqrt{m}$-consistent. In Theorem 2, we show that, with probability approaching 1, the objective function evaluated at the resulting estimate reaches a maximum when the zero parameters as estimated as $\mathbf{0}$. This proves the sparsity quality of the estimator. In Theorem 3, we show that our estimator has an asymptotically normal distribution.

For the non-negative regularization parameter λ_m, define the expanded objective function to be

$$
\begin{aligned}
\mathbf{Q}_{(m)}\left(\mathbf{\Phi}\right) &= l(\mathbf{\Phi}) - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi \in G_k} \frac{|\phi|}{|\bar{\phi}|} \right) \\
&= -\frac{1}{2}\log\left|\tilde{\mathbf{V}}\right| - \frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)' \tilde{\mathbf{V}}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) \\
&\quad - \lambda_m \left(\sum_{r1=1}^{p1} \frac{|\beta_{r1}|}{|\bar{\beta}_{r1}|} + \frac{\left|\tilde{\beta}_1^-\right|}{|\bar{\beta}_1|} \left(\frac{|d_1|}{|\bar{d}_1|} + \sum_{r2=p1+1}^{p1+p2} \frac{|\beta_{r2}|}{|\bar{\beta}_{r2}|} \right) \right. \\
&\quad \left. + \sum_{r1=2}^{p1} \frac{\left|\tilde{\beta}_{r1}^-\right|}{|\bar{\beta}_{r1}|} \left(\frac{|d_{r1}|}{|\bar{d}_{r1}|} + \sum_{r2=p1+1}^{p1+p2} \frac{\left|\tilde{\beta}_{r2}^-\right|}{|\bar{\beta}_{r2}|} \frac{|\beta_{r1,r2}|}{|\bar{\beta}_{r1,r2}|} \right) \right),
\end{aligned} \tag{3.12}
$$

where $\tilde{\phi}_t^-$ represents the regularized estimate of a parameter ϕ_t from the previous iteration of the quadratic program. Subscripted indices in the format $(r1, r2)$ are cross-level interactions of terms with subscripted indices $r1$ at level one and $r2$ at level two. All other components are as previously defined.

Theorem 1. For the parameter set $\mathbf{\Phi} = \left(\mathbf{\Phi}_{0^c}', \mathbf{0}'\right)'$, letting $\lambda_m/\sqrt{m} \to 0$, there exists a local maximizer $\tilde{\mathbf{\Phi}} = \left(\tilde{\mathbf{\Phi}}_{0^c}', \mathbf{0}'\right)'$ of $\mathbf{Q}\left(\mathbf{\Phi}_{0^c}', \mathbf{0}'\right)'$ such that $\left\|\tilde{\mathbf{\Phi}}_{0^c} - \mathbf{\Phi}_{0^c}^*\right\| = O_p(m^{-\frac{1}{2}})$ and $\tilde{\mathbf{\Phi}}_{0^c}$ is $\sqrt{m}$-consistent for $\mathbf{\Phi}_{0^c}^*$.

Proof. It is of interest to show that there is a neighborhood of the true parameter set that contains the estimate with probability 1. That is,

$$
P\left\{ \sup_{\|\mathbf{u}\|=C} \mathbf{Q}\begin{pmatrix} \mathbf{\Phi}_{0^c}^* + \alpha_m \mathbf{u} \\ \mathbf{0} \end{pmatrix} < \mathbf{Q}\begin{pmatrix} \mathbf{\Phi}_{0^c}^* \\ \mathbf{0} \end{pmatrix} \right\} \geq 1 - \epsilon,
$$

for some $\epsilon > 0$, letting $\alpha_m = m^{-\frac{1}{2}} = \frac{1}{\sqrt{m}}$ and $\mathbf{u} \neq \mathbf{0}$. Define the difference in objective functions as

$$
\begin{aligned}
D_m(\mathbf{u}) &\equiv \mathbf{Q}\left(\mathbf{\Phi}^*_{0^c} + \alpha_m \mathbf{u}\right) - \mathbf{Q}\left(\mathbf{\Phi}^*_{0^c}\right) \\
&= \left[l\left(\mathbf{\Phi}^*_{0^c} + \alpha_m \mathbf{u}\right) - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k} \frac{|\phi^*_r + \alpha_m \mathbf{u}|}{|\overline{\phi_r}|} \right) \right] \\
&\quad - \left[l\left(\mathbf{\Phi}^*_{0^c}\right) - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k} \frac{|\phi^*_r|}{|\overline{\phi_r}|} \right) \right].
\end{aligned}
$$

Take the Taylor series expansion of $\mathbf{Q}\left(\mathbf{\Phi}^*_{0^c} + \alpha_m \mathbf{u}\right)$ about $\mathbf{\Phi}^*_{0^c}$. Note that $\left(\mathbf{\Phi}^*_{0^c} + \alpha_m \mathbf{u} - \mathbf{\Phi}^*_{0^c}\right) = \alpha_m \mathbf{u}$. Let R represent the remainder. The expansion can be represented as,

$$
\begin{aligned}
\mathbf{Q}\left(\mathbf{\Phi}^*_{0^c} + \alpha_m \mathbf{u}\right) &= \mathbf{Q}\left(\mathbf{\Phi}^*_{0^c}\right) + \left(\mathbf{Q}'\left(\mathbf{\Phi}^*_{0^c}\right)\right)' \alpha_m \mathbf{u} + \frac{1}{2}\alpha_m^2 \mathbf{u}'\mathbf{Q}''\left(\mathbf{\Phi}^*_{0^c}\right)\mathbf{u} + R \\
&= l\left(\mathbf{\Phi}^*_{0^c}\right) - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \mathbf{\Phi}^*_{0^c}} \frac{|\phi^*_r|}{|\overline{\phi_r}|} \right) + l'\left(\mathbf{\Phi}^*\right) \alpha_m \mathbf{u} \\
&\quad - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \mathbf{\Phi}^*_{0^c}} \frac{\operatorname{sgn}(\phi^*_r)}{|\overline{\phi_r}|} \right) \alpha_m \mathbf{u} \\
&\quad + \frac{1}{2}\alpha_m^2 \mathbf{u}'\mathbf{l}''\left(\mathbf{\Phi}^*_{0^c}\right)\mathbf{u} + 0 + R,
\end{aligned}
$$

noting that the derivative of the absolute value function is the sign function when its argument is nonzero, and the derivative of the sign function is 0 when its argument is nonzero (i.e., higher order derivatives vanish). Because these are functions of the true parameters and are only considered on those terms that are nonzero, these statements hold. Therefore,

$$\begin{aligned}
D_m(\mathbf{u}) =& l\left(\boldsymbol{\Phi}_{0^c}^*\right) - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \boldsymbol{\Phi}_{0^c}^*} \frac{|\phi_r^*|}{|\overline{\phi_r}|}\right) + l'\left(\boldsymbol{\Phi}_{0^c}^*\right) \alpha_m \mathbf{u} \\
& - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \boldsymbol{\Phi}_{0^c}^*} \frac{\operatorname{sgn}(\phi_r^*)}{|\overline{\phi_r}|}\right) \alpha_m \mathbf{u} + \frac{1}{2}\alpha_m^2 \mathbf{u}'\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\mathbf{u} + R \\
& - l\left(\boldsymbol{\Phi}_{0^c}^*\right) + \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \boldsymbol{\Phi}_{0^c}^*} \frac{|\phi_r^*|}{|\overline{\phi_r}|}\right) \\
=& l'\left(\boldsymbol{\Phi}_{0^c}^*\right) \alpha_m \mathbf{u} + \frac{1}{2}\alpha_m^2 \mathbf{u}'\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\mathbf{u} \\
& - \lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \boldsymbol{\Phi}_{0^c}^*} \frac{\operatorname{sgn}(\phi_r^*)}{|\overline{\phi_r}|}\right) \alpha_m \mathbf{u} + R.
\end{aligned}$$

The remainder consists of a function of the third partial derivatives of the log-likelihood. This is bounded by assumption. We examine each remaining term of the final right hand side of the equation in sequence.

1. $l'\left(\boldsymbol{\Phi}_{0^c}^*\right)\alpha_m = \frac{1}{\sqrt{m}} l'\left(\boldsymbol{\Phi}_{0^c}^*\right) \xrightarrow{p} \mathcal{N}\left(\mathbf{0}, \mathcal{I}(\boldsymbol{\Phi}_{0^c}^*)\right)$. Because the information matrix is finite by assumption on the true parameter set when the true zero parameters are known, the score function evaluated at the true parameters is bounded in probability (i.e., $\frac{1}{\sqrt{m}} l'\left(\boldsymbol{\Phi}_{0^c}^*\right) = O_p(1)$.)

2. $-\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)$ represents the observed information, and

$$\frac{1}{m}\left(-\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\right) \xrightarrow{p} E\left(-\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\right) = \mathcal{I}(\boldsymbol{\Phi}_{0^c}^*),$$

with $\mathcal{I}(\mathbf{\Phi}_{0^c}^*)$ the Fisher (expected) information. Therefore,

$$\frac{1}{m}\left(-\mathbf{l}''\left(\mathbf{\Phi}_{0^c}^*\right)\right) = \mathcal{I}(\mathbf{\Phi}_{0^c}^*) + o_p(1),$$

where a sequence of random variables $A_n = o_p(1) \Leftrightarrow A_n \xrightarrow{p} 0$. Note that the Fisher information at the true parameter set, when the true zero parameters are known, is finite and positive definite by our assumptions.

3. The third term can be rewritten and expanded, considering only true nonzero terms, as

$$\begin{aligned}
&-\lambda_m \left(\sum_{k=1}^{K} \zeta_k \sum_{\phi_r \in G_k \cap \mathbf{\Phi}_{0^c}^*} \frac{\operatorname{sgn}(\phi_r^*)}{\left|\overline{\phi_r}\right|} \right) \alpha_m \mathbf{u}_r \\
&\quad = -\frac{\lambda_m}{\sqrt{m}} \left(\sum_{k=1}^{K} \prod_{t \in P(\mathcal{G}_k)} \frac{\left|\tilde{\phi}_t^-\right|}{\left|\overline{\phi_t}\right|} \sum_{\phi_r \in G_k \cap \mathbf{\Phi}_{0^c}^*} \frac{\operatorname{sgn}(\phi_r^*)}{\left|\overline{\phi_r}\right|} \right) \mathbf{u}_r \\
&\quad = -\frac{\lambda_m}{\sqrt{m}} \sum_{r1=1}^{p1_{0^c}} \frac{\operatorname{sgn}\left(\beta_{r1}^*\right)}{\left|\overline{\beta_{r1}}\right|} \mathbf{u}_{r1} \\
&\qquad - \frac{\lambda_m}{\sqrt{m}} \frac{\left|\tilde{\beta}_1^-\right|}{\left|\overline{\beta_1}\right|} \left(\frac{\operatorname{sgn}\left(d_1^*\right)}{\left|\overline{d_1}\right|} \mathbf{u}_{q1} + \sum_{r2=p1+1}^{p1+p2_{0^c}} \frac{\operatorname{sgn}\left(\beta_{r2}^*\right)}{\left|\overline{\beta_{r2}}\right|} \mathbf{u}_{r2} \right. \\
&\qquad \left. + \sum_{r1=2}^{p1_{0^c}} \frac{\left|\tilde{\beta}_{r1}^-\right|}{\left|\overline{\beta_{r1}}\right|} \left(\frac{\operatorname{sgn}\left(d_{r1}^*\right)}{\left|\overline{d_{r1}}\right|} \mathbf{u}_{qr1} + \sum_{r2=p1+1}^{p1+p2_{0^c}} \frac{\left|\tilde{\beta}_{r2}^-\right|}{\left|\overline{\beta_{r2}}\right|} \frac{\operatorname{sgn}(\beta_{(r1,r2)}^*)}{\left|\overline{\beta_{(r1,r2)}}\right|} \mathbf{u}_{(r1,r2)} \right) \right),
\end{aligned}$$

where $P(\mathcal{G}_k)$ represent the parent term(s) for the kth term group $(\mathcal{G}_k)$ and $\tilde{\phi}_t^-$ represents the regularized estimate of a parameter ϕ_t from the previous iteration of the quadratic program.

The first term in the expansion (call it Ⓐ), behaves the same way as the estimates generated by the M-ALASSO method of Bondell *et al.* because of the analogous weighting term. The parameter's weight in the penalty term is just the reciprocal of the absolute value of the MLE. Therefore, by the same arguments, if we have $\frac{\lambda_m}{\sqrt{m}} \to 0$ as $m \to \infty$, and we know that $\frac{1}{|\bar{\beta}_{r1}|} \xrightarrow{p} \frac{1}{|\beta_{r1}|}$ for $r1 = 1, 2, \ldots, p1$ the level-one fixed terms (because the MLE is functionally invariant), we have Ⓐ$\xrightarrow{p} 0$.

The second part of the expansion, Ⓑ, consists of terms (second-level fixed terms and the random intercept) that are weighted by the estimate of the fixed intercept from Ⓐ. This is essentially an M-ALASSO estimate. The M-ALASSO estimates were shown in Bondell *et al.* to be $\sqrt{m}$-consistent and to converge in probability to their respective true parameters. The ratio of convergent random sequences converges due to the property of preservation of convergence in products. The MLE is functionally invariant. Therefore, $\frac{|\tilde{\beta}_1^-|}{|\bar{\beta}_1|} \xrightarrow{p} \frac{|\beta_1|}{|\beta_1|} = 1$. The weight goes to 1 when considering the true parameter set. The estimates are again left in the structure of Ⓐ, which were shown to converge to 0. Because we are considering the zero terms as fixed at $\mathbf{0}$, if the fixed intercept is 0 then the weight will also be 0 and the term will drop out.

The third term, Ⓒ, contains the cross-level interactions and random terms other than the intercept. For the random terms, the weights are estimates from Ⓐ, which were shown to converge to 0. For the fixed terms, the weights are products of estimates from Ⓐ and Ⓑ. Each were shown to converge to 0. Therefore, the product of the terms also converges to 0. Because we are considering the zero terms as fixed at $\mathbf{0}$, if either parent term is 0 then the weight will also be 0 and the term will drop out.

Taking all parts together, $D_m(\mathbf{u}) = O_p(1)\mathbf{u} - \frac{1}{2}\mathbf{u}' \left[\mathcal{I}(\mathbf{\Phi}^*_{0^c}) + o_p(1)\right]\mathbf{u} - o_p(1)$. Letting C be large enough, with $||\mathbf{u}|| \leq C$, the second term is of order C^2 and dominates the other terms. This proves that there is, for large enough C, a local maximum that exists in the ball $\left\{\left(\mathbf{\Phi}^*_{0^c} + \frac{1}{\sqrt{m}}\mathbf{u}\right) : ||\mathbf{u}|| \leq C\right\}$ with probability at least $1-\epsilon$ (i.e., a local maximizer $\tilde{\mathbf{\Phi}} = (\tilde{\mathbf{\Phi}}_{0^c}, \mathbf{0})$ of $\mathbf{\Phi}^* = (\mathbf{\Phi}^*_{0^c}, \mathbf{0})$ exists such that $||\tilde{\mathbf{\Phi}}_{0^c} - \mathbf{\Phi}^*_{0^c}|| = O_p(m^{-\frac{1}{2}})$). $\square$

Theorem 2. If $\lambda_m \to \infty$, then with probability approaching 1, for any $\mathbf{\Phi}_{0^c}$ such that

$$||\mathbf{\Phi}_{0^c} - \mathbf{\Phi}^*_{0^c}|| \leq Cm^{-\frac{1}{2}} \tag{3.13}$$

with C a positive constant,

$$\mathbf{Q}\left\{\begin{pmatrix}\mathbf{\Phi}_{0^c} \\ \mathbf{0}\end{pmatrix}\right\} = \max_{||\mathbf{\Phi}_0|| \leq Cm^{-\frac{1}{2}}} \mathbf{Q}\left\{\begin{pmatrix}\mathbf{\Phi}_{0^c} \\ \mathbf{\Phi}_0\end{pmatrix}\right\}.$$

Proof We show, with probability approaching 1 as $m \to \infty$ for any $\mathbf{\Phi}_{0^c}$ satisfying the stated condition and small enough $\epsilon_m = Cm^{-\frac{1}{2}}$ that, for all true zero terms $\phi_r \ni \phi_r \in \mathbf{\Phi}_0^*$ with $\frac{\partial}{\partial \phi_r}\mathbf{Q}(\mathbf{\Phi}) = \frac{\partial}{\partial \phi_r} l(\mathbf{\Phi}) - \lambda_m \zeta_k \frac{1}{|\phi_r|} \operatorname{sgn}(\phi_r)$ for $\zeta(\phi_r)$ the heredity weight applied to term ϕ_r,

$$\frac{\partial}{\partial \phi_r}\mathbf{Q}(\mathbf{\Phi}) < 0$$

when $0 < \phi_r < \epsilon_m$ and

$$\frac{\partial}{\partial \phi_r}\mathbf{Q}(\mathbf{\Phi}) > 0$$

when $-\epsilon_m < \phi_r < 0$. In words, the partial derivative of the objective function with respect to a zero term has a different sign than the term. Therefore, the objective function reaches a maximum at $\mathbf{0}$ for the zero terms. Note that this range of zero terms implies that

$$\|\mathbf{\Phi}_0 - \mathbf{\Phi}_0^*\| \leq Cm^{-\frac{1}{2}} \tag{3.14}$$

with C a positive constant. By the Triangle Inequality for Euclidean norms,

$$\|\mathbf{\Phi} - \mathbf{\Phi}^*\| \leq Cm^{-\frac{1}{2}}. \tag{3.15}$$

It is sufficient to examine the behavior of the objective function's gradient for $\phi_r \in (\boldsymbol{\beta}_0', \mathbf{d}_0')'$. Estimating the zero parameters in $\mathbf{d}$ as 0 will cause their associated parameters in $\boldsymbol{\gamma}$ to be estimated as 0.

Beginning with $\phi_r \in \boldsymbol{\beta}_0$, we use a Taylor series expansion around $\boldsymbol{\Phi}^*$, with $\boldsymbol{\Phi}_*$ between $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^*$,

$$\frac{\partial}{\partial \phi_r}\mathbf{Q}\left(\boldsymbol{\Phi}\right) = \frac{\partial}{\partial \phi_r} l\left(\boldsymbol{\Phi}^*\right) + \sum_{s=1}^{T} \frac{\partial^2}{\partial \phi_r \, \partial \phi_s} l\left(\boldsymbol{\Phi}^*\right)\left(\phi_s - \phi_s^*\right) \tag{3.16}$$

$$+ \sum_{s=1}^{T}\sum_{t=1}^{T} \frac{\partial^3}{\partial \phi_r \, \partial \phi_s \, \partial \phi_t} l\left(\boldsymbol{\Phi}_*\right)\left(\phi_s - \phi_s^*\right)(\phi_t - \phi_t^*) \tag{3.17}$$

$$- \lambda_m \zeta(\phi_r) \frac{\operatorname{sgn}\left(\beta_r\right)}{\left|\beta_r\right|}, \tag{3.18}$$

because the derivative of the sign function is 0 and higher order derivatives vanish.

Starting with β_r and from the first and the second partial derivatives given previously, with $\|\boldsymbol{\Phi}_* - \boldsymbol{\Phi}^*\| \leq \|\boldsymbol{\Phi} - \boldsymbol{\Phi}^*\|$,

$$\begin{aligned}
\frac{\partial}{\partial \beta_r}\mathbf{Q}\left(\boldsymbol{\Phi}\right) = \frac{\partial}{\partial \beta_r} l\left(\boldsymbol{\Phi}\right) &= \mathbf{X}_r' \tilde{\mathbf{V}}^{-1*}\left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\right) + \sum_{s=1}^{p} \mathbf{X}_r' \tilde{\mathbf{V}}^{-1*}\mathbf{X}_s\left(\beta_s - \beta_s^*\right) \\
&- \sum_{s=p+1}^{p+q} \mathbf{X}_r' \left(-\tilde{\mathbf{V}}^{-1*}\tilde{\mathbf{S}}^{(s)*}\tilde{\mathbf{V}}^{-1*}\right)\left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\right)\left(d_s - d_s^*\right) \\
&- \sum_{s=p+q+1}^{p+q(q-1)/2} \mathbf{X}_r' \left(-\tilde{\mathbf{V}}^{-1*}\tilde{\mathbf{T}}^{(s)*}\tilde{\mathbf{V}}^{-1*}\right)\left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\right)\left(\gamma_s - \gamma_s^*\right) \\
&- \sum_{s=1}^{T}\sum_{t=1}^{T} \frac{\partial^3}{\partial \beta_r \, \partial \phi_s \, \partial \phi_t} l\left(\boldsymbol{\Phi}_*\right)\left(\phi_s - \phi_s^*\right)(\phi_t - \phi_t^*) \\
&- \lambda_m \zeta(\phi_r) \frac{\operatorname{sgn}\left(\beta_r\right)}{\left|\beta_r\right|}.
\end{aligned}$$

The third derivatives are bounded by assumption, and we have

$$\begin{aligned}\frac{1}{\sqrt{m}}\frac{\partial}{\partial\beta_r}\mathbf{Q}\left(\mathbf{\Phi}\right) = \frac{1}{\sqrt{m}}\Bigg(& O_p\left(\sqrt{m}\right) - m\sum_{s=1}^{p}\left\{\mathbf{X}_r'\tilde{\mathbf{V}}_0^{-1}\mathbf{X}_s + o_p(1)\right\}(\beta_s-\beta_s^*) \\ & -m\sum_{s=p+1}^{p+q} o_p(1)(d_s-d_s^*) \\ & -m\sum_{s=p+q+1}^{p+q(q-1)/2} o_p(1)(\gamma_s-\gamma_s^*) - \lambda_m\zeta(\phi_r)\frac{\operatorname{sgn}\left(\beta_r\right)}{\left|\beta_r\right|}\Bigg).\end{aligned}$$

Under the assumption (3.15), $(\mathbf{\Phi}-\mathbf{\Phi}^*) = O_p(m^{-\frac{1}{2}})$. Therefore,

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial\phi_r}\mathbf{Q}\left(\mathbf{\Phi}\right) = \frac{1}{\sqrt{m}}O_p(1) - \frac{\lambda_m}{\sqrt{m}}\zeta(\phi_r)\frac{\operatorname{sgn}\left(\beta_r\right)}{\left|\beta_r\right|}.$$

As show in the proof of Theorem 1, $\boldsymbol{\zeta} \xrightarrow{p} \mathbf{1}$. Under the assumption (3.14), $(\mathbf{\Phi}_0-\mathbf{\Phi}_0^*) = O_p(m^{-\frac{1}{2}})$ (i.e., $\frac{1}{\sqrt{m}}\left(\mathbf{\Phi}_0-\mathbf{\Phi}_0^*\right) = \frac{1}{\sqrt{m}}\left(\mathbf{\Phi}_0-\mathbf{0}\right) = \frac{1}{\sqrt{m}}\mathbf{\Phi}_0 = O_p(1)$). We have

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial\phi_r}\mathbf{Q}\left(\mathbf{\Phi}\right) = \frac{1}{\sqrt{m}}O_p(1) - \lambda_m\frac{\operatorname{sgn}\left(\beta_r\right)}{O_p(1)},$$

and, as $\lambda_m\to\infty$, the sign of β_r completely determines the (opposite) sign of $\mathbf{Q}\left(\mathbf{\Phi}\right)$.

When evaluating terms d_r, it should first be noted that these terms, diagonal entries of a modified-Cholesky decomposition, have a nonnegative parameter space. However, under full maximum likelihood estimation, the likelihood function may explore the negative parameter space. From the first and the second derivatives given previously, with $\|\mathbf{\Phi}_* - \mathbf{\Phi}^*\| \leq \|\mathbf{\Phi} - \mathbf{\Phi}^*\|$,

$$\begin{aligned}\frac{1}{\sqrt{m}}\frac{\partial}{\partial\phi_r}\mathbf{Q}(\mathbf{\Phi}) = \frac{1}{\sqrt{m}}\Bigg(& O_p\left(\sqrt{m}\right) - m\sum_{s=1}^{p}\left\{\mathbf{X}_r'\mathbf{V}_0^{-1}\mathbf{X}_s + o_p(1)\right\}(\beta_s - \beta_s^*) \\ & - m\sum_{s=p+1}^{p+q} o_p(1)(d_s - d_s^*) - m\sum_{s=p+q+1}^{p+q(q-1)/2} o_p(1)(\gamma_s - \gamma_s^*) \\ & - \lambda_m\zeta_k\frac{\operatorname{sgn}(\beta_r)}{|\beta_r|}\Bigg)\end{aligned}$$

Under the assumption 3.15, $(\mathbf{\Phi} - \mathbf{\Phi}^*) = O_p(m^{-\frac{1}{2}})$. Therefore,

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial\phi_r}\mathbf{Q}(\mathbf{\Phi}) = \frac{1}{\sqrt{m}}O_p(1) - \frac{\lambda_m}{\sqrt{m}}\zeta_k\frac{\operatorname{sgn}(\beta_r)}{|\beta_r|}.$$

As show in the proof of Theorem 1, $\boldsymbol{\zeta} \xrightarrow{p} \mathbf{1}$. Under the assumption 3.14, $(\mathbf{\Phi}_0 - \mathbf{\Phi}_0^*) = O_p(m^{-\frac{1}{2}})$ (i.e., $\frac{1}{\sqrt{m}}(\mathbf{\Phi}_0 - \mathbf{\Phi}_0^*) = \frac{1}{\sqrt{m}}(\mathbf{\Phi}_0 - \mathbf{0}) = \frac{1}{\sqrt{m}}\mathbf{\Phi}_0 = O_p(1)$. We have

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial\phi_r}\mathbf{Q}(\mathbf{\Phi}) = \frac{1}{\sqrt{m}}O_p(1) - \lambda_m\frac{\operatorname{sgn}(\beta_r)}{O_p(1)},$$

and, as $\lambda_m \to \infty$, the sign of d_r completely determines the (opposite) sign of $\mathbf{Q}(\mathbf{\Phi})$.

Therefore, with $\phi_r \in (\boldsymbol{\beta}', \mathbf{d}')$, the objective function changes sign around the true zero parameters at $\mathbf{0}$. The objective function is negative when $0 < \phi_r < \epsilon_m$ and positive when $-\epsilon_m < \phi_r < 0$. The maximum for truly zero parameters is found at $\mathbf{0}$. The zeroes of $\boldsymbol{\gamma}$ are determined by those of $\mathbf{d}$. $\square$

Theorem 3. As $\lambda_m \to \infty$ and $\frac{\lambda_m}{\sqrt{m}} \to 0$,

$$\sqrt{m}\mathcal{I}\left(\mathbf{\Phi}_{0^c}\right)\left(\left(\tilde{\mathbf{\Phi}_{0^c}} - \mathbf{\Phi}_{0^c}^*\right) - \frac{1}{m}\lambda_m \mathcal{I}\left(\mathbf{\Phi}_{0^c}\right)^{-1}\mathbf{h}\left(\tilde{\mathbf{\Phi}}_{0^c}^*\right)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}\left(\mathbf{\Phi}_{0^c}\right)\right),$$

with $\mathbf{h}\left(\tilde{\mathbf{\Phi}_{0^c}}\right)$ the vector consisting of the $p^* + q^*(q^*+1)/2$ (truly nonzero parameter) entries of the form $\zeta(\phi_r)\frac{\text{sgn}(\phi_r)}{\overline{\phi_r}}$ for $\phi_r \in \mathbf{\Phi}_{0^c}$, with $\zeta(\phi_r)$ the heredity weight applied to term ϕ_r.

Proof. The previous proofs have demonstrated that $\tilde{\mathbf{\Phi}_{0^c}}$, shown to be a local maximizer of $\mathbf{Q}\left(\mathbf{\Phi}_{0^c}\right)$, exists with $\|\mathbf{\Phi}_{0^c} - \mathbf{\Phi}_{0^c}^*\| = O_p(m^{-\frac{1}{2}})$ (Theorem 1). Further, the parameter set consisting of this maximizer for nonzero terms and $\mathbf{0}$ for the zero terms satisfies:

$$\frac{\partial}{\partial \mathbf{\Phi}_{0^c}}\mathbf{Q}\left(\mathbf{\Phi}\right)\Bigg|_{\mathbf{\Phi}=\begin{pmatrix}\tilde{\mathbf{\Phi}}_{0^c}\\ \mathbf{0}\end{pmatrix}} = \frac{\partial}{\partial \mathbf{\Phi}_{0^c}}\mathbf{l}\left(\mathbf{\Phi}\right)\Bigg|_{\mathbf{\Phi}=\begin{pmatrix}\tilde{\mathbf{\Phi}}_{0^c}\\ \mathbf{0}\end{pmatrix}} - \lambda_m \sum_{\phi_r \in \mathbf{\Phi}} \zeta(\phi_r)\frac{\text{sgn}(\phi_r)}{\overline{\phi_r}}.$$

Note that $\text{sgn}(\phi_r) = 0$ for $\phi_r \notin \mathbf{\Phi}_{0^c}$. Let $\mathbf{h}\left(\tilde{\mathbf{\Phi}_{0^c}}\right)$ be the vector consisting of the $p^* + q^*(q^*+1)/2$ (truly nonzero parameter) entries of the form $\zeta(\phi_r)\frac{\text{sgn}(\phi_r)}{\overline{\phi_r}}$ for $\phi_r \in \mathbf{\Phi}_{0^c}$.

From the Taylor series expansion of $\frac{\partial}{\partial \mathbf{\Phi}_{0^c}}\mathbf{Q}\left(\mathbf{\Phi}\right)$ around $\mathbf{\Phi}_{0^c}^*$,

$$\mathbf{\Phi}_{0^c} = l'\left(\mathbf{\Phi}_{0^c}^*\right) + \sum_{\phi_r \in \mathbf{\Phi}_{0^c}} \mathbf{l}''\left(\mathbf{\Phi}_{0^c}^*\right)\left(\tilde{\mathbf{\Phi}_{0^c}} - \mathbf{\Phi}_{0^c}^*\right) - \lambda_m \sum_{\phi_r \in \mathbf{\Phi}} \zeta(\phi_r)\frac{\text{sgn}(\phi_r^*)}{|\overline{\phi_r}|} = 0.$$

Therefore,

$$\frac{1}{m}l'\left(\boldsymbol{\Phi}_{0^c}^*\right)+\frac{1}{m}\sum_{\phi_r\in\boldsymbol{\Phi}_{0^c}}\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\left(\tilde{\boldsymbol{\Phi}_{0^c}}-\boldsymbol{\Phi}_{0^c}^*\right)-\frac{1}{m}\lambda_m\mathbf{h}\left(\tilde{\boldsymbol{\Phi}}_{0^c}^*\right)=0,$$

and

$$\frac{1}{\sqrt{m}}l'\left(\boldsymbol{\Phi}_{0^c}^*\right)=\sqrt{m}\left(\frac{1}{m}\sum_{\phi_r\in\boldsymbol{\Phi}_{0^c}}\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\left(\tilde{\boldsymbol{\Phi}}_{0^c}-\boldsymbol{\Phi}_{0^c}^*\right)-\frac{1}{m}\lambda_m\mathbf{h}\left(\tilde{\boldsymbol{\Phi}}_{0^c}^*\right)\right). \tag{3.19}$$

The score vector evaluated at the true parameter has $E\left(l'\left(\boldsymbol{\Phi}_{0^c}^*\right)\right)=\mathbf{0}$ and variance $\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)$, so that $\frac{1}{\sqrt{m}}\left(l'\left(\boldsymbol{\Phi}_{0^c}\right)\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)\right)$.

From 3.19 and with $-\frac{1}{m}\sum_{\phi_r\in\boldsymbol{\Phi}}\mathbf{l}''\left(\boldsymbol{\Phi}_{0^c}^*\right)\xrightarrow{p}\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)$,

$$\sqrt{m}\left(\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)\left(\tilde{\boldsymbol{\Phi}_{0^c}}-\boldsymbol{\Phi}_{0^c}^*\right)-\frac{1}{m}\lambda_m\mathbf{h}\left(\tilde{\boldsymbol{\Phi}}_{0^c}^*\right)\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)\right),$$

and

$$\sqrt{m}\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)\left(\left(\tilde{\boldsymbol{\Phi}_{0^c}}-\boldsymbol{\Phi}_{0^c}^*\right)-\frac{1}{m}\lambda_m\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)^{-1}\mathbf{h}\left(\tilde{\boldsymbol{\Phi}}_{0^c}^*\right)\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathcal{I}\left(\boldsymbol{\Phi}_{0^c}^*\right)\right).\square$$

3.4 Illustrations

The HM-ALASSO is illustrated in this section using simulated data and an application to a subset of the HS&B data. Simulated data give insight into true model recovery and comparative performance of estimation methods. However, it is important that the algorithm can handle real-data sample sizes and unbalanced groups. Computational challenges (e.g., lack of convergence) are common when data sets have many observations or when models have many parameters. Simulation studies for single-level heredity-constrained methods have primarily focused on models that were generated to adhere to heredity principles. We simulated data sets using several models that do not conform in order to create a more realistic illustration. This allows us also to distinguish whether the zero components are being found due to heredity or due to their truly zero natures. This means that model-data fit is not as optimally captured. Therefore, several selection procedures were evaluated to find one that selected the best model that did adhere to the principles and was still a reasonable model (as would be needed in practice).

Selection Criteria

Letting $\widehat{\boldsymbol{\phi}} = \left(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{d}}\right)$ be those components of the parameter estimate set, Bondell et al. counted the degrees of freedom as

$$\mathrm{df} = \sum_{r=1}^{p} I\left(\widehat{\beta}_r \neq 0\right) + \frac{1}{2}\sum_{s=1}^{q} I\left(\widehat{d}_s \neq 0\right)\left(\sum_{s=1}^{q} I\left(\widehat{d}_s \neq 0\right) + 1\right),$$

in which $I(\cdot)$ represents the indicator function. Common selection criteria include those of the forms

$$\text{BIC} = -2l(\widehat{\boldsymbol{\phi}}) + \ln N * \text{df}, \tag{3.20}$$

$$\text{AIC} = -2l(\widehat{\boldsymbol{\phi}}) + 2 * \text{df},$$

and

$$\text{AIC}c = -2l(\widehat{\boldsymbol{\phi}}) + 2 * \text{df} + \frac{(2 * \text{df} * (\text{df} + 1))}{(N - \text{df} - 1)}.$$

The AICc, a version of AIC that is corrected for small sample sizes, has been suggested as a default over AIC. This is because it mitigates issues that occur when the sample size is small or the ratio of sample size to number of parameters is small. Further, AICc produces selection behavior similar to AIC when sample sizes are larger. The sample sizes in use for this study result in very close AICc and AIC values. Note that these definitions do not count the level-one variance parameter, which is assumed to be non-zero in all candidate models.

Values of these criteria are often very similar between models. For comparison purposes, for each criterion the differences between values for candidate models (Δ_l for the lth candidate model) are of interest. Difference between the criterion values of the model with minimum value and another candidate model are calculated. A difference of at most 2 means that the support for selecting either of the two models is essentially similar (Burnham & Anderson, 2004). Differences between 4 and 7 indicate some support for choosing the alternative model. Differences greater than 10 indicate that there is little support for choosing the alternative model. Bondell *et al.* (2010) and Schelldorfer *et al.* (2011) used (3.20) to select the fractional threshold that produces the best set of regularized parameter estimates. In addition to the likelihood-based criteria, CV1 PMSE estimate was compared in the current study.

These methods may identify the model or models that are the best fit to the data. However, they do not take into account the heredity structure. The applicaton of heredity rules is based on principles of model building. This may be in conflict with traditional model-fit statistics. Applying heredity principles may lead to some model terms that contribute significantly to model-data fit being disallowed. Therefore, it is inappropriate to base heredity-focused criteria on formulations that include those terms. Because adherence to the heredity structure is the focus of this dissertation, an additional set of selection criteria were evaluated. These criteria attempt to select models with reasonable fit, no excessive dimensionality, and no hierarchical gap. The criteria are based on the BIC, AIC, and AICc criteria. Instead of using the candidate model likelihood, they use the likelihood based only on the terms in the candidate model that would be allowed under heredity. The full dimensionality penalty (df) given by the active set was applied to take into account all terms that were estimated by the procedure. The criteria calculated from the reduced model are labeled BICr, AICr, and AICcr.

To select the best criterion, a measure of the Kullback-Leibler divergence (KLD) was evaluated. For simulated data, the true data-generating model is known. In that case, the (non-negative and non-symmetric) distance between the true model and each estimated model is typically evaluated. In the current study, the true generating model did not necessarily adhere to the principles of strong heredity. Therefore, we also estimated a heredity-adjusted true model (called the "heredity" model). The heredity model parameters were found by estimating the parameters using only the terms that would be allowed in the true model under the principles of strong heredity. We evaluated each candidate model compared with the true and heredity model values. The former value provides the information lost for each estimated model based on model-data fit with respect to the true model. The latter value provides a measure of information lost for each estimated model in terms of the information that would have been available in the heredity model. Our KLD measure is an average over J groups of the multivariate Gaussian distribution of the outcome in each group, where the distribution is described in (2.32) and (2.33). In the equation for KLD, subscript $j : (t, h)$ denotes a term in the jth group under the true or heredity model. Subscript $j : e$ denotes a term in the jth group for the candidate model. The KLD for the jth group is defined as

$$\mathrm{KLD}_{(t,h):e} = \frac{1}{2}\left(\mathrm{tr}(\mathbf{V}_{j:e}^{-1}\mathbf{V}_{j:(t,h)}) + (\boldsymbol{\mu}_{j:e} - \boldsymbol{\mu}_{j:(t,h)})'\mathbf{V}_{j:e}^{-1}(\boldsymbol{\mu}_{j:e} - \boldsymbol{\mu}_{j:(t,h)}) - \ln\frac{|\mathbf{V}_{j:(t,h)}|}{|\mathbf{V}_{j:e}|} - n_j\right).$$

3.4.1 Simulation Study

Simulations were evaluated for parameter recovery and model error. The model was described in (3.1) and (3.2). Five structural cases were simulated with predictors modeled following a standard approach Busing & te Leiden (1993). Level-one and level-two covariates are independently $\mathcal{N}(0,1)$. Random effects are simulated from $\mathcal{N}_q(0,\mathbf{\Sigma})$. Level-one errors are distributed as $N(0, 0.0625)$. In each case, we compare the HM-ALASSO to the M-ALASSO. The consistency of the M-ALASSO has been demonstrated theoretically. M-ALASSO has performed very well empirically for multilevel models. We focused on the evaluation of whether HM-ALASSO could improve upon the M-ALASSO in terms of addressing the heredity structure of the model.

Three cases varied according to which controlling effects were simulated to be zero: 1) the intercept, 2) another level-one effect, 3) a level-two effect. The fourth case involved a null random-effect variance component. The final case consisted of data that were simulated using a true model that does adhere to strong heredity principles. This was done in order to provide a direct measure between methods.

In the literature on multilevel regularization reviewed in this dissertation, no simulation conditions exceeded 1500 individuals. The simulation data sets generated for the current study included 20 individuals in each group and 150 groups (3000 total individuals). This somewhat larger sample size was employed to better ensure the power to detect cross-level interaction effects Mathieu *et al.* (2012). There were 50 iterations run for each case. Heredity model parameters were obtained using `lmer` with only the terms that would be nonzero in each case. Then, 50 instances of each model were simulated from the resulting parameters. The average parameter estimates were used for comparison to our candidate models. The true and heredity model parameters for the five cases are given in Table 3.4. Note that the heredity parameters are means of the estimates over 50 simulations. All standard errors were 0 up to 2 or more decimal points and are not reported.

The error in the parameter sets was evaluated using the mean square error for the fixed components as

$$ME(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

A quadratic loss measures were evaluated for all random-effects variance ma-

Table 3.4: Simulation Conditions.

Case	Statistic	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	$\mathbf{\Sigma}_{11}$	$\mathbf{\Sigma}_{12}$	$\mathbf{\Sigma}_{13}$	$\mathbf{\Sigma}_{22}$	$\mathbf{\Sigma}_{23}$	$\mathbf{\Sigma}_{33}$
1	True	0.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.20	0.04	1.00	0.20	1.00
	Heredity	0.00	1.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	True	1.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.20	0.04	1.00	0.20	1.00
	Heredity	1.05	0.97	0.00	1.12	0.00	0.99	0.00	0.00	0.00	0.97	0.21	0.00	1.12	0.00	0.00
3	True	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.20	0.04	1.00	0.20	1.00
	Heredity	1.03	0.93	1.05	0.00	1.05	0.00	1.10	0.00	1.15	0.96	0.26	0.19	2.04	1.08	1.82
4	True	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.20	0.00	1.00	0.20	0.00
	Heredity	0.88	0.00	0.00	1.01	0.84	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.00	0.00
5	True	1.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.20	0.00	1.00	0.20	0.00
	Heredity	1.08	1.01	0.00	0.88	0.00	1.07	0.00	0.00	0.00	1.03	0.26	0.00	0.96	0.00	0.00

trix parameters as

$$ME1(\widehat{\boldsymbol{\Sigma}}) = \operatorname{tr}\left(\left(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right)\left(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right)'\right).$$

A quadratic loss measure was evaluated for the variances only as

$$ME2(\widehat{\boldsymbol{\Sigma}}) = \left((\operatorname{diag}(\boldsymbol{\Sigma})) - \operatorname{diag}\left(\widehat{\boldsymbol{\Sigma}}\right)\right)'\left((\operatorname{diag}(\boldsymbol{\Sigma})) - \operatorname{diag}\left(\widehat{\boldsymbol{\Sigma}}\right)\right).$$

For each case, the means and standard errors of the KLD, hierarchical gaps, parameter estimates, and parameter discrepancies for the models chosen using the various selection criteria are reported.

The group structure may lead to a loss of information due to clustering. Therefore, the BIC, BICr, AICc, and AICcr criteria were also calculated using the effective sample size as given in (2.12). The actual sample size used in estimation of the model was 2850 in each case. This was due to the leave-one-out (per group, with 150 groups) cross-validation. The effective sample sizes ranged from 284 for the model with a null random variance component to 589 for the model in which a level-two predictor was excluded. These effective sample sizes were still reasonably large for the number of parameters. In the cases examined in this study, the results for the effective sample sizes matched those of the total samples and are not reported separately.

The descriptive statistics of overall model criteria for the chosen iterations

Table 3.5: Simulation Study: Case 1 Criteria.

Method	Tuning	Statistic	KLD$_t$	KLD$_h$	Hierarchical Gap
HM-ALASSO	AIC	Mean	15.05	42.98	0.00
		SE	0.51	1.12	0.00
HM-ALASSO	AICc	Mean	15.05	42.98	0.00
		SE	0.51	1.12	0.00
HM-ALASSO	BIC	Mean	15.07	42.89	0.00
		SE	0.52	1.12	0.00
HM-ALASSO	PMSE	Mean	16.84	47.19	0.00
		SE	0.16	0.55	0.00
M-ALASSO	AIC	Mean	16.60	46.64	1.67
		SE	0.13	0.66	0.26
M-ALASSO	AICc	Mean	16.60	46.64	1.67
		SE	0.13	0.66	0.26
M-ALASSO	BIC	Mean	16.52	46.02	2.37
		SE	0.13	0.64	0.26
M-ALASSO	PMSE	Mean	21.79	32.68	3.08
		SE	0.94	2.15	0.28

from each case are displayed in Tables 3.5, 3.6, 3.7, 3.8, and 3.9. It should be clear from all cases that using the AIC, AICc, and AIC led to the choice of similar models. The same was true for the heredity-adjusted criteria AICr, AICcr, and BICr. There was no noticeable difference between the two sets of criteria. Therefore, parameter-related results are reported only for the models chosen using BIC and PMSE. Results are reported in Tables 3.10, 3.11, 3.12 and 3.13, 3.14, 3.15. The HM-ALASSO provided comparable parameter estimates in all cases and also resulted in similar or smaller model errors.

Observations based on the simulation results can be summarized as follows. First the methods and selection criteria were evaluated based on KLD_t, KLD_h, and hierarchical gap.

- Case 1:The largest differences in KLD between estimation methods were

Table 3.6: Simulation Study: Case 2 Criteria.

Method	Tuning	Statistic	KLD_t	KLD_h	Hierarchical Gap
HM-ALASSO	AIC	Mean	17.03	22.46	0.02
		SE	0.36	0.52	0.02
HM-ALASSO	AICc	Mean	17.03	22.46	0.02
		SE	0.36	0.52	0.02
HM-ALASSO	BIC	Mean	16.94	22.33	0.02
		SE	0.40	0.56	0.02
HM-ALASSO	PMSE	Mean	17.75	22.58	0.04
		SE	0.27	0.71	0.04
M-ALASSO	AIC	Mean	16.73	24.28	0.85
		SE	0.11	0.20	0.13
M-ALASSO	AICc	Mean	16.73	24.28	0.85
		SE	0.11	0.20	0.13
M-ALASSO	BIC	Mean	16.80	24.40	1.10
		SE	0.12	0.23	0.16
M-ALASSO	PMSE	Mean	20.45	23.68	2.14
		SE	0.75	0.37	0.21

Table 3.7: Simulation Study: Case 3 Criteria.

Method	Tuning	Statistic	KLD_t	KLD_h	Hierarchical Gap
HM-ALASSO	AIC	Mean	17.81	30.79	0.04
		SE	0.32	1.00	0.04
HM-ALASSO	AICc	Mean	17.81	30.79	0.04
		SE	0.32	1.00	0.04
HM-ALASSO	BIC	Mean	17.81	30.79	0.04
		SE	0.32	1.00	0.04
HM-ALASSO	PMSE	Mean	22.09	24.57	0.04
		SE	0.65	0.94	0.04
M-ALASSO	AIC	Mean	16.71	33.53	0.00
		SE	0.11	0.41	0.00
M-ALASSO	AICc	Mean	16.71	33.53	0.00
		SE	0.11	0.41	0.00
M-ALASSO	BIC	Mean	16.71	33.53	0.00
		SE	0.11	0.41	0.00
M-ALASSO	PMSE	Mean	21.93	31.65	1.55
		SE	0.86	0.61	0.19

Table 3.8: Simulation Study: Case 4 Criteria.

Method	Tuning	Statistic	KLD_t	KLD_h	Hierarchical Gap
HM-ALASSO	AIC	Mean	16.84	151.26	0.00
		SE	0.12	0.62	0.00
HM-ALASSO	AICc	Mean	16.84	151.26	0.00
		SE	0.12	0.62	0.00
HM-ALASSO	BIC	Mean	16.78	150.87	0.00
		SE	0.12	0.59	0.00
HM-ALASSO	PMSE	Mean	16.52	148.05	0.04
		SE	0.21	1.47	0.04
M-ALASSO	AIC	Mean	16.58	149.95	1.56
		SE	0.07	0.42	0.20
M-ALASSO	AICc	Mean	16.58	149.95	1.56
		SE	0.07	0.42	0.20
M-ALASSO	BIC	Mean	16.54	149.81	1.88
		SE	0.09	0.49	0.20
M-ALASSO	PMSE	Mean	19.72	137.44	1.88
		SE	0.62	2.59	0.21

Table 3.9: Simulation Study: Case 5 Criteria.

Method	Tuning	Statistic	KLD_t	KLD_h	Hierarchical Gap
HM-ALASSO	AIC	Mean	123.51	16.79	0.14
		SE	0.51	0.11	0.05
HM-ALASSO	AICc	Mean	123.42	16.77	0.15
		SE	0.51	0.11	0.05
HM-ALASSO	BIC	Mean	123.29	16.74	0.25
		SE	0.53	0.12	0.07
HM-ALASSO	PMSE	Mean	122.67	18.98	0.10
		SE	0.69	0.58	0.06
M-ALASSO	AIC	Mean	122.76	16.73	0.86
		SE	0.48	0.11	0.09
M-ALASSO	AICc	Mean	122.76	16.73	0.86
		SE	0.48	0.11	0.09
M-ALASSO	BIC	Mean	122.67	16.70	0.86
		SE	0.51	0.12	0.09
M-ALASSO	PMSE	Mean	118.55	21.01	0.74
		SE	1.27	0.91	0.13

Table 3.10: Simulation Study: Cases 1 and 2 Fixed Effects and Discrepancies.

Case	Method	Tuning	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\beta_9}$	$ME_t(\widehat{\boldsymbol{\beta}})$	$ME_h(\widehat{\boldsymbol{\beta}})$
1	HM-ALASSO	BIC	Mean	−0.01	1.00	0.01	1.00	1.00	1.00	0.01	0.01	0.99	0.05	4.05
1			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.05
1		PMSE	Mean	−0.02	1.00	0.00	0.98	1.00	0.97	0.01	0.01	0.97	0.09	3.94
1			SE	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.07
1	M-ALASSO	BIC	Mean	0.01	0.98	0.01	1.00	1.02	0.98	−0.01	−0.01	0.99	0.05	4.05
1			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.05
1		PMSE	Mean	0.00	0.60	−0.01	0.58	0.65	0.57	0.01	0.00	0.66	1.70	2.73
1			SE	0.01	0.06	0.01	0.07	0.06	0.06	0.00	0.01	0.05	0.28	0.20
2	HM-ALASSO	BIC	Mean	1.01	1.01	0.00	0.98	0.01	0.99	−0.01	0.00	0.80	0.22	0.85
2			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.05	0.05	0.05
2		PMSE	Mean	1.01	1.01	0.00	0.98	0.01	0.99	−0.01	0.00	0.73	0.29	0.77
2			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.06	0.06	0.06
2	M-ALASSO	BIC	Mean	0.99	1.00	−0.02	0.98	0.03	0.99	0.02	0.02	0.98	0.06	1.05
2			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.02
2		PMSE	Mean	0.60	0.60	−0.01	0.60	0.01	0.63	0.00	0.00	0.65	1.62	2.05
2			SE	0.06	0.06	0.01	0.06	0.01	0.06	0.00	0.00	0.05	0.29	0.20

Table 3.11: Simulation Study: Cases 3 and 4 Fixed Effects and Discrepancies.

Case	Method	Tuning	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\beta_9}$	$ME_t(\widehat{\boldsymbol{\beta}})$	$ME_h(\widehat{\boldsymbol{\beta}})$
3	HM-ALASSO	BIC	Mean	0.99	0.99	0.99	0.00	1.01	0.82	0.97	0.83	0.99	0.40	1.75
3			SE	0.01	0.01	0.01	0.01	0.01	0.05	0.01	0.05	0.01	0.11	0.11
3		PMSE	Mean	0.83	0.71	0.74	0.00	0.83	0.33	0.66	0.32	0.72	2.41	1.93
3			SE	0.04	0.06	0.05	0.01	0.04	0.06	0.06	0.06	0.05	0.32	0.23
3	M-ALASSO	BIC	Mean	1.02	1.01	1.01	−0.02	0.98	0.99	0.98	0.99	1.00	0.07	2.07
3			SE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.04
3		PMSE	Mean	0.63	0.58	0.61	0.00	0.59	0.58	0.54	0.57	0.56	2.85	3.46
3			SE	0.06	0.06	0.06	0.01	0.06	0.06	0.06	0.06	0.06	0.45	0.27
4	HM-ALASSO	BIC	Mean	0.98	−0.01	0.00	1.00	1.01	0.99	0.01	0.00	1.00	0.04	2.05
4			SE	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02
4		PMSE	Mean	0.98	0.00	0.00	0.99	1.00	0.98	0.00	0.00	1.00	0.04	2.05
4			SE	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.03
4	M-ALASSO	BIC	Mean	0.99	−0.01	0.00	1.01	0.98	0.98	0.00	0.00	1.00	0.04	2.03
4			SE	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02
4		PMSE	Mean	0.68	0.00	0.00	0.67	0.65	0.67	−0.01	0.00	1.00	1.14	2.33
4			SE	0.06	0.01	0.00	0.06	0.06	0.06	0.01	0.00	0.00	0.22	0.09

Table 3.12: HM-ALASSO Simulation Study: Case 5 Fixed Effects and Discrepancies.

Case	Method	Tuning	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\beta_9}$	$ME_t(\widehat{\boldsymbol{\beta}})$	$ME_h(\widehat{\boldsymbol{\beta}})$
5	HM-ALASSO	BIC	Mean	0.98	0.99	0.00	0.98	−0.01	0.99	0.00	0.00	0.00	0.03	0.06
5			SE	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01
5		PMSE	Mean	0.82	0.85	0.00	0.80	0.01	0.80	0.00	0.00	0.00	0.47	0.51
5			SE	0.03	0.03	0.00	0.04	0.01	0.04	0.01	0.00	0.00	0.10	0.11
5	M-ALASSO	BIC	Mean	0.98	0.98	0.00	1.01	−0.02	0.99	0.00	0.00	0.00	0.04	0.07
5			SE	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01
5		PMSE	Mean	0.74	0.73	0.00	0.76	−0.01	0.74	−0.01	0.00	0.00	0.89	0.94
5			SE	0.05	0.05	0.00	0.05	0.01	0.05	0.01	0.00	0.00	0.19	0.19

Table 3.13: Simulation Study: Cases 1 and 2 Random-effects Covariance Matrix Terms and Discrepancies.

Case	Method	Tuning	Stat	$\widehat{\boldsymbol{\Sigma}}_{11}$	$\widehat{\boldsymbol{\Sigma}}_{12}$	$\widehat{\boldsymbol{\Sigma}}_{13}$	$\widehat{\boldsymbol{\Sigma}}_{22}$	$\widehat{\boldsymbol{\Sigma}}_{23}$	$\widehat{\boldsymbol{\Sigma}}_{33}$	$ME1_t(\widehat{\boldsymbol{\Sigma}})$	$ME1_h(\widehat{\boldsymbol{\Sigma}})$	$ME2_t(\widehat{\boldsymbol{\Sigma}})$	$ME2_t(\widehat{\boldsymbol{\Sigma}})$
1	HM-ALASSO	BIC	Mean	0.95	0.21	−0.02	0.96	0.24	1.01	0.64	3.69	0.30	3.15
1			SE	0.03	0.01	0.04	0.03	0.04	0.06	0.33	0.45	0.14	0.26
1		PMSE	Mean	0.93	0.19	0.04	0.94	0.19	0.92	0.14	2.85	0.09	2.66
1			SE	0.02	0.01	0.01	0.02	0.01	0.03	0.02	0.09	0.02	0.08
1	M-ALASSO	BIC	Mean	0.98	0.20	0.06	0.97	0.20	1.00	0.08	3.14	0.04	2.94
1			SE	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.06	0.00	0.05
1		PMSE	Mean	0.98	0.20	0.05	0.96	0.18	0.89	0.17	2.98	0.14	2.79
1			SE	0.03	0.01	0.01	0.03	0.01	0.02	0.02	0.13	0.02	0.12
2	HM-ALASSO	BIC	Mean	0.99	0.22	0.04	0.96	0.20	1.12	0.31	1.66	0.25	1.52
2			SE	0.02	0.01	0.01	0.02	0.01	0.06	0.07	0.17	0.06	0.16
2		PMSE	Mean	0.96	0.20	0.04	0.93	0.17	1.07	0.29	1.53	0.24	1.42
2			SE	0.02	0.01	0.01	0.02	0.01	0.06	0.06	0.15	0.06	0.15
2	M-ALASSO	BIC	Mean	0.97	0.19	0.03	0.99	0.19	0.98	0.08	1.12	0.04	1.00
2			SE	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.03	0.00	0.03
2		PMSE	Mean	0.90	0.18	0.04	0.94	0.18	0.84	0.19	0.95	0.15	0.84
2			SE	0.03	0.01	0.01	0.03	0.01	0.03	0.03	0.04	0.03	0.03

Table 3.14: Simulation Study: Cases 3 and 4 Random-effects Covariance Matrix Terms and Discrepancies.

Case	Method	Tuning	Stat	$\widehat{\mathbf{\Sigma}}_{11}$	$\widehat{\mathbf{\Sigma}}_{12}$	$\widehat{\mathbf{\Sigma}}_{13}$	$\widehat{\mathbf{\Sigma}}_{22}$	$\widehat{\mathbf{\Sigma}}_{23}$	$\widehat{\mathbf{\Sigma}}_{33}$	$ME1_t(\widehat{\mathbf{\Sigma}})$	$ME1_h(\widehat{\mathbf{\Sigma}})$	$ME2_t(\widehat{\mathbf{\Sigma}})$	$ME2_t(\widehat{\mathbf{\Sigma}})$
3	HM-ALASSO	BIC	Mean	0.96	0.22	0.05	1.16	0.27	1.14	0.43	2.98	0.33	1.53
3			SE	0.01	0.01	0.01	0.06	0.02	0.05	0.11	0.16	0.10	0.10
3		PMSE	Mean	0.96	0.25	0.07	1.50	0.37	1.48	1.07	1.97	0.89	0.81
3			SE	0.02	0.02	0.02	0.07	0.02	0.06	0.14	0.17	0.12	0.11
3	M-ALASSO	BIC	Mean	0.99	0.21	0.04	1.00	0.19	0.98	0.08	3.51	0.04	1.83
3			SE	0.02	0.01	0.01	0.02	0.01	0.02	0.01	0.07	0.01	0.04
3		PMSE	Mean	0.94	0.21	0.05	1.02	0.20	1.02	0.20	3.49	0.15	1.83
3			SE	0.02	0.01	0.01	0.03	0.01	0.03	0.03	0.15	0.03	0.12
4	HM-ALASSO	BIC	Mean	0.96	0.18	0.00	0.97	0.00	0.00	0.05	1.10	0.03	1.02
4			SE	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.03
4		PMSE	Mean	0.92	0.21	0.00	0.91	0.00	0.00	0.27	1.20	0.15	0.99
4			SE	0.04	0.04	0.00	0.04	0.00	0.00	0.16	0.24	0.06	0.11
4	M-ALASSO	BIC	Mean	1.00	0.20	0.00	0.95	0.00	0.00	0.05	1.10	0.04	1.01
4			SE	0.02	0.01	0.00	0.02	0.00	0.00	0.01	0.04	0.00	0.03
4		PMSE	Mean	0.98	0.19	0.00	0.84	0.00	0.00	0.12	0.92	0.10	0.83
4			SE	0.03	0.01	0.00	0.03	0.00	0.00	0.02	0.05	0.02	0.05

Table 3.15: Simulation Study: Case 5 Random-effects Covariance Matrix Terms and Discrepancies.

Case	Method	Tuning	Stat	$\widehat{\mathbf{\Sigma}}_{11}$	$\widehat{\mathbf{\Sigma}}_{12}$	$\widehat{\mathbf{\Sigma}}_{13}$	$\widehat{\mathbf{\Sigma}}_{22}$	$\widehat{\mathbf{\Sigma}}_{23}$	$\widehat{\mathbf{\Sigma}}_{33}$	$ME1_t(\widehat{\mathbf{\Sigma}})$	$ME1_h(\widehat{\mathbf{\Sigma}})$	$ME2_t(\widehat{\mathbf{\Sigma}})$	$ME2_t(\widehat{\mathbf{\Sigma}})$
5	HM-ALASSO	BIC	Mean	0.96	0.19	0.00	0.99	0.00	0.00	0.04	0.06	0.03	0.03
5			SE	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00
5		PMSE	Mean	0.96	0.20	0.00	0.96	0.00	0.00	0.06	0.07	0.05	0.05
5			SE	0.02	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.01	0.01
5	M-ALASSO	BIC	Mean	0.97	0.19	0.00	0.99	0.00	0.00	0.05	0.06	0.03	0.03
5			SE	0.02	0.01	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00
5		PMSE	Mean	0.90	0.17	0.00	0.90	0.00	0.00	0.11	0.12	0.09	0.10
5			SE	0.02	0.01	0.00	0.03	0.00	0.00	0.02	0.02	0.02	0.02

found when PMSE was used as a criterion. HM-ALASSO produced lower average KLD_t and higher average KLD_h. For all other criteria, the HM-ALASSO produced either similar or smaller (particularly with respect to the heredity model) KLD values.

- Case 2: HM-ALASSO always produced similar or smaller (particularly with respect to the heredity model) KLD values for all criteria.

- Case 3: HM-ALASSO produced smaller KLD_h values for all criteria. KLD_t values for all likelihood-based criteria were slightly less favorable for HM-ALASSO (mean=17.82, SE=0.32) than the M-ALASSO (mean=16.7, SE=0.11).

- Case 4: HM-ALASSO outperformed M-ALASSO in terms of KLD_t when PMSE was used. The methods produced similar KLD_t results for all other criteria and similar KLD_h for BIC. M-ALASSO performed better than HM-ALASSO in terms of KLD_h when the other likelihood criteria or PMSE were used.

- Case 5: HM-ALASSO outperformed M-ALASSO with respect to KLD_h when PMSE was used for selection. The reverse was true with respect to KLD_t. For all other criteria in both model comparisons, the methods performed similarly.

- The average hierarchical gap was demonstrably smaller in all cases for

HM-ALASSO than for M-ALASSO.

- Evaluating only candidate models for which the hierarchical gap was closed, HM-ALASSO always had lower KLD_h than M-ALASSO. Only Case 1 with PMSE and Case 2 with all criteria led to lower KLD_t for M-ALASSO.

After examining KLD results for all criteria, BIC and PMSE were chosen for candidate-model selection.

- For candidate models with the same non-zero terms, all individual fixed-effect and random-effects covariance parameter estimates were comparable for both methods. This was true using either selection criterion.

- $ME(\widehat{\boldsymbol{\beta}})$: HM-ALASSO performed similarly to, or better than, M-ALASSO in all cases and for both criteria, with the exceptions of Case 1 with PMSE in terms of KLD_h and Cases 2 and 3 for BIC with respect to KLD_t. BIC led to smaller model errors in all cases and for both methods, with the exception of case 1 for both methods when evaluating KLD_h.

- $ME1(\widehat{\boldsymbol{\Sigma}})$, $ME2(\widehat{\boldsymbol{\Sigma}})$: No clear patterns were evident with respect to method, criterion, or to which model the comparison was being made. However, while all errors were small, there was slightly better performance overall for M-ALASSO.

The simulation study provides evidence that supports the claim that HM-ALASSO provides parameter estimates that are comparable to those from an estimator that has proven to be theoretically and empirically sound. Further, HM-ALASSO directly addresses the heredity requirements in Table 3.1. However, variation in results between the structural cases in the simulation studies indicate that the heredity model structure has a non-negligible effect on the challenges of constrained estimation.

Figures 3.1, 3.2, 3.3, 3.4, and 3.5 show all fixed-effect estimates for each of the five cases for HM-ALASSO and M-ALASSO, respectively. For the level-one fixed-effect parameters β_1, β_2, and β_3, the symbol size is proportional to the size of each effect's associated random-variance component. The Figures show an overall trend of a more monotone shrinkage in the M-ALASSO estimates across fractional thresholds as would be expected due to the heredity control that HM-ALASSO invokes in addition to shrinkage. It should be noted that each fractional threshold starts anew with the full GLS estimates bounded by a stricter fraction of the penalty term. This is, in part, because the `quadprog` program does not allow for the specification of initial estimates. The process could be optimized through the use of a more robust quadratic programming that does provide this option.

Figures 3.6, 3.7, 3.8, 3.9, and 3.10 display importance ratios for each of the five cases for HM-ALASSO and M-ALASSO, respectively. In these Figures,

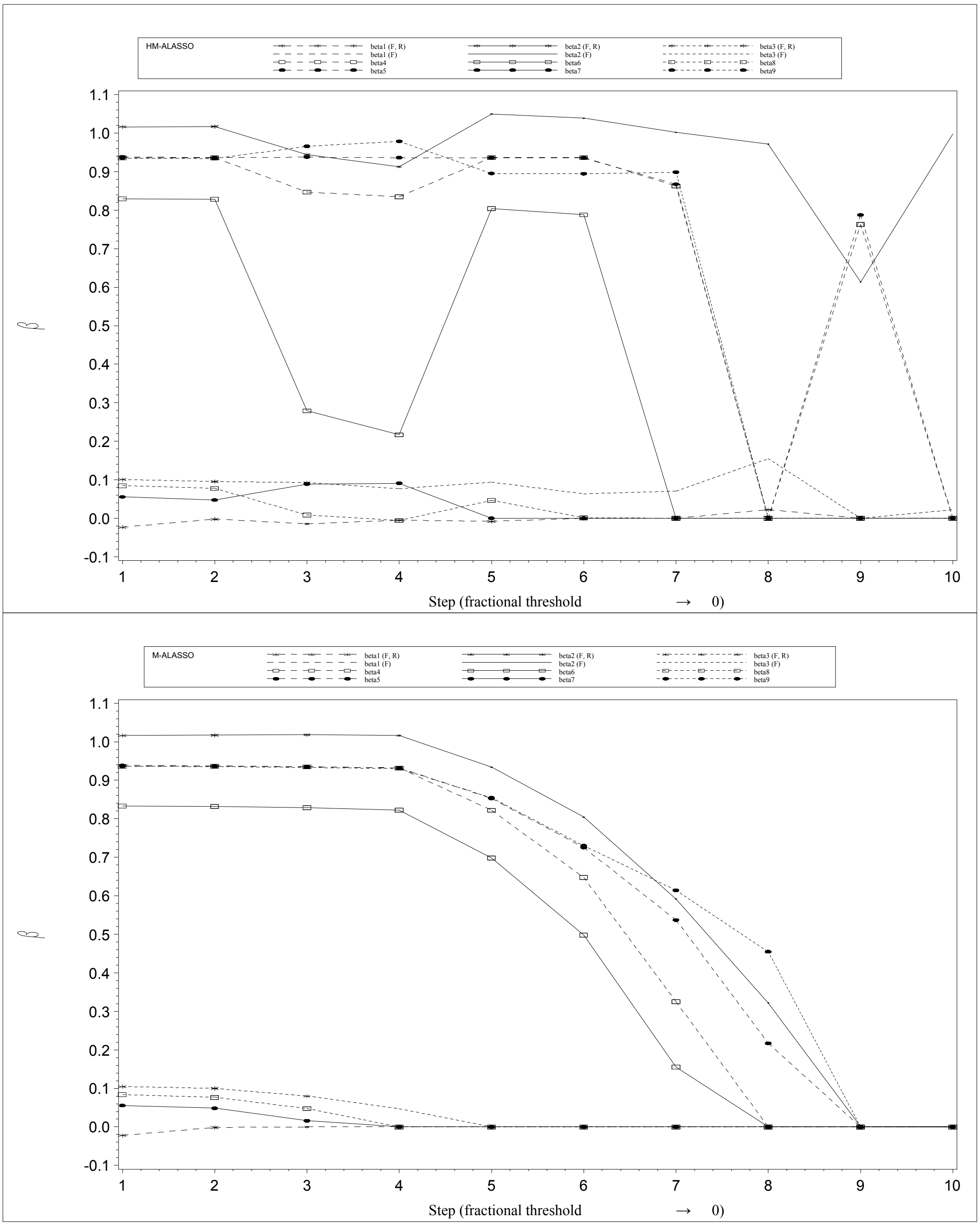

Figure 3.1: Coefficient Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 1

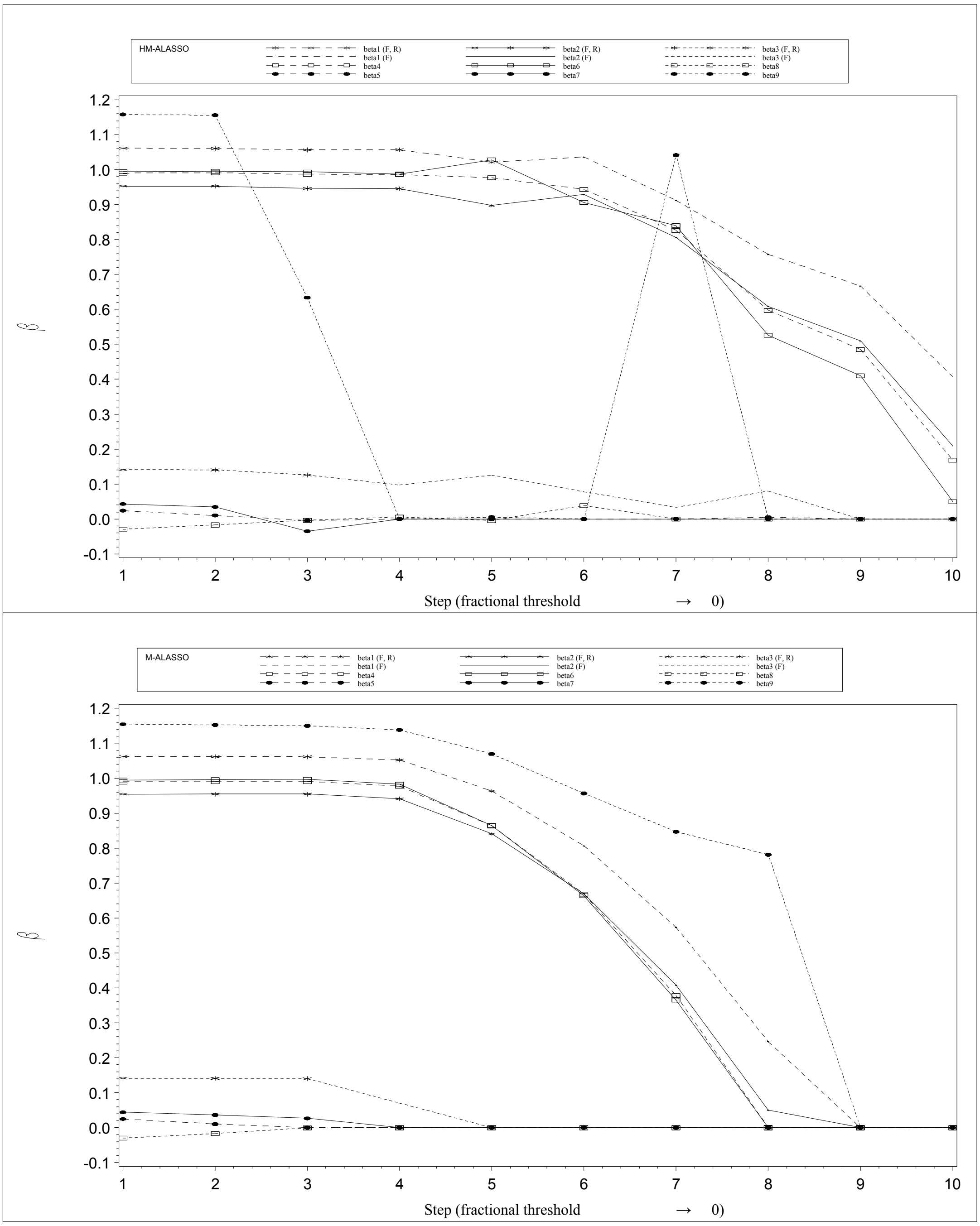

Figure 3.2: Coefficient Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 2

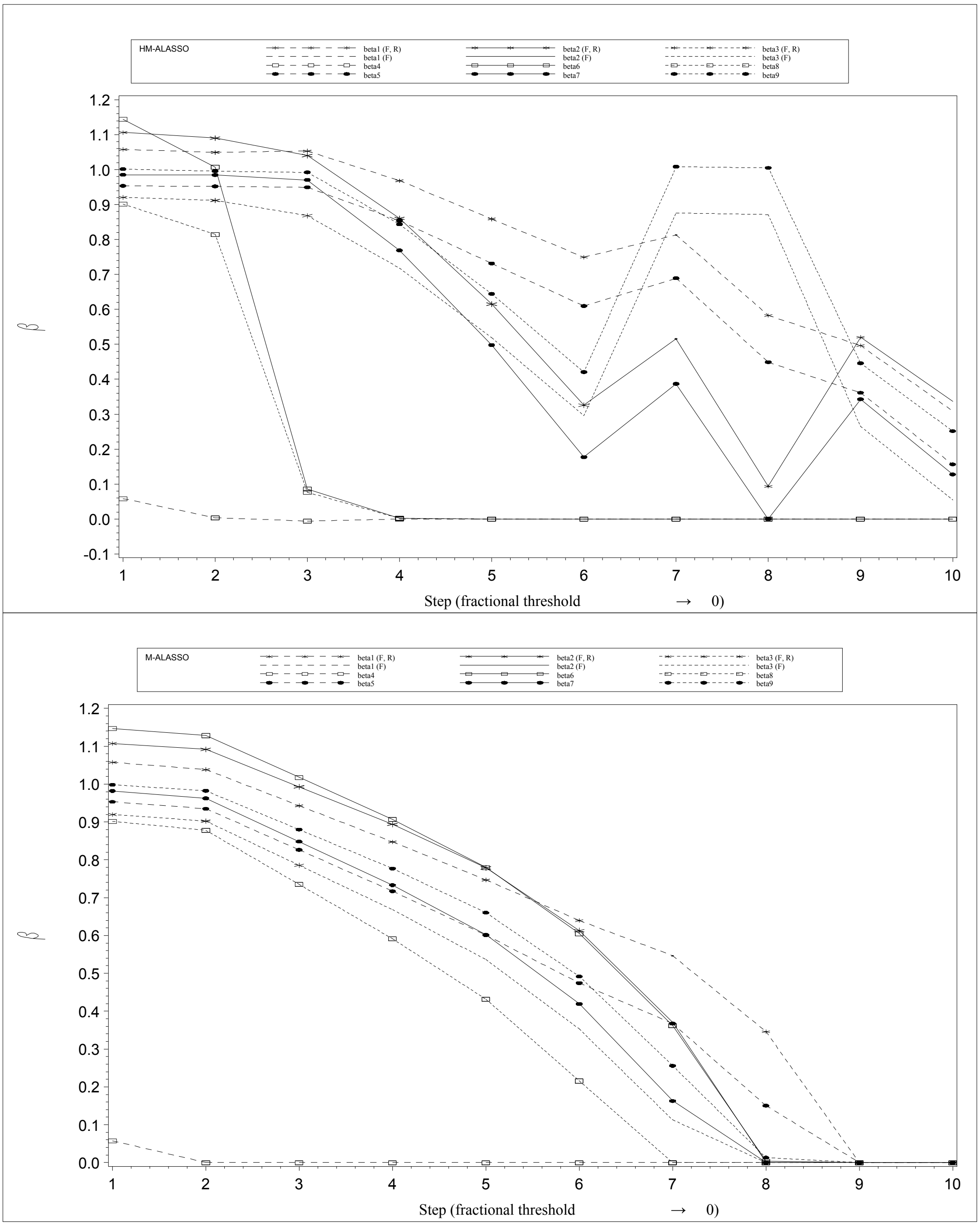

Figure 3.3: Coefficient Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 3

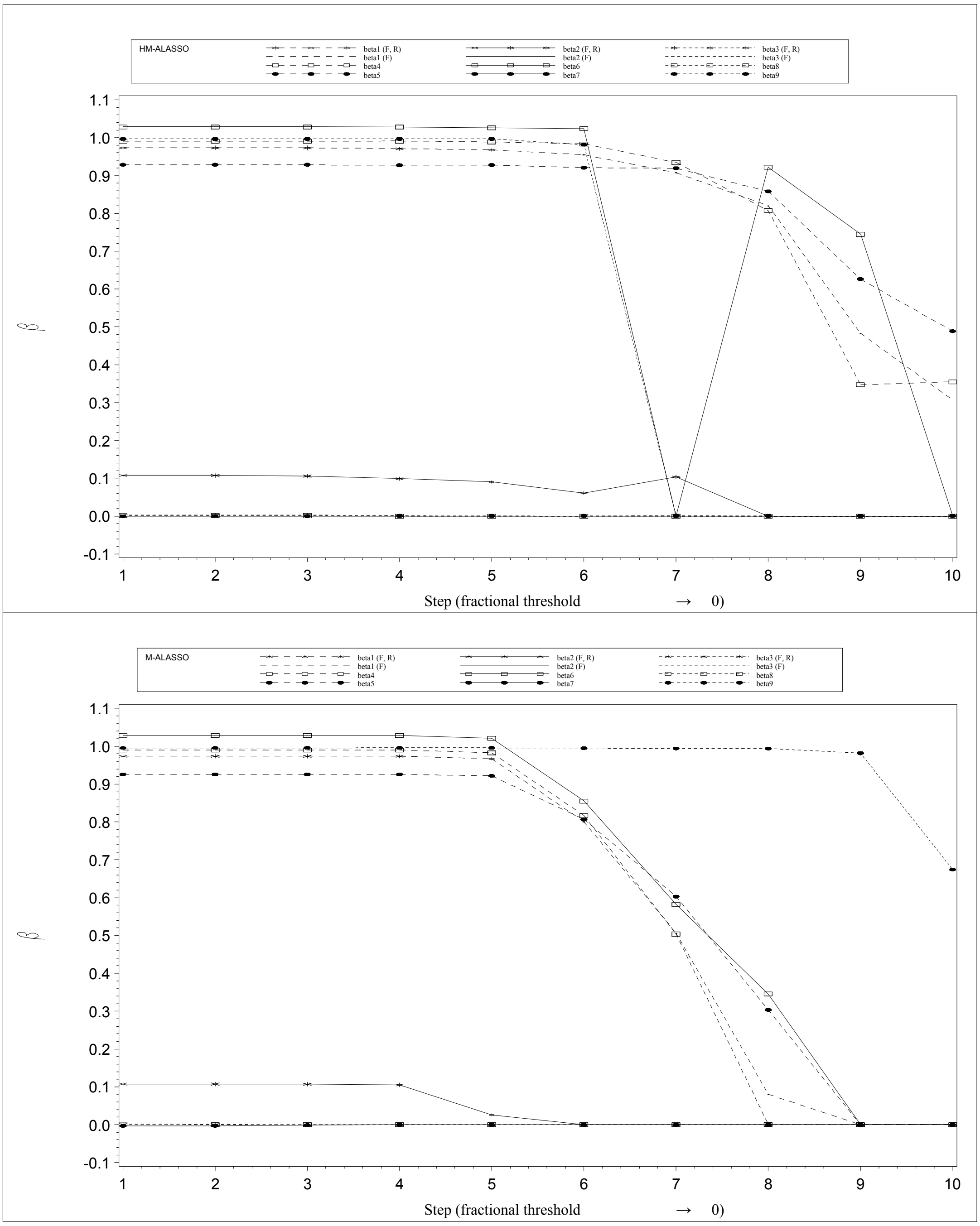

Figure 3.4: Coefficient Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 4

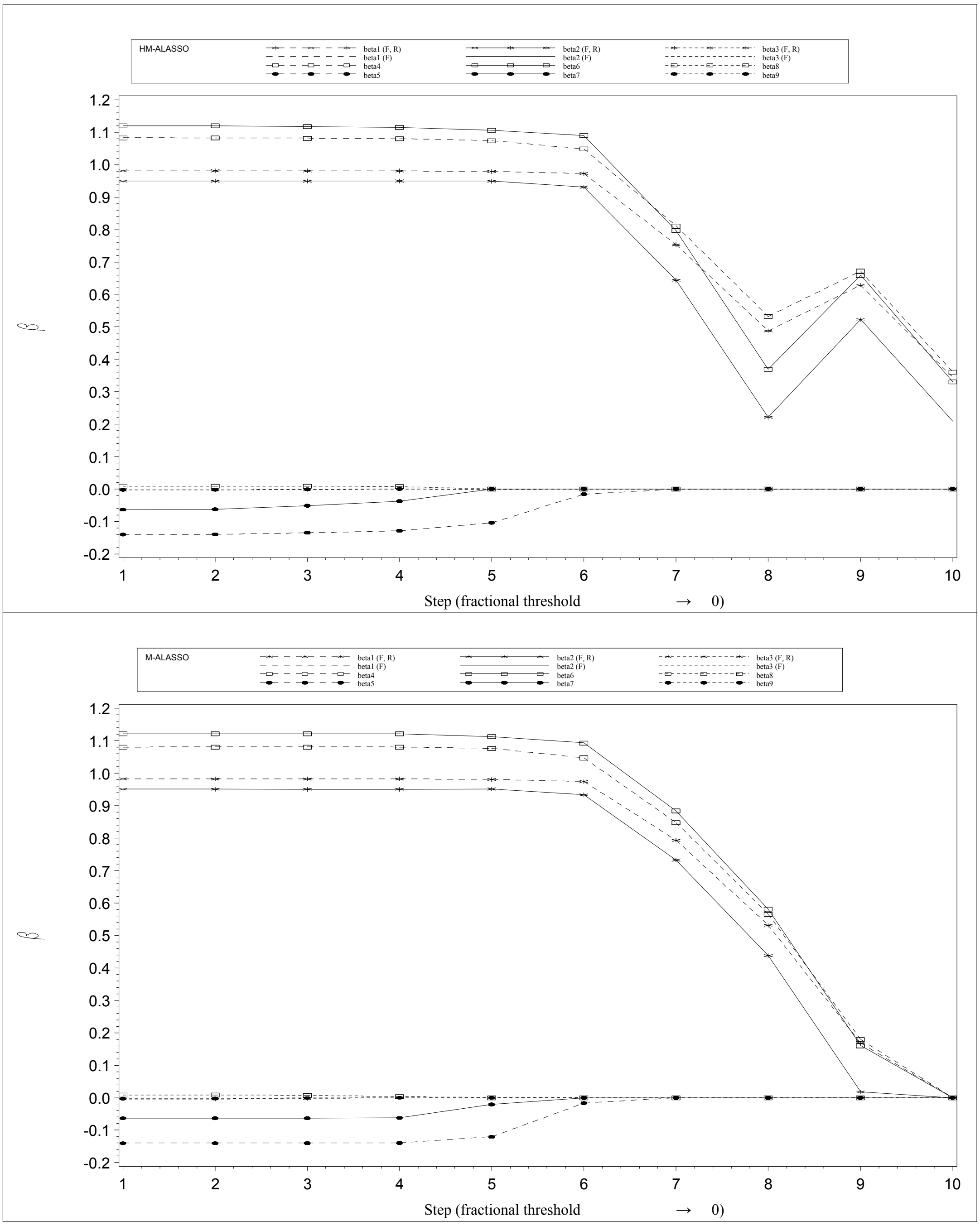

Figure 3.5: Coefficient Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 5

the importance ratio is a ratio of the absolute value of the shrunken estimate to the absolute value of the GLS estimate. The Figures show that the HM-ALASSO better prevents parent terms from being selected out before their descendants. Each plot represents one iteration from each simulated case. The left side of each plot shows the importance ratios for HM-ALASSO. The right side of each plot shows the importance rations for M-ALASSO. Each row plot displays results for the controlled groups in Table 3.2. Under the heredity rules in Table 3.1, we would not want to see a parent term (”parent”, empty symbol) reach 0 before one of its descendants (”child”, solid symbol). The importance ratio plots show that the heredity rules are violated for M-ALASSO. For example, the Case 1 plot (3.6) shows that the group 2 parent term (fixed intercept) is excluded at step 2, while its descendants remain in the model. This behavior is very consistent throughout the groups and cases.

3.4.2 Application to HS&B Data

This study applied the HM-ALASSO and M-ALASSO to the HS&B data subset described previously. To investigate the proposed process, a subset of the full data set with features of interest was chosen. This evaluation of this subset in `lmer` produced a model with an intercept and `ses` at level one, and `sector`, `meanses`, and `disclim` at level two. The intercept and `ses` were specified as random in order to induce cross-level interactions.

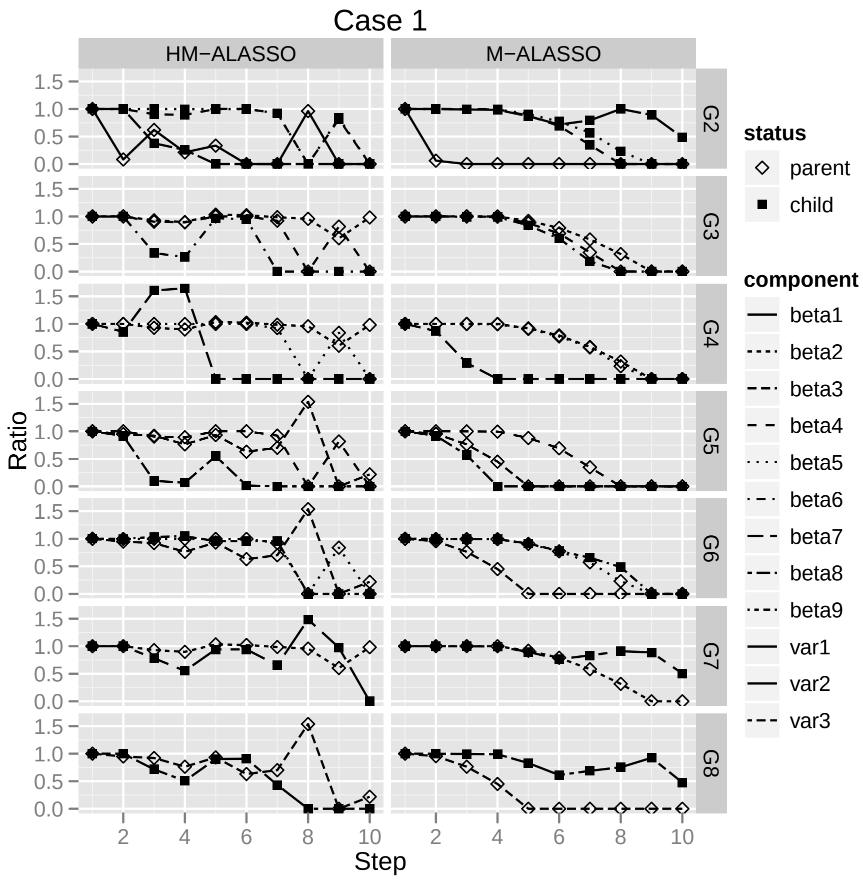

Figure 3.6: Importance Ratio Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 1

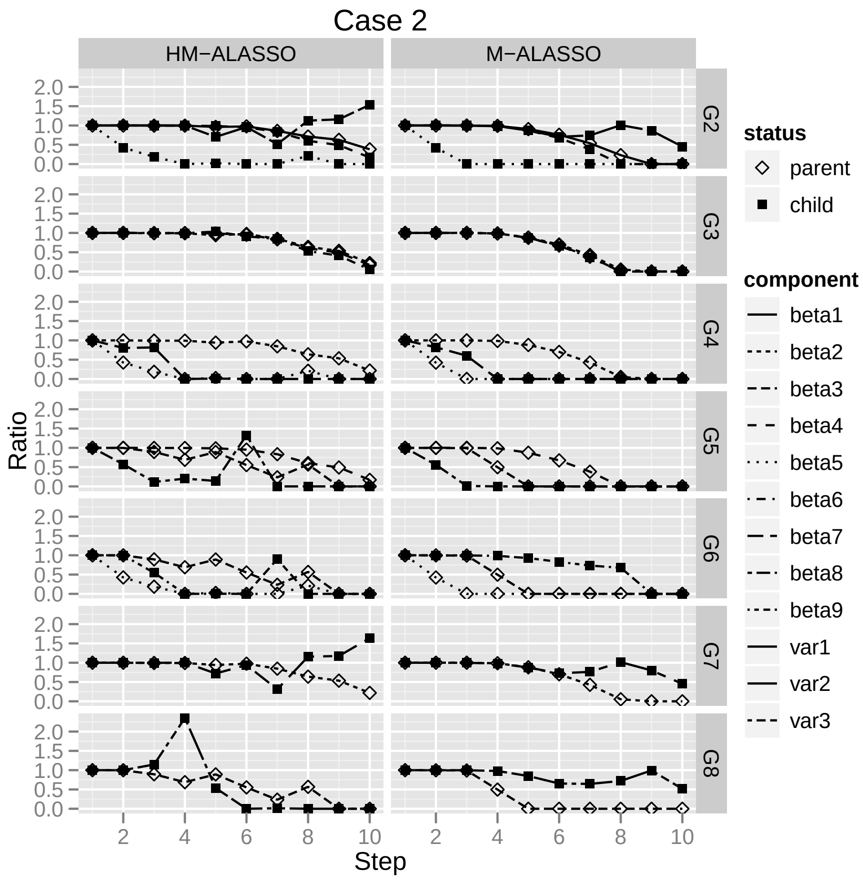

Figure 3.7: Importance Ratio Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 2

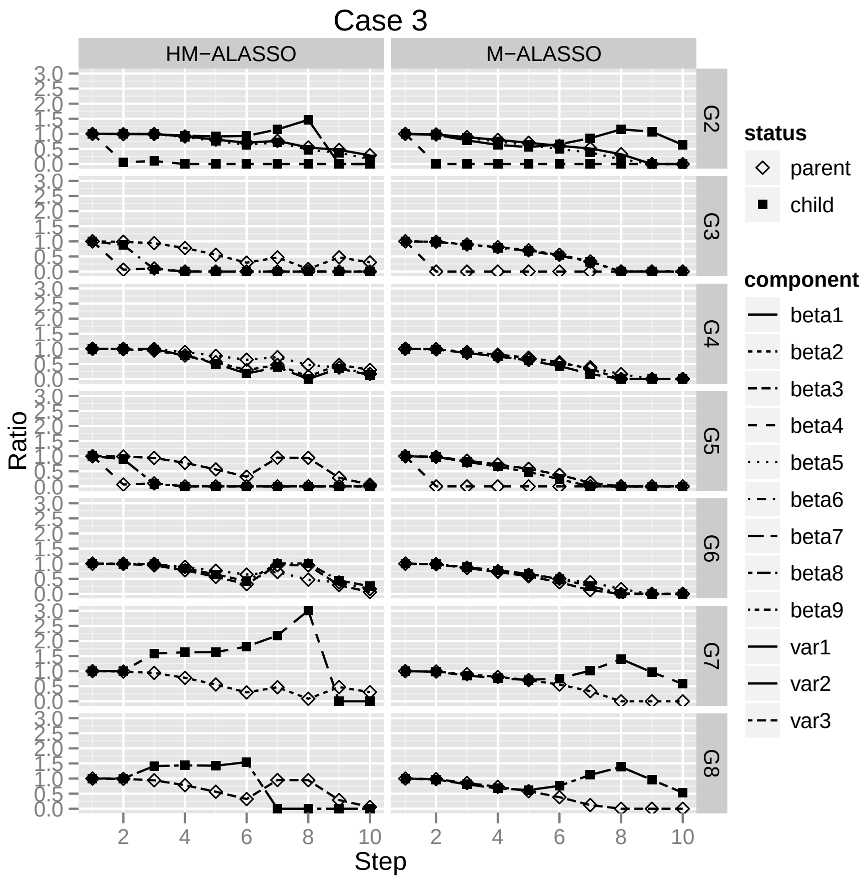

Figure 3.8: Importance Ratio Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 3

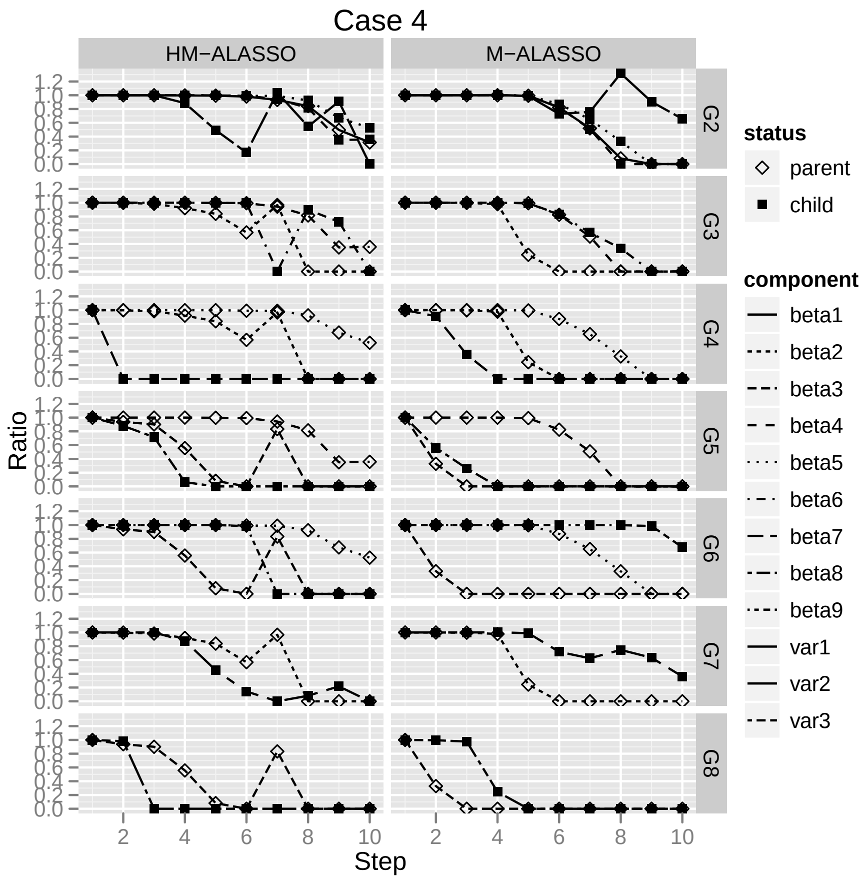

Figure 3.9: Importance Ratio Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 4

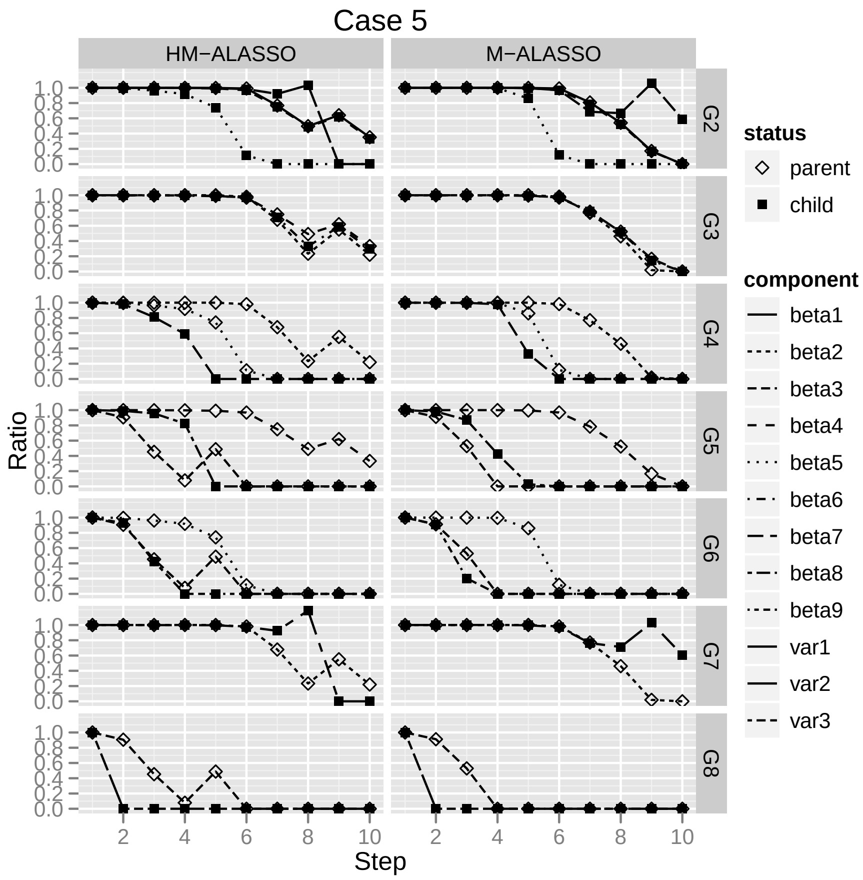

Figure 3.10: Importance Ratio Plots of Fixed-Effects Estimates at Each Regularization Step: Simulation Case 5

The combined model involving these predictors is

$$
\begin{aligned}
Y_{ij} =& \beta_1 + \beta_2(\texttt{ses}_{ij}) + \beta_3(\texttt{meanses}_{ij}) + \beta_4(\texttt{sector}_{ij}) + \beta_5(\texttt{disclim}_{ij}) \\
&+ \beta_6(\texttt{meanses}_{ij})(\texttt{ses}_{ij}) + \beta_7(\texttt{sector}_{ij})(\texttt{ses}_{ij}) + \beta_8(\texttt{disclim}_{ij})(\texttt{ses}_{ij}) \\
&+ B_{1j} + B_{2j}(\texttt{ses}_{ij}) + R_{ij},
\end{aligned}
$$

with all elements as previously defined.

From the initial analysis of the data using `lmer`, the model contains at least one main effect that would not be considered significant in terms of t-ratio (β_4, estimated as 0.61 with t-ratio 1.50). It also contains a non-significant interaction term (β_8, estimated as 0.25 with t-ratio 1.44). Therefore, it provides a relatively simple, but challenging, real-data illustration of the methods in a low-dimensional context with little sparsity. The full model has basic structural similarity to that in the simulated Case 3, in which a cross-level interaction was significant while its group-level parent term was not.

Table 3.16 contains the fixed-effect and random-effect-covariance matrix estimates for the HS&B data subset using HM-ALASSO and M-ALASSO. The BIC and CV1 PMSE were used for selection. Although CV did not always choose the more parsimonious model, CV provides a way of accounting for out-of-sample behavior and model stability. For BIC, any model with $\Delta_{BIC} \leq 2$ was included.

Each estimation and tuning method combination chose the full GLS estimates. Using BIC, each method was able to detect and select out β_8. From 3.17, it is clear that both methods did select out components as the fractional thresholds decreased. In fact, both methods selected β_4 at the threshold just after the strictest one chosen by the criteria. However, this may have come at the cost of model-data fit as defined by commonly used criteria. Figure 3.11 displays the behaviors of the HM-ALASSO and M-ALASSO in shrinking and selecting out terms. Scaled coefficients representing each fixed-effect estimate as a portion of the total penalty term are plotted against steps (decreasing fractional thresholds, increasing λ). The scaled coefficient, rather than the coefficient itself, is used in this case because of the different scales of the predictors. In the simulation study, all non-zero terms were given equal ground. In the HS&B subset, it is clear that there are different scales for the terms. This could be addressed prior to estimation by applying centering or standardization to the data columns. However, this would change the explicit model being used to illustrate the procedure. The more direct movement toward 0 of the interaction terms in the HM-ALASSO illustrates the effects of the heredity constraints. For example, the fixed-effect term β_7 (interaction of `sector` and `ses`), should be excluded as soon as its parent term `sector` is excluded. This occurs for HM-ALASSO but does not occur for M-ALASSO.

Table 3.16: Parameter Estimates for Models Chosen by BIC and PMSE from a Subset of HS&B Data.

Method	Tuning	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\mathbf{\Sigma}}_{11}$	$\widehat{\mathbf{\Sigma}}_{12}$	$\widehat{\mathbf{\Sigma}}_{22}$
HM-ALASSO	BIC	12.39	2.80	5.19	0.60	−0.48	1.04	−1.31	0.23	2.23	0.22	0.06
		12.41	2.81	5.20	0.57	−0.49	1.01	−1.33	0.20	2.22	0.22	0.02
		12.48	2.81	5.24	0.40	−0.49	0.87	−1.38	0.10	2.20	0.00	0.00
		12.59	2.79	5.31	0.16	−0.50	0.66	−1.38	0.00	2.18	0.00	0.00
	PMSE	12.39	2.80	5.19	0.60	−0.48	1.04	−1.31	0.23	2.23	0.22	0.06
M-ALASSO	BIC	12.40	2.82	5.18	0.58	−0.51	1.10	−1.34	0.21	2.20	0.17	0.01
		12.42	2.83	5.19	0.54	−0.52	1.07	−1.37	0.18	2.19	0.18	0.02
		12.49	2.84	5.22	0.39	−0.53	0.95	−1.42	0.08	2.17	0.00	0.00
		12.62	2.81	5.29	0.10	−0.56	0.75	−1.39	0.00	2.16	0.00	0.00
	PMSE	12.40	2.82	5.18	0.58	−0.51	1.10	−1.34	0.21	2.20	0.17	0.01

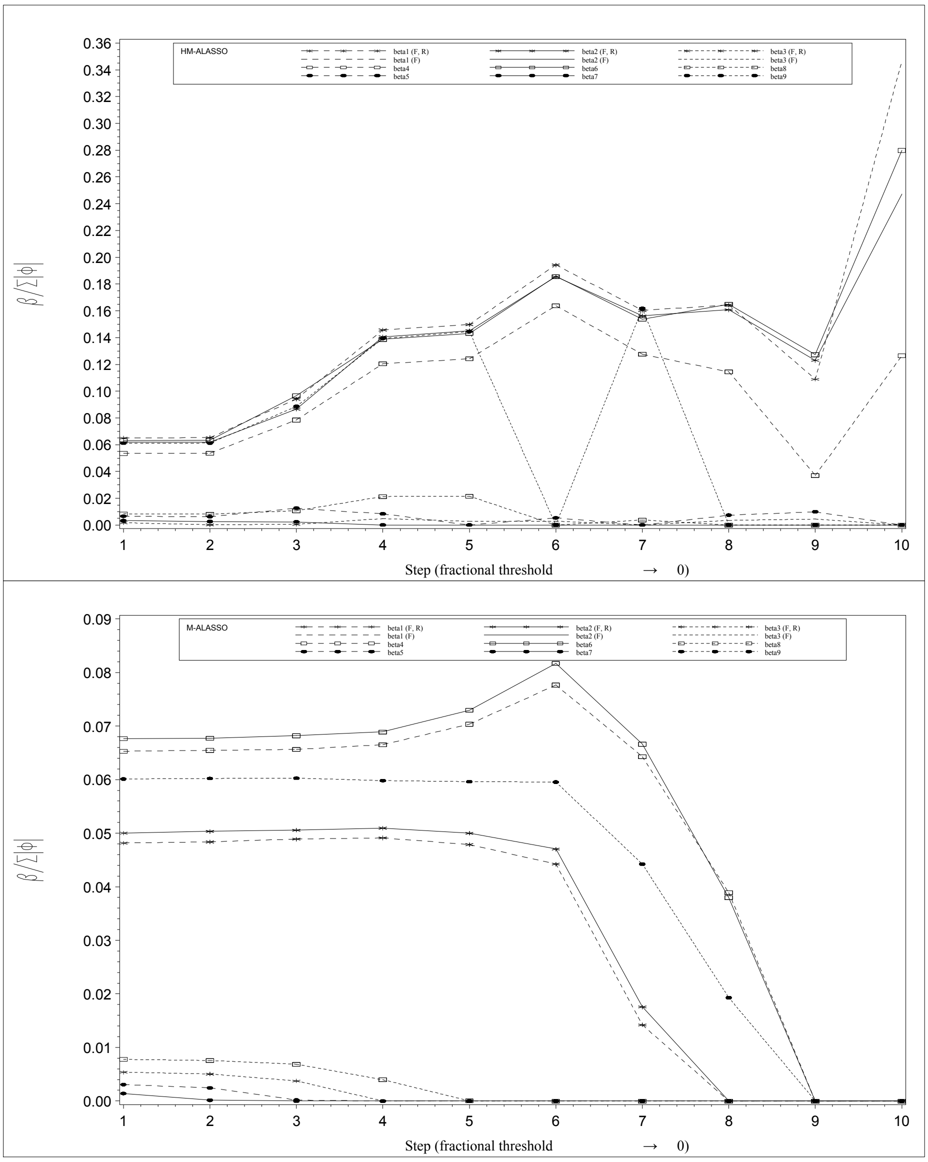

Figure 3.11: HS&B Scaled Coefficients: Scaled Fixed-Effects Estimates at Each Regularization Step

Table 3.17: Number of Thresholds (out of 10) with Nonzero Parameter Estimates for HS&B Data Subset.

Method	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\mathbf{\Sigma}}_{11}$	$\widehat{\mathbf{\Sigma}}_{12}$	$\widehat{\mathbf{\Sigma}}_{22}$
HM-ALASSO	10	9	10	4	7	5	4	5	10	2	2
M-ALASSO	10	9	10	4	6	5	7	3	8	2	2

CHAPTER 4

Conclusions and Directions for Future Work

This dissertation introduced a novel method of implementing heredity-constrained, regularized estimation for multilevel models. Multilevel models were defined in this dissertation to refer to models for data that have structures with at least two levels of variation. For multilevel models, one typically wishes to estimate fixed-effects and random-effects covariance-matrix parameters. Parameter estimates are typically produced through maximum-likelihood estimation. However, multilevel models do not have closed-form expressions for their likelihoods. Therefore, iterative procedures must be applied in order to produce the desired estimates. If the likelihood alone is maximized to produce the estimates, we refer to the estimation process as unconstrained.

In other words, no additional constraints are imposed that define the solution space.

Regularization is a penalized estimation process that we refer to as constrained. Regularization involves the use of one or more penalty functions. The penalty function is a function of the parameters that are to be estimated. The penalty function imposes a constraint on the possible solution space. The addition of the penalty function has the effect of reducing, or shrinking, the sizes of the resulting parameter estimates. Some penalty functions allow estimates to be shrunken to 0. Such functions can be used for selection of model components, reducing model overfit.

Regularization approaches may add bias to the estimated coefficients, but the standard errors of the estimates are decreased. Therefore, regularization reduces the instability in parameter estimates. Regularized estimation is useful for analyzing data with complex structures. Regularized approaches to multilevel regression have only recently been a research focus. They have not, to date, been designed with the specific intention of building models that adhere to principles of heredity. The estimator proposed in this dissertation, the heredity-constrained M-ALASSO (HM-ALASSO), includes a regularization penalty that capitalizes on the heredity structure of the model.

The model-building principles on which this disseration focused were strong versions of fixed heredity (e.g., including main effects when their interactions are included) and random heredity (e.g., including fixed effects when their random terms are included). We refer to the controlling terms (main effects and fixed effects, respectively) as parent terms. We refer to the controlled terms (interactions and random terms, respectively) as descendant or child terms. One measure of whether a candidate model is formulated in accordance with heredity is the hierarchical gap. The term *hierarchical gap* was coined by Zhao *et al.* (2009) as a measure of the distance between the candidate model structure and a minimally suitable hierarchical structure. The hierarchical gap indicates the number of terms that would need to be added to a model to meet heredity requirements. From the perspective of parsimony, descendant terms should be excluded if their parent terms are not included.

Strong heredity, in which all parent terms must be in the model, is related to the principle of marginality (see, e.g., McCullagh & Nelder, 1989, pp. 69–70). The core marginality argument does not allow the inclusion of higher-order terms without their lower-order counterparts. The HM-ALASSO enforces both fixed and random heredity, closing the hierarchical gap for data with a multilevel structure.

The goals in developing this method were to create a procedure that provided reasonable estimates of all parameters, adhered to fixed and random heredity principles, resulted in a parsimonious model, was theoretically justifiable, and was able to be implemented and used in available software (e.g., in

`R`). This last criterion has been cited as important for realizing the benefits of new procedures (Hesterberg *et al.*, 2008).

The HM-ALASSO incorporates heredity-constrained selection of fixed and random components directly into the estimation process. The HM-ALASSO algorithm invokes the heredity constraints by applying heredity weights to groups of components specified by the researcher. Heredity weights are based on the importance ratios of terms for which the group terms are considered descendants. Importance ratios were defined for each controlling model component as the ratio of the absolute value of its penalized estimate to the absolute value of its generalized-least-squares (GLS) estimate. The importance ratio indicates how important a controlling term is (i.e., how much or how little shrinkage was applied to it).

The M-ALASSO `Pen.LME` function (Bondell *et al.*, 2010) applies an adaptive Lasso penalty for joint selection of fixed and random components but does not take into account the structure of the data. The adaptive Lasso penalty weights the absolute values of penalty terms by the absolute values of their respective GLS estimates. The M-ALASSO estimator is implemented with an EM-type algorithm using a linearly constrained quadratic programming approach to optimize, or solve for, the desired parameter estimates. Successively tighter bounds on the size of the penalty term (referred to in this dissertation as fractional thresholds), constrain the sizes of the resulting estimates. As with other regularized estimators of its kind, the HM-ALASSO estimates could be solved for using various optimization procedures. For the analyses in this dissertation, the HM-ALASSO penalty term was implemented by modifying the `Pen.LME` function with relatively little extra code.

The approach described herein produces comparable estimates to those from available software (`Pen.LME`) for candidate models with the same structure (i.e., the same zero and non-zero terms). Candidate models with different structures arise because of the different penalty functions that are applied by different regularized estimation procedures. The HM-ALASSO candidate models are closer in fit to their respective data-generating models when taking heredity structure into account.

The HM-ALASSO procedure addresses heredity issues that arise when cross-level interactions are induced in a two-level, cross-sectional multilevel model with random coefficients and level-two covariates. A strict set of heredity rules were applied for this structure, and a particular model structure was used for illustration. However, this versatile conceptual method can be adapted to invoke other rule-set constraints under different model parameterizations.

The quality of the HM-ALASSO was demonstrated theoretically and empirically. The asymptotic proofs in this dissertation demonstrate that HM-ALASSO enjoys the properties of consistency, sparsity, and asymptotic normality that would be desired for a regularized estimator. The empirical analyses with simulated data and a subset of the High School and Beyond (HS&B) data set provided challenging low-dimensional contexts for which the true data-generating process by which the data were generated may not adhere to strong heredity. In these applications, the performance of the HM-ALASSO was compared to that of the M-ALASSO. The M-ALASSO estimator has been demonstrated to be of high quality. However, M-ALASSO does not take data structure into account. Therefore, the empirical analyses were undertaken to determine whether heredity structure could be imposed effectively using the HM-ALASSO estimator while producing estimates of similarly high quality.

Two types of graphical elements were introduced to illustrate visually the comparison of the behavior of the two estimators. Importance ratio plots displaying importance ratios based on final estimates at each step were used to show that HM-ALASSO prevents parent terms from being eliminated prior to their descendants. Coefficient and scaled-coefficient plots were included to show how the fixed-effects coefficients changed at each fractional threshold. For level-one fixed-effects plot lines, symbols represent the presence of the associated variance component. This approach allows both the fixed and random structures to be visualized jointly.

The simulation study included four structural cases with data generated from models that violate heredity. A fifth case with a data-generating model that did conform to heredity was included as a baseline for comparison of estimates. Parameter estimates and model error were found to be comparable between methods, and the hierarchical gap was demonstrably smaller for models produced by HM-ALASSO. The results demonstrate that HM-ALASSO performs well for a variety of heredity-violating data-generating model structures.

The HM-ALASSO and M-ALASSO were applied to a subset of the HS&B data set. The full model had basic structural similarity to that in a simulated case in which a cross-level interaction was significant while its group-level parent term was not (referred to in the dissertation as Case 3). The HM-ALASSO was able to exclude the cross-level interaction term as soon its parent term was eliminated, while M-ALASSO retained the cross-level interaction in the model.

In the course of this dissertation research, several ways to incorporate heredity were investigated. The linearly constrained quadratic programming framework proved to be a useful way to implement the HM-ALASSO heredity penalty. However, future research on the approach described in this dissertation will focus on optimization and model selection.

In terms of optimization, the quadratic programming estimation mechanism should be evaluated to determine whether it can be made more efficient. There is little guidance in the literature about how to determine the number of fractional thresholds, and the grain of those thresholds, that should be used to produce candidate models. This is analogous to the question of how to specify the λ tuning parameter(s) that function as regularization parameters. Efficiency could be improved by better capitalizing on the highly structured nature of the data for which the HM-ALASSO was developed. Nonlinear constraints such as indicator functions may be better suited for controlling inclusion of terms. However, software packages that provide nonlinear quadratic programming solutions do not yet have widespread availability. This approach is, therefore, less useful to practitioners at the current time.

Once candidate models have been produced, it is critical to be able to select a final model that meets heredity goals. The analyses undertaken in this dissertation demonstrated that commonly used criteria are not sufficient for this purpose. Selection criteria typically focus only on model-data fit or prediction quality. Therefore, they are not built to identify candidate models that also conform to heredity rules. The commonly used criteria that were applied to select from candidate models in this dissertation suffered from that deficiency. Although the researcher can evaluate a small number of candidate models in a brute-force fashion, the automation in the regularization process can produce many viable candidate models. The benefit of having more models available to evaluate is offset by the lack of a criterion that can adequately select models on its own. Future research to develop criteria that take the data structure into account will be crucial.

Research to optimize the estimation and selection processes is important due to the computational burdens that accompany larger data sets and more complicated models. Improvements in these areas will be critical when these methods are extended to models with high dimensional contexts. The HM-ALASSO concept could be adapted for use in other settings. In particular, extensions to models with different fixed and random structures (e.g., effects related by serial or spatio-temporal correlation), alternative link functions, and more levels of variation are viable areas for future research and application.

BIBLIOGRAPHY

Afshartous, D. & de Leeuw, J. (2004). An Application of Multilevel Model Prediction to NELS: 88. *Behaviormetrika*, **31**, 43–66.

Afshartous, D. & de Leeuw, J. (2005a). Decomposition of Prediction Error in Multilevel Models. *Communications in Statistics—Simulation and Computation*, **34**, 909–928.

Afshartous, D. & de Leeuw, J. (2005b). Prediction in Multilevel Models. *Journal of Educational and Behavioral Statistics*, **30**, 109–139.

Ahn, M. (2010). *Random Effect Selection in Linear Mixed Models (PhD Thesis)*. North Carolina State University.

Albert, J. (2009). Is Roger Clemens' WHIP Trajectory Unusual? *CHANCE*, **22**, 9–20.

Armagan, A. & Dunson, D. (2011). Sparse Variational Analysis of Linear Mixed Models for Large Data Sets. *Statistics & Probability Letters*, **81**, 1056–1062.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed-effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language*, **59**, 390–412.

Barcikowski, R. S. (1981). Statistical Power with Group Mean as the Unit of Analysis. *Journal of Educational and Behavioral Statistics*, **6**, 267–285.

Bartels, B. L. (2008). Beyond "Fixed versus Random Effects": A Framework for Improving Substantive and Statistical Analysis of Panel, Time-Series Cross-Sectional, and Multilevel Data. *The Society for Political Methodology Working Paper*.

Bates, D. M. (2006). lmer, p-values and All That. `http://www.stat.ethz.ch/pipermail/r-help/2006-May/094765.html`.

Bates, D. M. & Pinheiro, J. C. (1998). Computational Methods for Multilevel Modelling. `http://stat.bell-labs.com/NLME/CompMulti.pdf`.

Berkhof, J. & Snijders, T. A. B. (2001). Variance Component Testing in Multilevel Models. *Journal of Educational and Behavioral Statistics*, **26**, 133–152.

Bondell, H. D., Krishna, A. & Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, **66**, 1069–1077.

Breiman, L. (1995). Better Subset Regression using the Nonnegative Garrote. *Technometrics*, **37**, 373–384.

Breiman, L. (2001a). Random Forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with discussion). *Statistical Science*, **16**, 199–215.

Browne, W. J. (1998). *Applying MCMC Methods to Multi-level Models (PhD Thesis)*. University of Bath.

Browne, W. J. (2005). *MCMC Estimation in MLwiN: Version 2.0*. Centre for Multilevel Modelling, University of Bristol.

Browne, W. J. & Steele, F. (2009). Editorial: Recent Advances in Multilevel Modelling Methodology and Applications. *Journal of the Royal Statistical Society, Ser. A*, **172**, 535–536.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**, 261–304.

Busing, F. M. T. A. & te Leiden, R. (1993). *Distribution characteristics of variance estimates in two-level models: a Monte Carlo study*. Department of Psychology, University of Leiden.

Chen, Z. & Dunson, D. B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, **59**, 762–769.

Crainiceanu, C. (2008). Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. In D. B. Dunson, ed., *Random Effect and Latent Variable Model Selection*, pp. 3–17. New York: Springer Verlag.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D. & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, **79**, 69–102.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, **32**, 407–451.

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

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London, Ser. A*, **222**, 309–368.

Frank, I. E. & Friedman, J. H. (1993). A Statistical View of Some Chemometric Regression Tools. *Technometrics*, **35**, 109–135.

Gelman, A. (2006a). Multilevel Modeling: What it Can and Can't Do. *Technometrics*, **48**, 241–251.

Gelman, A. (2006b). Number of Parameters in a Model and DIC. `http://www.stat.columbia.edu/~cook/movabletype/archives/2006/07/number_of_param.html`.

Gelman, A. (2008). Beyond 'Fixed versus Random Effects'. `http://www.stat.columbia.edu/~cook/movabletype/archives/2008/10/beyond_fixed_ve.html`.

Gelman, A. (2011). Multilevel Regression with Shrinkage for "Fixed" Effects. `http://www.stat.columbia.edu/~cook/movabletype/archives/2011/04/multilevel_regr.html`.

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, A. B., Carlin, J. S., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd edition. Boca Raton: Chapman & Hall/CRC Press.

George, E. I. (2000). The Variable Selection Problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Gilbert, J. A. & Shultz, K. S. (1998). Multilevel Modeling in Industrial and Personnel Psychology. *Current Psychology*, **17**, 287–300.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*, pp. 1–19. Boca Raton: Chapman & Hall/CRC Press.

Goldfarb, D. & Idnani, A. (1983). A Numerically Stable Dual Method for Solving Strictly Convex Quadratic Programs. *Mathematical Programming*, **27**, 1–33.

Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (1989). Restricted Unbiased Iterative Generalized Least-squares Estimation. *Biometrika*, **76**, 622–623.

Goldstein, H. & Kounali, D. (2009). Multilevel Multivariate Modelling of Childhood Growth, Numbers of Growth Measurements and Adult Characteristics. *Journal of the Royal Statistical Society, Ser. A*, **172**, 599–613.

Goodman, A. C. & Thibodeau, T. G. (1998). Housing Market Segmentation. *Journal of Housing Economics*, **7**, 121–143.

Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, **54**, 93–108.

Harville, D. A. (1974). Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, **61**, 383–385.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320–338.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition. Springer Series in Statistics.

Hernández-Lloreda, M. V., Colmenares, F. & Arias, R. M. (2003). Application of Hierarchical Linear Modelling to the Study of Trajectories of Behavioural Development. *Animal Behaviour*, **66**, 607–613.

Hesterberg, T., Choi, N. H., Meier, L. & Fraley, C. (2008). Least Angle and l_1-penalized Regression: A Review. *Statistics Surveys*, **2**, 61–93.

Hoffman, D. A., Izenman, A. J. & Lidicker, J. R. (2007). Docketology, District Courts, and Doctrine. *Washington University Law Review*, **85**, 681–751.

Hox, J. J. (1998). Multilevel Modeling: When and why. In I. Balderjahn, R. Mathar & M. Schader, eds., *Classification, Data Analysis, and Data Highways*, pp. 147–154. New York: Springer Verlag.

Huang, J., Ma, S., Xie, H. & Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

Ibrahim, J. G., Zhu, H., Garcia, R. I. & Guo, R. (2010). Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Verlag.

Kinney, S. K. & Dunson, D. B. (2008). Bayesian Model Uncertainty in Mixed Effects Models. In D. B. Dunson, ed., *Random Effect and Latent Variable Model Selection*, pp. 37–62. New York: Springer Verlag.

Kreft, I. G. G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage.

Lan, L. (2006). *Variable Selection in Linear Mixed Model for Longitudinal Data (PhD Thesis)*. North Carolina State University.

Leckie, G. (2009). The Complexity of School and Neighbourhood Effects and Movements of Pupils on School Differences in Models of Educational Achievement. *Journal of the Royal Statistical Society, Ser. A*, **172**, 537–554.

Leyland, A. H. & Næss, Ø. (2009). The Effect of Area of Residence over the Life Course on Subsequent Mortality. *Journal of the Royal Statistical Society, Ser. A*, **172**, 555–578.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford: Oxford University Press.

Maas, C. J. M. & Hox, J. J. (2004). Robustness Issues in Multilevel Regression Analysis. *Statistica Neerlandica*, **58**, 127–137.

Mathieu, J. E., Aguinis, H., Culpepper, S. & Chen, G. (2012). Understanding and Estimating the Power to Detect Cross-level Interaction Effects in Multilevel Modeling. *J. Applied Psych.*, **97**, 951–966.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Boca Raton: Chapman & Hall.

McMurray, K. (2010). *A Monte Carlo Study of Fit Indices in Hierarchical Linear Models (PhD Thesis)*. Georgia State University.

Merlo, J., Bengtsson-Boström, K., Lindblad, U., Råstam, L. & Melander, O. (2006). Multilevel Analysis of Systolic Blood Pressure and ACE Gene I/D Polymorphism in 438 Swedish Families – A Public Health Perspective. *Medical Genetics*, **7**, 14–22.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. Springer Science+ Business Media.

Patterson, H. D. & Thompson, R. (1971). Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrika*, **58**, 545–554.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer Verlag.

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*. Stata Corp.

Rasbash, J. & Browne, W. J. (2007). Non-hierarchical Multilevel Models. In J. de Leeuw & E. Meijer, eds., *Handbook of Multilevel Analysis*, pp. 301–334. New York: Springer.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd edition. Thousand Oaks: Sage.

Rowe, A. K., Lama, M., Onikpo, F. & Deming, M. S. (2002). Design Effects and Intraclass Correlation Coefficients from a Health Facility Cluster Survey in Benin. *International Journal for Quality in Health Care*, **14**, 521.

Schelldorfer, J., Bühlmann, P. & van de Geer, S. (2011). Estimation for High-dimensional Linear Mixed-effects Models Using l_1-penalization. *Scandinavian Journal of Statistics*, **38**.

Schochet, P. J. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics*, **33**, 62–87.

Shor, B., Bafumi, J., Keele, L. & Park, D. (2007). A Bayesian Multilevel Modeling Approach to Time-series Cross-sectional Data. *Political Analysis*, **15**, 165–181.

Short, J. C., Jr, D. J. K., Bennett, N. & du Toit, M. (2006). An Examination of Firm, Industry, and Time Effects on Performance Using Random Coefficients Modeling. *Organizational Research Methods*, **9**, 259–284.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, **86**, 420–428.

Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Snijders, T. A. B. (2005). Power and Sample Size in Multilevel Modeling. *Encyclopedia of Statistics in Behavioral Science*, **3**, 1570–1573.

Snijders, T. A. B. & Berkhof, J. (2007). Diagnostic Checks for Multilevel Models. In J. de Leeuw & E. Meijer, eds., *Handbook of Multilevel Analysis*, pp. 141–176. New York: Springer.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **64**, 583–639.

Stone, E. A. & Izenman, A. J. (2012). A Review of Multilevel Model Selection. *Technical Report, Department of Statistics, Temple University.*

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, Ser. B*, **67**, 91–108.

Trautwein, U., Köller, O., Schmitz, B. & Baumert, J. (2002). Do Homework Assignments Enhance Achievement? A Multilevel Analysis in 7th-grade Mathematics. *Contemporary Educational Psychology*, **27**, 26–50.

Tseng, P. & Yun, S. (2009a). A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization. *Mathematical Programming*, **117**, 387–423.

Tseng, P. & Yun, S. (2009b). Block-Coordinate Gradient Descent Method for Linearly Constrained Nonsmooth Separable Optimization. *Journal of Optimization Theory and Applications*, **140**, 513–535.

Wagner, T., Hayes, D. B. & Bremigan, M. T. (2006). Accounting for Multilevel Data Structures in Fisheries Data Using Mixed Models. *Fisheries*, **31**, 180–187.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large. *Transactions of the American Mathematical Society*, **54**, 426–482.

Wang, S., Song, P. & Zhu, J. (2010). Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-effects Models. *The University of Michigan Department of Biostatistics Working Paper Series Paper 89*.

Willems, T., Takkenberg, J., Steyerberg, E., Kleyburg-Linkers, V., Roelandt, J., Bos, E. & van Herwerden, L. (2001). Human Tissue Valves in Aortic Position Determinants of Reoperation and Valve Regurgitation. *Circulation*, **103**, 1515–1521.

Wolfinger, R. (1993). Covariance Structure Selection in General Mixed Models. *Communications in Statistics—Simulation and Computation*, **22**, 1079–1106.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. S. (2001). Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal Computational Biology*, **8**, 625–637.

Yuan, M., Joseph, V. R. & Zou, H. (2009). Structured Variable Selection and Estimation. *Annals of Applied Statistics*, **3**, 1738–1757.

Yuan, M. & Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Ser. B*, **68**, 49–67.

Zahs, D., Pedlow, S., Morrissey, M., Marnell, P. & Nichols, B. (1995). High School and Beyond Fourth Follow-Up Methodology Report, NCES 95426. *National Center for Education Statistics*.

Zhao, P., Rocha, G. & Yu, B. (2009). The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection. *Annals of Statistics*, **37**, 3468–3497.

Zhou, N. & Zhu, J. (2010). Group Variable Selection via a Hierarchical Lasso and its Oracle Property. *Arxiv preprint arXiv:1006.2871*.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "Degrees of Freedom" of the Lasso. *Annals of Statistics*, **35**, 2173–2192.