ABSTRACT

Title of thesis: SELECTION OF FIXED AND RANDOM EFFECTS IN LINEAR MIXED EFFECTS MODELS WITH APPLICATIONS TO THE TRIAL OF ACTIVITY IN ADOLESCENT GIRLS

Edward Grant, Master of Public Health, 2013

Directed by: Professor Tong Tong Wu
Department of Epidemiology and Biostatistics

Linear mixed effect (LME) models have become popular in modeling data in a wide variety of fields, particularly in public health. These models are beneficial because they are able to account for both the means as well as the covariance structure of clustered or longitudinal data. However, as studies are able to collect an increasing amount of data for large numbers of predictors, a major challenge has been the selection of only important variables to create a more interpretable, parsimonious model. Previous methods for LME models have been inefficient in variable selection, but three new methods attempt to select and estimate both important fixed and important random effects simultaneously. The models are compared through analysis of simulated longitudinal data. Additionally, as an example of the important applications to public health, the methods are applied to the Trial of Activity in Adolescent Girls (TAAG) study, to determine important predictors for Moderate to Vigorous Physical Activity (MVPA).

SELECTION OF FIXED AND RANDOM EFFECTS IN LINEAR MIXED EFFECTS MODELS WITH APPLICATIONS TO THE TRIAL OF ACTIVITY IN ADOLESCENT GIRLS

by

Edward Grant

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Public Health
2013

Advisory Committee:
Professor Tong Tong Wu, Chair, Advisor
Professor Shuo Chen
Professor Brit Saksvig

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
BIC	Bayes Information Criterion
EM	Expectation-Maximization
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LME	Linear Mixed Effects
ML	Maximum Likelihood
MVPA	Moderate to Vigorous Physical Activity
PA	Physical Activity
PE	Physical Education
OLS	Ordinary Least Squares
OSCAR	Octagonal Shrinkage and Clustering Algorithm for Regression
REML	Restricted Maximum Likelihood
SCAD	Smoothed Clipped Absolute Deviance
TAAG	Trial of Activity in Adolescent Girls

Chapter 1

Introduction

Linear mixed-effects (LME) models (Laird and Ware 1982) are statistical models that are used in the analysis of clustered or longitudinal data. LME models estimate the relationship between the dependent variable and the predictors included in the model, accounting for both the fixed effects and the random effects of the independent variables. Compared with linear regression models without considering clustering or temporal effects, LME models are able to more accurately estimate the fixed effects by estimating the covariance structure through the inclusion individual-specific random effects. Ignoring the covariance structure has been shown to lead to biased estimates (Lange and Laird 1998).

Improvements in technology have enabled researchers to collect and store data on an increasing number of predictors. However, inferences and predictions of an LME model that includes all predictors become too complex or infeasible as the number of predictors, all of which include fixed and random components, increases. One challenge in LME models is choosing a parsimonious model that selects only the significant covariates, while excluding variables that have no true effect on the outcome. Many methods have been published on model selection, but three new methods have been introduced which, unlike many previous approaches, can estimate both fixed and random effects simultaneously. First, a method developed by

Li et al. (2012), optimizes a regularization problem with two separate penalization methods for fixed and random effects. Next, Bondell et al. (2010) select and estimate fixed and random effects simultaneously by maximizing a jointly penalized regularization problem. Finally, Fan and Li (2012) use a proxy matrix to account for covariance structure in maximizing a penalized profile likelihood for fixed and random effects separately. These methods present more practical ways of selecting important fixed and random effects of LME models compared to previous methods. The goal of this thesis will be to compare the efficacy and accuracy of these new variable selection methods though analysis of data simulation studies.

Additionally, a comparison of these methods will be performed through a real public health data set. As an example of the power of these methods, we consider the study of the Trial of Activity for Adolescent Girls 2 (TAAG 2), which determined the predictors of physical activity among adolescent girls from 6 schools in Maryland (Young et al. 2013). Data for 65 multilevel variables from 551 girls were collected at two time points, 2006 and 2009, when the girls were enrolled in the 8th and 11th grade, respectively. Using traditional methods, building a parsimonious model for this data would be tedious and could introduce bias. However, this data can be analyzed efficiently using these methods to determine which variables truly have a relationship with the outcome variable of interest, moderate to vigorous physical activity (MVPA). This data analysis will demonstrate the important applications that these methods can have in the public health field.

The rest of this thesis will proceed as follows. Section 2 will discuss previous methods used for variable selection and give an introduction to the TAAG trial.

Section 3 will introduce the three methods used for variable selection in LME models. In Section 4, simulation studies will be performed to compare the effectiveness of the three methods. Section 5 will carry out the analysis of the TAAG data set. Finally, in section 6, the strengths and weaknesses of the methods will be discussed, along with brief implications of their use on high-dimensional data.

Chapter 2

Background

2.1 Linear Mixed-Effects Models for Longitudinal Data

Consider a longitudinal study with n subjects. Each subject i, $(i = 1, ..., n)$ has m observations. The number of observations can be generalized to m_i so that the number of observations can vary across subjects, for a total of $\sum_{i=1}^{n} m_i = N$ observations. Suppose there are p covariates associated with the fixed effects, $X_1, ..., X_p$ and q random effects associated with the random effects $Z_1, ..., Z_q$. For subject i, let $\boldsymbol{x}_{ij}$ be the vector of p predictors for fixed effects for $j \in \{1, \ldots, p\}$ and $\boldsymbol{z}_{ik}$ be the vector of predictors for q random effects for $k \in \{1, \ldots, q\}$. For the outcome Y at observation t, the linear mixed-effects model can then be written as:

$$Y_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \boldsymbol{z}_{it}^T \boldsymbol{b}_i + \epsilon_{ij}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector for fixed effects, $\boldsymbol{b}_i$ is the $q \times 1$ parameter vector for random effects, and ϵ_{ij} represents the error term which is independently and identically distributed to $N(0, \sigma^2)$. The random effects for each subject are independently and identically distributed to multivariate normal distribution $MVN(0, \sigma^2 \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the $m \times m$ covariance matrix. By combining $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im})^T, \boldsymbol{X}_i^T = (x_{i1}, ..., x_{im})$, and $\boldsymbol{Z}_i^T = (z_{i1}, ..., z_{im})$, the LME mode can be simplified to:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

where $\boldsymbol{X}_i$ is the $m \times p$ design matrix for the fixed effects and $\boldsymbol{Z}_i$ is the $m \times q$ design matrix for the random effects for subject i. The error term $\boldsymbol{\epsilon}_i$ is independently and identically distributed $N(0, \sigma^2\boldsymbol{I}_m)$. The subject-specific random effects $\boldsymbol{b}_i$ are independent of the population specific fixed effects $\boldsymbol{\beta}$. It can be seen that $\boldsymbol{\beta}$ is associated with the fixed effects and is used for predicting the mean, while $\boldsymbol{b}$ is associated with the random effects and accounts for much of the variance in the model.

2.2 Penalization Methods for Selection of Fixed Effects

Other methods have become widely used for performing model selection more efficiently, notably penalized regression algorithms. Many different penalization methods have been used to estimate parameters in a regression model with outcome $\boldsymbol{y}_i$, design matrix $\boldsymbol{X}_i$, and coefficients $\boldsymbol{\beta}$. The penalization methods reduce the number of dimensions of the model by setting the coefficients of unimportant predictors to 0, leading to a more parsimonious and interpretable model. For parameter vector $\theta = (\mu, \beta_1, \ldots, \beta_p)$, penalization methods can be summarized as

$$\min f(\theta) = g(\theta) + P_\lambda(\beta), \tag{2.2}$$

where $g(\theta)$ is a loss function and $P_\lambda(\beta)$ is a penalty function on β with tuning parameter λ.

2.2.1 Lasso Penalty

When there are a large number of parameters to be estimated, the "least absolute shrinkage and selection operator," or lasso (Tibshirani 1996), is a commonly used regularization method to reduce the number of parameters. A simple example is lasso penalized linear regression, in which the penalized objective function (2.2) can be written as:

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|, \tag{2.3}$$

where $\lambda > 0$. Lasso applies a constraint on the residual sum of squares when finding the Ordinary Least Squares (OLS) estimate. Placing this constraint is equivalent to adding the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the OLS estimates. The regularization problem (2.3) can then be rewritten as, in dual formation, the sum of the absolute value of the coefficients constrained to less than a certain tuning parameter $t \geq 0$:

$$\min \left\{ \sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 \right\}, \text{ subject to} \sum_j |\beta_j| \leq t,$$

When t is large, the magnitude of the constraint placed on the estimates is minimal, resulting in solutions close to the OLS estimates. When t is small, the constraint placed on the solution causes shrinkage in the OLS estimates towards zero. Equivalently, a large λ corresponds to a small t, resulting in a larger penalty on the

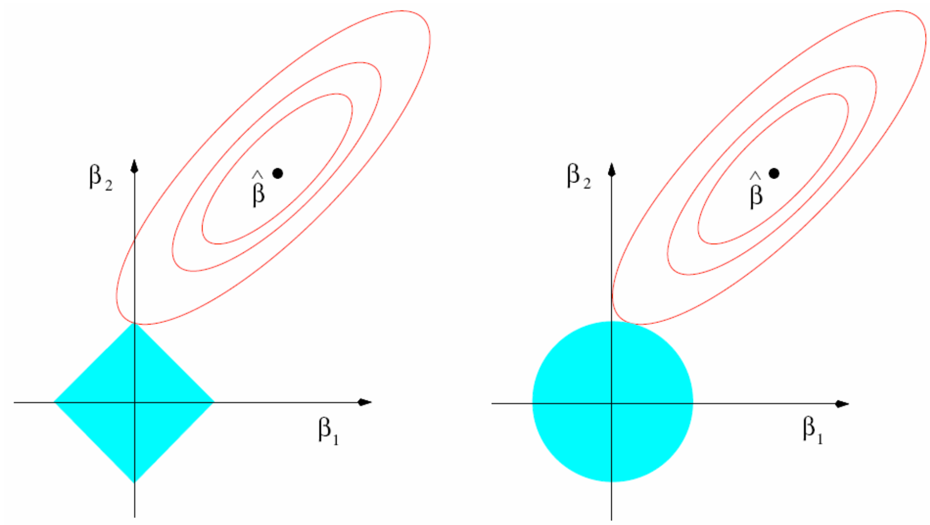

Figure 2.1: Geometric Interpretation of the Lasso Penalization (Hastie et al. 2009). The figure on the left is the Lasso Penalty, while the figure on the right is the Ridge Penalty. The shaded region represents the constraint region for the coefficients. The contours represent the area the objective function has coefficient values.

parameters. As the value of λ increases, a greater magnitude of shrinkage will be placed on the coefficient estimates. An advantage of the lasso is its ability to not only shrink these coefficient estimates towards the origin, but also to perform model selection of the important variables by setting unimportant coefficients to exactly zero.

Figure 2.1 represents a simple problem with two coefficients, β_1 and β_2, showing a geometric interpretation of the lasso (left) compared with the ridge penalty (right). The ridge penalty, where $P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$, is a method that is known to only shrink coefficients without setting any to zero. Similar to the lasso, this penalty is equivalent to a constraint region for the estimation of the parameters, given by:

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min \left\{ \sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2 \right\}, \text{ subject to } \sum_j \beta_j^2 \leq t,$$

However, the shaded regions in the figure show that the penalties result in different shapes for their constraint regions. The region for the ridge's penalty $P_\lambda(\beta) = \lambda(\beta_1^2+\beta_2^2)$ is represented as a circular shape, while the lasso's $P_\lambda(\beta) = \lambda(|\beta_1|+|\beta_2|)$ is represented as a diamond shape with corners at 0. The elliptical contours correspond to the quadratic form of the loss function $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$, where the center of the contour shape is the non-penalized OLS solution for β_1 and β_2. The minimum of the sum of $g(\beta) + P_\lambda(\beta)$ will produce the optimal value of β_1 and β_2, and this minimum occurs at the point of intersection of the contours and the shaded region.

In the figure, the solution on the left shows that β_1 will to be set to 0. In contrast, in the ridge plot, it can be seen that this will almost never happen due to the rounded shape of both the loss function and of penalty function. This demonstrates the important benefit of the lasso penalty: due to the shape of its constraint region, it is more likely to eliminate unimportant predictors and perform model selection (Hastie et al. 2009).

A simple example of the model selection attribute of the lasso can be seen in Table 1, which compares estimates of six predictors using ordinary least squares, estimates after ridge penalization, and estimates after lasso penalization with the true model of three predictors. For sample size $n = 30$ the true model is:

$$y = \beta_0 + 0.5X_1 + 1X_3 + 0.5X_4 + \epsilon,$$

where $\beta_0 = 1.5$, $\boldsymbol{X} \sim N(0, 1)$, and $\epsilon \sim N(0, 1)$. The true predictors in the model are X_1, X_3, and X_4, while X_2, X_5, and X_6 are noise variables. While the least squares offers estimates of the coefficients, the unimportant predictors, as expected, remain in the model with small, nonzero coefficients. The ridge penalty shrinks the estimates compared to the least squares estimates, but fails to eliminate the noise variables. Conversely, the lasso is able to eliminate the false predictors X_2, X_5, and X_6, while retaining the true predictors X_1, X_3, and X_4. This example shows that using the lasso provides a more efficient way to perform model selection. However, it is important to note that the lasso penalty often produces biased estimates, since all coefficients, even important ones, shrink towards the origin. The magnitude of selected lasso coefficients will be underestimated due to this shrinkage.

An extension of the lasso, the adaptive lasso (Zou 2006), seeks to minimize this bias. The adaptive lasso applies a weight $\bar{\boldsymbol{w}}$ to the lasso penalization, seeking to minimize:

$$\sum_{i=1}^{N}(y_i - \sum_j \boldsymbol{\beta}_i \boldsymbol{x}_{ij})^2 + \lambda \sum_j \bar{\boldsymbol{w}}|\beta_j|$$

where $\bar{\boldsymbol{w}} = 1/|\hat{\beta}|^{\gamma}$ and $\hat{\beta}$ is usually the ordinary least squares estimate and $\gamma > 0$. It can be seen that, with this weight, β_j's with small values will be further penalized towards 0, further reducing the number of parameters in the model. Conversely, large and important coefficients will be minimally penalized. The adaptive lasso has been shown to have have the oracle properties defined by Fan and Li (2001): the adaptive lasso consistently selects true variables in a known model and has asymptotic normality.

2.2.2 Smoothly Clipped Absolute Deviation (SCAD) Penalty

Consider the penalized regression problem (2.2). Due to the biased results of lasso, Fan and Li (2001) sought to create a penalty function that gave unbiased and sparse results, and was a continuous function. The Smoothly Clipped Absolute Deviation (SCAD) penalty for coefficient vector $\boldsymbol{\beta}$ by its derivative

$$P'_\lambda(\beta) = \begin{cases} \text{sgn}(\beta)\lambda, & \text{if } x < 0 \\ \text{sgn}(\beta)\frac{(a\lambda-|\beta|}{a-1}, & \text{if } \lambda < |\beta| \leq a\lambda. \\ 0, & \text{if } |\beta| > a\lambda \end{cases}$$

for $a > 2$ and $\lambda > 0$. The resulting penalty function is:

$$P'_\lambda(\beta) = \begin{cases} \lambda|\beta|, & \text{if } x < 0 \\ \frac{-(\beta^2-2a\lambda|\beta|+\lambda^2)}{2(a-1)}, & \text{if } \lambda < |\beta| \leq a\lambda. \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta| > a\lambda \end{cases}$$

The penalty function is a quadratic spline function, dependent on two tuning parameters a and λ. The results were found to be relatively insensitive to the parameter a. Through cross-validation, $a = 3.7$ was found give satisfactory results consistently. The second tuning parameter λ can also be found through cross validation for given data.

The advantages of the SCAD penalty are that it is continuous at all points except 0 and that it produces results with low bias. Similar to the adaptive lasso, when β is large, there is little penalization compared to when β is small. This will ensure that unimportant variables will be penalized heavily while leaving important variables relatively unpenalized. The SCAD penalty particularly outperforms the lasso penalty in selecting important variables while eliminating unimportant variables when the variance of the data is large. Additionally, the SCAD penalty was also shown to have oracle properties.

2.2.3 Review of Other Penalization Methods

Similar to the lasso and SCAD regression, there have been many penalized regularization methods to select and estimate fixed effects. Zou and Hastie (2005) created the elastic net, which combines lasso and Ridge penalization methods and proposed the algorithm to solve the elastic net efficiently. Bondell and Reich (2008) created the octagonal shrinkage and clustering algorithm for regression (OSCAR). This method has the ability to select important variables among a set of highly correlated predictors. However, these methods for selection of fixed effects do not take both the correlation and covariance structure of the random effects into account. Ignoring or underestimating covariance structure can lead to biased results of the variance estimates for the ordinary least squares of fixed effects (Lange and Laird 1998).

2.3 Review of Methods for Selection of Random Effects

In order to select important random effects in a model, Stram and Lee (1994) discuss the use of likelihood ratio tests in testing for nonzero variance components linear mixed effects models. Lin (1997) proposed a global score test to test the null hypothesis that all variance components were equal to 0, then individual score tests are determined for each random effect and estimation of the variance components can be made. Hall and Praestgaard (2001) place constraints on the score tests to select and estimate important random effects. Chen and Dunson (2003) use a Cholesky decomposition of the covariance matrix and Bayesian methods to select and estimate random effects variances. Foster et al. (2009) develop a lasso-based method for selection of random effects. These methods, however, only consider the random effects and do not have the ability to select or estimate the fixed effects.

2.4 Information Criteria for Model Selection

Information criteria methods are among the most popular methods of model selection, as they have the ability to select both fixed and random effects in a linear mixed effects model. Two of these methods are Akaike Information Criteria (AIC, Akaike 1973) and Bayes Information Criteria (BIC, Schwartz 1978). For these methods, the likelihood L is found for every combination of parameters in the model. From there, a penalization for the number of parameters in the model is added to $-2\ln L$. The goal is to find the minimum of the following:

$$\text{AIC} = -2\ln L + 2k$$

$$\text{BIC} = -2 \ln L + k \ln(n)$$

where k is the number of parameters, and n is the number of observations. The BIC method places a larger penalization on the number of parameters than AIC. The information criteria seek to find a balance between creating a model with good fit for the data and with having a small set of parameters.

While these methods are effective in choosing a parsimonious set of fixed and random effects that give the best fit, they can be burdensome as the number of possible parameters increases. For p fixed effect parameters and q random effect parameters, the number of models that need to be compared for information criteria is 2^{p+q}. As p and q increase, the number of possible models increases exponentially. Thus, for a large number of predictors, AIC and BIC are inefficient methods of variable selection.

2.5 Summary of Previous Model Selection Methods

While these methods are effective in performing model selection, they are not ideal for use in LME models. The penalization methods for fixed effects do not take the random effects into account, and can lead to inaccurate estimates. The random effects methods do not consider the fixed effects. Information criteria can find models with fixed and random effects, but are extremely inefficient for problems with a large number of predictors. Recently, there have been other methods used to select both fixed and random effects more efficiently. Jiang et al. (2008) consider a method to select models with important predictors that doesn't rely on minimizing a criterion function. In this method, a statistical ”fence” is created to eliminate incorrect models. From there, an optimal model is selected from the remaining models on the right side of this fence. This method eases the burden that is common with the information criterion methods.

Three new methods have been created in recent years to simultaneously select fixed and random effects. The latter sections of this thesis will describe and evaluate these three methods.

2.6 Trial of Activity in Adolescent Girls (TAAG)

Physical inactivity has been identified as a risk factor for obesity, or a high percentage of body fat, especially in adolescents (Pietiläinen et al. 2008). The prevalence of childhood overweight has been increasing drastically in the United States (Ogden et al. 2002). The resulting health problems that can occur from physical inactivity and obesity, such as type 2 diabetes, high blood pressure, and sleep disorders, have been on the rise in children and adolescents in recent years (Daniels et al. 2005). It has been recommended by the Council on Sports Medicine and Fitness and Council on School Health (2006) that increasing physical activity in children and adolescents can be effective in reducing the prevalence of obesity and the resulting health problems later on in life.

Among black and white girls, physical activity declines as a child ages through adolescence (Kimm et al. 2002). This decline in physical activity is more prevalent in girls than in boys (Sallis et al. 1996). Previous school-based interventions targeted at boys and girls have not been extremely successful. The Trial of Activity in Adolescent Girls (TAAG) was a school and community-based, multisite, interventional trial targeted at girls in order to lessen the typical declines in physical activity. The study was conducted at 26 sites across six geographically diverse areas in the United States, consisting of California, Minnesota, Maryland, Louisiana, South Carolina, and Arizona. Data was collected at two time points in the spring of 2003, for girls in the 6th grade and in the spring of 2005, when the girls were in 8th grade. The intervention and control groups were assigned randomly in 2003. The program was designed to create environments in schools and the surrounding community that encouraged physical activity and to give cues or messages that incentivize physical activity (Webber et al. 2008). The purpose of the intervention was to reduce the declines in Moderate to Vigorous Physical Activity (MVPA) that normally occurs in adolescent girls.

As a part of the TAAG study, data was collected to assess the sustainability of the program in the spring of 2006 in a new group of 8th grade girls. In the spring of 2009, followup data was collected from only the girls at the six Maryland TAAG sites for the Trial of Activity in Adolescent Girls 2 (TAAG 2). This data was collected when the 2006 8th grade group was in 11th grade. The purpose of the TAAG 2 study was determining factors at the individual, social, school, and neighborhood levels that may influence levels of MVPA in adolescent girls (Young et al. 2013). The analysis in this paper will use data from only these 2006 and 2009 time points. The main outcome of interest for this paper will be average MVPA minutes per day in the TAAG 2 study. The goal will be to select important fixed and random effects of interest from the TAAG data. The methods will be discussed in the latter sections of this thesis.

Chapter 3

Methods

3.1 Method 1 - Double Penalization

This section describes the method created by Li et al. (2012). It selects and estimates for the parameters of an LME through a regularization problem with two penalization functions: one for the fixed effects and one for the random effects.

Model Consider the LME given by equation (2.1) that is standardized to have a mean equal to 0 and a Euclidian norm equal to 0. The fixed effect intercept is removed from the model, but a random intercept b_{i0} remains in the model. The mean and variance of $\boldsymbol{Y}_i$ are $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ and $Var(\boldsymbol{Y}_i) = \sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T + \boldsymbol{I}_m)$.

Maximum Likelihood Estimation For $N > p$, a modified log-likelihood incorporates the restricted log-likelihood (Harville 1974). To maximize for β, Ψ, and σ^2,

$$\begin{aligned} \ell_{nM}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) &= -\frac{1}{2}\sum_{i=1}^{n} \log|\sigma^2 \boldsymbol{V}_i| - \frac{1}{2}\log\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \end{aligned} \tag{3.1}$$

where $\boldsymbol{V}_i = \boldsymbol{I}_m + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T$, the covariance structure of $\boldsymbol{Y}_i$. When $N \leq p$, the restricted term in (3.1) becomes singular. Therefore, when $N \leq p$, the following full log-likelihood to be maximized is:

$$\ell_{nF}(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n}\log|\sigma^2\boldsymbol{V}_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \quad (3.2)$$

Objective Function and Penalization A general formula for the objective function can be written:

$$Q_n(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) = \ell_n(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) - \lambda_1 P_1(\boldsymbol{\beta}) - \lambda_2 P_2(\boldsymbol{\Psi})$$

where P_1 is the penalization for the fixed effects, $P_2(\boldsymbol{\Psi})$ is the penalization for the random effects, and λ_1 and λ_2 are their non-negative tuning parameters, respectively. The log likelihood $\ell_n(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$ will be (3.1) or (3.2) when $N > p$ and $N \leq p$, respectively.

For the fixed effects, an adaptive L_1-norm, or adaptive lasso, penalty $J_1(\mathbf{fi})$ is applied (Zou 2006), where

$$P_1(\beta_j) = \sum_{j=1}^{p}\boldsymbol{w}_j|\beta_j|,$$

where $\boldsymbol{w}_j = \frac{1}{|\beta_j|}$ is a weight given by dividing by the estimated coefficient.

For the random effects, first, a Cholesky Decomposition is used to break $\boldsymbol{\Psi}$ into $\boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{L}^T$. The Cholesky factor of $\boldsymbol{\Psi}$, $\boldsymbol{L}$, is a unique lower triangular matrix with positive diagonals. Penalization will then be performed on $\boldsymbol{L}$. For any given $k \in \{1, \ldots, q\}$, finding a nonzero row (k) in $\boldsymbol{L}$, or $\boldsymbol{L}_{(k)}$ (and therefore the nonzero diagonal element Ψ_{kk}), will select the corresponding random effect b_k. Conversely, if $\boldsymbol{L}_{(k)}$ is equal to 0, then the corresponding Ψ_{kk} will equal 0, effectively removing the k^{th} random effect from the model. An adaptive weight is added to an L_2-norm penalty and this penalty is applied to the random effects, shrinking towards the coefficients toward zero:

$$P_2(\boldsymbol{L}) = \sum_{k=2}^{q} \mathbf{w_k} \sqrt{\mathbf{L_{k1}^2} + ...\mathbf{L_{kq}^2}},$$

where $\mathbf{w_k} = \frac{\mathbf{1}}{||\mathbf{L_{(k)}}||}$ is the weight given by dividing by the norm of the estimated coefficient. Again, this adaptive weight will help shrink small coefficients further towards zero, while leaving the important predictors unpenalized.

Algorithm First, σ^2 can be estimated (Lindstrom and Bates 1988) by:

$$\hat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

for $N > p$ and

$$\hat{\sigma}^2 = \tfrac{1}{N} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

for $N \leq p$. By inserting the estimated $\hat{\sigma}^2$ into (3.1) and (3.2), respectively, the objective function can then be solved for one fewer parameter.

The algorithm of estimating $\boldsymbol{\beta}$ and $\boldsymbol{L}$ is done in iterations, through maximizing the simplified objective function:

$$Q_n(\boldsymbol{\beta}, \boldsymbol{L}) = P_R(\boldsymbol{\beta}, \boldsymbol{L}) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \sqrt{L_{k1}^2 + ... + L_{kq}^2}$$

where $P_R(\beta, \boldsymbol{L})$ is the updated log-likelihood functions (3.1) or (3.2) with $\hat{\sigma}^2$ substituted into the equation.

The algorithm updates in iterations between two quadratic components until convergence. First, $\boldsymbol{L}$ is fixed and $\boldsymbol{\beta}$ is estimated, then $\boldsymbol{\beta}$ is fixed and $\boldsymbol{L}$ is updated. This is repeated until convergence. The step when $\boldsymbol{L}$ is fixed is similar to a lasso problem. However, when $\boldsymbol{\beta}$ is fixed, the random component must be split into two problems, estimating $\boldsymbol{L}$ and new parameter γ, which is updated from $\boldsymbol{L}$ estimates. The algorithm is completed as follows:

1. Initialize the parameters $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{L}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$

2. Update L_{kj} for iteration r by finding the maximum of the first quadratic component:

$$L_{kj}^{(r)} = \arg\max_{L_{kj}} P_R(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{L}) - \frac{\lambda_2^2}{4} \sum_{k=1}^{q} \frac{1}{(\gamma_k^{(r-1)})^2} \sum_{j=1}^{k} L_{kj}^2$$

3. Update γ_k:

$$\gamma_k^{(r)} = \sqrt{\frac{\lambda_2}{2} ||L_k^{(r)}||_2}$$

4. Update $\boldsymbol{\beta}$, using the LARS algorithm (Efron et al. 2004), the second quadratic component.

5. If the difference between $L_{kj}^{(r)}$ and $L_{kj}^{(r-1)}$ and between $\beta_j^{(r)}$ and $\beta_j^{(r-1)}$ are smaller than a specified amount, usually 10^{-5}, then the algorithm can end and estimates are obtained. If not, the process from step 2 can be continued for iteration $r + 1$.

3.2 Method 2 - Joint Penalization

This section describes the method introduced by Bondell et al. (2010). This method simultaneously selects and estimates fixed and random effects in an LME model using one joint penalization function for fixed and random effects.

Model Consider the LME model in equation (2.1). Using a modified Cholesky Decomposition (Chen and Dunson 2003), the covariance matrix $\mathbf{\Psi}$ is factorized as

$$\mathbf{\Psi} = \boldsymbol{D}\mathbf{\Gamma}\mathbf{\Gamma}'\boldsymbol{D}$$

where $\mathbf{\Gamma}$ is a $q \times q$ lower triangular matrix with 1's on the diagonal and whose (l, r)th element is given by γ_{lr} and $\boldsymbol{D} = \text{diag}(d_1, d_2, \ldots, d_q)$ is a diagonal matrix. After this decomposition, the LME model can be written

$$\boldsymbol{Y_i} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{D}\mathbf{\Gamma}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{3.3}$$

where it is now assumed that $\boldsymbol{Y}_i$ has been centered so that $\boldsymbol{X}_i^T\boldsymbol{X}_i$ and $\boldsymbol{Z}_i^T\boldsymbol{Z}_i$ represent correlation matrices and $\boldsymbol{b}_i$ is independently and identically distributed to $MVN(0, \sigma^2\boldsymbol{I}_m)$. The covariance matrix of $\boldsymbol{b}_i$ is now expressed in terms of vector $\boldsymbol{d} = (d_1, d_2, \ldots, d_q)^T$ and of the free elements of $\mathbf{\Gamma}$, denoted by vector $\gamma = (\gamma_{lr} : l = 1, \ldots, q : r = l + 1, \ldots, q)^T$. Setting any $d_l = 0$ will set the corresponding lth row and column of the covariance matrix $\mathbf{\Psi}$ to 0 and therefore remove the lth random effect from the model.

In the new model in (3.3), $\boldsymbol{Y}_i$ follows a normal distribution with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and variance $\boldsymbol{V}_i = \sigma^2(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{D}\boldsymbol{Z}_i + \boldsymbol{I}_m)$.

Maximum Likelihood Estimation Given $\boldsymbol{Y}$ and by treating $\boldsymbol{b}_i$ as observed, the log-likelihood function for the LME model is:

$$\ell_F(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{b}) = -\frac{N - nq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{\Gamma}}\boldsymbol{b}||^2 + \boldsymbol{b}^T\boldsymbol{b} \tag{3.4}$$

where $\boldsymbol{Z}$ is a block diagonal matrix of $\boldsymbol{Z}_i$, and $\widetilde{\boldsymbol{D}} = \boldsymbol{I}_n \otimes \boldsymbol{D}$ and $\widetilde{\boldsymbol{\Gamma}} = \boldsymbol{I}_n \otimes \boldsymbol{D}$, where $\otimes$ is the Kronecker product.

Objective Function and Penalization By minimizing the $||\boldsymbol{Y} - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{\Gamma}}\boldsymbol{b}||$ term in (3.4), the log-likelihood will be maximized. Therefore the objective function is:

$$Q_n(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{b}) = ||\boldsymbol{Y} - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{\Gamma}}\boldsymbol{b}||^2 + P_{\lambda_1}(\boldsymbol{\beta}, \boldsymbol{d})$$

where $P_{\lambda_1}(\boldsymbol{\beta}, \boldsymbol{d})$ is chosen to be an adaptive lasso penalty function with tuning parameter λ_1 such that:

$$P_{\lambda_1}(\boldsymbol{\beta}, \boldsymbol{d}) = \lambda_1\left(\sum_{j=1}^{p}\frac{|\beta_j|}{|\hat{\beta}_j|} + \sum_{k=1}^{k}\frac{|d_k|}{|\hat{d}_k|}\right)$$

where $\hat{\beta}$ and $\hat{d}$ are the ordinary least squares estimates. Rearranging the terms, the joint penalized objective function to be minimized in the algorithm is:

$$Q_F(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{b}) = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathrm{Diag}(\widetilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\boldsymbol{1}_q \otimes \boldsymbol{I}_n)\boldsymbol{d}||^2 + \lambda_1\left(\sum_{j=1}^{p}\frac{|\beta_j|}{|\hat{\beta}_j|} + \sum_{k=1}^{k}\frac{|d_k|}{|\hat{d}_k|}\right)$$

where $\boldsymbol{1}_q$ is a $q \times 1$ column vector of 1's.

Algorithm To solve for $\boldsymbol{\beta}, \boldsymbol{d}$, and $\boldsymbol{\gamma}$, the expectation-maximization (EM) algorithm (Laird and Ware 1982) is used. The algorithm consists of two steps. First, the conditional expectation of $Q_F(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{b})$ is taken (E-Step), then the objective function is minimized (M-Step) with respect to $(\boldsymbol{\beta}^T, \boldsymbol{d}^T, \boldsymbol{\gamma}^T)^T$. The overall process is as follows:

1. Let $\phi = (\boldsymbol{\beta}^T, \boldsymbol{d}^T, \boldsymbol{\gamma}^T)^T$, the vector of parameters and $\phi^{(r)}$ be the estimate of parameters at the r^{th} step. For $r = 0$, the REML estimates are chosen for the parameters

2. For the r^{th} step, first take the E-step, or find the conditional expectation of the objective function, assuming the random effects are unobserved:

$$g(\boldsymbol{\phi}|\boldsymbol{\phi}^{(r)}) = E_{\boldsymbol{b}|y,\phi^{(r)}}\left\{||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\text{Diag}(\widetilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\boldsymbol{1}_q \otimes \boldsymbol{I}_n)\boldsymbol{d}||^2\right\} + \lambda_1\left(\sum_{j=1}^{p}\frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{k=1}^{k}\frac{|d_k|}{|\bar{d}_k|}\right)$$

3. Complete the M-step by minimizing $g(\boldsymbol{\phi}|\boldsymbol{\phi}^{(r)})$ with respect to ϕ. This is completed by iterating between γ and $(\boldsymbol{\beta}, \boldsymbol{d})$.

4. The process is completed for step $r+1$ at step 2, unless convergence has occurred.

3.3 Method 3 - Independent Selection with Proxy Matrix

This section describes the method created by Fan and Li (2012). This method solves for the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}$ separately. A proxy matrix is substituted for the unknown true covariance structure during the selection and estimation of fixed and random effects.

Model Consider the model in (2.1). By stacking $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{b}_i$, and $\boldsymbol{\epsilon}_i$, notate $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{b}$, and $\boldsymbol{\epsilon}$. Let $\boldsymbol{Z} = \text{diag}\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n\}$ and $\tilde{\boldsymbol{\Psi}} = \text{diag}\{\boldsymbol{\Psi}, ..., \boldsymbol{\Psi}\}$ be block diagonal matrices. The fixed effect predictors $\boldsymbol{X}$ are standardized so that each column has norm $\sqrt{n}$. The LME model becomes:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

Maximum Likelihood Estimation of Fixed Effects The MLE for fixed effects can be found by maximizing the joint density function of $\boldsymbol{Y}$ and $\boldsymbol{b}$:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{b}) = {} & (2\pi\sigma)^{-(n+qm)/2} |\tilde{\boldsymbol{\Psi}}|^{-1/2} \\ & \times \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z} - \frac{1}{2} \boldsymbol{b}^T \tilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{b} \right\} \end{aligned}$$

Expressing the MLE for $\boldsymbol{b}$ in terms of a given $\boldsymbol{\beta}$, is $\hat{\boldsymbol{b}}(\boldsymbol{\beta}) = \boldsymbol{B}_z(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, where $\boldsymbol{B}_z = (\boldsymbol{Z}^T\boldsymbol{Z} + \sigma^2 \tilde{\boldsymbol{\Psi}}^{-1})^{-1} \boldsymbol{Z}^T$. By inserting $\hat{\boldsymbol{b}}(\boldsymbol{\beta})$, the MLE for $\boldsymbol{b}$ in terms of $\boldsymbol{\beta}$, the likelihood function for the fixed effects $\boldsymbol{\beta}$ can be expressed as:

$$\ell_n(\boldsymbol{\beta}, \hat{\boldsymbol{b}}(\boldsymbol{\beta})) = \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{P}_z (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \tag{3.5}$$

where $\boldsymbol{P}_z = (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{B}_z)^T(\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{B}_z) + \sigma^2\boldsymbol{B}_z^T\tilde{\boldsymbol{\Psi}}^{-1}\boldsymbol{B}_z$. Finding the $\boldsymbol{\beta}$ that maximizes (3.5) will give the fixed effects solution.

Objective Function and Penalization for Fixed Effects A general formula for the objective function of fixed effects is written:

$$Q_n(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{P}_z(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{p} P_{\lambda_1}(|\boldsymbol{\beta}|) \tag{3.6}$$

where the goal is to minimize $Q_n(\boldsymbol{\beta})$. It is required that the penalty function $P_{\lambda_1}(|\boldsymbol{\beta}|)$ is concave and increasing, so a smoothly clipped absolute deviation (SCAD, Fan and Li 2001) penalty function is chosen with tuning parameter $\lambda_1 > 0$.

Proxy Matrix for Fixed Effects Because $\boldsymbol{P}_z$ is dependent on the unknown covariance matrix $\tilde{\boldsymbol{\Psi}}$ and unknown variance σ^2, a proxy matrix $\widetilde{\boldsymbol{P}}_z = (\boldsymbol{I} + \boldsymbol{Z}\mathcal{M}\boldsymbol{Z}^T)$ is substituted for $\boldsymbol{P}_z$ with $\mathcal{M} = \log(N)\boldsymbol{I}$. Using this $\mathcal{M}$, the proxy matrix $\boldsymbol{P}_z$ satisfies a condition of decreasing minimal signal decay strength as sample size increases. It also satisfies constraints placed on the proxy matrix $\widetilde{\boldsymbol{P}}_z$ to ensure that the model selection has the oracle property. With this proxy matrix substituted into (3.6), the optimization problem becomes a quadratic problem which can be solved using the LARS algorithm (Efron et al. 2004).

Objective Function and Penalization for Random Effects The number of random effects q is allowed to increase with sample size n. For $\boldsymbol{P}_x = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, the objective function for the random effects is

$$Q_n(\boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{Zb})^T \boldsymbol{P}_x (\boldsymbol{y} - \boldsymbol{Zb}) + \sigma^2 \boldsymbol{b}^T \tilde{\boldsymbol{\Psi}}^+ \boldsymbol{b}$$

where $\tilde{\boldsymbol{\Psi}}^+$ is the Moore-Penrose generalized inverse of $\tilde{\boldsymbol{\Psi}}$. Adding a penalty, the regularization problem is created:

$$Q_n(\boldsymbol{b}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{Zb})^T \boldsymbol{P}_x (\boldsymbol{y} - \boldsymbol{Zb}) + \frac{1}{2}\sigma^2 \boldsymbol{b}^T \tilde{\boldsymbol{\Psi}}^+ \boldsymbol{b} + n \sum_{k=1}^{q_n} P_{\lambda_2}(b_k) \tag{3.7}$$

where $P_{\lambda_2}(b_k)$ is the SCAD penalty function with parameter λ_2. In reality, the covariance matrix $\tilde{\boldsymbol{\Psi}}$ and the variance σ^2 are unknown, so again, a proxy matrix $\mathcal{M}$ is substituted for $\sigma^{-2}\tilde{\boldsymbol{\Psi}}$ so the regularization problem becomes:

$$Q_n(\boldsymbol{b}) = \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{Zb})^T \boldsymbol{P}_x (\boldsymbol{y} - \boldsymbol{Zb}) + \tfrac{1}{2}\sigma^2 \boldsymbol{b}^T \mathcal{M}^{-1} \boldsymbol{b} + n \textstyle\sum_{k=1}^{q_n} P_{\lambda_2}(b_k)$$

Minimizing this equation gives an estimate of the random effects parameter vector $\hat{\boldsymbol{b}}$. Note that once the proxy matrix is substituted into the objective function, this method does not require knowledge of the fixed effect parameter $\boldsymbol{\beta}$.

Proxy Matrix for Random Effects Again, $\mathcal{M} = (\log n)\boldsymbol{I}$ is chosen to satisfy constraints placed on the proxy matrix. Substituting this proxy matrix into (3.7) creates a quadratic optimization problem similar to the the adaptive elastic net (Zou and Hastie 2005). This allows the problem to be solved using existing quadratic algorithms.

It should be noted that using $(\log n)\boldsymbol{I}$ ignores correlations among the random effects, which could introduce bias into the estimation of the covariance matrix. However, although there may be a biased covariance matrix estimate, it avoids errors caused by estimating a large number of parameters. The authors argue that the overall error caused by the accumulation of these errors from each parameter estimate would give poorer results than by using the proxy matrix.

Chapter 4

Analysis of Simulated Data

Experiments of six simulated data situations will be conducted to compare the effectiveness and accuracy of the three methods. All of the simulations will represent data sets where the number of observations N are greater than the number of predictors p. For all methods, tuning parameters are chosen through grid search to find the λ's that result in the lowest BIC. Each simulation consisted of 50 replicates.

4.1 Simulation 1

Setting This simulation generates a small study population of $n = 30$ clusters with $m = 5$ observations within each cluster, for observation $l \in \{1, \ldots, m\}$. There are 10 predictors in consideration, of which only four are important fixed effects and three of which are important random effects. The random effects will be selected from the same 10 predictors as the fixed effects, so $p = 10$ and $q = 10$. The true model is given by:

$$y_{il} = (1 + b_{i0}) + (3 + b_{i1l})x_{i1l} + (1.5 + b_{i2l})x_{i2l} + (2 + 0)x_{i5l} + (2 + b_{i10l})x_{i10l} \quad (4.1)$$

with $x_{ijl} \sim N(0,1)$ and $Corr(x_{ijl}, x_{ijl'}) = 0.5^{l-l'}$. The random effects $(b_{i0l}, b_{i1l}, b_{i2l}, b_{i10l})$ are generated from $MVN(0, \sigma^2\mathbf{R})$, with $\sigma = 0.8$ and

$$\boldsymbol{R} = \begin{pmatrix} 1.0 & 0.5 & 0.3 & 0.2 \\ 0.5 & 1.0 & 0.5 & 0.3 \\ 0.3 & 0.5 & 1.0 & 0.5 \\ 0.2 & 0.3 & 0.5 & 1.0 \end{pmatrix} \tag{4.2}$$

Result Simulation 1 results are in Tables 2 and 3. All methods correctly select all true fixed effects 100% of the time, with the exception of β_2 in Method 3, which was correctly selected 98 percent of the time. Only Method 1 selects all true random effects 100 percent of the time. Methods 2 and 3 still perform well, selecting the all correct random effects 92.67 percent and 70.67 percent of the time, respectively. Method 3 eliminates predictors the most, resulting in the smallest average model sizes for both fixed and random effects. In fact, Method 3's average model size is consistently less than the true model size, so when using Method 3, it is probable that true random effects are not selected.

4.2 Simulation 2

Setting Simulation 2 has the same true LME equation as Simulation 1, seen in (4.1). However, this will be a a larger data set, where $m = 8$ observations within $n = 200$ clusters for $p = 100$ and $q = 50$ predictors. There are still only four important fixed effects and three important random effects, as in equation (4.1).

All $\beta_i = 0$ for $\beta > 10$. The random effects are chosen from the first 50 fixed effects x_{ij}, where $j = \{1, \ldots, 50\}$, so $p = 100$ and $q = 50$.

Results The results of simulation 2 can be found in Tables 4 and 5. Method 2 was unable to complete analysis due to lack of memory, resulting in error "Error: cannot allocate vector of size 793.8 Mb." Method 3 was unable to complete analysis for the random effects, running for hours and then force closing MATLAB without results. Method 1 was able to select the true fixed and random effects in 100 percent of the simulations, while Method 3 selected the true fixed effects 100 percent of the time. Method 1 performed well at eliminating random effects, including false predictors in the model only 0.469 percent of the time. Both methods were able to eliminate noise variables well, only selecting false fixed effects less than one percent of the time.

4.3 Simulation 3

Setting This simulation generates a small study population of $n = 60$ clusters with $m = 3$ observations taken at within the cluster, with observation $l \in \{1, \ldots, m\}$. X_{ijl} and Z_{ikl} are generated from N(0,1) with $Corr(X_{ijl}, X_{ijl'}) = 0.5^{l-l'}$ and $Corr(Z_{ikl}, Z_{ikl'}) = 0.8^{l-l'}$. Random effects b_{ikl} are generated from $MVN(0, \sigma^2 R)$, where $\sigma = 0.8$ and

$$\boldsymbol{R} = \begin{pmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.5 \\ 0.3 & 0.5 & 1.0 \end{pmatrix} \tag{4.3}$$

There are $p = 10$ predictors for fixed and $q = 5$ predictors for random effects to choose from, with four true fixed effects and two true random effects. The true model is given by:

$$y_{il} = [1 + 3x_{i1l} + 1.5x_{i2l} + 2x_{i5l} + 2x_{i10l}] + [b_{i0l} + b_{i1l}z_{i1l} + b_{i5l}z_{i5l}] \qquad (4.4)$$

Results For fixed and random effects selected from different sets of predictors, the simulation 3 results are found in Tables 6 and 7. Again, all methods perform well when selecting fixed effects, correctly keeping true fixed effects 100 percent of the time. The methods do not perform as well selecting random effects as in simulation 1, but still correctly select true variables, on average, more than 70 percent of the time. Method 1 performs the best in this regard at 83 percent. Methods 2 and 3 both eliminate random predictors more heavily, resulting in average random model sizes that are less than the true model size.

4.4 Simulation 4

Setting Simulation 4 has the same true LME equation as Simulation 3, seen in (4.4). However, this will be a a larger data set, where $n = 600$ clusters with $m = 3$ observations within the clusters for $p = 50$ and $q = 10$ predictors. There are still only four important fixed effects and two important random effects, as in equation (4.4). All $\beta_i = 0$ for $i > 10$. All $b_i = 0$ for $i > 5$.

Results For larger data sets, simulation 4 results are found in Tables 8 and 9. Again, due to computational limitations, Method 2 was unable to complete both fixed and random effects, while Method 3 was unable to complete the analysis for the random effects. Methods 1 and 3 selected the true fixed effects 100 percent of the time. Method 3 was able to eliminate false fixed effects 100 percent of the time. Method 1 effectively selected true random effects an average of 100 percent of the time, while eliminating false random effects almost 94 percent of the time.

4.5 Simulation 5

Setting This simulation generates a small study population of $n = 60$ clusters with observations taken at $m = 3$ points within the clusters (observation $l \in \{1, \ldots, m\}$). However, this will simulate a multilevel study, where the fixed effects are selected at the individual level and the random effects are selected from group level predictors. At the individual level, $\boldsymbol{X}_{ijl}$ is generated from N(0,1) with $Corr(X_{ijl}, X_{ijl'}) = 0.5^{l-l'}$. At the group level, the predictors Z_{ikl} are nested within $g = 6$ groups, so all members of group g will have the same set of responses Z_{gl}. Z_{gkl} is generated from N(0,1) with $Corr(Z_{gkl}, Z_{gkl'}) = 0.8^{l-l'}$. Random effects $\boldsymbol{b}_{igkl}$ are generated from $MVN(0, \sigma^2 R)$, where $\sigma = 0.8$ and $\boldsymbol{R}$ is equation 4.3 above. The random effects have a subject-specific intercept b_{i0l} and predictor-associated $\boldsymbol{b}_{gk}$, for $k \in \{1, \ldots, q\}$. There are $p = 10$ predictors for fixed and $q = 5$ predictors for random effects to choose from, with four true fixed effects and two true random effects. The true model is given by:

$$y_{igl} = [1 + 3x_{i1l} + 1.5x_{i2l} + 2x_{i5l} + 2x_{i10l}] + [b_{i0l} + b_{g1l}z_{g1l} + b_{g5l}z_{g5l}] \tag{4.5}$$

Results For results from nested clustered designs, Simulation 5 results can be found in Tables 10 and 11. All methods again perform well in the identification of fixed effects, selecting true fixed effects 100 percent of the time. The methods do not perform as well selecting random effects as in simulations 1 and 3. Methods 1 and 2 perform the best at selecting random effects, and among the two, Method 1 is also better at eliminating noise random effects. Method 3 again eliminates random effects from the model the most. Method 3's models average just over one random effect in each model, about half the true size.

4.6 Simulation 6

Setting Simulation 6 has the same true LME equation as Simulation 5, seen in (4.6). However, this will be a a larger data set, where $n = 600$ clusters with $m = 3$ observations within clusters for $p = 50$ and $q = 20$ predictors. There are still only four important fixed effects and two important random effects, as in equation (4.6). All $\beta_i = 0$ for $i > 10$. All $b_i = 0$ for $i > 5$.

Results Simulation 6 results are listed in Tables 12 and 13. Again Methods 2 and 3 were limited by computational power. Methods 1 and 3 selected the true fixed effects 100 percent of the time, and both were able to eliminate noise fixed effects 100 percent of the time. Method 1 effectively selected true random effects an average of 100 percent of the time, while eliminating false random effects more than 95 percent of the time.

4.7 Simulation 7

Setting This simulation generates a small study population of $n = 60$ individuals with observations taken at $m = 3$ time points. It will simulate a multilevel study, where the fixed effects are selected at the individual level and the random effects are selected from group level predictors. It will also simulate a longitudinal study, with time $t = (1, 2, 3)$.

At the individual level $X_{ij}(t)$, for individual i and predictor j, $X_1 = t$ and $\boldsymbol{X}_{ij}$ for $j \in (2, \ldots, p)$ are generated from N(0,1) with $Corr(X_{ij}(t), X_{ij'}(t)) = 0.5^{t-t'}$. At the group level, the predictors Z_{gk} are nested within $g = 6$ groups, so all members of group g will have the same set of responses $Z_{gk}(t)$. $Z_{gk}(t)$ is generated from $N(0, 1)$ with $Corr(Z_{gk}(t), Z_{gk}(t')) = 0.8^{t-t'}$. Random effects $\boldsymbol{b}$ are generated from $MVN(0, \sigma^2 R)$, where $\sigma = 0.8$ and $\boldsymbol{R}$ is equation 4.3 above. The random effects have a subject-specific intercept $b_{i0}(t)$ and predictor-associated $\boldsymbol{b}_{igk}$, for $k \in \{1, \ldots, q\}$. There are $p = 10$ predictors for fixed and $q = 5$ predictors for random effects to choose from, with four true fixed effects and two true random effects. The true model is given by:

$$
\begin{aligned}
y_{ig}(t) &= 1 + \boldsymbol{\beta}\boldsymbol{X}_{ij}(t) + \boldsymbol{b}_{i0}(t) + \boldsymbol{b}_{igk}\boldsymbol{Z}_{gk}(t) \\
&= [1 + 3x_1(t) + 1.5x_{i2}(t) + 2x_{i5}(t) + 2x_{i10}(t)] + [b_{i0}(t) + b_{ig1}z_{g1}(t) + b_{ig5}z_{g5}(t)]
\end{aligned}
$$

Results For results from multilevel longitudinal designs, Simulation 7 results can be found in Tables 14 and 15. All methods again perform well in the identification of fixed effects, selecting true fixed effects 100 percent of the time. Methods 2 and 3 overestimate β_1, the coefficient for the time variable. Method 1 performs the best at selecting random effects. Method 2 selects one of the true random effects often while eliminating the other often. Both are selected more often than the noise random effects, however. Method 3 again eliminates random effects from the model the most. Method 3's models are on average less than the true model size. Again, Method 1 performs the best overall.

4.8 Summary of Simulation Studies

Overall, Method 1 was the most effective across data sets of different structures and sizes. It tends to underestimate the true parameter values of fixed and random effects, but this can be expected from penalized optimization problems. This can be remedied through re-estimating the selected model without penalization. Method 3 performed very well in the selection of fixed effects in all settings, but could not perform selection of random effects well for large data sets. Method 2 performed the worst. While able to select fixed effects accurately, it performed poorly for random effects in nested settings. It was also unable to complete any analysis for large data sets.

Chapter 5

Real Data Analysis

5.1 Data Description and Model Formuation

The data used in this analysis will include only the 2006 and 2009 time points of the TAAG 2 Maryland data set. The outcome variable of interest is the average minutes of MVPA per day in the adolescent girls at each time point. In total, there were 66 variables of interest for 551 subjects at two time points. A reference table of the variables can be found in Table 18.

Data was collected at the individual, social, school, and neighborhood levels. Examples of predictors at the individual level include BMI, percent body fat, self esteem measures, enjoyment of physical activity, and depression, among others. At the social level, predictors include measures of peer and family support, such as amount of encouragement received from members in the household or time spent home alone. At the school level, variables include policies for items such as physical education and transportation, as well as metrics regarding the schools' performance academically. At the neighborhood level, variables include proximity to their school, parks, and physical activity facilities as well as measures of safety in the neighborhood (Young et al. 2013).

Because the variable selection methods require that no data is missing, it is required that the subjects have measures at both time points. Those who were not present at both time points were removed from the set. Of the remaining data, missing data is imputed using the Sequential Regression Imputation Method (Raghunathan et al. 2001) through IVEware (Raghunathan et al. 2002) where possible. It is not possible to impute factors at the neighborhood level that used GIS data, so most neighborhood level variables were not included in the selection procedures. Questions regarding the subjects' perceptions of their neighborhood, however, are included.

For the i^{th} girl, the fixed effects variables $\boldsymbol{X}_{ij}$ to be considered will be from the individual, social, and neighborhood levels, where $j \in \{1 \ldots p\}$. Time will also be included for longitudinal consideration in the model. The predictors associated with the random effects will be used to generalize the variance components of the model. For girls in school g ($1 \leq g \leq 6$), the random effects will be selected from the school-level variables $\boldsymbol{Z}_{gk}$, where $k \in \{1 \ldots q\}$. For the school-level variables, only the data from the 8th grade middle school time point will be considered. Therefore, these predictors are not time-dependent.

For the outcome $\boldsymbol{y}_{ig}(t)$, or average daily MVPA at time $t \in (0, 1)$ for girl i in school g, the LME model is represented by

$$\boldsymbol{y}_{ig}(t) = \beta_0 + \theta t + \boldsymbol{\beta} \boldsymbol{X}_i + \boldsymbol{b}_{i0} + \boldsymbol{b}_{g0} + \boldsymbol{b}_i \boldsymbol{Z}_g + \epsilon_{ig},$$

where β_0 is the fixed effects intercept, θ is the parameter associated with time, $\boldsymbol{b}_i$ are the individual-specific random effects with intercept $\boldsymbol{b}_{i0}$, and $\boldsymbol{b}_{g0}$ is the school-specific random intercept. Individual-specific random effects $\boldsymbol{b}_i$ are distributed to

$N(0, \sigma^2 \mathbf{\Sigma}_i)$. School-specific random intercept $\boldsymbol{b}_{g0}$ is distributed to $N(0, \sigma^2 \mathbf{\Sigma}_g)$. The error term is distributed $\epsilon \sim N(0, 1)$.

5.2 Results of Real Data Analysis

Method 1 and Method 3 were able to complete analysis, while Method 2 was not able to due to incomplete rank of the Z matrix. The results of Methods 1 and 3 can be found in Tables 16 and 17, with, for comparison, results of selected fixed and random effects from the individual time points analyzed by Young et al. (2013). For Method 3, the fixed effects $\boldsymbol{X}$ were scaled so each column had a norm of $\sqrt{n}$. Following this, the method selected 40 fixed effects and 1 random effect. Only 6 fixed effects were eliminated from the model, while 18 random effects were eliminated. In the simulations, the random effects model sizes were generally smaller than the true model size. Therefore, it is likely true random effects were eliminated from this model.

Method 1 was more successful at creating a more parsimonious model. First, the data was standardized such that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are scaled that have zero mean and unit Euclidean norm. Of the original 47 fixed effects and 19 random effects, Method 1 selected three fixed effects and two random effects. The fixed effects selected were (1) self-management strategies (MSQBOD_F), (2) perceived barriers (MSQBOD_I), and (3) support from friends (MSQBOD_OB). Because of the tendency of Method 1 to underestimate the parameter values, the selected model was updated to a non-penalized fitted LME model using the lme4 package in R, with corresponding p-values for the fixed effects. Although not selected, time was included in the re-estimated model to determine the longitudinal effects of time on MVPA.

After re-estimation, the results suggest that there is a positive association between self-management strategies and friend support with MVPA (β_{MSQBOD_F} = 0.12 and $\beta_{MSQBOD\_OB} = 0.43$). There was also a negative association between perceived barriers and MVPA ($\beta_{MSQBOD\_F} = -0.21$). Perceived barriers ($p =< .001$) and friend support ($p < .001$) were both highly significant, and self-management strategies was significant at the $\alpha = 0.05$ significance level ($p = .05$). There was negative, but nonsignificant, effect of time on MVPA ($\beta_{time} = -0.52$, $p = .35$). All three fixed effects were selected in the previous study, and the results of perceived barriers and friend support show the same direction of association. Based on the results, it would be suggested to offer programs or interventions for improving self-management strategies, for reducing barriers to physical activity, and for encouraging peers to give each other support in participating in physical activities.

One random effect was selected out of the 18 original variables. The random effect selected was PPIC19, indicating whether the schools offered interscholastic and intramural physical activity programs. This was selected for the 2009 11th grade time point in Young et al. (2013), but not in the 2006 8th grade model. Following re-estimation, the results from this analysis suggest that there is substantial variation in MVPA from girl to girl associated with whether their middle school offered interscholastic or intramural physical activity programs programs ($\hat{\sigma} = 6.68$).

Chapter 6

Conclusion

This thesis has presented three new methods for variable selection in LME models. The method proposed by Bondell et al. (2010) was one of the first to simultaneously select fixed and random effects in LME models. While it can effectively select fixed effects, it performs less accurately in selecting random effects when the data becomes nested. Further, it cannot perform analysis when the data is time independent, as was the case of the random effects in the TAAG data. Also, the EM algorithm that Method 2 uses is an inefficient way to solve optimization problems. As data sets get larger through increases in sample size or number of predictors, the slow rate of convergence of the EM algorithm becomes inefficient and even implausible with limited computing resources. An option for high-dimensional data, where $N \leq p$, would be to reduce the number of fixed effect parameters using previous methods, such as the lasso, while ignoring the random effects. Following this, the method could be applied to the random effects and the selected fixed effects. However, due to its slow rate of convergence, this method and its use of the EM algorithm would not be able ideal for use on high-dimensional data,

Method 3, proposed by Fan and Li (2012) can accurately and quickly select fixed effects in LME models. In simulations it performed excellently in not only selecting true predictors, but it is very effective at removing noise fixed effects variables from the model as well. However, the performance with the TAAG data was inconsistent with the sparse results displayed in the simulations. The use of the proxy matrix requires certain conditions to be satisfied. Notably, for fixed effects $\boldsymbol{X}$ and random effects $\boldsymbol{Z}$, the signal and noise variables must not be highly correlated. By using a proxy matrix, the correlation between variables is ignored. In cases of highly correlated signal and noise predictors, the use of the proxy matrix could introduce bias that can hinder the model selection oracle property. There are potentially many correlated variables in the TAAG data set that could violate this condition set in order to use the proxy matrix, which may have caused the poor results. Additionally, the performance of Method 3 in selecting random effects can be troublesome, as it tends to under-select true models. This can lead to models that are missing important random effects.

For high-dimensional data, it is necessary to first reduce the number of fixed effects parameters while ignoring the random effects through previous regularization methods. Next the random effects can be selected using the chosen fixed effects from the previous step. Finally, these fixed effects can be selected and re-estimated using the selected random effects from the second step.

Based on the results of the simulations, Method 1 by Li et al. (2012) is clearly the optimal method of the three. It selects the true model consistently while eliminating noise variables effectively. Additionally, it's new algorithm for solving the optimization problem is much more efficient than previous methods, such as the EM algorithm. By splitting the optimization problem into two penalized quadratic algorithms, convergence can be reached much quicker than previous methods. Additionally, this method can be used with high-dimensional data. All that is needed is to use the maximum likelihood approach in equation (3.2), instead of the REML-modified equation (3.1).

The benefits of these methods can surely prove invaluable to researchers. This is especially true in the field of public health, where longitudinal data is often used and is vital for understanding temporal trends of health outcomes. The temporal trends can provide a deeper understanding of biological, social, or environmental processes that can lead to progress in the discovery and improvement of health risks. With the methods introduced in this thesis, it is possible to efficiently and select important fixed and random effects from large, complex sets of predictors. This can aid and advance the field of public health data greatly in the future, especially as technology and data collection methods improve.

Table 1: Example of Model Selection Using Lasso

Variable	True Value	Least Squares	Ridge	Lasso
Intercept	1	1.04	0.96	0.90
X_1	0.5	0.69	0.58	0.52
X_2	0	0.13	0.06	0.00
X_3	1.5	1.44	1.36	1.26
X_4	0.5	0.39	0.43	0.41
X_5	0	0.11	0.04	0.00
X_6	0	-0.16	-0.003	0.00

Table 2: Simulation 1 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.37	0.32	100.00	0.53	0.42
2	100.00	0.89	0.26	100.00	0.49	0.37
3	26.00	0.09	0.05	28.00	0.09	0.10
4	22.00	0.05	0.03	18.00	0.07	0.06
5	100.00	1.68	0.18	56.00	0.17	0.18
6	14.00	0.02	0.01	18.00	0.06	0.06
7	6.00	0.01	0.00	18.00	0.02	0.02
8	14.00	0.03	0.02	20.00	0.03	0.03
9	14.00	0.03	0.02	16.00	0.05	0.04
10	100.00	1.32	0.30	100.00	0.48	0.38
Method 2						
1	100.00	3.02	0.18	100.00	0.90	0.42
2	100.00	1.50	0.21	100.00	0.90	0.46
3	62.00	0.03	0.13	56.00	0.60	0.47
4	70.00	-0.03	0.11	50.00	0.63	0.46
5	100.00	2.00	0.14	40.00	0.64	0.49
6	72.00	-0.02	0.12	36.00	0.59	0.44
7	66.00	0.01	0.10	52.00	0.58	0.44
8	66.00	-0.03	0.10	28.00	0.58	0.40
9	54.00	-0.00	0.07	42.00	0.65	0.48
10	100.00	2.02	0.19	78.00	0.87	0.60
Method 3						
1	100.00	3.03	0.25	74.00	-	-
2	98.00	1.52	0.30	72.00	-	-
3	0.00	0.00	0.00	2.00	-	-
4	0.00	0.00	0.00	10.00	-	-
5	100.00	1.94	0.18	10.00	-	-
6	0.00	0.00	0.00	2.00	-	-
7	0.00	0.00	0.00	2.00	-	-
8	0.00	0.00	0.00	4.00	-	-
9	0.00	0.00	0.00	14.00	-	-
10	100.00	1.99	0.21	66.00	-	-

Table 3: Simulation 1 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.96	7.88	3.98
Avg Model Size Random (True = 3)	4.74	5.8	2.56
Percent True β Included	100.00	100.00	99.33
Percent True D Included	100.00	92.67	70.67
Percent False β Included	16.00	64.58	0.00
Percent False D Included	21.75	35.00	6.29

Table 4: Simulation 2 Results - Parameter Estimates

Covariate	Fixed Effects $\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	Random Effects $\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.52	0.15	100.00	0.53	0.06
2	100.00	0.97	0.14	100.00	0.51	0.06
3	4.00	0.02	0.00	8.00	0.01	0.00
4	10.00	0.01	0.00	0.00	0.00	0.00
5	100.00	1.89	0.05	32.00	0.04	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00
7	2.00	0.01	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	2.00	0.01	0.00
10	100.00	1.47	0.14	100.00	0.54	0.09
11	2.00	0.01	0.00	4.00	0.01	0.00
12-50	0.00	0.00	0.00	0.00	0.00	0.00
51-100	0.00	0.00	0.00	-	-	-
Method 2						
1-50	*	*	*	*	*	*
51-100	*	*	*	-	-	-
Method 3						
1	100.00	3.01	0.08	*	*	*
2	100.00	1.48	0.07	*	*	*
3	2.00	0.14	0.02	*	*	*
5	100.00	2.00	0.05	*	*	*
8	2.00	-0.17	0.02	*	*	*
9	4.00	0.01	0.04	*	*	*
10	100.00	1.99	0.07	*	*	*
15	2.00	-0.16	0.02	*	*	*
22	4.00	0.03	0.03	*	*	*
28	2.00	-0.13	0.02	*	*	*
29	2.00	0.15	0.02	*	*	*
36	4.00	-0.01	0.03	*	*	*
40	2.00	0.12	0.02	*	*	*
45	2.00	0.14	0.02	*	*	*
49	2.00	0.14	0.02	*	*	*
54	2.00	0.14	0.02	-	-	-
56	2.00	0.17	0.02	-	-	-
66	2.00	0.14	0.02	-	-	-
72	2.00	0.12	0.02	-	-	-
73	2.00	-0.13	0.02	-	-	-
75	2.00	0.13	0.02	-	-	-

Table 4 – Continued

Covariate	$\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
76	2.00	-0.17	0.02	-	-	-
77	2.00	0.13	0.02	-	-	-
81	2.00	0.15	0.02	-	-	-
88	2.00	-0.13	0.02	-	-	-
89	2.00	-0.15	0.02	-	-	-
90	2.00	-0.15	0.02	-	-	-
98	2.00	-0.15	0.02	-	-	-

*Could not complete due to computational limitations

Table 5: Simulation 2 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Method 3			
Avg Model Size Fixed (True = 4)	4.22	*	4.54
Avg Model Size Random (True = 3)	3.46	*	*
Percent True β Included	100.00	*	100.00
Percent True D Included	100.00	*	*
Percent False β Included	0.229	*	0.563
Percent False D Included	0.469	*	*

*Could not complete due to computational limitations

Table 6: Simulation 3 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.52	0.18	90.00	0.33	0.30
2	100.00	1.03	0.15	46.00	0.20	0.20
3	32.00	0.00	0.00	54.00	0.19	0.22
4	36.00	0.00	0.01	36.00	0.18	0.17
5	100.00	1.36	0.16	86.00	0.35	0.30
6	4.00	0.00	0.00	-	-	-
7	4.00	0.00	0.00	-	-	-
8	2.00	0.00	0.00	-	-	-
9	14.00	0.00	0.00	-	-	-
10	100.00	1.24	0.16	-	-	-
Method 2						
1	100.00	3.01	0.14	94.00	0.94	0.37
2	100.00	1.45	0.18	20.00	0.71	0.33
3	24.00	-0.02	0.04	2.00	1.18	0.17
4	20.00	0.03	0.03	14.00	0.72	0.26
5	100.00	1.97	0.14	56.00	0.90	0.47
6	14.00	0.08	0.05	-	-	-
7	16.00	0.04	0.04	-	-	-
8	10.00	0.04	0.04	-	-	-
9	18.00	-0.00	0.06	-	-	-
10	100.00	1.94	0.13	-	-	-
Method 3						
1	100.00	3.04	0.23	66.00	-	-
2	100.00	1.44	0.24	18.00	-	-
3	6.00	0.19	0.12	12.00	-	-
4	6.00	-0.49	0.12	20.00	-	-
5	100.00	2.01	0.20	76.00	-	-
6	8.00	-0.01	0.15	-	-	-
7	2.00	0.39	0.05	-	-	-
8	4.00	0.03	0.10	-	-	-
9	4.00	0.39	0.08	-	-	-
10	100.00	1.95	0.16	-	-	-

Table 7: Simulation 3 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.92	5.02	4.30
Avg Model Size Random (True = 2)	2.44	1.86	1.26
Percent True β Included	100.00	100.00	100.00
Percent True D Included	83.00	75.00	71.00
Percent False β Included	15.33	17.00	5.00
Percent False D Included	26.00	18.00	25.00

Table 8: Simulation 4 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}$%	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}$%	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.93	0.03	100.00	0.69	0.13
2	100.00	1.44	0.03	2.00	0.04	0.00
3	12.00	0.00	0.00	8.00	0.01	0.00
4	4.00	0.00	0.00	8.00	0.02	0.00
5	100.00	1.91	0.03	100.00	0.63	0.00
6	0.00	0.00	0.00	10.00	0.02	0.10
7	0.00	0.00	0.00	4.00	0.02	0.00
8	4.00	0.01	0.01	4.00	0.02	0.00
9	10.00	0.00	0.00	2.00	0.03	0.00
10	100.00	1.89	0.03	12.00	0.01	0.00
11	4.00	0.00	0.00	-	-	-
12-50	0.00	0.00	0.00	-	-	-
Method 2						
1-10	*	*	*	*	*	*
11-50	*	*	*	-	-	-
Method 3						
1	100.00	3.00	0.05	*	*	*
2	100.00	1.50	0.05	*	*	*
3	0.00	0.00	0.00	*	*	*
4	0.00	0.00	0.00	*	*	*
5	100.00	2.00	0.05	*	*	*
6	0.00	0.00	0.00	*	*	*
7	0.00	0.00	0.00	*	*	*
8	0.00	0.00	0.00	*	*	*
9	0.00	0.00	0.00	*	*	*
10	100.00	2.00	0.04	*	*	*
11-50	0.00	0.00	0.00	-	-	-

* Could not complete due to computational limitations

Table 9: Simulation 4 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.34	*	4.00
Avg Model Size Random (True = 2)	2.50	*	*
Percent True β Included	100.00	*	100.00
Percent True D Included	100.00	*	*
Percent False β Included	3.09	*	0.00
Percent False D Included	5.55	*	*

*Could not complete due to computational limitations

Table 10: Simulation 5 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}$%	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}$%	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.46	0.19	78.00	0.41	0.21
2	100.00	1.04	0.21	36.00	0.23	0.15
3	34.00	0.01	0.01	36.00	0.25	0.16
4	42.00	0.01	0.01	32.00	0.24	0.14
5	100.00	1.31	0.15	60.00	0.32	0.19
6	12.00	0.00	0.00	-	-	-
7	2.00	0.00	0.00	-	-	-
8	0.00	0.00	0.00	-	-	-
9	6.00	0.00	0.00	-	-	-
10	100.00	1.21	0.15	-	-	-
Method 2						
1	100.00	3.00	0.12	94.00	0.91	0.59
2	100.00	1.47	0.16	50.00	0.71	0.48
3	44.00	0.01	0.07	44.00	0.81	0.48
4	42.00	0.03	0.07	26.00	0.73	0.36
5	100.00	1.99	0.10	42.00	0.79	0.42
6	32.00	0.04	0.06	-	-	-
7	38.00	-0.00	0.07	-	-	-
8	48.00	0.00	0.08	-	-	-
9	28.00	-0.02	0.05	-	-	-
10	100.00	1.98	0.10	-	-	-
Method 3						
1	100.00	2.97	0.17	50.00	-	-
2	100.00	1.50	0.18	16.00	-	-
3	2.00	0.41	0.06	20.00	-	-
4	0.00	0.00	0.00	16.00	-	-
5	100.00	2.00	0.14	52.00	-	-
6	2.00	0.39	0.06	-	-	-
7	0.00	0.00	0.00	-	-	-
8	2.00	-0.39	0.06	-	-	-
9	0.00	0.00	0.00	-	-	-
10	100.00	1.99	0.13	-	-	-

Table 11: Simulation 5 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.96	6.32	4.06
Avg Model Size Random (True = 2)	2.38	2.56	1.04
Percent True β Included	100.00	100.00	100.00
Percent True D Included	69.00	68.00	51.00
Percent False β Included	16.00	38.67	1.00
Percent False D Included	34.67	40.00	17.33

Table 12: Simulation 6 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.77	0.04	90.00	0.45	0.32
2	100.00	1.30	0.04	24.00	0.12	0.08
3	0.00	0.00	0.00	12.00	0.20	0.12
4	0.00	0.00	0.00	14.00	0.12	0.06
5	100.00	1.69	0.05	88.00	0.48	0.34
6	0.00	0.00	0.00	12.00	0.18	0.13
7	0.00	0.00	0.00	6.00	0.13	0.05
8	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00
10	100.00	1.66	0.05	0.00	0.00	0.00
12	0.00	0.00	0.00	2.00	0.17	0.04
14	0.00	0.00	0.00	2.00	0.13	0.04
15	0.00	0.00	0.00	2.00	0.10	0.04
19	0.00	0.00	0.00	6.00	0.23	0.13
20	0.00	0.00	0.00	4.00	0.12	0.05
21-50	0.00	0.00	0.00	-	-	-
Method 2						
1-20	*	*	*	*	*	*
21-50	*	*	*	-	-	-
Method 3						
1	100.00	3.01	0.05	*	*	*
2	100.00	1.49	0.04	*	*	*
3	0.00	0.00	0.00	*	*	*
4	0.00	0.00	0.00	*	*	*
5	100.00	1.99	0.04	*	*	*
6	0.00	0.00	0.00	*	*	*
7	0.00	0.00	0.00	*	*	*
8	0.00	0.00	0.00	*	*	*
9	0.00	0.00	0.00	*	*	*
10	100.00	2.01	0.04	*	*	*
11-20	0.00	0.00	0.00	*	*	*
21-50	0.00	0.00	0.00	-	-	-

*Could not complete due to computational limitations

Table 13: Simulation 6 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.00	*	4.00
Avg Model Size Random (True = 2)	2.60	*	*
Percent True β Included	100.00	*	100.00
Percent True D Included	89.00	*	*
Percent False β Included	0.00	*	0.00
Percent False D Included	4.67	*	*

*Could not complete due to computational limitations

Table 14: Simulation 7 Results - Parameter Estimates

	Fixed Effects			Random Effects		
Covariate	$\hat{\beta}\%$	$\hat{\beta}$	$\hat{\beta}$ **Error**	$\hat{\sigma}\%$	$\hat{\sigma}$	$\hat{\sigma}$ **Error**
Method 1						
1	100.00	2.70	0.14	86.00	0.15	0.14
2	100.00	1.24	0.13	52.00	0.03	0.04
3	18.00	0.08	0.05	34.00	0.02	0.01
4	24.00	0.06	0.04	42.00	0.03	0.03
5	100.00	1.76	0.12	82.00	0.19	0.17
6	10.00	0.05	0.02	-	-	-
7	2.00	0.15	0.02	-	-	-
8	10.00	0.05	0.02	-	-	-
9	10.00	0.05	0.02	-	-	-
10	100.00	1.72	0.10	-	-	-
Method 2						
1	100.00	3.39	0.06	84.00	0.86	0.42
2	100.00	1.42	0.17	20.00	0.69	0.35
3	18.00	0.02	0.05	6.00	0.58	0.15
4	14.00	0.03	0.04	10.00	0.75	0.24
5	100.00	1.96	0.11	42.00	0.89	0.48
6	14.00	0.02	0.04	-	-	-
7	16.00	-0.02	0.05	-	-	-
8	18.00	0.06	0.04	-	-	-
9	18.00	0.00	0.03	-	-	-
10	100.00	1.94	0.14	-	-	-
Method 3						
1	100.00	3.44	0.10	64.00	-	-
2	100.00	1.49	0.19	16.00	-	-
3	8.00	0.21	0.14	14.00	-	-
4	2.00	-0.59	0.08	14.00	-	-
5	100.00	2.03	0.15	56.00	-	-
6	2.00	0.78	0.11	-	-	-
7	2.00	-0.59	0.08	-	-	-
8	0.00	0.00	0.00	-	-	-
9	0.00	0.00	0.00	-	-	-
10	100.00	2.02	0.16	-	-	-

Table 15: Simulation 7 Results - Summary

	Method 1 Double Penalty	**Method 2** Joint Penalty	**Method 3** Independent Selection
Avg Model Size Fixed (True = 4)	4.74	*	4.00
Avg Model Size Random (True = 2)	2.96	*	1.60
Percent True β Included	100.00	*	100.00
Percent True D Included	84.00	*	60.00
Percent False β Included	12.33	*	0.00
Percent False D Included	42.67	*	14.67

Table 16: TAAG 2 Data - Fixed Effects

	Method 1	Method 1 Re-estimate	Method 3	Young et al. (2013) Results	
				8^{th} Grade	11^{th} grade
	$\hat{\beta}$	$\hat{\beta}$ (p-value)	$\hat{\beta}$	$\hat{\beta}$ (p-value)	$\hat{\beta}$ (p-value)
time	-	-0.52(.35)*	1.01	-	-
COMBPAREDUC	-	-	-0.23	-	-
MSQBA5A	-	-	0.71	-	-
MSQBA5B	-	-	-1.22	-	-
MSQBA7	-	-	0.43	-	-
MSQBC1	-	-	-0.78	-	-
MSQBC2	-	-	-	-	-
MSQBC3	-	-	-0.32	-	-
MSQBM1	-	-	-0.19	0.39(.27)	-0.15(.72)
MSQBM2	-	-	-0.45	-0.10(.76)	-1.40(<.001)
MSQBM3	-	-	-0.23	-0.03(.86)	-0.04(.86)
MSQBM4	-	-	-	0.78(.05)	0.80(.08)
MSQBM5	-	-	0.58	-	-
MSQBM6	-	-	0.21	-	-
MSQBM7	-	-	-0.74	-	-
MSQBM8	-	-	0.74	-	-
MSQBM9	-	-	-0.14	-	-
MSQBM10	-	-	-0.77	-	-
MSQBR1	-	-	1.51	-	-
MSQBR2	-	-	-0.43	-	-
r1	-	-	-1.15	1.15(.36)	0.31(.83)
r2	-	-	1.18	2.25(.12)	-0.73(.68)
r3	-	-	-0.30	0.02(.99)	-0.83(.66)
BMI	-	-	-2.47		
PFAT3	-	-	1.71	-0.09(.08)	-0.06(.43)
MSQBA_DAD_MOM	-	-	-	-	-
MSQBOD_B	-	-	0.51	0.12(.04)	0.05(.43)
MSQBOD_DA	-	-	0.89	-	-
MSQBOD_DB	-	-	0.31	0.31(.20)	0.00(.99)
MSQBOD_E	-	-	0.25	-	-
MSQBOD_F	0.05	0.12(.05)	-	-0.04(.68)	0.01(.95)
MSQBOD_G	-	-	-	0.09(.34)	0.26(.03)
MSQBOD_H	-	-	1.08	-0.00(0.92)	-0.31(0.01)
MSQBOD_I	-0.09	-0.21(< .001)	-0.84	-0.20(.04)	−0.37(< .001)
MSQBOD˙JA	-	-	-0.15	-	-
MSQBOD_JB	-	-	0.17	0.00(0.92)	-0.02(0.06)
MSQBOD_K	-	-	0.24	0.57(.14)	-0.00(.99)

Table 14 – Continued

	Method 1 $\hat{\beta}$	Method 1 Re-estimate $\hat{\beta}$ (p-value)	Method 3 $\hat{\beta}$	**Young et al. (2013) Results** 8^{th} Grade $\hat{\beta}$ (p-value)	11^{th} grade $\hat{\beta}$ (p-value)
MSQBOD_LA	-	-	-0.26	-0.19(.35)	-0.11(0.65)
MSQBOD_LB	-	-	-0.75	-0.13(.39)	0.38(.04)
MSQBOD_LC	-	-	-0.52	-	-
MSQBOD_N	-	-	-0.43	-	-
MSQBOD_OA	-	-	-0.58	-	-
MSQBOD_OB	0.22	0.43(< .001)	1.31	0.32(.08)	0.28(.22)
MSQBOD_OC	-	-	-0.39	-0.08(.45)	0.07(.61)
MSQB80P	-	-	-	0.01(.80)	0.01(.89)
MSQBQ1	-	-	0.51	-	-
MSQBR34SUM	-	-	-0.52	-	-

* Not selected but included in re-estimated model

Table 17: TAAG 2 Data - Random Effects

	Method 1	Method 1 Re-estimate	Method 3	Young et al. (2013) Results	
				8^{th} Grade	
	$\hat{\sigma}$	$\hat{\sigma}$	Selected	$\hat{\beta}$ (p-value)	$\hat{\beta}$ (p-value)
Individual Level					
Intercept	3.97	7.61	-	-	-
MSMA4	-	-	-	-1.45($<$ .001)	0.11(.53)
MSMA5A	-	-	-	-0.85(.20)	0.32(.28)
MSMA5B	-	-	Yes	-	-
PPIC1C2	-	-	-	-0.41(.22)	-
PPIC18A	-	-	-	-5.63(.19)	
PPIC19	1.47	6.68	-	-	7.35(.01)
PPIC21	-	-	-	16.76($<$ .001)	3.08(.15)
PPIC22	-	-	-	-2.93(0.21)	-
PPIC34	-	-	-	-2.93(0.21)	5.79(.07)
School Level					
Intercept	0.09	0.93	-	-	-

Table 18: Reference Table for Fixed and Random Effects Predictors

Variable	Description
Fixed Effects	
time	time = (0,1) for 8th grade and 11th grade respectively
COMBPAREDUC	Parents' education combined
MSQBA5A	Employment status: father
MSQBA5B	Employment status: mother
MSQBA7	Receive free or low-cost lunches at school
MSQBC1	Difficulty getting home from school-based activity
MSQBC2	Difficulty getting to community-based activity
MSQBC3	Difficulty getting home from community-based activity
MSQBM1	Perceived places to go within walking distance of home
MSQBM2	Perceived sidewalks in neighborhood
MSQBM3	Perceived bike/walking trails in neighborhood
MSQBM4	Perceived safety to walk/jog in neighborhood
MSQBM5	Perceived walkers/bikers easily seen in neighborhood
MSQBM6	Perceived traffic in neighborhood
MSQBM7	Perceived frame in neighborhood
MSQBM8	Perceived seeing kids outside playing in neighborhood
MSQBM9	Perceived interesting things to look at in neighborhood
MSQBM10	Perceived well-lit neighborhood
MSQBR1	Grade began current middle school
MSQBR2	Currently taking PE
r1	Race: white
r2	Race: black
r3	Race: hispanic
BMI	BMI
BMI85	BMI above 85th percentile
BMI95	BMI above 95th percentile
PFAT3	Percent Fat
MSQBA_DAD_MOM	Number of parents living with
MSQBOD_B	Average time alone per week
MSQBOD_DA	Sports team participation at school
MSQBOD_DB	Sports team participation outside school
MSQBOD_E	Enjoyment of PA classes/lessons
MSQBOD_F	Self-management strategies
MSQBOD_G	Self-efficacy
MSQBOD_H	Enjoyment of PA
MSQBOD_I	Perceived barriers

Table 16 – Continued

Variable	Description
MSQBOD_JA	Outcome expectancy
MSQBOD_JB	Outcome expectancy value
MSQBOD_K	Enjoyment of PE
MSQBOD_LA	Positive PA school climate for teachers
MSQBOD_LB	Positive PA school climate for boys
MSQBOD_LC	PA norms
MSQBOD_N	Access to recreational facilities
MSQBOD_OA	Provides social support
MSQBOD_OB	Friend support
MSQBOD_OC	Family support
MSQB80p	Sum score on depressive scale
MSQBQ1	Ever tried cigarettes
MSQBR34_SUM	Sum of PE class taking
	Random Effects (School Level)
MSMA3E	Percent white
MSMA4	Percent free/reduced lunch
MSMA5A	Percent passing state math test
MSMA5B	Percent passing state English/reading test
PDHA1	PE class size
PPIC1C2	Required weeks of PE per year
PPIC2	Percent students not meeting requirements
PPIC17	PA school events this year
PPIC19	Interscholastic and Intramural PA programs
PPIC21	School ground changes in past year
PPIC22	Policy changes that encourage PA
PPIC24	Budget change positive for PA
PPIC30	Percent bike/walk to school
PPIC34	Unstructured free play before school
PPIC35	Unstructured free play during school
PPIC36	Unstructured free play after school
PSB_Numprog	Number of programs in school
MVPA	MVPA at school

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory.*

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

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–23.

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

Council on Sports Medicine and Fitness and Council on School Health (2006), "Active healthy living: prevention of childhood obesity through increased physical activity," *Pediatrics*, 117, 1834–42.

Daniels, S. R., Arnett, D. K., Eckel, R. H., Gidding, S. S., Hayman, L. L., Kumanyika, S., Robinson, T. N., Scott, B. J., Jeor, S. S., and Williams, C. L. (2005), "Overweight in Children and Adolescents : Pathophysiology, Consequences, Prevention, and Treatment," *Circulation*, 111, 1999–2012.

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

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–60.

Fan, Y. and Li, R. (2012), "Variable Selecion in Linear Mixed Effects Models," *Annals of Statistics*, 40, 2043–2068.

Foster, S. D., Verbyla, A. P., and Pitchford, W. S. (2009), "Estimation, Prediction and Inference for the Lasso Random Effects Model," *Australian and New Zealand Journal of Statistics*, 51, 43–61.

Hall, D. B. and Praestgaard, J. T. (2001), "Order-restricted Score Tests for Homogeneity in Generalised Linear and Nonlinear Models," *Biometrika*, 88, 739–751.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer, 2nd ed.

Jiang, J. J., Rao, J. S., Gu, Z., and Nguyen, T. (2008), "Fence Models for Mixed Model Selection," *The Annals of Statistics*, 36, 1669–92.

Kimm, S. Y., Glynn, N. W., Kriska, A. M., Barton, B. A., Kronsberg, S. S., Daniels, S. R., Crawford, P. B., Sabry, Z. I., and Liu, K. (2002), "Decline in physical activity in black girls and white girls during adolescence," *New England Journal of Medicine*, 347, 709–15.

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

Lange, N. and Laird, N. M. (1998), "The Effects of Covariance Structure on Variance Estimation in Balanced Growth-Curve Models," *Journal of the American Statistical Association*, 84, 241–247.

Li, Y., Wang, S., Song, P. X.-K., Wang, N., and Zhu, J. (2012), "Doubly Regularized Estimation and Selection in Linear Mixed-Effects Models for High-Dimenstional Longitudinal Data," .

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

Lindstrom, M. J. and Bates, D. M. (1988), "Newton-Raphson and EM algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Ogden, C., Kuczmarski, R., Flegal, K., Mei, Z., Guo, S., Wei, R., Grummer-Strawn, L., Curtin, L., Roche, A., and Johnson, C. (2002), "Centers for Disease Control and Prevention 2000 growth charts for the United States: improvements to the 1977 National Center for Health Statistics Centers for Disease Control and Prevention 2000 growth charts for the United States: improvements to the 1977 National Center for Health Statistics Version," *Pediatrics*, 109.

Pietiläinen, K. H., Kaprio, J., Borg, P., Plasqui, G., Yki-Järvinen, H., Kujala, U. M., Rose, R. J., Westerterp, K. R., and Rissanen1, A. (2008), "Physical inactivity and obesity: A vicious circle," *Obesity*, 16, 409–14.

Raghunathan, T. E., Lepkowski, J., Solenberger, P. W., and Van Hoewyk, J. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, 27, 85–95.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002), "IVEware: Imputation and Variance Estimation Software," Computer Software.

Sallis, J. F., Zakarain, J. M., Howell, M. F., and Hofstetter, C. R. (1996), "Ethnic, Socioeconomic, and Sex Differences Ethnic, Socioeconomic, and Sex Differences in Physical Activity Among Adolescents," *Journal of Clinical Epidemiology*, 49, 125–34.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–4.

Stram, D. O. and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, 1171–7.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, 58, 267–88.

Webber, L. S., Catellier, D. J., Lytle, L. A., Murray, D. M., Pratt, C. A., Young, D. R., Elder, J. P., Lohman, T. G., Stevens, J., Jobe, J. B., Pate, R. R., and TAAG Collaborative Research Group (2008), "Promoting physical activity in middle school girls: Trial of Activity for Adolescent Girls," *Am J Prev Med*, 34, 173–84.

Young, D., Saksvig, B., Wu, T., Zook, K., Li, X., Champaloux, S., Grieser, M., Lee, S., and Treuth, M. (2013), "Multilevel Predictors of Physical Activity For Early, Mid, and Late Adolescent Girls," *Journal of Physical Activity and Health (In press)*.

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

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B (Methodological)*, 67, 301–320.