

ABSTRACT

KRISHNA, ARUN. Shrinkage-Based Variable Selection Methods for Linear Regression and Mixed-Effects Models. (Under the direction of Professors H. D. Bondell and S. K. Ghosh).

In this dissertation we propose two new shrinkage-based variable selection approaches. We first propose a Bayesian selection technique for linear regression models, which allows for highly correlated predictors to enter or exit the model, simultaneously. The second variable selection method proposed is for linear mixed-effects models, where we develop a new technique to jointly select the important fixed and random effects parameters. We briefly summarize each of these methods below.

The problem of selecting the correct subset of predictors within a linear model has received much attention in recent literature. Within the Bayesian framework, a popular choice of prior has been Zellner's g -prior which is based on the inverse of empirical covariance matrix of the predictors. We propose an extension of Zellner's g -prior which allow for a power parameter on the empirical covariance of the predictors. The power parameter helps control the degree to which correlated predictors are smoothed towards or away from one another. In addition, the empirical covariance of the predictors is used to obtain suitable priors over model space. In this manner, the power parameter also helps to determine whether models containing highly collinear predictors are preferred or avoided. The proposed power parameter can be chosen via an empirical Bayes method which leads to a data adaptive choice of prior. Simulation studies and a real data example are presented to show how the power parameter is well determined from the degree of cross-correlation within predictors. The proposed modification compares favorably to the standard use of Zellner's prior and an intrinsic prior in these examples.

We propose a new method of simultaneously identifying the important predictors that correspond to both the fixed and random effects components in a linear mixed-effects model. A reparameterized version of the linear mixed-effects model using a modified Cholesky decomposition is proposed to aid in the selection by dropping out the random effect terms whose corresponding variance is set to zero. We propose a

penalized joint log-likelihood procedure with an adaptive penalty for the selection and estimation of the fixed and random effects. A constrained EM algorithm is then used to obtain the final estimates. We further show that our penalized estimator enjoys the Oracle property, in that, asymptotically it performs as well as if the true model was known beforehand. We demonstrate the performance of our method based on a simulation study and a real data example.

Shrinkage-Based Variable Selection Methods for Linear Regression and
Mixed-Effects Models

by
Arun Krishna

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

Statistics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Hao H. Zhang

Dr. Lexin Li

Dr. Howard D. Bondell
Chair of Advisory Committee

Dr. Sujit K. Ghosh
Co-Chair of Advisory Committee

DEDICATION

To my mom and dad.

BIOGRAPHY

Arun Krishna was born on July 1st, 1980 in Bangalore, India. He graduated from high school in 1997 and joint Manipal Institute of Technology, Manipal, India, to pursue a degree in Electrical Engineering, he obtained his bachelors degree in 2001. He then decided to fulfill his dream and come to America.

In January of 2002, Arun was accepted into the Graduate program in the school of engineering at Texas A & M. During his stay there he met Dr Margaret Land a professor in the department of mathematics who introduced him to vast and challenging area of computational and mathematical Statistics. Needless to say he was instantly mesmerized and decided to pursue a degree in Statistics. He was accepted into the statistics department at North Carolina State University in the fall of 2003. Arun received his masters degree in statistics in May 2005. The craving of knowing more propelled him to continue his stay in same department to pursue a Ph.D degree.

During his stay he felt that it was important to simultaneously gain some work experience. He worked for 2 years as a graduate industrial trainee at SAS Institute Inc., Cary, NC, in the Econometrics and Time Series group, where he gained some valuable experience in the field of computational statistics. He also held an internship in the PK/PD group at Pfizer R & D, Groton, CT. Arun is currently working as a Senior Biostatistician at Novartis Oncology, New Jersey.

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisors Dr. Howard Bondell and Dr. Sujit Ghosh, without whose direction and guidance this dissertation would not exist. I would also like to thank my committee members Dr. Helen Zhang and Dr. Lexin Li whose helpful suggestions and comments helped to strengthen this dissertation.

I would like to thank the Department of Statistics, North Carolina State University for their support. My advisors and I would like to thank the executive editor and the three anonymous referees for their useful comments to our paper, “Bayesian Variable Selection using Adaptive Powered Correlation Prior”. We would also like to thank Steven Howard at United States Environmental Protection Agency (EPA) for processing and formatting the CASTNet data for our application.

I would like to thank my friends Dr. Brian Reich in the department of statistics at North Carolina State University and Dr. Curtis McKay for their helpful comments and stimulating discussions. I would also like to thank Dr. Susan Willavize at Pfizer for giving me the opportunity to gain some valuable experience in pharmacokinetics and systems biology during my stay at Pfizer. I would like to thank my girlfriend Cristina who has seen me through thick and thin, and without whose support, motivation and understanding this dissertation would not have been written. Lastly, I would like to thank my parents who have endured a lot of hardship to get me where I am today, and I owe it all to them.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	ix
1 Introduction	1
1.1 Variable Selection in Linear Regression Models	2
1.1.1 Subset Selection	3
1.1.2 Penalized Least Squares	4
1.1.3 Bayesian Variable Selection	7
1.2 Introduction to Linear Mixed-Effects Models	11
1.2.1 The Maximum Likelihood and Restricted Maximum Likelihood Estimation	13
1.2.2 Numerical Algorithms	14
1.3 Variable Selection in LME models	17
1.3.1 Subset Selection	18
1.3.2 Bayesian Variable Selection for LME Models	20
1.3.3 The Likelihood Ratio Test	21
1.3.4 Selection with Shrinkage Penalty	22
1.4 Plan of Dissertation	23
2 Bayesian Variable Selection Using Adaptive Powered Correlation Prior	25
2.1 Introduction	25
2.2 The Adaptive Powered Correlated Prior	27
2.3 Model Specification	30
2.3.1 Choice of g	32
2.4 Model Selection using Posterior Probabilities	33
2.5 Simulations and Real Data	35
2.5.1 Simulation Study	35
2.5.2 Real Data Example	37
2.6 Discussion and Future Work	38
3 Joint Variable Selection of Fixed and Random Effects in Linear Mixed-Effects Model and its Oracle Properties	47
3.1 Introduction	47
3.2 The Reparameterized Linear Mixed Effects Model	51

3.2.1	The Likelihood	53
3.3	Penalized Selection and Estimation for the Reparameterized LME model	54
3.3.1	The Shrinkage Penalty	54
3.3.2	Computation and Tuning	55
3.4	Asymptotic Properties	60
3.5	Simulation Study and Real Data Analysis	62
3.5.1	Simulation Study	63
3.6	Real Data Analysis	65
3.7	Discussion and Future Work	68
Bibliography		77
APPENDICES		83
Appendix A. Description of NCAA Data		84
Appendix B. Description of CASTnet Data		85
Appendix C. Asymptotic Properties: Regularity conditions and proofs		86
C.1	Regularity Condition	86
C.2	Proof of Theorem 1	87
C.3	Proof of Theorem 2	89
C.4	Proof of Theorem 3	93

LIST OF TABLES

Table 1.1 Penalized Regression methods and their corresponding penalty terms .	24
Table 2.1 For Case 1, Comparing Average Posterior Probabilities, corresponding to the case 1: $p = 4$, averaged across 1,000 simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (2.19). PoCor represents our proposed modification as in (2.14). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)	43
Table 2.2 Case 2, Comparing Average Posterior Probabilities, corresponding to the case 2: $p = 12$, averaged over 1,000 Simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (2.19). PoCor represents our proposed modification as in (2.14). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)	44
Table 2.3 Comparing Posterior Probabilities and average prediction errors for the models of the NCAA Data. The entries in parenthesis are the standard errors obtained by 1000 replications.	45
Table 2.4 Posterior Probabilities and average prediction errors for the models of the NCAA Data using the Intrinsic Priors. The entries in parenthesis are the standard errors obtained by 1000 replications.	46
Table 3.1 Comparing the median KullbackLeibler discrepancy (KLD) from the true model, along with the percentage of the times the true model was selected (%Correct) for each method, across 200 datasets. R.E represents the relative efficiency compared to the oracle model. %CF, %RF corresponds to the percentage of times the correct fixed and random effects were selected , respectively.	73
Table 3.2 Variables selected for the fixed and the random components for the CASTnet data.	74

Table 3.3 Penalized Likelihood estimates for regression coefficients and the random effects variances for the model selected using our proposed method.	75
Table 3.4 Spearman Rank Correlation Coefficients between Observed Values	76

LIST OF FIGURES

Figure 2.1 Plot of λ vs. $\text{Log}[m(\mathbf{y} \mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 1: $p = 4$. Averaged over 1,000 simulations. The vertical line represents the location of the global maximum.....	40
Figure 2.2 Plot of λ vs. $\text{Log}[m(\mathbf{y} \mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 2: $p = 12$, averaged over 1,000 simulations. The vertical line represents the location of the global maximum.....	41
Figure 2.3 Plot of λ vs. $\text{Log}[m(\mathbf{y} \mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to the NCAA Dataset. The vertical line represents the location of the global maximum.	42
Figure 3.1 The location of the 15 Sites that were used for our analysis. The ∇ represents the 4 sites used for the overlay plots in Figure 3.4.....	69
Figure 3.2 Site (individual) profile plot to assess the seasonal trend in Nitrate concentration over each 12 month period, for the CASTnet dataset.....	70
Figure 3.3 Box Plot to assess heterogeneity among the 15 sites in the measured Nitrate concentration corresponding to the CASTnet dataset.....	71
Figure 3.4 Plot of the observed $\text{LOG}(\text{TNO})_3$ concentration represented by the solid line, overlaid with the fitted values for the fixed effects model selected using our proposed method for 4 centers, for the CASTnet dataset.	72

Chapter 1

Introduction

In the statistical literature, the problem of selecting variables/predictors has received considerable attention over the years. Variable selection has two main goals, easy model interpretation (sparse representation) and prediction accuracy. In practice, statistical data often consist of a large number of potential predictors or explanatory variables, denoted by $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$. Usually not all these variables contribute to the explanation of a particular quantity of interest, the response variable \mathbf{y} . The variable selection problem arises when it is of particular interest to identify the subset $p' \leq p$ of relevant predictor variables. This then results in removing the noninformative variables in order to improve the predictability of the models and parsimoniously describe the relationship between the outcome and the predictive variables.

The remainder of the introduction is as follows. In Section 1, we describe in detail the different selection methods proposed in the literature for linear regression models. In Section 2, we introduce the theory behind mixed models and the different techniques used to estimate the fixed effects and the variance components. The selection methods proposed in the literature for linear mixed-effects models are discussed in Section 3. Finally, in Section 4 we end the chapter with an overview of what follows in the rest of the dissertation.

1.1 Variable Selection in Linear Regression Models

Most statistical problems involve determining the relationship between the response variable and the set of predictors. This can often be represented in a linear regression framework as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1.1)$$

where \mathbf{y} is an $n \times 1$ vector of responses and $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_p)'$ is an $n \times p$ matrix of known explanatory variables, with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ a $p \times 1$ vector of unknown regression parameters, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$, where $N(0, \sigma^2 \mathbf{I})$ denotes a multivariate normal distribution with mean vector zero and variance-covariance matrix $\sigma^2 \mathbf{I}$. We assume throughout that the \mathbf{X} 's and \mathbf{y} 's have been centered so that the intercept may be omitted from the model. Under the above regression model, it is assumed that only an unknown subset of the coefficients is nonzero, so that the variable selection problem is to identify this unknown subset.

In the context of variable selection we begin by indexing each candidate model with one binary vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$, where each element δ_j takes the value 1 or 0 depending on whether a predictor is included or excluded from the model. Given $\boldsymbol{\delta} \in \{0, 1\}^p$, the linear regression model assumes

$$\mathbf{y} | \boldsymbol{\delta}, \boldsymbol{\beta}_{\boldsymbol{\delta}}, \sigma^2 \sim N(\mathbf{X}_{\boldsymbol{\delta}} \boldsymbol{\beta}_{\boldsymbol{\delta}}, \sigma^2 \mathbf{I}), \quad (1.2)$$

where $\mathbf{X}_{\boldsymbol{\delta}}$ and $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ are the design matrix and the regression parameters corresponding to the non-zero elements of $\boldsymbol{\delta}$, respectively, and σ^2 is the unknown error variance for model $\boldsymbol{\delta}$. Notice that $\mathbf{X}_{\boldsymbol{\delta}}$ is a matrix of dimension $n \times p_{\boldsymbol{\delta}}$ where $p_{\boldsymbol{\delta}} = \sum_{j=1}^p \delta_j$ and $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ is of size $p_{\boldsymbol{\delta}} \times 1$. Numerous selection methods have been proposed in the literature for linear regression models. In this Section, we review a few of these methods.

1.1.1 Subset Selection

Subset selection methods (Miller, 2002) such as all subsets, forward selection, backward elimination and STEPWISE have been widely used by statisticians to select significant variables. The all subsets selection method performs an exhaustive search over all possible subsets in model space. For instance, given the linear model in (1.1) the total number of possible sub-models is 2^p . After enumerating all possible models in consideration, the all subset method then selects the best model using a given criterion. The main drawback of this method is that it is computationally intensive and could be time consuming, especially if the number of predictors (p) is large.

To obtain the ‘best’ model using subset selection, Mallows (1973) proposed a C_p statistic (we denote it by $C_{\boldsymbol{\delta}}$) which involves estimating the out-of-sample prediction error for each model indexed by $\boldsymbol{\delta}$. We use this statistic as a criterion to compare different subsets of regression models. The criterion is computed as

$$C_{\boldsymbol{\delta}} = \frac{\|\mathbf{y} - \mathbf{X}_{\boldsymbol{\delta}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\delta}}\|^2}{\hat{\sigma}^2} - n + 2p_{\boldsymbol{\delta}}, \quad (1.3)$$

where $\hat{\sigma}^2$ is the unbiased estimate of the error variance based on the full model. The model with the minimum value of $C_{\boldsymbol{\delta}}$ is then termed the ‘best’ model.

Information-based criterion approaches attempt to find the model with the smallest Kullback-Leibler (Kullback and Leibler, 1951) divergence from the true but unknown data generating process. Akaike Information Criterion (Akaike, 1973) is one such estimate of this distance. Ignoring the constant terms the AIC for model $\boldsymbol{\delta}$ in a linear regression framework can be computed as

$$AIC_{\boldsymbol{\delta}} = n \log \left(\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \right) + 2p_{\boldsymbol{\delta}} \quad (1.4)$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\delta}}$ is the vector of estimated regression coefficient for the predictors $\mathbf{X}_{\boldsymbol{\delta}}$. The model with the smallest AIC value is selected as the ‘best’ model. The drawback of this approach is that it tends to overfit. Schwartz (1978) proposed a consistent selection method called the Bayesian Information Criterion (BIC) which puts a heavier

penalty on the degrees of freedom. The BIC criterion is given as

$$BIC_{\boldsymbol{\delta}} = n \log \left(\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \right) + \log(n) \times p_{\boldsymbol{\delta}}, \quad (1.5)$$

where n is the total number of observations. Again, the model which minimizes (1.5) is selected. The BIC is consistent for model selection. I.e., if the true model p_0 is in the class of all models considered, the value of $p_{\boldsymbol{\delta}}$ that minimizes (1.5) converges to the true model (p_0) as $n \rightarrow \infty$.

To avoid the burden of enumerating all possible (2^p) models, methods such as forward selection or backward elimination are used to find the best subset, either by adding or deleting one variable at a time or by a combination of the two, such as in STEPWISE. In each step the best predictor is included or excluded according to some criterion, e.g., the F-statistic, AIC or BIC. However, these methods are extremely unstable and are very sensitive to small changes due to their inherent discreteness (Breiman, 1996).

1.1.2 Penalized Least Squares

Though subset selection is practical and widely used as a method to find the best model, it has many drawbacks. Breiman (1996) discusses the lack of stability of the subset selection procedures, where a small change in the data could result in a large change in their predictive error and can result in very different models being selected. In practice, these methods could be very time consuming due to being computationally intensive, especially when the number of predictors is large.

To overcome these obstacles, the use of penalized regression, or regression shrinkage, has emerged as a highly successful method to tackle this problem. The idea behind penalized least squares is to obtain the estimates for the regression coefficients by minimizing the penalized sum of squared residual

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_{\lambda}(|\beta_j|), \quad (1.6)$$

where $\sum_{j=1}^p p_\lambda(|\beta_j|)$ is the penalty term corresponding to the regression coefficients, and λ is a non-negative tuning parameter. Several forms of the penalty function have been proposed. These are summarized in Table 1.1.

The penalized least square technique first proposed by Hoerl and Kennard (1970), called Ridge regression, minimizes the residual sum of squares by imposing a bound on the l_2 norm of the regression coefficients. By introducing this penalty they succeeded in shrinking the least square estimates continuously toward zero and achieving a better predictive performance than the Ordinary Least Squares (OLS). Due to the continuous shrinking process it results in being more stable and not sensitive to small changes in the data as in subset selection. Though this method does not perform variable selection as it does not set the coefficients to zero, it is sometimes used in combination with other penalty terms which do perform variable selection.

Adopting the good quality (continuous shrinkage) of the ridge regression (Hoerl and Kennard, 1970) penalty and combining it with the good features of subset selection, Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO). This method introduces an l_1 penalty ($\sum_{j=1}^p |\beta_j|$) on the regression coefficients. Due to the geometric nature of the l_1 penalty the LASSO does both continuous shrinkage as well as variable selection by having the ability the regression coefficients to zero. The LASSO estimates for the regression parameters are defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j|, \quad (1.7)$$

where λ is a non-negative tuning/regularization parameter. As λ increases, the coefficients are continuously shrunk towards zero, and some coefficients are exactly shrunk to zero for a sufficiently large λ . A plethora of penalized regression techniques have been proposed by modifying the LASSO penalty to accommodate other desirable properties. We shall discuss a few.

Recently, Efron, Hastie, Johnstone and Tibshirani (2004) proposed a model selection algorithm called Least Angle Regression (LARS). They showed that the algorithm can be used to obtain the entire solution path for LASSO estimates while being

computationally efficient.

In the presence of highly correlated predictors it has been shown empirically that the performance of the LASSO is sub-par compared to the ridge regression in predictive performance (Tibshirani, 1996). With this in mind, Zou and Hastie (2005) proposed the ‘Elastic-Net’ procedure, whose penalty term is a convex combination of the LASSO penalty (Tibshirani, 1996) and the ridge regression (Hoerl and Kennard, 1970) penalty, as shown in Table (1.1). The ‘Elastic-Net’ simultaneously performs automatic variable selection with the help of the l_1 penalty while encouraging a grouping effect by selecting groups of highly correlated predictors, rather than simply eliminating some of them from the model arbitrarily.

Along the same lines as the ‘Elastic-net’, Bondell and Reich (2008) proposed a clustering algorithm called the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) which uses a convex combination of the LASSO penalty and a pairwise L_∞ penalty (Table 1.1). This method not only shrinks the unimportant coefficients to zero with the help of the LASSO penalty; it also encourages highly correlated predictors that have a similar effect on the response to form predictive clusters by setting them equal to one another. Hence, the OSCAR eliminates the unimportant predictors while performing supervised clustering on the important ones.

To overcome the drawback that the LASSO shrinkage method produces biased estimates for the large coefficients, Fan and Li (2001) proposed a variable selection approach via non-concave penalized likelihood where a Smoothly Clipped Absolute Deviation (SCAD) penalty is used (see Table 1.1). They proposed a general theoretical setting to understand the asymptotic behavior of these penalized methods called the “Oracle property”. The formal definition in the context of a linear regression framework is as follows.

Definition 1. *Let the true value of β be given by*

$$\beta_0 = (\beta_{10}, \beta_{20}, \dots, \beta_{p_0})' = (\beta'_{10}, \beta'_{20})', \quad (1.8)$$

where β_{10} are the true non-zero components and $\beta_{20} = \mathbf{0}$. Let $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)'$ de-

note the penalized regression estimates obtained. The penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ enjoys the ‘Oracle’ property if the following hold true.

- (i) The true zero coefficients are estimated as zero with probability tending to 1, that is, $\lim_{n \rightarrow \infty} P(\hat{\boldsymbol{\beta}}_2 = 0) = 1$.
- (ii) The penalized regression estimates of the true non-zero coefficients are \sqrt{n} consistent and asymptotically normal; that is, $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \rightarrow_d N(0, I^{-1}(\boldsymbol{\beta}_{10}))$, where $I(\boldsymbol{\beta}_{10})$ is the Fisher information matrix, knowing that $\boldsymbol{\beta}_2 = \mathbf{0}$.

Hence an estimator that satisfies these conditions performs as well as if the true model were known beforehand. Zou (2006) proposed another penalized method called the adaptive LASSO in which adaptive weights are used to penalize the different regression coefficients in the l_1 penalty. That is, large amount of shrinkage are applied to the zero-coefficients while smaller amounts are used for the non-zero coefficients. This then results in an estimator with improved efficiency, as opposed to the LASSO which gives the same amount of shrinkage to all the coefficients. The adaptive LASSO estimates are defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p \bar{w}_j |\beta_j|, \quad (1.9)$$

where \bar{w}_j are the adaptive weights, typically $\bar{w}_j = 1/|\bar{\beta}_j|$, with $\bar{\beta}_j$ the ordinary least squares estimate. The adaptive LASSO can now be thought of as a convex optimization problem with an l_1 constraint. Zou demonstrated that the adaptive LASSO estimates can be obtained by using the highly efficient LARS algorithm (Efron, Hastie, Johnstone and Tibshirani, 2004) by making a small change to the design matrix. Zou further showed that the adaptive LASSO estimates enjoys the ‘Oracle’ property (Definition 1).

1.1.3 Bayesian Variable Selection

One approach to Bayesian variable selection in a linear regression framework has been to pick a model with the highest posterior probability among all candidate

models. Given the conditional distribution of \mathbf{y} (1.2), the approach then proceeds by assigning a prior probability to all possible models indexed by $\boldsymbol{\delta}$ along with a prior probability distribution to the model parameters $(\boldsymbol{\beta}_{\boldsymbol{\delta}}, \sigma^2)$ in a hierarchical fashion

$$P(\boldsymbol{\beta}_{\boldsymbol{\delta}}, \sigma^2, \boldsymbol{\delta}) = P(\boldsymbol{\beta}_{\boldsymbol{\delta}} | \sigma^2, \boldsymbol{\delta}) P(\sigma^2 | \boldsymbol{\delta}) P(\boldsymbol{\delta}). \quad (1.10)$$

Posterior model probabilities are then used to select the ‘best’ model. With this set-up in mind, many variable selection techniques have been proposed by using different and innovative priors for the model parameters and the prior over model space. We shall discuss a few of these methods.

Mitchell and Beauchamp (1988) introduced the ‘spike and slab’ prior for $\boldsymbol{\beta}$ ’s. They assume that the $\boldsymbol{\beta}$ ’s are independent of σ and that the β_j are also mutually independent. More explicitly, the prior for each β_j is

$$P(\beta_j | \delta_j) = (1 - \delta_j) \mathbf{1}_{\{0\}}(\beta_j) + \delta_j \frac{1}{2a} \mathbf{1}_{[-a, a]}(\beta_j).$$

The prior for the inclusion indicators $\boldsymbol{\delta}$ is given an independent Bernoulli distribution with parameter w_j . The hyperparameters are obtained using a Bayesian cross-validation and ‘best’ model is obtained using posterior probabilities.

George and McCulloch (1993, 1997) propose a mixture normal prior for the regression coefficients and an inverse-gamma conjugate prior on σ^2 . The difference between this method and the ‘spike and slab’ prior is that they do not implicitly put a probability mass for $\beta_j = 0$. Instead, for $\delta_j = 0$, the corresponding prior on β_j has a variance close to its mean (zero). As for the prior over model space, George and McCulloch suggests

$$p(\boldsymbol{\delta}) \propto \pi^{p_{\boldsymbol{\delta}}} (1 - \pi)^{p - p_{\boldsymbol{\delta}}}, \quad (1.11)$$

where $p_{\boldsymbol{\delta}} = \sum_{j=1}^p \delta_j$ is the number of predictors in the model defined by $\boldsymbol{\delta}$, and π is the prior inclusion probability for each covariate. We can see this being equivalent to placing Bernoulli (π) priors on δ_j and thereby giving equal weight to any pair of equally-sized models. Setting $\pi = 0.5$ results in the popular uniform prior over model

space

$$P(\boldsymbol{\delta}) = \left(\frac{1}{2}\right)^p. \quad (1.12)$$

However, the drawback of using this prior is that it puts most of its mass on models of size $p/2$. For large p it is nearly impossible to enumerate all 2^p possible models. With this in mind George and McCulloch (1993,1997) introduce the SVSS (Stochastic Search Variable Selection) algorithm to search through model space by utilizing Gibbs sampling to indirectly sample from the posterior distribution on a set of possible subset choices. Models with high posterior probabilities can then be identified as the ones which most frequently appeared in the Gibbs sample.

More recently, Yuan and Lin (2005) showed that under certain conditions that the model with the highest posterior probability selected using their method will also be the model selected by LASSO. To show this, they proposed a double exponential prior distribution on the regression coefficients. As for the prior over model space they proposed

$$P(\boldsymbol{\delta}) \propto \pi^{p\delta} (1 - \pi)^{p-p\delta} |\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}}|^{1/2}, \quad (1.13)$$

where $|\cdot|$ denotes the determinant, and $|\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}}| = 1$ if $p_{\boldsymbol{\delta}} = 0$. Since $|\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}}|$ is small for models with highly collinear predictors, this prior discourages these models and arbitrary select one variable from the group.

In a linear regression framework, Zellner (1986) suggested a particular form of a conjugate Normal-Gamma family called the g -prior, which can be expressed as

$$\begin{aligned} \boldsymbol{\beta} | \sigma^2, g &\sim N\left(0, \frac{\sigma^2}{g} (\mathbf{X}' \mathbf{X})^{-1}\right) \\ \sigma^2 &\sim IG(a_0, b_0), \end{aligned} \quad (1.14)$$

where $g > 0$ is a known scaling factor and $a_0 > 0$, $b_0 > 0$ are known parameters of the Inverse Gamma distribution with mean $\frac{a_0}{b_0 - 1}$. The prior covariance matrix of $\boldsymbol{\beta}$ is the scalar multiple σ^2/g of the inverse Fisher information matrix, which concurrently depends on the observed data through the design matrix \mathbf{X} . In the context of variable

selection the ‘g-priors’ are conditioned on $\boldsymbol{\delta}$ to give

$$\boldsymbol{\beta}_{\boldsymbol{\delta}} | \boldsymbol{\delta}, g, \sigma^2 \sim N\left(0, \frac{\sigma^2}{g} (\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{-1}\right).$$

This particular prior has been widely adopted in the context of Bayesian variable selection due to its closed form calculations of all marginal likelihoods which is suitable for rapid computations over a large number of submodels. Its simple interpretation can be derived from the idea of a likelihood for a pseudo-data set with the same design matrix \mathbf{X} as the observed sample.

George and Foster (2003) used the above ‘g-prior’ to establish a connection between models selected using posterior probabilities and model selection criteria such as the AIC (Akaike, 1973), BIC (Schwartz, 1978) and the RIC (Foster and George, 1994). They showed that the ordering induced using the posterior probabilities is the same as the ordering obtained using any of the model selection criteria. They proposed an empirical Bayes method by maximizing the marginal likelihood of the data to obtain the estimates for the hyperparameters g and π . Similar results were also established by Fernandez et. al (2001) by using $P(\sigma^2) \propto 1/\sigma^2$. This representation avoids the need for choosing a_0, b_0 . To avoid the need to specify hyperparameters, Casella and Moreno (2006) proposed a fully automatic Bayesian variable selection procedure where posterior probabilities are computed using intrinsic priors (Berger and Pericchi, 1996). Final model comparisons are made based on Bayes factors.

But the main drawback of all these methods is that they are not designed to handle correlated predictors. As including groups of highly correlated predictors together in the model can improve predictor accuracy as well as model interpretation (Zou and Hastie, 2005; Bondell and Reich, 2008). In Chapter 2 we propose a modification of Zellner’s g-prior (1.14) by replacing $(\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{-1}$ with $(\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{\lambda}$ where the power $\lambda \in \mathbb{R}$, controls the amount of smoothing of the regression coefficients of collinear predictors towards or away from each other accordingly as $\lambda > 0$ or $\lambda < 0$, respectively. In addition, the empirical covariance of the predictors, along with the hyperparameter λ , is used to obtain suitable priors over model space. Hence coupled together, our

proposed method encourages groups of correlated predictors to enter or exit the model simultaneously.

1.2 Introduction to Linear Mixed-Effects Models

Linear mixed-effects (LME) models are a class of statistical models which are used directly in many fields of application. LME models are used to describe the relationship between the response and a set of covariates that are grouped according to one or more classification factors. Mixed models have been commonly used to analyze clustered or repeated measures data. In this section we shall provide the reader with a general overview on the theory behind mixed models. We will also briefly describe the techniques commonly used to obtain the estimates for the fixed effects parameters and the variance components.

Consider a study consisting of m subjects, with response from each subject $i = 1, 2, \dots, m$ measured n_i times, and let $N = \sum_{i=1}^m n_i$ be the total number of observations. Let \mathbf{y}_i be a $n_i \times 1$ vector of the response variable for subject i . Let \mathbf{X}_i be the $n_i \times p$ design matrix of explanatory variables, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ be the vector of unknown regression parameters which are assumed to be fixed. Let $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})'$ be a $q \times 1$ vector of unknown subject-specific random effects, and \mathbf{Z}_i is the $n_i \times q$ known design matrix corresponding to the random effects. A general class of LME models can be written as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i. \quad (1.15)$$

The random effects \mathbf{b}_i and $\boldsymbol{\epsilon}_i$ are independent of each other and generally assumed to come from a normal distribution as shown

$$\begin{bmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_i} \end{bmatrix} \right). \quad (1.16)$$

Treating $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_m)$ as an $N \times 1$ vector, the LME model given in (1.15) can

also be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim N(\mathbf{0}, \sigma^2 \tilde{\boldsymbol{\Psi}}), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1.17)$$

where $\mathbf{X} = [\mathbf{X}'_1, \dots, \mathbf{X}'_m]'$ denotes a $N \times p$ stacked matrix of \mathbf{X}_i , and \mathbf{Z} and $\tilde{\boldsymbol{\Psi}}$ denote block diagonal matrices which are given by

$$\tilde{\boldsymbol{\Psi}} = \begin{bmatrix} \boldsymbol{\Psi} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Psi} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\Psi} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_m \end{bmatrix}. \quad (1.18)$$

Given (1.17) and $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_m)'$, the hierarchical formulation is written as

$$\begin{aligned} \mathbf{y}|\mathbf{b} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2 \mathbf{I}), \\ \mathbf{b} &\sim N(\mathbf{0}, \sigma^2 \tilde{\boldsymbol{\Psi}}). \end{aligned} \quad (1.19)$$

Integrating out the random effects, it can be shown that the marginal distribution of \mathbf{y} follows

$$\begin{aligned} \mathbf{y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \tilde{\mathbf{V}}) \\ \text{where, } \tilde{\mathbf{V}} &= \sigma^2(\mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}' + \mathbf{I}), \end{aligned} \quad (1.20)$$

where $\tilde{\mathbf{V}} = \text{Diag}(\mathbf{V}_1, \dots, \mathbf{V}_m)$ is an $N \times N$ block diagonal matrix of \mathbf{V}_i 's. The fixed effects and the variance components in the LME model are typically estimated using maximum likelihood (ML) or restricted maximum likelihood (REML) proposed by Patterson and Thompson (1971). We shall briefly describe each of these methods.

1.2.1 The Maximum Likelihood and Restricted Maximum Likelihood Estimation

Let us denote $\boldsymbol{\theta} = (\boldsymbol{\beta}', \text{Vech}'(\boldsymbol{\Psi}), \sigma^2)'$ to be a $s \times 1$ combined vector of unknown parameters, where $\text{Vech}(\boldsymbol{\Psi})$ represents the $q(q+1)/2$ free parameters of $\boldsymbol{\Psi}$. Hence the total dimension of $\boldsymbol{\theta}$ is $1 + p + q(q+1)/2$. Let Θ denote the parameter space for $\boldsymbol{\theta}$ such that

$$\Theta = \{\boldsymbol{\theta} : \boldsymbol{\beta} \in \mathbb{R}^p, \sigma^2 > 0, \boldsymbol{\Psi} \text{ is non-negative definite}\}. \quad (1.21)$$

Dropping out the constant terms, the log-likelihood function based on the LME model (1.17) is given by

$$L_{ML}(\boldsymbol{\theta}) = -\frac{1}{2} \left\{ N \times \log(\sigma^2) + \log|\mathbf{I} + \mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}'| + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} + \mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}')^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}, \quad (1.22)$$

where $\tilde{\boldsymbol{\Psi}} = \text{Diag}(\boldsymbol{\Psi}, \dots, \boldsymbol{\Psi})$ is a block diagonal matrix of $\boldsymbol{\Psi}$'s. The maximum likelihood estimate (MLE) maximizes (1.22) over the parameter space Θ given in (1.21). Note that in certain situations the ML estimates may fall outside the boundary of the parameter space Θ , one could avoid this one would need to deal with constrained maximization, we shall discuss in detail at the end of section 1.2.2. For a known $\boldsymbol{\Psi}$, the estimate for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'(\mathbf{I} + \mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}')^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} + \mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}')^{-1}\mathbf{y}, \quad (1.23)$$

where the variance of $\hat{\boldsymbol{\beta}}$ is given by $\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X}$. For $\boldsymbol{\Psi} = \mathbf{0}$, notice that the estimate for $\boldsymbol{\beta}$ reduces to the OLS (ordinary least square) estimate. Again, for a known $\boldsymbol{\Psi}$ and replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ given in (1.23), and taking the derivative of (1.22) with respect to σ^2 we see that the log-likelihood function is maximized at

$$\hat{\sigma}^2 = \frac{1}{N} \left\{ (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{I} + \mathbf{Z}\tilde{\boldsymbol{\Psi}}\mathbf{Z}')^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \right\}. \quad (1.24)$$

When Ψ is unknown we obtain an estimate of it by maximizing (1.22) after replacing β and σ^2 by (1.23) and (1.24), respectively.

Maximum Likelihood estimators are functions of sufficient statistics and are \sqrt{n} consistent and asymptotically normal (see Searle, Casella and McCulloch, 1992). But the estimates of the variance components are biased downwards. The bias arises because the method does not take into consideration the degrees of freedom lost due to the estimation of the fixed effects. To overcome this, Patterson and Thompson (1971) proposed the restricted maximum likelihood (REML) method to obtain unbiased estimates of the variance components in an LME framework. The REML log-likelihood for the LME model in (1.15) is

$$L_R(\theta) = -\frac{1}{2} \left\{ (N-p)\log(\sigma^2) + \log|\mathbf{I} + \mathbf{Z}\tilde{\Psi}\mathbf{Z}'| + \log|\mathbf{X}'(\mathbf{I} + \mathbf{Z}\tilde{\Psi}\mathbf{Z}')^{-1}\mathbf{X}| \right. \\ \left. + \frac{1}{\sigma^2} \left\{ (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{I} + \mathbf{Z}\tilde{\Psi}\mathbf{Z}')^{-1}(\mathbf{y} - \mathbf{X}\beta) \right\} \right\}. \quad (1.25)$$

For a known Ψ , taking the derivative with respect to σ^2 results in maximizing the function (1.25) at

$$\hat{\sigma}_R^2 = \frac{1}{N-p} \left\{ (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{I} + \mathbf{Z}\tilde{\Psi}\mathbf{Z}')^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \right\}, \quad (1.26)$$

where $\hat{\beta}$ is as given in (1.23), and p is the number of fixed effects. We see that the degrees of freedom for REML estimates for σ^2 have been adjusted by the number of fixed effects.

1.2.2 Numerical Algorithms

For an unknown Ψ , the estimate of $\theta = (\beta', \text{vech}'(\Psi), \sigma^2)'$ is obtained by joint maximization of (1.22) or (1.25) with respect to all the parameters simultaneously. In general there is no closed form solution for these estimates and must be determined by numerical algorithms such as, Expectation-maximization (Laird, Lange and Stram

(1987), Newton-Raphson (Lindstrom and Bates, 1988) and Fishers scoring algorithm (see McLachlan and Krishnan (1994) and Demidenko (2004) for a comprehensive review of each of these methods). In this section we shall discuss the EM algorithm in detail.

The EM Algorithm

Dempster, Laird and Rubin (1977) proposed a general class of algorithm to compute the maximum likelihood estimates of parameters in the presence of missing data. Laird and Ware (1982) and Laird, Lange and Stram (1987) adopted this formulation to LME model by treating the random effects as unobserved (missing) data. Hence, in the general EM setting we shall call the observed data \mathbf{y} as incomplete data, and (\mathbf{y}, \mathbf{b}) as the complete data.

The EM algorithm has two main steps: the E-step where we compute the conditional expectation of the complete data log-likelihood assuming that the latent variables (i.e. random effects) are unobserved; and the M-step, where we maximize the expected complete-data log-likelihood with respect to the parameters in the model, using numerical optimization techniques. This process is then repeated iteratively to obtain the final converged estimates.

Given the hierarchical setup in (1.19) and dropping out the constant terms, the complete-data log-likelihood is given as

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{b}) = -\frac{1}{2} \left\{ (N + mq)\log(\sigma^2) + \log|\mathbf{I} + \mathbf{Z}\tilde{\Psi}\mathbf{Z}'| + \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2 + \mathbf{b}'(\tilde{\Psi})^{-1}\mathbf{b}}{\sigma^2} \right\}. \quad (1.27)$$

Given (1.27), the conditional distribution of \mathbf{b} given $\boldsymbol{\theta}$ and \mathbf{y} is $\mathbf{b}|\mathbf{y}, \boldsymbol{\theta} \sim N(\hat{\mathbf{b}}, \mathbf{G})$, where the mean and the conditional variance are given by

$$\begin{aligned} \hat{\mathbf{b}} &= ((\Psi)^{-1} + \mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{G} &= \sigma^2(\mathbf{Z}'\mathbf{Z} + (\Psi)^{-1})^{-1}. \end{aligned} \quad (1.28)$$

The EM algorithm starts with an initial value $\boldsymbol{\theta}^{(0)}$. Let $\boldsymbol{\theta}^{(\omega)}$ denote the estimate of

$\boldsymbol{\theta}$ at the ω^{th} iteration. Given $\boldsymbol{\theta}^{(\omega)}$, we compute the conditional expectation $\left(\mathbb{E}_{\mathbf{b}|\mathbf{y},\boldsymbol{\theta}^{(\omega)}}\right)$ of the complete-data log-likelihood as

$$\begin{aligned} g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\omega)}) &= \int L(\boldsymbol{\theta}|\mathbf{y},\mathbf{b})f(\mathbf{b}|\mathbf{y},\boldsymbol{\theta}^{(\omega)})d\mathbf{b}, \quad \text{for all } \boldsymbol{\theta} \\ &= \mathbb{E} \left\{ L(\boldsymbol{\theta}|\mathbf{y},\mathbf{b})|\mathbf{y},\boldsymbol{\theta}^{(\omega)} \right\}, \end{aligned} \quad (1.29)$$

where $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{b})$ is as given in (1.27). This defines the E-step. For the M-step, the updates estimate for $\boldsymbol{\theta}$ is obtained by maximizing the objective function $\left(g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\omega)})\right)$ to obtain $\boldsymbol{\theta}^{(\omega+1)}$ such that

$$g(\boldsymbol{\theta}^{(\omega+1)}|\boldsymbol{\theta}^{(\omega)}) \geq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\omega)}). \quad (1.30)$$

The process is now repeated iteratively until convergence to obtain the final estimates. The main drawback of the EM algorithm is its slow convergence, when the variance-covariance matrix $\boldsymbol{\Psi}$ is in the neighborhood of zero or in certain cases if ML/REML estimates of $\boldsymbol{\Psi}$ are on the boundary of the parameter space. Many different extensions and modifications have been proposed to overcome the slow rate of convergence. For examples see Meng and Rubin (1993); Baker (1992); Liu and Rubin (1994); McLachlan and Krishnan (1997); Meng (1997); Lu, Rubin and Wu (1998); Meng and Van Dyk (1998).

The advantage of the EM algorithm is that it will produce positive/semi-positive definite matrices if the starting values $(\boldsymbol{\Psi}^{(0)})$ are positive/semi-positive definite. In contrast the Newton-Raphson (Lindstrom and Bates, 1988) and Fishers scoring algorithms which allow the estimates of $\boldsymbol{\Psi}$ to converge outside the boundary of the parameter space, that is, $\boldsymbol{\Psi}$ could be negative definite. In such situations one would need to deal with constrained maximization, i.e., force the matrix $\boldsymbol{\Psi}$ to be non-negative. Another method is to replace $\boldsymbol{\Psi}$ with a non-negative definite matrix. Pinhero and Bates (1996) suggested using the Cholesky decomposition of $\boldsymbol{\Psi} = \mathbf{L}\mathbf{L}'$, where \mathbf{L} denotes a lower triangular matrix. Replacing $\boldsymbol{\Psi}$ by $\mathbf{L}\mathbf{L}'$ in the log-likelihood function guarantees $\boldsymbol{\Psi}$ to be non-negative definite. This decomposition is also convenient numerically

as it involves unconstrained parameters for numerical optimization. See Lindstrom and Bates (1988); Pinhero and Bates (1996); Bates and DebRoy (2003).

1.3 Variable Selection in LME models

Variable selection for LME models is a challenging problem which has its application in many disciplines. As a motivational example, let us consider a recent study (Lee and Ghosh, 2008) of the association between total nitrate concentration in the atmosphere and a set of measured predictors using the U.S. EPA CASTnet data. The dataset consists of multiple sites with repeated measurements of nitrate concentration along with a set of potential covariates on each site. To analyze this data one could use an LME model, which takes into account the possible heterogeneity among the different study sites by introducing one or more site-specific random intercept and slopes, this model specification allows us the flexibility to model both the means as well as the covariance structure. But one would need to take care to fit the appropriate covariance structure of the random effects, as underfitting it might result in the underestimation of the standard errors for the fixed effects. Overfitting the covariance structure on the other hand, could result in near singularities (Lange and Laird, 1989). Therefore, an important problem in LME models is how to simultaneously select the important fixed and random effect components, as changing one of the two types of effects greatly affects the other.

Given the LME model (1.15), it is assumed that only an unknown subset of fixed effects and random effects are nonzero. The variable selection problem is now to identify the two unknown subsets. The problem of variable selection in this context of LME models has received little or no attention. Few methods have been proposed in the literature and we shall discuss some of them in detail.

1.3.1 Subset Selection

Information-based criteria are the most commonly used method to identify the important fixed as well as the important random effects in the model. This is done by computing the AIC (Akaike, 1973) and BIC (Schwartz, 1978) for all plausible models in consideration. Let \mathbb{M} represent the set of candidate models. Let M denote a candidate model such that, $M \in \mathbb{M}$. Let $\boldsymbol{\theta}_M = (\boldsymbol{\beta}'_M, \text{vech}'(\boldsymbol{\Psi}_M), \sigma^2_M)'$ denote the vector of parameters under model M . Given M , The AIC and BIC is computed as

$$\begin{aligned} \text{AIC}_M &= -2L_{ML}(\hat{\boldsymbol{\theta}}_M) + 2 \times \dim(\boldsymbol{\theta}_M), \\ \text{BIC}_M &= -2L_{ML}(\hat{\boldsymbol{\theta}}_M) + \log(N) \times \dim(\boldsymbol{\theta}_M), \end{aligned} \quad (1.31)$$

where $\hat{\boldsymbol{\theta}}_M$ is the maximizer of $L_{ML}(\boldsymbol{\theta})$ given in (1.22) under model M , and $\dim(\boldsymbol{\theta}_M)$ is the dimension of $\boldsymbol{\theta}_M$. Given the LME model (1.15), the total number of possible sub-models includes all possible combinations resulting from the mean and the covariance structures, that is 2^{p+q} . Hence, this method could result in being extremely time consuming as well as computationally demanding especially if p and q are very large.

In order to slightly reduce the computation time Wolfinger (1993) and Diggle, Liang and Zeger, (1994) proposed the Restricted Information Criterion (which we denote by REML.IC). Using the most complex mean structure i.e. including all possible fixed effects in the model, selection is first performed on the variance-covariance structure of the random effects by computing the AIC/BIC value for every possible covariance structure using the restricted log-likelihood given in (1.25). The degrees of freedom corresponds to the number of free variance-covariance parameters in M_q . Given the ‘best’ covariance structure, selection is then performed on the fixed effects. Hence, the number of possible sub-models evaluated for this procedure is the sum of the mean structures and the covariance structures in consideration. That is, given the LME model (1.15) the total possible models to enumerate is $2^p + 2^q$.

Along the same lines, Pu and Niu (2006) proposed an extension of the GIC (Rao and Wu, 1989) procedure called the EGIC (Extended Generalized Information Criterion) where selection is first performed on the mean structure of the model by

including all the random effects in the model, using the BIC. Once the fixed effects are chosen selection is then performed on the random effects. They further showed that this procedure is consistent and asymptotically efficient. But as seen from their simulation study, this method performs poorly when it is used to select the random effects.

To avoid enumerating a large number of possible sub-models, Morell, Pearson and Brant (1997) proposed a backward selection criteria to identify the important fixed and random effects. Given the full model, elimination is first performed on the fixed effect which do not have matching random effects. Once all the highest-order fixed effects are statistically significant, backward elimination of the fixed and the random effects is performed starting with the highest-order factors. At each step of the elimination of the fixed effect, the corresponding random effect can be considered for elimination. Though this method is much faster than enumerating all possible models it faces the drawback of being extremely unstable due to its inherent discreteness (Breiman, 1996).

Recently, Jiang, Rao, Gu, Nguyen (2006) proposed a ‘fence’ method in a more general mixed models setting, a special case being the LME model. The idea is to isolate a sub group of models called the ‘correct’ model by using a ‘statistical fence’ which eliminates the ‘incorrect’ models. The optimal model is then selected from this group within the fence according to some criteria., the statistical fence is constructed in the following way,

$$-L_{ML}(\hat{\boldsymbol{\theta}}_M) \leq -L_{ML}(\hat{\boldsymbol{\theta}}_{\tilde{M}}) + c_n \times \hat{\sigma}_{(M, \tilde{M})}, \quad (1.32)$$

where $\tilde{M} \in \mathbb{M}$ represents the full model, and $(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{\tilde{M}})$ are the local maximizers of $L_{ML}(\hat{\boldsymbol{\theta}}_M)$ and $L_{ML}(\hat{\boldsymbol{\theta}}_{\tilde{M}})$ respectively, and where $\hat{\sigma}_{(M, \tilde{M})} = \sqrt{(\dim(\boldsymbol{\theta}_{\tilde{M}}) - \dim(\boldsymbol{\theta}_M)) / 2}$. Jiang et al. (2008) showed that this difference is small for the correct models and large for an incorrect model. Once this fence has been constructed and has identified a set of candidate models a simple stepwise algorithm is used to identify the optimal model. Though these techniques avoid searching through the entire model

space, it could still result in being computationally intensive especially if the number of predictors (p, q) is large.

1.3.2 Bayesian Variable Selection for LME Models

A common approach to Bayesian variable selection in LME models is as follows. Again, let \mathbb{M} denote the model space and let M denote a candidate model such that, $M \in \mathbb{M}$. Given M , we have

$$\mathbf{y}|M, \boldsymbol{\beta}_M, \sigma_M^2, \boldsymbol{\Psi}_M \sim N(\mathbf{X}_M \boldsymbol{\beta}_M, \tilde{\mathbf{V}}_M), \quad (1.33)$$

where $\tilde{\mathbf{V}}_M$ is a block diagonal matrix of $\mathbf{V}_{i,M} = \sigma_M^2(\mathbf{I}_{n_i} + \mathbf{Z}_{i,M} \boldsymbol{\Psi}_M \mathbf{Z}'_{i,M})$ for $i = 1, \dots, m$, under model M .

Given (1.33), the Bayesian selection approach is to assign a prior distribution to the model parameters $(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$ for each model indexed by M in a hierarchical fashion. Such that

$$P(\boldsymbol{\beta}_M, \boldsymbol{\Psi}_M, \sigma^2, M) = P(\boldsymbol{\beta}_M|M, \boldsymbol{\Psi}_M, \sigma^2)P(\boldsymbol{\Psi}_M|M)P(\sigma^2|M)P(M). \quad (1.34)$$

Bayes theorem is then used to calculate posterior probabilities.

Using this hierarchical formulation, Weiss, Wang and Ibrahim (1997) proposed a Bayesian method with emphasis on the selection of the fixed effects using Bayes factors by keeping the number of random effects fixed. They use a predictive approach to specify the priors with the emphasis on selecting the fixed effects. Similar to the Zellner's g-priors but in a LME model setting, Weiss et al. (1997) proposed a conjugate Normal-Gamma prior given as

$$\begin{aligned} \boldsymbol{\beta}_M|M, \tilde{\mathbf{V}}_M &\sim N\left(\boldsymbol{\mu}_M, \frac{(\mathbf{X}'_M \tilde{\mathbf{V}}_M^{-1} \mathbf{X}_M)^{-1}}{g}\right) \\ \sigma^2 &\sim IG(a_0, b_0), \end{aligned} \quad (1.35)$$

where $\boldsymbol{\mu}_M = (\mathbf{X}'_M \tilde{\mathbf{V}}_M^{-1} \mathbf{X}_M)^{-1} \mathbf{X}'_M \tilde{\mathbf{V}}_M^{-1} \boldsymbol{\mu}_0$, and $\boldsymbol{\mu}_0$ is a constant. In addition, $g > 0$

is a known scaling factor, and $a_0 > 0$, $b_0 > 0$ are known parameters of the Inverse Gamma distribution with mean $\frac{a_0}{b_0-1}$. As for the prior on Ψ they use a Wishart distribution. Final model selection is made based on Bayes factors.

Recently, Chen and Dunson (2003) proposed a hierarchical Bayesian approach to select the important random effects in a linear mixed-effects model. They reparameterized the linear mixed models using a modified Cholesky decomposition such that they can employ standard Bayesian variable selection techniques. They choose a mixture prior with point mass at zero for the variance components parameters. This allows for the random effects to effectively drop out of the model. They then employ Gibbs sampler to sample from the full conditional distribution to select the non-zero random effects. This was later extended to logistic models by Kinney and Dunson (2006).

1.3.3 The Likelihood Ratio Test

The likelihood ratio tests (LRT) are classical statistical tests for making a decision between two hypotheses: the null hypothesis, $H_0 \in \Theta_0$, and the alternative, $H_a \in \Theta$, where Θ is as given in (1.21). The LRT is a commonly used method to compare models with different means as well as different covariance structures in a mixed model setting. The LRT is computed as

$$-2\log(c_n) = -2 \left(L_{ML}(\hat{\theta}_0) - L_{ML}(\hat{\theta}) \right), \quad (1.36)$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are the maximum likelihood estimates obtained by maximizing $L_{ML}(\theta)$ given in (1.22) under H_0 and H_a , respectively. Under certain regularity conditions, $-2\log(c_n)$ asymptotically follows a χ^2 distribution, under the null hypothesis, with degrees of freedom equal to the difference in the dimension of Θ_0 and Θ .

Non-standard likelihood ratio based methods proposed by Self and Liang (1987) have been used for making inference when the true parameters are allowed to be on the boundary of the parameter space. Stram and Li (1994) describe in detail the

use of non-standard LRT to test for non-zero variance components under the null hypothesis that such a component is zero and to decide whether or not to include that specific random effect. They showed that the test under the null hypothesis that a single variance component is zero does not follow the classical χ^2 distribution but rather a 50:50 mixture of chi-square distributions. Furthermore, by not accounting for this discrepancy they also showed that the p-values will be overestimated and we would accept the null more often than we should. However, the main drawback of these procedures is that simultaneous testing of multiple random effects becomes prohibitively impossible when the number of predictors is moderately large.

1.3.4 Selection with Shrinkage Penalty

Recently, Lan (2006) extended the shrinkage penalty approach using the SCAD penalty (Fan and Li, 2001) to variable selection in LME models. They assume that the covariance structure for the random effects can be specified, and their main interest was to select the subset of variables associated with the fixed effects. Given the LME model (1.17) and an initial estimate for the variance components, they obtain the penalized likelihood estimates for the regression coefficients by minimizing

$$G(\boldsymbol{\beta}) = -L_{ML}(\boldsymbol{\theta}) + \sum_{j=1}^p p_{\lambda}(|\beta_j|), \quad (1.37)$$

where $L_{ML}(\boldsymbol{\theta})$ is as given in (1.22) and $p_{\lambda}(|\beta_j|)$ for the SCAD penalty is as given in Table 1.1. Given the penalized likelihood estimates for $\boldsymbol{\beta}$, the estimates for the variance components are obtained using the traditional REML procedure. The same idea was then extended to generalized linear mixed models (GLMM) by Yang (2007)

Notice that the method proposed by Lan (2006) does not perform selection on the random effects but rather assumes that the number of random effects in the model is fixed. A difficulty in defining a shrinkage approach to random effects selection is that an entire row and column of $\boldsymbol{\Psi}$ must be eliminated to successfully remove a random

effect. This leads to complications in how to perform the shrinkage appropriately. In Chapter 3 we propose a new method to simultaneously identify the subset of $p' \leq p$ and $q' \leq q$ important predictors that corresponds to the fixed and the random components in the LME model (1.17), respectively. Our proposed method is based on a reparameterization of the LME model obtained by the modified Cholesky decomposition of $\Psi = \mathbf{D}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}$ (Chen and Dunson, 2003), where \mathbf{D} is a diagonal matrix that is proportional to the standard deviations of the random effects, and $\mathbf{\Gamma}$ is a lower triangular matrix with 1's on the diagonal that relates to the correlation among the random effects. This factorization aids us in the selection of the random effects by dropping out the random effects terms which have zero variance (Chapter 3, Section 3.2). We propose a penalized joint log-likelihood procedure with an adaptive penalty for the selection of the fixed and random effects. We obtain the estimates for our parameters by using a constrained EM algorithm, by first computing the conditional expectation of the penalized joint log-likelihood and maximizing the objective function using an optimization routine, to obtain the final penalized likelihood estimates. We also show that our penalized likelihood estimates enjoys the ‘Oracle’ property (Definition 1) and performs asymptotically as well as if the true model was known before hand. This is shown both empirically as well as theoretically. We show that our method outperforms the commonly used methods describe in this section via a simulation study and a real data example.

1.4 Plan of Dissertation

The remainder of this dissertation is as follows. We present our research in the form of two papers that have been submitted to academic journals. In Chapter 2 we present the Adaptive Powered Correlation Prior which allows highly correlated predictors to enter or exit the model simultaneously. In Chapter 3 we present our method for simultaneously selecting the important fixed and random effects within a LME model framework.

Table 1.1: Penalized Regression methods and their corresponding penalty terms

Shrinkage Penalties	
Method	Penalty
Ridge Regression	$\lambda \sum_{j=1}^p \beta_j^2$
Bridge Regression	$\lambda \sum_{j=1}^p \beta_j ^q$
LASSO	$\lambda \sum_{j=1}^p \beta_j $
Adaptive LASSO	$\lambda \sum_{j=1}^p \bar{w}_j \beta_j $
Elastic-Net	$\lambda_1 \sum_{j=1}^p \beta_j + \lambda_2 \sum_{j=1}^p \beta_j^2$
OSCAR	$\lambda_1 \sum_{j=1}^p \beta_j + \lambda_2 \sum_{j \leq k} \max(\beta_j , \beta_k)$
SCAD	$\sum_{j=1}^p p_\lambda(\beta_j)$ Where,

$$p_\lambda(|\omega|) = \begin{cases} |\omega| & |\omega| \leq \lambda \\ -\frac{1}{(a-1)\lambda} (|\omega|^2 - 2a\omega + \lambda^2) & \lambda \leq |\omega| \leq a\lambda \\ \frac{1}{2}(a+1)\lambda & |\omega| \geq a\lambda \end{cases}$$

Chapter 2

Bayesian Variable Selection Using Adaptive Powered Correlation Prior

2.1 Introduction

Consider the linear regression model with n independent observations and let $\mathbf{y} = (y_1, \dots, y_n)'$ be the vector of response variables. The canonical linear model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2.1)$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ matrix of explanatory variables with $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})'$ for $j = 1, \dots, p$. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ be the corresponding vector of unknown regression parameters, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$. Throughout this chapter, we assume \mathbf{y} to be empirically centered to have mean zero, while the columns of \mathbf{X} have been standardized to have mean zero and norm one, so $\mathbf{X}'\mathbf{X}$ will be the empirical correlation matrix.

Under the above regression model, it is assumed that only an unknown subset of the coefficients are nonzero, so that the variable selection problem is to identify this unknown subset. Bayesian approaches to the problem of selecting variables/predictors

within a linear regression framework has received considerable attention over the years, for example see, Mitchell and Beauchamp (1988), Geweke (1996), George and McCulloch (1993, 1997), Brown, Vannucci and Fearn (1998), George (2000) and Chipman, George and McCulloch (2001) and Casella and Moreno (2006).

For the linear model, Zellner (1986) suggested a particular form of a conjugate Normal-Gamma family called the g -prior which can be expressed as

$$\begin{aligned}\boldsymbol{\beta}|\sigma^2, \mathbf{X} &\sim N\left(0, \frac{\sigma^2}{g}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ \sigma^2 &\sim IG(a_0, b_0),\end{aligned}\tag{2.2}$$

where $g > 0$ is a known scaling factor and $a_0 > 0$, $b_0 > 0$ are known parameters of the Inverse Gamma distribution with mean $\frac{a_0}{b_0-1}$. The prior covariance matrix of $\boldsymbol{\beta}$ is the scalar multiple σ^2/g of the inverse Fisher information matrix, which concurrently depends on the observed data through the design matrix \mathbf{X} .

This particular prior has been widely adopted in the context of Bayesian variable selection due to its closed form calculations of all marginal likelihoods which is suitable for rapid computations over a large number of submodels, and its simple interpretation that it can be derived from the idea of a likelihood for a pseudo- data set with the same design matrix \mathbf{X} as the observed sample (see, Zellner (1986), George and Foster (2000), Smith and Kohn (1996), Fernandez, Ley and Steel (2001)).

In this chapter, we point out a drawback of using Zellner's prior on $\boldsymbol{\beta}$ particularly when the predictors (\mathbf{x}_j) are highly correlated. The conditional variance of $\boldsymbol{\beta}$ given σ^2 and \mathbf{X} is based on the inverse of the empirical correlation of predictors and puts most of its prior mass in the direction that causes the regression coefficients of correlated predictors to be smoothed away from each other. So when coupled with model selection, Zellner's prior discourages highly collinear predictors to enter the models simultaneously by inducing a negative correlation between the coefficients.

We propose a modification of Zellner's g -prior by replacing $(\mathbf{X}'\mathbf{X})^{-1}$ by $(\mathbf{X}'\mathbf{X})^\lambda$ where the power $\lambda \in \mathbb{R}$, controls the amount of smoothing of collinear predictors towards or away from each other accordingly as $\lambda > 0$ or $\lambda < 0$, respectively. For $\lambda > 0$,

the new conditional prior variance of β puts more prior mass in the direction that corresponds to a strong prior smoothing of regression coefficients of highly collinear predictors towards each other. Therefore, by choosing $\lambda > 0$ our proposed modification in contrast, forces highly collinear predictors entering or exiting the model simultaneously (see Section 2). Hence, the use of the power hyperparameter λ to the empirical correlation matrix helps us to determine whether models with high collinear predictors are preferred or not.

The hyperparameter λ is further incorporated into the prior probabilities over model space with the same intentions of encouraging or discouraging the inclusion of groups of correlated predictors. We adopt a Bayesian Hierarchical framework with the new prior specifications. The choice of hyperparameter is obtained via an empirical Bayes approach and the inference regarding model selection is then made based on the posterior probabilities. By allowing the power parameter λ to be chosen by the data, we let the data decide whether to include collinear predictors or not.

The remainder of the Chapter is structured as follows. In Section 2, we describe in detail the Powered Correlation Prior and provide a simple motivating example, when $p = 2$. Section 3, describes the choice of new prior specifications for model selection. The Bayesian hierarchical model and the calculation of posterior probabilities are presented in Section 4. The superior performance of using the Powered Correlation Prior over Zellner's g-priors is illustrated with the help of simulation studies and real data examples in Section 5. Finally, in Section 6 we conclude with a discussion.

2.2 The Adaptive Powered Correlated Prior

Consider again a normal regression model as in (2.1), where $\mathbf{X}'\mathbf{X}$ represents the correlation matrix. Let $\mathbf{X}'\mathbf{X} = \mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}'$ be the spectral decomposition, where the columns of $\mathbf{\Gamma}$ are the p orthonormal eigenvectors and \mathbf{D} is the diagonal matrix with eigenvalues $d_1 \geq \dots \geq d_p \geq 0$ as the diagonal entries. The powered correlation prior

for $\boldsymbol{\beta}$ conditioned on σ^2 and \mathbf{X} is defined as

$$\boldsymbol{\beta}|\sigma^2, \mathbf{X} \sim N\left(0, \frac{\sigma^2}{g}(\mathbf{X}'\mathbf{X})^\lambda\right), \quad (2.3)$$

where $(\mathbf{X}'\mathbf{X})^\lambda = \boldsymbol{\Gamma}\mathbf{D}^\lambda\boldsymbol{\Gamma}'$, with $g > 0$ and $\lambda \in \mathbb{R}$ controlling the strength and the shape, respectively, of the prior covariance matrix, for a given $\sigma^2 > 0$.

There are several priors which are special cases of the powered correlation prior. For instance, $\lambda = -1$ produces the Zellner's g -prior (2.2). By setting $\lambda = 0$ we have $(\mathbf{X}'\mathbf{X})^0 = I$ which gives us the ridge regression model of Hoerl and Kennard (1970), under this model β_j are given independent $N(0, \sigma^2/g)$ priors. Next we illustrate how λ controls the model's response to collinearity which is the main motivation for using the powered correlation prior.

Let $\mathbf{T} = \mathbf{X}\boldsymbol{\Gamma}$ and $\boldsymbol{\theta} = \boldsymbol{\Gamma}'\boldsymbol{\beta}$. The linear model can be written in terms of the principal components as:

$$\mathbf{y} \sim N(\mathbf{T}\boldsymbol{\theta}, \sigma^2) \text{ with } \boldsymbol{\theta} \sim N\left(0, \frac{\sigma^2}{g}\mathbf{D}^\lambda\right). \quad (2.4)$$

The columns of the new design matrix \mathbf{T} are the principal components, and so the original prior on $\boldsymbol{\beta}$ can be viewed as independent mean zero normal priors on the principal component regression coefficients, with prior variance proportional to the power of the corresponding eigenvalues, $d_1^\lambda, \dots, d_p^\lambda$. Principal components with d_i near zero indicate a presence of a near-linear relationship between the predictors, and the direction determined by the corresponding eigenvectors are those which are uninformative from the data. A classical frequentist approach to handle collinearity is to use principal component regression, and eliminate those dimensions with very small eigenvalues. Then transform back to the original scale, so that no predictors are actually removed. Along the same lines, we shall illustrate on how changing the value of λ would affect the prior correlation and demonstrate the intuition behind our proposed modification. For a simple illustration consider the case with $p = 2$ with a

positive correlation ρ between them so that

$$(\mathbf{X}'\mathbf{X})^\lambda = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^\lambda. \quad (2.5)$$

It easily follows that in this case,

$$\mathbf{\Gamma} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and,} \quad \mathbf{D}^\lambda = \begin{bmatrix} (1 + \rho)^\lambda & 0 \\ 0 & (1 - \rho)^\lambda \end{bmatrix}. \quad (2.6)$$

The first principal component of our new design matrix \mathbf{T} can be written as the sum of the predictors and the second as the difference

$$\mathbf{T} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{x}_1 + \mathbf{x}_2 \\ \mathbf{x}_1 - \mathbf{x}_2 \end{bmatrix}' \quad \text{with,} \quad \boldsymbol{\theta} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{\sigma^2}{g} \begin{bmatrix} (1 + \rho)^\lambda & 0 \\ 0 & (1 - \rho)^\lambda \end{bmatrix} \right), \quad (2.7)$$

for $\lambda > 0$ the prior on the coefficient for the sum has mean zero and variance $(1 + \rho)^\lambda$, while the prior on the coefficient for the difference has mean zero and variance $(1 - \rho)^\lambda$.

As ρ in (2.7) increases, a smaller prior variance is given to the coefficient for the difference of the two predictors, and hence introduces more shrinkage to the principal component directions that are associated with small eigenvalues. So that larger λ forces the difference to be more likely closer to the prior mean (zero). On the original $\boldsymbol{\beta}$ scale this corresponds to strong prior smoothing of regression parameters corresponding to highly collinear predictors.

On the other hand $\lambda < 0$ places a large prior variance on the coefficient for the difference, and a smaller variance on the coefficient of the sum, thereby shrinking those directions which correspond to large eigenvalues. This has an effect of smoothing the regression parameters corresponding to highly collinear predictors away from each other, forcing the two predictors to be negatively correlated.

Hence in dimensions greater than two, in the presence of collinear predictors, λ has the flexibility to introduce more shrinkage in the directions that correspond to the small eigenvalues. This behavior motivates us to allow for the possibility of choosing

alternative values for λ . In particular, we allow the data to determine the choice of λ using an empirical Bayes approach.

2.3 Model Specification

The main focus here is to use this powered correlation prior in a model selection problem. For the linear regression model in (2.1), it is typically the case that only an unknown subset of the coefficients β_j are non-zero, so in the context of variable selection we begin by indexing each candidate model with one binary vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$ where each element δ_j takes the value 1 or 0 depending on whether it is included or excluded from the model. More specifically, let

$$\delta_j = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ is included in the model,} \\ 0 & \text{if } \mathbf{x}_j \text{ is excluded from the model.} \end{cases} \quad (2.8)$$

We now rewrite the linear regression model, given $\boldsymbol{\delta}$ as

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\delta}}\boldsymbol{\beta}_{\boldsymbol{\delta}} + \boldsymbol{\epsilon}, \quad (2.9)$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, and $\mathbf{X}_{\boldsymbol{\delta}}$ and $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ are the design matrix and the regression parameters of the model only including the predictors with $\delta_j = 1$. In the context of variable selection we can write the powered correlation prior as

$$\begin{aligned} \boldsymbol{\beta}_{\boldsymbol{\delta}} | \boldsymbol{\delta}, \sigma^2, \mathbf{X} &\sim N\left(0, \frac{\sigma^2}{g} (\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{\lambda}\right) \\ \text{with } (\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{\lambda} &= \boldsymbol{\Gamma}_{\boldsymbol{\delta}} \mathbf{D}_{\boldsymbol{\delta}}^{\lambda} \boldsymbol{\Gamma}'_{\boldsymbol{\delta}}, \end{aligned} \quad (2.10)$$

where $\boldsymbol{\Gamma}_{\boldsymbol{\delta}}$ is the matrix of eigenvectors and $\mathbf{D}_{\boldsymbol{\delta}}^{\lambda}$ is a diagonal matrix with diagonal entries as the eigenvalues of $(\mathbf{X}'_{\boldsymbol{\delta}} \mathbf{X}_{\boldsymbol{\delta}})^{\lambda}$.

Now that we have defined the prior for the coefficients given the model we now incorporate the same idea into the choice of prior for the inclusion indicators. With

respect to Bayesian variable selection, a common prior for the inclusion indicators is, $p(\boldsymbol{\delta}) \propto \pi^{p_\delta}(1 - \pi)^{p-p_\delta}$ (George and McCulloch, 1993,1997; George and Foster, 2000) where $p_\delta = \sum_{j=1}^p \delta_j$ is the number of predictors in the model defined by $\boldsymbol{\delta}$, and π is the prior inclusion probability for each covariate. We can see this being equivalent to placing Bernoulli (π) priors on δ_j and thereby giving equal weight to any pair of equally-sized models. Setting $\pi = 1/2$ yields the popular uniform prior over model space formed by considering all subsets of predictors and, under this prior the posterior model probability is proportional to the marginal likelihood. A drawback of using this prior is that it puts most of its mass on models of size $\simeq p/2$ and it does not take into account the correlation between the predictors. Yuan and Lin (2005) proposed an alternative prior over model space.

$$P(\boldsymbol{\delta}|\pi) \propto \pi^{p_\delta}(1 - \pi)^{p-p_\delta} |\mathbf{X}'_\delta \mathbf{X}_\delta|^{1/2}, \quad (2.11)$$

where $|\cdot|$ denotes the determinant, and $|\mathbf{X}'_\delta \mathbf{X}_\delta| = 1$ if $p_\delta = 0$. Since $|\mathbf{X}'_\delta \mathbf{X}_\delta|$ is small for models with highly collinear predictors, this prior discourages these models. We follow Yuan and Lin in that we use the information from the design matrix to build a prior for the model space. However, we do not necessarily want to penalize models with collinear predictors. We propose to incorporate the power parameter λ into a prior for $\boldsymbol{\delta}$ that could encourage or discourage inclusion of groups of correlated predictors. So we propose the following prior on model space:

$$P(\boldsymbol{\delta}|\lambda, \pi) \propto \pi^{p_\delta}(1 - \pi)^{p-p_\delta} |\mathbf{X}'_\delta \mathbf{X}_\delta|^{-\lambda/2}. \quad (2.12)$$

So for large values of λ , the prior puts more of its mass on models with highly collinear predictors; while for $\lambda < 0$, penalizing models with collinear predictors. Hence coupled with the powered correlation prior, positive (negative) λ encourages (discourages) highly collinear predictors to enter the model simultaneously. Note that $\lambda = -1$ gives us Zellner's prior on the coefficients coupled with the prior of Yuan and Lin on the models.

2.3.1 Choice of g

The parameter g defines the strength of the powered correlation prior. The choice of g is complicated in that large values of g will result in the prior dominating the likelihood, and small values of g would favor the null model (George and Foster, 2000). Various choices of g have been proposed over the years. For example, Smith and Kohn (1996) performed variable selection involving splines with a fixed value of $g = .01$. However the choice of g may also depend on the sample size n , or the number of predictors p . George and Foster (2000) propose an empirical Bayes method for estimating g from its marginal likelihood. Foster and George (1994) recommended using $g = 1/p^2$ based on a Risk Inflation Criterion (RIC). Kass and Wasserman (1995) suggests the unit information prior, where the amount of information in the prior corresponds to the amount of information in one observation, leading to $g = 1/n$. This leads to the Bayes factor as an approximation of the BIC. Fernandez *et. al.* (2001) suggest $g = 1/\max(n, p^2)$ called the Benchmark Prior (BRIC), which is a combination of RIC and BIC. More recently Liang *et. al.*(2008) suggest a mixture of g -priors as an alternative to the default g -priors.

Since the scale of $(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda}$ will depend on λ , we first standardize so that we may separate out the scale of g from that of $(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda}$. To do so we modify (2.10) as

$$\begin{aligned} \boldsymbol{\beta}_{\boldsymbol{\delta}}|\boldsymbol{\delta}, \sigma^2, \mathbf{X} &\sim N(0, \frac{\sigma^2}{g}k(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda}) \\ \text{with } k \equiv k(\lambda, \boldsymbol{\delta}, \mathbf{X}) &= \frac{\text{Tr}[(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}]}{\text{Tr}[(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda}]} \end{aligned} \tag{2.13}$$

This has an effect of setting the trace of $(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda}$ to be equal to that of using $\lambda = -1$ regardless of the choice of λ . We then choose $g = 1/n$, as in the unit information prior (Kass and Wasserman, 1995).

For $(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{\lambda} = \boldsymbol{\Gamma}_{\boldsymbol{\delta}}\mathbf{D}_{\boldsymbol{\delta}}^{\lambda}\boldsymbol{\Gamma}_{\boldsymbol{\delta}}$, k can be considered as the ratio of the average eigenvalues with those of $\lambda = -1$, $\frac{\sum_j \mathbf{D}_{\delta_j}^{-1}}{n} / \frac{\sum_j \mathbf{D}_{\delta_j}^{\lambda}}{n}$. Instead of the trace one could have

opted to choose the determinant, i.e. the ratio of the product of the eigenvalues. An advantage of using the average of the eigenvalues is that it provides more stability and in turn helps prevent the prior from dominating the likelihood. We note that other choices of standardization and choice of g are possible and are left for future investigations.

2.4 Model Selection using Posterior Probabilities

In the Bayesian framework, a set of prior distributions is specified on the parameters $\boldsymbol{\theta}_\delta = (\boldsymbol{\beta}_\delta, \sigma^2)$ for each model, along with a meaningful set of prior model probabilities $P(\boldsymbol{\delta}|\lambda, \pi)$ over the class of all models. Model selection is then done based on the posterior probabilities. Using the set of priors defined in the previous section, we can now construct a hierarchical Bayesian model to perform variable selection

$$\begin{aligned}
 \mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X} &\sim N(\mathbf{X}_\delta \boldsymbol{\beta}_\delta, \sigma^2 I) \\
 \boldsymbol{\beta}_\delta|\boldsymbol{\delta}, \sigma^2, \mathbf{X} &\sim N(\boldsymbol{\beta}_0, \frac{\sigma^2}{g} k(\mathbf{X}'_\delta \mathbf{X}_\delta)^\lambda), \\
 \sigma^2 &\sim IG(\frac{\gamma_o}{2}, \frac{\gamma_o}{2}), \\
 P(\boldsymbol{\delta}|\lambda, \pi) &\propto \pi^{p_\delta} (1 - \pi)^{p - p_\delta} |\mathbf{X}'_\delta \mathbf{X}_\delta|^{-\lambda/2},
 \end{aligned} \tag{2.14}$$

where k is as defined in (2.13) The key idea in computing the posterior model probabilities is to obtain the marginal likelihood of the data under model $\boldsymbol{\delta}$ by integrating out the model parameters

$$P(\mathbf{y}|\boldsymbol{\delta}, \mathbf{X}) = \int P(\mathbf{y}|\boldsymbol{\theta}_\delta, \boldsymbol{\delta}, \mathbf{X}) P(\boldsymbol{\theta}_\delta|\boldsymbol{\delta}, \mathbf{X}) d\boldsymbol{\theta}_\delta. \tag{2.15}$$

The choice of conjugate priors allows us to analytically compute the above integral. Using the hierarchical model and integrating out $\boldsymbol{\theta}_\delta$ we obtain the conditional distri-

bution of \mathbf{y} given $\boldsymbol{\delta}$ and \mathbf{X} ,

$$\mathbf{y}|\boldsymbol{\delta}, \mathbf{X} \sim t_{(\gamma_o+n)} \left\{ \mathbf{X}_\boldsymbol{\delta} \boldsymbol{\beta}_0, \left(\mathbf{I}_n + \frac{k}{g} \mathbf{X}_\boldsymbol{\delta} (\mathbf{X}'_\boldsymbol{\delta} \mathbf{X}_\boldsymbol{\delta})^\lambda \mathbf{X}'_\boldsymbol{\delta} \right) \right\}. \quad (2.16)$$

Then model comparison is done via the posterior probabilities,

$$P(\boldsymbol{\delta}|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\boldsymbol{\delta}, \mathbf{X})P(\boldsymbol{\delta}|\lambda, \pi) \quad (2.17)$$

In order to fully specify our prior distribution we need to specify g , γ_o , π and λ . We choose $g = 1/n$ the unit information prior proposed by Kass and Wasserman (1995). For γ_o , after trying various choices, we saw that the model selected was not sensitive to the value of γ_o chosen, and since there is little or no information about this hyperparameter we decided to set γ_o to a constant, which has led to reasonable results as pointed out by George and McCulloch (1997). Following these lines, we set $\gamma_o = 0.01$ for the rest of the article, which corresponds to placing a non-informative prior on σ^2 .

The parameters (λ, π) are very influential and informative with respect to the model selected and it is of utmost importance that we choose them carefully. Thus, we propose an empirical Bayes approach to select $\pi \in (0, 1)$ and $\lambda \in \mathbb{R}$ by marginalizing over $\boldsymbol{\delta}$ and maximizing the marginal likelihood function given by

$$m(\mathbf{y}|\mathbf{X}, \pi, \lambda) = \sum_{\boldsymbol{\delta}} P(\mathbf{y}|\boldsymbol{\delta}, \mathbf{X})P(\boldsymbol{\delta}|\lambda, \pi). \quad (2.18)$$

When the number of predictors, p is of moderate size (e.g., $p \leq 20$) the above sum can be computed by evaluating (2.15) for each model via complete enumeration for a given (π, λ) . Numerical optimization is then used to maximize $m(\mathbf{y}|\mathbf{X}, \pi, \lambda)$ defined in (2.18) to obtain the pair $(\hat{\pi}, \hat{\lambda})$. Specifically, we fix λ on a fine grid and for each λ , we maximize over $\pi \in (0, 1)$, and obtain $\hat{\pi}(\lambda)$ and obtain $\hat{\lambda} = \operatorname{argmax} m(\mathbf{y}|\mathbf{X}, \hat{\pi}(\lambda), \lambda)$.

2.5 Simulations and Real Data

We shall now compare our proposed method to the standard use of Zellner's prior with a uniform prior over model space, i.e. the common approach

$$\begin{aligned}\beta_{\boldsymbol{\delta}}|\sigma^2, \boldsymbol{\delta}, \mathbf{X} &\sim N(0, \sigma^2(\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}})^{-1}/g) \quad \text{where, } g = 1/n, \\ \sigma^2 &\sim IG\left(\frac{\gamma_o}{2}, \frac{\gamma_o}{2}\right), \quad \gamma_o = .01, \\ P(\boldsymbol{\delta}) &= (1/2)^p.\end{aligned}\tag{2.19}$$

We also compared our method to the fully automatic Bayesian variable selection procedure proposed by Casella and Moreno (2006) where posterior probabilities are computed using intrinsic priors (Berger and Pericchi, 1996) which eliminates the need for tuning parameters. However, we note that this procedure was not specifically designed to handle correlated predictors.

So, in this section we evaluate the performance of using our proposed method in selecting the correct subset of predictors as compared to the two above mentioned methods, based on a simulated data involving highly collinear predictors. Comparisons are also presented for one real dataset.

2.5.1 Simulation Study

For the simulated example, we consider the true model

$$y = x_1 + x_2 + \epsilon \quad \text{where } \epsilon \sim N(0, 1).\tag{2.20}$$

We generate p predictors from a multivariate normal with $\text{Cov}(\mathbf{x}_j, \mathbf{x}_k) = \rho^{|j-k|}$, for $\rho = 0.9$. For Case one, we fixed $p = 4$, while for Case two we used $p = 12$, so that in the first case there were 2 unimportant predictors, while in case two, there were 10. For both cases, we generated 1000 datasets each with $n = 30$ observations.

Case 1: p=4

Using the empirical Bayes approach mentioned earlier we compute the optimal pair $\hat{\lambda} = 1.6$ and $\hat{\pi} = .15$ which maximizes the marginal likelihood obtained under complete enumeration of all possible $2^4 - 1$ models. The estimates $(\hat{\lambda}, \hat{\pi})$ are the values obtained after averaging over the 1000 replications.

Figure (2.1) goes here.

From Table 2.1, the performance of our proposed method appears quite good compared to the other two methods. We see that Zellner's as well as the intrinsic prior's chooses single variable models with over half of its posterior probability. In contrast, using the powered correlation prior smoothes the regression parameters of the correlated variables towards each other, by giving more prior information in the direction that are less determined by the data, and selects the correct model, $(\mathbf{x}_1, \mathbf{x}_2)$ with an overwhelming 0.622 posterior probability.

Table 2.1 also lists the number of times (in %) each model was selected as the model with highest posterior probability out of 1000 replications by the three methods. We see that the Powered Correlation Prior based method picks the correct model, $(\mathbf{x}_1, \mathbf{x}_2)$, 68.7 % of the times.

Table () goes here.

Case 2: p=12

Similar to the previous case, the optimal values ($\hat{\lambda} = 1.7, \hat{\pi} = 0.12$) were obtained by averaging over 1000 replications (Figure 2.2) which maximizes the marginal likelihood function. The performance of the powered correlation priors in terms of selecting correlated predictors is also similar to the previous case.

Figure (2.2) goes here.

From Table 2.2 it is clear that Zellner's prior penalizes models with high collinearity, thereby putting more posterior mass on single variable models. In contrast, the

powered correlation prior method favors the true model $(\mathbf{x}_1, \mathbf{x}_2)$ with maximum average posterior probability of 0.145 and the correct model was selected 46.1 % of the time. For the intrinsic prior, the maximum posterior model is the model including only \mathbf{x}_1 and the correct model is selected only 9.8% of the times.

Table (2.2) goes here.

2.5.2 Real Data Example

We consider a real dataset to demonstrate the performance of our method. For our real data example we use the data on NCAA graduation rates (Mangold, Bean and Adams, 2003) where there are 97 observations and 19 predictors. The response variable is the average graduation rates for each of the 97 colleges (see Appendix A for a description of the dataset). Mangold, Bean and Adams used this dataset with the goal of showing that successful sports programs raise graduation rates. This dataset is of specific interest to us, due to the presence of high correlation among the variables. We fit a main effects only model with the 19 possible predictors.

For this dataset we obtain the optimal values of $\hat{\lambda} = 1.9$ and $\hat{\pi} = 0.19$ (Figure 2.3). Posterior probabilities are computed using these optimal values by complete enumeration of all $2^{19} - 1$ possible models.

Figure (2.3) goes here.

In Table 2.3 and 2.4 we compare the posterior model probabilities by using our proposed method to those obtained using the standard Zellners g-prior and the fully automatic intrinsic priors. The highest posterior model selected using the powered correlation prior to predict the average graduation rates is a 6 variable model, as compared to Zellners which selects a 5 variable model by dropping \mathbf{x}_{17} (Acceptance Rate) from the model chosen by our approach. This could be attributed to the high correlation between \mathbf{x}_2 and \mathbf{x}_{17} ($\rho = .81$). The intrinsic prior approach picks out a simpler (fewer predictors) model as its highest posterior probability model.

Table (2.3), (2.4) goes here.

Model comparison and validation are now made based on the average predicted error, where the parameter estimates are obtained by computing the posterior mean of β_{δ} for each given configuration of δ and \mathbf{y} for each of the three methods. Table (2.3) and (2.4) reports the average mean square predictive error along with their standard errors obtained using 5-fold crossvalidation (CV), whose estimates are first averaged across 10 cross-validation splits to reduce variability, and then replicated 1000 times. We see that the top two models picked out by the powered correlation prior's posterior probabilities has a significantly lower prediction error than that of the models selected using the two other methods. Hence, both the simulation and the real data example show strong support for the use of our proposed powered correlation prior.

2.6 Discussion and Future Work

In this Chapter we have demonstrated that within a linear model framework the powered correlation prior helps to resolve the problem of selecting subsets using a suitable modification of Zellners g-prior when the predictors are highly correlated. By using simulated and the real data examples we have illustrated that the powered correlation prior tends to perform better in terms of choosing the correct model than the standard Zellners prior and the intrinsic prior for correlated predictors. The choice of hyperparameter λ obtained using a empirical Bayes method controls the degree of smoothing of correlated predictors towards or away from each other.

For a large number of predictors (e.g. $p > 30$), a attractive feature of this prior is that all the parameters can be integrated out analytically to obtain a closed form for the unnormalized posterior model probabilities. Hence a simple Gibbs sampler over model space (George and McCulloch, 1997) can be implemented to approximate the marginal for each pair (λ, π) . This can be implemented on a two dimensional grid, and although it may take significant computation time, it remains feasible.

Model averaging for linear regression models has received considerable attention (Raftery, Madigan and Hoeting, 1997). This method accounts for model uncertainty

by averaging over all possible models. It is possible to extend the use of our proposed prior to perform model averaging via the use of posterior probabilities.

There has also been considerable interest in Bayesian variable selection for generalized linear models. The selection criteria are based on extensions of Bayesian methods used in linear regression framework. One can extend the powered correlation prior used here to generalized linear models.

Another possible application of our proposed method could be to use it for non-parametric curve fitting where it is of interest to choose suitable basis functions.

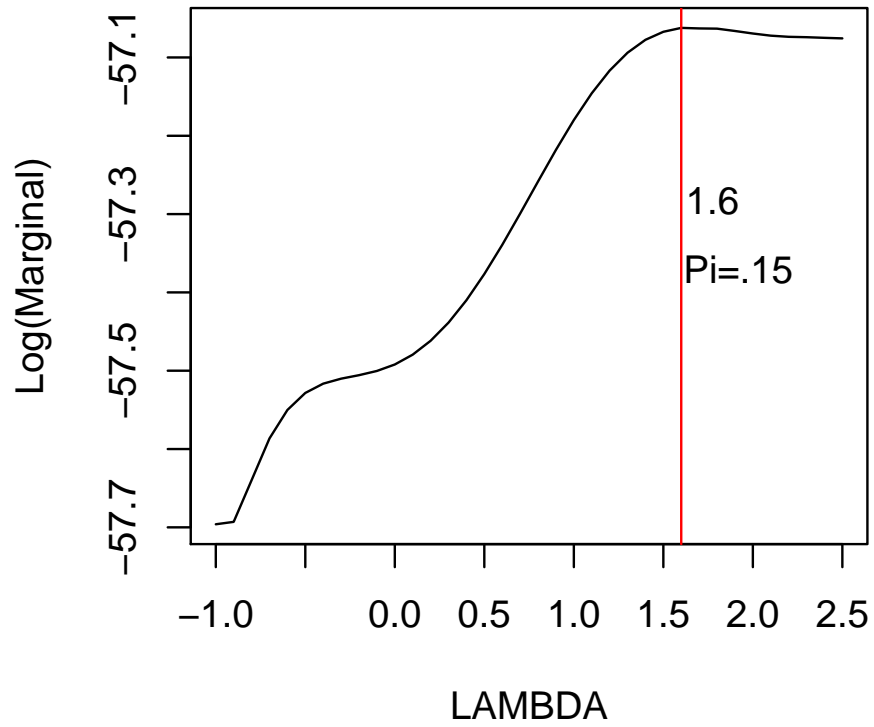


Figure 2.1: Plot of λ vs. $\text{Log}[m(\mathbf{y}|\mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 1: $p = 4$. Averaged over 1,000 simulations. The vertical line represents the location of the global maximum.

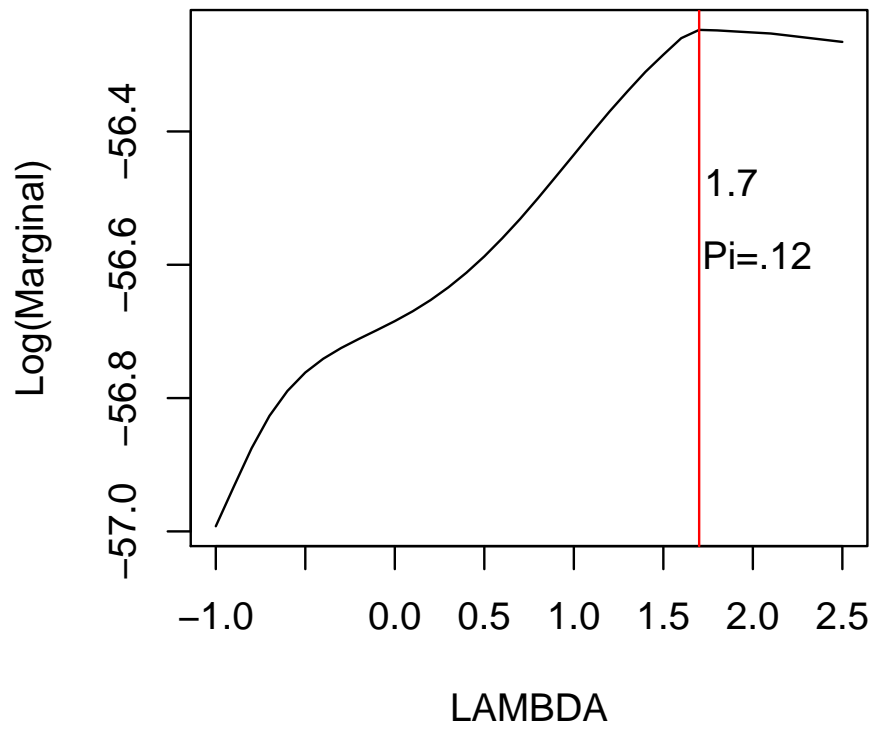


Figure 2.2: Plot of λ vs. $\text{Log}[m(\mathbf{y}|\mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 2: $p = 12$, averaged over 1,000 simulations. The vertical line represents the location of the global maximum.

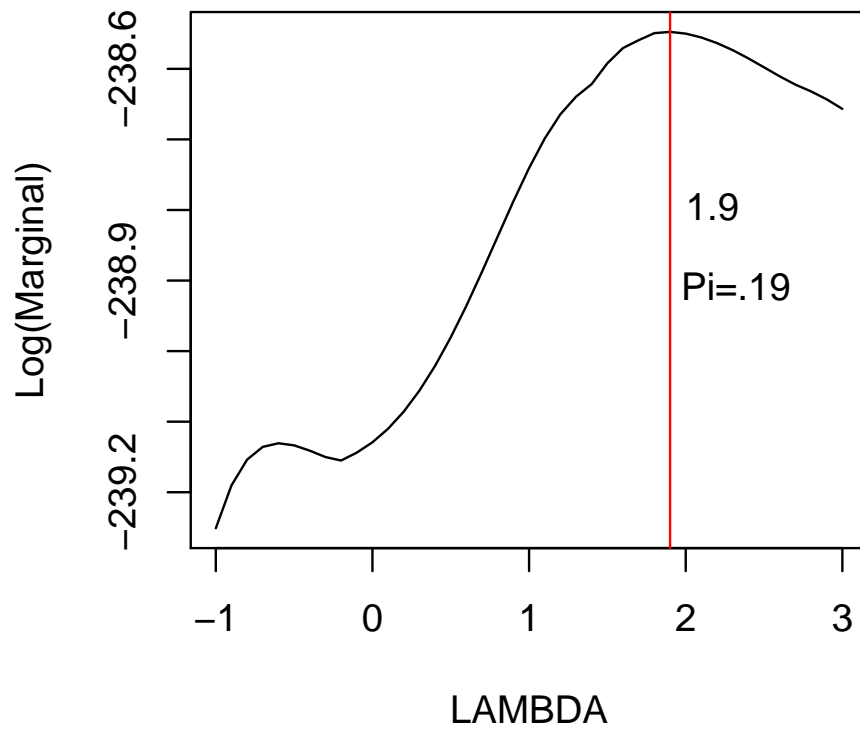


Figure 2.3: Plot of λ vs. $\text{Log}[m(\mathbf{y}|\mathbf{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to the NCAA Dataset. The vertical line represents the location of the global maximum.

Table 2.1: For Case 1, Comparing Average Posterior Probabilities, corresponding to the case 1: $p = 4$, averaged across 1,000 simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (2.19). PoCor represents our proposed modification as in (2.14). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)

Zellner's			PoCor			Intrinsic Prior		
Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected
x_2	.321	37.6	x_1, x_2	.622	68.7	x_1	.468	51.9
x_1	.221	20.4	x_2	.116	10.4	x_2	.410	44.2
x_1, x_2	.105	11.2	x_1, x_2, x_3	.089	3.4	x_1, x_2	.039	.62
x_1, x_2, x_4	.074	5.9	x_1, x_2, x_4	.040	3.1	x_1, x_3	.020	.36
x_1, x_2, x_3	.055	4.4	x_1, x_2, x_3, x_4	.039	2.5	x_3	.020	.24
x_2, x_4	.052	2.8	x_1	.030	2.3	x_1, x_4	.017	1.07
x_2, x_3	.045	2.6	x_1, x_3	.009	1.4	x_2, x_3	.010	.26
x_1, x_3	.035	2.1	x_2, x_3	.003	1.0	x_2, x_4	.008	.20
x_1, x_3, x_4	.024	1.4	x_2, x_3, x_4	.002	.9	x_4	.002	.12
x_1, x_2, x_3, x_4	.019	1.1	x_1, x_3, x_4	.002	.6	x_1, x_2, x_4	.003	.08

Table 2.2: Case 2, Comparing Average Posterior Probabilities, corresponding to the case 2: $p = 12$, averaged over 1,000 Simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (2.19). PoCor represents our proposed modification as in (2.14). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)

Zellner's			PoCor			Intrinsic Prior		
Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected
x_1	.068	28.9	x_1, x_2	.145	46.1	x_1	.039	33.2
x_1, x_2	.051	10.6	x_1, x_2, x_3	.109	12.4	x_2	.016	14.5
x_1, x_3	.025	10	x_1, x_3	.090	7.8	x_1, x_2	.013	9.8
x_1, x_4	.016	5.9	x_2, x_3	.078	5.5	x_1, x_2, x_3	.011	5.4
x_1, x_5	.0133	5.4	x_1	.056	5.1	x_1, x_3	.009	2.3
x_1, x_2, x_5	.013	4.4	x_1, x_2, x_4	.020	4.7	x_1, x_2, x_4	.008	.89
x_1, x_2, x_6	.012	3.5	x_1, x_2, x_3, x_4	.016	3.6	x_1, x_3, x_4	.008	.76
x_2, x_2, x_8	.011	3.5	x_1, x_2, x_5	.010	3.4	x_1, x_2, x_5	.007	.54

Table 2.3: Comparing Posterior Probabilities and average prediction errors for the models of the NCAA Data. The entries in parenthesis are the standard errors obtained by 1000 replications.

Zellner's			PoCor		
Subset	Post Prob	C.V. Pred Err	Subset	Post Prob	C.V. Pred Err
x_2, x_3, x_4, x_5, x_7	.042	54.38 (0.615)	$x_2, x_3, x_4, x_5, x_7, x_{17}$.036	51.53 (0.561)
x_2, x_3, x_4, x_7	.041	55.57 (0.646)	$x_2, x_3, x_4, x_5, x_7, x_{17}, x_{18}$.030	52.38 (0.619)
x_2, x_3, x_4, x_5	.028	56.74 (0.599)	$x_1, x_2, x_3, x_4, x_5, x_7$.028	53.46 (0.608)
$x_2, x_3, x_4, x_5, x_7, x_{18}$.017	55.17 (0.609)	$x_2, x_3, x_4, x_5, x_7, x_{18}$.021	54.08 (0.572)
$x_2, x_3, x_4, x_5, x_7, x_{17}$.015	54.89 (0.623)	x_2, x_3, x_4, x_5, x_7	.018	54.51 (0.568)
$x_1, x_2, x_3, x_4, x_5, x_7$.013	55.64 (0.611)	x_1, x_2, x_3, x_4, x_5	.015	56.13 (0.601)
$x_2, x_3, x_4, x_7, x_8, x_9$.011	56.54 (0.628)	$x_2, x_3, x_4, x_5, x_7, x_{10}$.009	54.75 (0.554)

Table 2.4: Posterior Probabilities and average prediction errors for the models of the NCAA Data using the Intrinsic Priors. The entries in parenthesis are the standard errors obtained by 1000 replications.

Intrinsic Prior		
Subset	Post Prob	C.V. Pred Err
x_2, x_4, x_7	.066	53.97 (0.530)
x_2, x_3, x_4, x_5, x_7	.040	52.94 (0.541)
x_2, x_3, x_4, x_5	.028	54.09 (0.602)
x_2, x_4, x_5, x_7	.022	54.82 (0.576)
x_2, x_3, x_4	.016	58.71 (0.617)
x_2, x_4, x_7, x_{11}	.011	58.67 (0.594)
x_2, x_4, x_{11}	.010	60.08 (0.581)

Chapter 3

Joint Variable Selection of Fixed and Random Effects in Linear Mixed-Effects Model and its Oracle Properties

3.1 Introduction

Linear mixed-effects (LME) models (Laird and Ware, 1982) are a class of statistical models used to describe the relationship between the responses and the covariates, based on grouped data. Some common examples of grouped data are repeated measures data and nested designs. By introducing one or more subject-specific random effects, the LME model allows us the flexibility to model both the means as well as the covariance structure of the grouped data.

Consider a study consisting of m subjects, with response from each subject $i = 1, 2, \dots, m$ each measured n_i times, and let $N = \sum_{i=1}^m n_i$ be the total number of observations. Let \mathbf{y}_i be an $n_i \times 1$ vector of the response variable for subject i . Let \mathbf{X}_i be the $n_i \times p$ design matrix of explanatory variables, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ be the vector of unknown regression parameters which are assumed to be fixed. Let

$\mathbf{b}_i^* = (b_{i1}^*, \dots, b_{iq}^*)'$ be a $q \times 1$ vector of unknown subject-specific random effects with $\mathbf{b}_i^* \sim N(0, \sigma^2 \mathbf{\Psi})$, and are assumed to be independently distributed across subjects. Denote \mathbf{Z}_i as the $n_i \times q$ design matrix corresponding to the random effects. Often one sets $\mathbf{Z}_i = \mathbf{X}_i$, but it is not necessary. Then, a general class of LME models can be written as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i^* + \boldsymbol{\epsilon}_i, \quad (3.1)$$

where the random errors $\boldsymbol{\epsilon}_i$'s are independently distributed $N(0, \sigma^2 \mathbf{I}_{n_i})$ and are assumed to be independent of the \mathbf{b}_i^* 's.

Greenland (2000) argued that models with random coefficients in repeated measures analysis offer a more scientifically defensible framework than just using fixed effects models, as some models need to be complex to capture the uncertainty about the relationship. For example, consider a recent study of the association between total nitrate concentration in the atmosphere and a set of measured predictors using the U.S EPA CASTnet data (Lee and Ghosh, 2008). The dataset consists of multiple sites with repeated measurements of nitrate concentration along with a set of 16 potential covariates on each site. To analyze this data one could use a LME model where additional random components are needed due to capture the variations across sites. Lange and Laird (1989) showed that in these situations, underfitting the covariance structure for the random effects would lead to bias in the estimated variance of the fixed effects. On the other hand, including unnecessary random effects could also lead to a near singular random effect covariance matrix. Hence, it is of practical importance to identify not only the fixed effects, but also the subset of q random effects which contribute to the heterogeneity among the groups. The main goal of this paper is to simultaneously identify the subset of $p^* \leq p$ and $q^* \leq q$ important predictors that correspond to the fixed and the random components in the LME model, respectively.

The motivation behind variable selection is to identify a model which is both meaningful i.e., identifies the important set of predictors, as well as easy to interpret (sparse representation) while ensuring high prediction accuracy. In the statistical

literature, the problem of selecting variables/predictors has received considerable attention over the years, and a large number of methods have been proposed where the main goal is to identify the predictors with non-zero coefficients. See for example, Miller (2002) for a comprehensive review. Traditional methods such as forward selection and backward elimination have been used to select important covariates, but such methods are highly unstable because of its inherent discreteness (Breiman, 1996). More recently, penalized regression, or regression shrinkage, has emerged as a highly-successful method to tackle this problem, for example see, Tibshirani (1996), Fan and Li (2001), Efron, Hastie, Johnstone and Tibshirani (2004), Zou and Hastie (2005), Zou (2006), Bondell and Reich (2008). The Bayesian approaches to selecting variables is typically based on the idea of choosing a model with high posterior probability (either jointly, or marginally for each predictor). Some examples include, Mitchell and Beauchamp (1988), Geweke (1996), George and McCulloch (1996, 1997) and Liang et. al. (2008).

Although the variable selection problem has received much attention, the selection of random effects together with the fixed effects in the LME model has received little attention. An important problem in linear mixed-effects models is how to simultaneously select the important fixed and random effect components, as changing one of the two types of effects greatly affects the other. Few procedures have been proposed in the literature. Model selection criteria such as AIC (Akaike, 1973), BIC (Schwartz, 1978), GIC (Rao and Wu, 1989) have been used to compare a list of plausible models under consideration, but the number of possible candidate models increases exponentially as the number of predictors increase, making the process computationally demanding. For instance given the LME model in (3.1) the total number of possible sub-models is, 2^{p+q} .

To slightly reduce the computational demand, Niu and Pu (2007) proposed the EGIC (Extended GIC), while Wolfinger (1993) and Diggle, Liang and Zeger (1994) proposed the Restricted Information Criteria, where selection is first performed on either the mean or the covariance structure while fixing the other at the full model, although ideally one would like to select them simultaneously. This then results in

the number of possible sub-models evaluated to be equal to the sum of the mean and the random effects structures in consideration, that is, $2^p + 2^q$. A more detailed description of these two methods are presented in Section 5.

Forward or Backward selection techniques can also be used to avoid enumerating all possible models (for example, see Morell, Pearson and Brant, 1997) however the discrete nature of these procedures makes them extremely unstable. Non-standard methods (Self and Liang, 1987; Stram and Li, 1994; Lin, 1997) have been used to test for non-zero variance components under the null hypothesis that such a component is zero. More recently, Jiang, Rao, Gu and Nguyen (2006) proposed a ‘fence’ method to select predictors in a general mixed model setting where a “statistical fence” is constructed to eliminate incorrect models. Although these methods may avoid the need to search through the entire model space to find the optimal model, it could still be computationally intensive when the total number of predictors are large.

To identify the important random effects in an LME model, Chen and Dunson (2003) proposed a hierarchical Bayesian method, by choosing a mixture of priors with point mass at zero for the random effect variances, which is then extended to logistic mixed effects model (Kinney and Dunson, 2006). Another Bayesian approach to model selection in the linear mixed effects model was proposed by Weiss, Wang and Ibrahim (1997), and a shrinkage selection method using the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001) was proposed by Lan (2006), but both methods only identify the important fixed effects, keeping the set of random effects fixed.

A difficulty in defining a shrinkage approach to random effects selection is that an entire row and column of Ψ must be eliminated to successfully remove a random effect. This leads to complications in how to perform the shrinkage appropriately. In this article we propose a new method for simultaneously selecting the fixed and the random effects parameters. As opposed to the previous approaches, the selection is done for both types of effects in a combined penalized procedure. Our proposed method is based on a re-parametrization of the LME model obtained by a modified Cholesky decomposition of Ψ (Chen and Dunson, 2003). This modified factorization

aids us in the selection of the random effects by dropping out the random effects terms which have zero variance.

The SCAD (Fan and Li, 2001) and the adaptive LASSO (Zou, 2006) showed that asymptotically these penalized estimators perform as well as the 'Oracle' estimators which knows the true model beforehand. Motivated by the oracle properties of the adaptive LASSO estimates we use an adaptive penalty on the reparameterized model that simultaneously selects the fixed and the random effects, based on a single tuning parameter. To obtain the penalized likelihood estimates of our parameters we use a constrained version of the EM algorithm (Dempster, Laird and Rubin, 1977; Laird and Ware, 1982; Laird, Lange and Stram, 1987).

The remainder of the paper is structured as follows. In Section 2, we describe the re-parameterized linear mixed models and its properties. Section 3, describes our method which automatically selects the important variables for the fixed as well as the random effects. In Section 4 we show that our penalized estimators have desirable theoretical properties such as the Oracle property. We illustrate the performance of our method with a simulation study and the CASTnet data in Section 5. Finally, in Section 6 we conclude with a discussion. All proofs are given in Appendix A.

3.2 The Reparameterized Linear Mixed Effects Model

The Cholesky decomposition has been extensively used in efficiently estimating the covariance matrix of the random effects within a LME model framework (for example, see Lindstrom and Bates, 1988; Smith and Kohn, 2002). Pinhero and Bates (1996) pointed out that this decomposition is convenient numerically as it involves unconstrained parameters for numerical optimization. However the parameters in the Cholesky decomposition does not allow for the direct elimination of random effects as the covariance matrix is a function of all of these parameters. In this section, we adopt a modified Cholesky decomposition as in Chen and Dunson (2003), where we

factorize the symmetric covariance matrix, Ψ , of the random effects as

$$\Psi = \mathbf{D}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}, \quad (3.2)$$

where $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_q)$ is a diagonal matrix with its elements proportional to the standard deviations of the random effects, and $\mathbf{\Gamma}$, whose $(l, r)^{th}$ element is denoted by γ_{lr} , is a $q \times q$ lower triangular matrix with 1's on the diagonal and the off-diagonal relates to the correlation among the random effects. This decomposition in terms of \mathbf{D} and $\mathbf{\Gamma}$ is unique, and leads to a non-negative definite matrix Ψ .

Given the decomposition in (3.2), the reparameterized LME model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{D}\mathbf{\Gamma}\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (3.3)$$

where \mathbf{y}_i has been centered and the predictors have been standardized so that, $\mathbf{X}_i'\mathbf{X}_i$ and $\mathbf{Z}_i'\mathbf{Z}_i$ represents the correlation matrices, and $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})'$ for $l = 1, \dots, q$, is a $q \times 1$ vector of independent $N(0, \sigma^2\mathbf{I}_{n_i})$. The covariance matrix of \mathbf{b}_i^* , is now expressed as a function of $\mathbf{d} = (d_1, d_2, \dots, d_q)'$ with each element d_l proportional to the standard deviation of the b_{il}^* random effect, and the $q(q-1)/2$ free elements of $\mathbf{\Gamma}$, denoted by the vector $\boldsymbol{\gamma} = (\gamma_{lr} : l = 1, \dots, q : r = l+1, \dots, q)'$. We denote $\boldsymbol{\phi} = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')$, a $k \times 1$ combined vector of unknown parameters, where $k = p + \frac{q(q+1)}{2}$.

Random effects selection can be difficult due to the fact that removal of a random effect corresponds to setting an entire row and column of Ψ to zero. However, with this convenient decomposition (3.2), setting $d_l = 0$ is equivalent to setting all the elements in the l^{th} column and l^{th} row of Ψ to zero and creating a new sub-matrix by completely removing the corresponding row and column. Hence under this reparameterization a single parameter controls the inclusion/exclusion of the random effects allowing for a more straight forward procedure. This is the key initial step to defining a simultaneous selection approach.

3.2.1 The Likelihood

For the reparameterized linear model (3.3) assume that the data $\{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{y}_i\}$ are collected independently. Conditioning on \mathbf{X}_i and \mathbf{Z}_i the marginal distribution of \mathbf{y}_i obtained by integrating out the random effects \mathbf{b}_i follows a normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$, and variance given by

$$\mathbf{V}_i = \sigma^2(\mathbf{Z}_i\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}\mathbf{Z}_i' + \mathbf{I}_{n_i}). \quad (3.4)$$

Dropping out the constant terms, the log-likelihood function based on the LME model (3.3) is given by

$$L(\boldsymbol{\phi}) = -\frac{1}{2}\log|\tilde{\mathbf{V}}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\tilde{\mathbf{V}})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (3.5)$$

where $\tilde{\mathbf{V}} = \text{Diag}(\mathbf{V}_1, \dots, \mathbf{V}_m)$ a block diagonal matrix of \mathbf{V}_i 's, and $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$, $\mathbf{X} = [\mathbf{X}'_1, \dots, \mathbf{X}'_m]'$ are the stacked vectors and matrices of \mathbf{y}_i and \mathbf{X}_i , respectively.

A large literature is available on the estimation of the parameters in the linear mixed models (e.g., Patterson and Thompson, 1971; Harville, 1974, 1977; Laird and Ware, 1982; Lindstrom and Bates, 1988). In this paper, we use penalized/shrinkage techniques together with the EM algorithm (Dempster, Laird and Rubin, 1977) to obtain the penalized estimates of the fixed effects as well as the variance components.

By treating $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_m)'$ as unobserved, and dropping out the constants we can now write the joint log-likelihood function with respect to the complete data (\mathbf{y}, \mathbf{b}) given $\boldsymbol{\phi}$ as

$$L_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b}) = -\frac{1}{2\sigma^2} \left\{ \|\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\boldsymbol{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2 + \mathbf{b}'\mathbf{b} \right\}, \quad (3.6)$$

where $\tilde{\mathbf{D}}$, $\tilde{\mathbf{\Gamma}}$ and \mathbf{Z} represent block diagonal matrices given by

$$\tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D} & 0 & \dots & 0 \\ 0 & \mathbf{D} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{D} \end{bmatrix}, \tilde{\mathbf{\Gamma}} = \begin{bmatrix} \mathbf{\Gamma} & 0 & \dots & 0 \\ 0 & \mathbf{\Gamma} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{\Gamma} \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_m \end{bmatrix}. \quad (3.7)$$

We now maximize the conditional expectation of (3.6) along with a penalty function, with respect to $\boldsymbol{\beta}$ and \mathbf{d} , to decide whether to include or exclude a predictor in the model. However, dropping out the terms which do not involve either $\boldsymbol{\beta}$ or \mathbf{d} is then equivalent to minimizing the conditional expectation of $\|\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\mathbf{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2$ plus the penalty term, as described in the following section.

3.3 Penalized Selection and Estimation for the Reparameterized LME model

3.3.1 The Shrinkage Penalty

Recently, Zou (2006) proposed the Adaptive LASSO where adaptive weights are used to penalizing the different regression coefficients in the l_1 penalty. That is, we wish to have a large amount of shrinkage applied to the zero-coefficients while smaller amounts are used for the non-zero ones which then results in an estimator with improved efficiency and selection properties. The Adaptive LASSO estimate for the linear regression model is defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^p \bar{w}_j |\beta_j|, \quad (3.8)$$

where λ_n is a non-negative regularization/tuning parameter, \bar{w}_j are the adaptive weights, typically $\bar{w}_j = 1/|\bar{\beta}_j|$, with $\bar{\beta}_j$ the ordinary least squares estimate. So intuitively, as λ_n increases, the coefficients are continuously shrunk towards zero and,

some coefficients are exactly shrunk to zero for a sufficiently large λ_n . Adopting this adaptive penalty, we propose our method of selecting the important fixed and random effect components.

Given the LME model (3.3) and the complete data log-likelihood (3.6), we can define our penalized criterion under the l_1 penalty with the adaptive weights, jointly for the fixed and random effects as

$$\mathbf{Q}_c(\phi|\mathbf{y}, \mathbf{b}) = \|\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\mathbf{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_n \left\{ \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|} + \sum_{j=1}^q \frac{|d_j|}{|\tilde{d}_j|} \right\}. \quad (3.9)$$

Here $\tilde{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$, and $\tilde{\mathbf{d}}$ is obtained by decomposition of the estimated covariance matrix obtained by REML procedure.

Rearranging the terms, the equation given in (3.9) can also be written as

$$\mathbf{Q}_c(\phi|\mathbf{y}, \mathbf{b}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\text{Diag}(\tilde{\mathbf{\Gamma}}\mathbf{b})\mathbf{d}\|^2 + \lambda_n \left\{ \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|} + \sum_{j=1}^q \frac{|d_j|}{|\tilde{d}_j|} \right\}. \quad (3.10)$$

The re-expressed form in (3.10) will be useful in obtaining the estimates for \mathbf{d} . By weighting the coefficients by a \sqrt{n} consistent estimator for the penalty term, a single tuning parameter is sufficient to perform variable selection for both the fixed and the random effects. One could add a penalty term to the elements of $\boldsymbol{\gamma}$ to additionally shrink the off diagonal terms to zero if desired, but our main motivation in this paper is to identify the important random effects.

3.3.2 Computation and Tuning

The Constrained EM Algorithm

In an incomplete-data setting, Dempster, Laird and Rubin (1977) proposed a very general computing algorithm called Expectation-Maximization (EM). Laird and Ware (1982) and Laird, Lange and Stram (1987) used the EM algorithm in the context of repeated measures data, where the complete data consists of the observed data \mathbf{y}_i 's plus the unobserved random parameters (missing data) specified in the model. Here

we adopt the EM algorithm (Laird and Ware, 1982, Laird, Lange and Stram, 1987), in that, we first compute the conditional expectation of $\mathbf{Q}_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b})$ assuming the random effects are unobserved (E-step). Then we minimize the conditional expectation of $\mathbf{Q}_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b})$ to obtain the updated penalized likelihood estimates of our parameters (M-step). This process is then repeated iteratively until convergence.

Given (3.6), the conditional distribution of \mathbf{b} given $\boldsymbol{\phi}$ and \mathbf{y} is, $\mathbf{b}|\mathbf{y}, \boldsymbol{\phi} \sim N(\hat{\mathbf{b}}, \mathbf{U})$ where the conditional mean and variance are given by,

$$\begin{aligned} \hat{\mathbf{b}}^{(\omega)} &= (\tilde{\boldsymbol{\Gamma}}^{(\omega)} \tilde{\mathbf{D}}^{(\omega)} \mathbf{Z}' \mathbf{Z} \tilde{\mathbf{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)} + \mathbf{I})^{-1} (\mathbf{Z} \tilde{\mathbf{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)})' (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(\omega)}) \\ \text{and } \mathbf{U}^{(\omega)} &= \sigma^2 (\tilde{\boldsymbol{\Gamma}}^{(\omega)} \tilde{\mathbf{D}}^{(\omega)} \mathbf{Z}' \mathbf{Z} \tilde{\mathbf{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)} + \mathbf{I})^{-1}, \end{aligned} \quad (3.11)$$

respectively, and where ω indexes the iterations and $\omega = 0$ refers to the initial estimates, which in our case are chosen to be the ML estimates for $\boldsymbol{\beta}$ and the decomposed estimated variance-covariance matrix obtained from REML for \mathbf{d} and $\boldsymbol{\gamma}$.

Let $\boldsymbol{\phi}^{(\omega)}$ be the estimate of $\boldsymbol{\phi}$ at the ω^{th} iteration. We first compute the conditional expectation $\left(\mathbb{E}_{\mathbf{b}|\mathbf{y}, \boldsymbol{\phi}^{(\omega)}} \right)$ of $\mathbf{Q}_c(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b})$ as

$$\mathbf{g}(\boldsymbol{\phi}|\boldsymbol{\phi}^{(\omega)}) = \mathbb{E}_{\mathbf{b}|\mathbf{y}, \boldsymbol{\phi}^{(\omega)}} \left\{ \|\mathbf{y} - \mathbf{Z} \tilde{\mathbf{D}} \tilde{\boldsymbol{\Gamma}} \mathbf{b} - \mathbf{X} \boldsymbol{\beta}\|^2 \right\} + \lambda_n \left\{ \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|} + \sum_{j=1}^q \frac{|d_j|}{|\tilde{d}_j|} \right\}. \quad (3.12)$$

For the M-step, we minimize the objective function $\left(\mathbf{g}(\boldsymbol{\phi}|\boldsymbol{\phi}^{(\omega)}) \right)$ with respect to $\boldsymbol{\beta}$ and the full covariance matrix, which is a function of \mathbf{d} and $\boldsymbol{\gamma}$, iteratively, to obtain the updated estimates.

We first fix the covariance matrix by fixing \mathbf{d} and $\boldsymbol{\gamma}$ by its most recent update. Given the covariance matrix, and omitting the terms that do not involve $\boldsymbol{\beta}$, the conditional expectation of (3.9) is given as

$$\mathbf{g}(\boldsymbol{\beta}|\boldsymbol{\phi}^{(\omega)}) = \boldsymbol{\beta}' [\mathbf{X}' \mathbf{X}] \boldsymbol{\beta} - 2 \left[(\mathbf{y} - \mathbf{Z} \tilde{\mathbf{D}} \tilde{\boldsymbol{\Gamma}} \hat{\mathbf{b}}^{(\omega)})' \mathbf{X} \right] \boldsymbol{\beta} + \lambda_n \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|}, \quad (3.13)$$

where $\hat{\mathbf{b}}^{(\omega)}$ is as given in (3.11). We now minimize the above objective function (3.13)

to obtain the updated estimates for β (details are given in the next section). Given the updated value for β , the updated covariance matrix is obtained by minimizing $g(\phi|\mathbf{y}, \mathbf{b})$ with respect to \mathbf{d} and γ iteratively.

Fixing β and γ at its most recent update, we now obtain the expression for the objective function for \mathbf{d} , the vector that relates to the standard deviation of the random effects. Expanding the first term in (3.10), we have

$$\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\text{Diag}(\tilde{\Gamma}\mathbf{b})\mathbf{d}\|^2 = \mathbf{d}'[\text{Diag}(\tilde{\Gamma}\mathbf{b})\mathbf{W}\text{Diag}(\tilde{\Gamma}\mathbf{b})]\mathbf{d} - 2\left[(\mathbf{y} - \mathbf{X}\beta)' \mathbf{Z}\text{diag}(\tilde{\Gamma}\mathbf{b})\right]\mathbf{d}, \quad (3.14)$$

where $\mathbf{W} = \mathbf{Z}'\mathbf{Z}$ denotes a symmetric block diagonal matrix. After some matrix manipulation, the first term in (3.14) can be written as

$$\mathbf{d}'[\text{Diag}(\tilde{\Gamma}\mathbf{b})\mathbf{W}\text{Diag}(\tilde{\Gamma}\mathbf{b})]\mathbf{d} = \mathbf{d}'[\mathbf{W} \bullet \Gamma\text{Diag}(\mathbf{b})\mathbf{1}\mathbf{1}'\text{Diag}(\mathbf{b})\Gamma']\mathbf{d}, \quad (3.15)$$

where \bullet represents the Hadamard (element by element) product operator, and $\mathbf{1}$ represents an $mq \times 1$ vector of ones. Computing the outer product $\text{Diag}(\mathbf{b})\mathbf{1}\mathbf{1}'\text{Diag}(\mathbf{b})$, the expression given in (3.15) further simplifies to $\mathbf{d}'[\mathbf{W} \bullet \Gamma\mathbf{b}\mathbf{b}'\Gamma']\mathbf{d}$.

Using this simplification and omitting the terms that do not involve \mathbf{d} , the conditional expectation of (3.10) can be expressed in a quadratic form in \mathbf{d} is given as

$$g(\mathbf{d}|\phi^{(\omega)}) = \mathbf{d}'\left[\mathbf{W} \bullet \tilde{\Gamma}\hat{\mathbf{G}}^{(\omega)}\tilde{\Gamma}'\right]\mathbf{d} - 2\left[(\mathbf{y} - \mathbf{X}\beta)' \mathbf{Z}\text{diag}(\tilde{\Gamma}\hat{\mathbf{b}}^{(\omega)})\right]\mathbf{d} + \lambda_n \sum_{l=1}^q \frac{|d_l|}{|\bar{d}_l|}, \quad (3.16)$$

where $\hat{\mathbf{G}}^{(\omega)} = E(\mathbf{b}\mathbf{b}') = \mathbf{U}^{(\omega)} + \hat{\mathbf{b}}^{(\omega)}\hat{\mathbf{b}}'^{(\omega)}$, and $\mathbf{U}^{(\omega)}$ and $\hat{\mathbf{b}}^{(\omega)}$ are as given in (3.11). For a fixed β and γ , we now minimize the objective function (3.16) to obtain the updated estimates for \mathbf{d} .

We can then obtain a closed form expression for the estimate of γ , the vector that relates to the correlation between the random effect parameters. If $\gamma = \mathbf{0}$, then the random effects are mutually independent with the random effect covariance matrix Ψ being reduced to a simple diagonal form. Furthermore, if $\mathbf{d}_l = 0$ then $\gamma_{lr} = 0$ for

$r = l+1, \dots, q$. Hence, the elements of $\boldsymbol{\Gamma}$ and \mathbf{d} are functionally related. Fixing $\boldsymbol{\beta}$ and \mathbf{d} at its most recent update, we first rewrite $\|\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\boldsymbol{\Gamma}}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2$ in a quadratic form for $\boldsymbol{\gamma}$ and then compute its conditional expectation. After some matrix manipulation, and omitting terms not involving $\boldsymbol{\gamma}$, we have

$$g(\boldsymbol{\gamma}|\hat{\boldsymbol{\phi}}^{(\omega)}) = \boldsymbol{\gamma}'\mathbf{P}^{(\omega)}\boldsymbol{\gamma} - 2 \left[\{(\mathbf{Z}\tilde{\mathbf{D}}\hat{\mathbf{b}}^{(\omega)})' - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\} \mathbf{R}^{(\omega)} \right] \boldsymbol{\gamma}, \quad (3.17)$$

where $\mathbf{P}^{(\omega)} = E_{\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\phi}}^{(\omega)}}[\mathbf{A}'\mathbf{A}]$ and $\mathbf{R}^{(\omega)} = E_{\mathbf{b}|\mathbf{y},\hat{\boldsymbol{\phi}}^{(\omega)}}(\mathbf{A})$. Where $\mathbf{A} = [\mathbf{A}'_1, \dots, \mathbf{A}'_m]'$ represents a stacked matrix of \mathbf{A}_i , with each \mathbf{A}_i an $n_i \times q(q-1)/2$ matrix, whose elements in each row are given by $A_{ij} = (b_{il}d_r z_{ijr} : l = 1, \dots, q, r = l+1, \dots, q)$. The minimizer of (3.17) is then given by

$$\boldsymbol{\gamma}_* = \left(\mathbf{P}^{(\omega)}\right)^{-1} \left[\{(\mathbf{Z}\tilde{\mathbf{D}}\hat{\mathbf{b}})' - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\} \mathbf{R}^{(\omega)} \right]'. \quad (3.18)$$

The optimization problem is now solved by minimizing the quadratic forms (3.13) and (3.16), along with explicit solution (3.18) iteratively to complete the M-step. The final penalized likelihood estimates, $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}, \hat{\mathbf{d}}, \hat{\boldsymbol{\gamma}})$ are obtained by successive EM steps.

Computation of the M-step

Recently, Efron, Hastie, Johnstone and Tibshirani (2004) proposed the LARS (Least Angle Regression) algorithm, and showed that it can be used to obtain the entire solution path for LASSO estimates, while being computationally efficient. Zou (2006) showed that with minor changes to the design matrix, the LARS algorithm can be implemented to obtain the estimates for the regression coefficients under the adaptive LASSO penalty. Although we can use the LARS algorithm to minimize the penalized quadratic form in our M-step via a pseudo “design matrix”, it is not as advantageous to obtain the entire solution path here. This is due to the fact that the “design matrix” changes with every iterative step of the EM algorithm. Hence, we propose the use of a standard quadratic programming technique to obtain the

penalized likelihood estimates for our parameters at each iterative step.

Given $\phi = \phi^{(\omega)}$ and a tuning parameter λ_n , we write $\beta = \beta^+ - \beta^-$ with both β^+ and β^- being non-negative, and only one is non-zero, and $|\beta| = \beta^+ + \beta^-$, the optimization problem for β given in (3.13) is equivalent to

$$\begin{aligned} & \text{minimize} \quad \begin{bmatrix} \beta^+ \\ \beta^- \end{bmatrix}' \begin{bmatrix} \mathbf{X}'\mathbf{X} & -\mathbf{X}'\mathbf{X} \\ -\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{X} \end{bmatrix} \begin{bmatrix} \beta^+ \\ \beta^- \end{bmatrix} \\ & -2 \left(\begin{bmatrix} (\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\Gamma}\hat{\mathbf{b}}^{(\omega)})'\mathbf{X} \\ -(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{D}}\tilde{\Gamma}\hat{\mathbf{b}}^{(\omega)})'\mathbf{X} \end{bmatrix}' + \lambda_n \left[\frac{1}{\beta_1}, \dots, \frac{1}{\beta_p}, \frac{1}{\beta_1}, \dots, \frac{1}{\beta_p} \right] \right) \begin{bmatrix} \beta^+ \\ \beta^- \end{bmatrix} \\ & \text{subject to} \\ & \beta^+ \geq 0, \beta^- \geq 0. \end{aligned} \tag{3.19}$$

The minimization of β with respect to the expanded parameter (β^+, β^-) is now a quadratic programming problem with $2p$ total parameters and $2p$ total linear constraints. This expansion of the parameters is not necessary with respect to \mathbf{d} as it can be viewed as a constrained maximization problem where we restrict $\mathbf{d} \geq 0$. Hence it is a direct quadratic programming problem.

Choice of tuning parameter

The choice of the tuning parameter λ_n can be accomplished via minimizing any of the standard criteria such as AIC, BIC, GIC, Generalized Cross-Validation (GCV), or via k-fold Cross-Validation. It is known that under general conditions, BIC is consistent for model selection if the true model belongs to the class of models considered, while although AIC is minimax optimal, it is not consistent for selection (Shao, 1997; Yang, 2005). BIC is given by

$$BIC_{\lambda_n} = -2L(\hat{\phi}) + \log(N) \times (df_{\lambda_n}) \tag{3.20}$$

where $L(\hat{\phi})$ is the maximized value of $L(\phi)$ given in (3.5) for the estimated model. We take the degrees of freedom, df_{λ_n} , as the number of non-zero coefficients in $\hat{\phi}$.

Since it has been shown for the linear model that this is an unbiased estimate of the degrees of freedom (Zou, Hastie and Tibshirani, 2007) we adopt this for this setting as well. Note that, we obtain an estimate for σ^2 by maximizing the log-likelihood after replacing ϕ by $\hat{\phi}$, i.e. $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})' \left(\mathbf{I} + \mathbf{Z}(\hat{\mathbf{D}}\hat{\Gamma}\hat{\mathbf{D}})\mathbf{Z}' \right)^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) / N$.

3.4 Asymptotic Properties

In this section we shall study the asymptotic properties of our estimator $\hat{\phi}$. Consider again $\phi = (\beta, \mathbf{d}, \gamma)'$ a $k \times 1$ vector and let $\bar{\phi}$ denote an initial \sqrt{n} consistent estimator of ϕ . Let $\mathbf{Q}(\phi)$ denote the penalized log-likelihood function with $L(\phi)$ is as given in (3.5), then

$$\mathbf{Q}(\phi) = L(\phi) - \lambda_n \sum_{j=1}^k \bar{w}_j (|\phi_j|)$$

$$\text{where, } \bar{w}_j = \begin{cases} 0, & \text{for, } \phi_j \in \gamma \\ 1/\bar{\phi}_j, & \text{Otherwise} \end{cases} . \quad (3.21)$$

Denote the true value of ϕ as

$$\phi_0 = (\phi_{10}, \dots, \phi_{k0})' = (\phi'_{10}, \phi'_{20})' \quad (3.22)$$

where $\phi_{10} = (\beta_{10}, \mathbf{d}_{10}, \gamma_{10})'$ is an $s \times 1$ vector whose components are non-zero and let ϕ_{20} be the $(k-s)$ remaining components of ϕ_0 , so that $\phi_{20} = 0$, we can also decompose ϕ itself as $\phi = (\phi'_1, \phi'_2)'$. We now show that the penalized likelihood estimator $\hat{\phi}_1$ of the true non-zero coefficients is \sqrt{n} consistent and asymptotically normal with covariance matrix $I^{-1}(\phi_{10})$, where $I(\phi_{10})$ is the Fisher information matrix knowing that $\phi_{20} = 0$. Further, we show that our method is consistent in selection, in that, the true zero coefficients are estimated as zero with probability tending to 1. Hence, our penalized likelihood estimator satisfies the ‘Oracle’ property (Fan and Li, 2001) and asymptotically performs as well as the true model structure was known before hand. We shall next state our theorems, while the proofs and regularity conditions

are given in Appendix A.

For the penalized log-likelihood given in (3.21), let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, 0)$, that is fixing $\boldsymbol{\phi}_2 = 0$. Let $L(\boldsymbol{\phi}_1)$, $\mathbf{Q}(\boldsymbol{\phi}_1)$ denote the log-likelihood and the penalized log-likelihood of the first s components of $\boldsymbol{\phi}$ given by

$$\begin{aligned} L(\boldsymbol{\phi}_1) &\equiv L \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix} \right\} = -\frac{1}{2} \log |\tilde{\mathbf{V}}_{(1)}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}_{(1)} \boldsymbol{\beta}_1)' (\tilde{\mathbf{V}}_{(1)})^{-1} (\mathbf{y} - \mathbf{X}_{(1)} \boldsymbol{\beta}_1), \\ \mathbf{Q}(\boldsymbol{\phi}_1) &\equiv \mathbf{Q} \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix} \right\} = L(\boldsymbol{\phi}_1) - \lambda_n \sum_{j=1}^s \bar{w}_j (|\phi_j|), \end{aligned} \quad (3.23)$$

where $\tilde{\mathbf{V}}_{(1)} = \mathbf{Z}_{(1)} \{ \tilde{\mathbf{D}}_1 \tilde{\boldsymbol{\Gamma}}_1, \tilde{\boldsymbol{\Gamma}}_1' \tilde{\mathbf{D}}_1 \} \mathbf{Z}_{(1)}' + \mathbf{I}$, is the block diagonal matrix corresponding to the non-zero components $(\mathbf{d}_1, \boldsymbol{\gamma}_1)$ and $\mathbf{X}_{(1)}$ and $\mathbf{Z}_{(1)}$ are the corresponding design matrices.

Theorem 1. *Let $\boldsymbol{\phi} = (\boldsymbol{\phi}'_1, \mathbf{0}')'$, and the observations follow the LME model (3.3) satisfying conditions i and ii given in Appendix A. If $\lambda_n/\sqrt{n} \rightarrow 0$, then there exists a local maximizer $\hat{\boldsymbol{\phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_1 \\ \mathbf{0} \end{pmatrix}$ of $\mathbf{Q} \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \mathbf{0} \end{pmatrix} \right\}$ such that $\hat{\boldsymbol{\phi}}_1$ is \sqrt{n} consistent for $\boldsymbol{\phi}_{10}$.*

Theorem 2. *Let the observations follow the LME model (3.3) satisfying conditions i and ii given in Appendix A. If $\lambda_n \rightarrow \infty$, then with probability tending to 1 for any given $\boldsymbol{\phi}_1$ satisfying $\|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_{10}\| \leq Mn^{-1/2}$ and some constant $M > 0$,*

$$\mathbf{Q} \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix} \right\} = \max_{\|\boldsymbol{\phi}_2\| \leq Mn^{-1/2}} \mathbf{Q} \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \right\}. \quad (3.24)$$

Remark 1. From Theorem 1 we see that we are able to get into a \sqrt{n} neighborhood, while Theorem 2 shows that, with probability tending to 1, there exists a local maximizer in that neighborhood with $\hat{\boldsymbol{\phi}}_2 = \mathbf{0}$. Hence, combining the two, we see that our penalized likelihood estimator can consistently identify the true model.

Theorem 3. *Let the observations follow the LME model (3.3) satisfying conditions*

i and ii given in Appendix A. Then as $\lambda_n \rightarrow \infty$ and $\lambda_n/\sqrt{n} \rightarrow 0$, we have

$$\sqrt{n}I(\phi_{10})\{(\hat{\phi}_1 - \phi_{10}) + \left(\frac{\lambda_n}{n}\right)I^{-1}(\phi_{10})\mathbf{h}(\phi_{10})\} \rightarrow_d N\{0, I(\phi_{10})\} \quad (3.25)$$

where $\mathbf{h}(\phi_{10}) = (\bar{w}_1 \text{sgn}(\phi_{10}), \dots, \bar{w}_s \text{sgn}(\phi_{s0}))'$ an $s \times 1$ vector.

Remark 2. From Theorem 2 and 3 as $\lambda_n \rightarrow \infty$ and $\lambda_n/\sqrt{n} \rightarrow 0$, we can say that our penalized estimator enjoys the oracle property in that asymptotically it performs as well as the oracle estimators, which in our case are the maximum likelihood estimates of ϕ_1 knowing $\phi_2 = 0$. In particular, to first order, $\sqrt{n}(\hat{\phi}_1 - \phi_{10}) \rightarrow N(\mathbf{0}, I^{-1}(\phi_{10}))$.

3.5 Simulation Study and Real Data Analysis

In order to avoid complete enumeration of all possible (2^{p+q}) models in a LME framework, Wolfinger (1993) and Diggle, Liang and Zeger (1994) recommended the Restricted Information Criterion, in that, by using the most complex mean structure, selection is first performed on the variance-covariance structure by computing the AIC and/or BIC, obtained via the restricted (REML) log-likelihood. Given the best covariance structure, selection is then performed on the fixed effects. Alternatively, Pu and Niu (2006) proposed the EGIC (Extended GIC) for the LME model, where using the BIC, selection is first performed on the fixed effects by including all of the random effects into the model. Once the fixed effects model is chosen, selection is then performed on the random effects.

In this section, for the simulation study, we compare our proposed method to the Restricted Information Criterion (which we denote by REML.IC) as well as the EGIC where selection of the fixed and the random effects are done by enumerating ($2^p + 2^q$) set of models. Given the selected random effects model by using the REML.IC, further comparisons are also shown for the cases where the selection on the fixed effects are performed using the LASSO, adaptive LASSO, and the stepwise procedure which allows movement in either direction, in that, starting of in a backward approach by choosing the least significant variable to drop and then re-considering all dropped

variables (except the most recently dropped) for re-introduction into the model.

For the real data example, we have $p = q = 16$, so we select the random effects by using the backward elimination procedure where the inclusion/exclusion of a random effect is based on the results of a Likelihood Ratio Test, due to the computational difficulty in enumerating all possible models.

3.5.1 Simulation Study

We shall evaluate the performance of our method and other competing methods by comparing them to the ‘Oracle’ model which knows beforehand the true underlying model. We show that our penalized likelihood estimators outperforms the other methods with respect to estimation as well as selection. Its performance is close to that of the ‘Oracle’ estimator which is the ML/REML estimates of ϕ_1 knowing $\phi_2 = 0$.

Three scenarios are considered in the simulation. In each example, 200 datasets were simulated from a multivariate normal density

$$\mathbf{y}_i \sim N\{\mathbf{X}_i\boldsymbol{\beta}, \sigma^2(\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i' + \mathbf{I}_{n_i})\}. \quad (3.26)$$

The three scenarios are given by:

1. Example 1: We consider $m = 30$ subjects and $n_i = 5$ observations per subject, for a particular case where $p = 9$ and $q = 4$. We consider the true model

$$y_{ij} = b_{i1} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{i2} z_{ij1} + b_{i3} z_{ij2} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 1), \quad (3.27)$$

with true values $(\beta_1, \beta_2) = (1, 1)'$ and true variance-covariance matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} 9 & 4.8 & .6 \\ 4.8 & 4 & 1 \\ .6 & 1 & 1 \end{bmatrix}, \quad (3.28)$$

such that there are 7 unimportant predictors for the fixed effects and 1 unim-

portant predictor with respect to the random effects. The covariates x_{ijk} for, $k = 1, \dots, 9$ and z_{ijl} , for $l = 1, 2, 3$ are generated from a uniform $(-2, 2)$ distribution, along with a vector of $\mathbf{1}$'s for the subject-specific intercept.

2. Example 2: The setup for the second scenario is the same as the first, except we increase the number of observation to $m = 60$ subjects and $n_i = 10$ observations per subject. This allows us to investigate the performance in a larger sample.
3. Example 3: We now set $m = 60$ subjects and $n_i = 5$ observations per subject, for a particular case where $p = 9$ and $q = 10$. The covariates x_{ijk} for $k = 1, \dots, 9$, are then generated from a uniform $(-2, 2)$ distribution. We set $\mathbf{Z}_i = \mathbf{X}_i$ plus a random intercept term, this model specification allows each regression coefficient including the intercept to vary for different subjects. The true model is then given by

$$y_{ij} = b_{i1} + (\beta_1 + b_{i2})x_{ij1} + b_{i3}x_{ij2} + \beta_3x_{ij3} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 1), \quad (3.29)$$

the true parameter values are $(\beta_1, \beta_3) = (1, 1)'$, and the true covariance matrix is the same as example one, such that there are 7 unimportant predictors corresponding to the fixed as well as the random effects.

For the simulation study, model comparisons and validation are made based on the Kullback-Leibler discrepancy (Kullback and Leibler, 1954) given by

$$KLD = E \left\{ L(\phi_0) - L(\hat{\phi}) \right\}. \quad (3.30)$$

Here $L(\phi)$ is given by (3.5), ϕ_0 are the true parameters, and $\hat{\phi}$ are the estimates obtained for the selected models for each methods under comparison and the expectation is taken with respect to the true model.

***** TABLE 3.1 GOES HERE *****

Table 3.1 compares our proposed method (denoted by M-ALASSO) tuned via the BIC to 5 variable selection algorithms: EGIC (Niu and Pu, 2006), REMLIC

(Wolfinger, 1993; Diggle, Liang and Zeger, 1994), stepwise procedure (denoted by STEPWISE), LASSO (Tibshirani, 1996) and the adaptive LASSO (Zou, 2006), all of which are tuned using either AIC and/or BIC. Note that, the LASSO, adaptive LASSO and STEPWISE are used to perform selection only on the fixed effects, given the random effects selected using REML.IC. Comparisons are also shown for the true model (denoted by Oracle) and the full model (denoted by ML/REML), where the coefficients β are obtained using the ML method, and the covariance matrix is then estimated using REML.

Column 4 lists the median Kullback-Leibler discrepancy (KLD) along with its bootstrap standard error's over 200 simulations from the true model for all three examples. In column 5 we report the relative-efficiency (RE) which is computed as the ratio of the median KLD of the 'Oracle' to the median KLD obtained for each method. We see that for all the three scenarios the relative KLD between our method denoted by M-ALASSO with the BIC-type tuning criterion is the closest to the 'Oracle' with a relative-efficiency upward of 0.75. We also notice that as the sample increases from example one to example two, the relative KLD between our method and the 'Oracle' model becomes much smaller, as the theoretical results would suggest. Column six (% Correct) in Table 3.1 gives us the percentage of times the true model (fixed and random effects combined) is selected, while columns seven (% CF) and column eight (% CR) corresponds to the percentage of times the correct fixed and the correct random effects are selected by each method, respectively. In all the three examples we see that our method, which jointly selects the fixed and random effects in a single step, outperforms all the other competing method by correctly identifying the true model the highest percentage of times.

3.6 Real Data Analysis

In this section we demonstrate the performance of our method on a real dataset. For our real data example we consider the U.S. EPA CASTnet (Clean Air Status and Trend Network) data. This data has been widely used in air quality models

to simulate the levels of various air pollutants in the atmosphere. Recently, Lee and Ghosh (2008) used this data to capture the relationship between total Nitrate concentration (TNO_3) and a set of measured predictors. The data used here is a subset of the dataset obtained from 15 relevant sites of NO_x ($\text{NO}_2 + \text{NO}$) emission from 2000 to 2004 in the eastern portion of the United States. The 15 sites are shown in Figure 3.1.

***** FIGURE 3.1 GOES HERE *****

The data has been averaged to create monthly observations. The sites vary in the number of observations that they have over a 5 year period. There are a total of 826 observations among the 15 sites. The response variable is taken as $\log(\text{TNO}_3)$ rather than TNO_3 , as in the previous analysis. The responses \mathbf{y} have been centered and the predictors have been standardized, hence the fixed intercept can be removed.

***** FIGURE 3.2 GOES HERE *****

We see from Figure 3.2 that the behavior of $\log(\text{TNO}_3)$ concentration seems to exhibit seasonality. In order to account for this periodic effect we include trigonometric functions $s_j(t) = \text{Sin}(\frac{2\pi jt}{T})$ and $c_j(t) = \text{Cos}(\frac{2\pi jt}{T})$ for $T=12$ (months) and $j = 1, 2, 3$, as potential predictors to capture this seasonal effect. In addition there seem to be a possible overall downward trend over the 5 year period. The data now consists of 9 quantitative predictors, 6 constructed predictors, plus a covariate (denoted by $l(t)$) to capture any linear trend, making it a total of 16 variables (see Appendix B for a description of the dataset).

This data is of specific interest to us due to the possible heterogeneity in the measured value of TNO_3 (see, Figure 3.3) and in the association between the response and the set of 16 predictors among the 15 sites. Hence, we apply our approach to not only identify the predictors that have an significant overall population-averaged effect, but also the set of predictors that contribute to the heterogeneity among the sites. To achieve this, we fit a linear mixed-effects model by setting $\mathbf{Z}_i = \mathbf{X}_i$ (as

in example 3, section 5.1) along with a random intercept. This model specification allows each regression coefficient as well as the intercept to vary across the sites.

***** TABLE 3.2 GOES HERE *****

Table 3.2 compares the variables selected using the different methods, for both fixed and random effects. For the mean structure of our model, we see that our method (M-ALASSO) selects variables x_2, x_6, x_7, x_9 and months (denoted by $l(t)$) along with first order harmonics $s_1(t)$ and $c_1(t)$ which has a period of 12, to have an overall population-averaged effect and shrinks the regression coefficients of the other predictors to zero. In general the other approaches tend to select some additional predictors, often being an additional harmonic term. Note that again none of these methods combine the selection of fixed and random effects into a single step.

***** FIGURE 3.4 GOES HERE *****

Figure 3.4 is the plot of predicted mean overlaid with the observed values of $\log(TNO_3)$ using the penalized likelihood estimates obtained for the fixed effects using our method, for 4 specific sites of interest. Though our selected model seems to fit well, we see that in some sites it tends to underestimate (DCP 114) while overestimate (PNF 126) in others. Hence, this further reiterates that the random effects are important, in that one would need to account for possible heterogeneity between the sites by introducing subject-specific slopes. We can also see from Figure 3.4 that a single cycle ($s_1(t), c_1(t)$) seems to be sufficient to describe the seasonal trend.

To assess the site to site variability, our method selects the random-intercept which accounts for the heterogeneity among the 15 sites in the measured $\log(TNO_3)$ along with the subject-specific slopes for the covariates $x_2, x_6, l(t), s_1(t)$ and $c_1(t)$, and shrinks the variance of the other random coefficients to zero. On the other hand, a backward elimination procedure (while keeping the mean structure fixed at the full model) includes random coefficients for x_3 and $s_2(t)$ and leaves out $s_1(t)$ from the model.

***** TABLE 3 GOES HERE *****

Table 3 list the estimated regression coefficients and the estimated variance of the random effects for the model selected using our proposed method.

3.7 Discussion and Future Work

In this paper we have demonstrated that within an LME model framework our proposed method successfully identifies the important predictors that correspond to the fixed and the random effects, simultaneously. We have shown that the proposed reparameterized LME model using the modified Cholesky decomposition of the covariance matrix aids us in the efficient selection of the random effects by using just a single tuning parameter. By using simulated and real data we have illustrated that the proposed penalized method can outperform the commonly used methods with respect to both selection and estimation. We have also shown both theoretically as well as empirically that our penalized likelihood estimators asymptotically performs as well as the ‘Oracle’ model.

There has been considerable interest in variable selection in a generalized linear mixed effects model (GLMM) setting (for example, see, Jiang at al., 2006; Kinney and Dunson, 2008). Zou (2006) proposed an extension of the adaptive LASSO to generalized linear models. It is possible to extend the use of our proposed shrinkage method to simultaneously identify the important fixed and random effects in a GLMM setting. However, it is not straightforward to directly do so, and further research in this direction is warranted.



Figure 3.1: The location of the 15 Sites that were used for our analysis. The ▽ represents the 4 sites used for the overlay plots in Figure 3.4.

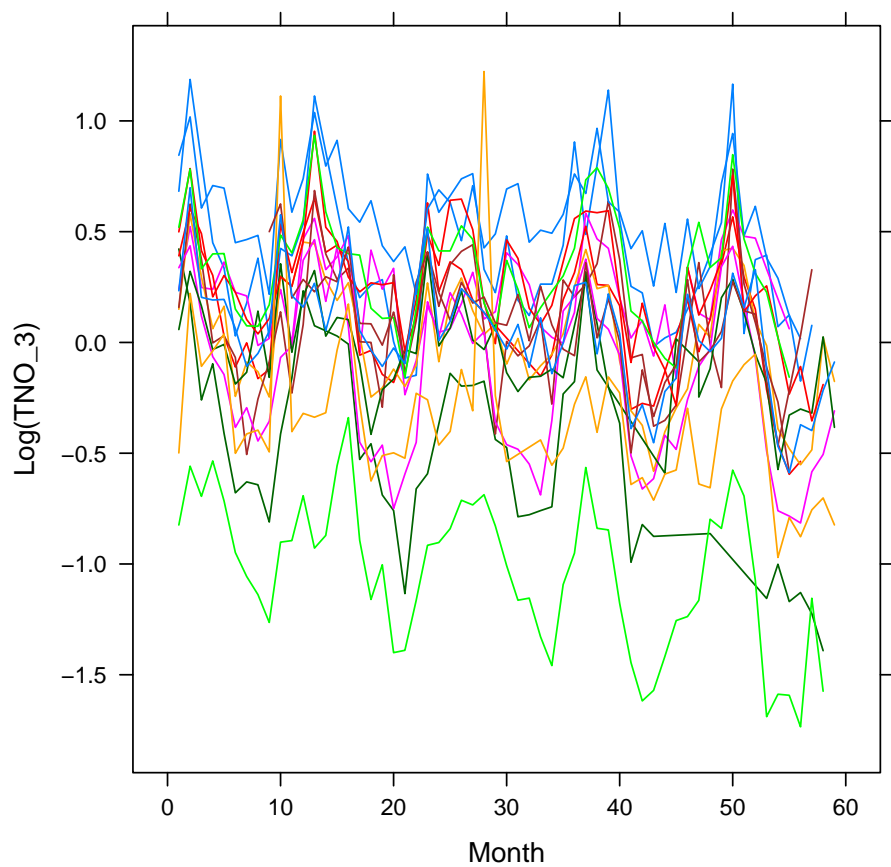


Figure 3.2: Site (individual) profile plot to assess the seasonal trend in Nitrate concentration over each 12 month period, for the CASTnet dataset.

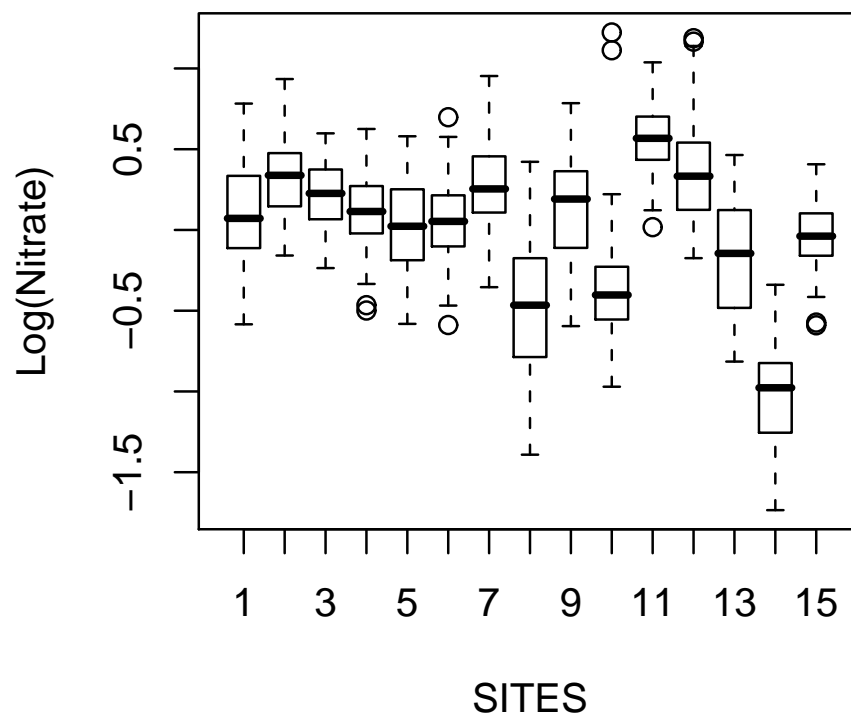


Figure 3.3: Box Plot to assess heterogeneity among the 15 sites in the measured Nitrate concentration corresponding to the CASTnet dataset.

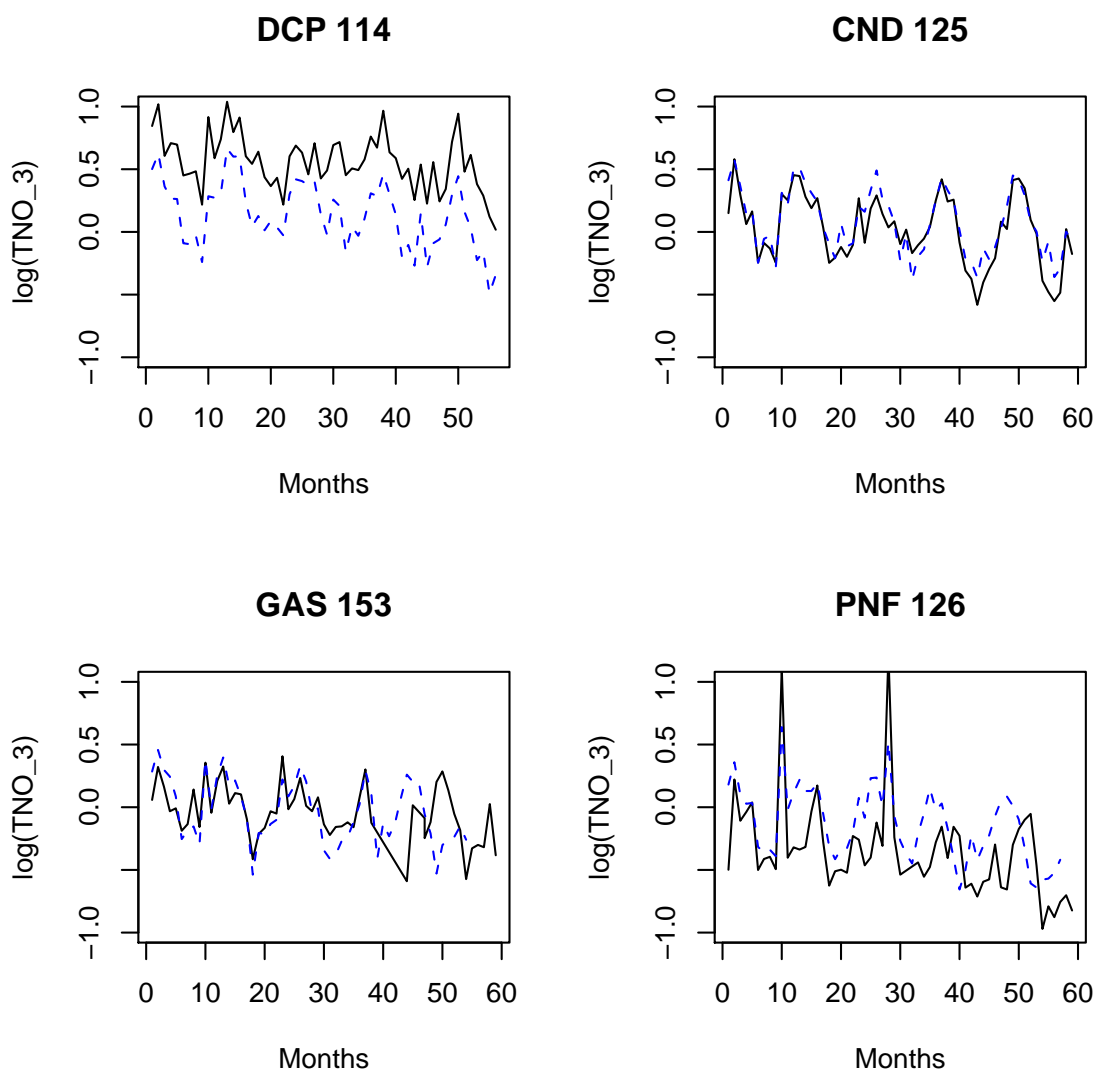


Figure 3.4: Plot of the observed $\text{LOG}(\text{TNO})_3$ concentration represented by the solid line, overlaid with the fitted values for the fixed effects model selected using our proposed method for 4 centers, for the CASTnet dataset.

Table 3.1: Comparing the median KullbackLeibler discrepancy (KLD) from the true model, along with the percentage of the times the true model was selected (%Correct) for each method, across 200 datasets. R.E represents the relative efficiency compared to the oracle model. %CF, %RF corresponds to the percentage of times the correct fixed and random effects were selected, respectively.

Example	Method	Tuning	$D_{KL}(S.E)$	—R.E—	%Correct	%CF	%CR
1	Oracle	-	9.70 (0.343)	-	-	-	-
	M-ALASSO	BIC	10.94(0.475)	0.88	71	73	79
	EGIC	BIC	13.91(0.583)	0.69	47	56	52
	RIC	AIC	15.51(0.567)	0.63	19	21	62
	RIC	BIC	12.48(0.642)	0.77	59	59	68
	STEPWISE	AIC	16.01(0.611)	0.60	13	15	62
	STEPWISE	BIC	12.91(0.584)	0.75	51	53	68
	LASSO	AIC	13.52(0.489)	0.71	17	21	62
	LASSO	BIC	12.87(0.414)	0.75	45	47	68
	ALASSO	AIC	13.03(0.399)	0.74	21	24	62
	ALASSO	BIC	12.12(0.414)	0.80	62	63	68
	ML/REML	-	20.71 (0.513)	0.47	0	0	0
2	Oracle	-	7.84(0.326)	-	-	-	-
	M-ALASSO	BIC	7.98(0.341)	0.98	83	83	89
	EGIC	BIC	12.55(0.581)	0.63	48	59	53
	RIC	AIC	11.93(0.432)	0.72	31	34	74
	RIC	BIC	10.18(0.415)	0.77	77	79	81
	STEPWISE	AIC	12.87(0.501)	0.61	26	28	74
	STEPWISE	BIC	10.71(0.438)	0.73	68	69	81
	LASSO	AIC	12.53(0.388)	0.63	29	29	74
	LASSO	BIC	11.44(0.419)	0.69	59	61	81
	ALASSO	AIC	11.12(0.443)	0.69	39	41	74
	ALASSO	BIC	9.41 (0.420)	0.83	74	75	81
	ML/REML	-	15.47 (0.476)	0.51	0	-	-
3	Oracle	-	13.34(0.912)	-	-	-	-
	M-ALASSO	BIC	17.45(0.961)	0.76	61	63	84
	EGIC	BIC	24.89(2.013)	0.53	41	43	59
	RIC	AIC	28.87(2.231)	0.46	12	14	68
	RIC	BIC	23.39(1.872)	0.57	53	54	73
	STEPWISE	AIC	29.58(2.893)	0.47	8	11	68
	STEPWISE	BIC	25.66(2.011)	0.52	38	40	73
	LASSO	AIC	22.97(1.031)	0.58	11	15	68
	LASSO	BIC	21.08(1.176)	0.63	22	25	73
	ALASSO	AIC	21.69(0.958)	0.62	27	29	68
	ALASSO	BIC	20.23(0.961)	0.66	52	55	73
	ML/REML	-	38.52(2.172)	.27	0	0	0

Table 3.2: Variables selected for the fixed and the random components for the CASTnet data.

Method	Tuning	Variables Selected	
		—Fixed—	—Random—
M-ALASSO	BIC	$x_2, x_6, x_7, x_9, l(t), s_1(t), c_1(t)$	Intercept, $x_2, x_6, l(t), s_1(t), c_1(t)$
Repl.IC	AIC	$x_1, x_2, x_3, x_7, x_6, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
Repl.IC	BIC	$x_1, x_2, x_3, x_6, x_9, l(t), s_1(t), c_2(t), s_3(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
STEPWISE	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
STEPWISE	BIC	$x_1, x_2, x_3, x_6, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
LASSO	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
LASSO	BIC	$x_2, x_3, x_6, x_9, l(t), s_1(t), c_1(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
ALASSO	AIC	$x_1, x_2, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$
ALASSO	BIC	$x_2, x_7, x_9, l(t), s_1(t), c_1(t)$	Intercept, $x_2, x_3, x_6, l(t), c_1(t), s_2(t)$

Table 3.3: Penalized Likelihood estimates for regression coefficients and the random effects variances for the model selected using our proposed method.

Variables	Int	SO ₄	NH ₄	O ₃	<i>T</i>	T _d	<i>RH</i>	<i>SR</i>	<i>WS</i>	<i>P</i>	<i>l(t)</i>	<i>s</i> ₁ (<i>t</i>)	<i>c</i> ₁ (<i>t</i>)	<i>s</i> ₂ (<i>t</i>)	<i>c</i> ₂ (<i>t</i>)	<i>s</i> ₃ (<i>t</i>)	<i>c</i> ₃ (<i>t</i>)
Fixed	-	0	4.28	0	0	0	-1.81	1.07	0	-0.84	-1.43	3.97	4.08	0	0	0	0
Random	0.97	0	2.23	0	0	0	2.13	0	0	0	1.87	2.55	1.81	0	0	0	0

Table 3.4: Spearman Rank Correlation Coefficients between Observed Values

	TNO ₃	SO ₄	NH ₄	O ₃	SR	<i>T</i>	T _d	<i>WS</i>	<i>RH</i>	<i>P</i>
TNO ₃	1.00									
SO ₄	0.27	1.00								
NH ₄	0.58	0.83	1.00							
O ₃	0.09	0.66	0.41	1.00						
SR	0.06	0.58	0.35	0.89	1.00					
<i>T</i>	-0.09	0.69	0.40	0.80	0.80	1.00				
T _d	-0.16	0.63	0.36	0.64	0.63	0.91	1.00			
<i>WS</i>	0.50	-0.17	0.07	-0.07	-0.03	-0.22	-0.28	1.00		
<i>RH</i>	0.00	0.29	0.30	-0.05	-0.09	0.16	0.28	0.00	1.00	
<i>P</i>	-0.14	-0.10	-0.01	0.08	0.03	0.18	0.20	0.05	0.39	1.00

Bibliography

- Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Second international symposium on information theory*, eds. Petrov, B. N. and Csaki, F.
- Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Biometrics*, **50**, 821-826.
- Bates, D. M. and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**, 1-17.
- Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.*, **91**, 109-122.
- Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection. *Ann. Statist.*, **24**, 2350-2383.
- Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc. B*, **60**, 627-642.
- Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, **64**, 115-123.
- Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.*, **101**, 157-167.
- Chipman, H., George, E. I. and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection, vol. 38 of *IMS Lecture Notes - Monograph Series*

- Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762-769.
- Dimendenko, E. (2004). *Mixed Models: Theory and Applications*. New York: John Wiley & Sons.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete Data via the EM Algorithm. *J. R. Statist. Soc. B*, **39**, 1-38.
- Diggle, P, Liang, K, Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford Press.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32**, 407-499.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.
- Fernandez, C., Ley, E. and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics*, **100**, 381-427.
- Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**, 1947-1975.
- George, E. I(2000). The variable selection problem. *J. Amer. Statist. Assoc.*, **95**, 1304-1308.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **7**, 881-889.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica*, **7**, 339-374.
- Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, (eds. J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith), pp. 609-620. Oxford Univ. Press.
- Ghosh, S. K., Bhave, P. V., Davis, J. M. and Lee H. (2008). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models, (under review).

- Greenland, S.(2000). When should Epidemiologic Regressions Use Random Effects? *Biometrics*, **56**, 915-921.
- 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. *J. Amer. Statist. Assoc.*, **72**, 320-338.
- Hoerl, A. E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.
- Yang., H (2007). *Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis*. Ph.D Thesis, North Carolina State University.
- Jiang, J., Rao, J., Gu, Z. and Nguyen, T. (2008). Fence method for mixed models selection. *Ann. Statist.*, **36**, 1669-1692.
- Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothesis and its relationship to the schwarz criterion. *J. Amer. Statist. Assoc.*, **90**, 928-934.
- Kinney, S. K. and Dunson, D. B.(2007). Fixed and Random Effects selection in Linear and Logistic Models. *Biometrics*, **63**, 690-698.
- Kullback, S. and Leiber, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 72-86.
- Laird, N. M. and Ware, J. L. (1982). Random-Effects model for Longitudinal Data. *Biometrics*, **38**, 963-974.
- Laird, N. M., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Amer. Statist. Assoc.*, **82**, 97-105.
- Lan, L. (2006). *Variable selection for linear mixed models for longitudinal data*. Ph.D. thesis, North Carolina State University.

- Lange, N. and Laird, N. M. (1989). The effect of Covariance Structures on Variance Estimation in Balance Growth-Curve Models with Random Parameters. *J. Amer. Statist. Assoc.*, **84**, 241-247.
- Lee, H. and Ghosh, S. K. (2008). A Reparametrization approach for dynamic space-time models. *Journal of Statistical Theory and Practice*, **2**, 1-14.
- Liang, F., Paulo, R., Molina, G., Clyde, C.A. and Berger, J.O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *J. Amer. Statist. Assoc.*, **103**, 410-423.
- Lin, X. (1997). Variance Components Testing in Generalized Linear Models with Random Effects. *Biometrika*, **84**, 309-9326.
- Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects models for Repeated Measures Data. *J. Amer. Statist. Assoc.*, **83**, 1014-1022.
- Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence. *Biometrika*, **81**, 633-648.
- Liu, C., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755-770.
- Mangold, W. D., Bean, L. and Adams, D. (2003). The Impact of Intercollegiate Athletics on Graduation Rates among Major NCAA Division I Universities: Implications for College Persistence Theory and Practice, *Journal Of Higher Education*, pp. 540-562.
- Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall/ CRC, 2nd ed.
- Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.*, **83**, 1023-1032.
- Meng, X. -L. and Rubin D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267-278.
- Meng, X. -L. (1997). The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research*, **86**, 899-909.

- Meng, X. -L. and van Dyk, D. (1998). Fast EM-type implementation for mixed effects models. *J. R. Statist. Soc. B*, **3**, 511-567.
- McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and Extensions*. New York: John Wiley & Sons.
- Morell, C. H., Pearson, J. D. and Brant, L. J.(1997). Linear Transformations of Linear Mixed-Effects Models. *The American Statistician*, **51**, 338-343.
- Niu, F. and Pu, P.(2006). Selecting mixed-effects models based on generalized information criterion. *Journal of Multivariate Analysis*, **97**, 733-758.
- Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block information when Block sizes and Unequal. *Biometrika*, **58**, 545-554.
- Pinheiro, J. and Bates, D.(1996). Unconstrained parameterizations for Variance-Covariance Matrices. *Stat. Computing* **6**, 289-286.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179-191.
- Rao, C. R. and Wu, Y.(1989). A Strongly Consistent Procedure for Model Selection in Regression Problems. *Biometrika*, **76**, 369-374.
- Schwarz, G. (1978). Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 461-464.
- Searle, S. R., Casella, G. and McCulloch C. E. (1992). *Variance Components*. John Wiley & Sons.
- Self, S. G. and Liang, K. Y. (1987). Asymptotic Properties of Maximum Likelihood Ratio Tests under Non-Standard conditions. *J. Amer. Statist. Assoc.*, **82**, 605-610.
- Shao, J. (1997). An asymptotic Theory for linear model selection. *Statistica Sinica*, **7**, 221-264.
- Stram, D. O. and Lee, J. W. (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, **50**, 1171-1177.

- Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 314-343.
- Smith, M. and Kohn, R. (2002). Parsimonious Covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.*, **97**, 1141-1153.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. R. Statist. Soc. B*, **58**, 267-288.
- Weiss, R. E., Wang, Y. and Ibrahim, J. G. (1997). Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors. *Biometrics*, **53**, 592-602.
- Wolfinger, R.D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, **22**, 1079- 1106.
- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937-950.
- Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models, *J. Amer. Statist. Assoc.*, **100**, 1215-1224.
- Zellner, A. (1986). *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*. In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, (eds. P. K. Goel and A. Zellner), pp. 233-243. North-Holland/Elsevier.
- Zou, H. (2006). Adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418-1429.
- Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net, *J. R. Statist. Soc. B*, **67**, 301-320.
- Zou, H., Hastie, T., and Tibshirani, R. (2007). On the Degree's of Freedom of the LASSO. *Ann. Statist.*, **35**, 2173-2192.

Appendices

Appendix A

Description of NCAA Data

Data from Mangold, Bean, Adams (2003), Journal Of Higher Education, p. 540-562, "The Impact of Intercollegiate Athletics on Graduation Rates Among Major NCAA Division I Universities." The data were taken from the 1996-99 editions of the US News "Best Colleges in America" and from the US Department of Education data and includes 97 NCAA Division 1A schools. The authors hoped to show that successful sports programs raise graduation rates. Here is a list describing briefly the response variable and 19 predictors.

<i>Y</i>	Average 6 yr graduation rate for 1996, 1997, 1998
<i>x1</i>	% Students in top 10 Percent HS
<i>x2</i>	ACT COMPOSITE 25TH
<i>x3</i>	% On living campus
<i>x4</i>	% First-time undergraduates
<i>x5</i>	Total Enrollment/1000
<i>x6</i>	% Courses taught by TAs
<i>x7</i>	Composite of basketball ranking
<i>x8</i>	In-state tuition/1000
<i>x9</i>	Room and board/1000
<i>x10</i>	Avg BB home attendance
<i>x11</i>	Full Professor Salary
<i>x12</i>	Student to faculty ratio
<i>x13</i>	% White
<i>x14</i>	Assistant professor salary
<i>x15</i>	Population of city where located
<i>x16</i>	% Faculty with PHD
<i>x17</i>	Acceptance rate
<i>x18</i>	% Receiving loans
<i>x19</i>	% Out of state

Appendix B

Description of CASTnet Data

A complete description of the data can be found on the EPA website: <http://www.epa.gov/castnet>. The data used here is a subset of the complete data and consists of 826 observation from 15 relevant sites of NO_X ($NO_2 + NO$) emission from 2000 to 2004 across the eastern United States. The map shown in Figure (??) marks the relevant sites we have used for this analysis. Here is a list describing briefly the response variable and 16 predictors.

Y	Log of Total Nitrate Concentration ($\mu\text{mol}/\text{m}^3$)
x_1	SO_4 , Sulphate Concentration ($\mu\text{mol}/\text{m}^3$)
x_2	NH_4 , Ammonia Concentration ($\mu\text{mol}/\text{m}^3$)
x_3	O_3 , Maximum Ozone (ppb, parts per billion)
x_4	T , Average Temperature (°)
x_5	T_d , Average Dew Point Temperature (°)
x_6	RH , Average Relative Humidity (%)
x_7	SR , Average Solar Radiation (W/m^2)
x_8	WS , Average Wind Speed (m/sec)
x_9	P , Total Precipitation ($mm/month$)
$l(t)$	Time of measurement in months ($1, \dots, 60$) from 2000-2004
$s_j(t)$	$\text{Sin}(\frac{2\pi jt}{T})$ for $T=12$ and $j = 1, 2, 3$
$c_j(t)$	$\text{Cos}(\frac{2\pi jt}{T})$ for $T=12$ and $j = 1, 2, 3$

Appendix C

Asymptotic Properties: Regularity conditions and proofs

C.1 Regularity Condition

Assume that the data $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{y}_i); i = 1, \dots, m\}$ is a random sample from a linear mixed-effects model (3.3) with probability density $f(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\phi})$ where $\boldsymbol{\phi} = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$ is a $k \times 1$ vector of unknown parameters. Let $L_i(\boldsymbol{\phi}) = \log(f(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\phi}))$ denote the contribution of observation i to the log-likelihood function, and is given by

$$L_i(\boldsymbol{\phi}) = -\frac{1}{2} \log |\mathbf{V}_i| - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' (\mathbf{V}_i)^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \quad (\text{C.1})$$

where $\mathbf{V}_i = \sigma^2 (\mathbf{Z}_i \mathbf{D} \boldsymbol{\Gamma} \boldsymbol{\Gamma}' \mathbf{D} \mathbf{Z}_i' + \mathbf{I}_{n_i})$. Let $L(\boldsymbol{\phi}) = \sum_{i=1}^n L_i(\boldsymbol{\phi})$ and $Q(\boldsymbol{\phi})$ denote the log-likelihood and the penalized log-likelihood as given in (3.5) and (3.21), respectively. To present the proof of the theorems the following regularity conditions are imposed:

- (i) The Fisher information matrix $I(\boldsymbol{\phi}_{10})$ knowing $\boldsymbol{\phi}_{20} = 0$ is finite and positive definite.
- (ii) There exists a subset Θ of \mathbb{R}^k , containing the true parameter $\boldsymbol{\phi}_0$ such that $L_i(\boldsymbol{\phi})$ given in (C.1) admits all third order derivatives. Specifically, for $\phi_j = \beta_j$ and $(\phi_l, \phi_m) = \{(d_l, \gamma_m), (d_l, \gamma_m), (\gamma_l, \gamma_m)\}$, there exists a function $M_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ such that

$$\left| \frac{\partial^3}{\partial \beta_j \partial \phi_l \partial \phi_m} L_i(\boldsymbol{\phi}) \right| = \left| \mathbf{X}'_{ij} \frac{\partial \mathbf{V}_i^{-1}}{\partial \phi_l \partial \phi_m} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}) \right| < M_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$

for all $\phi \in \Theta$, and $E_{\phi_0}[M_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)] < \infty$. For $(\phi_j, \phi_l) = (\beta_j, \beta_l)$ and ϕ_m is either d_m or γ_m there exists a function $N_{jlm}(\mathbf{y}, \mathbf{X}, \mathbf{Z})$ such that

$$\left. \begin{aligned} \left| \frac{\partial^3}{\partial \beta_j \partial \beta_l \partial d_m} L_i(\phi) \right| &= |\mathbf{X}'_{ij} (\mathbf{V}_i^{-1} \mathbf{S}_i^m \mathbf{V}_i^{-1}) \mathbf{X}_{il}| \\ \left| \frac{\partial^3}{\partial \beta_j \partial \beta_l \partial \gamma_m} L_i(\phi) \right| &= |\mathbf{X}'_{ij} (\mathbf{V}_i^{-1} \mathbf{T}_i^m \mathbf{V}_i^{-1}) \mathbf{X}_{il}| \end{aligned} \right\} < N_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$

for all $\phi \in \Theta$, and $E_{\phi_0}[N_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)] < \infty$. Here \mathbf{S}_i^m and \mathbf{T}_i^m denote the partial derivatives of \mathbf{V}_i with respect to d_m and γ_m , respectively, and are given by

$$\mathbf{S}_i^m = \mathbf{Z}_i \left\{ \frac{\partial}{\partial d_m} (\mathbf{D} \mathbf{\Gamma} \mathbf{\Gamma}' \mathbf{D}) \right\} \mathbf{Z}'_i, \quad \mathbf{T}_i^m = \mathbf{Z}_i \mathbf{D} \left\{ \frac{\partial}{\partial d \gamma_m} (\mathbf{\Gamma} \mathbf{\Gamma}') \right\} \mathbf{D} \mathbf{Z}'_i. \quad (\text{C.2})$$

For $\phi_j = d_j$ and $(\phi_l, \phi_m) = \{(d_l, d_m), (d_l, \gamma_m), (\gamma_l, \gamma_m)\}$,

$$\left| \frac{\partial^3}{\partial d_j \partial \phi_l \partial \phi_m} L_i(\phi) \right| < P_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$

for all $\phi \in \Theta$, and $E_{\phi_0}[P_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)] < \infty$.

Note that although it must be that $d_j \geq 0$ for all j , we allow the estimates to fall outside the boundary of the parameter space.

C.2 Proof of Theorem 1

Proof. Consider the penalized log-likelihood given in (3.21) in a neighborhood of the true value ϕ_{10} . Let $\alpha_n = n^{-1/2}$ with $\mathbf{u} \neq 0$, and $\phi_1 = \phi_{10} + \alpha_n \mathbf{u}$. Fixing $\phi_2 = 0$ we show that for a small enough $\epsilon > 0$ there exists a large constant C such that

$$P \left\{ \sup_{\|\mathbf{u}\|=C} \mathbf{Q} \left(\begin{array}{c} \phi_{10} + \alpha_n \mathbf{u} \\ 0 \end{array} \right) < \mathbf{Q} \left(\begin{array}{c} \phi_{10} \\ 0 \end{array} \right) \right\} \geq 1 - \epsilon.$$

Note that

$$\begin{aligned} D_n(\mathbf{u}) &\equiv \mathbf{Q}(\phi_1) - \mathbf{Q}(\phi_{10}) \\ &= \{L(\phi_{10} + \alpha_n \mathbf{u}) - L(\phi_{10})\} - \lambda_n \left[\sum_{j=1}^s \bar{w}_j (|\phi_{j0} + \alpha_n u_j| - |\phi_{j0}|) \right]. \end{aligned}$$

Using a Taylor series expansion we have

$$\begin{aligned} D_n(\mathbf{u}) &= \alpha_n (\nabla L(\phi_{10}))' \mathbf{u} + \frac{1}{2} \mathbf{u}' [\nabla^2 L(\phi_{10})] \mathbf{u} \alpha_n^2 - \lambda_n \sum_{j=1}^s \bar{w}_j \text{sgn}(\phi_{j0}) \alpha_n u_j \\ &= \frac{1}{\sqrt{n}} (\nabla L(\phi_{10}))' \mathbf{u} + \frac{1}{2n} \mathbf{u}' [\nabla^2 L(\phi_{10})] \mathbf{u} - \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^s \bar{w}_j \text{sgn}(\phi_{j0}) u_j, \end{aligned} \quad (\text{C.3})$$

where $\nabla L(\phi_{10})$, $\nabla^2 L(\phi_{10})$ denote the vector and matrix of the first and second order partial derivatives of $L(\phi_1)$, respectively, evaluated at ϕ_{10} . From regularity condition *ii* it follows that

$$\left(\frac{1}{6n^{3/2}} \right) \sum_{i=1}^n \frac{\partial}{\partial \phi_j \partial \phi_l \partial \phi_m} L_i(\phi) \Big|_{\phi_1 = \phi_{10}} \rightarrow 0, \text{ as } n \rightarrow \infty,$$

hence the remainder term vanishes. For $\nabla L(\phi_{10})$ the j^{th} partial derivative for each corresponding β_1 , \mathbf{d}_1 and γ_1 satisfies

$$\left. \begin{aligned} E \left\{ \frac{\partial}{\partial \beta_j} L(\phi_1) \right\} &= E \left[\mathbf{X}'_{(1)j} \tilde{\mathbf{V}}_{(1)}^{-1} (\mathbf{y} - \mathbf{X}_{(1)} \beta_1) \right] \\ E \left\{ \frac{\partial}{\partial d_j} L(\phi_1) \right\} &= E \left[\frac{1}{2} [\text{Tr}(\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{S}}_{(1)}^j) + (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)' (\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{S}}_{(1)}^j \tilde{\mathbf{V}}_{(1)}^{-1}) (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)] \right] \\ E \left\{ \frac{\partial}{\partial \gamma_j} L(\phi_1) \right\} &= E \left[\frac{1}{2} [-\text{Tr}(\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{T}}_{(1)}^j) + (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)' (\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{T}}_{(1)}^j \tilde{\mathbf{V}}_{(1)}^{-1}) (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)] \right] \end{aligned} \Big|_{\phi_1 = \phi_{10}} \right\} = 0,$$

where $\mathbf{X}_{(1)j}$ corresponds to the j^{th} column of stacked matrix $\mathbf{X}_{(1)}$, and $\tilde{\mathbf{S}}_{(1)}^j$ and $\tilde{\mathbf{T}}_{(1)}^j$ are block diagonal matrices of the partial derivatives of $\tilde{\mathbf{V}}_{(1)}$ and are given by

$$\tilde{\mathbf{S}}_{(1)}^j = \mathbf{Z}_{(1)} \left\{ \frac{\partial}{\partial d_j} (\tilde{\mathbf{D}}_1 \tilde{\Gamma}_1 \tilde{\Gamma}_1' \tilde{\mathbf{D}}_1) \right\} \mathbf{Z}'_{(1)} \text{ and } \tilde{\mathbf{T}}_{(1)}^j = \mathbf{Z}_{(1)} \tilde{\mathbf{D}}_1 \left\{ \frac{\partial}{\partial \gamma_j} (\tilde{\Gamma}_1 \tilde{\Gamma}_1') \right\} \tilde{\mathbf{D}}_1 \mathbf{Z}'_{(1)}.$$

From standard arguments we have

$$\left. \begin{aligned} \frac{1}{\sqrt{n}} \mathbf{X}'_{(1)j} \tilde{\mathbf{V}}_{(1)}^{-1} (\mathbf{y} - \mathbf{X}_{(1)} \beta_1) \\ \frac{1}{2\sqrt{n}} [\text{Tr}(\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{S}}_{(1)}^j) + (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)' (\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{S}}_{(1)}^j \tilde{\mathbf{V}}_{(1)}^{-1}) (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)] \\ \frac{1}{2\sqrt{n}} [-\text{Tr}(\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{T}}_{(1)}^j) + (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)' (\tilde{\mathbf{V}}_{(1)}^{-1} \tilde{\mathbf{T}}_{(1)}^j \tilde{\mathbf{V}}_{(1)}^{-1}) (\mathbf{y} - \mathbf{X}_{(1)} \beta_1)] \end{aligned} \Big|_{\phi_1 = \phi_{10}} \right\} = \mathbf{O}_p(1). \quad (\text{C.4})$$

For $\nabla^2 L(\phi_1)$ we have

$$\frac{1}{n} \nabla^2 L(\phi_{10}) \rightarrow_p -I(\phi_{10}), \quad (\text{C.5})$$

where $I(\phi_{10})$ is the Fisher information evaluated at ϕ_{10} . Using (C.4) and (C.5) the expansion in (C.3) becomes

$$D_n(\mathbf{u}) = \mathbf{O}_p(1)\mathbf{u} - \frac{1}{2}\mathbf{u}'\{I(\phi_{10}) + o_p(1)\}\mathbf{u} - \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^s \bar{w}_j \text{sgn}(\phi_{j0}) u_j.$$

Since $I(\phi_{10})$ is finite and positive definite (condition *i*), hence choosing a sufficiently large C , the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. For the penalty term, if $\lambda_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, and since $\bar{w}_j = 1/\hat{\phi}_j \rightarrow 1/\phi_j$, it follows that

$$\frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^s \bar{w}_j \text{sgn}(\phi_{j0}) u_j \rightarrow_p 0,$$

and thus is also dominated by the second term. Hence by choosing a sufficiently large C there exists a local maximum in the ball $\{(\phi_{10} + \alpha_n \mathbf{u}, 0)' : \|\mathbf{u}\| \leq C\}$ with probability with $1 - \epsilon$, and hence there exists a local maximizer $\hat{\phi} = (\hat{\phi}_1, 0)$ of $\phi_0 = (\phi_{10}, 0)$ such that $\|\hat{\phi}_1 - \phi_{10}\| = O_p(n^{-1/2})$. \square

C.3 Proof of Theorem 2

Let $\phi = (\beta', \mathbf{d}', \gamma)'$ denote the $k \times 1$ vector of unknown parameters, where $k = k_\beta + k_d + k_\gamma$, the sum of the lengths corresponding to each parameter. Let $\phi_2 = (\beta_2', \mathbf{d}_2', \gamma_2)'$ be a vector of length $k_2 = k - s$, corresponding to the true zero values, where $k_2 = k_{\beta_2} + k_{d_2} + k_{\gamma_2}$.

Proof. It is sufficient to show that with probability tending to 1 as $n \rightarrow \infty$, for any ϕ_1 satisfying $\|\phi_1 - \phi_{10}\| \leq Mn^{-1/2}$ and for some small $\epsilon_n = Mn^{-1/2}$ and for each $j = (s+1), \dots, (k_{\beta_2} + k_{d_2})$, we have that

$$\begin{aligned} \frac{\partial}{\partial \phi_j} \mathbf{Q}(\phi) &< 0 & \text{for } 0 < \phi_j < \epsilon_n, \\ \frac{\partial}{\partial \phi_j} \mathbf{Q}(\phi) &> 0 & \text{for } -\epsilon_n < \phi_j < 0. \end{aligned} \quad (\text{C.6})$$

Note that

$$\frac{\partial}{\partial \phi_j} \mathbf{Q}(\phi) = \frac{\partial}{\partial \phi_j} L(\phi) - \lambda_n \bar{w}_j \text{sgn}(\phi_j).$$

To show (C.6) consider the Taylor series expansion about $\partial L(\phi)/\partial \phi_j$, we have

$$\begin{aligned} \frac{\partial}{\partial \phi_j} \mathbf{Q}(\phi) = & \frac{\partial}{\partial \phi_j} L(\phi_0) - \sum_{l=1}^k \frac{\partial}{\partial \phi_j \partial \phi_l} L(\phi_0) (\phi_l - \phi_{l0}) \\ & + \frac{1}{2} \sum_{i=1}^n \sum_{l=1}^k \sum_{m=1}^k \frac{\partial^3}{\partial \phi_j \partial \phi_l \partial \phi_m} L_i(\phi_*) (\phi_l - \phi_{l0}) (\phi_m - \phi_{m0}) - \lambda_n \bar{w}_j \text{sgn}(\phi_j), \end{aligned} \quad (\text{C.7})$$

where ϕ_* lies between ϕ and ϕ_0 . Again the first order partial derivative for j^{th} term for each β and \mathbf{d} are given by

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta_j} L(\phi_0) &= \frac{1}{\sqrt{n}} \mathbf{X}'_j \mathbf{V}_0^{-1} (\mathbf{y} - \mathbf{X} \beta_0) = \mathbf{O}_p(1), \\ \frac{1}{\sqrt{n}} \frac{\partial}{\partial d_j} L(\phi_0) &= 0. \end{aligned}$$

where \mathbf{X}_j corresponds to the j^{th} column of the stacked matrix \mathbf{X} . The second order derivatives in (C.7) follows

$$\left(\frac{1}{n} \right) \nabla^2 L(\phi) \Big|_{\phi=\phi_0} \rightarrow E(\nabla^2 L(\phi)) \Big|_{\phi=\phi_0},$$

where $E(\nabla^2 L(\phi))$ is given as

$$E(\nabla^2 L(\phi)) = E \begin{bmatrix} L_{\beta\beta} & L_{\beta\mathbf{d}} & L_{\beta\gamma} \\ L'_{\beta\mathbf{d}} & L_{\mathbf{d}\mathbf{d}} & L_{\mathbf{d}\gamma} \\ L'_{\beta\gamma} & L'_{\mathbf{d}\gamma} & L_{\gamma\gamma} \end{bmatrix},$$

where $E(L_{\beta\beta}) = -\mathbf{X}' \tilde{\mathbf{V}}^{-1} \mathbf{X}$, and $E(L_{\beta\mathbf{d}}), E(L_{\beta\gamma})$ has j^{th} column

$$\left. \begin{aligned} E \left\{ L_{\beta\mathbf{d}} \right\}_j &= -E[\mathbf{X}'_j (\tilde{\mathbf{V}}^{-1} \tilde{\mathbf{S}}^j \tilde{\mathbf{V}}^{-1}) (\mathbf{y} - \mathbf{X} \beta)] \\ E \left\{ L_{\beta\gamma} \right\}_j &= -E[\mathbf{X}'_j (\tilde{\mathbf{V}}^{-1} \tilde{\mathbf{T}}^j \tilde{\mathbf{V}}^{-1}) (\mathbf{y} - \mathbf{X} \beta)] \end{aligned} \right|_{\phi=\phi_0} \Bigg\} = 0,$$

where $\tilde{\mathbf{S}}^j$ and $\tilde{\mathbf{T}}^j$ are block diagonal matrices of \mathbf{S}_i and \mathbf{T}_i given in (C.2). The expectation for the second order partial derivatives for \mathbf{d} and $\boldsymbol{\gamma}$ has $(j, l)^{th}$ term

$$\begin{aligned} E \{L_{\mathbf{d}\mathbf{d}}\}_{jl} &= -\text{Tr}(\tilde{\mathbf{V}}^{-1} \tilde{\mathbf{S}}^j \tilde{\mathbf{V}}^{-1} \tilde{\mathbf{S}}^l) \\ E \{L_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\}_{jl} &= -\text{Tr}(\tilde{\mathbf{V}}^{-1} \tilde{\mathbf{T}}^j \tilde{\mathbf{V}}^{-1} \tilde{\mathbf{T}}^l) \\ E \{L_{\mathbf{d}\boldsymbol{\gamma}}\}_{jl} &= -\text{Tr}(\tilde{\mathbf{V}}^{-1} \tilde{\mathbf{S}}^j \tilde{\mathbf{V}}^{-1} \tilde{\mathbf{T}}^l), \end{aligned} \quad (\text{C.8})$$

for $j = s + 1, \dots, (k_\beta + k_d)$, it can be shown that $\tilde{\mathbf{S}}^j$ or $\tilde{\mathbf{T}}^j$ when evaluated at $\phi_j = 0$ are zero matrices and the set of equations given in (C.8) simplifies to zero.

First, consider $\phi_j = \beta_j$ the expansion given in (C.7) yields

$$\begin{aligned} \frac{1}{\sqrt{n}} \left(\frac{\partial}{\partial \beta_j} \mathbf{Q}(\phi) \right) &= \frac{1}{\sqrt{n}} \left(O_p(n^{1/2}) - n \sum_{l=1}^{k_\beta} \{ \mathbf{X}'_j \mathbf{V}_0^{-1} \mathbf{X}_l + o_p(1) \} (\beta_l - \beta_{l0}) \right. \\ &\quad - n \sum_{l=k_\beta+1}^{k_d} o_p(1) (d_l - d_{l0}) - n \sum_{l=k_d+1}^{k_\gamma} o_p(1) (\gamma_l - \gamma_{l0}) \\ &\quad + \sum_{i=1}^n \sum_{l=1}^{k_\beta} \sum_{m=k_\beta+1}^{k_d} \mathbf{X}'_{ij} (\mathbf{V}_{i*}^{-1} \mathbf{S}_{i*}^m \mathbf{V}_{i*}^{-1}) \mathbf{X}_{il} (\beta_l - \beta_{l0}) (d_m - d_{m0}) \\ &\quad + \sum_{i=1}^n \sum_{l=1}^{k_\beta} \sum_{m=k_d+1}^{k_\gamma} \mathbf{X}'_{ij} (\mathbf{V}_{i*}^{-1} \mathbf{T}_{i*}^m \mathbf{V}_{i*}^{-1}) \mathbf{X}_{il} (\beta_l - \beta_{l0}) (\gamma_m - \gamma_{m0}) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{l=k_\beta+1}^{k_d} \sum_{m=k_\beta+1}^{k_d} \mathbf{X}_{ij} \frac{\partial \mathbf{V}_i^{-1}}{\partial d_l \partial d_m} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_*) (d_l - d_{l0}) (d_m - d_{m0}) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{l=k_d+1}^{k_\gamma} \sum_{m=k_d+1}^{k_\gamma} \mathbf{X}_{ij} \frac{\partial \mathbf{V}_i^{-1}}{\partial \gamma_l \partial \gamma_m} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_*) (\gamma_l - \gamma_{l0}) (\gamma_m - \gamma_{m0}) \\ &\quad \left. + \sum_{i=1}^n \sum_{l=k_\beta+1}^{k_d} \sum_{m=k_d+1}^{k_\gamma} \mathbf{X}_{ij} \frac{\partial \mathbf{V}_i^{-1}}{\partial d_l \partial \gamma_m} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_*) (d_l - d_{l0}) (\gamma_m - \gamma_{m0}) - \lambda_n \bar{w}_j \text{sgn}(\beta_j) \right), \end{aligned} \quad (\text{C.9})$$

where $\boldsymbol{\phi}_*$ lies between $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_0$. Since we are considering $\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| \leq Mn^{-1/2}$, (C.9) gives

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \beta_j} \mathbf{Q}(\phi) = -\lambda_n \frac{\bar{w}_j}{\sqrt{n}} \text{sgn}(\beta_j) + O_p(1).$$

Since for $\beta_{j0} = 0$, we have $\bar{w}_j/\sqrt{n} = |\sqrt{n}\bar{\beta}_j|^{-1} = O_p(1)$, and $\lambda_n \rightarrow \infty$, the sign of the derivative is completely determined by that of β_j .

Now consider $\phi_j = d_j$. The Taylor series expansion in (C.7) gives

$$\begin{aligned}
\frac{1}{\sqrt{n}} \left(\frac{\partial}{\partial d_j} \mathbf{Q}(\phi) \right) = & \frac{1}{\sqrt{n}} \left(\mathbf{0} - n \sum_{l=1}^{k_\beta} o_p(1)(\beta_l - \beta_{l0}) \right. \\
& - n \sum_{l=k_\beta+1}^{k_d} o_p(1)(d_l - d_{l0}) - n \sum_{l=k_d+1}^{k_\gamma} o_p(1)(\gamma_l - \gamma_{l0}) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{l=k_\beta+1}^{k_d} \sum_{m=k_\beta+1}^{k_d} \frac{\partial L_i(d_j)}{\partial d_l \partial d_m} (d_l - d_{l0})(d_m - d_{m0}) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{l=k_d+1}^{k_\gamma} \sum_{m=k_d+1}^{k_\gamma} \frac{\partial L_i(d_j)}{\partial \gamma_l \partial \gamma_m} (\gamma_l - \gamma_{l0})(\gamma_m - \gamma_{m0}) \\
& + \sum_{i=1}^n \sum_{l=1}^{k_\beta} \sum_{m=1}^{k_\beta} \mathbf{X}'_{il} (\mathbf{V}_{i*}^{-1} \mathbf{S}_{i*}^j \mathbf{V}_{i*}^{-1}) \mathbf{X}_{il} (\beta_l - \beta_{l0})(\beta_m - \beta_{m0}) \\
& + \sum_{i=1}^n \sum_{l=k_\beta+1}^{k_d} \sum_{m=k_d+1}^{k_\gamma} \frac{\partial L_i(d_j)}{\partial d_l \partial \gamma_m} (d_l - d_{l0})(\gamma_m - \gamma_{m0}) \\
& + \sum_{i=1}^n \sum_{l=1}^{k_\beta} \sum_{m=k_d+1}^{k_\gamma} \frac{\partial L_i(d_j)}{\partial \beta_l \partial \gamma_m} (\beta_l - \beta_{l0})(\gamma_m - \gamma_{m0}) \\
& \left. + \sum_{i=1}^n \sum_{l=1}^{k_\beta} \sum_{m=k_\beta+1}^{k_d} \frac{\partial L_i(d_j)}{\partial \beta_l \partial d_m} (\beta_l - \beta_{l0})(d_m - d_{m0}) \right), \quad (\text{C.10})
\end{aligned}$$

where

$$L_i(d_j) = \frac{\partial}{\partial d_j} L_i(\phi_*) = \frac{1}{2} [-\text{Tr}(\mathbf{V}_{i*}^{-1} \mathbf{S}_{i*}^j) + (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_*)' (\mathbf{V}_{i*}^{-1} \mathbf{S}_{i*}^j \mathbf{V}_{i*}^{-1}) (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_*)],$$

and ϕ_* lies between ϕ and ϕ_0 . As above, (C.10) simplifies to

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial d_j} \mathbf{Q}(\phi) = -\lambda_n \frac{\bar{w}_j}{\sqrt{n}} \text{sgn}(d_j).$$

Hence, for $d_{j0} = 0$, as $\bar{w}_j/\sqrt{n} = |\sqrt{n}\bar{d}_j|^{-1} = O_p(1)$, the sign of the derivative is again completely determined by that of d_j . This completes the proof. \square

C.4 Proof of Theorem 3

Proof. We have from Theorem 1 shown that there exists a $\hat{\phi}_1$ that is a local maximizer of $\mathbf{Q}(\phi_1)$ such that $\|\hat{\phi}_1 - \phi_{10}\| = O_p(n^{-1/2})$, and satisfies the set of penalized likelihood equations

$$\frac{\partial}{\partial \phi_1} \mathbf{Q}(\phi) \Big|_{\phi=(\hat{\phi}_1, 0)'} = \frac{\partial}{\partial \phi_1} L(\phi) \Big|_{\phi=(\hat{\phi}_1, 0)'} - \lambda_n \mathbf{h}(\hat{\phi}_1) = 0,$$

where $\mathbf{h}(\hat{\phi}_1) = (\bar{w}_1 \text{sgn}(\hat{\phi}_1), \dots, \bar{w}_s \text{sgn}(\hat{\phi}_s))'$ an $s \times 1$ vector where $\bar{w}_j = 0$ for $\phi_j = \gamma_j$. Using the Taylor series expansion and multiplying throughout by $1/n$, we have

$$\begin{aligned} \frac{1}{n} \nabla L(\phi_{10}) - \{I(\phi_{10}) + o_p(1)\}(\hat{\phi}_1 - \phi_{10}) - \frac{\lambda_n}{n} \mathbf{h}(\phi_{10}) &= 0 \\ \sqrt{n} \left\{ (\hat{\phi}_1 - \phi_{10}) I(\phi_{10}) + \frac{\lambda_n}{n} \mathbf{h}(\phi_{10}) \right\} &= \frac{1}{\sqrt{n}} \nabla L(\phi_{10}) \quad . \end{aligned}$$

Since $E\{\nabla L(\phi_1)\} = 0$ as in the proof of Theorem 1, it follows from the multivariate central theorem that

$$\frac{1}{\sqrt{n}} \nabla L(\phi_{10}) \rightarrow_d N(0, I(\phi_{10})),$$

where $I(\phi_{10})$ is as given in the proof of Theorem 1. Therefore

$$\sqrt{n} I(\phi_{10}) \left\{ (\hat{\phi}_1 - \phi_{10}) + \frac{\lambda_n}{n} I(\phi_{10})^{-1} \mathbf{h}(\phi_{10}) \right\} \rightarrow_d N\{0, I(\phi_{10})\}.$$

This completes the proof. □