ADAPTIVE LASSO FOR MIXED MODEL SELECTION VIA PROFILE LOG-LIKELIHOOD

Juming Pan

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2016

Committee:

Junfeng Shang, Advisor

Lewis Fulcher,
Graduate Faculty Representative

Hanfeng Chen

John Chen

ABSTRACT

Junfeng Shang, Advisor

Linear mixed models describe the relationship between a response variable and some predictors for data that are grouped according to one or more clustering factors. A linear mixed model consists of both fixed effects and random effects. Fixed effects are the conventional linear regression coefficients, and random effects are associated with units which are drawn randomly from a population. By accommodating such two types of parameters, linear mixed models provide an effective and flexible way of representing the means as well as the covariance structure of the data, therefore have been primarily used to model correlated data, and have received much attention in a variety of disciplines including agriculture, biology, medicine, and sociology.

Due to the complex nature of the linear mixed models, the selection of only important covariates to create an interpretable model becomes challenging as the dimension of fixed or random effects increases. Thus, determining an appropriate structural form for a model to be used in making inferences and predictions is a fundamental problem in the analysis of longitudinal or clustered data using linear mixed models.

This dissertation focuses on selection and estimation for linear mixed models by integrating the recent advances in model selection. More specifically, we propose a two-stage penalized procedure for selecting and estimating important fixed and random effects. Compared with the traditional subset selection approaches, penalized methods can enhance the predictive power of a model, and can significantly reduce computational cost when the number of variables is large (Fan and Li, 2001). Our proposed procedure is different from the existing ones in the literature mainly in two aspects. First, the proposed method is composed of two stages to separately choose the parameters of interests, therefore can respect and accommodate the distinct properties between the random and fixed effects. Second, the usage of the profile log-likelihoods in the selection process can make the computation more efficient and stable due to a smaller number of dimensions involved.

In the first stage, we choose the random effects by maximizing the penalized restricted profile log-likelihood, and the maximization is completed by the Newton-Raphson algorithm. Observe

that if a random effect is a noise variable, then the corresponding variance components should be all zero. Thus, we first estimate the covariance matrix of random effects using the adaptive LASSO penalized method and then identify the vital ones based on the estimated covariance matrix. In the view of such a selection procedure, the selected random effects are invariant to the selection of the fixed effects. When a proper model for the covariance is adopted, the correct covariance structure will be obtained and valid inferences for the fixed effects can then be achieved in the next stage. We further study the theoretical properties of the proposed procedure for random effects selection. We prove that, with probability tending to one, the proposed procedure surely identifies all true random effects.

After the completion of the random effects selection, in the second stage, we select the fixed effects through the maximization of the penalized profile log-likelihood, which only involves the regression coefficients. The optimization of the penalized profile log-likelihood can be solved by the Newton-Raphson algorithm. We then investigate the sampling properties of the resulting estimate of fixed effects. We show that the resulting estimate enjoys model selection oracle properties, indicating that asymptotically the proposed approach can discover the subset of significant predictors. After finishing the two-stage penalized procedure, the best linear mixed model can subsequently be determined and be applied to handle correlated data in a number of fields.

To illustrate the performance of the proposed method, numerous simulation studies have been conducted. The simulation results demonstrate that the proposed technique is quite efficient in selecting the best covariates and random covariance structure in linear mixed models and outperforms the existing selection methodologies in general. We finally apply the method to two real data applications for further examining its effectiveness in mixed model selection.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deep gratitude and respect to my advisor, Dr. Junfeng Shang. She taught me the delight of studying statistics and motivated me to investigate the area of mixed model selection. She inspired me to be an independent researcher and helped me realize the power of critical thinking. She was always there to cheer me up and to give me sound advice whenever I had a hard time with either research or career. I appreciate all her contributions of ideas, guidance, and encouragement to make my Ph.D experience productive and stimulating.

My sincere thanks must also go to other members of my committee: Dr. Hanfeng Chen, Dr. John Chen, and Dr. Lewis Fulcher. They generously gave their time to offer me valuable suggestions toward improving my work.

I gratefully acknowledge Ohio Supercomputer Center (OSC) for the computing resources and technical support which help to implement the simulation studies and real data applications in this manuscript.

I want to extend my appreciation to the Department of Mathematics and Statistics, and the Department of Applied Statistics and Operations Research, where I have been provided with qualified education, excellent teaching and research environment, and generous financial assistantship during the six years at Bowling Green State University.

Last but not least, I thank with love to my family. I am very grateful to my parents and parents-in-law, for their faith in me and allowing me to follow my ambition. Very special thanks to my wife, Yan Wang, for being my best friend, my lover, and my soulmate. I could never have accomplished this dissertation without her constant understanding, patience, and support. Finally, to my lovely children, Leo and Zoey. They make my life so much happier and more meaningful than I could ever imagine.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Motivation of Dissertation

Conceptually, data are generated from a particular process or system. For an observed data, it contains a certain amount of information about such process or system, and we wish to use a statistical model to express this information in a realistic, yet also interpretable and concise form. Burnham and Anderson (2002) considered modeling of information in data as a change in coding like a change in language. A perfect translation for a poem or an article expressed in one language (e.g., Chinese) to another language (e.g., English) should maintain all the exactness. Similarly, the ultimate objective of model selection is to obtain a transfer such that no information is lost from the data to a model.

Unfortunately, the idealized goal is unachievable since reality cannot be described or predicted with complete accuracy. However, we can try to select a model of the data that is optimal in the sense that the model misses as little information as possible. Box (1976) stated that "all models are wrong, but some are useful". While no model can reflect all of reality, models could be ordered from useful to useless. Therefore, in practice, the task of model selection is to choose the best model from a candidate set, and the ranking of these models depends on the method or the criterion that is utilized.

Guyon and Elisseeff (2003) discussed three targets of model selection. The first target is to improve performance of inferences and predictions. Any statistical inference relies on assumptions. A model is a set of assumptions regarding the generation of the observed data. Given an appropriate model, then methods exist that are objective and optimal for model parameter estimation. Furthermore, models are widely used to predict future outcomes. For example, in survival analysis, statistical models are often utilized to predict the probability that a patient with a set of characteristics will experience a health outcome. Precise perditions based these models can help in clinical decision making of patients' treatment. On the other hand, inappropriate model spec-

ification may significantly impact both the estimators of the model parameters and the predicted values of the response. There are two types of model misspecification, underfitting and overfitting. Underfitting refers to a scenario that the model is too simplistic with the key components are excluded. Underfitting a model may induce severely biased results. Overfitting relates to a situation that extraneous variables are involved in the model. In addition to the unnecessarily complicate form, overfitting a model might lead to high variability. To avoid model misspecification, the determination of a suitable model structure is crucial. The following example best demonstrates the necessity of model selection in terms of inferences and predictions.

The body fat data (Kutner *et al.*, 2004) is obtained from 20 healthy females 25-34 years old, and is utilized to examine the relation of amount of body fat (Y) to three possible predictors. The possible predictors are triceps skinfold thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). It would be helpful if a regression model with some or all of the predictors could yield reliable estimates and accurate predictions.

Table 1.1: Regression coefficients under different models for the body fat data.

Model	Predictor	Intercept	b_1	b_2	b_3
Model 1	X_1	-1.496	0.857	–	–
Model 2	X_1, X_2	-19.174	0.222	0.659	–
Model 3	X_1, X_2, X_3	117.085	4.334	-2.857	-2.186

Table 1.1 lists the regression coefficients for X_1, X_2, and X_3 under three possible models. We can see that the value of b_1, the estimated regression coefficient for X_1, varies significantly under different modes. Indeed, b_2, the estimated regression coefficient for X_2, even flips the sign under different models. In terms of prediction, if a new observation with $X_1 = 20$, $X_2 = 50$ and $X_3 = 30$ is sampled, the predicted values of amount of body fat under the three models are 15.644, 18.216, and -4.665, respectively. For the measurement of body fat, such three values are completely dissimilar, especially, a negative prediction value is unreasonable.

From the body fat data, the conclusion we can draw is that model choices may markedly affect regression coefficients and consequently may provide quite different predicted values. Thus, this

example illustrates that in order to have reliable estimation and accurate prediction, selecting an appropriate model is of importance.

The second target of model selection is to provide faster and more cost-effective models. As one of the integral parts of the statistical methodology application, model selection should always be driven by the consideration of cost, looking for the subset of predictors resulting in the best compromise between the performance of the model and cost of the analysis. More specifically, in the presence of a group of useful but highly correlated predictors, only one predictor should be retained in the model. With large data sets, computation time for model selection should also be taken into account. Computational efficiency is another indicator for a good model selection approach. A desirable model can trade off a small decrease in performance for a large reduction in cost or time. Focusing on the cost, many new approaches, such as the LASSO (Tibshirani, 1996) and LARS (Efron *et al.*, 2004), are proposed to select an econometric model from a large candidate set. These novel strategies can simultaneously select and estimate the significant predictors to construct a model that is not only good at prediction but also cost efficient.

The third target of model selection is to build interpretable models that can present easy and clear results for describing the data. On one hand, if many unimportant variables are selected, the model will lose its predictive power and the results will be difficult to interpret. On the other hand, parsimonious and compact representations of the data allow a better understanding of the underlying process that generated the data. For example, in the genome sequencing initiative, the number of genes in the raw microarray data ranges from 6000 to 60000, which challenges data analysis and interpretation tasks. The interpretation of these data and the structuring in form of compact models are vital to access these data in an effective manner (e.g., Lee *et al.*, 2003; Baragatti, 2011). In practice, an advantageous model selection procedure should select variables consistently and result in a succinct structure.

Over the last two decades, model selection has received increasing attention, motivated by the desire to understand structure in massive data sets that are now frequently confronted across many applications. For example, insurance companies have gathered a vast amount of data in their data

warehouses. When actuaries build a predictive model, they are encountered datasets involving thousands of variables. With many variables of interest, there is a high potential that the model efficiency is reduced. First of all, we could easily overfit the data when a lot of predictors are highly correlated, then the capability of parameter estimation of the model is lessened. Furthermore, it is harder to have an explainable model when there are many redundant predictors. A simpler selected model is much easier understood and interpreted than a complex one. Finally, creating models with all possible predictors is exhausted, it even would take indefinite time when there are thousands of predictors. There exists a substantial literature devoted to model selection problems for big data, and a comprehensive overview can be found in Fan and Lv (2010).

With the advances in high speed computers and computing technologies, it is becoming more and more feasible to apply model selection methods for larger datasets and richer classes of models. For example, the bootstrap is a resampling statistical technique which is widely applicable and allows the treatment of more realistic models. It involves a relatively simple procedure, but it repeated so many times that bootstrap method is intensively dependent on computer calculation. The idea of applying the bootstrap to improve the performance of model selection was introduced by Efron (1983, 1986), and was exhaustively discussed by Efron and Tibshirani (1993, pp. 237-253). Shang and Cavanaugh (2008) advocated two bootstrap-corrected variants of the Akaike information criterion for the purpose of small-sample mixed model selection. Pan and Le (2001) proposed a bootstrap approach to estimating the predictive mean squared error (PMSE) and then used the PMSE for model selection in generalized linear models. Markov chain Monte Carlo (MCMC) is another computing statistical technique that has been widely used in modern statistics. Since the presence of Green (1995), there has been great effort to use MCMC for model selection problems. For instance, Carlin and Chib (1995) proposed an approach to Bayesian model selection based on the use of conventional Markov chains. A summary of the theories and examples of model selection by MCMC computation can be found in Andrieu *et al.* (2001).

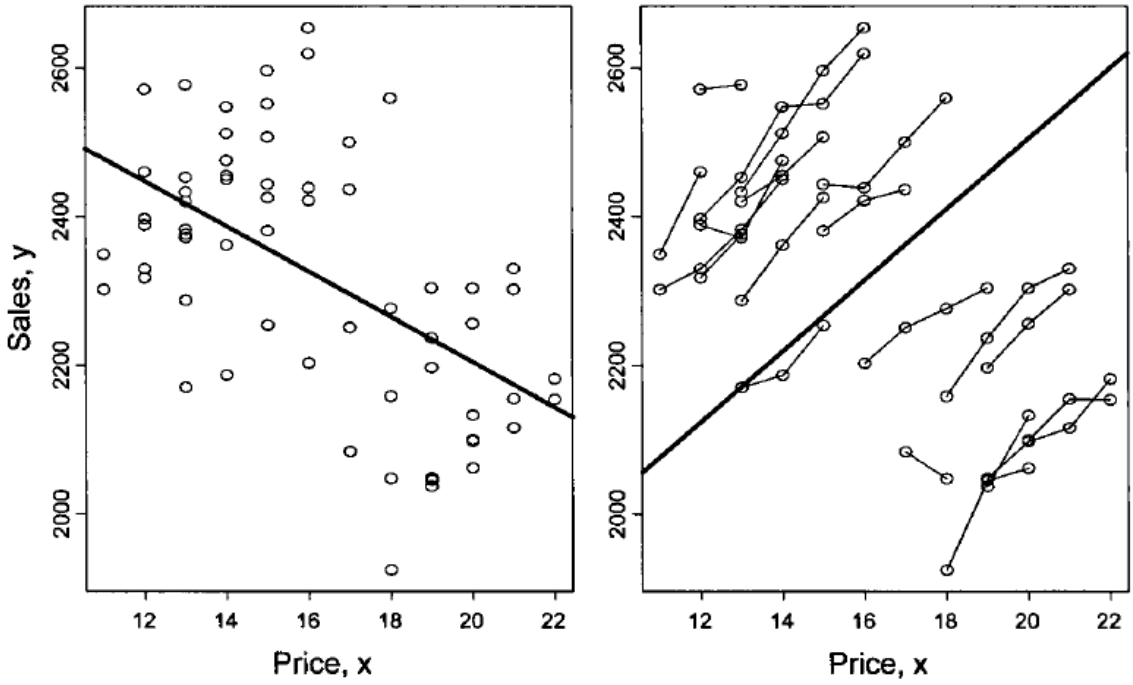
There is an extensive development on model selection in a wide range of data sets over the years. One popular area of investigation is concerned with model selection that incorporates cor-

related data which occurs frequently in the fields of agriculture, biology, economics, medicine, and sociology. For correlated data, the correlation may arise from clustering of subjects. For example, in social surveys, respondents are often clustered under blocks, neighborhoods or other geographical regions. Individuals in the same cluster tend to be more similar than those in different clusters. Members of the same block tend to have similar political views; residents of the same county tend to have similar opinions of the quality of medical care. Therefore, it is reasonable to assume that observations within the same cluster are correlated. The correlation may also be due to repeated measurements on the same subject over time or space. For instance, in clinical research we are often able to take several measurements on the same patient, and the repeated observations are usually (positively) correlated. To choose the best model for correctly analyzing the data, the correlation needs to be acknowledged and taken into account.

A linear mixed model (Laird and Ware, 1982) provides a general and flexible tool for the analysis of correlated data. The term of mixed model refers to the use of both fixed and random effects in the same analysis. Fixed effects have levels that are of primary interest and are defined by differences from the population mean. For example, if a pharmaceutical company focused on the effectiveness of two brands of a medicine, brand would be a fixed effect if the experimenter's concern did not go beyond the two medicine brands. Random effects, on the other hand, are utilized to account for the correlation in the data. The levels of random effects are assumed to be chosen from a population with a distribution having a certain variance. As a consequence, the outputs for random effects are estimates of the variances and rather than differences from a mean. Examples of random effects include regions in a multi-site experiment, and classrooms in an educational research. By allowing the correlation, or the variance-covariance structure to be explicitly modeled, linear mixed models are well suited for modeling data with correlated structure, such as longitudinal data and clustered data.

To illustrate that a mixed model captures the nature of correlated data, we can consider a study of exploring the relationship between price (x) and sales (y). In this study, a sample of observations was collected on price and sales for several commodities (Demidenko, 2013). Assuming that the

Figure 1.1: Linear regression model (left) and mixed model (right) lead to reverse conclusions.



observations are independent, the left hand panel in Figure 1.1 reveals a negative relationship by the classical simple linear regression, and the straight line shows simple regression estimated by ordinary least squares. However, one may argue that each commodity represents a cluster, thus the data has a clustered structure. Using a linear mixed model, observations for each commodity are connected and a positive relationship between prices and sales is obtained, as shown in the right hand panel in Figure 1.1. The straight line shows that the mixed model with population averaged slope and commodity specific intercept. As we can see, the linear regression model and the linear mixed model lead to completely reverse conclusions about the regression relationship between price and sales of the commodities. Apparently, the results obtained from the linear mixed model are more trustworthy, since it is reasonable to assume that the sales of each commodity are correlated.

This example demonstrates that ignoring dependent structure of correlated data may result in false analysis. By including subject-specific random effects in the regression model to account for within-subject dependency, the linear mixed model provides an effective and flexible way of

describing the means as well as the covariance structure of data, therefore has been applied in a variety of disciplines including health research and political science.

With the comprehensive applicability of linear mixed models in practice, determining an appropriate structural form for a model to be used in making inferences and predictions has been a fundamental problem in the analysis of longitudinal or clustered data. However, the selection of mixed model is challenging due to the complex nature of the model. In mixed model selection, not only the proper mean structure but also the correct covariance structure should be identified, yet both have distinct properties and different relative importance. Particularly, selection on the covariance structure is not straightforward due to computational issues and boundary problems arising from positive semidefinite constraints on covariance matrices (Müller *et al.*, 2013). In addition, the selection of only important covariates to create an interpretable model become challenging as the dimension of fixed or random effects increases. To choose the most proper model for correlated data, this dissertation focuses on the selection and estimation of the fixed and random effects in linear mixed models. We believe that investigating mixed model selection techniques not only serves as an appealing exploration in statistical modeling, but also is a very applicative subject that can be utilized to a broad range of data in real life.

1.2 Objectives of Dissertation

To facilitate the mixed model selection, the traditional information criteria such as Akaike information criterion (AIC, Akaike, 1973, 1974), Bayesian information criterion (BIC, Schwarz, 1978), Generalized information criterion (GIC, Rao and Wu, 1989), and Mallows' Cp (Mallows, 1973, 1995) have been utilized. In general, they are applied by finding the model which has the minimum score among the fitted candidate models. However, the estimators based on these selection procedures suffer from lack of stability because of the inherited discreteness (Breiman, 1996). Meanwhile, these selection procedures have to select the most appropriate model from a candidate family where all the reasonable models are considered, yet the number of candidate models increases exponentially with the number of predictors, hence it is computationally infeasible when the number of candidate predictors is large. Some other criteria, such as the extended GIC (EGIC,

Pu and Niu, 2006) and the restricted information criterion (RIC, Wolfinger, 1993), have been proposed to reduce computation cost. By first selecting on either fixed effects or random effects while fixing the other at full model, the number of possible models is considerably reduced, but may still be large. Instead of trying to find an optimal model that minimizes a criterion function, Jiang *et al.* (2008) proposed a fence method by constructing a statistical fence, or barrier, to carefully eliminate incorrect models. Although the fence algorithm does not search over all the candidate models, it is computationally very demanding.

The penalized methods (also known as shrinkage methods) have been introduced for model selection. Along with the LASSO method, brought into the field of selection literature by Tibshirani in 1996, the SCAD (Fan and Li, 2001) and the adaptive LASSO (Zou, 2006) developed as the seeds of growing and maturing the new idea in model selection. Through continuously shrinking the coefficients of certain predictors in the model toward zero, and also because of computational feasibility and statistical precision, the penalized methods hold both selection and estimation in one procedure and usually with the sacrifice of the estimation accuracy to improve the estimation precision, they are therefore turning out to be of more interests for statisticians.

Starting from the linear regression setting, further development has been rapidly extended to linear mixed models. Some papers have applied the penalized methods for selecting the fixed effects (e.g., Foster *et al.*, 2007, Ni *et al.*, 2010, Schelldorfer *et al.*, 2011). Some papers have discussed the random effects selection (e.g., Ahn *et al.*, 2012, Pan and Huang, 2014). Meanwhile, using the penalized methods to jointly select and estimate both fixed and random effects has also received much attention. Bondell *et al.* (2010) and Ibrahim *et al.* (2011) respectively proposed a joint selection procedure, although the differences existed in incorporating tuning parameters, both methods utilized the penalized maximum likelihood (ML) and applied EM algorithm to estimate the parameters. Lin *et al.* (2013) employed a two-step penalized method based upon the restricted maximized likelihood (REML) and pathwise coordinate optimization for the random and fixed effects selection. Peng and Lu (2012) adopted an iterative method without the restriction of distributions to perform model selection.

The main objective of this dissertation is to develop methodologies for effectively selecting the proper predictors and the correct covariance structure by integrating the recent advances in mixed model selection. More specifically, we propose to use the penalized approach to select and estimate fixed and random effects. Compared with the traditional selection procedures, the novel penalized methods can enhance the predictive power of a model, and can significantly reduce computational cost when the number of fixed or random effects increases. We wish our proposed procedure can further improve the behavior of the existing ones in mixed model selection from both theoretical and practical perspectives.

First of all, we aim to introduce a selection method that can reflect and accommodate the distinct properties between the random effects and fixed effects. Here, fixed effects refer to those with coefficients that affect the population mean, and random effects refer to those whose coefficients vary among subjects. The natures of these parameters are quite dissimilar, jointly selecting both the fixed and random effects would seem unnatural. Therefore, we propose a two-stage procedure for separately choosing the two types of effects. In the first stage, we choose the random effects by excluding all the fixed effects. After the completion of the random effects selection, in the second stage, our procedure for selecting the fixed effects only involves the regression coefficients. By discretely selecting the parameters of interests in the two stages, the proposed method can successfully respect the different features between the random effects and fixed effects, and therefore is effective in identifying the important covariates. Moreover, the dimension in each of the two stages is lower than the combined dimension of both fixed and random effects, and this makes the computation more effective and steady. Notice that we choose the random effects first, then select the fixed effects, that is because the ultimate object of linear mixed models is to describe the relationship between a response and the fixed effects. When an appropriate model for the covariance is adopted, the correct covariance structure will be obtained and valid inferences for the fixed effects can then be made.

Our second objective is to improve selection accuracy and computational efficiency. We address this goal by using the penalized profile log-likelihoods to choose both random and fixed

effects. In the first stage, the penalized restricted profile log-likelihood is utilized to choose the random effects; in the second stage, after the random effects are determined, the penalized profile log-likelihood is applied to select the fixed effects. Compared with other log-likelihood-based approaches, the profile log-likelihoods not only possess the theoretical advantages including robustness and unbiasedness, but also possess the more efficient and stable computation due to a smaller number of parameters involved.

In linear mixed models, a challenging problem is that both fixed and random effects are included in the estimation so that there is no closed form for the solutions of either parts. Accordingly, choosing a suitable numerical method to maximize the targeted quantity is of importance. In each stage, we apply the Newton-Raphson algorithm to implement parameter estimation. While comparing with the Expectation-Maximization and the other well-known optimization algorithms, the Newton-Raphson algorithm we adopt converges steadfast and speedy with a good initial value.

To assist the mixed model selection, we employ the adaptive LASSO penalized term to individually penalize the restricted profile log-likelihood and profile log-likelihood. The adaptive LASSO is computationally appealing due to its concave form, that is, the absolute maximizer can be efficiently solved without suffering from the multiple local maximal issue.

To evaluate the effectiveness of the proposed selection procedure, we conduct various simulation studies, and we compare the results with those for the existing selection approaches. We measure the performance of model selection with regard to correct selection frequencies, computation times, and three model accuracy measurements including the Kullback-Leibler discrepancy, the mean square error, and the quadratic loss error. We further illustrate the proposed procedures via two real data examples.

Our third goal is to provide a theoretical foundation for the proposed selection approach. We systematically study the sampling properties of the resulting estimate of both random and fixed effects. We establish conditions on the asymptotic analysis and show that the resulting estimate enjoys estimation consistency and model selection oracle properties, indicating that asymptotically the right covariance structure and predictors are surely selected. The proofs theoretically solidify

the promising performance of the proposed method and establish theoretical contribution in model selection.

1.3 Outline of Dissertation

The rest of this paper is organized as follows. Chapter 2 gives background materials that serve as a foundation for the remainder of the dissertation. It includes a discussion of two major classes of model selection approaches in linear regression models, along with an overview of selection criteria that are used to choose the optimal model. To further improve the existing selection procedures, we propose an adaptive penalty procedure with weighted ridge estimator at the end of this chapter, and we conduct two simulation studies to examine the performance of the proposed procedure.

Chapter 3 describes of the model selection framework in linear mixed models. We define the notations of linear mixed models, which are consistently used in the remaining parts of the dissertation. Some existing mixed model selection methods are introduced, including likelihood based approaches and distribution free procedures.

In Chapter 4, we employ the adaptive LASSO penalized term to propose a two-stage model selection procedure for the purpose of selecting both the random and fixed effects. In the first stage, we utilize the penalized restricted profile log-likelihood to choose the random effects; in the second stage, after the random effects are determined, we apply the penalized profile log-likelihood to select the fixed effects. In each stage, the Newton-Raphson algorithm is performed to complete the parameter estimation. We prove that the proposed procedure is consistent and possesses the oracle properties, indicating that asymptotically the proposed procedure surely selects the true model. Since the performance of penalized methods highly relies on the tuning parameters, we propose three tuning parameter candidates to be used for balancing between model fitting and model complicity.

Numerical experiments are conducted in Chapter 5 and 6. In Chapter 5, we illustrate the effectiveness of the proposed procedure via numerous simulation studies, and we compare the results with those for the existing selection approaches. In Chapter 6, two real applications are

presented to further investigate the performance of the proposed method.

We conclude in Chapter 7 with an overall discussion of our proposed model selection procedure in the mixed model. Some future research plans are provided in this chapter as well.

Selected R programs of the simulation studies are attached in the Appendix.

CHAPTER 2 MODEL SELECTION IN LINEAR REGRESSION MODELS

Linear regression is one of the most common data analysis techniques for modeling the relationship between a response variable and a set of predictors. A key part in the regression analysis of data is model selection. Over the years quite a number of selection techniques have been proposed in the setting of linear regression models. Linear regression models can be viewed as special cases of linear mixed models, so methods proposed for selecting linear regression models are helpful to exploit approaches in mixed model selection, and actually some model selection methodologies in linear mixed models are initiated from linear regression models.

In this chapter, we will review model selection methods and criteria in linear regression models. These reviews give the background knowledge that serves as a foundation for the remaining chapters of the dissertation. Additionally, to further improve the behavior of the existing selection procedures, we will propose a two-stage adaptive penalty approach with weighted ridge estimator in the last section.

Given the dataset of N observations, a linear regression model takes the form of

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, 2, \ldots, N, \tag{2.1}$$

where y_i is the response in the ith trial, $x_{i1}, \ldots, x_{ip}$ are the predictors, and ϵ_i is the error term. Typically, we assume the error terms are independent of the predictors, and are normally distributed with zero mean and constant variance σ^2. $\beta_1, \ldots, \beta_p$ are the regression coefficients, and statistical estimation and inference in linear regression focuses on these coefficients.

Alternatively, in matrix form, model (2.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}.$$

For simplicity, we assume the variance σ^2 is known, then for model (2.2), the log-likelihood function, ignoring constant terms, is given by

$$\ell(\boldsymbol{\beta}) = -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.3}$$

In general, the methods of maximum likelihood (ML) and least squares (LS) are utilized to estimate the regression coefficients. By maximizing the log-likelihood function in (2.3) with respect to $\boldsymbol{\beta}$, the former method of estimation defines a maximum likelihood estimator (MLE) as

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.4}$$

For the method of least squares, we define the residual sum of squares

$$RSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.5}$$

The method chooses $\hat{\boldsymbol{\beta}}$ to minimize the RSS. Note that maximizing $\ell(\boldsymbol{\beta})$ in function (2.3) is equivalent to minimizing the RSS in (2.5), thus, under the typical normal error assumption, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ and the least squares estimator (LSE) $\hat{\boldsymbol{\beta}}_{\text{LSE}}$ are the same.

In the practice of model fitting, the ML and LS techniques are widely used due to their ease of implementation. However, both the MLE and LSE suffer from large variance, then resulting in poor predictions on future observations. Moreover, they are not able to discover an important

subset from a large number of predictors in the model, while including all the predictors leads to model selection unachievable. As a result, alternative procedures have been proposed to improve prediction accuracy and to accomplish the purpose of model selection. Two major classes of model selection methods: subset selection and penalized selection, will be introduced in Section 2.1 and 2.2.

2.1 Subset Selection

Generally, subset selection is the process of selecting a subset of relevant predictors for use in model construction. In this section, we will go over the procedures that are in common use.

Best Subset Selection

The idea of best subset selection is to fit a separate least squares regression for each possible combination of the p predictors. That is, we fit all models that contain one predictor, all models that contain two predictors, and so forth. We then compare all candidate models, and choose the best model by using one of the model selection criteria, which will be described in Section 2.3.

While best subset selection is a simple and useful approach, it suffers from computational burden. There exist 2^p candidate models if we have p predictors. As p increases, the number of possible models grows rapidly. In general, best subset selection becomes infeasible when the number of predictors is greater than 30. Furthermore, it tends to overfit a model with redundant variables, and the final model would be very unstable. We present three computationally efficient surrogates to best subset selection next.

2.1.1 Forward Selection

Forward selection begins with no predictors in the model, then predictors are added to the model one at a time. At each step, each predictor that is not already in the model is tested for inclusion in the model. The most significant of these predictors is added to the model. The procedure is continued until no predictor is significant at a pre-set level.

Unlike best subset selection which includes fitting 2^p models, forward selection contains $1+\frac{p(p+1)}{2}$ models and therefore owns computational advantage over best subset selection. For instance, when

$p = 10$, best subset selection requires fitting 1024 models, whereas forward selection searches through only 56 models.

2.1.2 Backward Elimination

As a reverse process of forward selection, backward elimination starts with all predictors in the model, then iteratively removes the least useful predictor one at a time, and continues until every remaining variable is significant at a cut-off level.

Like forward selection, backward elimination requires fitting only $1+\frac{p(p+1)}{2}$ models, so it provides another efficient alternative to best subset selection. However, there is no guarantee that backward elimination and forward selection will arrive at the same final model.

2.1.3 Stepwise Selection

The stepwise selection can be considered as a hybrid approach of forward selection and backward elimination. Analogous to forward selection, predictors are added to the model sequentially in stepwise selection. However, after adding each new predictor, the method may also remove any predictors that no longer significant at some level. Such an approach intends to imitate best subset selection while holding the computational advantages of forward selection and backward elimination.

As we can see, the subset selection methods described in this section are conceptually appealing and easy to perform. Nevertheless, subset selection suffers from unstable selection results and highly variable due to the innate discreetness (Breiman, 1996; Fan and Li, 2001); that is, predictors are either retained or discarded from the model. Slight changes in the data may result in completely different models and it inhibits prediction accuracy. When we have correlated predictors or a large number of predictors (or both), the instability of subset selection could be even more problematic (Harrell, 2001). In the next section, we will review penalized selection methods which are proposed to address the weaknesses of subset selection.

2.2 Penalized Selection

In contrast to subset selection methods, penalized approaches do not explicitly select the predictors, instead they maximize the likelihood function by using a penalty on the size of the regression coefficients. The penalty shrinks the coefficient estimates towards zero, and some small coefficients will become exactly zero, to reach the purpose of model selection, and that is why penalized approaches are also known as shrinkage methods. Indeed, penalized selection allows us to achieve the same objective as subset selection, but in a more stable, continuous, and computationally efficient fashion. In general, the penalized likelihood function takes the form of

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{\beta}\right) - \lambda * Pen\left(\boldsymbol{\beta}\right) \right\}, \tag{2.6}$$

where $\ell\left(\boldsymbol{\beta}\right)$ is the log-likelihood function in (2.3), $Pen\left(\boldsymbol{\beta}\right)$ is a penalty function which determines the type of shrinkage, and $\lambda \geq 0$ is a tuning parameter which controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. We can select the proper value of λ through the model selection criteria, which will be described in Section 2.3. In the following, we will introduce some renowned penalized methods.

2.2.1 Ridge Regression

The ridge regression (Hoerl and Kennard, 1970) estimate is defined by

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{\beta}\right) - \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{2.7}$$

The solution to the ridge regression problem is given by

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.8}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix.

The benefits of ridge regression are most striking in the presence of collinearity. The maximum

likelihood estimates are asymptotically unbiased estimators, but they may be far from the true values in small and moderate samples when the predictors are correlated. By trading off a small increase in bias for a large decrease in variance, ridge regression shrinks the estimates toward zero and provides a power tool to address the problem of collinearity in the data.

Even though the ridge regression shrinks coefficients continuously to zero and hence is a stable procedure, however, it is not proper for model selection because it does not set any coefficients exact zero. In the recent two decades, many other penalized approaches have been proposed to retain the good features of ridge regression but be able to produce parsimonious models.

2.2.2 LASSO

The LASSO, for "least absolute shrinkage and selection operator", was proposed by Tibshirani in 1996, which later on turned out to be the root of the growing tree of penalized model selection. The LASSO estimate is defined by

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{\beta}\right) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.9}$$

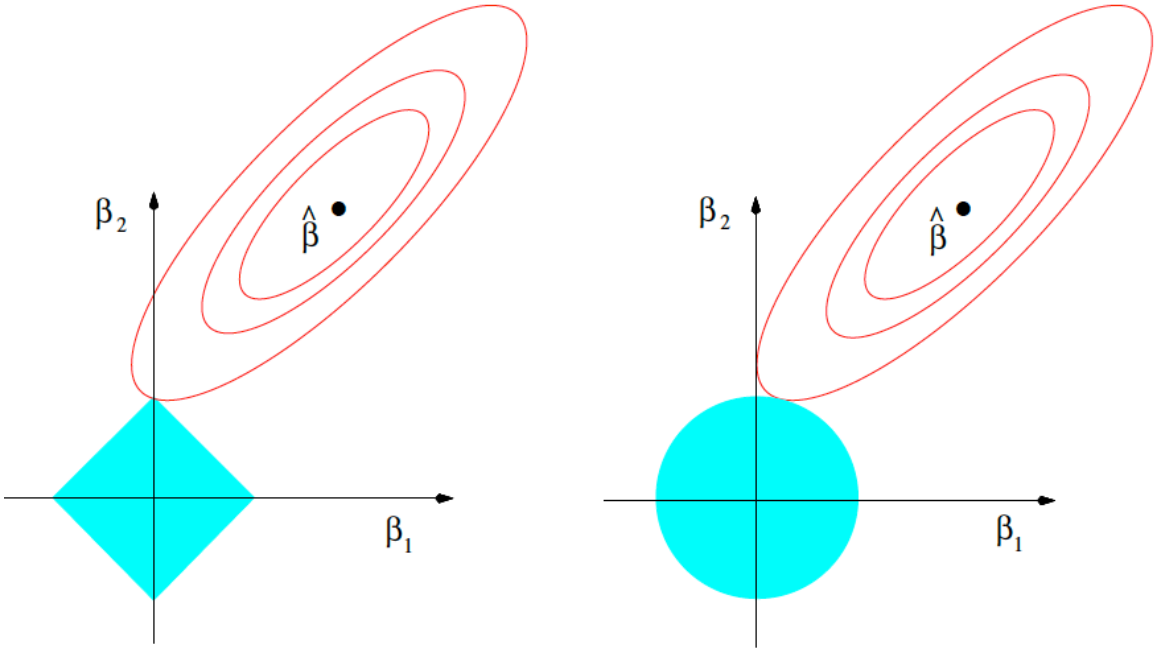
We can also write the LASSO problem (2.9) in the equivalent form

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \ell\left(\boldsymbol{\beta}\right), \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t,$$

where $t \geq 0$ is a tuning parameter.

Unlike ridge regression which involves a $\ell 2$ penalty $\sum_{j=1}^{p} \beta_j^2$, by adding a $\ell 1$ penalty $\sum_{j=1}^{p} |\beta_j|$ to the log-likelihood function, the LASSO forces some predictors to have zero as coefficients, inherently performing model selection. Figure 2.1 illustrates the difference between the LASSO and ridge regression when there are only two predictors in the model. The likelihood function has elliptical contours, centered at the maximum likelihood estimate. The constraint region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t$, while that for the LASSO is the diamond $|\beta_1| + |\beta_2| \leq t$. Both methods find the first point where the elliptical contours hit the constraint region. The diamond

Figure 2.1: Estimation picture for the LASSO (left) and ridge regression (right).



has corners, so the LASSO will shrink the coefficient to zero if the solution occurs at a corner. While there are no corners for the contours to hit the disk, so ridge regression can never shrink the coefficient to zero.

While the LASSO and its variants are very useful for model selection, the LASSO solution does not usually have a closed form expression. Various algorithms for the computation of the LASSO estimators have been studied. Fu (1998) proposed a shooting algorithm, which is straightforward and fast to solve the LASSO problems, but often proves to be too greedy in solution search. The least angle regression (LARS, Efron *et al.*, 2004) is another efficient algorithm and is crucial to the rapid spread of the LASSO within the statistics community. More recently, the coordinate descent algorithm (Friedman *et al.*, 2007; Wu and Lange, 2008) which updates the estimator in a coordinate-wise way until convergence is reached, has been proposed for rapidly solving the LASSO problems.

The LASSO has attracted a lot of attention because of its ability to yield sparse models. However, with high dimensional data, the LASSO is not satisfactory either, since it can not choose more predictors than the number of observations. Further, the LASSO fails to do group selection.

It inclines to select one predictor from a group and ignore the others. The other disadvantage of the LASSO is that it tends to shrink coefficients more than expected when predictors are highly correlated.

2.2.3 SCAD

Fan and Li (2001) argued that a good selection procedure δ should have the oracle properties, namely, the estimator $\hat{\boldsymbol{\beta}}(\delta)$ satisfies the following conditions,

1. Identifies the right subset model, $\{j : \hat{\beta}(\delta)_j \neq 0\} = \{j : \beta_j \neq 0\}$.

2. Has the optimal estimation rate, $\sqrt{n}\,(\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \Sigma)$, where Σ is the covariance matrix knowing the true subset model.

In other words, for an oracle procedure, the covariates with nonzero coefficients will be identified with probability tending to one, and the estimates of nonzero coefficients have the same asymptotic distribution as the true model.

It has been shown that the LASSO suffers from some drawbacks, due to the lack of oracle properties. To improve the performance of the LASSO, Fan and Li (2001) proposed an oracle selection method referred to as the "smoothly clipped absolute deviation" (SCAD). The idea of the SCAD is to penalize small coefficients heavily and large coefficients lightly, and its penalty function $p_{\mathrm{SCAD}}(\boldsymbol{\beta})$ is defined by

$$p_{\mathrm{SCAD}}(\boldsymbol{\beta}) = \begin{cases} \lambda|\boldsymbol{\beta}| & \text{if} |\boldsymbol{\beta}| \leq \lambda; \\ -\frac{|\boldsymbol{\beta}|^2 - 2a\lambda|\boldsymbol{\beta}| + \lambda^2}{2(a-1)} & \text{if} \lambda < |\boldsymbol{\beta}| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2} & \text{if} |\boldsymbol{\beta}| > a\lambda, \end{cases}$$

where $a > 2$ and $\lambda > 0$.

For the SCAD penalty, despite its good asymptotic properties, the corresponding optimization problem is non-concave, and as a result much harder to solve since there is no guarantee that the local maximum of the penalized likelihood is the global maximum. Additionally, the SCAD is computationally difficult due to its complex form.

2.2.4 Elastic Net

Zou and Hastie (2005) introduced an elastic net penalty which linearly combines the LASSO and ridge regression penalties. The elastic net estimate is given by

$$\hat{\boldsymbol{\beta}}_{\text{elastic}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{\beta}\right) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

By bridging the LASSO and ridge regression, the elastic net tries to maintain the good features of both methods. The LASSO penalty generates a parsimonious model, while the ridge regression penalty removes the limitation on the number of selected predictors, encourages group selection, and stabilizes the selection process. As a consequence, the elastic is particularly useful when there are more parameters than observations or there is a group of predictors that have high pairwise correlations.

2.2.5 Adaptive LASSO

Zou (2006) showed that the LASSO could be inconsistent in model selection and studied a necessary condition for the consistency. He also proposed the "adaptive LASSO", which is a development of the LASSO. The adaptive LASSO estimate is defined by

$$\hat{\boldsymbol{\beta}}_{\text{ALASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell\left(\boldsymbol{\beta}\right) - \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}, \tag{2.10}$$

where $\mathbf{w} = (w_1, \ldots, w_p)$ is a known weight vector chosen adaptively by the data. For the adaptive LASSO, the choice of the weights is very important, and it is often suggested that $\mathbf{w} = 1/|\hat{\boldsymbol{\beta}}|$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. By incorporating relatively larger penalties for insignificant predictors and smaller penalties for significant predictors, the adaptive LASSO attempts to reduce the estimation bias and improve model selection accuracy.

One advantage of the adaptive LASSO relies on its a concave optimization property, thus the absolute maximizer can be efficiently solved without suffering from the multiple local maximal issue. Further, if the weights are cleverly chosen, Zou (2006) showed that the method is selection

consistent and processes the oracle properties. Compared to other oracle procedures such as the SCAD, the adaptive LASSO is computationally more attractive, since its entire solution path can be obtained effectively.

Considering all the optimal properties of the adaptive LASSO in linear regression models, we employ it as the penalty term in our proposed procedure later in Chapter 4, and we wish it is also efficient in mixed model selection.

2.3 Selection Criteria

Just as the subset selection approaches considered in Section 2.1 require a criterion to determine which of the models under consideration is the best, implementing penalized procedures in Section 2.2 desires a rule for properly selecting a value for the tuning parameter in function (2.6) among the candidate values. A model selection criterion can be used to assign scores to each of the fitted candidate models in order to assist the analyst in choosing the best model. We give a brief review on those widely used criteria in the following.

2.3.1 AIC

The Akaike information criterion (AIC, Akaike, 1973, 1974) is generally accepted as the first model selection criterion, and remains the most popular tool for model selection. It is derived as an estimator of the expected Kullback discrepancy between the fitted model and the truth. Along with BIC, which will be described right after, AIC belongs to the family of information criteria which are likelihood-based measures of model fit including a penalty for complexity. In general, the AIC is defined as

$$\mathrm{AIC} = -2\ell\left(\hat{\boldsymbol{\beta}}\right) + 2 * p,$$

where $\ell\left(\hat{\boldsymbol{\beta}}\right)$ is the log-likelihood function in (2.3) evaluated at the estimate $\hat{\boldsymbol{\beta}}$, and p is the number of estimated parameters. In a model selection application, the optimal fitted model is identified by the minimum value of AIC.

Originally justified in asymptotic situations, AIC is applicable in a broad array of modeling frameworks. However, in settings where the sample size is small, AIC may favor the overfitted

models, which reduces its effectiveness as a model selection criterion. To address the deficiency of AIC, the "corrected' AIC, AICc, has been proposed as

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{N-p-1}.$$

Initially suggested for linear regression by Sugiura (1978), AICc has been extended to a number of additional modeling frameworks (e.g., Hurvich *et al.*, 1990; Hurvich and Tsai, 1993; Azari *et al.*, 2006). Hurvich and Tsai (1989) demonstrated that AICc outperforms AIC as a selection criterion in small sample applications. However, because the derivation of AICc depends upon the form of the candidate model class, AICc is less generally applicable than AIC.

2.3.2 BIC

BIC, the Bayesian information criterion, was introduced by Schwarz (1978) as a competitor to AIC. BIC serves as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. By choosing the fitted candidate model corresponding to the minimum value of BIC, the candidate model has the highest Bayesian posterior probability is selected. BIC is formally defined as

$$\text{BIC} = -2\ell\left(\hat{\boldsymbol{\beta}}\right) + p * \log(N).$$

Note that for $N \geq 8$, $p * \log(N)$ exceeds $2 * p$, so in general BIC has a more stringent penalty term than AIC. Consequently, BIC tends to choose smaller models than AIC, and such difference in selected models might be especially noticeable in large sample applications. Theoretically, BIC is a consistent criterion for model selection, that is, asymptotically BIC selects the fitted candidate model having the correct structure with probability one.

Nevertheless, BIC is not without drawbacks. For example, BIC is not asymptotically efficient, namely, it will not asymptotically select the fitted candidate model which minimizes the mean squared error of prediction. Therefore, BIC may not be advocated if the primary goal of the

modeling application is predictive. Chen and Chen (1999) showed that BIC may perform poorly with a moderate sample size but a huge number of covariates. For other criticisms for BIC, see Weakliem (1999).

2.3.3 Mallows' Cp

The statistic Cp (Mallows, 1973, 1995) is designed to estimate the Gauss discrepancy between the true model and the candidate model. Similar to other discrepancy-based model selection criteria such as AIC, Cp consists of a goodness-of-fit term and a penalty, and it is given by

$$\mathrm{Cp} = \frac{RSS}{\sigma^2} - N + 2p,$$

where RSS is the residual sum of squares of the given model, p is the number of predictors in the model. For this criterion, we desire models with Cp close to or smaller than p.

Mallows (1973) noted that one advantage of using Cp is that it can be clearly plot, and therefore a simple plot of Cp versus p can be used to choose among models. One limitation with the Cp criterion is that we need to decide an estimate of σ^2 since it is usually unknown. Typically, the estimate is from the full model, but it may not be an appropriate estimation of σ^2 for all the fitted models. Moreover, the CP statistic can be affected by outliers, that may lead to deterioration of the quality of this criterion. Fujikoshi and Satoh (1997) proposed a modified variant of Mallows' Cp, which improves of Cp on selecting the correct models from the pool.

2.3.4 Cross-Validation

Cross-validation (CV) is a widespread strategy for evaluating and selecting models by randomly dividing data into K groups, or folds, of approximately equal size. For $k = 1, 2, \ldots, K$, we let the validation set be the kth fold of the data , and let the training set be the remaining $K - 1$ folds. We fit the model to the training set, and compute the prediction error of the fitted model with the validation set. For each model, the process is repeated K times so that each fold is used once to be the validation set. As a result, each model results in K estimates of the prediction error, and the CV is computed by averaging these values. Under this criterion, the best model is the one

with the smallest value of CV. For the purpose of balancing between variance and bias, 5 or 10 fold cross-validation are recommended (Breiman and Spector, 1992; Kohavi, 1995).

An apparent shortcoming of cross-validation is that we must have a large enough sample to enable it to be divided into K groups. If the sample size is small, cross-validation may yield quite unstable results. The computational cost is usually mentioned as the other drawback of CV. For a K fold cross-validation, each candidate model will have K estimates of the prediction error, such process is quite time consuming. To obviate the need for the extensive computations, Tibshirani and Tibshirani (2009) proposed a bias correction for the minimum error rate in cross-validation. Bernau *et al.* (2013) suggested another bias correction cross-validation method to lower the computational price.

2.3.5 Generalized Cross Validation

The generalized cross validation (GCV, Craven and Wahba, 1979) provides a modified form and computational shortcut for cross-validation. The GCV statistic is defined by

$$\mathrm{GCV} = \frac{1}{N}\frac{RSS}{[1-p/N]^2}.$$

Just as the other criteria, we choose the optimal model by minimizing the GCV value.

Equivalently, in linear regression models, the GCV can be also given as

$$\mathrm{GCV} = -\frac{1}{N}\frac{\ell\left(\hat{\boldsymbol{\beta}}\right)}{[1-p/N]^2}.$$

GCV alleviates the computational burden of CV and thus is more popular to be used as a model selection tool. However, Wang *et al.* (2007) indicated that GCV performs similar to AIC, and the resulting model selected by GCV tends to overfit. Consequently, it has been argued that choice between GCV and CV should be based upon statistical rather than computational grounds.

2.4 Adaptive Penalty with Weighted Ridge Estimator

In Section 2.2, we have introduced the adaptive LASSO in (2.10) as a reliable penalized methodology for simultaneous parameter estimation and model selection. For guaranteeing the optimality of the solution, the chosen values for the weight w_j's are important. In general, $\mathbf{w} = 1/|\hat{\boldsymbol{\beta}}_{\mathrm{ML}}|$ is utilized as the weight vector, where $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ defined in (2.4). It is known that $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ are consistent estimators, so their values well reflect the relative importance of the covariates.

When collinearity occurs, the variances of maximum likelihood estimates are large so they may be far from the truth. The ridge solution in (2.8), on the other hand, is often suggested as a remedy for estimator variance, therefore may be more appropriate for weights in the adaptive LASSO penalty.

To improve the performance of the adaptive LASSO in linear regression models, we propose a two-stage adaptive LASSO model selection procedure with weighted ridge estimator. In the first stage, we find the optimal ridge solution $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ in (2.8) via one of the model selection criteria discussed in Section 2.3. In the second stage, we select the best model via the adaptive LASSO procedure in (2.10) using the ridge solution obtained in the first stage as weights.

In what follows, we examine the performance of the proposed procedure under two simulation settings, and compare the simulated results with those for the LASSO and the adaptive LASSO with weighted maximum likelihood estimator. All of the simulated data are generated from model (2.2), and BIC is used as the turning parameter. The R code for the simulation studies are available in the appendix.

Example 2.1. In this example, we inspect the performance of the proposed method in a low-dimensional model with a few large effects. For the true model, we let $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and let the predictors $\mathbf{X}_i$ $(i = 1, \ldots, N)$ be i.i.d normal vectors. The pairwise correlation between x_{ij} and x_{ik} is 0.5, $j, k = 1, \ldots, p$. The datasets are generated under different scenarios: $N = 20, 60$, and $\sigma = 1, 3$.

Example 2.2. In this example, we investigate the performance of the proposed method in a

higher-dimensional model. For the true model, let $\boldsymbol{\beta} = (0, \ldots, 0, 2, \ldots, 2)$ with 20 repeats in each block, and let $x_{ij} = z_{ij} + z_i$, where z_{ij} and z_i are independent standard normal variates. The datasets are generated under different cases: $N = 60, 100$, and $\sigma = 1, 3$.

Table 2.1: Simulation results for Example 2.1.

Case	σ = 1			σ = 3		
	MRPE	MCS	MIS	MRPE	MCS	MIS
$N = 20$						
LASSO	0.34	4.82	1.82	0.31	5.02	2.32
ALASSOM	0.28	3.50	0.54	0.32	4.04	1.70
ALASSOR	0.27	3.42	0.44	0.29	3.98	1.56
$N = 60$						
LASSO	0.11	4.70	1.70	0.11	4.84	1.88
ALASSOM	0.07	3.22	0.22	0.11	3.86	0.92
ALASSOR	0.06	3.18	0.18	0.10	3.70	0.80

Table 2.2: Simulation results for Example 2.2.

Case	σ = 1			σ = 3		
	MRPE	MCS	MIS	MRPE	MCS	MIS
$N = 60$						
LASSO	0.61	29.46	9.46	0.61	30.20	10.22
ALASSOM	0.40	21.10	1.10	0.51	24.34	4.34
ALASSOR	0.39	20.86	0.86	0.51	23.62	3.64
$N = 100$						
LASSO	0.40	28.05	8.05	0.33	31.26	11.26
ALASSOM	0.23	20.35	0.35	0.25	22.84	2.84
ALASSOR	0.22	20.35	0.35	0.25	22.76	2.76

For each example, we generate 100 datasets for each combination of (N, σ), and measure the performance in terms of model prediction and selection accuracy using the proposed method (ALASSOR). For prediction accuracy, we calculate the relative prediction error (RPE) which takes the form of

$$\mathrm{RPE} = \mathrm{E}[(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^2]/\sigma^2.$$

Figure 2.2: Boxplots of RPE, CS and IS for (N, σ) = (20, 1) in Example 2.1.

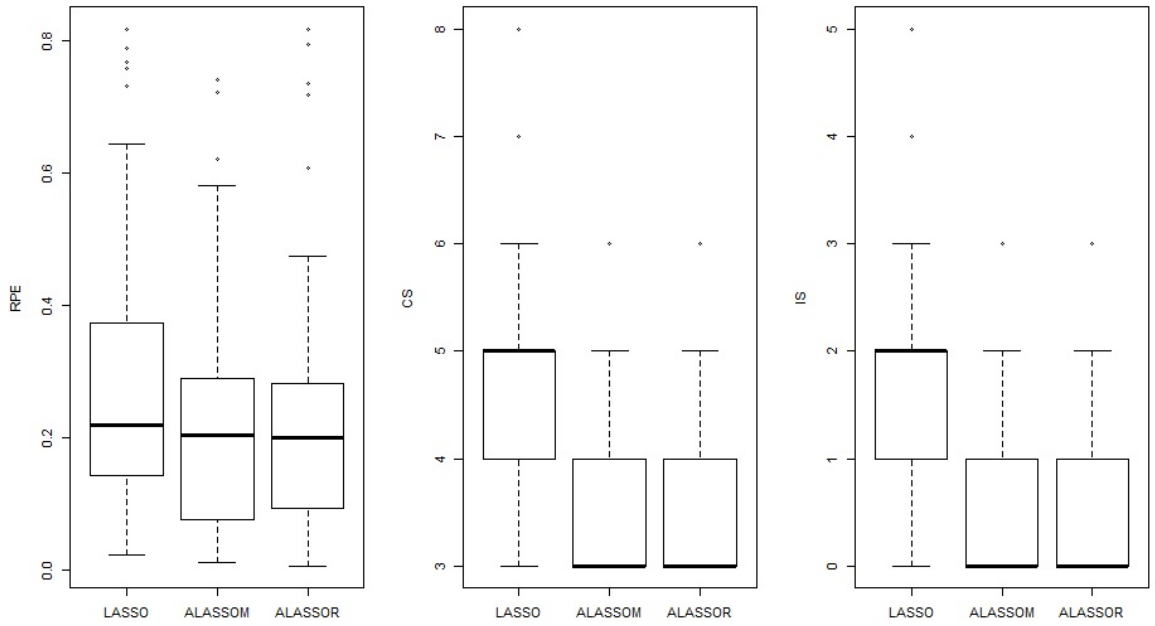
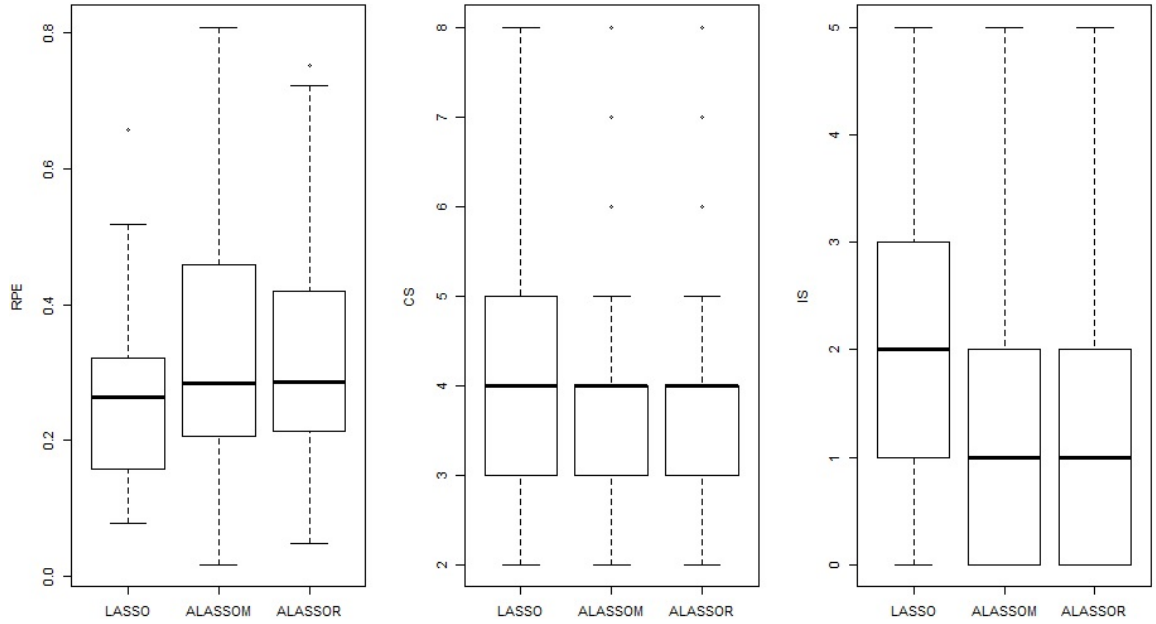


Figure 2.3: Boxplots of RPE, CS and IS for (N, σ) = (20, 3) in Example 2.1.



Small values of the RPE indicate that the fitted model is more accurate in predicting future data. For model selection accuracy, we calculate number of selected nonzero components (CS) and number of zero components incorrectly selected into model (IS). We expect the estimates of IS be close to 0, the estimates of CS be close to 3 and 20 in Example 2.1 and Example 2.2, respectively. Based on

Figure 2.4: Boxplots of RPE, CS and IS for (N, σ) = (60, 1) in Example 2.1.

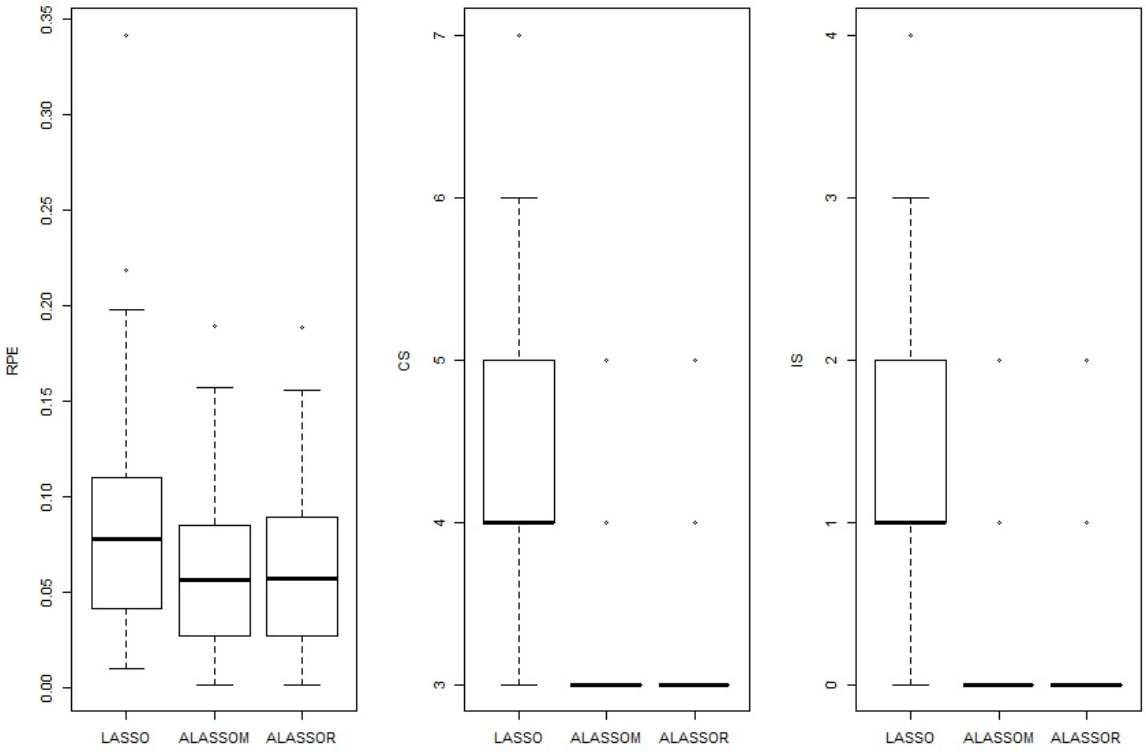
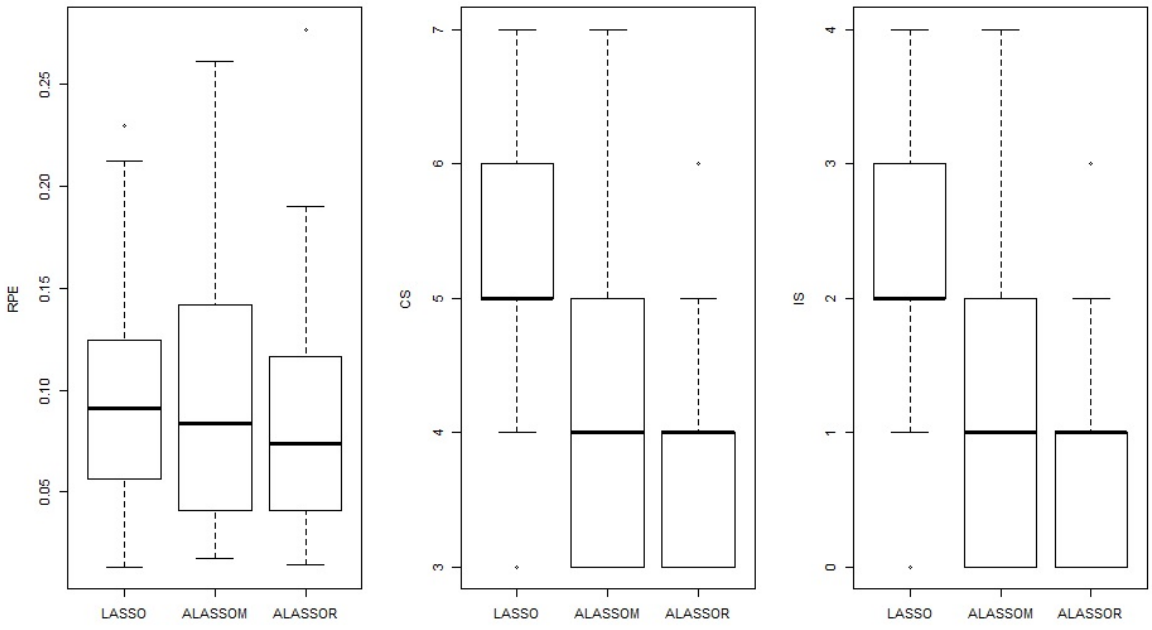
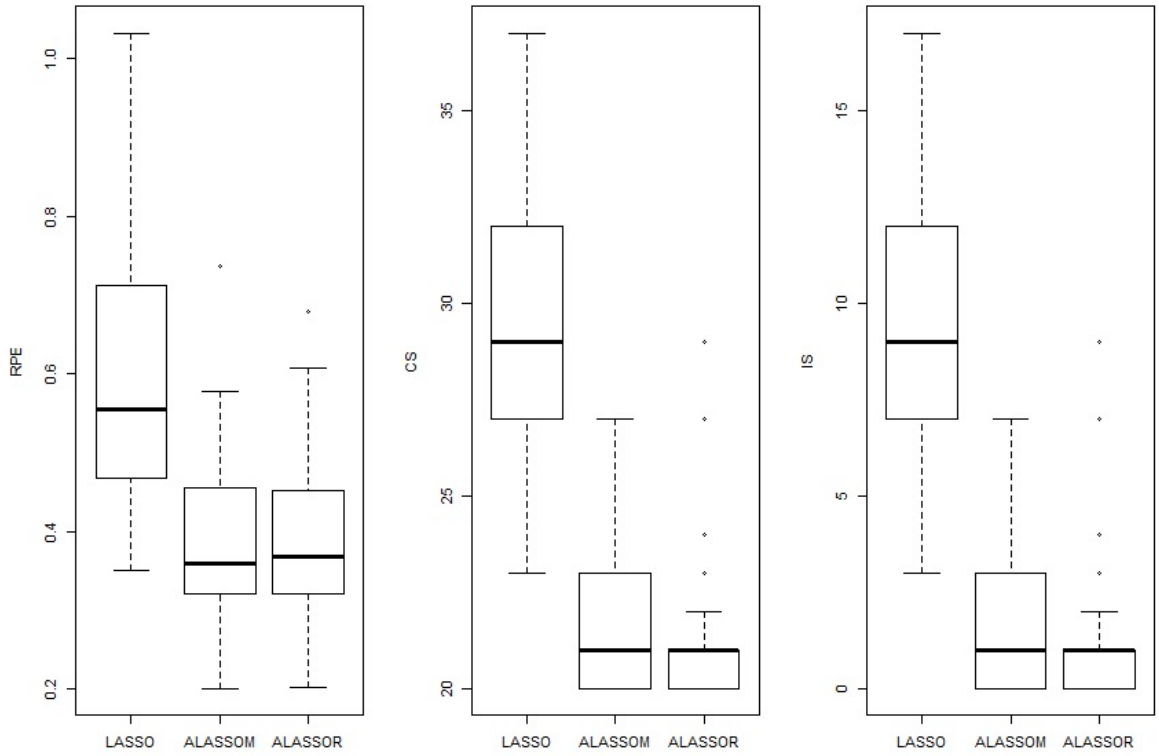


Figure 2.5: Boxplots of RPE, CS and IS for (N, σ) = (60, 3) in Example 2.1.



the 100 replications, we compute the mean of each of the three quantities and name them MRPE, MCS, and MIS, then we compare them with those for the LASSO and the adaptive LASSO with

Figure 2.6: Boxplots of RPE, CS and IS for (N, σ) = (60, 1) in Example 2.2.



weighted maximum likelihood estimator (ALASSOM).

Table 2.1 and Table 2.2 individually summarize the simulation results of Example 2.1 and Example 2.2. Although the two tables present the simulation results for the simulated data which are generated from different structures, three similar observations can be observed. First, all the three methods tend to perform better when the variance decreases. For example, when $N = 20$ in Table 2.1, as σ decreases from 3 to 1, the MRPE obtained from our method drops from 0.29 to 0.27, illustrating that the model prediction accuracy raises with less noise. Meanwhile, the MCS and MIS decline from 3.98 and 1.56 to 3.42 and 0.44, respectively, meaning that the model structure can be identified more accurately with smaller variance.

Second, as the sample size increases, each of all the methods can better identify the true model. For instance, when $\sigma = 3$ in Table 2.2, as sample size grows from 60 to 100, the MRPE, MCS, and MIS obtained from our method drop from 0.51, 23.62, and 3.64 to 0.25, 22.76, and 2.76, individually. It is reasonable that the behaviors of model selection methods are improved in larger

Figure 2.7: Boxplots of RPE, CS and IS for (N, σ) = (60, 3) in Example 2.2.

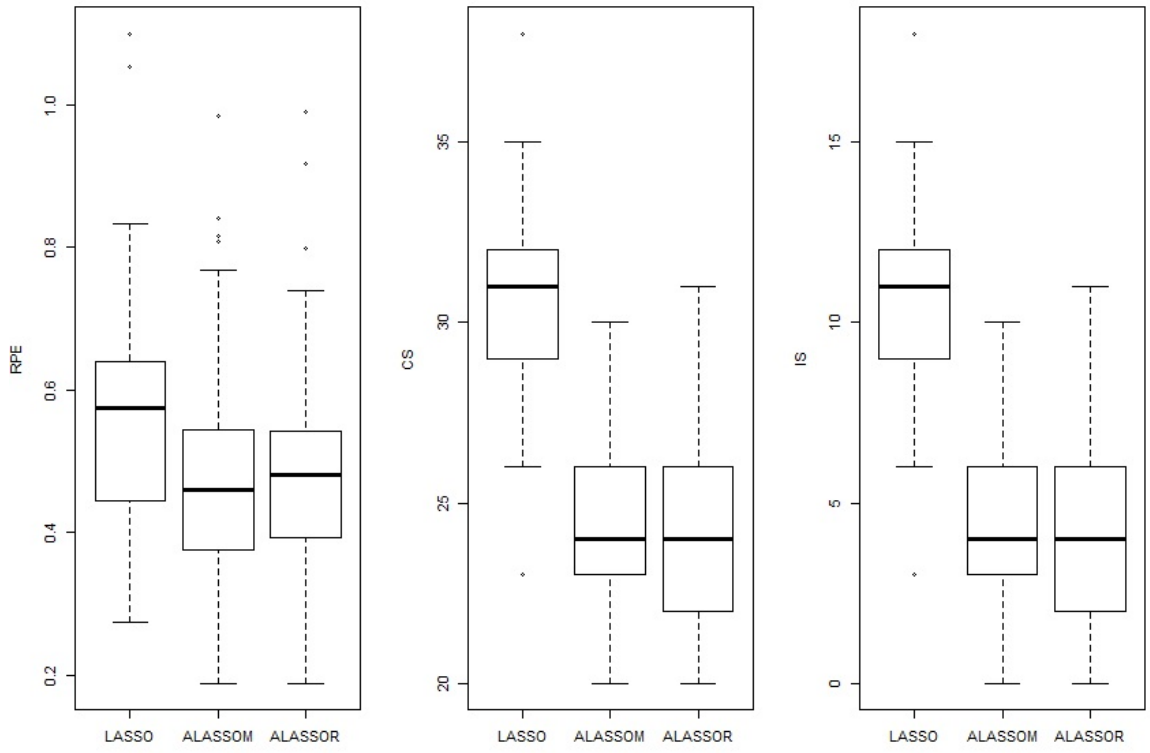


Figure 2.8: Boxplots of RPE, CS and IS for (N, σ) = (100, 1) in Example 2.2.

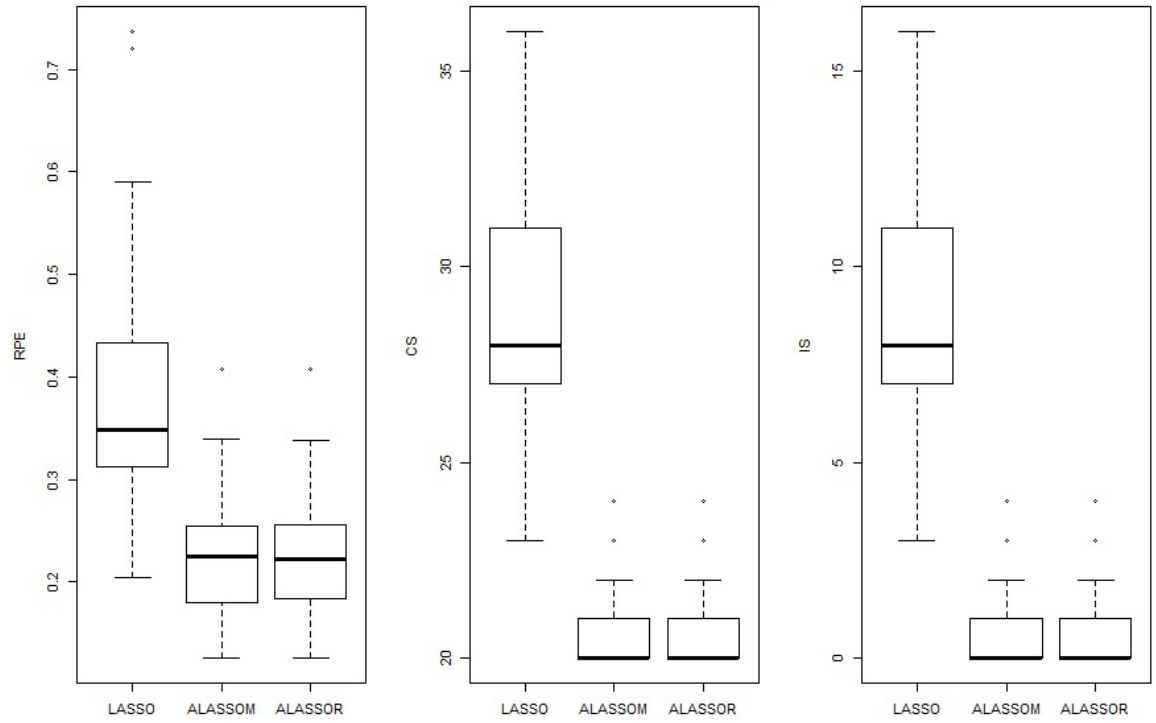
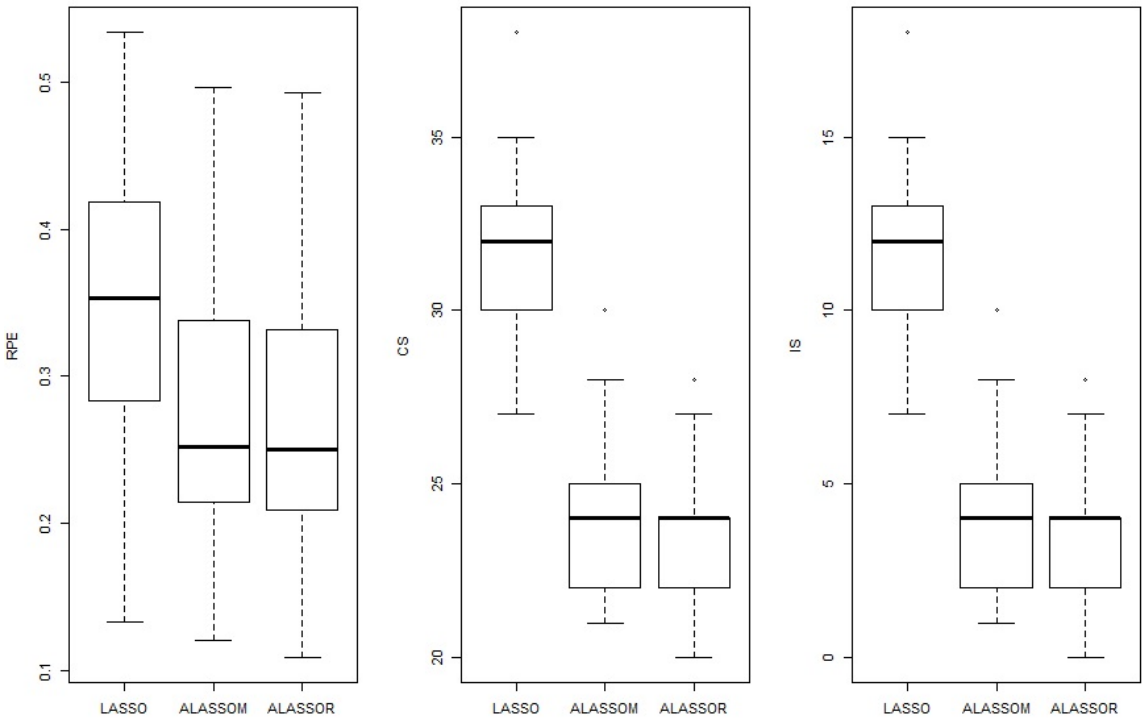


Figure 2.9: Boxplots of RPE, CS and IS for (N, σ) = (100, 3) in Example 2.2.



data sets, since increasing sample size produces more information about the true model.

Finally, the proposed method ALASSOR yields smaller MRPE, MCS and MIS values than the other two methods almost across all cases, ALASSOM is the follower, and the LASSO performs worst in general. In contract to the LASSO, both ALASSOR and ALASSOM are the adaptive LASSO procedures. By involving the weight vector in the penalty function, the adaptive LASSO imposes more penalties on insignificant predictors and less penalties on significant ones, yet the LASSO gives them the same amount of penalization, therefore it is not surprising that the adaptive LASSO procedures have better performance than the LASSO. While for the comparison of the two adaptive LASSO methods, our ALASSOR is superior than ALASSOM, since the values of ridge regression estimates better reflect the relative importance of the predictors than those of the maximum likelihood estimates as the predictors are correlated in both examples.

In addition to measuring the mean, we also graphically report the median, another middle value. Figure 2.2 - 2.9 are the box plots of the RPE, CS and IS values of all the three methods for different combinations of (N, σ) in Example 2.1 and 2.2, and the black horizontal line in each box

is the median. It is observed that ALASSOR produces smaller median values of RPE, CS and IS than the other two methods almost across all scenarios, demonstrating that the proposed approach has the smallest errors among the three procedures.

In conclusion, the proposed method improves the penalized methods especially the adaptive LASSO with regard to prediction and selection accuracy. By employing the ridge regression estimate in the process, the proposed adaptive LASSO procedure has a more proper weight vector and is therefore exceptionally effective in identifying the correct model when high correlations among the predictors are presented.

CHAPTER 3 MODEL SELECTION IN LINEAR MIXED MODELS

In Chapter 2, we have reviewed a variety of model selection methods and criteria in linear regression models, where the responses are assumed independent. Starting from the linear regression setting, we will extend model selection to linear mixed models, where the observations are dependent.

Typically, a linear mixed model contains both fixed effects and random effects. Fixed effects are the traditional linear regression coefficients, and random effects are associated with units which are chosen randomly from a population. By involving such two types of parameters, linear mixed models are primarily used to describe the regression relationship between a response variable and some possibly related covariates in the data that are grouped, and therefore have been extensively applied in a variety of disciplines including social sciences, medicine and biology (Demidenko, 2013; Jiang, 2007).

Because of the extensive applicability of linear mixed models, selecting the most appropriate model is of importance. As extensions of linear regression models, methods for mixed model selection can be recognized as expansions of methods developed for linear regression models. Nevertheless, model selection in linear mixed models is much more complicated than it in linear regression models, because both the fixed effects and the random effects need to be correctly identified, and the selection of informative covariates to construct an interpretable model is challenging as the number of fixed or random effects grows.

In this chapter, we will provide basic notations for a linear mixed model, which will be consistently used in the rest of the chapters. Furthermore, we will introduce some cutting-edge mixed model selection methodologies. In Chapter 4, we will propose a novel selection approach for linear mixed models by integrating the recent advances.

3.1 Model Setting and Notations

If we have $N = \sum_{i=1}^{n} n_i$ number of observations with n clusters, each of which has n_i measurements, where i refers to the ith cluster, a separate linear mixed model can be fitted to each cluster as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \ldots, n, \tag{3.1}$$

where $\mathbf{y}_i$ is an $n_i \times 1$ vector of responses for cluster i, $\boldsymbol{\beta}$ is a $p \times 1$ vector for fixed effects, the $n_i \times p$ matrix $\mathbf{X}_i$ is its associated design matrix and is assumed to be of full rank. The $n_i \times q$ design matrix $\mathbf{Z}_i$ is related with the $q \times 1$ vector of random effects $\mathbf{b}_i \sim \text{N}(0, \sigma^2\mathbf{D})$, and the matrix $\mathbf{D}$ is positive definite. The error term $\boldsymbol{\epsilon}_i \sim \text{N}\ (0, \sigma^2\mathbf{I}_{n_i})$, and $\boldsymbol{\epsilon}_i$ is independent with $\mathbf{b}_i$. Thus, $\mathbf{y}_i$ is distributed as $\text{N}\ (\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{V}_i(\boldsymbol{\theta}))$, and the matrix $\mathbf{V}_i(\boldsymbol{\theta}) = \mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$, where $\boldsymbol{\theta}$ denotes the vector consisting of $k = \frac{q(q+1)}{2}$ unique covariance parameters in $\mathbf{D}$. We use the notation $\mathbf{V}_i(\boldsymbol{\theta})$ to emphasize the dependence of $\mathbf{V}_i$ on $\boldsymbol{\theta}$. For the sake of brevity, we will often write $\mathbf{V}_i$ in short. In particular, a linear mixed model with no random effects reduces to the linear regression model in (2.2).

In linear mixed models, the magnitudes of the coefficients are concerned for the fixed effects. Conversely, for the random effects, researchers are interested in the distribution rather than the actual sizes of coefficients. Therefore, the aim of mixed model selection is to select and estimate the parameter vector $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$, which consists of the fixed effects and variance components of the random effects.

For parameter estimation in linear mixed models, maximum likelihood (ML, Hartley and Rao, 1967) and restricted maximum likelihood (REML, Thompson, 1962; Patterson and Thompson, 1971) are the two most commonly used techniques. For model (3.1), the log-likelihood function, ignoring constant terms, is given by

$$\ell_F\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma\right) = -\frac{1}{2}\sum_{i=1}^{n}\log\left|\sigma^2\mathbf{V}_i\right| - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i, \tag{3.2}$$

where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$.

Maximization of $\ell_F(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma)$ yields the ML estimators (MLE) of unknown parameters. When $\boldsymbol{\theta}$ is known, the MLE of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i \right). \tag{3.3}$$

In practice when $\boldsymbol{\theta}$ is unknown, $\mathbf{V}_i$ is simply replaced with its estimate, $\hat{\mathbf{V}}_i$. However, both fixed effects and variance components are involved in ML estimation so that there is no closed form solution of either part, and this causes computational challenges. In order to obtain the MLE, numerical methods such as Expectation-Maximization algorithm and Newton-Raphson algorithm are often desired. Moreover, ML method treats $\boldsymbol{\beta}$ as fixed but unknown parameters when $\boldsymbol{\theta}$ is estimated, but does not take into account the degrees of freedom lost by estimating the fixed effects, hence the MLE of the variance components $\boldsymbol{\theta}$ is biased.

REML estimation, on the other hand, is preferred when interest lies in accurate estimates of the variance components. Harville (1974) showed that the restricted log-likelihood function, dropping constant terms, is given by

$$\ell_R(\boldsymbol{\theta}, \sigma) = \ell_F\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \sigma\right) - \frac{1}{2} \log \left| \sigma^{-2} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right|, \tag{3.4}$$

where $\tilde{\boldsymbol{\beta}}$ is of the form given in (3.3).

Maximizing $\ell_R(\boldsymbol{\theta}, \sigma)$ produces the restricted maximum likelihood estimates of the variance components $\boldsymbol{\theta}$. Then we can obtain the REML estimator of $\mathbf{V}_i$, denoted as $\hat{\mathbf{V}}_{Ri}$, and the REML estimator of $\boldsymbol{\beta}$ as

$$\tilde{\boldsymbol{\beta}}_R = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{V}}_{Ri}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{V}}_{Ri}^{-1} \mathbf{y}_i \right).$$

Compared with ML method, REML is recommended to estimate the variance components in linear mixed models for several reasons. First, it accounts for the degrees of freedom lost by estimating the fixed effects, and makes a less biased estimation of variance components (Jiang,

2007). Second, REML estimators of $\boldsymbol{\theta}$ are invariant to the value of $\boldsymbol{\beta}$ and are more robust to outliers (Verbyla, 1993; McCulloch *et al.*, 2008). Third, the dimension involved in REML is lower than ML, so the computational costs are cheaper. However, one advantage of ML over REML is that it is able to compare two models with different fixed and random effects terms, while REML estimates only allow us to compare two models with identical fixed effects and are nested in their random effects terms.

To make full use of the log-likelihood functions' strength in estimation, quite a few mixed model selection procedures rely on either ML or REML, assuming the random effects and the error term follow normal distribution. Meanwhile, there exist some robust methods for non-normal data. We categorize the different methods into three broad classes and discuss each class in its own section.

3.2 Mixed Model Selection by ML

AIC is the most widely used model selection criterion, so no surprise that it is developed in the framework of linear mixed models. Sugiura (1978) proposed a marginal AIC (mAIC) which is derived by the marginal form of linear mixed models. By taking the covariance structure into account, it is defined as

$$\mathrm{mAIC} = -2\ell\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right) + 2a_N\left(p+q\right),$$

where p is the number of estimated fixed effects, q is the number of estimated random effects, $a_N = 1$ or $a_N = \frac{N}{N-p-q-1}$. However, Greven and Kneib (2010) showed that mAIC is positively biased for the marginal Akaike information.

Vaida and Blanchard (2005) showed that the general AIC is not appropriate for the linear mixed model, and they proposed instead the conditional AIC (cAIC) based on the conditional log-likelihood $\ell\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}|\hat{\mathbf{b}}\right)$. In general, cAIC is expressed as

$$\mathrm{cAIC} = -2\ell\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}|\hat{\mathbf{b}}\right) + a_N\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right),$$

where $\hat{\mathbf{b}}$ is often estimated by the best linear unbiased predictor (BLUP, Henderson, 1950)

$$\hat{\mathbf{b}} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathrm{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{3.5}$$

There are several versions of cAIC with different proposed penalty terms $a_N\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right)$. For example, Vaida and Blanchard (2005) suggested using $a_{N,VB}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right) = 2\left(\rho(\hat{\boldsymbol{\theta}}) + 1\right)$, while Burnham and White (2002) proposed using $a_{N,BW}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right) = 2\left(\rho(\hat{\boldsymbol{\theta}}) + q\right)$. Here, $\rho(\hat{\boldsymbol{\theta}})$ is the effective degrees of freedom used in estimating $\boldsymbol{\beta}$ and $\mathbf{b}$ (Hodges and Sargent, 2001).

As another most commonly used criterion, BIC in linear mixed models is obtained by taking mAIC and replacing $2a_N$ by $\log(N)$, so we have

$$\mathrm{mBIC} = -2\ell\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right) + \log(N)\,(p+q)\,.$$

Compared with mAIC, the increased weight in the penalty term should encourage mBIC to favor more parsimonious models. Jones (2011) proposed a variant of mBIC, which considers the effect of dependent structure and has an alternative measure of the effective sample size.

The generalized information criterion (GIC, Rao and Wu, 1989), is a generalization of AIC and BIC. Pu and Niu (2006) extended GIC to select linear mixed models with the form of

$$\mathrm{EGIC} = -2\ell\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}\right) + \lambda_N\,(p+q)\,.$$

This criterion allows for greater flexibility in choosing λ_N. Different choices of λ_N include $\lambda_N = 2$ for mAIC, $\lambda_N = \log(N)$ for mBIC, $\lambda_N = 2\log\log(N)$ for Hannan and Quinn (1979) penalty, and $\lambda_N = \log(N) + 1$ for Bozdogan (1987) penalty. Pu and Niu (2006) proved that, under mild conditions, the EGIC is consistency and asymptotic loss efficiency. They also suggested implementing EGIC into two stages, in the first stage, the fixed effects $\boldsymbol{\beta}$ are selected by fixing the random effects $\boldsymbol{\theta}$, then in the second stage, the random effects can be chosen after the fixed effects are selected.

We now review penalized methods in mixed models. While there is an extensive literature focusing on fixed effects or random effects selection, few references discuss about jointly selection on both fixed and random effects. Only recently, Bondell *et al.* (2010) and Ibrahim *et al.* (2011) respectively proposed a joint selection procedure, both of which used penalized maximum likelihood and Cholesky parameterizations.

Ibrahim *et al.* (2011) proposed to maximize the penalized maximum likelihood

$$\ell\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma\right) - n\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|) - n\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(||\gamma_k||),$$

where γ_k contains of all nonzero of the kth row of Γ, and Γ is the Cholesky factor of $\mathbf{D}$. The authors considered both the SCAD and the adaptive LASSO for the penalty functions. For instance, the adaptive LASSO penalties for fixed and random effects are individually defined as

$$\phi_{\lambda_j}(|\beta_j|) = \lambda_j \frac{|\beta_j|}{|\hat{\beta}_j|} \qquad j = 1, 2, \ldots, p,$$

$$\phi_{\lambda_{p+k}}(||\gamma_k||) = \lambda_{p+k} \frac{||\gamma_k||}{||\hat{\gamma}_k||} \qquad k = 1, 2, \ldots, q,$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are unpenalized maximum likelihood estimators.

Bondell *et al.* (2010) adopted a modified Cholesky decomposition by factorizing the covariance matrix of the random effects $\mathbf{D}$ as $\mathbf{D}^*\Gamma\Gamma^T\mathbf{D}^*$, where $\mathbf{D}^* = diag(d_1, \ldots, d_q)$ is a diagonal matrix, and Γ is a $q \times q$ lower triangular matrix. Then the vector $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^T, \mathbf{d}^T, \boldsymbol{\gamma}^T)^T$ was defined, where $\mathbf{d} = (d_1, \ldots, d_q)^T$, and $\boldsymbol{\gamma}$ is the vector consists the $\frac{q(q-1)}{2}$ elements of Γ. Thereafter the authors proposed to maximize the penalized maximum likelihood with the adaptive LASSO penalty terms. The penalized maximum likelihood is given by

$$\ell\left(\boldsymbol{\beta}, \mathbf{d}, \boldsymbol{\gamma}\right) - \lambda_n \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\hat{d}_j|} \right),$$

where λ_n is a tuning parameter, $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$, and $\hat{\mathbf{d}}$ can be

obtained by decomposing the estimated covariance matrix $\hat{\mathbf{D}}$.

Although the approaches of Bondell *et al.* (2010) and Ibrahim *et al.* (2011) have some contents in common, such as relying on the maximum likelihood function, using Cholesky decomposition, and adapting the EM algorithm to carry out parameter estimation, they are different in incorporating tuning parameters. Bondell *et al.* (2010) employed the same λ to penalize both the fixed and random effects, while Ibrahim *et al.* (2011) applied two sets of tuning parameters in the objective function. The use of the same λ in the scale of two kinds of parameters decreases computational burden in searching solutions, but compromises the proficiency of the method in identifying the true model. In the next section, we will go through REML based selection methods.

3.3 Mixed Model Selection by REML

The REML function in (3.4) is not a function of $\boldsymbol{\beta}$, so it seems to indicate that REML is not useful to select the fixed effects. However, selections of $\mathbf{X}$ in (3.4) associate with choices of $\boldsymbol{\beta}$, meaning that REML could be used in mixed model selection.

Vaida and Blanchard (2005) considered conditional AIC using REML. The criterion is of the form

$$\mathrm{cAIC}_R = -2\ell_R\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}|\hat{\mathbf{b}}\right) + a_N\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}\right),$$

where $\hat{\mathbf{b}}$ the best linear unbiased predictor of $\mathbf{b}$ defined in (3.5), and the penalty term is defined as

$$a_N\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}\right) = \frac{2(N-P-1)}{N-p-2}\{\rho(\hat{\boldsymbol{\theta}}) + 1 + \frac{p+1}{N-p-1}\},$$

where $\rho(\hat{\boldsymbol{\theta}})$ is the effective degrees of freedom used in estimating $\boldsymbol{\beta}$ and $\mathbf{b}$.

Marginal AIC also has a version based on the REML which is given by

$$\mathrm{mAIC}_R = -2\ell_R\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}\right) + 2a_N * q,$$

where $a_N = \frac{N-p}{N-p-q-1}$.

With regard to the penalized methods, Lin *et al.* (2013) proposed a two-stage mixed model

selection procedure based upon REML. In the first stage, the penalized restricted log-likelihood is employed to select the important random effects, and it is carried out by a Newton-type algorithm. Next, in the second stage, the penalized log-likelihood is utilized to select the proper fixed effects, the selection and estimation is accomplished by the pathwise coordinate optimization (Friedman *et al.*, 2007).

Compared to the maximum likelihood based approaches, the method in Lin *et al.* (2013) tries to make full use of REML's strength in variance component parameters estimation, and therefore is particularly more efficient in random effects selection. After the appropriate random effects are chosen in the first stage, selecting the fixed effects only involves the regression coefficients. The usage of two-stage selection respects the different natures of the fixed and random effects and improves the computational efficiency.

Note that all the above procedures require the normality for the random effects and the error terms, thus the efficacy of their inferences are limited if the distribution is not normal. To adjust for this weakness, some robust approaches without distribution assumption have been studied, and we will review them in the next section.

3.4 Robust Mixed Model Selection Methods

To select and estimate both fixed and random effects, Peng and Lu (2010) proposed a two-step distribution free penalized procedure.

In the first step, the covariance matrix $\mathbf{D}$ is estimated by minimizing the following penalized least squares

$$\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{b}_i\mathbf{Z}_i)^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{b}_i\mathbf{Z}_i) + N\sum_{k=1}^{q} p(\sqrt{|\mathbf{D}_{kk}|}),$$

where $p(\cdot)$ is the SCAD penalty function, and $\mathbf{D}_{kk}$ is the kth diagonal element of $\hat{\mathbf{D}}$. The solution of the above function $\hat{\mathbf{b}}_i$ can be updated based on ridge regression, and an estimate of $\mathbf{D}$ can be updated as

$$\hat{\mathbf{D}} = \frac{\sum_{i=1}^{n}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T}{n\hat{\sigma}^2} - \frac{\sum_{i=1}^{n}\hat{\mathbf{Z}}_i\hat{\mathbf{Z}}_i^T}{n}.$$

After $\hat{\mathbf{D}}$ is obtained, in the second step, the selection of fixed effects is achieved by minimizing

the following penalized least squares

$$\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T(\mathbf{I}_{n_i} + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) + N\sum_{k=1}^{p} p(|\beta_k|).$$

The $\hat{\boldsymbol{\beta}}$ is updated based on ridge regression, and the process is continued iterating between step one and two until convergence.

As a robust selection method, the method does not rely on normality assumption, so it is expected to have promising performance against non-normality of the data. Another advantage of this method is its computational stability since the complex constrained optimization problem of the covariance matrix is prevented. However, the method needs sufficient number of observations within each cluster. When the cluster size is small, it performs worse than the likelihood based procedures (Peng and Lu, 2012, page 119-120). Moreover, when the errors are known to be normally distributed, the procedure is shown to be less efficient.

Ahn *et al.* (2012) suggested to use a second-order moment loss function for estimating the covariance matrix of the random effects, then the random effects selection can be achieved by minimizing the penalized loss function. Two types of shrinkage penalties including a hard thresholding operator and a sandwich-type soft thresholding penalty are proposed for random effects selection. Furthermore, the procedure is extended to the selection of fixed effects. In the view of such moment-based method, the estimators of this procedure do not need normality assumption of the error terms, and hence are more robust for non-normal correlated data.

CHAPTER 4 ADAPTIVE LASSO FOR MIXED MODEL SELECTION VIA PROFILE LOG-LIKELIHOOD

Though some good strategies were identified for mixed model selection as we have introduced in Chapter 3, there is still much room for further improvement of selection accuracy and computation efficiency. To further improve the performance of the existing methodologies, we propose a two-stage selection procedure for linear mixed models in this chapter. In the two stages, the random effects and fixed effects are selected employing the adaptive LASSO penalized term via the restricted profile log-likelihood and the profile log-likelihood function, respectively.

Our proposal is different from the existing ones in the literature mainly in two aspects. First, the proposed method is composed of two stages to separately choose the parameters of interests, therefore can respect and accommodate the distinct properties between the random and fixed effects. Second, the usage of the profile log-likelihoods in the selection process can make the computation more efficient and stable due to a smaller number of dimensions involved.

Further, we study the theoretical properties of the proposed procedure including estimation consistency and the oracle properties (Fan and Li, 2001). We prove that, with probability tending to one, the proposed procedure surely selects the true mixed model.

Now, recall the linear mixed model defined in (3.1) takes the form of

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \ldots, n.$$

For this model, the log-likelihood function in (3.2) is

$$\ell_F\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma\right) = -\frac{1}{2}\sum_{i=1}^{n}\log\left|\sigma^2\mathbf{V}_i\right| - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i,$$

where $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$.

The restricted log-likelihood function in (3.4) is

$$\ell_R(\boldsymbol{\theta}, \sigma) = \ell_F\left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \sigma\right) - \frac{1}{2}\log\left|\sigma^{-2}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right|,$$

where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ defined in (3.3).

Note that the model variance parameter σ is involved in both the log-likelihood function in (3.2) and the restricted log-likelihood function in (3.4), while it is neither a fixed effect nor a random effect. In other words, with regard to the fixed and random effects selection, we can consider σ as a nuisance parameter. Therefore, on one hand, profiling out σ can still catch enough and primary information for the model in (3.1) (e.g., see Fan and Li, 2012). On the other hand, removing σ from the selection procedure makes the computation more effective and steady, since the dimension involved in the profiled log-likelihood is lower than the log-likelihood and the restricted log-likelihood.

Considering the points discussed above, we substitute the maximum likelihood (ML) of σ^2 $\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i$, and restricted maximum likelihood (REML) estimators of σ^2 $\hat{\sigma}^2_{\mathrm{REML}} = \frac{1}{N-p}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i$ into equations (3.2) and (3.4), respectively, and then the profile log-likelihood and restricted profile log-likelihood (Lindstrom and Bates, 1988) can be obtained as

$$p_F(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}_i| - \frac{N}{2}\log\left(\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i\right), \tag{4.1}$$

and

$$\begin{aligned} p_R(\boldsymbol{\theta}) \;=\; & -\frac{1}{2}\log\left|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right| \\ & -\frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}_i| \\ & -\frac{1}{2}(N-p)\log\left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}\right)^T\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}\right)\right]. \end{aligned} \tag{4.2}$$

By explicitly profiling $\boldsymbol{\beta}$ and σ out of the log-likelihood, the restricted profile log-likelihood

function in equation (4.2) depends only on the variance components $\boldsymbol{\theta}$, we then will have the penalized restricted profile log-likelihood function to select important random effects in the next section. Then accordingly, the penalized method will be established.

4.1 Selection of Random Effects via the Penalized Restricted Profile Log-Likelihood

The selection of the random effects in model (3.1) plays an important role in model estimation and inference. Lange and Laird (1989) argued that overfitting random effects could result in a near singular covariance matrix and then could make the fitting process more difficult. They also argued that underfitting random effects would lead to bias in the variance estimates of the fixed effects. From this perspective, the correct selection of the random effects will undoubtedly increase efficiency or precision with which the fixed effects can be estimated (e.g., see Fitzmaurice *et al.*, 2011, p.165), eventually improve prediction accuracy for future data.

A challenge in using a penalized approach for random effects selection is that an entire row and column of $\mathbf{D}$ must be removed to eliminate a random effect. This leads to difficulties in how to conduct the penalization suitably.

In the first stage, we aim to choose the proper random covariance structure by maximizing the penalized restricted profile log-likelihood in the first step. Observe that if a random effect is a noise variable, then the corresponding variance components should be all zero. Thus, we first estimate the covariance matrix of random effects using the adaptive LASSO penalized method and then identify the vital ones based on the estimated covariance matrix.

To facilitate the random effects selection, we factorize the vector $\boldsymbol{\theta}$ as ($\mathbf{d}$, $\boldsymbol{\gamma}$), where $\mathbf{d} = (d_1, d_2, ..., d_q)$ is a vector consisting of the diagonal elements of $\mathbf{D}$ and $\boldsymbol{\gamma}$ is the vector of free parameters. Now, we have the penalized restricted profile log-likelihood function as

$$Q_R(\boldsymbol{\theta}) = p_R(\boldsymbol{\theta}) - \lambda_{1n} \sum_{j=1}^{q} w_{1j} \left| d_j \right|, \tag{4.3}$$

where $p_R(\boldsymbol{\theta})$ is the restricted profile log-likelihood defined in function (4.2), λ_{1n} is the tuning parameter that controls the model complexity, d_j is jth element of the vector $\mathbf{d}$, and the positive

weights vector $\mathbf{w}_1 = (w_{11}, ..., w_{1q})^T$ is chosen adaptively by data. The chosen values for w_{1j}'s are important for guaranteeing the optimality of the solution. We propose to use $\mathbf{w}_1 = 1/|\tilde{\mathbf{d}}|$, where $\tilde{\mathbf{d}} = \left(\tilde{d}_1, ..., \tilde{d}_q\right)^T$ is a root-n consistent estimator of $\mathbf{d}$, since the values of consistent estimators well reflect the relative importance of the covariates.

With the penalized restricted profile log-likelihood in equation (4.3), penalizing any d_j to be zero is equivalent to setting the entire jth row and jth column of $\mathbf{D}$ to zero, and generating a new submatrix without the corresponding row and column.

To maximize $Q_R(\boldsymbol{\theta})$ in equation (4.3), we apply the Newton-Raphson algorithm, which is well known for fast convergence properties with a proper starting value and takes the iterative updating form of

$$\boldsymbol{\theta}_{b+1} = \boldsymbol{\theta}_b - \mathbf{M}_{\theta\theta}^{-1}\mathbf{sc}_\theta, \qquad b = 0, 1, \ldots, \tag{4.4}$$

where $\boldsymbol{\theta}_b$ is the current step value, and $\mathbf{sc}_\theta$ is a $k \times 1$ vector of the first derivative, and $\mathbf{M}_{\theta\theta}$ is a $k \times k$ matrix of the second derivative. Both are derived with respect to the penalized restricted profile log-likelihood in equation (4.3). We then have $\boldsymbol{\theta}_{b+1}$ in equation (4.4) updated for the next step.

The first and second derivatives of $p_R(\boldsymbol{\theta})$ in equation (4.3) can be calculated from Lindstrom and Bates (1988), then we have

$$\begin{aligned}
\frac{\partial p_R(\boldsymbol{\theta})}{\partial \theta_j} &= \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{X}_i^T\mathbf{A}_{ij}\mathbf{X}_i\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \theta_j}\right) \\
&\quad +\frac{1}{2}(N-p)\frac{\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{A}_{ij}\mathbf{r}_i}{\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i},
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 p_R(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_j} &= \frac{1}{2}\mathrm{tr}\left[\mathbf{H}^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i^T\mathbf{A}_{il}\mathbf{X}_i\right) * \mathbf{H}^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i^T\mathbf{A}_{ij}\mathbf{X}_i\right)\right] \\ &+\frac{1}{2}\mathrm{tr}\left[\mathbf{H}^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i^T\frac{\partial\mathbf{A}_{ij}}{\partial\theta_l}\mathbf{X}_i\right)\right] \\ &-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left[-\mathbf{A}_{il}^T\frac{\partial\mathbf{V}_i}{\partial\theta_j}+\mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\theta_l\partial\theta_j}\right] \\ &+\frac{1}{2}(N-p)\frac{\sum_{i=1}^{n}\mathbf{r}_i^T\frac{\partial\mathbf{A}_{ij}}{\partial\theta_l}\mathbf{r}_i * \sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i + \left(\frac{\sum_{i=1}^{n}-\mathbf{r}_i^T\mathbf{A}_{ij}\mathbf{r}_i}{\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}\right)^2}{\left(\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i\right)^2},\end{aligned}$$

respectively, where

$$\mathbf{A}_{ij} = \mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1},$$

$$\frac{\partial\mathbf{A}_{ij}}{\partial\theta_l} = -\mathbf{V}_i^{-1}\left(\frac{\partial\mathbf{V}_i}{\partial\theta_l}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j} - \frac{\partial^2\mathbf{V}_i}{\partial\theta_l\partial\theta_j} + \frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_l}\right)\mathbf{V}_i^{-1},$$

and

$$\mathbf{H} = \sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i.$$

Note that the Newton-Raphson algorithm can not be directly applied because the penalty term in function (4.3) is non-differentiable at the origin. Motivated by Li and Liang (2008), we can approximate the penalty in (4.3) by a local quadratic approximation at every iteration step as

$$|d_j| \approx \frac{1}{2}\left|d_j^{(0)}\right| + \frac{1}{2}\frac{d_j^2}{\left|d_j^{(0)}\right|}, \qquad \text{for } d_j \approx d_j^{(0)}, \tag{4.5}$$

where $\mathbf{d}^0$ is an initial value close to the maximizer of function (4.3). With the aid of the local quadratic approximation in (4.5), we can find the derivatives of the penalized restricted profile log-likelihood function in (4.3). The Newton-Raphson algorithm then can be utilized to search for the solution of maximizing the penalized restricted profile likelihood function in (4.3), and the process is repeated until the convergence is reached. The converged $\hat{\boldsymbol{\theta}}$ is the penalized restricted profile

likelihood estimator in (4.3), and then the covariance matrix of random effects $\mathbf{V}$ can be estimated by $\hat{\mathbf{V}}$ based on $\hat{\boldsymbol{\theta}}$.

4.2 Selection of Fixed Effects via the Penalized Profile Log-Likelihood

Using the proper estimation of covariance matrix of the random effects from the previous section can help us investigate on the selection and estimation of important fixed effects. We will select the fixed effects with the utility of the penalized profile log-likelihood, and then determine the final model.

After the covariance matrix of random effects $\mathbf{V}$ is estimated by $\hat{\mathbf{V}}$, dropping constant terms, the profile likelihood function in (4.1) is given by

$$p_F(\boldsymbol{\beta}) = -\frac{N}{2}\log\left(\sum_{i=1}^{n}\mathbf{r}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i\right). \tag{4.6}$$

From Lindstrom and Bates (1988), we can have the first and second derivatives of $p_F(\beta)$ in (4.6) as

$$\frac{\partial p_F(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = N\frac{\sum_{i=1}^{n}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i}{\sum_{i=1}^{n}\mathbf{r}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i},$$

and

$$\frac{\partial^2 p_F(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T\partial \boldsymbol{\beta}} = -\frac{N}{2}\frac{\sum_{i=1}^{n}2\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i * \sum_{i=1}^{n}\mathbf{r}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i - \left(\frac{\sum_{i=1}^{n}-2\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i}{\sum_{i=1}^{n}\mathbf{r}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i}\right)^2}{\left(\sum_{i=1}^{n}\mathbf{r}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{r}_i\right)^2}.$$

To determine the set of covariates for fixed effects, we can maximize the penalized profile log-likelihood, which has the form of

$$Q_F(\boldsymbol{\beta}) = p_F(\boldsymbol{\beta}) - \lambda_{2n}\sum_{j=1}^{p} w_{2j}\left|\beta_j\right|, \tag{4.7}$$

where $p_F(\boldsymbol{\beta})$ is the profile log-likelihood defined in (4.6), λ_{2n} is the tuning parameter for the fixed effects selection, and the positive weights vector $\mathbf{w}_2$ is data dependent. We suggest to use $\mathbf{w}_2 = 1/|\tilde{\boldsymbol{\beta}}|$, where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ in (3.3) using the estimated covariance matrix $\hat{\mathbf{V}}$. Since $\tilde{\boldsymbol{\beta}}$

is consistent estimator, its values well reflect the relative importance of the covariates.

To maximize $Q_F(\boldsymbol{\beta})$ in equation (4.7), we again apply the Newton-Raphson algorithm as

$$\boldsymbol{\beta}_{b+1} = \boldsymbol{\beta}_b - \mathbf{M}_{\beta\beta}^{-1}\mathbf{sc}_\beta, \qquad b = 0, 1, \ldots, \tag{4.8}$$

where $\boldsymbol{\beta}_b$ is the current step value, and $\mathbf{sc}_\beta$ is a $p \times 1$ vector of the first derivative, and $\mathbf{M}_{\beta\beta}$ is a $p \times p$ matrix of the second derivative. Both are derived with respect to the penalized profile log-likelihood in (4.7). We then have $\boldsymbol{\beta}_{b+1}$ in equation (4.8) updated for the next step.

By the same argument in equation (4.5), we can approximate the penalty term in (4.7) by a local quadratic approximation at every iteration step as

$$|\beta_j| \approx \frac{1}{2}\left|\beta_j^{(0)}\right| + \frac{1}{2}\frac{\beta_j^2}{\left|\beta_j^{(0)}\right|}, \qquad \text{for } \beta_j \approx \beta_j^{(0)}, \tag{4.9}$$

where $\boldsymbol{\beta}^0$ is an initial value close to the maximizer of (4.7). The Newton-Raphson algorithm then is utilized to find the maximizer of the penalized profile likelihood in (4.7), and the iteration is continued until the convergence is achieved. The converged $\hat{\boldsymbol{\beta}}$ is the penalized profile likelihood estimator in (4.7).

After finishing the two-stage penalized procedure, with the proper choice of the tuning parameters, the appropriate linear mixed model can finally be identified.

4.3 Selection of Tuning Parameters

The performance of penalized methods highly relies on the tuning parameters that balance the trade-off between model fitting and model sparsity (Sun *et al.*, 2013). To implement the model selection procedure, the tuning parameter λ has to be properly selected among the candidate values. The selection of λ can be carried out by minimizing one of the commonly used selection criteria, such as AIC, BIC, Mallows' Cp, CV, and GCV, which were all discussed in Section 2.3. In this dissertation, we propose three criteria for mixed model selection.

First, we employ the BIC-type criteria given by

$$\mathrm{BIC_R} = -2 * p_R\left(\hat{\boldsymbol{\theta}}\right) + \log\left(N\right) * df_R, \tag{4.10}$$

and

$$\mathrm{BIC_F} = -2 * p_F(\hat{\boldsymbol{\beta}}) + \log(N) * df_F, \tag{4.11}$$

for the random and fixed effects, respectively.

Second, the AIC-type criteria are given by

$$\mathrm{AIC_R} = -2 * p_R\left(\hat{\boldsymbol{\theta}}\right) + 2 * df_R, \tag{4.12}$$

and

$$\mathrm{AIC_F} = -2 * p_F(\hat{\boldsymbol{\beta}}) + 2 * df_F, \tag{4.13}$$

for the random and fixed effects, respectively.

We also suggest the GCV criteria expressed as

$$\mathrm{GCV_R} = -\frac{1}{N}\frac{p_R\left(\hat{\boldsymbol{\theta}}\right)}{[1 - df_R/N]^2}. \tag{4.14}$$

and

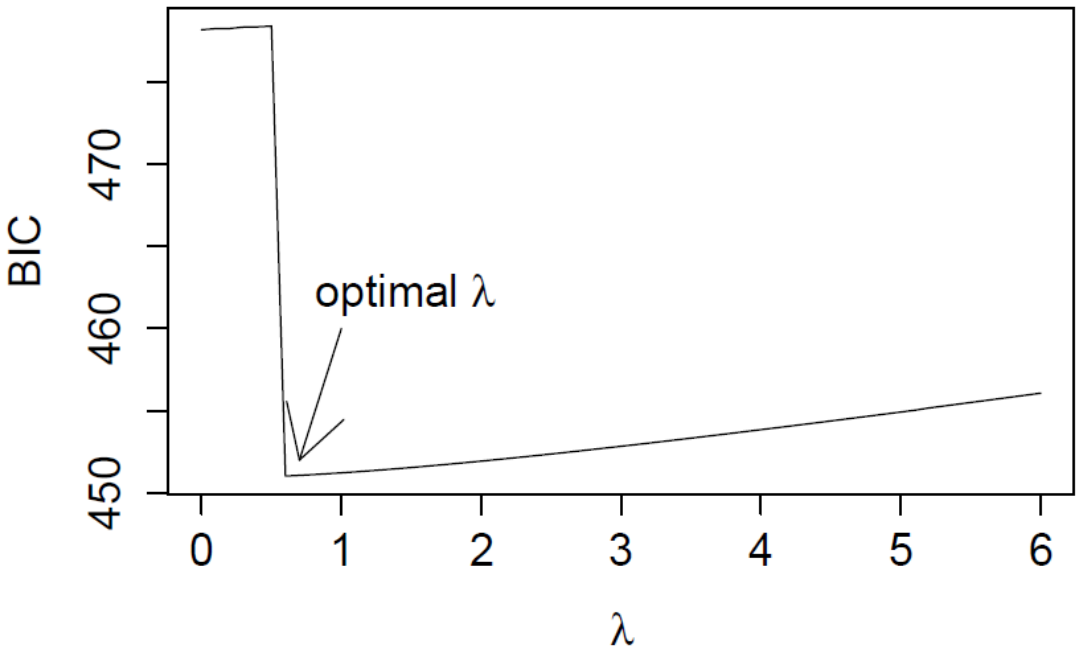
$$\mathrm{GCV_F} = -\frac{1}{N}\frac{p_F(\hat{\boldsymbol{\beta}})}{[1 - df_F/N]^2}. \tag{4.15}$$

for the random and fixed effects, respectively.

Note that $p_R(\hat{\boldsymbol{\theta}})$ is the restricted profile log-likelihood in (4.2) evaluated at $\hat{\boldsymbol{\theta}}$, and $p_F(\hat{\boldsymbol{\beta}})$ is the profile log-likelihood in (4.6) evaluated at $\hat{\boldsymbol{\beta}}$, df_R and df_F are the dimensions of nonzero parts of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, respectively. The λ minimizes the value of criterion (BIC, AIC, or GCV) is chosen as the optimal tuning parameter, as shown in Figure (4.1) using BIC selector. We will compare the performance of the three groups of criteria in Chapter 5.

.

Figure 4.1: Choosing optimal tuning parameter λ by minimizing BIC.



4.4 Theoretical Properties

It has been argued that a good selection procedure should have the oracle properties (Fan and Li, 2001), namely, asymptotically the procedure will choose the true model with probability one. In this section, we show that with a proper choice of the tuning parameter, the proposed estimator is consistent and owns the oracle properties.

Let the true value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}^T, \boldsymbol{\theta}_{20}^T)^T$, where $\boldsymbol{\theta}_{10} = (\mathbf{d}_{10}^T, \boldsymbol{\gamma}_{10}^T)^T$ is an $s \times 1$ vector whose components are nonzero and $\boldsymbol{\theta}_{20}$ is the $(k - s)$ remaining zero components. Denote the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{10}$ is a $v \times 1$ vector whose components are nonzero and $\boldsymbol{\beta}_{20}$ is the $(p - v)$ remaining zero components. Correspondingly, we write the maximizer of $Q_R(\boldsymbol{\theta})$ in equation (4.3) as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T)^T$ and the maximizer of $Q_F(\boldsymbol{\beta})$ in equation (4.7) as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$. Also, let $\mathbf{I}_R(\boldsymbol{\theta}_0)$ be the Fisher information matrix based on the restricted profile log-likelihood in equation (4.2) and let $\mathbf{I}_R(\boldsymbol{\theta}_{10}) = \mathbf{I}_R(\boldsymbol{\theta}_{10}, 0)$ be the Fisher information knowing $\boldsymbol{\theta}_{20} = 0$. Similarly, define $\mathbf{I}_F(\boldsymbol{\beta}_0)$ be the Fisher information matrix based on the profile log-likelihood in equation (4.6) and $\mathbf{I}_F(\boldsymbol{\beta}_{10}) = \mathbf{I}_F(\boldsymbol{\beta}_{10}, 0)$ be the Fisher information knowing $\boldsymbol{\beta}_{20} = 0$.

We assume $\mathbf{I}_R(\boldsymbol{\theta}_0), \mathbf{I}_R(\boldsymbol{\theta}_{10}), \mathbf{I}_F(\boldsymbol{\beta}_0)$, and $\mathbf{I}_F(\boldsymbol{\beta}_{10})$ are all finite and positive definite.

Moreover, assume that there exists a subset Θ of R^k, containing the true parameter θ_0 such that $p_R(\boldsymbol{\theta_0})$ in equation (4.2) admits all second order derivatives. Let $\bigtriangledown p_R(\boldsymbol{\theta_0}) = \frac{\partial p_R(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta_0}}$ and $\bigtriangledown^2 p_R(\boldsymbol{\theta_0}) = \frac{\partial \bigtriangledown p_R(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta_0}^T}$ be the vector and matrix of the first and second order partial derivatives of $p_R(\boldsymbol{\theta}_0)$, respectively. Assume that there exists a subset Ω of R^p, containing the true parameter β_0 such that $p_F(\boldsymbol{\beta_0})$ in equation (4.6) admits all second order derivatives. Denote $\bigtriangledown p_F(\boldsymbol{\beta_0}) = \frac{\partial p_F(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta_0}}$ and $\bigtriangledown^2 p_F(\boldsymbol{\beta_0}) = \frac{\partial \bigtriangledown p_F(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta_0}^T}$ to be the vector and matrix of the first and second order partial derivatives of $p_F(\boldsymbol{\beta}_0)$, respectively.

Theorem 1. *(Consistency for random effects estimation). If* $\frac{\lambda_{1n}}{n} = O_p(1)$*, then there exists a local maximizer* $\hat{\boldsymbol{\theta}}$ *of* $Q_R(\boldsymbol{\theta})$ *in equation (4.3) such that* $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p\left(n^{-\frac{1}{2}}\right)$.

Proof of Theorem 1:

To show $\hat{\boldsymbol{\theta}}$ is a $\sqrt{n}-$consistent estimator of $\boldsymbol{\theta}_0$, it suffices to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q_R\left(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}\right) < Q_R(\boldsymbol{\theta}_0)\right\} \geq 1 - \epsilon. \tag{4.16}$$

This implies with probability $1-\epsilon$, there exists a local maximizer of $Q_R(\boldsymbol{\theta})$ in the ball $\left\{\boldsymbol{\theta} : \boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}, \|\mathbf{u}\| \leq \mathrm{C}\right\}$. Hence, the maximizer $\hat{\boldsymbol{\theta}}$ must satisfy $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p\left(n^{-\frac{1}{2}}\right)$.

For $p_R(\boldsymbol{\theta_0})$, $\mathrm{E}\left[\bigtriangledown p_R(\boldsymbol{\theta_0})\right] = 0$ and $\mathrm{E}\left[-\bigtriangledown^2 p_R(\boldsymbol{\theta_0})\right] = \mathbf{I}_R(\boldsymbol{\theta}_0)$, then $\frac{\bigtriangledown p_R(\boldsymbol{\theta_0})}{\sqrt{n}} = O_p(1)$ and $-\frac{\bigtriangledown^2 p_R(\boldsymbol{\theta_0})}{n} = \mathbf{I}_R(\boldsymbol{\theta}_0) + o_p(1)$, by the second order Taylor expansion, we have

$$\begin{aligned} p_R\left(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}\right) - p_R(\boldsymbol{\theta}_0) &= \left[\bigtriangledown p_R(\boldsymbol{\theta}_0)\right]^T n^{-\frac{1}{2}}\mathbf{u} + \frac{1}{2}\mathbf{u}^T\frac{\bigtriangledown^2 p_R(\boldsymbol{\theta_0})}{n}\mathbf{u} + \mathbf{u}^T o_p(1)\,\mathbf{u} \\ &= -\frac{1}{2}\mathbf{u}^T\left[\mathbf{I}_R(\boldsymbol{\theta}_0) + o_p(1)\right]\mathbf{u} + O_p(1)\,\mathbf{u}. \end{aligned}$$

Now we define $D_n(\mathbf{u}) \equiv \frac{1}{n}\left[Q_R(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - Q_R(\boldsymbol{\theta}_0)\right]$, where $\mathbf{u} = (u_1, \ldots, u_q)^T$, then we

have

$$\begin{aligned} D_n(\mathbf{u}) =& \frac{1}{n}\left[p_R\left(\boldsymbol{\theta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_R\left(\boldsymbol{\theta}_0\right)-\lambda_{1n}\sum_{j=1}^{q}\left(\frac{\left|d_{j0}+n^{-\frac{1}{2}}u_j\right|}{\left|\tilde{d}_j\right|}-\frac{|d_{j0}|}{\left|\tilde{d}_j\right|}\right)\right] \\ \leq& \frac{1}{n}\left[p_R\left(\boldsymbol{\theta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_R\left(\boldsymbol{\theta}_0\right)-\lambda_{1n}\sum_{j=1}^{q_1}\left(\frac{\left|d_{j0}+n^{-\frac{1}{2}}u_j\right|}{\left|\tilde{d}_j\right|}-\frac{|d_{j0}|}{\left|\tilde{d}_j\right|}\right)\right] \\ \leq& \frac{1}{n}\left[p_R\left(\boldsymbol{\theta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_R\left(\boldsymbol{\theta}_0\right)+\frac{\lambda_{1n}}{\sqrt{n}}\sum_{j=1}^{q_1}\frac{|u_j|}{\left|\tilde{d}_j\right|}\right] \\ =& -\frac{1}{2n}\mathbf{u}^T\left[\mathbf{I}_R(\boldsymbol{\theta}_0)+o_p\left(1\right)\right]\mathbf{u}+\frac{1}{n}O_p\left(1\right)\mathbf{u}+\frac{\lambda_{1n}}{n\sqrt{n}}\sum_{j=1}^{q_1}\frac{|u_j|}{\left|\tilde{d}_j\right|}, \end{aligned} \tag{4.17}$$

where q_1 is the dimension of $\mathbf{d}_{10}$.

Since $\tilde{\mathbf{d}}$ is root-n consistent estimator of $\mathbf{d}$, $\|\tilde{\mathbf{d}}-\mathbf{d}\|=O_p\left(n^{-\frac{1}{2}}\right)$, then for $1\leq j\leq q_1$, by the first order Taylor expansion, we have

$$\frac{1}{|\tilde{d}_j|}=\frac{1}{|d_{j0}|}-\frac{\text{sign}(d_{j0})}{\left|d_{j0}^2\right|}\left(\tilde{d}_j-d_{j0}\right)+o_p\left(\left|\tilde{d}_j-d_{j0}\right|\right)=\frac{1}{|d_{j0}|}+\frac{O_p\left(1\right)}{\sqrt{n}}.$$

In addition, since $\frac{\lambda_{1n}}{n}=O_p\left(1\right)$, we have

$$\begin{aligned} \frac{\lambda_{1n}}{n\sqrt{n}}\sum_{j=1}^{q_1}\frac{|u_j|}{\left|\tilde{d}_j\right|} =& \frac{\lambda_{1n}}{n\sqrt{n}}\sum_{j=1}^{q_1}\left(\frac{|u_j|}{|d_{j0}|}+\frac{|u_j|}{\sqrt{n}}O_p\left(1\right)\right) \\ & \leq \mathrm{C}n^{-1}\left(n^{-1}\lambda_{1n}\right)O_p\left(1\right)=\mathrm{C}n^{-1}O_p\left(1\right). \end{aligned}$$

Since $\mathbf{I}_R(\boldsymbol{\theta}_0)$ is finite and positive definite, therefore in (4.17), if we choose a sufficient large C, the first term is of the order C^2n^{-1} the second and third terms are of the order of $\mathrm{C}n^{-1}$, which are dominated by the first term. Thus (4.16) holds and it completes the proof.

Theorem 2. *(Oracle properties for random effects selection). If* $\lambda_{1n}\to\infty$ *and* $\frac{\lambda_{1n}}{\sqrt{n}}\to 0$ *as n* $\to\infty$*, then with probability tending to 1, the root-n consistent local maximizer* $\hat{\boldsymbol{\theta}}$ *in Theorem 1 must satisfy*

1. *Sparsity:* $\hat{\boldsymbol{\theta}}_2 = 0$.

2. *Asymptotic normality:* $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\right) \xrightarrow{d} N\left(0, \mathbf{I}_R^{-1}\left(\boldsymbol{\theta}_{10}\right)\right)$.

Proof of Theorem 2:

(1). Here we show that $\hat{\boldsymbol{\theta}}_2 = 0$. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\theta}_1$ satisfying $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_p\left(n^{-\frac{1}{2}}\right)$, and for some small $\epsilon_n = \mathrm{C}n^{-1/2}$ and for $j = s+1, \ldots, k$,

$$\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta}) < 0 \quad \text{for} \quad 0 < \theta_j < \epsilon_n,$$

$$\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta}) > 0 \quad \text{for} \quad -\epsilon_n < \theta_j < 0.$$

For $j = s+1, \ldots, k$, we have

$$\begin{aligned}\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta}) =& \frac{\partial}{\partial \theta_j} p_R(\boldsymbol{\theta}) - \lambda_{1n} \frac{\text{sign}(\theta_j)}{|\tilde{\theta}_j|} I(\theta_j \in \mathbf{d}) \\ =& O_p\left(n^{\frac{1}{2}}\right) - \lambda_{1n} n^{\frac{1}{2}} \frac{\text{sign}(\theta_j)}{\left|n^{\frac{1}{2}} \tilde{\theta}_j\right|} I(\theta_j \in \mathbf{d}).\end{aligned}$$

Note that for $j = s+1, \ldots, k$, $n^{\frac{1}{2}}\left(\tilde{\theta}_j - 0\right) = O_p(1)$, so that we have

$$\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta}) = n^{\frac{1}{2}} \left[O_p(1) - \lambda_{1n} \frac{\text{sign}(\theta_j)}{|O_p(1)|} I(\theta_j \in \mathbf{d}) \right].$$

Since $\lambda_{1n} \to \infty$, the sign of $\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta})$ is completely determined by the sign of θ_j when n is large. This completes the proof.

(2). Here we show the asymptotic normality of $\hat{\boldsymbol{\theta}}_1$. From the proof of Theorem 1, we have that there exists a root-n local maximizer $\hat{\boldsymbol{\theta}}_1$ of $Q_R\left\{\begin{pmatrix} \boldsymbol{\theta}_1 \\ 0 \end{pmatrix}\right\}$, i.e.

$$\frac{\partial}{\partial \theta_j} Q_R(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1, \mathbf{0})^T} = \frac{\partial}{\partial \theta_j} p_R(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1, \mathbf{0})^T} - \lambda_{1n} \frac{\text{sign}(\hat{\theta}_j)}{|\tilde{\theta}_j|} I(\theta_j \in \mathbf{d}) = 0.$$

By the Taylor series expansion, we have

$$\begin{aligned}\mathbf{0} &= \bigtriangledown\, p_R(\boldsymbol{\theta}_{10}) - \hat{\mathbf{I}}_R(\boldsymbol{\theta}_*)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \\ &\quad - \lambda_{1n}\left(\frac{\text{sign}(\hat{\theta}_1)}{\left|\tilde{\theta}_1\right|} I(\theta_1 \in \mathbf{d}), \ldots, \frac{\text{sign}(\hat{\theta}_t)}{\left|\tilde{\theta}_t\right|} I(\theta_t \in \mathbf{d})\right)^T \\ &= \bigtriangledown p_R(\boldsymbol{\theta}_{10}) - \hat{\mathbf{I}}_R(\boldsymbol{\theta}_*)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \\ &\quad - \lambda_{1n}\left(\frac{\text{sign}(\theta_{10})}{\left|\tilde{\theta}_1\right|} I(\theta_1 \in \mathbf{d}), \ldots, \frac{\text{sign}(\theta_{t0})}{\left|\tilde{\theta}_t\right|} I(\theta_t \in \mathbf{d})\right)^T,\end{aligned}$$

where $\boldsymbol{\theta}_*$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, $t = q_1(q_1+1)/2$, and $\hat{\mathbf{I}}_R(\boldsymbol{\theta})$ is the first $t \times t$ sub-matrix of $\bigtriangledown^2 p_R(\boldsymbol{\theta})$. The last equation is implied by sign($\hat{\theta}_{jn}$) = sign(θ_{j0}) when n is large, since $\hat{\theta}$ is a root-n consistent estimator of $\boldsymbol{\theta}_0$.

By the the multivariate central limit theorem and the law of large numbers, we can prove that

$$\frac{\bigtriangledown p_R(\boldsymbol{\theta}_{10})}{\sqrt{n}} \xrightarrow{d} N\left(0, \mathbf{I}_R(\boldsymbol{\theta}_{10})\right), \text{and} \quad \frac{\hat{\mathbf{I}}_R(\boldsymbol{\theta}_*)}{n} \xrightarrow{p} \mathbf{I}_R(\boldsymbol{\theta}_{10}).$$

If $\frac{\lambda_{1n}}{\sqrt{n}} \to \lambda_0$, a nonnegative constant, by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \xrightarrow{d} N\left(-\lambda_0 \mathbf{I}_R^{-1}(\boldsymbol{\theta}_{10}) b_1, \mathbf{I}_R^{-1}(\boldsymbol{\theta}_{10})\right),$$

where $b_1 = \left(\frac{\text{sign}(\theta_{10})}{\left|\tilde{\theta}_1\right|} I(\theta_1 \in \mathbf{d}), \ldots, \frac{\text{sign}(\theta_{t0})}{\left|\tilde{\theta}_t\right|} I(\theta_t \in \mathbf{d})\right)^T$.

In particular, if $\frac{\lambda_{1n}}{\sqrt{n}} \to 0$, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_R^{-1}(\boldsymbol{\theta}_{10})\right).$$

This completes the proof.

Remark 1. Theorem 1 and 2 state the asymptotic properties of the proposed procedure for random effects selection and estimation. From Theorem 1 we see that our penalized restricted profile

likelihood estimator $\hat{\boldsymbol{\theta}}$ is root-n consistent of $\boldsymbol{\theta}$, and Theorem 2 shows that this estimator possesses the oracle properties, including selection consistency and asymptotic normality.

Theorem 3. *(Consistency for fixed effects estimation). If* $\frac{\lambda_{2n}}{n} = O_p(1)$*, then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_F(\boldsymbol{\beta})$ *in equation (4.7) such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(n^{-\frac{1}{2}}\right)$.

Proof of Theorem 3:

To show $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}-$consistent estimator of $\boldsymbol{\beta}_0$, it suffices to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q_F\left(\boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{u}\right) < Q_F\left(\boldsymbol{\beta}_0\right)\right\} \geq 1 - \epsilon. \tag{4.18}$$

This implies with probability $1-\epsilon$, there exists a local maximizer of $Q_F(\boldsymbol{\beta})$ in the ball $\left\{\boldsymbol{\beta} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{u}, \|\mathbf{u}\| \leq \mathrm{C}\right\}$. Hence, the maximizer $\hat{\boldsymbol{\beta}}$ must satisfy $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(n^{-\frac{1}{2}}\right)$.

For $p_F(\boldsymbol{\beta}_0)$, $\mathrm{E}\left[\bigtriangledown p_F\left(\boldsymbol{\beta}_\mathbf{0}\right)\right] = 0$ and $\mathrm{E}\left[-\bigtriangledown^2 p_F\left(\boldsymbol{\beta}_\mathbf{0}\right)\right] = \mathbf{I}_F(\boldsymbol{\beta}_0)$, then $\frac{\bigtriangledown p_F(\boldsymbol{\beta}_\mathbf{0})}{\sqrt{n}} = O_p(1)$ and $-\frac{\bigtriangledown^2 p_F(\boldsymbol{\beta}_\mathbf{0})}{n} = \mathbf{I}_F(\boldsymbol{\beta}_0) + o_p(1)$, by the second order Taylor expansion, we have

$$\begin{aligned} p_F\left(\boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{u}\right) - p_F\left(\boldsymbol{\beta}_0\right) &= \left[\bigtriangledown p_F\left(\boldsymbol{\beta}_0\right)\right]^T n^{-\frac{1}{2}}\mathbf{u} + \frac{1}{2}\mathbf{u}^T\frac{\bigtriangledown^2 p_F\left(\boldsymbol{\beta}_\mathbf{0}\right)}{n}\mathbf{u} + \mathbf{u}^T o_p(1)\,\mathbf{u} \\ &= -\frac{1}{2}\mathbf{u}^T\left[\mathbf{I}_F(\boldsymbol{\beta}_0) + o_p(1)\right]\mathbf{u} + O_p(1)\,\mathbf{u}. \end{aligned}$$

Now we define $M_n(\mathbf{u}) \equiv \frac{1}{n}\left[Q_F(\boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - Q_F(\boldsymbol{\beta}_0)\right]$, where $\mathbf{u} = (u_1, \ldots, u_p)^T$, then we

have

$$\begin{aligned} M_n(\mathbf{u}) =& \frac{1}{n}\left[p_F\left(\boldsymbol{\beta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_F\left(\boldsymbol{\beta}_0\right)-\lambda_{2n}\sum_{j=1}^{p}\left(\frac{\left|\beta_{j0}+n^{-\frac{1}{2}}u_j\right|}{\left|\tilde{\beta}_j\right|}-\frac{|\beta_{j0}|}{\left|\tilde{\beta}_j\right|}\right)\right] \\ \leq& \frac{1}{n}\left[p_F\left(\boldsymbol{\beta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_F\left(\boldsymbol{\beta}_0\right)-\lambda_{2n}\sum_{j=1}^{v}\left(\frac{\left|\beta_{j0}+n^{-\frac{1}{2}}u_j\right|}{\left|\tilde{\beta}_j\right|}-\frac{|\beta_{j0}|}{\left|\tilde{\beta}_j\right|}\right)\right] \\ \leq& \frac{1}{n}\left[p_F\left(\boldsymbol{\beta}_0+n^{-\frac{1}{2}}\mathbf{u}\right)-p_F\left(\boldsymbol{\beta}_0\right)+\frac{\lambda_{2n}}{\sqrt{n}}\sum_{j=1}^{v}\frac{|u_j|}{\left|\tilde{\beta}_j\right|}\right] \\ =& -\frac{1}{2n}\mathbf{u}^T\left[\mathbf{I}_F(\boldsymbol{\beta}_0)+o_p\left(1\right)\right]\mathbf{u}+\frac{1}{n}O_p\left(1\right)\mathbf{u}+\frac{\lambda_{2n}}{n\sqrt{n}}\sum_{j=1}^{v}\frac{|u_j|}{\left|\tilde{\beta}_j\right|}, \end{aligned} \tag{4.19}$$

where v is the dimension of $\boldsymbol{\beta}_{10}$.

Since $\tilde{\boldsymbol{\beta}}$ is root-n consistent estimator of $\boldsymbol{\beta}$, $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|=O_p\left(n^{-\frac{1}{2}}\right)$, then for $1\leq j\leq v$, by the first order Taylor expansion, we have

$$\frac{1}{|\tilde{\beta}_j|}=\frac{1}{|\beta_{j0}|}-\frac{\operatorname{sign}(\beta_{j0})}{\left|\beta_{j0}^2\right|}\left(\tilde{\beta}_j-\beta_{j0}\right)+o_p\left(\left|\tilde{\beta}_j-\beta_{j0}\right|\right)=\frac{1}{|\beta_{j0}|}+\frac{O_p\left(1\right)}{\sqrt{n}}.$$

In addition, since $\frac{\lambda_{2n}}{n}=O_p\left(1\right)$, we have

$$\begin{aligned} \frac{\lambda_{2n}}{n\sqrt{n}}\sum_{j=1}^{v}\frac{|u_j|}{\left|\tilde{\beta}_j\right|} =& \frac{\lambda_{2n}}{n\sqrt{n}}\sum_{j=1}^{v}\left(\frac{|u_j|}{|\beta_{j0}|}+\frac{|u_j|}{\sqrt{n}}O_p\left(1\right)\right) \\ & \leq \mathrm{C}n^{-1}\left(n^{-1}\lambda_{2n}\right)O_p\left(1\right)=\mathrm{C}n^{-1}O_p\left(1\right). \end{aligned}$$

Since $\mathbf{I}_F(\boldsymbol{\beta}_0)$ is finite and positive definite, therefore in (4.19), if we choose a sufficient large C, the first term is of the order C^2n^{-1} the second and third terms are of the order of $\mathrm{C}n^{-1}$, which are dominated by the first term. Thus (4.18) holds and it completes the proof.

Theorem 4. *(Oracle properties for fixed effects selection). If $\lambda_{2n}\to\infty$ and $\frac{\lambda_{2n}}{\sqrt{n}}\to 0$ as $n\to\infty$, then with probability tending to 1, the root-n consistent local maximizer $\hat{\boldsymbol{\beta}}$ in Theorem 3 must satisfy*

1. *Sparsity:* $\hat{\boldsymbol{\beta}}_2 = 0$.

2. *Asymptotic normality:* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right) \xrightarrow{d} N\left(0, \mathbf{I}_F^{-1}\left(\boldsymbol{\beta}_{10}\right)\right)$.

Proof of Theorem 4:

(1). Here we show that $\hat{\boldsymbol{\beta}}_2 = 0$. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p\left(n^{-\frac{1}{2}}\right)$, and for some small $\epsilon_n = \mathrm{C}n^{-1/2}$ and for $j = v+1, \ldots, p$,

$$\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta}) < 0 \quad \text{for} \quad 0 < \beta_j < \epsilon_n,$$

$$\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta}) > 0 \quad \text{for} \quad -\epsilon_n < \beta_j < 0.$$

For $j = v+1, \ldots, p$, we have

$$\begin{aligned}\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta}) =& \frac{\partial}{\partial \beta_j} p_F(\boldsymbol{\beta}) - \lambda_{2n} \frac{\text{sign}(\beta_j)}{|\tilde{\beta}_j|} \\ =& O_p\left(n^{\frac{1}{2}}\right) - \lambda_{2n} n^{\frac{1}{2}} \frac{\text{sign}(\beta_j)}{\left|n^{\frac{1}{2}} \tilde{\beta}_j\right|}.\end{aligned}$$

Note that for $j = v+1, \ldots, p$, $n^{\frac{1}{2}}\left(\tilde{\beta}_j - 0\right) = O_p(1)$, so that we have

$$\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta}) = n^{\frac{1}{2}} \left[O_p(1) - \lambda_{2n} \frac{\text{sign}(\beta_j)}{|O_p(1)|} \right].$$

Since $\lambda_{2n} \to \infty$, the sign of $\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta})$ is completely determined by the sign of β_j when n is large. This completes the proof.

(2). Here we show the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$. From the proof of Theorem 3, we have that there exists a root-n local maximizer $\hat{\boldsymbol{\beta}}_1$ of $Q_F\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ 0 \end{pmatrix}\right\}$, i.e.

$$\frac{\partial}{\partial \beta_j} Q_F(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_1, \mathbf{0})^T} = \frac{\partial}{\partial \beta_j} p_F(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_1, \mathbf{0})^T} - \lambda_{2n} \frac{\text{sign}(\hat{\beta}_j)}{|\tilde{\beta}_j|} = 0.$$

By the Taylor series expansion, we have

$$\begin{aligned}\mathbf{0} &= \bigtriangledown\, p_F(\boldsymbol{\beta}_{10}) - \hat{\mathbf{I}}_F(\boldsymbol{\beta}_*)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) - \lambda_{2n}\left(\frac{\text{sign}(\hat{\beta}_1)}{\left|\tilde{\beta}_1\right|}, \ldots, \frac{\text{sign}(\hat{\theta}_v)}{\left|\tilde{\beta}_v\right|}\right)^T \\ &= \bigtriangledown p_F(\boldsymbol{\beta}_{10}) - \hat{\mathbf{I}}_F(\boldsymbol{\beta}_*)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) - \lambda_{2n}\left(\frac{\text{sign}(\beta_{10})}{\left|\tilde{\beta}_1\right|}, \ldots, \frac{\text{sign}(\beta_{v0})}{\left|\tilde{\theta}_v\right|}\right)^T,\end{aligned}$$

where $\boldsymbol{\beta}_*$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and $\hat{\mathbf{I}}_F(\boldsymbol{\beta})$ is the first $v \times v$ sub-matrix of $\bigtriangledown^2 p_F(\boldsymbol{\beta})$. The last equation is implied by sign($\hat{\beta}_{jn}$) = sign(β_{j0}) when n is large, since $\hat{\beta}$ is a root-n consistent estimator of $\boldsymbol{\beta}_0$.

By the the multivariate central limit theorem and the law of large numbers, we can prove that

$$\frac{\bigtriangledown p_F(\boldsymbol{\beta}_{10})}{\sqrt{n}} \xrightarrow{d} N\left(0, \mathbf{I}_F(\boldsymbol{\beta}_{10})\right), \text{and} \quad \frac{\hat{\mathbf{I}}_F(\boldsymbol{\beta}_*)}{n} \xrightarrow{p} \mathbf{I}_F(\boldsymbol{\beta}_{10}).$$

If $\frac{\lambda_{2n}}{\sqrt{n}} \to \lambda_0$, a nonnegative constant, by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{d} N\left(-\lambda_0 \mathbf{I}_F^{-1}(\boldsymbol{\beta}_{10}) b_1, \mathbf{I}_F^{-1}(\boldsymbol{\beta}_{10})\right),$$

where $b_1 = \left(\frac{\text{sign}(\beta_{10})}{\left|\tilde{\beta}_1\right|}, \ldots, \frac{\text{sign}(\beta_{v0})}{\left|\tilde{\beta}_v\right|}\right)^T$.

In particular, if $\frac{\lambda_{2n}}{\sqrt{n}} \to 0$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_F^{-1}(\boldsymbol{\beta}_{10})\right).$$

This completes the proof.

Remark 2. Theorem 3 and 4 present the asymptotic properties of the proposed procedure for fixed effects selection and estimation, given the estimation of $\boldsymbol{\theta}$ in the first step. By appropriately choosing the tuning parameter λ, our penalized profile likelihood estimator $\hat{\boldsymbol{\beta}}$ is root-n consistent, asymptotically normal and holds the sparsity property, that is, it performs as well as the oracle estimators, knowing $\boldsymbol{\beta}_2 = 0$.

CHAPTER 5 SIMULATION STUDIES

After proposing the two-stage procedure for mixed model selection and deriving its large sample theories in Chapter 4, we examine the performance of the proposed procedure under three simulation studies, and compare the simulated results with those for the existing selection approaches. All of the simulated data are generated from model (3.1). The R code for the simulation studies are available in the appendix.

5.1 Simulation 1

This simulation study follows the setting in Ahn (2010). We are particularly interested in model performance in the following aspects: first, the performance of the proposed method under different design structures of the input covariates and the error term distributions; second, the behavior of the proposed method using different tuning parameters; third, the comparison of the simulation results with those for some existing selection approaches, in terms of correct selection frequencies and computation times. For this model setting, we do not include fixed intercept and random intercept in the model.

Consider the true model with $p = 5$ for fixed effects and $q = 5$ for random effects, the true parameter vector $\boldsymbol{\beta} = (1, 2, 2, 0, 0)^T$, and the true covariance matrix

$$\mathbf{D} = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

For this setting, we consider five cases as follows:

- Case 1. Assume the error term $\boldsymbol{\epsilon}_i \sim$ N $(0, \sigma^2\mathbf{I}_{n_i})$, $\mathbf{X}_i \sim$ N $(0, \mathbf{I}_p)$, and $\mathbf{X}_i = \mathbf{Z}_i$.
- Case 2. Assume the error term $\boldsymbol{\epsilon}_i \sim$ N $(0, \sigma^2\mathbf{I}_{n_i})$, $\mathbf{X}_i \sim$ N $(0,$ cov$(\mathbf{X}_i))$, cov$(\mathbf{X}_i)$ is compound

symmetry with variance 1 and covariance 0.5, and $\mathbf{X}_i = \mathbf{Z}_i$.

- Case 3. Assume the error term $\boldsymbol{\epsilon}_i \sim$ N (0, $\sigma^2\mathbf{I}_{n_i}$), $\mathbf{X}_i \sim$ N (0, cov($\mathbf{X}_i$)), cov($\mathbf{X}_i$) has autoregressive covariance structure with $\rho = 0.5$, that is, the covariance between x_j and x_k is $0.5^{|j-k|}$, and $\mathbf{X}_i = \mathbf{Z}_i$.

- Case 4. Assume the error term $\boldsymbol{\epsilon}_i \sim$ N (0, $\sigma^2\mathbf{I}_{n_i}$), $\mathbf{X}_i$ and $\mathbf{Z}_i$ are independently generated from N (0, $\mathbf{I}_p$).

- Case 5. We have $\mathbf{X}_i \sim$ N (0, $\mathbf{I}_p$), and $\mathbf{X}_i = \mathbf{Z}_i$, but here we assume $\boldsymbol{\epsilon}_i$ follows a t distribution with 5 degrees of freedom.

In all the cases, we let the error variance $\sigma^2 = 1$. Regarding the number of subjects, we consider $n = 50, 100$, and 200, and we set $n_i = 5$ observations per subject for all i. We generate 100 datasets for each case and then report the average performance.

First, we explore the performance of the proposed method using different tuning parameters defined in Chapter 4. For random effects selection, we consider three criteria: $\mathrm{BIC_R}$ in (4.10), $\mathrm{AIC_R}$ in (4.12), and $\mathrm{GCV_R}$ in (4.14). For fixed effects selection, we also consider three criteria: $\mathrm{BIC_F}$ in (4.11), $\mathrm{AIC_F}$ in (4.13), and $\mathrm{GCV_F}$ in (4.15). We present some measures for evaluating the selection performance. Note that "CZR" shows the number of zero random effects correctly estimated to be zero, "IZR" denotes the number non-zero random effects incorrectly set to zero, and "CR" provides the frequency that the correct random effects structure is selected. Also note that "CZF" represents the number of zero fixed effects correctly estimated to be zero, "IZF" signifies the number non-zero fixed effects incorrectly set to zero, and "CF" denotes the frequency that the correct fixed effects is selected. For the overall performance, "C" shows the frequency that the correct true model is selected, indicating that both the random effects and fixed effects are correctly identified. Since they are three non-zero fixed effects and two non-zero random effects, the true values of the above measures are CZR = 3, IZR = 0, CR = 100, CZF = 2, IZF = 0, CF = 100, and C = 100.

Tables 5.1 - 5.5 individually show the selection results for Case 1- 5 in Simulation 1. For random effects selection, $\mathrm{BIC_R}$ and $\mathrm{GCV_R}$ perform identical by having same values of CZR, IZR, and CR, and they are slightly better than $\mathrm{AIC_R}$. For fixed effects selection, $\mathrm{AIC_F}$ and $\mathrm{GCV_F}$ perform similar by having very close values of CZF, IZF, and CF, but both of them are dominated by $\mathrm{BIC_F}$, which has CZF closer to 2 and CF closer to 100. In general, BIC outperforms AIC and GCV by having the highest values of "C", which is the frequency that selecting the correct true model. Therefore, we recommend BIC as the tuning parameter criterion for selecting and estimating both fixed and random effects. Moreover, we can observe that all three criteria perform better as sample size grows, for example, in Case 1 using AIC as tuning parameter, when $n = 50$, the proposed method has CZR = 2.19, IZR = 0, CR = 63, CZF = 1.55, IZF = 0, CZF = 64, and C = 40; when $n = 100$, the proposed method has CZR = 2.7, IZR = 0, CR = 78, CZF = 1.6, IZF = 0, CZF = 72, and C = 58, when $n = 200$, the proposed method has CZR = 2.78, IZR = 0, CR = 83, CZF = 1.69, IZF = 0, CZF = 75, and C = 64, all the values are getting closer to the true values as the sample size increases. Another interesting observation from the tables is that, no matter which criterion is employed, the values of IZF and IZR are always zero, meaning the important fixed and random effects can be always identified using the proposed selection method.

Table 5.1: Simulation results for Simulation 1 Case 1, using different tuning parameters.

n	Random Effects				Fixed Effects				Model
	Criterion	CZR	IZR	CR	Criterion	CZF	IZF	CF	C
	Truth	3	0	100	Truth	2	0	100	100
50	$\mathrm{BIC_R}$	2.2	0	64	$\mathrm{BIC_F}$	1.9	0	92	59
	$\mathrm{AIC_R}$	2.19	0	63	$\mathrm{AIC_F}$	1.55	0	64	40
	$\mathrm{GCV_R}$	2.2	0	64	$\mathrm{GCV_F}$	1.55	0	64	41
100	$\mathrm{BIC_R}$	2.72	0	80	$\mathrm{BIC_F}$	1.93	0	94	75
	$\mathrm{AIC_R}$	2.7	0	78	$\mathrm{AIC_F}$	1.6	0	72	58
	$\mathrm{GCV_R}$	2.72	0	80	$\mathrm{GCV_F}$	1.59	0	72	59
200	$\mathrm{BIC_R}$	2.8	0	83	$\mathrm{BIC_F}$	1.98	0	98	83
	$\mathrm{AIC_R}$	2.78	0	81	$\mathrm{AIC_F}$	1.69	0	75	62
	$\mathrm{GCV_R}$	2.8	0	83	$\mathrm{GCV_F}$	1.69	0	75	64

Tables 5.6 - 5.10 compare the linear mixed model selection in our proposed method (denoted

Table 5.2: Simulation results for Simulation 1 Case 2, using different tuning parameters.

n	Random Effects				Fixed Effects				Model
	Criterion	CZR	IZR	CR	Criterion	CZF	IZF	CF	C
	Truth	3	0	100	Truth	2	0	100	100
50	$\mathrm{BIC_R}$	2.21	0	63	$\mathrm{BIC_F}$	1.91	0	92	59
	$\mathrm{AIC_R}$	2.19	0	62	$\mathrm{AIC_F}$	1.62	0	64	40
	$\mathrm{GCV_R}$	2.21	0	63	$\mathrm{GCV_F}$	1.61	0	63	40
100	$\mathrm{BIC_R}$	2.72	0	82	$\mathrm{BIC_F}$	1.96	0	96	79
	$\mathrm{AIC_R}$	2.7	0	80	$\mathrm{AIC_F}$	1.62	0	64	50
	$\mathrm{GCV_R}$	2.72	0	82	$\mathrm{GCV_F}$	1.62	0	64	52
200	$\mathrm{BIC_R}$	2.78	0	79	$\mathrm{BIC_F}$	1.99	0	99	78
	$\mathrm{AIC_R}$	2.77	0	78	$\mathrm{AIC_F}$	1.67	0	74	58
	$\mathrm{GCV_R}$	2.78	0	79	$\mathrm{GCV_F}$	1.67	0	74	58

Table 5.3: Simulation results for Simulation 1 Case 3, using different tuning parameters.

n	Random Effects				Fixed Effects				Model
	Criterion	CZR	IZR	CR	Criterion	CZF	IZF	CF	C
	Truth	3	0	100	Truth	2	0	100	100
50	$\mathrm{BIC_R}$	2.14	0	64	$\mathrm{BIC_F}$	1.93	0	94	61
	$\mathrm{AIC_R}$	2.13	0	63	$\mathrm{AIC_F}$	1.66	0	74	47
	$\mathrm{GCV_R}$	2.14	0	64	$\mathrm{GCV_F}$	1.65	0	73	47
100	$\mathrm{BIC_R}$	2.73	0	81	$\mathrm{BIC_F}$	1.96	0	96	78
	$\mathrm{AIC_R}$	2.69	0	77	$\mathrm{AIC_F}$	1.69	0	76	58
	$\mathrm{GCV_R}$	2.73	0	81	$\mathrm{GCV_F}$	1.67	0	74	60
200	$\mathrm{BIC_R}$	2.78	0	83	$\mathrm{BIC_F}$	1.96	0	96	80
	$\mathrm{AIC_R}$	2.76	0	82	$\mathrm{AIC_F}$	1.68	0	74	63
	$\mathrm{GCV_R}$	2.78	0	83	$\mathrm{GCV_F}$	1.68	0	74	64

by OUR) with two existing selection procedures: AHN (Ahn, 2010) which is a distribution free method, and BKG (Bondell *et al.*, 2010) which is a maximum log-likelihood based approach. Tables 5.1 - 5.5 show that BIC works well for choosing the best λ for parameter tuning in linear mixed model selection, we therefore only report the results using BIC as tuning for all three procedures, and BIC is used for the proposed method in the rest of the dissertation.

In Cases 1 - 4, in terms of random effects selection, when $n = 50$, the proposed method provides smaller values of CZR than those for AHN and BKG, yet as the sample size increases to 100 and 200, the method has larger CZR values than those for the other two approaches. Mean-

Table 5.4: Simulation results for Simulation 1 Case 4, using different tuning parameters.

n	Random Effects				Fixed Effects				Model
	Criterion	CZR	IZR	CR	Criterion	CZF	IZF	CF	C
	Truth	3	0	100	Truth	2	0	100	100
50	$\mathrm{BIC_R}$	2.26	0	68	$\mathrm{BIC_F}$	1.96	0	96	66
	$\mathrm{AIC_R}$	2.24	0	66	$\mathrm{AIC_F}$	1.69	0	72	50
	$\mathrm{GCV_R}$	2.26	0	68	$\mathrm{GCV_F}$	1.68	0	72	51
100	$\mathrm{BIC_R}$	2.82	0	87	$\mathrm{BIC_F}$	1.99	0	99	86
	$\mathrm{AIC_R}$	2.79	0	84	$\mathrm{AIC_F}$	1.73	0	76	67
	$\mathrm{GCV_R}$	2.82	0	87	$\mathrm{GCV_F}$	1.73	0	76	67
200	$\mathrm{BIC_R}$	2.77	0	80	$\mathrm{BIC_F}$	2	0	100	80
	$\mathrm{AIC_R}$	2.75	0	78	$\mathrm{AIC_F}$	1.6	0	69	52
	$\mathrm{GCV_R}$	2.77	0	80	$\mathrm{GCV_F}$	1.6	0	69	53

Table 5.5: Simulation results for Simulation 1 Case 5, using different tuning parameters.

n	Random Effects				Fixed Effects				Model
	Criterion	CZR	IZR	CR	Criterion	CZF	IZF	CF	C
	Truth	3	0	100	Truth	2	0	100	100
50	$\mathrm{BIC_R}$	1.71	0	35	$\mathrm{BIC_F}$	1.87	0	88	33
	$\mathrm{AIC_R}$	1.67	0	32	$\mathrm{AIC_F}$	1.49	0	57	23
	$\mathrm{GCV_R}$	1.71	0	35	$\mathrm{GCV_F}$	1.48	0	57	24
100	$\mathrm{BIC_R}$	2.12	0	44	$\mathrm{BIC_F}$	1.94	0	96	44
	$\mathrm{AIC_R}$	2.08	0	40	$\mathrm{AIC_F}$	1.61	0	70	32
	$\mathrm{GCV_R}$	2.12	0	44	$\mathrm{GCV_F}$	1.6	0	70	34
200	$\mathrm{BIC_R}$	2.33	0	50	$\mathrm{BIC_F}$	1.98	0	98	48
	$\mathrm{AIC_R}$	2.28	0	46	$\mathrm{AIC_F}$	1.6	0	67	31
	$\mathrm{GCV_R}$	2.33	0	50	$\mathrm{GCV_F}$	1.57	0	64	32

while, our method outperforms the other two procedures by showing smaller values of IZR and large values of CR, meaning that the proposed method performs better in random effects selection. With regard to fixed effects selection, our method consistently performs better than AHN and BKG by showing higher values of CZF and CF, and lower values of IZF, illustrating that our method has a higher frequency of identifying the correct model structure for the fixed effects. For true model selection, the proposed procedure constantly has higher values of C than those for AHN and BKG, which means our method is more effective in identifying the correct structure of linear mixed model under normal assumption. On the other hand, the results for Case 5 is not surprising,

as both our method and BKG are likelihood based methods which depend on the normal assumption, while AHN is a robust method does not require any distributional assumption on the random effects and error terms, that is why AHN dominates our method and BKG in this case.

In general, the above results imply that when the error term is normally distributed, our method is the best choice regarding both random effects and fixed effects selection. However, the proposed method does not perform effectively if the normality assumption is violated.

Table 5.11 compares one iteration computation times (in minute) for implementing our proposed method and the BKG method for each case. We note that the computation time of our method is substantially shorter than that of BKG in almost every case, and the difference is even more significant when n is large. The results are reasonable since σ^2 is not included in the profile log-likelihoods, our method therefore involves lower dimension than all the other methods and the solution search is accordingly faster.

Table 5.6: Simulation results for Simulation 1 Case 1, using different selection methods.

n	Method	Random Effects			Fixed Effects			Model
		CZR	IZR	CR	CZF	IZF	CF	C
	Truth	3	0	100	2	0	100	100
50	OUR	2.2	0	64	1.9	0	92	59
	AHN	2.55	0.04	63	1.91	0.01	90	57
	BKG	2.30	0	40	1.89	0	89	37
100	OUR	2.72	0	80	1.93	0	94	75
	AHN	2.66	0	74	1.97	0	97	71
	BKG	2.4	0	52	1.86	0	86	49
200	OUR	2.8	0	83	1.98	0	98	83
	AHN	2.56	0	73	1.98	0	98	71
	BKG	2.59	0	69	1.75	0	75	54

5.2 Simulation 2

To further compare the simulation results with those for the existing selection approaches, we follow the setting in Bondell *et al.* (2010).

Consider the true model with $p = 9$ for fixed effects and $q = 4$ for random effects, the true parameter vector $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0, 0, 0, 0)^T$, and the true covariance matrix

Table 5.7: Simulation results for Simulation 1 Case 2, using different selection methods.

n	Method	Random Effects			Fixed Effects			Model
		CZR	IZR	CR	CZF	IZF	CF	C
	Truth	3	0	100	2	0	100	100
50	OUR	2.21	0	63	1.91	0	92	59
	AHN	2.48	0.05	56	1.91	0.06	90	52
	BKG	2.44	0	51	1.87	0	89	49
100	OUR	2.72	0	82	1.96	0	96	79
	AHN	2.57	0	68	1.95	0.03	93	65
	BKG	2.55	0	59	1.89	0	89	56
200	OUR	2.78	0	79	1.99	0	99	78
	AHN	2.37	0	61	1.98	0	98	59
	BKG	2.64	0	70	1.87	0	87	63

Table 5.8: Simulation results for Simulation 1 Case 3, using different selection methods.

n	Method	Random Effects			Fixed Effects			Model
		CZR	IZR	CR	CZF	IZF	CF	C
	Truth	3	0	100	2	0	100	100
50	OUR	2.14	0	64	1.93	0	94	61
	AHN	2.52	0.04	61	1.91	0.08	88	56
	BKG	2.42	0	51	1.87	0	88	47
100	OUR	2.73	0	81	1.96	0	96	78
	AHN	2.52	0	66	1.97	0.05	95	64
	BKG	2.50	0	57	1.89	0	90	54
200	OUR	2.78	0	83	1.96	0	96	80
	AHN	2.48	0	65	1.98	0	98	63
	BKG	2.73	0	78	1.81	0	81	63

$$\mathbf{D} = \begin{pmatrix} 9 & 4.8 & 0.6 & 0 \\ 4.8 & 4 & 1 & 0 \\ 0.6 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For this setting, the first column of $\mathbf{Z}_i$ consists of $\mathbf{1}'s$ for the subject-specific intercept, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ except the first column of $\mathbf{Z}_i$ are independently generated from a uniform $(-2, 2)$ distribution.

Table 5.9: Simulation results for Simulation 1 Case 4, using different selection methods.

n	Method	Random Effects			Fixed Effects			Model
		CZR	IZR	CR	CZF	IZF	CF	C
	Truth	3	0	100	2	0	100	100
50	OUR	2.26	0	68	1.96	0	96	66
	AHN	2.43	0.01	60	1.90	0.08	890	54
	BKG	2.38	0	41	1.71	0	74	31
100	OUR	2.82	0	87	1.99	0	99	86
	AHN	2.57	0	68	1.96	0.05	96	65
	BKG	2.51	0	56	1.72	0	72	36
200	OUR	2.77	0	80	2	0	100	80
	AHN	2.59	0	69	1.98	0	98	68
	BKG	2.83	0	86	1.83	0	83	71

Table 5.10: Simulation results for Simulation 1 Case 5, using different selection methods.

n	Method	Random Effects			Fixed Effects			Model
		CZR	IZR	CR	CZF	IZF	CF	C
	Truth	3	0	100	2	0	100	100
50	OUR	1.71	0	35	1.87	0	88	33
	AHN	2.72	0.12	66	1.98	0.08	98	65
	BKG	2.18	0.02	35	1.9	0	90	33
100	OUR	2.12	0	44	1.94	0	96	44
	AHN	2.7	0.01	74	1.94	0	94	70
	BKG	2.47	0	53	1.88	0	89	49
200	OUR	2.33	0	50	1.98	0	98	48
	AHN	2.79	0	79	1.98	0	98	77
	BKG	2.46	0	55	1.81	0	81	47

We further assume the variance $\sigma^2 = 1$. Two cases are considered:

- Case 1. In this case, we investigate the behavior of the proposed procedure in moderate samples, here we use $n = 30$ subjects and $n_i = 5$ observations per subject.

- Case 2. For this case, we examine the performance of the proposed procedure in larger samples, and therefore increase the sample size to $n = 60$, n_i = 10.

We generate 100 datasets for each example and calculate the rates in selecting the correct true model, fixed effects and random effects using the proposed method, denoted by OUR, then we

Table 5.11: Comparison of computation times (in minute) for each case in Simulation 1.

Case	Method	$n = 50$	$n = 100$	$n = 200$
Case 1	OUR	1.5	2.7	4.5
	BKG	2.9	9.0	29.4
Case 2	OUR	1.8	3.2	5.4
	BKG	4.3	11.9	27.9
Case 3	OUR	1.7	3.7	6.0
	BKG	3.7	10.0	27.6
Case 4	OUR	2.0	3.4	4.5
	BKG	1.4	3.3	10.4
Case 5	OUR	2.2	4.6	8.3
	BKG	3.4	5.6	26.7

Table 5.12: Simulation results for Simulation 2.

Method	%C	%CF	%CR	%C	%CF	%CR
(Case 1)				(Case 2)		
OUR	73	81	88	92	92	100
LPJ	61	79	79	88	91	97
PL	19	49	35	86	86	100
BKG	71	73	79	83	83	89
EGIC	47	56	52	48	59	53
RIC	59	59	68	77	79	81

compare them with those for the existing selection procedures in Table 5.12: LPJ (Lin *et al.*, 2013), PL (Peng and Lu, 2012), BKG (Bondell *et al.*, 2010), EGIC (Pu and Niu, 2006), RIC (Wolfinger, 1993). For fairness of comparison, BIC-selector is used for all the methods. The results for LPJ are obtained by running the R code the authors provided, and the results of PL, BKG, EGIC, and RIC are copied from Table 2 in Peng and Lu (2012) and Table 1 in Bondell *et al.* (2010).

Let $\%$C, $\%$CF and $\%$CR be the percentages of times that the correct true model, fixed effects and random effects are selected. Table 5.12 shows that in Case 1, that is, $n = 30$, $n_i = 5$, our method selects the true model, fixed effects and random effects with 73%, 81% and 88%, respectively. For most of the other methods, the corresponding rates are much smaller than these values. For instance, the proportions of correctly choosing the true model, fixed effects and random effects using PL are respectively only 19%, 49% and 35%. With regard to these small selection

rates, Peng and Lu (2012) remarked that in small samples, they were caused by large standard errors of random effects b_i. In the same scenario, the simulation results in Table 5.12 show that our method performs better than the other ones.

As the sample size grows in Case 2, that is, $n = 60$, $n_i = 10$, each of all the methods identifies the correct model, fixed effects and random effects with increasing rates. Our method selects the correct random effects with 100%, and selects the correct fixed effects selection with 92%, indicating that the proposed method outperforms all the other approaches.

5.3 Simulation 3

In addition to calculating the rates of selecting the correct true model, fixed effects and random effects, we introduce three model accuracy measures to inspect the performance of the proposed method.

First, the Kullback-Leibler discrepancy (KLD, Kullback and Leibler, 1951) is adopted to measure the discrepancy between the true model and the candidate model. Small values of the KLD indicate that the fitted model is close to the data-generating model. The KLD is expressed as

$$\mathrm{KLD} = \mathrm{E}\left\{\log f\left(\mathbf{Y}, \mathbf{X}, \mathbf{Z}|\boldsymbol{\phi}\right) - \log f(\mathbf{Y}, \mathbf{X}, \mathbf{Z}|\hat{\boldsymbol{\phi}})\right\}.$$

Second, the mean square error (MSE) is employed to quantify the difference between the fixed effects parameters and their estimates. Small values of MSE show that the obtained estimates of fixed effects are satisfactory. The MSE is given by

$$\mathrm{MSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathrm{E}(\mathbf{X}\mathbf{X}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

We also consider the quadratic loss error (QLE) to assess the difference between the covariance matrix and its estimate. Small values of QLE imply that the obtained estimates of random effects

are effective. The QLE is defined as

$$\mathrm{QLE} = [\mathrm{tr}\left(\hat{\mathbf{D}} - \mathbf{D}\right)^2]^{1/2}.$$

For this setting, we set $n_i = 12$ observations per subject and have the true model with $p = 8$ for fixed effects and $q = 5$ for random effects. The true parameter vector is set as $\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)^T$, and the true covariance matrix

$$\mathbf{D} = \begin{pmatrix} 1 & 0.5 & 0.25 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 \\ 0.25 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We consider two cases as follows:

- Case 1. The generated data have an autoregressive covariance structure. The pairwise correlation between x_j and x_k in the 12×8 matrix $\mathbf{X}_i$ is $\rho^{|j-k|}$, and the pairwise correlation between z_l and z_m in the 12×5 matrix $\mathbf{Z}_i$ is $\rho^{|l-m|}$, where ρ is the first order correlation coefficient. The data are generated under different scenarios: $n = 30, 60, 90$, $\rho = 0.3, 0.5, 0.8$, and $\sigma^2 = 1, 2$. We further assume $\mathbf{X}_i$ and $\mathbf{Z}_i$ follow multivariate normal distribution with mean zero.

- Case 2. In this case, the data are generated from a compound symmetry covariance structure. We set the pairwise correlation between x_j and x_k in the 12×8 matrix $\mathbf{X}_i$, and the pairwise correlation between z_l and z_m in the 12×5 matrix $\mathbf{Z}_i$ to be ρ.

Let $\mathrm{KLD_p}$ be the KLD of the selected model by the proposed penalized method, and $\mathrm{KLD_f}$ be the KLD of the full model obtained by REML. Similarly, we define $\mathrm{MSE_p}$, $\mathrm{MSE_f}$, $\mathrm{QLE_p}$, and $\mathrm{QLE_f}$. We then compute the ratios $\frac{\mathrm{KLD_p}}{\mathrm{KLD_f}}$, $\frac{\mathrm{MSE_p}}{\mathrm{MSE_f}}$, and $\frac{\mathrm{QLE_p}}{\mathrm{QLE_f}}$ for each of the 100 simulated datasets,

and calculate the median of the ratios, denoted as MKLD, MMSE and MQLE. In addition, to estimate the standard error of MKLD, we generate a 100 bootstrapped sample from the $\frac{\mathrm{KLD_p}}{\mathrm{KLD_f}}$ ratios, then calculate the bootstrapped sample median. We repeat this process 500 times. The estimated standard error is the standard deviation of the 500 bootstrapped sample medians. The standard errors of MMSE and MQLE can be obtained in the same way.

Table 5.13: Simulation results for Simulation 3 Case 1.

Case	MKLD	MMSE	MQLE	%C	%CF	%CR
$\rho = 0.3$						
$n = 30, \sigma^2 = 1$	.815(.059)	.407(.040)	1.139(.052)	70	92	73
$n = 30, \sigma^2 = 2$	.664(.304)	.648(.055)	.824(.101)	33	60	34
$n = 60, \sigma^2 = 1$	.836(.028)	.460(.034)	1.067(.028)	88	97	91
$n = 60, \sigma^2 = 2$	.601(.191)	.631(.057)	.882(.293)	35	76	36
$n = 90, \sigma^2 = 1$	.837(.017)	.355(.037)	1.076(.051)	88	95	92
$n = 90, \sigma^2 = 2$	.501(.125)	.511(.059)	1.574(.121)	31	74	31
$\rho = 0.5$						
$n = 30, \sigma^2 = 1$	.693(.085)	.347(.034)	1.114(.073)	72	94	75
$n = 30, \sigma^2 = 2$	1.119(.204)	.595(.058)	.518(.206)	31	62	33
$n = 60, \sigma^2 = 1$	.788(.030)	.357(.031)	1.044(.031)	86	96	90
$n = 60, \sigma^2 = 2$	.372(.093)	.620(.056)	1.461(.154)	24	59	27
$n = 90, \sigma^2 = 1$	.821(.038)	.379(.037)	1.086(.043)	87	98	89
$n = 90, \sigma^2 = 2$	.443(.114)	.522(.052)	1.769(.078)	24	77	24
$\rho = 0.8$						
$n = 30, \sigma^2 = 1$	.859(.046)	.494(.049)	1.105(.064)	56	83	63
$n = 30, \sigma^2 = 2$	1.122(.039)	.625(.088)	.467(.132)	36	64	39
$n = 60, \sigma^2 = 1$	.840(.037)	.354(.030)	1.121(.038)	82	94	86
$n = 60, \sigma^2 = 2$	1.136(.009)	.402(.033)	.278(.028)	64	88	68
$n = 90, \sigma^2 = 1$	.848(.026)	.319(.036)	1.051(.053)	88	97	91
$n = 90, \sigma^2 = 2$	1.126(.005)	.367(.027)	.196(.020)	66	90	68

Tables 5.13 and Table 5.14 individually summarize the simulation results of Case 1 and Case 2 in Simulation 3. Although the two tables feature the simulation results for the simulated data which are generated from different covariance structures, we observe that the selection rates of the true model are all very high. We can conclude that our procedure is robust to covariance structures. For the settings with high correlation, such as $\rho = 0.8$, the selection rates for the true model, correct fixed effects and random effects are all quite optimal, and we therefore claim that

Table 5.14: Simulation results for Simulation 3 Case 2.

Case	MKLD	MMSE	MQLE	%C	%CF	%CR
$\rho = 0.3$						
$n = 30, \sigma^2 = 1$	.776(.062)	.425(.043)	1.128(.078)	65	87	71
$n = 30, \sigma^2 = 2$	.477(.323)	.656(.122)	.881(.168)	28	54	34
$n = 60, \sigma^2 = 1$	.784(.037)	.356(.044)	1.128(.049)	85	96	88
$n = 60, \sigma^2 = 2$	1.123(.058)	.596(.065)	.288(.039)	48	76	49
$n = 90, \sigma^2 = 1$	.867(.026)	.371(.029)	1.064(.027)	85	95	90
$n = 90, \sigma^2 = 2$	1.117(.004)	.507(.044)	.189(.014)	64	81	65
$\rho = 0.5$						
$n = 30, \sigma^2 = 1$	.751(.056)	.405(.044)	1.123(.063)	78	92	79
$n = 30, \sigma^2 = 2$	1.147(.122)	.628(.094)	.427(.153)	43	65	44
$n = 60, \sigma^2 = 1$	.839(.041)	.421(.038)	1.123(.037)	80	96	84
$n = 60, \sigma^2 = 2$	1.123(.005)	.496(.047)	.233(.022)	54	79	57
$n = 90, \sigma^2 = 1$	.875(.033)	.376(.025)	1.011(.029)	86	99	87
$n = 90, \sigma^2 = 2$	1.118(.006)	.442(.034)	.203(.017)	60	83	63
$\rho = 0.8$						
$n = 30, \sigma^2 = 1$	.747(.066)	.400(.026)	1.272(.080)	57	78	65
$n = 30, \sigma^2 = 2$	1.153(.013)	.510(.052)	.485(.063)	43	66	49
$n = 60, \sigma^2 = 1$	.791(.031)	.378(.042)	1.074(.047)	79	96	83
$n = 60, \sigma^2 = 2$	1.139(.005)	.394(.039)	.223(.012)	71	91	74
$n = 90, \sigma^2 = 1$	.847(.032)	.333(.025)	1.079(.057)	89	97	91
$n = 90, \sigma^2 = 2$	1.124(.004)	.340(.026)	.199(.010)	79	97	81

the proposed procedure copes proficiently with the model selection for highly correlated data. We also observe that our method performs better as the sample size increases, which confirms the asymptotic properties we present in Chapter 4, that is, when the sample size is large enough, the method can identify the correct model with probability one.

Regarding model accuracy, we notice that all of the MMSE values are much smaller than one, meaning the obtained estimates of fixed effects from the proposed method have smaller errors to the true parameters, compared with the REML estimates from the full model. Moreover, most of the MKLD are less than one, which shows the fitted model from our approach only has short distance to the data-generating model. Finally, the MQLE values are less than one in half of the scenarios, indicating our random effects estimates are also satisfactory. In general, these values illustrate that our approach significantly reduces the model error, and the fitted model by the proposed method

is close to the true model. The small numbers of standard errors in the parentheses illustrate the stability of our estimates.

We also notice that the performance becomes relatively worse when the variance of the error σ^2 increases, which we believe is a global problem for all model selection procedures.

CHAPTER 6 APPLICATIONS

The results from the simulation studies in Chapter 5 have demonstrated that the proposed procedure is quite efficient in selecting the best covariates and random covariance structure in linear mixed models and outperforms the existing selection methodologies in general. To further examine its effectiveness in mixed model selection, the proposed penalized method is utilized in two applications of the Amsterdam growth and health study data (Kemper,1995) and the colon cancer data (Fisher *et al.*, 2003) in this chapter.

6.1 The Amsterdam Growth and Health Study Data

6.1.1 Data Description

The data were collected to explore the relationship between lifestyle and health in adolescence and young adulthood. In growing towards independence, the lifestyle habits of teenagers change substantially with respect to physical activity, food intake, tobacco smoking, etc. Accordingly, their health perspective may also change. Individual changes in growth and development can be studied by observing and measuring the same participant over a long period of time. The Amsterdam growth and health longitudinal study was designed to monitor the growth and health of teenagers and to develop future effective interventions for adolescence. A total of 147 subjects in the Netherlands participated in the study, and they were measured over 6 time points, thus the total number of observations is 882. The continuous response variable of interest was the total serum cholesterol expressed in mmol/l. The five predictors used were:

1. $fitness$: fitness level at baseline measured as maximal oxygen uptake on a treadmill.
2. $bodyfat$: body fat estimated by the sum of the thickness of four skinfolds.
3. $smoking$: whether the subject smokes or not, 0= “no”, 1 =“yes”.
4. $gender$: 0 = “female”, 1= “male”.
5. $time$: measurement time, coded as 1, . . ., 6.

Figure 6.1 is the boxplot of the response over subjects, and it shows heterogeneity among the

147 subjects. The presence of heterogeneity is frequently undertaken by using a mixed model. The QQ-plot of the response variable is shown in Figure 6.2, and it can be easily figured that the normality assumption in model (3.1) in valid, so it is reasonable to use the proposed method for this data set.

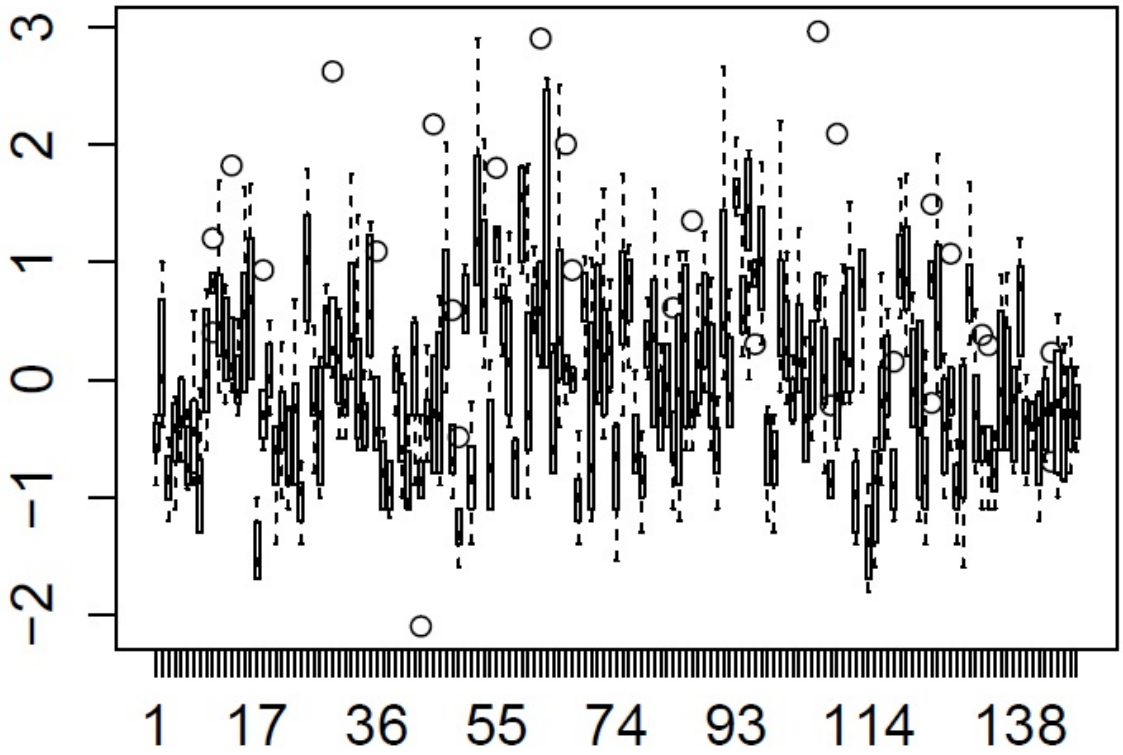
We will conduct mixed model selection on this data set using the proposed method, aiming to find the most appropriate mixed model for describing how the characteristics of the teenagers and the possibly existing random effects affect the total serum cholesterol.

6.1.2 Results Analysis

To inspect the relationship between the response and time, we plot the mean response profiles, mean response profiles by gender, and mean response profiles by smoking status, individually, all over time in Figure 6.3, 6.4 and 6.5. We can observe that the total serum cholesterol keeps decreasing during the first 4 time periods, and then goes up. In terms of gender effects, the total serum cholesterol of males are higher than females at the first time point, and after this time period, not significant different though, the total serum cholesterol of males consistently lower than females on average. Yet we did not detect any significant serum cholesterol difference between smokers and non-smokers.

Twisk (2003) studied this data by using various longitudinal data analysis techniques. Ahn *et al.* (2012) conducted the linear mixed model selection by two types of penalties, a hard thresholding operator (HARD) and a sandwich type soft thresholding penalty (SW). It has been shown that both HARD and SW methods are effective in identifying the correct mixed model structure with regard to selection accuracy and computation cost. For comparison, we follow this paper and center the response and then standardize all the predictors, so the fitted model does not allow an intercept for the fixed effects, but a random intercept is included. We then fit the model with all the five covariates for both the fixed and random effects by the proposed method (OUR), and compare our estimates with those for HARD, SW and REML methods. The REML estimates are obtained by the `lmer` function from `lme4` package in **R**, and the results of HARD and SW are copied from Table 5 in Ahn *et al.* (2012).

Figure 6.1: Boxplot of a response variable over subjects.



The estimation and selection results are summarized in Table 6.1. For the fixed effects selection, our method identifies $bodyfat$ and $time$ as significant, along with HARD, SW, and REML estimation. In the analysis of REML, such two variables have t-statistics of 5.73 and 7.31, which are the only two significant fixed effects. We can observe that the fitted fixed effects coefficients of the method are similar to those obtained by the other three methods. For the random effects selection, HARD selects $intercept$, $smoking$ and $gender$, SW chooses the $intercept$, $fitness$ and $gender$, and our approach recognizes $intercept$ and $gender$ as significant random effects, which contains the overlaps of HARD and SW estimates. The QQ-plot and histogram of the residuals for the model selected by the proposed method are plotted in Figure 6.6 and 6.7, the normality assumption approximately holds in the residuals. From this point, we remark that our method combines the strengths of such two methods and therefore identifies the most appropriate mixed model.

Figure 6.2: QQ plot of the response variable.

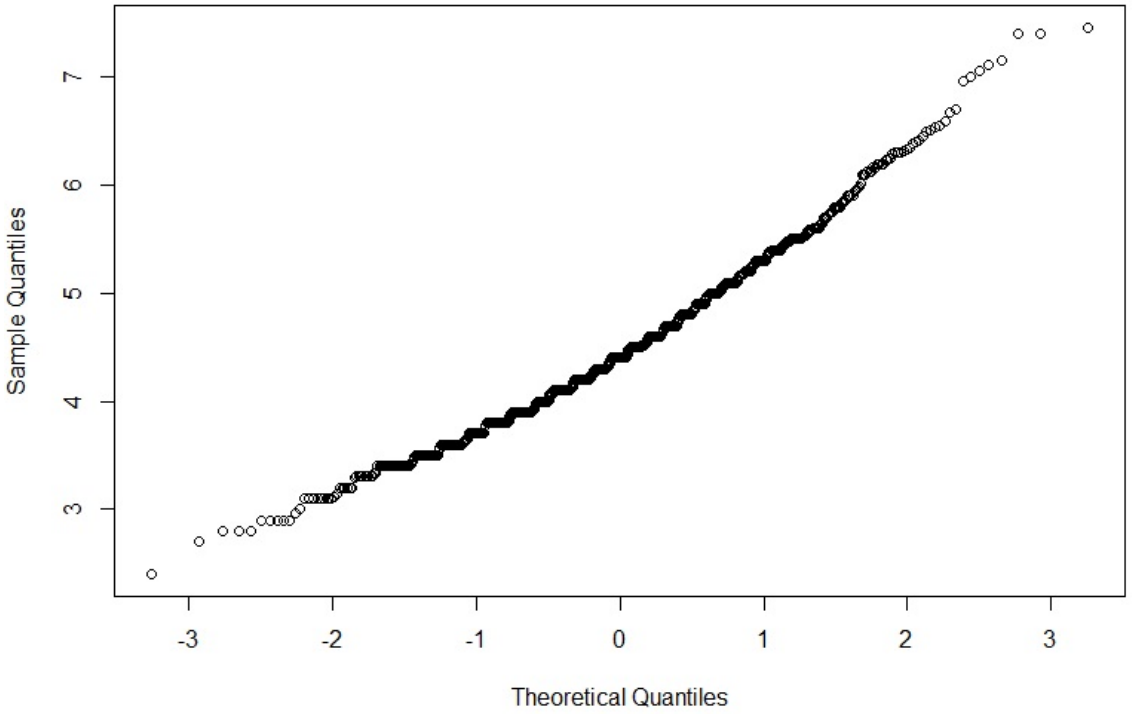


Figure 6.3: Plot of mean response profiles over time.

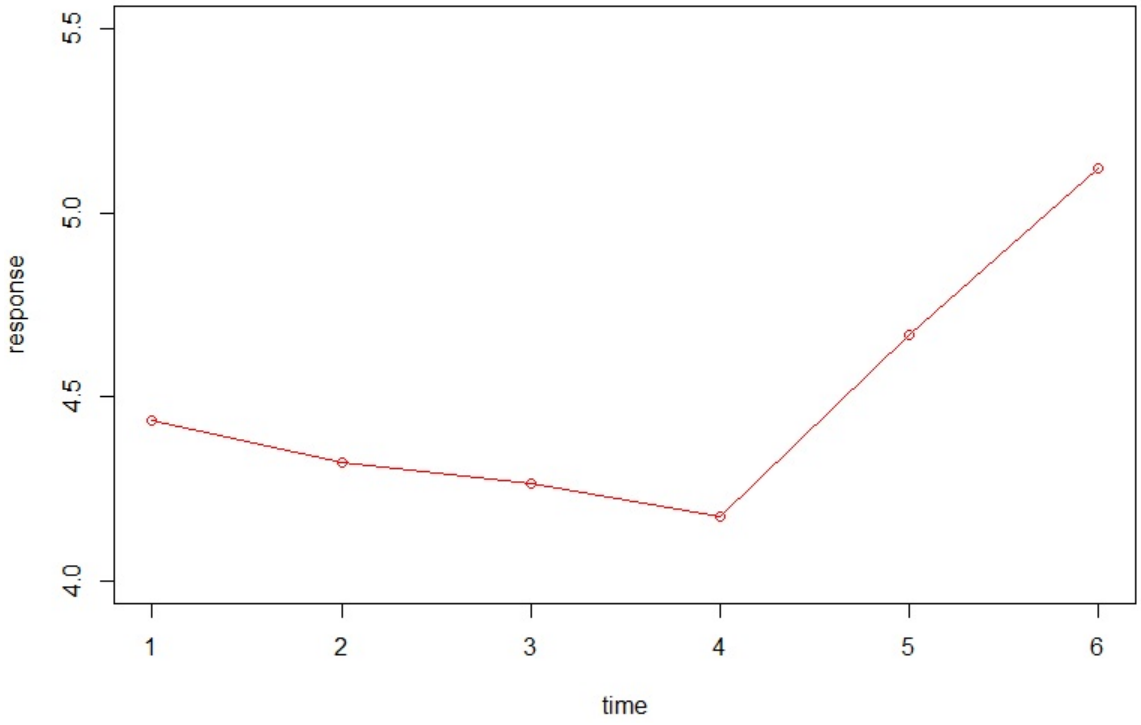


Figure 6.4: Plot of mean response profiles over time by gender.

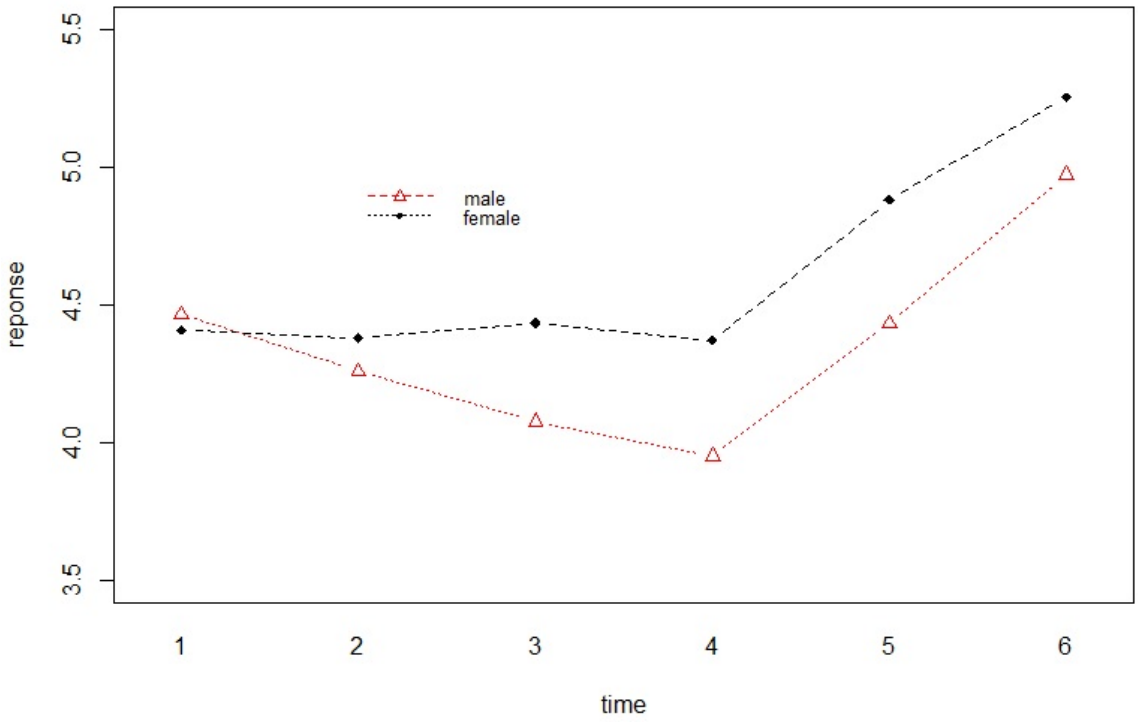


Figure 6.5: Plot of mean response profiles over time by smoking.

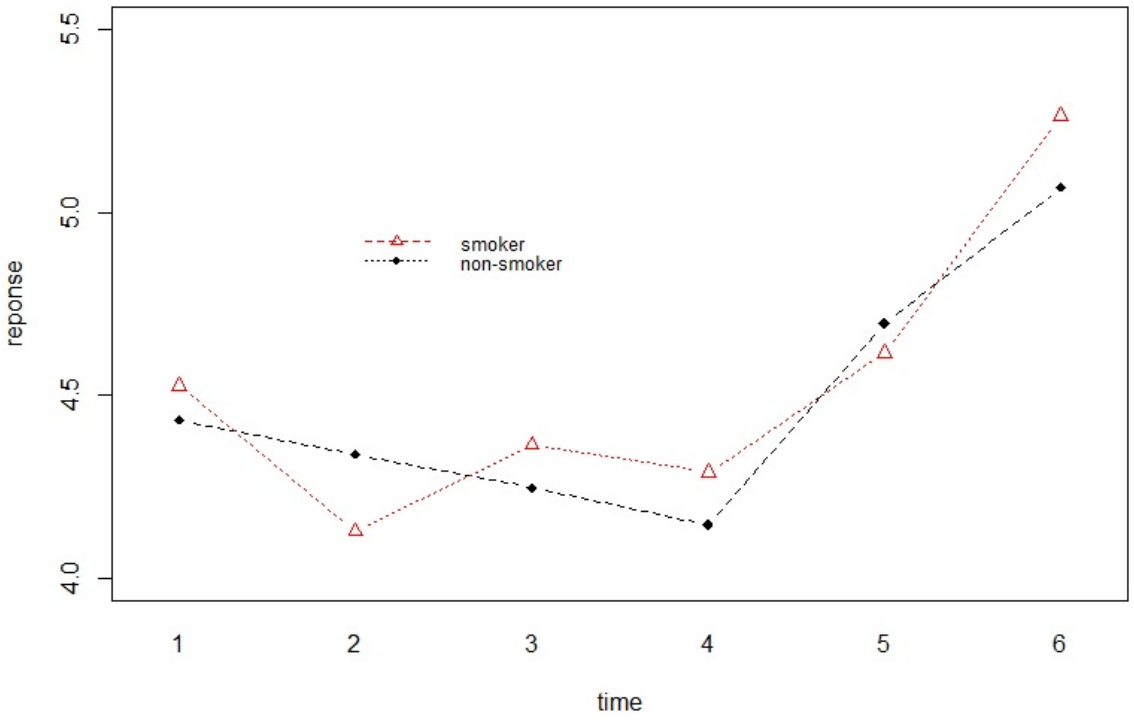


Table 6.1: Results for the Amsterdam growth and health study data.

Fixed Effects (β)	REML	HARD	SW	OUR
$fitness$	-0.039	0	0	0
$bodyfat$	0.194	0.174	0.165	0.170
$smoking$	-0.038	0	0	0
$gender$	0.083	0	0	0
$time$	0.165	0.156	0.167	0.165
Random Effects (**D**)	REML	HARD	SW	OUR
$intercept$	0.145	0.405	0.347	0.017
$fitness$	0.025	0	0.006	0
$bodyfat$	0.042	0	0	0
$smoking$	0.011	0.149	0	0
$gender$	0.249	0.668	0.624	0.888
$time$	0.037	0	0	0

Figure 6.6: QQ-plot of the residuals for the model selected by the proposed method.

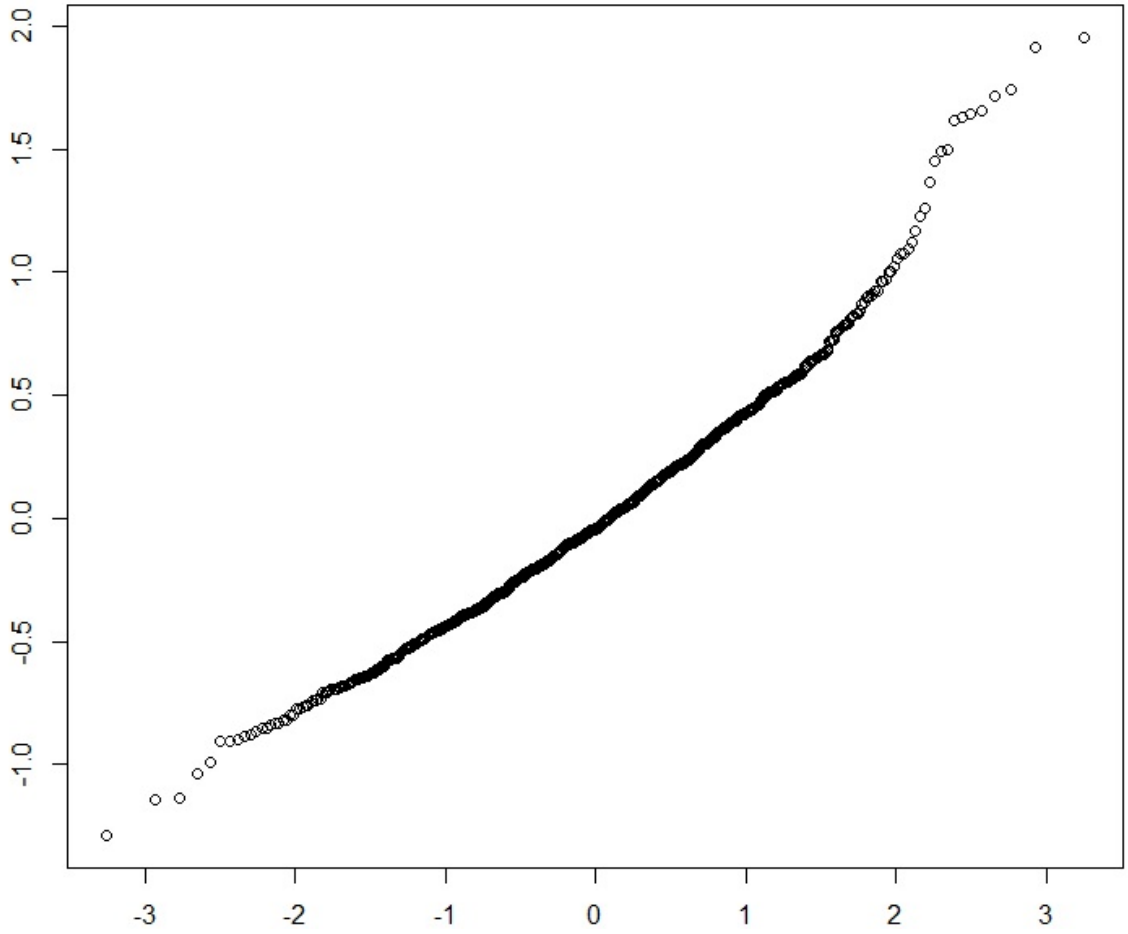
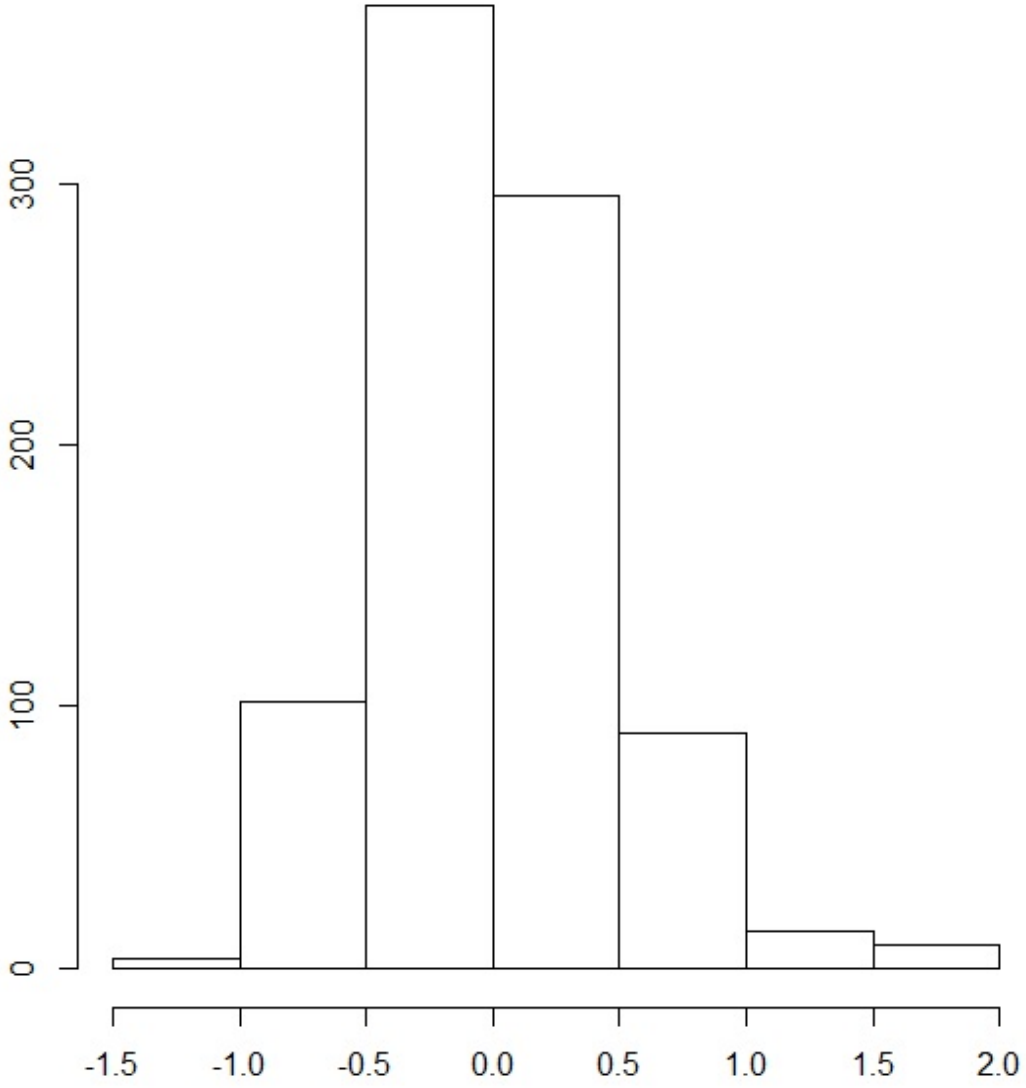


Figure 6.7: Histogram of the residuals for the model selected by the proposed method.



6.2 The Colon Cancer Data

6.2.1 Data Description

Colon cancer is formed in the tissues of the colon, which is the longest part of the large intestine. Most colon cancers are adenocarcinomas (cancers that begin in cells that make and release mucus and other fluids). By *United States Cancer Statistics: 1999-2011 Incidence and Mortality Web-based Report*, colon cancer is the second leading cause of cancer-related deaths and the third most common cancer in men and in women in the United States. Figure 6.8 shows the distribution

of colon cancer diagnosis by stage. It can be seen that most of the colon cancer patients (76 %) are in the first three stages. It has been studied that a Stage I cancer has a survival rate of 80-95 percent. Stage II tumors have survival rates ranging from 55 to 80 percent. A stage III colon cancer has about a 40 percent chance of cure and a patient in stage IV has only a 10 percent chance of recovery. Figure 6.8 demonstrates the colon cancer survival rates after initial diagnosis. We can observe that early detection and effective treatment are feasible and can often reduce mortality.

Figure 6.8: Colon cancer diagnosis by stage.

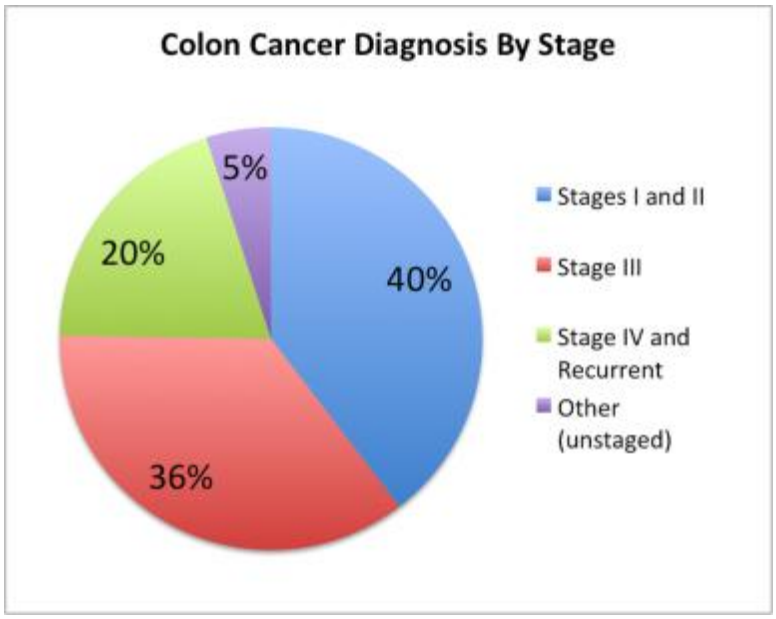
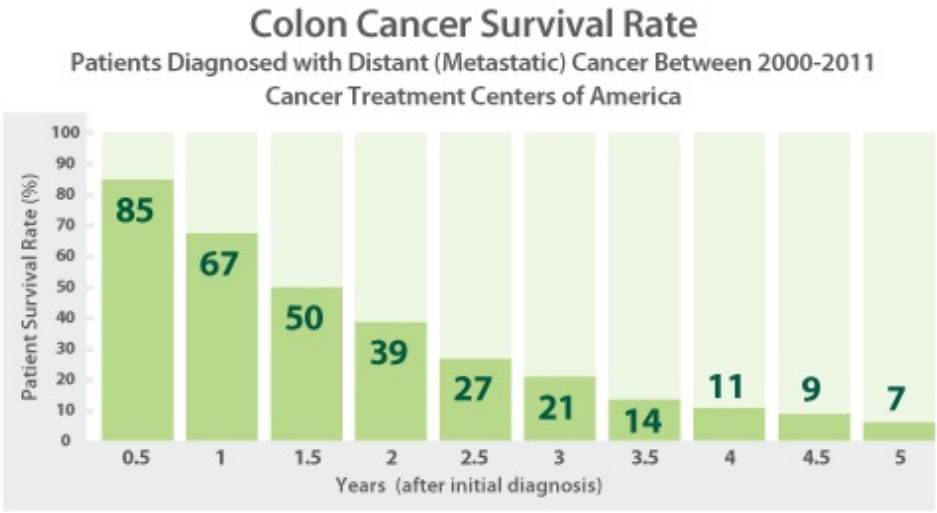


Figure 6.9: Colon cancer survival rate.



In recent years, the cost of colon cancer has been the topic of several scientific investigations. Precisely estimating the medical cost of colon cancer helps administrations policymakers making new policies for cancer prevention, screening, guidelines, and treatments.

The goal of our analysis is to estimate the cost attributable to colon cancer after initial diagnosis by cancer stage, comorbidity, treatment regimen, and other patient characteristics. The data reported aggregate Medicare spending on a cohort of 10,109 colon cancer patients up to 5 years after initial hospitalization, and these data are considered as the response for a linear mixed model. Congruently, the candidate predictors consist of characteristics of the patients, age, gender, race, stage, charlson comorbidity score, etc., and they were analogously reported. The measurements of regional medical intensity were also compatibly stated because overall spending on patients is quite possibly related to it.

In what follows, we will investigate mixed model selection on this data set using the proposed method, intending to figure out the most appropriate mixed model for describing how the characteristics of the patients and regional intensity of medical services and the possibly existing random effects affect the spending on colon cancer.

6.2.2 Results Analysis

The total number of the patients is $N = 10,109$, and for each patient, all the measurements are measured at 10 time points with a 6 month interval. It is hence easily assumed that a linear mixed model will fit the data. As shown in Chapter 4, the proposed penalized selection method will be adopted to choose the best mixed model, which is well-suited to the analysis of the relation between the response and the predictors and the random effects.

For the exploratory data analysis, we observe that the response data, total expense on colon cancer, are highly skewed to the right. We therefore use a log transformation to attenuate the skewness. The QQ-plots for both the total expense and its log transformation are all shown in Figure 6.10 and 6.11, and it can be easily figured that the natural log of the total expense is more appropriate for the assumption of normality presumed in model (3.1). Therefore, the response data for y_{it}, $i = 1, ..., N$, $t = 1, ..., 10$, are taken as the natural log of total spending rather than the raw

Figure 6.10: QQ plot of total expense.

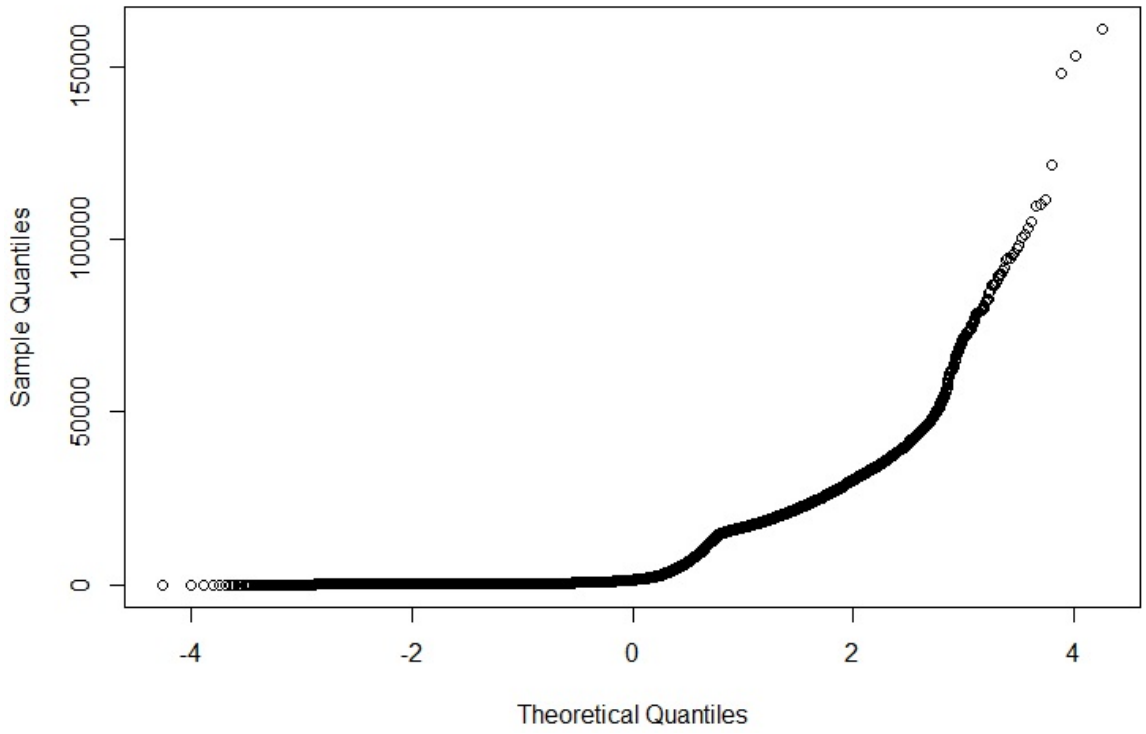
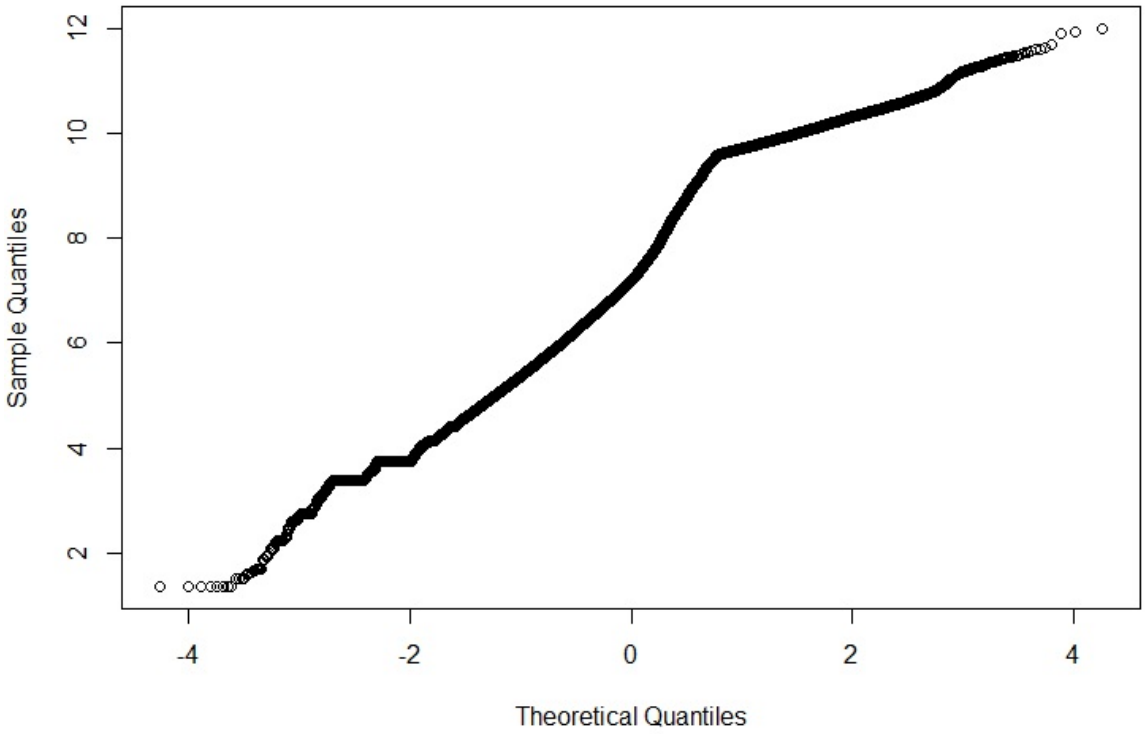


Figure 6.11: QQ plot of log total expense.



total spending.

To fit the linear mixed model, we consider the covariates as follows:

1. Gender: patient's gender, 1 = "female", 0 = "male".
2. Stage I: whether the cancer in the first stage, 1 = "yes", 0 = "no".
3. Stage II: whether the cancer in the second stage, 1 = "yes", 0 = "no".
4. Stage III: whether the cancer in the third stage, 1 = "yes", 0 = "no".
5. Charlson: patient's charlson comorbidity score.
6. HRR: regional medical intensity.
7. T1: whether the measurements are taken in the first 6 months, 1 = "yes", 0 = "no".
8. Time: in which intervals are cut by each 6 months and the measurements are taken.
9. Age: patient's age.
10. Race: patient's race, 1 = "African-American", 0 = "other".

To inspect the relationship between the spending and time, we plot the mean response profiles, mean response profiles by gender, and mean response profiles by race, individually, all over time in Figure 6.12, 6.13 and 6.14. We can observe that a large peak in medical spending during the first 6 months occurred after initial hospitalization, after that it went down smoothly. In terms of gender effects, both men and women spent almost the same in the first 6 months, and after this time period men spent consistently more than women on average. Yet we did not detect any significant expense difference between African-American and Non-African American.

First, we build a random intercept model to the data, which is a special case of linear mixed models. The random intercept model does not account for additional variation in spending across patients, which is expressed by

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \mathrm{Gender}_{it} + \beta_2 \mathrm{StageII}_{it} + \beta_3 \mathrm{StageIII}_{it} + \beta_4 \mathrm{Charlson}_{it} + \beta_5 \mathrm{HRR}_{it} \\ &\quad + \beta_6 \mathrm{T1}_{it} + \beta_7 \mathrm{Time}_{it} + \beta_8 \mathrm{Age}_{it} + \beta_9 \mathrm{Race}_{it} + b_i + \epsilon_{it}, \end{aligned} \tag{6.1}$$

where β's are the parameters coefficients for the fixed effects and the b_i is the random intercept.

Figure 6.12: Plot of mean response profiles over time.

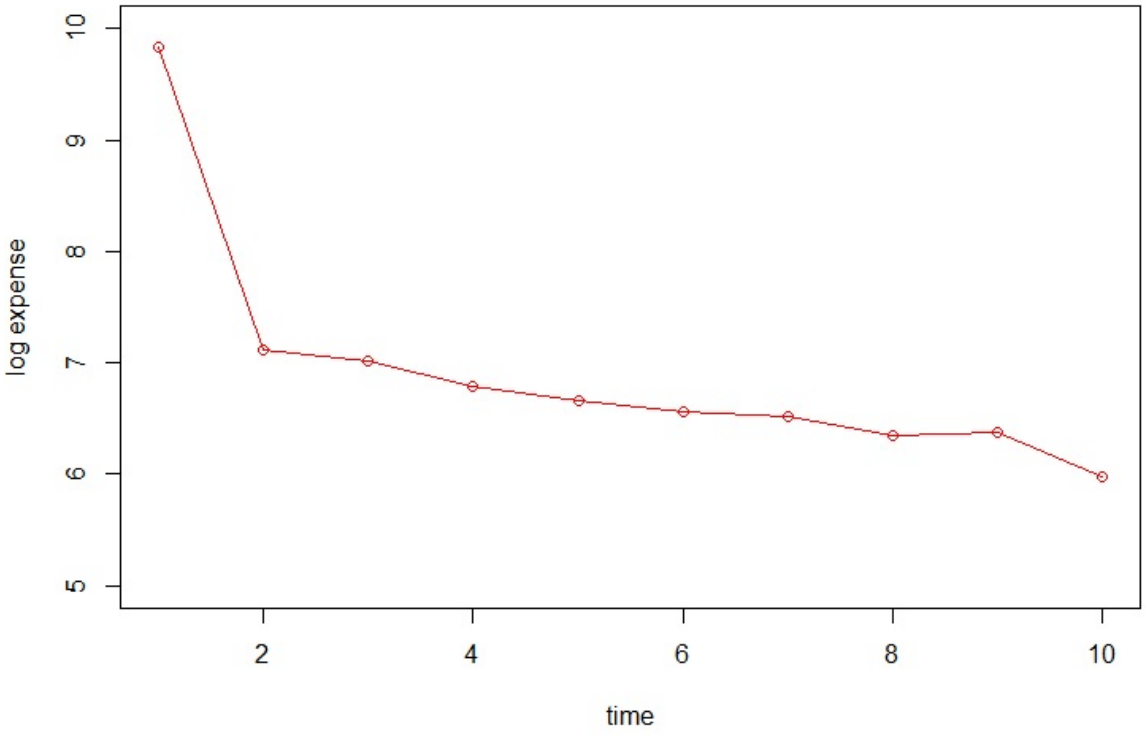


Figure 6.13: Plot of mean response profiles over time by gender.

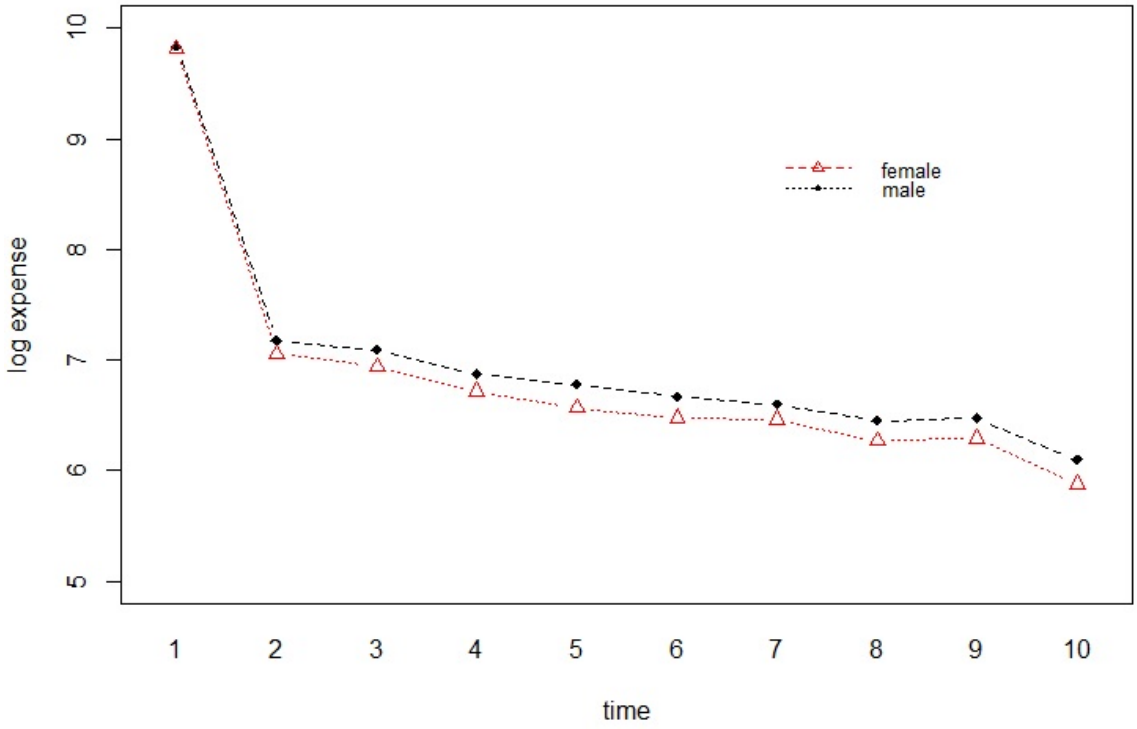
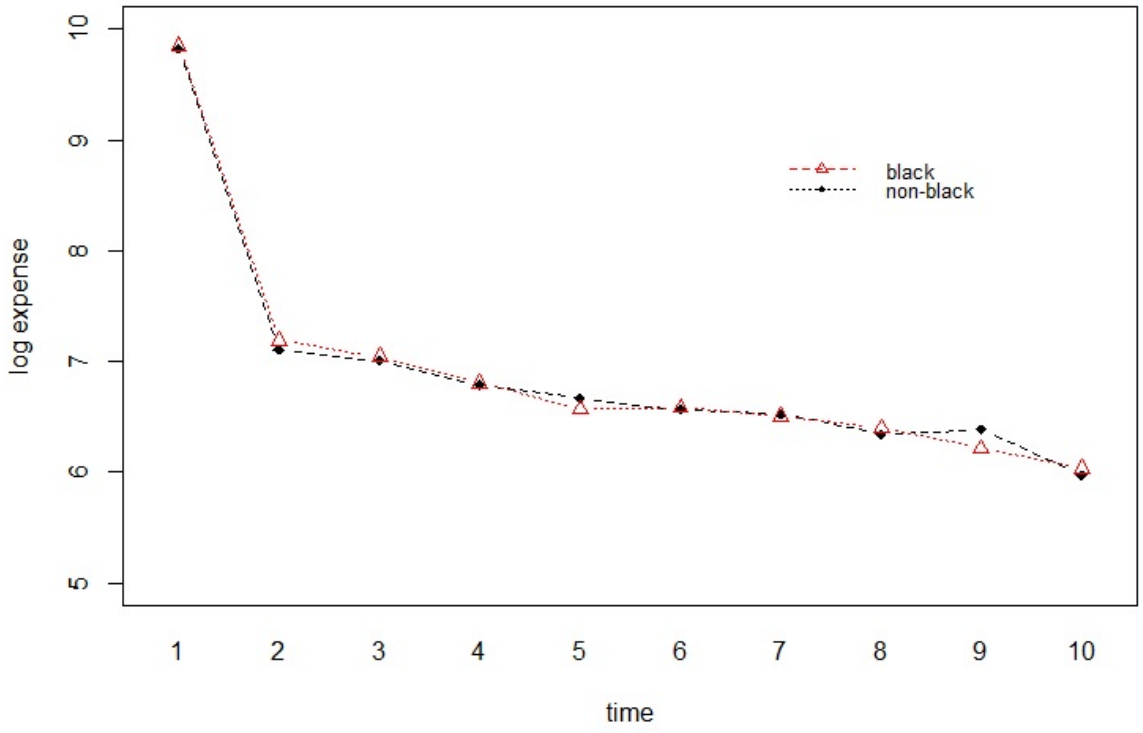


Figure 6.14: Plot of mean response profiles over time by race.



We note that none of the interaction term is included in the model because if the interaction term is significant, then the main effects are important as well. For simplicity, we only choose the important main effects.

Table 6.2 features both the REML estimates and the penalized profile log-likelihood estimates with the utility of the proposed method. It can be noticed that using the REML estimates for model (6.1), we obtain similar results as those in Demidenko and Stukel (2005), and all the predictors are significant based on the p-value. However, when we check the collinearity between the predictors, as shown in Table 6.3, we find that the collinearity occurs among the covariates, indicating that the REML results may not be the best choice for describing the data. On the contrary, the penalized method can attenuate the existing collinearity between the predictors, so for this colon cancer data, the proposed method will behave more effectively in settling the collinearity.

From Table 6.2, the results for the penalized method show that the coefficient of covariate Race is penalized to zero, meaning that race is a minor covariate and thus can be excluded for the selected model, which is compatible with the features from Figure 6.14. We also do not believe that

Table 6.2: Parameter estimates for fixed effect coefficients and random effect variance, using REML and the proposed method for model (6.1).

Method	REML		OUR
Fixed Effects ($\boldsymbol{\beta}$)	Coefficient	P-value	Coefficient
Intercept	6.673	0.000	6.667
Gender	-0.085	0.000	-0.073
Stage II	0.257	0.000	0.248
Stage III	0.402	0.000	0.400
Charlson	0.159	0.000	0.152
HRR	0.087	0.000	0.081
T1	2.609	0.000	2.607
Time	-0.115	0.000	-0.116
Age	-0.007	0.000	-0.006
Race	-0.070	0.006	0
Random Effect ($\mathbf{D}$)			
Intercept	0.005		0.004

the expense on curing a disease depends on race. From the values in Table 6.2, we can generally conclude that on average, comparing with Stage I patients, the medical spending is 25% higher for Stage II patients and is 40% higher for Stage III patients. In addition, the spending is 7% lower for females compared to that for males. From the coefficient for Time, we can see that the cost during the first 6 months after colectomy is extremely high; after this, the spending decreases by about 12% per 6-month interval. Independent of patient illness factors, the spending increases 8% with per $1,000 increase in the regional medical intensity. In regard to the related coefficients, the expense on colon cancer increases with severity of comorbidities and decreases with age.

Table 6.3: Correlations among predictors for the colon cancer data.

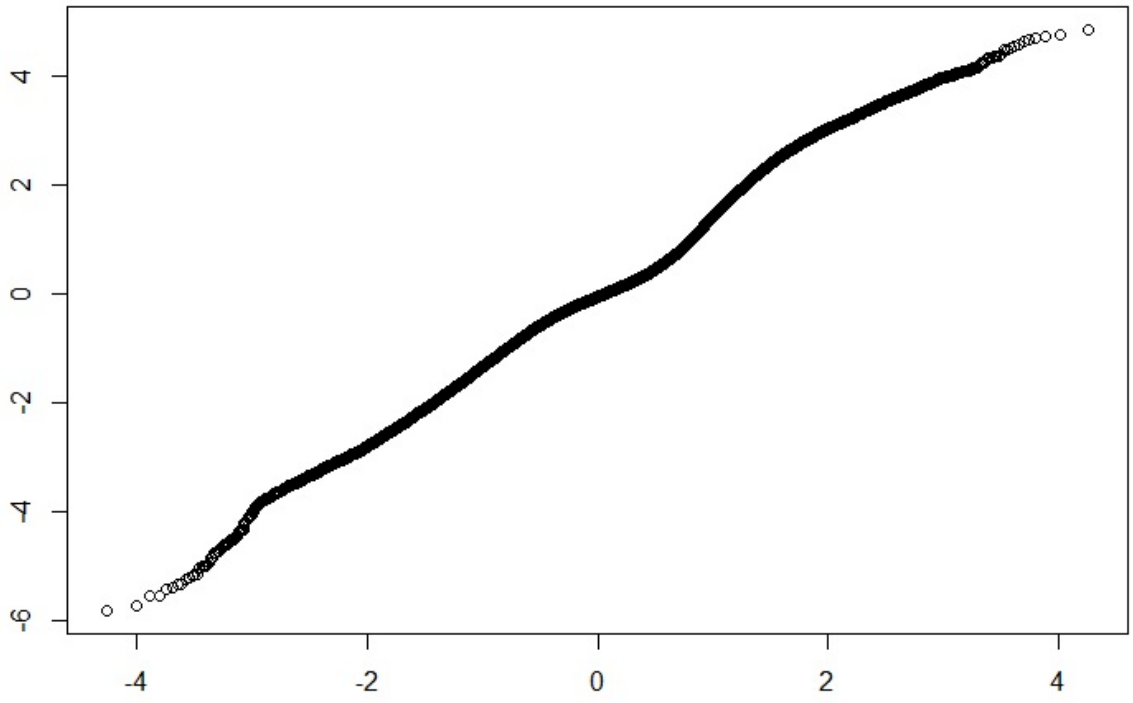
	Intercept	Gender	Race	StageII	StageIII	Age	Charlson	HRR	T1	Time
Intercept	1.00									
Gender	-0.22	1.00								
Race	-0.42	-0.04	1.00							
Stage II	-0.08	-0.01	0.02	1.00						
Stage III	-0.08	-0.02	0.03	0.21	1.00					
Age	-0.62	0.14	-0.04	0.05	0.05	1.00				
Charlson	-0.04	-0.12	0.02	0.10	0.09	0.02	1.00			
HRR	-0.36	-0.01	0.09	0.00	0.02	0.00	-0.01	1.00		
T1	-0.32	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	1.00	
Time	-0.50	0.00	0.00	0.01	0.02	0.01	0.01	0.01	0.53	1.00

Based on the penalized profile log-likelihood estimates in the fourth column of Table 6.2, we can build the model for describing the colon cancer expense as

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 \text{Gender}_{it} + \beta_2 \text{StageII}_{it} + \beta_3 \text{StageIII}_{it} + \beta_4 \text{Charlson}_{it} \\ &\quad + \beta_5 \text{HRR}_{it} + \beta_6 \text{T1}_{it} + \beta_7 \text{Time}_{it} + \beta_8 \text{Age}_{it} + b_i + \epsilon_{it}. \end{aligned} \tag{6.2}$$

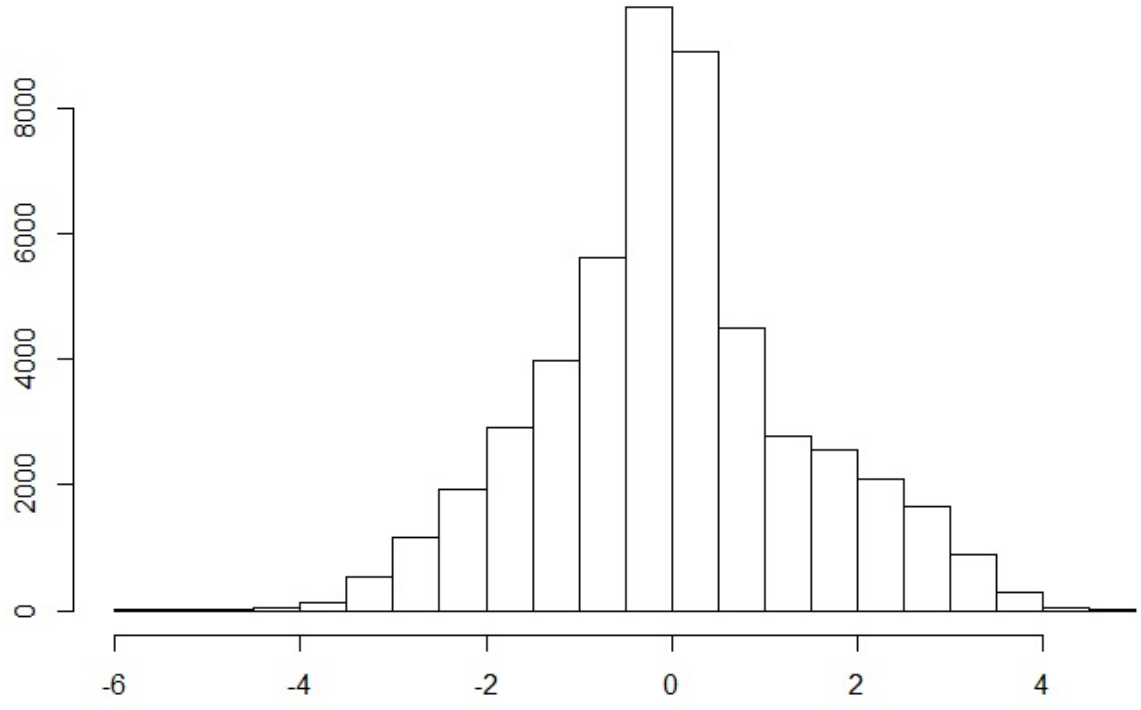
The QQ-plot and histogram of the residuals for model (6.2) are plotted in Figure 6.15 and 6.16, the normality assumption approximately holds with slight remaining skewness in the residuals.

Figure 6.15: QQ-plot of the residuals for model (6.2).



Second, we try another model by assuming that there are additional variations in spending across patients accounted for by gender, race, and age, and considering these three covariates of patient illness factors along with the intercept as the random effects, we have another linear mixed

Figure 6.16: Histogram of the residuals for model (6.2).



effects model written as

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 \text{Gender}_{it} + \beta_2 \text{StageII}_{it} + \beta_3 \text{StageIII}_{it} + \beta_4 \text{Charlson}_{it} \\ &\quad + \beta_5 \text{HRR}_{it} + \beta_6 \text{T1}_{it} + \beta_7 \text{Time}_{it} + \beta_8 \text{Age}_{it} + \beta_9 \text{Race}_{it} \\ &\quad + b_{i1} + b_{i2} \text{Gender}_{it} + b_{i3} \text{Age}_{it} + b_{i4} \text{Race}_{it} + \epsilon_{it}, \end{aligned} \tag{6.3}$$

where β's are the parameter coefficients for the fixed effects and the b's are the random effects.

The parameter estimates are reported in Table 6.4 for both the REML and proposed penalized methods for model (6.3). Table 6.4 shows that for model (6.3), in the REML method, compared to the other predictors, Age and Race are less important. So it is quite reasonable that in both the fixed effects and the random effects, the penalized method shrinks these two predictors or random slope parameter estimates to zero.

Under the proposed penalized method, it is interesting to compare the estimates with the random intercept model in Table 6.2. For the fixed effects, the parameters coefficients estimates are

rather similar except that the coefficient estimate for variable Age is penalized to zero. For the random effects, Race and Age become zero, yet only intercept and Gender remain in the final model. Note that Gender is selected in both the fixed and random effects, and the estimated coefficient for the fixed effect Gender and estimated variance for Gender's random slope indicate that the difference between gender exists, which is compatible with Figure 6.13.

Table 6.4: Parameter estimates for fixed effect coefficients and random effect variances, using REML and the proposed method for model (6.3).

Method	REML		OUR
Fixed Effects ($\boldsymbol{\beta}$)	Coefficient	P-value	Coefficient
Intercept	6.652	0.000	6.568
Gender	-0.099	0.000	-0.072
Stage II	0.257	0.000	0.254
Stage III	0.401	0.000	0.412
Charlson	0.159	0.000	0.152
HRR	0.087	0.000	0.050
T1	2.477	0.000	2.605
Time	-0.110	0.000	-0.116
Age	-0.006	0.019	0
Race	-0.071	0.006	0
Random Effects ($\mathbf{D}$)			
Intercept	0.336		0.336
Gender	0.003		0.004
Age	0.00006		0
Race	0.0002		0

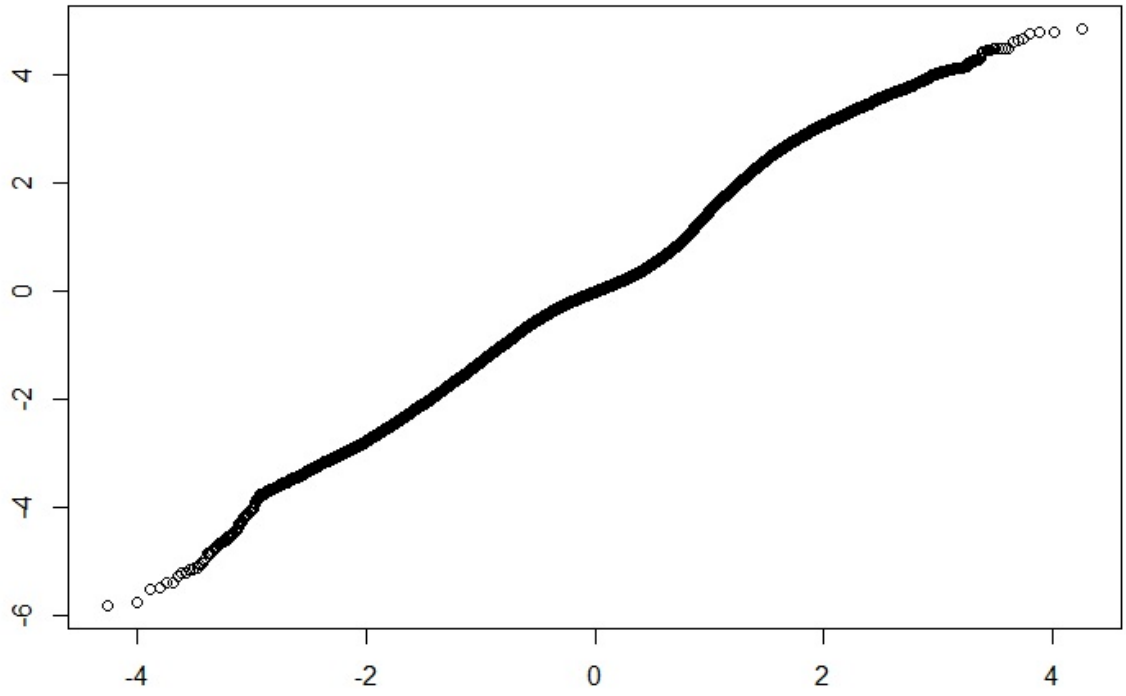
Based upon the penalized profile log-likelihood estimates in the fourth column of Table 6.4, we can build another model for describing the colon cancer expense as

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 \mathrm{Gender}_{it} + \beta_2 \mathrm{StageII}_{it} + \beta_3 \mathrm{StageIII}_{it} + \beta_4 \mathrm{Charlson}_{it} \\ &\quad + \beta_5 \mathrm{HRR}_{it} + \beta_6 \mathrm{T1}_{it} + \beta_7 \mathrm{Time}_{it} + b_{i1} + b_{i2} \mathrm{Gender}_{it} + \epsilon_{it}. \end{aligned} \tag{6.4}$$

Then we graph the QQ-plot and histogram of the residuals obtained for model (6.4), the plots demonstrate that the selected model fits the data even better than model (6.2). Thus, we prefer the model (6.4) as the selected model for describing the colon cancer expense, using the proposed

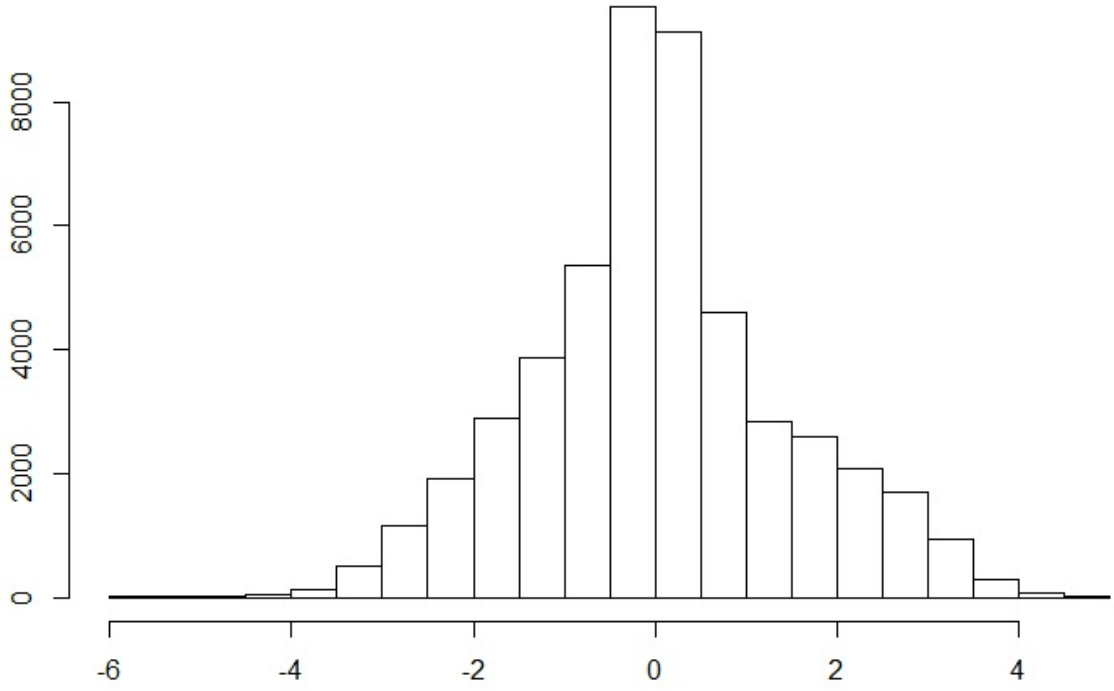
penalized mixed model selection approach.

Figure 6.17: QQ-plot of the residuals for model (6.4).



Regarding the selected model (6.4), we find that it is also supported by other research. For example, Lipska *et al.* (2006) indicated that comparing with women, men had higher incidence and prevalence of colorectal diseases, and greater risk of post-surgical anastomotic leakage, thus it was reasonable to conclude that on average the females spend less than the males for the colon cancer treatment. Additionally, Weichle *et al.* (2013) estimated the medical spending on colon cancer is 37% higher for Stage II patients, and 84% higher for Stage III patients, if Stage I was treated as reference. Although they had higher estimates than that in our approach for model (6.4) in Table 6.4, the same conclusion is drawn that the cost increases with the stage at diagnosis. We believe that with the disease develops to the further stage, more serious treatments will be applied and therefore causes a higher expense. They also showed that the cost is highly related to the comorbidity score, which is a method of categorizing the comorbidities of patients, and the higher this score was, the more likely the predicted outcome would result in mortality or higher resource use.

Figure 6.18: Histogram of the residuals for model (6.4).



CHAPTER 7 CONCLUSION REMARKS AND FUTURE RESEARCH

Linear mixed models involving both fixed effects and random effects are widely utilized to describe the complicatedly correlated data in a variety of fields. To aid the mixed model selection, we propose a two-stage model selection procedure by use of the adaptive LASSO penalized term for discretely selecting the random and fixed effects. To complete such a selection procedure, the restricted profile log-likelihood and profile log-likelihood functions are compatibly utilized. In this last chapter, we will provide summary and conclusions of the proposed method, and will discuss future possible research directions.

7.1 Conclusion Remarks

The proposed method is composed of two stages to separately penalize the parameters of interests, successfully respecting and accommodating the distinct properties between the random effects and fixed effects. In the first stage, the random effects are solidly selected; in the second stage, the fixed effects are selected. In the first stage of the proposed penalized method, the parameters estimators are obtained by maximizing the penalized restricted profile log-likelihood function, and they have the similar properties to those for the REML. The estimators in our approach for selecting the random effects are consequently more robust to outliers than those based on the ML methods (e.g., see Verbyla, 1993). Moreover, the equivalent REML method corrects the downward bias via taking into account the loss in degrees of freedom from the estimation of the fixed effects (e.g., see Lindstrom and Bates, 1988). When an appropriate model for the covariance is adopted, the correct covariance structure will be obtained and valid inferences for the fixed effects can then be made (e.g., see Fitzmaurice *et al.*, 2011, p. 165), and eventually prediction accuracy for future data is improved.

The profile log-likelihoods are adopted to select both the random effects and fixed effects in the proposed method. In contrast to the other log-likelihoods, the profile log-likelihood can not only catch enough and primary information for the model (e.g., see Fan and Li, 2012), but also requires

fewer iterations, the derivatives are somewhat simpler, and the convergence is more consistent in the Newton-Raphson optimization, since the variance σ^2 is not included in the iteration (e.g., see Lindstrom and Bates, 1988). Our method therefore involves lower dimension than all the other methods and thus owns the computation advantage.

We systematically study the theoretical properties of the proposed procedure. We prove that the procedure possesses the oracle properties in each stage, indicating that in each stage, for the corresponding parameter estimation, the parameter estimators asymptotically converge to normality in distribution with their true parameters as the mean. Moreover, the oracle property includes the possibility of actually performing model selection by shrinking the minor factors to zero. As the result of possessing the oracle properties, the right covariance structure and predictors are selected. The proofs theoretically solidify the optimal performance of the proposed procedure.

In practice, the proposed method improves computational efficiency and the quality of selection. In the setting of linear mixed model, to maximize the targeted quantity, e.g., the restricted profile log-likelihood for the random effects and the profile log-likelihood for the fixed effects, there is no closed form for the parameter estimation, all the methods accordingly use iterative way. While comparing with the Expectation-Maximization and the other well-known optimization algorithm, the Newton-Raphson algorithm we adopt converges steadfast and speedy. Dissimilar to the traditional methods, when the dimensions of fixed effects and random effects are quite large, the convergence in our approach still appears computational feasible and statistically accurate.

We employ the adaptive LASSO for the penalty term, and it is well known useful technique for simultaneous parameter estimation and variable selection, due to its simpler form and concave optimization property.

To investigate the behavior of the proposed model selection method, we conduct extensive simulation studies, and the results demonstrate that the proposed procedure outperforms the existing selection methodologies in terms of correct selection rates, computation time, the Kullback-Leibler discrepancy, the mean square error, and the quadratic loss error. It is worth mentioning that the proposed method performs quite efficaciously in selecting the most appropriate model for highly

correlated data and for different covariance structures. Further, although the optimality for the proposed method is derived in the asymptotic view, it still behaves noticeably in small to moderate sample sizes, as illustrated in the simulation studies.

We finally apply the proposed penalized method to the Amsterdam growth and health study data and the colon cancer data for further examining its effectiveness. The results illustrate that the proposed selection method can be employed to proficiently select and estimate the significant random and fixed effects for linear mixed model in real life.

7.2 Future Research

For the future work, we plan to continue working on topics in model selection. Our proposed method serves as a two-stage selection procedure to separately choose the important random effects and fixed effects, in the near future, we aim to develop a one-stage selection procedure for simultaneously selecting the significant random effects and fixed effects in the linear mixed model.

Recall the profile log-likelihood function in (4.1) is

$$p_F\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)=-\frac{1}{2}\sum_{i=1}^{n}\log\left|\mathbf{V}_i\right|-\frac{N}{2}\log\left(\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i\right).$$

Denote the $(p+k)\times 1$ parameter vector $\boldsymbol{\phi}=(\boldsymbol{\beta}^T,\boldsymbol{\theta}^T)^T$, then we can jointly select the fixed effects and random effects by maximizing the penalized profile log-likelihood. Two adaptive LASSO penalty terms can be used, one for the fixed effects, and one for the random effects. The penalized profile log-likelihood then is given by

$$Q_c(\boldsymbol{\phi})=p_F\left(\boldsymbol{\phi}\right)-\lambda\left(\sum_{j=1}^{p}w_{1j}\left|\beta_j\right|+\sum_{j=1}^{q}w_{2j}|d_j|\right). \tag{7.1}$$

To maximize $Q_c(\boldsymbol{\phi})$ in (7.1), the Newton-Raphson algorithm is applied as

$$\boldsymbol{\phi}_{b+1}=\boldsymbol{\phi}_b-\mathbf{M}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}\mathbf{sc}_{\boldsymbol{\phi}},\qquad b=0,1,\ldots,$$

where $\boldsymbol{\phi}_b$ is the current step value, and $\boldsymbol{\phi}_{b+1}$ is the updated value for the next step. $\mathbf{sc}_{\boldsymbol{\phi}}$ is $(p+k)\times$ 1 vector of the first derivative of $Q_c(\boldsymbol{\phi})$, and it is expressed as

$$\mathbf{sc}_{\boldsymbol{\phi}} = \left(\frac{\partial Q_c(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}}, \frac{\partial Q_c(\boldsymbol{\phi})}{\partial \theta} \right).$$

$\mathbf{M}_{\boldsymbol{\phi}\boldsymbol{\phi}}$ is $(p+k) \times (p+k)$ matrix of the second derivative of $Q_c(\boldsymbol{\phi})$, and it is given by

$$\mathbf{M}_{\boldsymbol{\phi}\boldsymbol{\phi}} = \begin{pmatrix} \frac{\partial^2 Q_c(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} & \frac{\partial^2 Q_c(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\theta}} \\ \frac{\partial^2 Q_c(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}} & \frac{\partial^2 Q_c(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} \end{pmatrix}.$$

The first derivative of the profile log-likelihood $p_F(\boldsymbol{\phi})$ in equation (7.1) is $\left(\frac{\partial p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}}, \frac{\partial p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \right)$, where

$$\frac{\partial p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}} = N \frac{\sum_{i=1}^n \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i},$$

and for $\theta_j, j = 1, 2, \ldots, k$,

$$\frac{\partial p_F(\boldsymbol{\phi})}{\partial \theta_j} = N \frac{\sum_{i=1}^n \mathbf{r}_i^T \mathbf{A}_{ij} \mathbf{r}_i}{\sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i} - \frac{1}{2} \sum_{i=1}^n \operatorname{tr} \left(\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_j} \right).$$

The second derivative of the profile log-likelihood $p_F(\boldsymbol{\phi})$ in equation (7.1) is given by

$$\begin{pmatrix} \frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} & \frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\theta}} \\ \frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\beta}} & \frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} \end{pmatrix},$$

where

$$\frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\frac{N}{2} \frac{\sum_{i=1}^n 2\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i * \sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i - \left(\frac{\sum_{i=1}^n -2\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i} \right)^2}{\left(\sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i \right)^2},$$

and for θ_j and θ_l, $j, l = 1, 2, \ldots, k$,

$$\frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}^T \partial \theta_j} = -\frac{N}{2} \frac{\sum_{i=1}^n \mathbf{X}_i^T (\mathbf{A}_{ij} + \mathbf{A}_{ij}^T) \mathbf{r}_i * \sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i - \sum_{i=1}^n \mathbf{r}_i^T \mathbf{A}_{ij} \mathbf{r}_i * 2\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\left(\sum_{i=1}^n \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i \right)^2},$$

and

$$\frac{\partial^2 p_F(\boldsymbol{\phi})}{\partial\theta_l\partial\theta_j} = -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left[-\mathbf{A}_{il}^T\frac{\partial\mathbf{V}_i}{\partial\theta_j}+\mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\theta_l\partial\theta_j}\right] - \frac{N}{2}\frac{\sum_{i=1}^{n}\mathbf{r}_i^T\frac{\partial\mathbf{A}_{ij}}{\partial\theta_l}\mathbf{r}_i * \sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i + \left(\frac{\sum_{i=1}^{n}-\mathbf{r}_i^T\mathbf{A}_{ij}\mathbf{r}_i}{\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}\right)^2}{\left(\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i\right)^2}.$$

For the penalty terms in function (7.1), we can use the local quadratic approximations discussed in (4.5) and (4.9), and then find the corresponding first and second derivatives. The Newton-Raphson algorithm thus can be utilized to search for the solution of maximizing the penalized profile likelihood function in (7.1). The process is repeated until the convergence is reached. The converged $\hat{\boldsymbol{\phi}}$ is the vector of selected and estimated fixed effects and random effects.

Compared with two-stage selection methods, the one-stage procedure owns at least two advantages. First, in a two-stage selection method, the incorrectly selected structure from the first step may affect the selection in the second step, yet such problem can be avoided in the one-stage procedure. Second, since the fixed and random effects are selected jointly, the computation cost and time are expected to be reduced.

Moreover, the proposed approach is based on the normality assumption for the data, and it may not perform effectively if the normality assumption is violated. In the future research, we might extend the current work to account for robustness to non-normality by using nonparametric methods.

We also note that missing values are quite common in longitudinal and cluster data, which leaves space for further research in model selection. Imputation handling will also be considered in addressing missing data in the future study.

In addition to studying on mixed model selection, we intend to further extend the current work to other modeling settings containing generalized linear mixed model and Cox proportional hazards model, both of which are widely used in biological and medical research.

Finally, we will develop other techniques to significantly improve model selection. We would

like to try different penalty terms and optimization algorithms which could increase correct selection rates and computational efficiency. For example, we look forward to utilizing eigenvalues to be the penalty term in the penalized method.

BIBLIOGRAPHY

[1] Ahn, M. (2010). Random effect selection in linear mixed models. *PhD thesis, North Carolina State University.*

[2] Ahn, M., Zhang, H. H. and Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica* **22**, 1539-1562.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267-81.

[4] Akaike, H. (1974). A new look at the model selection identification. *IEEE Trans. Automat. Control AC* **19**, 716-723.

[5] Andrieu, C., Djuric, P. M. and Doucet, A. (2001). Model selection by MCMC computation. *Signal Processing* **81**, 19-37.

[6] Azari, R., Li, L. and Tsai, C. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis* **50**, 3053-3066.

[7] Baragatti, M. (2011). Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Analysis* **6**, 209-230.

[8] Bernau, C., Augustin,T. and Boulesteix, AL. (2013). Correcting the optimal resamplingbased error rate by estimating the error rate of wrapper algorithms. *Biometrics* **69**, 693-702.

[9] Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069-1077.

[10] Bozdogan, H. (1987). Model selection and Akaikes information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

[11] Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**, 791-799.

[12] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

[13] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350-2383.

[14] Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: the X-random case. *International Statistical Review* **60**, 291-319.

[15] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

[16] Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference, 2nd edition.* Springer, New York.

[17] Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society Series B* **57**, 473-484.

[18] Chen, J. and Chen, Z. (1999). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

[19] Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762-769.

[20] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 317-403.

[21] Demidenko, E. (2013). *Mixed Models Theory and Applications*. Wiley, New York.

[22] Demidenko, E. and Stukel, T. A. (2005). Influence analysis for linear mixed-effects models. *Statistics in Medicine* **24**, 893-909.

[23] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.

[24] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461-470.

[25] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407-451.

[26] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[27] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

[28] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101-148.

[29] Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of Statistics* **40**, 2043-2068.

[30] Fisher, E. S., Wennberg, D. E., Stukel, T. A., Gottlieb, D. J., Lucas, F. L. and Pinder, E. L. (2003). The implications of regional variations in medicare spending. Part 1: the content, quality and accessibility of care. *Annals of Internal Medicine*, **138**, 273-287.

[31] Fitzmaurice, G., Laird, N. and Ware, J. (2011). *Applied Longitudinal Analysis, 2nd edition*. Wiley, New York.

[32] Foster, S. D., Verbyla, A. P. and Pitchford, W. S. (2007). Incorporating lasso effects into a mixed model for quantitative trait loci detection. *Journal of Agricultural, Biological, and Environmental Statistics* **12**, 300-314.

[33] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302-332.

[34] Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397-416.

[35] Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711-732.

[36] Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773-789.

[37] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B* **41**, 190-195.

[38] Harrell, F. E. (2001). *Regression Modeling Strategies.* Springer-Verlag, New York.

[39] Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93-108.

[40] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

[41] Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning; Data mining, Inference and Prediction*. Springer, New York.

[42] Henderson, C. R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics* **21**, 309-310.

[43] Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367-379.

[44] Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

[45] Hoerl, A. and Kennard, R. (1970). Ridge regression: application to nonorthogonal problems. *Technometrics* **12**, 69-82.

[46] Hurvich, C., Shumway. R. and Tsai, C. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709-719.

[47] Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

[48] Hurvich, C. and Tsai, C. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis* **14**, 271-279.

[49] Ibrahim, J. G., Zhu, H., Garcia, R. I. and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495-503.

[50] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

[51] Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36**, 1669-1692.

[52] Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine* **30**, 3050-3056.

[53] Kemper, H. (1995). The Amsterdam growth study: a longitudinal analysis of health, fitness and lifestyle. *HK Sport Science Monograph Series* **6**, Human Kinetics Publishers, Champaign IL.

[54] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, Morgan Kaufmann, pp. 1137-1143.

[55] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Statistics* **22**, 72-86.

[56] Kutner, M., Nachtsheim, C. and Neter, J. (2004). *Applied Linear Regression Models, 4th edition*. McGraw-Hill, New York.

[57] Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

[58] Lange, N. and Laird, N. (1989). The effect of covariance structures on variance estimation in balance-curve models with random parameters. *Journal of the American Statistical Association* **84**, 241-247.

[59] Lee, K., Sha, N., Dougherty, E., Vannucci, M. and Mallick, B. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19**, 90-97.

[60] Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics* **36**, 261-286.

[61] Lin, B., Pang, Z. and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics* **22**, 341-355.

[62] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measure data. *Journal of the American Statistical Association* **83**, 1014-1022.

[63] Lipska, M. A., Bisset, I. P., Parry, B. R. and Merrie, A. E. H. (2006). Anastomotic leakage after lower gastrointestinal anastomosis: men are at a higher risk. *ANZJ. Surg* **76**, 579-585.

[64] Mallows, C. L. (1973). Some comments on $\mathrm{C_p}$. *Technometrics* **15**, 661-675.

[65] Mallows, C. L. (1995). More comments on $\mathrm{C_p}$. *Technometrics* **37**, 362-372.

[66] McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models, 2nd edition*. Wiely, New York.

[67] Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science* **28**, 135-167.

[68] Ni, X., Zhang, D. and Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* **66**, 79-88.

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

[70] Pan, W. and Le, C. T. (2001). Bootstrap model selection in generalized linear models. *Journal of Agricultural, Biological and Environmental Statistics* **6**, 49-61.

[71] Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 273-289.

[72] Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis* **109**, 109-129.

[73] Pu, W. and Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis* **97**, 733-758.

[74] Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in regression problems. *Biometrika* **76**, 369-374.

[75] Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011). Estimation for high-dimensional linear mixed effects models using $\ell 1$-penalization. *Scandinavian Journal of Statistics* **38**, 197-214.

[76] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

[77] Shang, J. and Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis* **52**, 2004-2021.

[78] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B* **36**, 111-147.

[79] Sugiura, N. (1978). Further analysis of the data by Akaike information criterion and the finite corrections. *Communications in Statistics* **A7**, 13-26.

[80] Sun, W., Wang, J. and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research* **14**, 3419-3440.

[81] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 268-288.

[82] Tibshirani, RJ. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics* **3**, 822-829.

[83] Thompson, W. A. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics* **33**, 273-289.

[84] Twisk, J. W. (2003). *Applied Longitudinal Data Analysis for Epidemilogy-Practical Guide.* Cambridge University Press, New York.

[85] Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370.

[86] Verbyla, A. (1993). Modeling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society Series B* **55**, 493-508.

[87] Wang, H., Li, R. and Tsai C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

[88] Weakliem, D. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods and Research* **27**, 359-397.

[89] Weichle, T., Hynes, D. M., Durazo-Arvizu, R., Tarlov, E. and Zhang, Q. (2013). Impact of alternative approaches to assess outlying and influential observations on health care costs. *Springerplus* **2**, 614.

[90] Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation* **22**, 1079-1106.

[91] Wu, T. T. and Lange, K. (2008). Descent algorithms for Lasso penalized regression. *Annals of Applied Statistics* **2**, 224-244.

[92] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

[93] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301-320.

APPENDIX A SELECTED R PROGRAMS

- R code for simulation studies in Chapter 2.

```
library(MASS)
#########regular lasso######
lassoshoot=function(x,y,lambda){
  beta=lsfit(x,y)$coef[-1]
  #convergence flag
  status= 0
  # convergence tolerance
  tol= 1e-6;
  while( status == 0 )
  {
    #save current beta
    beta_old = beta
    # optimize elements of beta one by one
    for (j in 1:p){
      # optimize element j of beta
      # get jth col of X
      xj = x[,j]
      # get residual excluding ith col
      yj = (y - x%*%beta) + xj*beta[j]
      # calulate xj'*yj and see where it falls
      deltaj = t(yj)%*%xj
      if ( deltaj < -lambda ){
        beta[j] = ( deltaj + lambda )/(t(xj)%*%xj)
      } else if ( deltaj > lambda ){
        beta[j]= ( deltaj - lambda )/(t(xj)%*%xj)
```

```
      }else {beta[j] = 0}
    }
    #check difference between beta and beta_old
    if ( max(abs(beta - beta_old)) <= tol )
      break
  }
  return(beta)
}
############### adaptive lasso modified by pan, with weight is ols#######
panalassools=function(x,y,lambda){
  ######get ols estimate
  betaols=lsfit(x,y)$coef[-1]
  # convergence tolerance
  tol = 1e-6
  xx2 = t(x)%*%x*2
  xy2 = t(x)%*%y*2
  status = 0
  weight=1/abs(betaols)
  lams =lambda*weight
  beta=betaols
  while( status == 0 )
  {
    #save current beta
    beta_old = beta
    # optimize elements of beta one by one
    for (j in 1:p){
      # optimize element j of beta
      # get jth col of X
      xj = x[,j]
```

```
      # get residual excluding ith col
      yj = (y - x%*%beta) + xj*beta[j]
      # calulate xj'*yj and see where it falls
      deltaj = t(yj)%*%xj
      if ( deltaj < -lams[j] ){
        beta[j] = ( deltaj + lams[j] )/(t(xj)%*%xj)
      } else if ( deltaj > lams[j] ){
        beta[j]= ( deltaj - lams[j] )/(t(xj)%*%xj)
      }else {beta[j] = 0}
    }
    #check difference between beta and beta_old
    if ( max(abs(beta - beta_old)) <= tol )
      break
  }
  return(beta)
}
############### adaptive lasso modified by pan, with weight is ridge#######
panalassoridge=function(x,y,lambda){
  ########choose bestlambda in ridge regression and get the estimator
  lamb= seq(0, 10, length=101)
  gcv=numeric()
  for (i in 1:length(lamb)){
    beta.ridge=solve(t(x)%*%x+lamb[i]*diag(p))%*%t(x)%*%y
    gcv[i]=t(y-x%*%beta.ridge)%*%(y-x%*%beta.ridge)/
(n-sum(diag(solve(t(x)%*%x+lamb[i]*diag(p))%*%t(x)%*%x)))^2
  }
  ridgelambda=lamb[which.min(gcv)]
  ###ridge estimates
  ridgebeta=solve(t(x)%*%x+ridgelambda*diag(p))%*%t(x)%*%y
```

```
# convergence tolerance
tol = 1e-6
xx2 = t(x)%*%x*2
xy2 = t(x)%*%y*2
status = 0
weight=1/abs(ridgebeta)
lams =lambda*weight
beta=ridgebeta
while( status == 0 )
{
  #save current beta
  beta_old = beta
  # optimize elements of beta one by one
  for (j in 1:p){
    # optimize element j of beta
    # get jth col of X
    xj = x[,j]
    # get residual excluding ith col
    yj = (y - x%*%beta) + xj*beta[j]
    # calulate xj'*yj and see where it falls
    deltaj = t(yj)%*%xj
    if ( deltaj < -lams[j] ){
      beta[j] = ( deltaj + lams[j] )/(t(xj)%*%xj)
    } else if ( deltaj > lams[j] ){
      beta[j]= ( deltaj - lams[j] )/(t(xj)%*%xj)
    }else {beta[j] = 0}
  }
  #check difference between beta and beta_old
  if ( max(abs(beta - beta_old)) <= tol )
```

```
      break
  }
  return(beta)
}
##########generate data##############
####model 1 y=x*beta+sigma*epsilon###
###beta (3,1.5,0,0,2,0,0,0)
library(MASS)
# var cov matrix
sigma=matrix(0,8,8)
for (i in 1:8){
  for (j in 1:8){
    if (i==j) sigma[i,j]=1
    else sigma[i,j]=.5
  }
}
#define sigma
sig=3
## obs in each dataset
m=60
####truebeta
truebeta<-matrix(c(3,1.5,0,0,2,0,0,0),ncol=1)
lambda= exp(seq(0, 3, length=31))
Classo=Ilasso=Cpanalassools=Ipanalassools=numeric()
Cpanalassoridge=Ipanalassoridge=numeric()
rpelasso=rpelassopanalassools=rpepanalassoridge=numeric()
######simulate 100 replications
for ( b in 1:100){
  x=mvrnorm(n=m, numeric(8), sigma)
```

```
  e=rnorm(m)
  y=x%*%truebeta+sig*e
  n=length(t(y))
  p=ncol(x)
  ls=lsfit(x,y)$coeff[-1]
  b0=lsfit(x,y)$coeff[1]
  hsig=t(y-x%*%ls-b0)%*%(y-x%*%ls-b0)/(n-p)
  bicridge=bicols=biclasso=numeric()
  for (j in 1:length(lambda)){
    #########three procedures
    betapanalassoridge=panalassoridge(x,y,lambda[j])
    betapanalassools=panalassools(x,y,lambda[j])
    betapanlasso=lassoshoot(x,y,lambda[j])
    bicridge[j]=t(y-x%*%betapanalassoridge-b0)%*%
(y-x%*%betapanalassoridge-b0)/hsig
+log(n)*sum(betapanalassoridge!=0)
    bicols[j]=t(y-x%*%betapanalassools-b0)%*%
(y-x%*%betapanalassools-b0)/hsig
+log(n)*sum(betapanalassools!=0)
    biclasso[j]=t(y-x%*%betapanlasso-b0)%*%
(y-x%*%betapanlasso-b0)/hsig
+log(n)*sum(betapanlasso!=0)
  }
  #########get beta-hat
  betapanalassoridge=panalassoridge(x,y,lambda[which.min(bicridge)])
  betapanalassools=panalassools(x,y,lambda[which.min(bicols)])
  betalasso=lassoshoot(x,y,lambda[which.min(biclasso)])
  ##########get rpe for each iteration
  rpelasso[b]=mean((x%*%betalasso-x%*%truebeta)^2)/sig^2
```

```
  rpelassopanalassools[b]=mean((x%*%betapanalassools-x%*%truebeta)^2)/sig^2
  rpepanalassoridge[b]=mean((x%*%betapanalassoridge-x%*%truebeta)^2)/sig^2
  ####non zero components by lasso
  Classo[b]=sum(betalasso!=0)
  ###zero components incorrectly selected into the model by lasso
  Ilasso[b]=ifelse((betalasso[3]!=0),1,0)+ifelse((betalasso[4]!=0),1,0)
+ifelse((betalasso[6]!=0),1,0)+ifelse((betalasso[7]!=0),1,0)
+ifelse((betalasso[8]!=0),1,0)
  ####non zero component by panalassools
  Cpanalassools[b]=sum(betapanalassools!=0)
  ###zero components incorrectly selected into the model
  Ipanalassools[b]=ifelse((betapanalassools[3]!=0),1,0)
+ifelse((betapanalassools[4]!=0),1,0)
+ifelse((betapanalassools[6]!=0),1,0)
+ifelse((betapanalassools[7]!=0),1,0)
+ifelse((betapanalassools[8]!=0),1,0)
  ####non zero component by panalassoridge
  Cpanalassoridge[b]=sum(betapanalassoridge!=0)
  ###zero components incorrectly selected into the model
  Ipanalassoridge[b]=ifelse((betapanalassoridge[3]!=0),1,0)
+ifelse((betapanalassoridge[4]!=0),1,0)
+ifelse((betapanalassoridge[6]!=0),1,0)
+ifelse((betapanalassoridge[7]!=0),1,0)
+ifelse((betapanalassoridge[8]!=0),1,0)
}
#######mean rpe
mean(rpelasso)
mean(rpelassopanalassools)
mean(rpepanalassoridge)
```

```
###mean number of correct selection
mean(Classo)
mean(Cpanalassools)
mean(Cpanalassoridge)
###mean number of incorrect selection
mean(Ilasso)
mean(Ipanalassools)
mean(Ipanalassoridge)
```

- R code for simulation studies in Chapter 4.

```
library("mvtnorm")
library("MASS")
library("lme4")
options (warn=-1)

#define penalized restricted profile log likelihood########
prpll <- function(DDsig,beta,z,x,y,DDsig0,lambda,weight){
  n <- length(x)
  ni <- mapply(length,y)
  p <- ncol(x[[1]])
  q <- ncol(z[[1]])
  k <- q*(q+1)/2
  n.tot <- sum(ni)
  De <- matrix(0,nrow=q,ncol=q)
  De[lower.tri(De,diag=T)] <- DDsig
  if(ncol(De)>1){
    De <- De+t(De)-diag(diag(De))
  }
  log1 <- 0
```

```
  log2 <- 0
  log3 <- 0
  for(i in 1:n){
    si <- z[[i]]%*%De%*%t(z[[i]])+diag(1,ni[i])
    sii <- ginv(si)
    ei <- y[[i]]-x[[i]]%*%beta
    if(det(si)<=0){return(-9e10)
    }else {de1 <- log(det(si))}
    log1 <- log1+de1
    log2=log2+t(ei)%*%sii%*%ei
    log3=log3+t(x[[i]])%*%sii%*%x[[i]]
  }
  de1=log1
  if(log2<=0){return(-9e10)
  }else{de2 <- log(log2)}
  if(det(log3)<=0){return(-9e10)
  }else{de3 <- log(det(log3))}
  return(-.5*de1-.5*(n.tot-p)*de2-.5*de3-lambda*sum(abs(DDsig*weight/DDsig0)))
}
#define restricted profile log likelihood########
rpll <- function(DDsig,beta,z,x,y){
  n <- length(x)
  ni <- mapply(length,y)
  p <- ncol(x[[1]])
  q <- ncol(z[[1]])
  k <- q*(q+1)/2
  n.tot <- sum(ni)
  De <- matrix(0,nrow=q,ncol=q)
  De[lower.tri(De,diag=T)] <- DDsig
```

```
  if(ncol(De)>1){
    De <- De+t(De)-diag(diag(De))
  }
  log1 <- 0
  log2 <- 0
  log3 <- 0
  for(i in 1:n){
    si <- z[[i]]%*%De%*%t(z[[i]])+diag(1,ni[i])
    sii <- ginv(si)
    ei <- y[[i]]-x[[i]]%*%beta
    if(det(si)<=0){return(-9e10)
    }else {de1 <- log(det(si))}
    log1 <- log1+de1
    log2=log2+t(ei)%*%sii%*%ei
    log3=log3+t(x[[i]])%*%sii%*%x[[i]]
  }
  de1=log1
  if(log2<=0){return(-9e10)
  }else{de2 <- log(log2)}
  if(det(log3)<=0){return(-9e10)
  }else{de3 <- log(det(log3))}
  return(-.5*de1-.5*(n.tot-p)*de2-.5*de3)
}
#define penalized profile log likelihood########
ppll <- function(x,y,beta,sii,lambda,beta0){
  n <- length(x)
  ni <- mapply(length,y)
  n.tot <- sum(ni)
  ll<-0
```

```
  for(i in 1:n){
    ei <- y[[i]]-x[[i]]%*%beta
    ll=ll+t(ei)%*%sii[[i]]%*%ei
  }
  if(log(ll)<=0){return(-9e10)
  }else {lll=log(ll)}
  return(-.5*n.tot*lll-lambda*sum(abs(beta/beta0)))
}
#define profile log likelihood########
pll <- function(x,y,beta,sii){
  n <- length(x)
  ni <- mapply(length,y)
  n.tot <- sum(ni)
  ll<-0
  for(i in 1:n){
    ei <- y[[i]]-x[[i]]%*%beta
    ll=ll+t(ei)%*%sii[[i]]%*%ei
  }
  if(log(ll)<=0){return(-9e10)
  }else {lll=log(ll)}
  return(-.5*n.tot*lll)
}
##define trace function##
trace <- function(A)
{
  sum(diag(A))
}
##derivative of D
FdD=function(di,q){
```

```
  if(q>1){
    k <- q*(q+1)/2
    dD <- matrix(0,q,q)
    DD <- rep(0,k)
    DD[di] <- 1
    dD[lower.tri(dD,diag=T)] <- DD
    dD <- dD+t(dD)-diag(diag(dD))
  }else{
    dD <- matrix(1)
  }
  list(dD)
}
### derivative of V  V=ZDZ'+I
FdV <- function(r,z,dD,i){
  list(z[[i]]%*%dD[[r]]%*%t(z[[i]]))
}
### generalized least square estimator of beta
Fbeta <- function(x,y,z,De,sig){
  n <- length(y)
  p <- ncol(x[[1]])
  ni <- mapply(length,y)
  bet1 <- rep(0,p)
  H11 <- matrix(0,p,p)
  for(i in 1:n){
    si <- z[[i]]%*%De%*%t(z[[i]])+diag(1,ni[i])
    sii <- ginv(si)
    tss <- t(x[[i]])%*%sii
    H11 <- H11+tss%*%x[[i]]
    bet1 <- bet1+tss%*%y[[i]]
```

```
  }
  beta <- ginv(H11)%*%bet1
  beta
}
## sum (t(xi)*Aij*xi)
FA <- function(j,x,dV,sii,i){
  list(t(x[[i]])%*%sii[[i]]%*%dV[[i]][[j]]%*%sii[[i]]%*%x[[i]])
}
## first derivative of log |sum(XV^(-1)X)|
FD1 <- function(j,H00,p,n,XAijX){
  t2 <- matrix(0,p,p)
  for(i in 1:n){
    t2 <- t2+H00%*%XAijX[[i]][[j]]
  }
  -trace(t2)
}
## second derivative of log |sum(XV^(-1)X)|
SD1 <- function(k,j,x,n,H00,p,dV,sii,XAijX){
  t1 <- matrix(0,p,p)
  t2 <- matrix(0,p,p)
  t3 <- matrix(0,p,p)
  for(i in 1:n){
    derij <- dV[[i]][[j]]
    derik <- dV[[i]][[k]]
    t1 <- t1+XAijX[[i]][[k]]
    t2 <- t2+XAijX[[i]][[j]]
    t3 <- t3-t(x[[i]])%*%sii[[i]]%*%(derik%*%sii[[i]]%*%derij
+derij%*%sii[[i]]%*%derik)%*%sii[[i]]%*%x[[i]]
  }
```

```
  -trace(H00%*%t1%*%H00%*%t2)-trace(H00%*%t3)
}
## first derivative of sum(log|vi|)###
FD2 <- function(j,y,n,sii,dV){
  ni <- mapply(length,y)
  t8=matrix(0,ni,ni)
  for(i in 1:n){
    derij <- dV[[i]][[j]]
    t8=t8+sii[[i]]%*%derij
  }
  trace(t8)
}
## second derivative of sum(log|vi|)###
SD2 <- function(k,j,y,n,dV,sii){
  ni <- mapply(length,y)
  t9=matrix(0,ni,ni)
  for (i in 1:n){
    derij <- dV[[i]][[j]]
    derik <- dV[[i]][[k]]
    t9=t9-t(sii[[i]]%*%derik%*%sii[[i]])%*%derij
  }
  trace(t9)
}
## first derivative of log(sum(rV^(-1)r)) wrt theta###
FD3 <- function(j,n,dV,ei,sii){
  t5 <- matrix(0,1,1)
  t6 <- matrix(0,1,1)
  for(i in 1:n){
    derij <- dV[[i]][[j]]
```

```
    t5=t5-t(ei[[i]])%*%sii[[i]]%*%derij%*%sii[[i]]%*%ei[[i]]
    t6=t6+t(ei[[i]])%*%sii[[i]]%*%ei[[i]]
  }
  t5/t6
}
## second derivative of log(sum(rV^(-1)r)) wrt theta###
SD3 <- function(k,j,n,dV,ei,sii){
  t5 <- matrix(0,1,1)
  t6 <- matrix(0,1,1)
  t7 <- matrix(0,1,1)
  for(i in 1:n){
    derij <- dV[[i]][[j]]
    derik <- dV[[i]][[k]]
    t5=t5-t(ei[[i]])%*%sii[[i]]%*%derij%*%sii[[i]]%*%ei[[i]]
    t6=t6+t(ei[[i]])%*%sii[[i]]%*%ei[[i]]
    t7 <- t7+t(ei[[i]])%*%sii[[i]]%*%(derik%*%sii[[i]]%*%derij
+derij%*%sii[[i]]%*%derik)%*%sii[[i]]%*%ei[[i]]
  }
  (t7%*%t6-(t5/t6)^2)/(t6)^2
}
#### first derivative of log(sum(rV^(-1)r)) wrt beta
FD4 <- function(p,n,y,x,beta,sii){
  to <- matrix(0,1,1)
  tf <- matrix(0,p,1)
  for(i in 1:n){
    ei <- y[[i]]-x[[i]]%*%beta
    to=to+t(ei)%*%sii[[i]]%*%ei
    tf=tf-2*t(x[[i]])%*%sii[[i]]%*%ei
  }
```

```
  tf/as.numeric(to)
}
## second derivative of log(sum(rV^(-1)r)) wrt beta###
SD4 <- function(p,n,y,x,beta,sii){
  to <- matrix(0,1,1)
  tf <- matrix(0,p,1)
  ts <- matrix(0,p,p)
  for(i in 1:n){
    ei=y[[i]]-x[[i]]%*%beta
    to=to+t(ei)%*%sii[[i]]%*%ei
    tf=tf-2*t(x[[i]])%*%sii[[i]]%*%ei
    ts=ts+2*t(x[[i]])%*%sii[[i]]%*%x[[i]]
  }
  (ts*as.numeric(to)-(tf/as.numeric(to))
%*%t(tf/as.numeric(to)))/(as.numeric(to))^2
}
######### select optimal lambda for random effects based on BIC,AIC, and GCV
panran.lam.sel<- function(x,y,zz,D.init,eps,lam){
  BICR=numeric()
  AICR=numeric()
  GCVR=numeric()
  for (m in 1:length(lam)){
    lambda=lam[m]
    z=zz
    n <- length(x)
    ni <- mapply(length,y)
    p <- ncol(x[[1]])
    n.tot <- sum(ni)
    q <- ncol(z[[1]])
```

```
q0 <- q
kk <- q*(q+1)/2
De <- D.init
beta <- Fbeta(x,y,z,De)
dD <- sapply(1:kk,FdD,q)
weight <- diag(rep(1,q))
weight <- weight[lower.tri(weight,diag=T)]
DDsig0 <- D.init[lower.tri(D.init,diag=T)]
DD0 <- DDsig0
DDsignew <- DDsig0
step <- 1
maxstep <- 100
converge <- F
while(converge==F&&step<maxstep&&ncol(De)>1){
  DDsig <- DDsignew
  H0 <- matrix(0,nrow=p,ncol=p)
  sc <- rep(0,kk)
  H <- matrix(0,nrow=kk,ncol=kk)
  XAijX <- list(NA)
  length(XAijX) <- n
  si <- list(NA)
  length(si) <- n
  sii <- si
  dV <- list(NA)
  length(dV) <- n
  ei <- list(NA)
  length(ei) <- n
  for(i in 1:n){
    dV[[i]] <- sapply(1:kk,FdV,z,dD,i)
```

```
    si[[i]] <- z[[i]]%*%De%*%t(z[[i]])+diag(1,ni[i])
    sii[[i]] <- ginv(si[[i]])
    ei[[i]] <- y[[i]]-x[[i]]%*%beta
    H0 <- H0+t(x[[i]])%*%sii[[i]]%*%x[[i]]
    XAijX[[i]] <- sapply(1:kk,FA,x,dV,sii,i)
  }
  H00 <- ginv(H0)
  sc1=sapply(1:kk,FD1,H00,p,n,XAijX)
  sc2=sapply(1:kk,FD2,y,n,sii,dV)
  sc3=sapply(1:kk,FD3,n,dV,ei,sii)
  sc=-.5*(sc1+sc2+(n.tot-p)*sc3)

  he1=he2=he3=matrix(0,nrow=kk,ncol=kk)
  for (k in 1:kk){
    for (j in 1:kk){
      he1[k,j]=SD1(k,j,x,n,H00,p,dV,sii,XAijX)
      he2[k,j]=SD2(k,j,y,n,dV,sii)
      he3[k,j]=SD3(k,j,n,dV,ei,sii)
    }
  }
  H=-.5*(he1+he2+(n.tot-p)*he3)
  sc=sc-lambda*weight*sign(DDsig)/abs(DDsig0)
  H <- H-diag(lambda*weight/abs(DDsig*DDsig0))
  llold <- prpll(DDsig,beta,z,x,y,DDsig0,lambda,weight)
  llnew <- llold-1
  mm <- 1
  la <- 1
  gH <- ginv(H)%*%sc
  while(llnew<=llold&&mm<15){
```

```
    DDsignew <- DDsig-la*gH
    llnew <- prpll(DDsignew,beta,z,x,y,DDsig0,lambda,weight)
    la <- 1/2^mm
   mm <- mm+1
}
DDnew <- DDsignew
Dnew <- matrix(0,nrow=q,ncol=q)
Dnew[lower.tri(Dnew,diag=T)] <- DDnew
if(ncol(Dnew)>1)Dnew <- Dnew+t(Dnew)-diag(diag(Dnew))
ad <- abs(diag(Dnew))<=eps
Dnew[ad,] <- 0
Dnew[,ad] <- 0
DDsignew <- Dnew[lower.tri(Dnew,diag=T)]
for(j in 1:n){
   z[[j]] <- as.matrix(z[[j]][,!ad])
}
if(sum((DDsignew-DDsig)^2)<eps)converge <- TRUE
DDnew <- Dnew[lower.tri(Dnew,diag=T)]
DD <- DDnew[abs(DDnew)>0]
DD0 <- DD0[abs(DDnew)>0]
De <- Dnew[!ad,!ad]
De <- as.matrix(De)
DDsig0 <- DD0
beta <- Fbeta(x,y,z,De)
q <- dim(De)[2]
kk <- q*(q+1)/2
dD <- sapply(1:kk,FdD,q)
weight <- diag(rep(1,q))
weight <- weight[lower.tri(weight,diag=T)]
```

```
      DDsignew <- De[lower.tri(De,diag=T)]
      step=step+1
    }
    BICR[m]=-2*rpll(DDsig,beta,z,x,y)+log(n.tot)*length(De)
    AICR[m]=-2*rpll(DDsig,beta,z,x,y)+2*length(De)
    GCVR[m]=-rpll(DDsig,beta,z,x,y)/((1-length(De)/n.tot)^2*n.tot)
  }
  fit <- NULL
  fit$bic <- BICR
  fit$aic <- AICR
  fit$gcv <- GCVR
  return(fit)
}
#####select random effects using the optimal lambda
panran.sel<- function(lambda,x,y,zz,D.init,eps){
  z=zz
  n <- length(x)
  ni <- mapply(length,y)
  p <- ncol(x[[1]])
  n.tot <- sum(ni)
  q <- ncol(z[[1]])
  q0 <- q
  kk <- q*(q+1)/2
  De <- D.init
  beta <- Fbeta(x,y,z,De)
  dD <- sapply(1:kk,FdD,q)
  weight <- diag(rep(1,q))
  weight <- weight[lower.tri(weight,diag=T)]
  DDsig0 <- D.init[lower.tri(D.init,diag=T)]
```

```
DD0 <- DDsig0
DDsignew <- DDsig0
step <- 1
maxstep <- 100
converge <- F
record <- seq(q)
while(converge==F&&step<maxstep&&ncol(De)>1){
  DDsig <- DDsignew
  H0 <- matrix(0,nrow=p,ncol=p)
  sc <- rep(0,kk)
  H <- matrix(0,nrow=kk,ncol=kk)
  XAijX <- list(NA)
  length(XAijX) <- n
  si <- list(NA)
  length(si) <- n
  sii <- si
  dV <- list(NA)
  length(dV) <- n
  ei <- list(NA)
  length(ei) <- n
  for(i in 1:n){
    dV[[i]] <- sapply(1:kk,FdV,z,dD,i)
    si[[i]] <- z[[i]]%*%De%*%t(z[[i]])+diag(1,ni[i])
    sii[[i]] <- ginv(si[[i]])
    ei[[i]] <- y[[i]]-x[[i]]%*%beta
    H0 <- H0+t(x[[i]])%*%sii[[i]]%*%x[[i]]
    XAijX[[i]] <- sapply(1:kk,FA,x,dV,sii,i)
  }
  H00 <- ginv(H0)
```

```
sc1=sapply(1:kk,FD1,H00,p,n,XAijX)
sc2=sapply(1:kk,FD2,y,n,sii,dV)
sc3=sapply(1:kk,FD3,n,dV,ei,sii)
sc=-.5*(sc1+sc2+(n.tot-p)*sc3)

he1=he2=he3=matrix(0,nrow=kk,ncol=kk)
for (k in 1:kk){
  for (j in 1:kk){
    he1[k,j]=SD1(k,j,x,n,H00,p,dV,sii,XAijX)
    he2[k,j]=SD2(k,j,y,n,dV,sii)
    he3[k,j]=SD3(k,j,n,dV,ei,sii)
  }
}
H=-.5*(he1+he2+(n.tot-p)*he3)
sc=sc-lambda*weight*sign(DDsig)/abs(DDsig0)
H <- H-diag(lambda*weight/abs(DDsig*DDsig0))
llold <- prpll(DDsig,beta,z,x,y,DDsig0,lambda,weight)
llnew <- llold-1
mm <- 1
la <- 1
gH <- ginv(H)%*%sc
while(llnew<=llold&&mm<15){
  DDsignew <- DDsig-la*gH
  llnew <- prpll(DDsignew,beta,z,x,y,DDsig0,lambda,weight)
  la <- 1/2^mm
  mm <- mm+1
}
DDnew <- DDsignew
Dnew <- matrix(0,nrow=q,ncol=q)
```

```
    Dnew[lower.tri(Dnew,diag=T)] <- DDnew
    if(ncol(Dnew)>1)Dnew <- Dnew+t(Dnew)-diag(diag(Dnew))
    ad <- abs(diag(Dnew))<=eps
    Dnew[ad,] <- 0
    Dnew[,ad] <- 0
    DDsignew <- Dnew[lower.tri(Dnew,diag=T)]
    for(j in 1:n){
      z[[j]] <- as.matrix(z[[j]][,!ad])
    }
    record <- record[!ad]
    if(sum((DDsignew-DDsig)^2)<eps)converge <- TRUE
    DDnew <- Dnew[lower.tri(Dnew,diag=T)]
    DD <- DDnew[abs(DDnew)>0]
    DD0 <- DD0[abs(DDnew)>0]
    De <- Dnew[!ad,!ad]
    De <- as.matrix(De)
    DDsig0 <- DD0
    beta <- Fbeta(x,y,z,De)
    q <- dim(De)[2]
    kk <- q*(q+1)/2
    dD <- sapply(1:kk,FdD,q)
    weight <- diag(rep(1,q))
    weight <- weight[lower.tri(weight,diag=T)]
    DDsignew <- De[lower.tri(De,diag=T)]
    step=step+1
  }
  Df <- matrix(0,q0,q0)
  Df[record,record]=De
  fit <- NULL
```

```
  fit$beta <- beta
  fit$D <- Df
  return(fit)
}
######### select optimal lambda for fix effects based on BIC, AIC and GCV
panfix.lam.sel<- function(xx,y,zz,beta.init,D.init,eps,lam){
  BICF=numeric()
  AICF=numeric()
  GCVF=numeric()
  for (m in 1:length(lam)){
    lambda=lam[m]
    z=zz
    x=xx
    D=D.init
    beta=beta.init
    beta0=beta.init
    p <- ncol(x[[1]])
    p0=p
    n <- length(y)
    ni <- mapply(length,y)
    n.tot <- sum(ni)
    si <- list(NA)
    length(si) <- n
    sii <- si
    for(i in 1:n){
      #V#
      si[[i]] <- z[[i]]%*%D%*%t(z[[i]])+diag(1,ni[i])
      #V^(-1)
      sii[[i]] <- ginv(si[[i]])
```

```
}
converge <- F
step <- 1
maxstep <- 100
while(converge==F&&step<maxstep){
  beta.old <- beta
  sc=-.5*n.tot*FD4(p,n,y,x,beta.old,sii)
  H=-.5*n.tot*SD4(p,n,y,x,beta.old,sii)
  sc=sc-lambda*sign(beta.old)/abs(beta0)
  H <- H-diag(lambda/as.vector(abs(beta.old*beta0)))
  llold <- ppll(x,y,beta.old,sii,lambda,beta0)
  llnew <- llold-1
  mm <- 1
  la <- 1
  gH <- ginv(H)%*%sc
  while(llnew<=llold&&mm<15){
    beta <- beta.old-la*gH
    llnew <- ppll(x,y,beta,sii,lambda,beta0)
    la <- 1/2^mm
    mm <- mm+1
  }
  ad <- abs(beta)<=1e-3
  for(j in 1:n){
    x[[j]] <- as.matrix(x[[j]][,!ad])
  }
  if(abs(llnew-llold)<eps)converge <- TRUE
  p=ncol(x[[1]])
  beta=beta[!ad]
  beta0=beta0[!ad]
```

```
      step=step+1
    }
    BICF[m]=-2*pll(x,y,beta,sii)+log(n.tot)*length(beta)
    AICF[m]=-2*pll(x,y,beta,sii)+2*length(beta)
    GCVF[m]=-pll(x,y,beta,sii)/((1-length(beta)/n.tot)^2*n.tot)
  }
  fit <- NULL
  fit$bic <- BICF
  fit$aic <- AICF
  fit$gcv <- GCVF
  return(fit)
}
#####select fix effects using the optimal lambda###########
panfix.sel=function(xx,y,zz,beta.init,D.init,lambda,eps){
  x=xx
  z=zz
  D=D.init
  beta=beta.init
  beta0=beta.init
  p <- ncol(x[[1]])
  p0=p
  n <- length(y)
  ni <- mapply(length,y)
  n.tot <- sum(ni)
  si <- list(NA)
  length(si) <- n
  sii <- si
  for(i in 1:n){
    #V#
```

```
    si[[i]] <- z[[i]]%*%D%*%t(z[[i]])+diag(1,ni[i])
    #V^(-1)
    sii[[i]] <- ginv(si[[i]])
  }
  converge <- F
  step <- 1
  maxstep <- 100
  record <- seq(p)
  while(converge==F&&step<maxstep){
    beta.old <- beta
    sc=-.5*n.tot*FD4(p,n,y,x,beta.old,sii)
    H=-.5*n.tot*SD4(p,n,y,x,beta.old,sii)
    sc=sc-lambda*sign(beta.old)/abs(beta0)
    H <- H-diag(lambda/as.vector(abs(beta.old*beta0)))
    llold <- ppll(x,y,beta.old,sii,lambda,beta0)
    llnew <- llold-1
    mm <- 1
    la <- 1
    gH <- ginv(H)%*%sc
    while(llnew<=llold&&mm<15){
      beta <- beta.old-la*gH
      llnew <- ppll(x,y,beta,sii,lambda,beta0)
      la <- 1/2^mm
      mm <- mm+1
    }
    ad <- abs(beta)<=1e-3
    record <- record[!ad]
    for(j in 1:n){
      x[[j]] <- as.matrix(x[[j]][,!ad])
```

```
    }
    if(abs(llnew-llold)<eps)converge <- TRUE
    p=ncol(x[[1]])
    beta=beta[!ad]
    beta0=beta0[!ad]
    step=step+1
  }
  bet <- rep(0,p0)
  bet[record]=beta
  fit <- NULL
  fit$beta <- bet
  fit$D <- D
  return(fit)
}
corrx=matrix(0,5,5)
for (i in 1:5){
  for (j in 1:5){
    if (i==j) corrx[i,j]=1
    else corrx[i,j]=0
  }
}
###sequence of lambda
lam=exp(seq(-2, 2, length=20))
PANCRBIC=0  ##correct number of random selection for pan bic
PANCFBIC=0  ##correct number of fix selection for pan bic
PANCCBIC=0  ##correct number of random&fix selection for pan bic
PANCRAIC=0  ##correct number of random selection for pan aic
PANCFAIC=0  ##correct number of fix selection for pan aic
PANCCAIC=0  ##correct number of random&fix selection for pan aic
```

```
PANCRGCV=0  ##correct number of random selection for pan gcv
PANCFGCV=0  ##correct number of fix selection for pan gcv
PANCCGCV=0  ##correct number of random&fix selection for pan gcv
CZRBIC=CZFBIC=IZRBIC=IZFBIC=CZRAIC=CZFAIC=IZRAIC=IZFAIC
=CZRGCV=CZFGCV=IZRGCV=IZFGCV=numeric()
#########data input########
for (j in 1:100){
  sig <- 1
  ni  <-  5
  n  <-  50
  y  <-  NULL
  x  <-  NULL
  z  <-  NULL
  subject  <-  kronecker(1:n, rep(1,5))
  true.beta  <-  c(1,2,2,0,0)
  Dt  <-  matrix(c(1,.5,0,0,0,.5,1,rep(0,18)), nrow=5, ncol=5)
  for (i in 1:n)
  {
    x[[i]] <- mvrnorm(ni,numeric(5),corrx)
    z[[i]] <- x[[i]]
    S  <- sig*(z[[i]]%*%Dt%*%t(z[[i]]) + diag(ni))
    y.temp  <-  t(rmvnorm(1, x[[i]]%*%true.beta, S))
    y[[i]] <- y.temp
  }
  n <- length(y)
  y1 <- y[[1]]
  x1 <- x[[1]]
  z1 <- z[[1]]
  for(i in 2:n){
```

```
    y1 <- rbind(y1,y[[i]])
    x1 <- rbind(x1,x[[i]])
    z1 <- rbind(z1,z[[i]])
  }
  ob <- lmer(y1~x1-1+(0+z1|subject),
control = lmerControl(check.nobs.vs.nRE = "warning"))
  hh <- VarCorr(ob)
  D.init <- hh[[1]]
  beta.hat=as.matrix(fixef(ob))
  sig.init <- sig
  zz=z
  xx=x

  ###Using BIC
  aa=panran.lam.sel(x,y,zz,D.init,eps=1e-5,lam)
  bestbicr=aa$bic
  lambda1bic=lam[which.min(bestbicr)]
  estr.bic=panran.sel(lambda1bic,x,y,zz,D.init,eps=1e-5)
  bb1=panfix.lam.sel(xx,y,zz,estr.bic$beta,estr.bic$D,eps=1e-5,lam)
  bestbicf=bb1$bic
  lambda2bic=lam[which.min(bestbicf)]
  est.bic=panfix.sel(xx,y,zz,estr.bic$beta,estr.bic$D,lambda2bic,eps=1e-5)

  ####Using AIC
  bestaicr=aa$aic
  lambda1aic=lam[which.min(bestaicr)]
  estr.aic=panran.sel(lambda1aic,x,y,zz,D.init,eps=1e-5)
  bb2=panfix.lam.sel(xx,y,zz,estr.aic$beta,estr.aic$D,eps=1e-5,lam)
  bestaicf=bb2$aic
```

```
lambda2aic=lam[which.min(bestaicf)]
est.aic=panfix.sel(xx,y,zz,estr.aic$beta,estr.aic$D,lambda2aic,eps=1e-5)

####Using GCV
bestgcvr=aa$gcv
lambda1gcv=lam[which.min(bestgcvr)]
estr.gcv=panran.sel(lambda1gcv,x,y,zz,D.init,eps=1e-5)
bb3=panfix.lam.sel(xx,y,zz,estr.gcv$beta,estr.gcv$D,eps=1e-5,lam)
bestgcvf=bb3$gcv
lambda2gcv=lam[which.min(bestgcvf)]
est.gcv=panfix.sel(xx,y,zz,estr.gcv$beta,estr.gcv$D,lambda2aic,eps=1e-5)

#####CORRECT SELECTION
if (sum(est.bic$D!=0)==sum(Dt!=0))
  PANCRBIC=PANCRBIC+1
if (sum(est.bic$beta!=0)==sum(true.beta!=0))
  PANCFBIC=PANCFBIC+1
if(sum(est.bic$D!=0)==sum(Dt!=0) && sum(est.bic$beta!=0)==sum(true.beta!=0)
  PANCCBIC=PANCCBIC+1
if (sum(est.aic$D!=0)==sum(Dt!=0))
  PANCRAIC=PANCRAIC+1
if (sum(est.aic$beta!=0)==sum(true.beta!=0))
  PANCFAIC=PANCFAIC+1
if(sum(est.aic$D!=0)==sum(Dt!=0) && sum(est.aic$beta!=0)==sum(true.beta!=0)
  PANCCAIC=PANCCAIC+1
if (sum(est.gcv$D!=0)==sum(Dt!=0))
  PANCRGCV=PANCRGCV+1
if (sum(est.gcv$beta!=0)==sum(true.beta!=0))
  PANCFGCV=PANCFGCV+1
```

```
   if(sum(est.gcv$D!=0)==sum(Dt!=0) && sum(est.gcv$beta!=0)==sum(true.beta!=0)
     PANCCGCV=PANCCGCV+1

   CZFBIC[j]=ifelse((est.bic$beta[4]==0),1,0)+ifelse((est.bic$beta[5]==0),1,0)
   CZRBIC[j]=ifelse((est.bic$D[3,3]==0),1,0)+ifelse((est.bic$D[4,4]==0),1,0)
+ifelse((est.bic$D[5,5]==0),1,0)
   IZFBIC[j]=ifelse((est.bic$beta[1]==0),1,0)+ifelse((est.bic$beta[2]==0),1,0)
+ifelse((est.bic$beta[3]==0),1,0)
   IZRBIC[j]=ifelse((est.bic$D[1,1]==0),1,0)+ifelse((est.bic$D[2,2]==0),1,0)

   CZFAIC[j]=ifelse((est.aic$beta[4]==0),1,0)+ifelse((est.aic$beta[5]==0),1,0)
   CZRAIC[j]=ifelse((est.aic$D[3,3]==0),1,0)+ifelse((est.aic$D[4,4]==0),1,0)
+ifelse((est.aic$D[5,5]==0),1,0)
   IZFAIC[j]=ifelse((est.aic$beta[1]==0),1,0)+ifelse((est.aic$beta[2]==0),1,0)
+ifelse((est.aic$beta[3]==0),1,0)
   IZRAIC[j]=ifelse((est.aic$D[1,1]==0),1,0)+ifelse((est.aic$D[2,2]==0),1,0)

   CZFGCV[j]=ifelse((est.gcv$beta[4]==0),1,0)+ifelse((est.gcv$beta[5]==0),1,0)
   CZRGCV[j]=ifelse((est.gcv$D[3,3]==0),1,0)+ifelse((est.gcv$D[4,4]==0),1,0)
+ifelse((est.gcv$D[5,5]==0),1,0)
   IZFGCV[j]=ifelse((est.gcv$beta[1]==0),1,0)+ifelse((est.gcv$beta[2]==0),1,0)
+ifelse((est.gcv$beta[3]==0),1,0)
   IZRGCV[j]=ifelse((est.gcv$D[1,1]==0),1,0)+ifelse((est.gcv$D[2,2]==0),1,0)
}

## number of correct selection
PANCRBIC
PANCFBIC
PANCCBIC
```

PANCRAIC

PANCFAIC

PANCCAIC

PANCRGCV

PANCFGCV

PANCCGCV

mean (CZRBIC)

mean (CZFBIC)

mean (IZRBIC)

mean (IZFBIC)

mean (CZRAIC)

mean (CZFAIC)

mean (IZRAIC)

mean (IZFAIC)

mean (CZRGCV)

mean (CZFGCV)

mean (IZRGCV)

mean (IZFGCV)