ABSTRACT

LAN, LAN. Variable Selection in Linear Mixed Model for Longitudinal Data. (Under the direction of Dr. Daowen Zhang and Dr. Hao Helen Zhang.)

Fan and Li (JASA, 2001) proposed a family of variable selection procedures for certain parametric models via a nonconcave penalized likelihood approach, where significant variable selection and parameter estimation were done simultaneously, and the procedures were shown to have the oracle property. In this presentation, we extend the nonconcave penalized likelihood approach to linear mixed models for longitudinal data. Two new approaches are proposed to select significant covariates and estimate fixed effect parameters and variance components. In particular, we show the new approaches also possess the oracle property when the tuning parameter is chosen appropriately. We assess the performance of the proposed approaches via simulation and apply the procedures to data from the Multicenter AIDS Cohort Study.

Variable Selection in Linear Mixed Model for Longitudinal Data

by

Lan Lan

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

STATISTICS

Raleigh

2006

APPROVED BY:

Daowen Zhang Chair of Advisory Comittee	Hao Helen Zhang Co-Chair of Advisory Comittee
Marie Davidian	Dennis Boos

This work is dedicated to

Sheng and my parents

Biography

Lan Lan was born on September 10, 1972 and raised in Weiyuan, Sichuan, China. She obtained her B.S. in electronic engineering from the Southeast University in 1993. She served as an assistant engineer in an electronic company in Chengdu for two years before she received her M.S. in biomedical engineering from the University of Electronic Science and Technology of China in 1998. She worked as a software engineer in a biomedical device company in Chengdu from July, 1998 to December, 1999.

In January 2000, lan entered the graduate program in electronic engineering at University of Hawaii, Manoa. One year later, she came to the University of Nebraska at Lincoln, where she received a M.S. degree in statistics in August, 2002. She continued her studies towards a Ph.D degree in statistics at North Carolina State University.

While studying at North Carolina State University, she married Sheng Feng and they have shared a wonderful life together since.

Acknowledgements

I would like to express my deepest appreciation to my advisors, Dr. Daowen Zhang and Dr. Helen Zhang for their constant support, guidance and patience throughout my work. Thank you also to my committee members Dr. Dennis Boos and Dr. Marie Davidian for their suggestions and teaching during my graduate career.

Thank you to my fellow graduate students for their support and friendship, especially to Joe Boyer, Kirsten Doehler, Jimmi Doi, Sheng Feng, Changku Kang, Hyeyoung Lee, Amy Nail, Lavanya Ramanan, Michael Rubitski, Abdus Shakoor Wahed, Jing Wang, Weiwei Wu and Yujun Wu for the late nights spent studying in the Hillsborough office during the past four years.

My gratitude goes to my family. To my parents, Jiagui and Shuhua for giving me endless love and inspiration. To my elder brother Zheng, sister-in-law Min and my niece Lydia for their constant encouragement and prayers. To my parents-in-law, Dehua and Sishen for their understanding and caring. Especially to my husband Sheng, thank you for always being with me in this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Variable selection, or subset selection, is an important and interesting problem in linear regression. Variable selection has two main goals: ease in interpretation and accuracy of prediction.

In practice, the data often consist of a large number of potential explanatory variables or predictors $X_1, X_2, ..., X_d$, and not all of these variables may contribute to the explanation of the response variable Y. To give the data an appropriate interpretation, we should select a model that is both parsimonious and meaningful. In other words, we have to select from these variables to obtain a model that contains as few variables as possible while still being the "best" model. The variable selection problem studies how to pick the subset of predictor variables that leads to the "best" model.

Since the true relationship between the explanatory and response variables is unknown in most practical cases, we rely on the observed data to select important variables and calibrate the relationship that is used for future prediction.

1.1 Variable Selection in Linear Regression Models for Independent Data

Assume the observations $(\boldsymbol{X}_i, y_i)$ are independently and identically sampled from some distribution, $i = 1, ..., n$. In the context of the familiar linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{Y} = (y_1, y_2, ..., y_n)^T$ is a vector of n observations, $\boldsymbol{X}_i = (X_{i1}, ..., X_{id})$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, ..., \boldsymbol{X}_n^T)^T$ is the $n \times d$ design matrix, $\boldsymbol{\beta}$ is a $d \times 1$ coefficient vector possibly with some zero elements, $\boldsymbol{\epsilon}$ follows n-dimensional multivariate normal distribution $N(0, \sigma^2\boldsymbol{I})$, then $\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ is the prediction of $\boldsymbol{Y}$. The predictors with nonzero coefficients are called important variables, and unimportant variables otherwise.

In the statistical literature, many selection procedures have been developed, but most of them are for linear regression models only. Miller (1990) reviews various selection procedures. Classical methods include subset selection and selection with shrinkage penalty, which will be briefly discussed in Section (1.1.1) and Section (1.1.2). Most recently, Luo, Boos, and Stefanski (2006) proposed a variable selection approach by adding additional noise to the response variable that allows them to tune the stopping rule with the hope of finding an unbiased estimator of the total error variance σ^2. Wu, Boos, and Stefanski (2005) developed a new approach to variable selection by controlling the false selection rate (FSR) which is the proportion of unimportant variables included in the model. This approach is implemented by adding a known number of pseudo variables to the real data set and selecting variables by monitoring the proportion of pseudo variables falsely selected.

1.1.1 Subset Selection

The procedure of subset selection, such as all subsets, forward selection, and backward elimination, are widely used in combination with certain criteria.

All subsets approach is an exhaustive search, and it searches through all possible subsets and selects the subset with the smallest residual sum of squares. However, a drawback of this method is that it is very time consuming when the number of predictors d is large. This computational difficulty prevents the all subsets algorithm from being widely used when there are a large number of predictors in practical problems. To overcome this problem, stepwise methods have been proposed to search through a smaller number of subsets. Two basic ideas of stepwise selection are forward selection and backward elimination. Starting with no variables in the model, forward selection adds one variable at a time until a stopping criterion is satisfied. On the opposite end, backward selection starts with the full model, eliminates one variable at a time, and the variables are deleted one by one until a stopping criterion is satisfied. Particularly, forward selection is probably the most widely used subset selection method.

Most common selection criteria combine statistical measures with penalties for increasing number of predictors in the model. Reviews can be found in Hocking

(1976), Thompson (1978), Rao and Wu (2001), and Miller (2002). Basically, these criteria are classified into four categories: (1) Prediction criteria; (2) Information-based criteria; (3) Data-reuse procedures; (4) Bayesian variable selection.

To evaluates the prediction accuracy of a model, we use the prediction error (PE)

$$\begin{aligned} \mathrm{PE} &= \frac{1}{n}\mathrm{E}\left[(\boldsymbol{Y}^{\text{new}} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y}^{\text{new}} - \hat{\boldsymbol{Y}})\right] \\ &= \mathrm{ME} + \sigma^2, \end{aligned} \tag{1.2}$$

where the expectation is taken with respect to $\boldsymbol{Y}^{\text{new}}$ and the model error (ME) is defined as

$$\mathrm{ME} = \frac{1}{n}\mathrm{E}\left[(\hat{\boldsymbol{Y}} - \boldsymbol{\mu})^T(\hat{\boldsymbol{Y}} - \boldsymbol{\mu})\right], \tag{1.3}$$

where $\boldsymbol{Y}^{\text{new}}$ is a random vector independently and identically distributed as $\boldsymbol{Y}$. Notice that the term σ^2 in (1.2) is not affected by the variable selection procedure, therefore it can be ignored and only ME is used for model evaluation. However, as the underlying true $\boldsymbol{\mu}$ is unknown in practice, the estimate of ME needs to be constructed. Mallows (1973) proposed a statistic C_P as a criterion, which is defined as

$$C_P = \frac{RSS_q}{\hat{\sigma}_F^2} - n + 2q,$$

where RSS_q is the residual sum of squares from the least squares fit of q predictors, and $\hat{\sigma}_F^2 = \frac{RSS_q}{n-q}$ is the estimate of σ^2 for the full model, where RSS_F is the full model residual sum of squares and $n > q$. The selected model is the one with the minimum value of C_P. Mallows (1995) pointed out that $\mathrm{E}(n\mathrm{ME}) = \mathrm{E}(\hat{\sigma}_F^2 C_P)$, hence the expected ME can be estimated by $\frac{\hat{\sigma}_F^2 C_P}{n}$ and minimizing C_P is approximately equivalent to minimizing ME.

Information-based criteria are related to likelihood or information measures. The most popular criteria include AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria). AIC was proposed by Akaike (1970, 1974), and it selects the model that minimizes

$$\mathrm{AIC} = -2\ell + 2q, \tag{1.4}$$

where ℓ is the log-likelihood of the model, q is the number of predictors, or dimension of the covariate vector $\boldsymbol{X}$ in the model. BIC was proposed by Schwarz (1978) and has a similar form to AIC except that the log-likelihood is penalized by $q \log(n)$ instead of $2q$, it selects the model that minimizes

$$\mathrm{BIC} = -2\ell + q \log(n), \tag{1.5}$$

where n is number of observations.

Data-reuse procedures estimate the model error for each candidate subset of variables by replicating the data in a certain way, and selecting the subset with the smallest estimate of model error. Picard and Cook (1984) proposed cross-validation (CV), and Burman (1989) and Zhang (1993) studied a multifold cross-validation procedure. Efron (1977) first introduced the bootstrap technique for data resampling, which can be applied to regression model selection. Breiman (1992) proposed the little bootstrap procedure for variable selection.

The idea of Baysian model selection is to pick the model with the highest posterior probability among all candidate models. The problem of this procedure is how to choose the priors, since the specification of the prior strongly influences the posterior probabilities. Most studies focus on how to find good priors. For instance, Lindley

(1968) suggested using the uniform prior and a cost function to select variables, and Mitchell and Beauchamp (1988) introduced "spike and slab" priors. To avoid the prior difficulty, Berger and Pericchi (1996a, b) proposed the default Bayes factor criterion. More discussion on Bayesian variable selection is reviewed by George and McCulloch (1993), Clyde et al. (1998), Berger and Pericchi (2001), and Miller (2002).

1.1.2 Selection with Shrinkage Penalty

Although subset selection procedures are practically useful, their theoretic properties are somewhat hard to understand. In addition, the subset selection procedure ignores stochastic errors in the variable selection process. In practice, subset selection procedure is very time consuming, especially when the number of predictors is large.

To overcome these drawbacks, some researchers proposed a family of new variable selection methods based on shrinkage methods. The idea is to add a penalty function to either the residual sum of squares or to the negative log-likelihood and then minimize it. Such penalties have the feature that the small components of $\boldsymbol{\beta}$ are shrunk all the way to exactly zero. In effect, these components are deleted. These methods are different from traditional subset selection approaches in that they delete insignificant variables if their coefficients are estimated as zero under the selected tuning parameters. Thus, selecting significant variables and estimating regression coefficients are conducted simultaneously.

The penalized least squares approach, a member of this family, estimates coefficients and selects variables by minimizing the penalized least squares. Recently, a variety of penalty functions have been proposed in penalized least squares, such as bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996). The penalized least squares solve (1.6) and (1.7) respectively,

$$(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda\Sigma_{j=1}^{d}|\beta_j|^{\gamma}, \quad \gamma \geq 1 \tag{1.6}$$

$$(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda\Sigma_{j=1}^{d}|\beta_j|, \tag{1.7}$$

where λ is the tuning parameter. In fact, LASSO is a special case when $\gamma = 1$, and ridge regression (Hoerl and Kennard (1970a,b)) is the case when $\gamma = 2$. Knight and Fu (2000) studied the asymptotic properties of LASSO type estimators. Recently, Efron et al. (2004) proposed a new model selection approach, Least Angle Regression (LARS), and show that it can implement the LASSO by a simple modification.

Fan and Li (2001) recently proposed a variable selection approach via a nonconcave penalized likelihood. This approach adds a special penalty, named Smoothly Clipped Absolute Deviation (SCAD) penalty, to the log-likelihood function. The SCAD penalty procedure selects variables by shrinking the coefficients of unimportant variables to 0. Furthermore, under certain regular conditions, they show that the resulting estimators possess an oracle property. Namely, their method works as well as the underlying correct submodel when the submodel is known in advance. The LASSO and ridge regression do not possess this oracle property. Using the technique of nonconcave penalized likelihood, Fan and Li (2002) extended their research to the Cox proportional hazards model and to the frailty model. Very recently, Fan and Li (2004) expanded the methodology to semiparametric analysis for longitudinal data. However, they did not explicitly model the correlation in the longitudinal data, instead they used an independent working correlation matrix.

1.2 Variable Selection for Longitudinal Data

In practice, researchers often conduct experiments in which subjects are measured repeatedly over time, and the observations on the same subject are correlated. Such types of data are called longitudinal data. This feature is in violation of the basic assumption of independent in the linear regression model. A popular model for the longitudinal data is the linear mixed model that includes the fixed effects and the subject specific random effects.

It is well known that linear mixed models are widely applicable and practically important for longitudinal data. Fixed effect variable selection in linear mixed models has been one of the outstanding problems. For convenience, we call this problem as variable selection in linear mixed models throughout our study although there is a separate research on random effect variable selection. In our study, we assume the random effects and variance structure can be specified, and the main interest is selecting the subset of variables associated with fixed effects. In the literature, little work has been done in this area. Previous studies relied on subset selection procedures with certain criteria such as AIC and BIC. Similar to the general linear model situation, subset selection methods for longitudinal models have some drawbacks such as being computationally time-consuming, lack of theoretic properties, and ignoring stochastic errors.

Although SCAD, LASSO and ridge are very efficient in terms of computation compared with subset selection, none of them has been used in linear mixed models for longitudinal data. Our research goal is to propose a new variable selection method that overcomes the drawbacks of the subset selection methods in linear mixed models. Motivated by the good properties of SCAD penalty, we propose the penalized likelihood estimation procedure with the SCAD penalty. As an extension of Fan and Li's (2001) approach, our method estimates the fixed effects parameters and variance components jointly, and select variables at the same time. For longitudinal normal data, our method improves Fan and Li's (2001) by taking the random effects into account, which is shown by our simulation study. Two procedures for this methods are proposed, the simple procedure and the iterative procedure.

The organization of this thesis is as follows. In Chapter 2, we introduce our method and propose two estimation procedures and algorithms. In Chapter 3, we derive the theoretical results for the estimates obtained from the simple procedure. In Chapter 4, we conduct simulation studies: we first describe the simulation design, and then present and discuss the simulation results. The application of the proposed procedures to one real data set is reported in Chapter 5. Chapter 6 contains the conclusion remarks.

Chapter 2

Method and Materials

2.1 Background

2.1.1 Linear Mixed Models

In longitudinal studies, the primary endpoints are measured repeatedly over time within the same subject. A general framework to model the relationship between the primary response and the covariates of longitudinal profiles is to postulate a model that incorporates the underlying subject specific random effects. A popular model for longitudinal data is a linear mixed model. In particular, the linear mixed model assumes

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b_i} + \epsilon_{ij}, \tag{2.1}$$

where $i = 1, ..., m$, m represents the total number of subjects, $j = 1, ..., n_i$, n_i represents the total number of responses measured for subject i, Y_{ij} is the value of the response variable for the jth observation from the ith subject, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_d)^T$ is the vector of fixed effects that are identical for all subjects, $\boldsymbol{X}_{ij}$ is the fixed effect regressor vector for observation j in subject i, $\boldsymbol{Z}_{ij}$ is the random effect regressor vector for observation j in subject i, $\boldsymbol{b}_i \sim N(0, \boldsymbol{D}(\boldsymbol{\theta}))$ is the subject specific random effects measuring among-subject variation of Y_{ij}, $\boldsymbol{\theta}$ is the variance components parameters, ϵ_{ij} is the random error independent of $\boldsymbol{b}_i$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})^T$, $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i(\boldsymbol{\theta}))$ and for $i \neq k$, $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\epsilon}_k$.

With above assumptions, the variance of $\boldsymbol{Y}_i$ is, therefore,

$$\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{Z}_i^T + \boldsymbol{R}_i(\boldsymbol{\theta})$$

and $\boldsymbol{V} = \mathrm{Var}(\boldsymbol{Y})$, which is the block diagonal matrix of $\boldsymbol{V}_i$.

A problem associated with the use of mixed models is estimation of the parameters that determine the distribution of $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$, or which are the variance of the random effects, i.e. the estimation of the variance components $\boldsymbol{\theta}$. The maximum likelihood method is commonly used for estimating the parameter. For model (2.1), according to the assumption that $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent and normally distributed, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{1}{2} \log |\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{\theta}) (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.2}$$

where $\boldsymbol{W}(\boldsymbol{\theta}) = \boldsymbol{V}(\boldsymbol{\theta})^{-1}$.

Normally, the maximum likelihood estimators of variance structures are attractive since ML estimators are functions of sufficient statistics and are consistent, asymptotically normal, and efficient. However, the maximum likelihood estimator of variance components in linear mixed models is biased downwards because it takes no account of the degrees of freedom lost by estimating the regression coefficients. To correct this problem, Patterson and Thompson (1971) proposed the restricted maximum likelihood (REML) as a method of estimating variance components in the linear mixed model (2.1). Compared to the log-likelihood (2.2), the restricted log-likelihood for (2.1) is

$$\ell_R(\boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{2.3}$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}. \tag{2.4}$$

The REML estimates for $\boldsymbol{\theta}$ are the minimizer of the negative restricted log-likelihood $\ell_R(\boldsymbol{\theta}; \mathbf{Y})$. Specifically, the REML estimator $\hat{\boldsymbol{\theta}}$ solves $S(\boldsymbol{\theta}) = \mathbf{0}$, where $S_{\boldsymbol{\theta}}$ is the score function expressed as

$$S_{\theta_i} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}) + \frac{1}{2}\boldsymbol{Y}^T\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}\boldsymbol{P}\boldsymbol{Y},$$

with $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$. The REML estimate $\hat{\boldsymbol{\theta}}$ can be solved by an iterative scheme. Expanding $S_\theta(\boldsymbol{\theta})$ at the current estimate $\boldsymbol{\theta}_0$ results in

$$S_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) \approx S_{\boldsymbol{\theta}_0} + \frac{\partial S(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^T}(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Equating the right hand side to zero leads to the update formula for $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \left(\frac{\partial S(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^T}\right)^{-1} S(\boldsymbol{\theta}_0). \tag{2.5}$$

The estimates for the fixed effect parameters are the weighted least-square solution in (2.4) given REML estimation $\hat{\boldsymbol{\theta}}$. The estimated variance of $\hat{\boldsymbol{\beta}}$ is

$$(\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{X})^{-1}, \tag{2.6}$$

and the estimated variance of $\hat{\boldsymbol{\theta}}$ is

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}, \tag{2.7}$$

where $\boldsymbol{I}(\hat{\boldsymbol{\theta}})_{ij} = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_j})|_{\hat{\boldsymbol{\theta}}}$.

2.1.2 Penalized Likelihood

Many variable selection techniques are implemented by maximizing the penalized log-likelihood with various penalty functions. The idea is to subtract a penalty function $m\sum_{j=1}^{d} p_\lambda(|\boldsymbol{\beta}_j|)$ from the log-likelihood $\ell(\boldsymbol{\beta}; \boldsymbol{\theta}, \boldsymbol{Y})$ and maximize the penalized log-likelihood, where m is the number of subjects for longitudinal data, d is the dimension of the covariate vector $\boldsymbol{X}$, p_λ is the nonnegative penalty function, and λ is the tuning parameter. The penalized log-likelihood is defined as

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}) = \ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}) - m\sum_{j=1}^{d} p_\lambda(|\boldsymbol{\beta}_j|). \tag{2.8}$$

Specifically, if the observed longitudinal data follow the assumptions in linear mixed model (2.1), the penalized log-likelihood given $\boldsymbol{\theta}$ thus takes the form

$$\ell_p(\boldsymbol{\beta}; \boldsymbol{\theta}, \boldsymbol{Y}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - m\sum_{j=1}^{d} p_\lambda(|\boldsymbol{\beta}_j|). \tag{2.9}$$

Many well-known variable selection criteria are special cases of the penalized log-likelihood of (2.9). For instance, the AIC and BIC (Bayesian Information Criteria). The penalty functions correspond to AIC and BIC are $2d$ and $d\log(n)$ respectively, although these criteria were motivated from different principles.

Recently, many researchers have been working on penalized least squares or penalized log-likelihood with the shrinkage penalty on the coefficients. Frank and Friedman (1993) proposed bridge regression, the solution of penalized least square with the penalty $\lambda\Sigma_{j=1}^{d}|\beta_j|^{\gamma}$ $(\gamma \geq 1)$; Donoho and Johnstone (1994) studied the L_1 penalty $\lambda\Sigma_{j=1}^{d}|\beta_j|$ in the wavelet context; Tibshirani(1996, 1997) proposed the LASSO which is the penalized squares estimate with the L_1 penalty $\lambda\Sigma_{j=1}^{d}|\beta_j|$ in the general regression settings; ridge regression is the solution of penalized likelihood with the L_2 penalty $\lambda\Sigma_{j=1}^{d}|\beta_j|^2$. For variable selection purpose, the penalty has to be chosen in such a way that maximizing the penalized log-likelihood will shrink the coefficients of unimportant variables to zero.

2.1.3 Smoothly Clipped Absolute Deviation Penalty

Fan and Li (2001) suggested three principles that a good penalty function should possess: (1) unbiasedness such that there is no overpenalization for large parameters to avoid unnecessary modeling bias; (2) sparsity, the resulting estimator should follow a thresholding rule so that the insignificant parameters are automatically set to zero to reduce model complexity; (3) continuity, which is to avoid instability in model prediction. Penalty functions determine the sampling properties of the penalized likelihood estimators.

As the penalty functions such as LASSO and ridge do not satisfy above three principles simultaneously, Fan and Li (2001) proposed a continuous differentiable (except origin) penalty function called smoothly clipped absolute deviation penalty (SCAD). The SCAD was originally proposed by Fan & Li (2001) in linear regression models for independent data. The penalty is essentially a quadratic spline with two knots at λ and $a\lambda$ defined as

$$p_\lambda(\omega) = \lambda \begin{cases} \omega & \text{if } \ 0 < \omega \le \lambda, \\ -\frac{1}{2(a-1)\lambda}(\omega^2 - 2a\lambda\omega + \lambda^2) & \text{if } \ \lambda < \omega \le a\lambda, \\ \frac{1}{2}(a+1)\lambda & \text{if } \ \omega > a\lambda, \end{cases} \tag{2.10}$$

for $\omega \ge 0$. Here λ and a are tuning parameters and they need to be chosen in practice, and we will discuss how to select λ and a in Section (2.3). The first derivative of $p_\lambda(\omega)$ with respect to ω is

$$p_\lambda^{'}(\omega) = \begin{cases} \lambda & \text{if } \ \omega \le \lambda \\ \frac{a\lambda-\omega}{a-1} & \text{if } \ \lambda < \omega \le a\lambda \\ 0 & \text{if } \ \omega > a\lambda \end{cases} \tag{2.11}$$

Figure 2.1 depicts the SCAD penalty function. From Figure 2.1, the SCAD penalty is symmetric, nonconcave on $(0, \infty)$, and has singularity at the origin. Fan and Li (2001) pointed out the solution of the SCAD penalty satisfies all the three properties for a good penalty. This motivates us to consider SCAD as the penalty function for variable selection in linear mixed models for longitudinal data.

2.2 Penalized Likelihood for Linear Mixed Models

2.2.1 SCAD-Penalized Likelihood for Linear Mixed Models

Given the longitudinal data following the assumptions in Section (2.1.1), we propose to maximize the log-likelihood of the data subject to the SCAD penalty that takes the form

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - m\sum_{j=1}^{d} p_\lambda(|\boldsymbol{\beta}_j|), \tag{2.12}$$

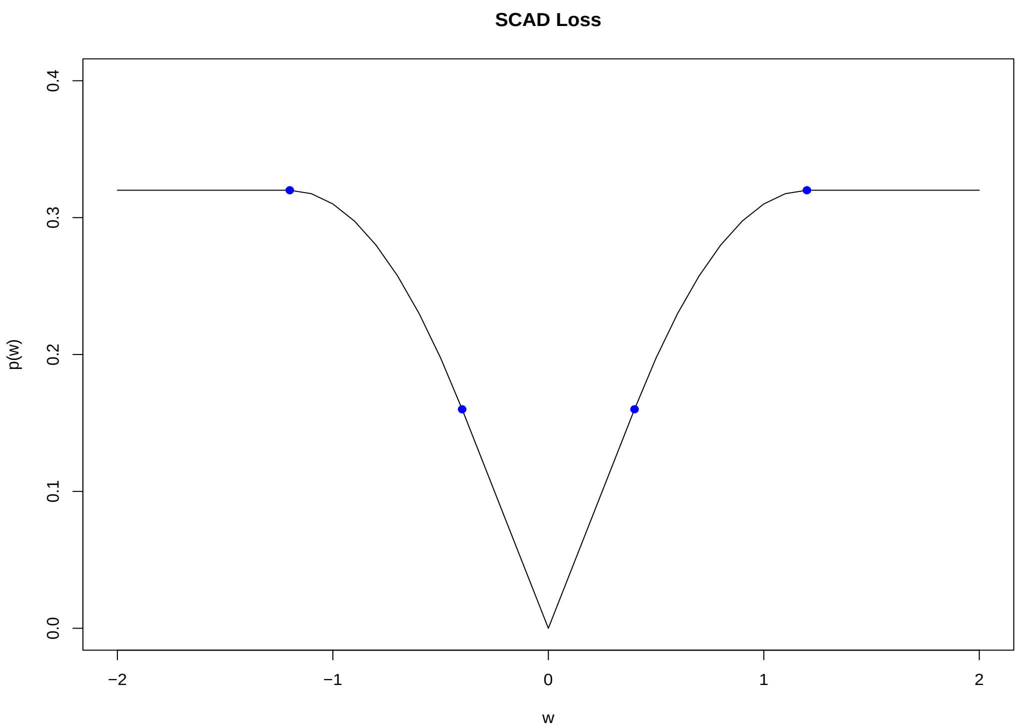

Figure 2.1: Plot of SCAD Function

where p_λ is the SCAD penalty, $\boldsymbol{V}$ is variance-covariance of the response vector $\boldsymbol{Y}$ and $\boldsymbol{W} = \boldsymbol{V}^{-1}$. Due to the extra effort on modeling the covariance structure, inference based on (2.12) is more complicated than that in Fan and Li (2001). Note when $\boldsymbol{\theta}$ is given, the optimization problem is equivalent to minimizing the objective function

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + m \sum_{j=1}^{d} p_\lambda(|\boldsymbol{\beta}_j|), \tag{2.13}$$

since $\boldsymbol{V}$ is a function of $\boldsymbol{\theta}$ only. The SCAD penalty is not convex in $\boldsymbol{\beta}$, thus standard optimization packages fail to solve (2.12).

2.2.2 Estimation of $\boldsymbol{\beta}$

It is difficult to obtain the exact solution of parameter estimates by minimizing the objective function (2.13) because the SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. To solve this optimization problem, following Fan and Li (2001), we apply the approximation below.

Suppose $\boldsymbol{\beta}_0$ is a current estimate of $\boldsymbol{\beta}$, if β_{j0} is very close to 0, then set $\hat{\beta}_j = 0$; otherwise when $\beta_{j0} \neq 0$, we use the local approximation

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j,$$

which implies

$$[p_\lambda(|\beta_j|)]'' \approx p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|. \tag{2.14}$$

Consequently, the SCAD penalty can be locally approximated by a quadratic function

$$p_\lambda(|\beta_j|) \;\; \approx \;\; p_\lambda(|\beta_{j0}|) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2). \tag{2.15}$$

Hence,

$$\sum_{j=1}^{d} p_\lambda(|\beta_j|) = \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta} + \text{term not involving } \boldsymbol{\beta},$$

where $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)$ =diag$\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, ..., p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}$. It follows that the objective function (2.13) can be locally approximated by the sum of two quadratic terms:

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T m \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}. \tag{2.16}$$

As a result, minimizing the objective function (2.13) is done by minimizing the quadratic form (2.16) iteratively.

Given $\boldsymbol{\theta}$, to solve $\boldsymbol{\beta}$ from the equation (2.16), we take the first derivative with respect to $\boldsymbol{\beta}$ and set it to $\mathbf{0}$, that is

$$\frac{dG(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta} = \mathbf{0}, \tag{2.17}$$

which gives the update

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)]^{-1} \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{Y}. \tag{2.18}$$

The update formula (2.18) for our method is more general than Fan and Li's (2001) in that (2.18) takes into account the variance components.

In summary, given $\boldsymbol{\theta}$, we can update $\boldsymbol{\beta}$ by iteratively solving the estimating equation (2.17) : starting from some initial value $\boldsymbol{\beta}^{(0)}$, from the kth iteration, obtain the $(k+1)$th iteration as

$$\hat{\boldsymbol{\beta}}^{(k+1)} = [\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)})]^{-1} \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{Y}. \tag{2.19}$$

We use the relative change as the convergence criterion. The criterion is

$$\max \frac{|\boldsymbol{\beta}_l^{(k+1)} - \boldsymbol{\beta}_l^{(k)}|}{|\boldsymbol{\beta}_l^{(k)}| + \eta} < \text{tol} \qquad l = 1, 2, ..., d \quad ,$$

where η is some smal positive number. This stopping rule implies the maximum relative change in all elements of $\boldsymbol{\beta}$ from iteration k to $k+1$ is smaller than some pre-specified tolerance.

2.2.3 Estimation of $\boldsymbol{\theta}$

Compared to Fan and Li's (2001) objective function, there is an extra term related to $\boldsymbol{W}(\boldsymbol{\theta})$ in (2.13), the inverse of covariance matrix $\boldsymbol{V}$. Thus another key point in our method is the estimation of variance component $\boldsymbol{\theta}$. We propose two procedures to estimate $\boldsymbol{\theta}$: the simple procedure and the iterative procedure.

In the simple procedure, we first ignore the correlation of the data and use Fan and Li's (2001) approach to estimate $\boldsymbol{\beta}$ and select the important covariates. Given this "naive" estimate $\hat{\boldsymbol{\beta}}_0$, we then estimate $\hat{\boldsymbol{\theta}}_0$ using the REML method. In the rest, we hold $\hat{\boldsymbol{\theta}}$ at $\hat{\boldsymbol{\theta}}_0$ and only update $\hat{\boldsymbol{\beta}}$ until convergence. At convergence, the REML is used again to update $\boldsymbol{\theta}$.

The iterative procedure, a more sophisticated one compared to the simple procedure, estimates $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ alternatively in each iteration and continues iteration until $\hat{\boldsymbol{\beta}}$ is convergent.

The detail of both procedures will be elaborated in Section (2.4).

2.3 Selection of Tuning parameters

The forgoing estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is based on fixed tuning parameters. Recall that there are two unknown tuning parameters a and λ in the SCAD function. How to choose these tuning parameters is crucial for the procedure. In theory, we may search the best pair (λ, a) over two dimensional grid of points with some criteria. However, such an implementation can be computationally expensive.

From the Bayesian statistical point view, Fan and Li (2001) suggested $a = 3.7$ is a good choice for various problems. Furthermore, Fan and Li (2001) demonstrated that the performance can not be significantly improved with a selected by data driven methods. Therefor $a = 3.7$ will be used throughout our study. For the other tuning parameter λ, Fan and Li (2001) suggested using the GCV score as a criterion for independent data. Since the objective functions are different between our method and Fan and Li's (2001), we extend this GCV to correlated data and develop a GCV by taking into account the correlation in the longitudinal data. In addition, we propose a new REML based score, and explore the use of BIC in this context.

2.3.1 GCV Criterion

In the statistical literature, there is plenty of work on selecting the tuning parameter λ when the observations are independent. The most commonly used methods include CV (cross validation) and GCV (generalized cross validation); see Wahba (1990) and Gu (2002).

The idea of cross validation is to use part of the data to fit the model, and use the rest to evaluate the fitted model. One particular case is the leave-one-out cross validation (LOOCV). If there are m data points in an independent data set, cross validation works by leaving one data point (x_i, y_i) out at a time, estimating the squared residual at x_i based on the model fitted with the remaining $m-1$ data points, and choosing λ to minimize the cross validation function

$$CV(\lambda) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_\lambda^{-i}(x_i))^2,$$

where $\hat{y}_\lambda^{-i}$ indicates the fit obtained by leaving out the ith data point. Normally, $\hat{y}_\lambda$ can be formulated as a linear combination of the responses as

$$\hat{\boldsymbol{y}}_\lambda = \boldsymbol{A}(\lambda)\boldsymbol{y},$$

for some matrix $\boldsymbol{A}(\lambda)$, which is often called a projection matrix or an influential matrix. Let a_{ii} be the diagonal elements of $\boldsymbol{A}(\lambda)$, then the CV function can be approximated by

$$CV(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{y_i - \hat{y}_\lambda(x_i)}{1-a_{ii}}\right]^2.$$

In practice, it is very time consuming to compute a_{ii}, so Wahba (1990) suggests the GCV as an alternative criterion. Wahba(1990) showed that the GCV score for regression model with independent error is

$$GCV(\lambda) = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_\lambda(x_i))^2}{(m - tr(\boldsymbol{A}(\lambda)))^2} = \frac{\frac{1}{m}||(\mathbf{I} - \boldsymbol{A}(\lambda))\boldsymbol{Y}||^2}{[\frac{1}{m}tr(\mathbf{I} - \boldsymbol{A}(\lambda))]^2}. \tag{2.20}$$

GCV is closely related to CV but it is much easier to compute. Wang (1998) extended the GCV for independent data to the correlated data.

For regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ with $\epsilon \sim N(0, \boldsymbol{W}^{-1})$, we multiply $\boldsymbol{W}^{\frac{1}{2}}$ on both sides of the model equation

$$\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X\beta} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\epsilon}. \tag{2.21}$$

Define $\boldsymbol{Y}^* = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y}$, $\boldsymbol{X}^* = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}$ and $\boldsymbol{\epsilon}^* = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\epsilon}$, (12) can be written as

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \qquad \text{with} \quad \boldsymbol{\epsilon}^* \sim N(0, I). \tag{2.22}$$

By this transformation, $\boldsymbol{Y}^*$ is independent and henceforth we can use the GCV function (2.20) for independent data.

Next, we derive the GCV score for the SCAD penalized likelihood method. Consider the penalized least-squared regression

$$\frac{1}{2}(\boldsymbol{Y}^* - \boldsymbol{X}^{\boldsymbol{*}}\boldsymbol{\beta})^T(\boldsymbol{Y}^* - \boldsymbol{X}^{\boldsymbol{*}}\boldsymbol{\beta}) + m\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.23}$$

The solution to minimize (2.23) is the one to minimize (2.13). As derived in Section (2.2.2), the minimizer of (2.13) satisfies $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0))^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$. Let $\boldsymbol{U}_\lambda(\boldsymbol{\beta}_0) = [\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)]^{-1}\boldsymbol{X}^T\boldsymbol{W}$. Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda &= [\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)]^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Y}^* \\
&= \boldsymbol{U}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Y}^*,
\end{aligned}$$

and $\hat{\boldsymbol{Y}}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_\lambda$.

Since the solution $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ at the iteration when the algorithm is convergent or when the current estimate of $\boldsymbol{\beta}$ is close enough to the previous iteration, we have

$$\begin{aligned}
\hat{\boldsymbol{Y}}^* &= \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_\lambda \\
&= \boldsymbol{X}^*\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}})\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Y}^* \\
&\equiv \boldsymbol{A}_\lambda\boldsymbol{Y}^*,
\end{aligned}$$

where $\boldsymbol{A}_\lambda$ is the projection matrix given by

$$\boldsymbol{A}_\lambda = X^*\boldsymbol{U}\lambda(\hat{\boldsymbol{\beta}})\boldsymbol{W}^{-\frac{1}{2}} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}})\boldsymbol{W}^{-\frac{1}{2}}.$$

Following Wang (1998), the GCV score for (2.13) with correlated errors is given by

$$GCV(\lambda) = \frac{\frac{1}{m}||(\mathbf{I} - \boldsymbol{A}(\lambda))\boldsymbol{Y}^*||^2}{[\frac{1}{m}tr(\mathbf{I} - \boldsymbol{A}(\lambda))]^2}. \tag{2.24}$$

Further note that

$$\begin{aligned}
(\boldsymbol{I} - \boldsymbol{A}_\lambda)\boldsymbol{Y}^* &= (\boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}})\boldsymbol{W}^{-\frac{1}{2}})\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Y} \\
&= (\boldsymbol{W}^{\frac{1}{2}} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))\boldsymbol{Y} \\
&= \boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))\boldsymbol{Y},
\end{aligned}$$

and

$$\begin{aligned}
tr(\boldsymbol{I} - \boldsymbol{A}_\lambda) &= tr(\boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}})\boldsymbol{W}^{-\frac{1}{2}}) \\
&= tr[\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))\boldsymbol{W}^{-\frac{1}{2}}] \\
&= tr(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}})).
\end{aligned}$$

Therefore the GCV score can be simplified as

$$\begin{aligned}
GCV(\lambda) &= \frac{\frac{1}{m}||\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))\boldsymbol{Y}||^2}{[\frac{1}{m}tr(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))]^2} \\
&= \frac{\frac{1}{m}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})'\boldsymbol{W}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})}{[\frac{1}{m}tr(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}))]^2}.
\end{aligned} \tag{2.25}$$

As previously shown, the matrix $\boldsymbol{U}_\lambda(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}))^{-1}\boldsymbol{X}^T\boldsymbol{W}$. We choose λ by minimizing the GCV score function given in (2.25).

2.3.2 REML Criterion

Alternatively, we propose a new criterion motivated by REML. As stated in section (2.1.1), REML was introduced by Patterson and Thompson (1971) as a method of estimating variance components in linear mixed models. Compared to ML, REML is considered superior because it produces unbiased estimating equations for variance components, while the maximum likelihood estimator of variance components is biased downwards because it takes no account of the degrees of freedom lost by estimating the regression coefficients.

For the linear mixed model (2.1), the restricted log-likelihood for the observed longitudinal data is

$$\ell_R(\boldsymbol{\theta}|\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})) = -\frac{1}{2}(\log|\boldsymbol{V}| + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}|) - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}), \quad (2.26)$$

where $\boldsymbol{\beta}$ is the vector including the important parameters only. Consider the approximated objective function (2.16), denoting $m\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)$ as $\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}_0$ is the current estimate of $\boldsymbol{\beta}$, we then have

$$G(\boldsymbol{\beta}) \approx \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \quad (2.27)$$

where we denote $\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0) = m\Sigma_\lambda(\boldsymbol{\beta}_0) =$diag$\{mp'_\lambda(|\beta_{10}|)/|\beta_{10}|, ..., mp'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}$.

Since for $j = 1, ..., d$,

$$p'_\lambda(|\beta_j|) = \begin{cases} \lambda & \text{if } |\beta_j| \le \lambda, \\ \frac{a\lambda-|\beta_j|}{a-1} & \text{if } \lambda < |\beta_j| \le a\lambda, \\ 0 & \text{if } |\beta_j| > a\lambda, \end{cases}$$

the diagonal matrix $\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0)$ may have some zero elements. We reorder $\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0)$ by moving all the zero diagonal elements to the left top and nonzero diagonal elements to the right bottom. Denote the reordered $\boldsymbol{\Sigma}_{\lambda m}(\boldsymbol{\beta}_0)$ as $\boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}$ has the form of $\boldsymbol{\Sigma} = \begin{bmatrix} 0 & \\ & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. Next we rearrange $\boldsymbol{\beta}$ corresponding to the order of $\boldsymbol{\Sigma}$, the correspondingly reordered $\boldsymbol{\beta}$ has the form of $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$, where $\boldsymbol{\beta_1}$ is a vector of parameters without penalty and $\boldsymbol{\beta}_2$ is a vector including those with penalty. Based on such reordering, the objective function (2.27) can be rewritten as

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}_2^T \boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_2. \tag{2.28}$$

Note that, if we regard $\boldsymbol{\beta}_2 \sim N(0, \boldsymbol{\Sigma}_{22}^{-1})$, the objective function (2.28) is essentially the negative joint log-likelihood of the model

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{2.29}$$

where $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the covariates corresponding to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively, $\boldsymbol{\beta}_2$ and $\boldsymbol{\epsilon}$ are independent.

Define $\boldsymbol{e} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$, the model (2.29) then becomes

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}, \tag{2.30}$$

where

$$\mathrm{Var}(\boldsymbol{e}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T + \sigma^2\boldsymbol{I} + \boldsymbol{X}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2^T = \boldsymbol{V} + \boldsymbol{X}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2^T = \boldsymbol{V}^*.$$

Following the equation (2.26), we define the REML value of the underlying model (2.29) as

$$\ell_{REML} = -\frac{1}{2}(\log|\boldsymbol{V}^*| + \log|\boldsymbol{X}_1^T\boldsymbol{V}^{*-1}\boldsymbol{X}_1|) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1)^T\boldsymbol{V}^{*-1}(\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1). \tag{2.31}$$

The REML function in (2.31) will be used as another criterion to determine λ, and the selected λ is the one associated with the largest REML value.

2.3.3 BIC_{scad} Criterion

In Section (1.1.1), we stated different types of information criteria. Among them, BIC is a widely used criterion for variable selection. Here, we explore the use of BIC_{scad} as a criterion to select variables for the SCAD penalized likelihood method. In particular, (2.32) gives the formula to calculate BIC based on log-likelihood.

$$\text{BIC}_{\text{scad}} = -2\ell + q \log m, \tag{2.32}$$

where ℓ is the log-likelihood at the SCAD penalized estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, q is the number of predicted variables which are the important variables after selection, m is the number of subjects. λ is chosen based on the smallest BIC_{scad}.

2.4 Computing Algorithms

To solve the SCAD penalized log-likelihood for longitudinal data, we introduce two computing algorithms: the simple procedure and the iterative procedure. The difference between these two procedures is the way to estimate the variance components $\boldsymbol{\theta}$. The generic idea underlying both procedures is to estimate $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and select the tuning parameter λ simultaneously, the variable selection is done automatically once the estimates are obtained.

2.4.1 Simple Procedure

We first introduce the algorithm for the simple procedure.

1, Initiation

Step (1.1) Given data, standardize the covariates $\boldsymbol{X}$ with their sample means and standard errors.

Step (1.2) Ignoring the data dependence, use Fan and Li's (2001) method to estimate $\boldsymbol{\beta}$ and select initial important variables. Then apply REML to estimate $\boldsymbol{\theta}_0$ by these initial important variables.

Step (1.3) Fix the $\hat{\boldsymbol{\theta}}_0$ obtained from Step (1.2). The estimated $\hat{\boldsymbol{\beta}}_0$ obtained from (1.2) will be used as the initial value for Step (3.2) in the Inner Loop.

2, Outer Loop

Step (2.1) Set a range for λ, grid λ into t levels $\lambda_1, \lambda_2, ..., \lambda_t$.

Step (2.2) For each level of λ_i, go to the inner loop to estimate $\boldsymbol{\beta}$, compute $\text{GCV}(\lambda_i)$, $\text{REML}(\lambda_i)$ or $\text{BIC}_{\text{scad}}(\lambda_i)$ with this $\hat{\boldsymbol{\beta}}$ and the fixed $\hat{\boldsymbol{\theta}}$ from (1.3). The selected λ_i is the one with the lowest GCV score, the smallest BIC_{scad} or the largest REML value.

Step (2.3) The final estimator for $\boldsymbol{\beta}$ is the one with the selected λ_i from (2.2), and we update the $\hat{\boldsymbol{\theta}}$ once more with the final $\hat{\boldsymbol{\beta}}$. The nonzero components of $\hat{\boldsymbol{\beta}}$ correspond to significant variable coefficients.

3, Inner Loop

Step (3.1) Given λ_i, form the weight matrix $\boldsymbol{W}$ by the fixed $\hat{\boldsymbol{\theta}}$.

Step (3.2) For the $(k+1)$th iteration, set the estimate of $\boldsymbol{\beta}$ from the kth iteration

$\hat{\boldsymbol{\beta}}^k$, re-estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = [\boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{X} + m \boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}^{(k)})]^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(k)} \boldsymbol{Y}.$$

Step (3.3) Repeat step (3.2) until $\hat{\boldsymbol{\beta}}$ is convergent.

Step (3.4) Adjust $\hat{\boldsymbol{\beta}}$ to the original scale.

This procedure is named the simple procedure because we update $\boldsymbol{\theta}$ only once for the whole procedure. Considering a more sophisticated way to estimate $\boldsymbol{\theta}$, we now propose an alternative procedure, the iterative procedure.

2.4.2 Iterative Procedure

Compared to the simple procedure, the iterative procedure is more complicated in that we estimate the variance components iteratively with the regression coefficients. The algorithm for this procedure is similar to that for the simple procedure. Likewise, the algorithm includes three parts: initiation, outer loop and inner loop.

The Initiation and Outer Loop steps for iterative procedure is a little bit different from the simple procedure in the following way:

1, Initiation

Step (1.1) Given data, standardize the covariates with their sample means and standard errors.

Step (1.2) Ignore the dependence, use Fan and Li's (2001) method to estimate $\boldsymbol{\beta}$ and select initial important variables, apply REML to estimate $\boldsymbol{\theta}$ by these initial important variables.

Step (1.3) The estimated $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ obtained from (1.2) are the initial values for (3.2) in the Inner Loop.

2, Outer Loop

Step (2.1) Set a range for λ, grid λ into t levels $\lambda_1, \lambda_2, ..., \lambda_t$.

Step (2.2) For each level of λ_i, go to the inner loop to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, compute GCV(λ_i), REML(λ_i) or $\text{BIC}_{\text{scad}}(\lambda_i)$ with the $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ from the inner loop.

Step (2.3) The selected λ_i is the one with the lowest GCV score or largest REML value.

Step (2.4) The final estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is the one with the selected λ_i from (2.2). The nonzero components of $\hat{\boldsymbol{\beta}}$ correspond to significant variable coefficients.

The Inner Loop for iterative procedure involves iteratively estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

3, Inner Loop

Step (3.1) Given λ_i, for the $(k+1)$th iteration, set $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\theta}}^{(k)}$, form the weight matrix $\hat{\boldsymbol{W}}^{(k)}$ by $\hat{\boldsymbol{\theta}}^{(k)}$, re-estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = [\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{X} + m\boldsymbol{\Sigma}_{\lambda_i}(\hat{\boldsymbol{\beta}}^{(k)})]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{Y}$$

Step (3.2) Set $k = k+1$, apply REML procedure to estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$ using $\hat{\boldsymbol{\beta}}^{(k+1)}$.

Step (3.3) Repeat step (3.1) and step (3.2) until $\hat{\boldsymbol{\beta}}$ is convergent.

Step (3.4) Adjust $\hat{\boldsymbol{\beta}}$ to the original scale.

2.5 Standard Error Formula for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$

Because we are estimating parameters and selecting variables at the same time, the standard errors for estimated parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be obtained. We will investigate the asymptotic normality of $\hat{\boldsymbol{\beta}}$ in Chapter 3. By the estimated asymptotic variance formula (3.16) for the nonzero coefficient $\boldsymbol{\beta}_1$, we have

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}_1) = \frac{1}{m}[\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}) + \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1)]^{-1}\boldsymbol{I}_1(\hat{\boldsymbol{\theta}})[\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}) + \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1)]^{-1},$$

where $\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}) = \frac{1}{m}\boldsymbol{X}_{\text{imp}}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{X}_{\text{imp}}$ and $\boldsymbol{X}_{\text{imp}}$ includes important covariate only, and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1) = \text{diag}\{p''_{\lambda_m}(|\hat{\boldsymbol{\beta}}_{11}|), ..., p''_{\lambda_m}(|\hat{\boldsymbol{\beta}}_{1d}|)\}$. By the approximation given in (2.14), $[p_\lambda(|\beta_j|)]'' \approx p'_\lambda(|\beta_{j0}|)/\beta_{j0}$ for the true value β_{j0}, thus when $\boldsymbol{\beta}$ is convergent, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1) =$ diag $\{p'_\lambda(|\boldsymbol{\beta}_{11}|)/|\boldsymbol{\beta}_{11}|, ..., p'_\lambda(|\boldsymbol{\beta}_{1d}|)/|\boldsymbol{\beta}_{1d}|\}$. Following the conventional technique in the likelihood settings, $\boldsymbol{I}_1(\hat{\boldsymbol{\theta}})$ can be estimated by $\frac{1}{m}\boldsymbol{X}_{\text{imp}}\boldsymbol{W}(\hat{\boldsymbol{\theta}})\boldsymbol{X}_{\text{imp}}$, it follows that the covariance for the nonzero coefficient $\hat{\boldsymbol{\beta}}_1$ obtained in our methods can be estimated by the corresponding sandwich formula

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}_1}) = \{\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\theta}})\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}(\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\theta}})\boldsymbol{X})\{\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\theta}})\boldsymbol{X} + m\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}, \tag{2.33}$$

where $\boldsymbol{X}$ includes important covariate only.

We estimate $\boldsymbol{\theta}$ through the REML approach and algorithm in the whole study, the variance formula for $\hat{\boldsymbol{\theta}}$ in (2.7) is used to calculate $\text{Var}(\boldsymbol{\theta})$. Hence, the estimate of $\text{Var}(\hat{\boldsymbol{\theta}})$ is

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})^{-1}, \tag{2.34}$$

where $\widehat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})_{ij} = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}(\hat{\boldsymbol{\theta}})\frac{\partial \boldsymbol{V}(\hat{\boldsymbol{\theta}})}{\partial\theta_i}\boldsymbol{P}(\hat{\boldsymbol{\theta}})\frac{\partial \boldsymbol{V}(\hat{\boldsymbol{\theta}})}{\partial\theta_j})$ is the (i,j)th element of $\widehat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})$ and

$$\boldsymbol{P}(\hat{\boldsymbol{\theta}}) = \boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1} - \boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1}.$$

In this chapter, a few variable selection procedures for linear mixed model were proposed via penalized likelihood approach. Some Criteria were presented to select the tuning parameters. Algorithms were proposed to estimate the regression coefficients and variance components by maximizing the SCAD penalized log-likelihood function.

The properties and the performance of the above procedures will be studied by theoretical proofs in Chapter 3 and empirical simulation studies in Chapter 4.

Chapter 3

Theoretical Results

3.1 Introduction

In this chapter we show that the proposed estimators $\boldsymbol{\beta}$ has nice theoretical properties such as the oracle property. Donoho and Johnstone (1994) defines estimators with oracle property if estimators work as well as if the correct submodel were known in advance.

The SCAD penalty function has a second order continuous derivative at nonzero components of $\boldsymbol{\beta}_0$ and is nonnegative with $p_\lambda(0) = 0$. Denote

$$a_m = \max\{|p'_{\lambda_m}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \tag{3.1}$$

$$h_m = \max\{|p''_{\lambda_m}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \tag{3.2}$$

$$\boldsymbol{\Sigma} = \mathrm{diag}\{p''_{\lambda_m}(|\beta_{10}|), ..., p''_{\lambda_m}(|\beta_{s0}|)\}, \tag{3.3}$$

$$\boldsymbol{h} = (p'_{\lambda_m}(|\beta_{10}|)\mathrm{sign}(\beta_{10}), ..., p'_{\lambda_m}(|\beta_{s0}|))\mathrm{sign}(\beta_{s0}))^T, \tag{3.4}$$

where s is the number of nonzero components in $\boldsymbol{\beta}_0$.

Let the true value of $\boldsymbol{\beta}$ be

$$\boldsymbol{\beta}_0 = (\beta_{10}, ..., \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T,$$

and the true value of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_0$, where the dimension of $\boldsymbol{\beta}_{10}$ is s. Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$ and all components of $\boldsymbol{\beta}_{10}^T$ are not equal to zero. Under some regularity conditions, we will show our SCAD estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ possesses the property that with probability tending to 1, under certain conditions, $\hat{\boldsymbol{\beta}}_2 = 0$ and $\hat{\boldsymbol{\beta}}_1$ is root-m consistent and asymptotically normal.

The theorems below are for the estimates from simple procedure, we conjecture similar theorems exist for the estimates from iterative procedure.

3.2 Regularity Conditions

To present the proofs of the theorems, we first state some regularity conditions. Denote by $\boldsymbol{\Omega}$ the parameter space for $\boldsymbol{\beta}$. We assume $\boldsymbol{\theta}$ is fixed and $\boldsymbol{\theta} \in \Theta$.

Regularity Conditions:

(1) The observations $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ are independently distributed each with the conditional normal density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\theta})$. Let $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}_i|\boldsymbol{X}_i)$ be the log-likelihood function of observations $\boldsymbol{Y}_i|\boldsymbol{X}_i$, and

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}_i|\boldsymbol{X}_i) = -\frac{1}{2}\log|\boldsymbol{V}_i(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

Furthermore, assume the number of observation n_i is bounded, and

$$\boldsymbol{A}(\boldsymbol{\theta}) = \lim_{m \to \infty} \frac{1}{m} \boldsymbol{X}^T \boldsymbol{V}^{-1}(\boldsymbol{\theta}) \boldsymbol{X} = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\theta}) \boldsymbol{X}_i$$

exists and is positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

(2) There exists an open subset of $\boldsymbol{\Omega}$ and $\boldsymbol{\Theta}$ containing the true parameter point $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ such that for almost all $\boldsymbol{Y}_i|\boldsymbol{X}_i$, the density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\theta})$ admits all third derivatives $\frac{\partial^3 \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_j \partial \theta_k \partial \theta_l} = \boldsymbol{X}_{ij}^T \frac{\partial^2 \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})$ and $\frac{\partial^3 \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_j \partial \beta_k \partial \theta_l} = \boldsymbol{X}_{ij}^T \frac{\partial \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_l} \boldsymbol{X}_{ik}$ for all $\boldsymbol{\beta} \in \boldsymbol{\Omega}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In addition, there exist functions M_{jkl} and N_{jkl} such that

$$\left| \frac{\partial^3}{\partial \beta_j \partial \theta_k \partial \theta_l} \log f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) \right| = \left| \boldsymbol{X}_{ij}^T \frac{\partial^2 \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \right| \leq M_{jkl}(\boldsymbol{Y}_i, \boldsymbol{X}_i)$$

and

$$\left| \frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \theta_l} \log f(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) \right| = \left| \boldsymbol{X}_{ij}^T \frac{\partial \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_l} \boldsymbol{X}_{ik} \right| \leq N_{jkl}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$$

for all $\boldsymbol{\beta} \in \boldsymbol{\Omega}$ and $\boldsymbol{\theta} \in \Theta$, and $m_{jkl} = E_{(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}[M_{jkl}(\boldsymbol{Y}_i|\boldsymbol{X}_i)] < \infty$, and $n_{jkl} = E_{(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}[N_{jkl}(\boldsymbol{Y}_i|\boldsymbol{X}_i)] < \infty$ for j, k, l.

3.3 Proof of Consistency

Let $\boldsymbol{X}_{\mathrm{FL}}$ be the covariates selected using Fan and Li's (2001) method and $\hat{\boldsymbol{\theta}}_m$ be the REML estimate based on the $\boldsymbol{X}_{\mathrm{FL}}$. Then the log-likelihood function of observations $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ is

$$\ell(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m) = -\frac{1}{2} \log |\boldsymbol{V}(\hat{\boldsymbol{\theta}}_m)| - \frac{1}{2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Denote $Q(\boldsymbol{\beta})$ as the SCAD penalized log-likelihood function

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m) - m\Sigma_{j=1}^{d} p_{\lambda_m}(|\beta_j|).$$

Before the proof of Theorem 1, we prove the consistency of $\hat{\boldsymbol{\theta}}_m$ which is estimated from the simple procedure.

Lemma 1, Let $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ be independently distributed each with the normal density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\boldsymbol{\beta},\boldsymbol{\theta})$ which satisfies regularity conditions (1) and (2). Assume $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ follow the linear mixed model (2.1). Let $\boldsymbol{X}_{\mathrm{FL}}$ be the covariates selected using Fan and Li's (2001) method and $\hat{\boldsymbol{\theta}}_m$ be the REML estimate based on $\boldsymbol{X}_{\mathrm{FL}}$. Then $\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 = O_P(m^{-\frac{1}{2}})$.

proof: Let $\boldsymbol{X}_0$ be the true covariates. By the consistency of REML estimate $\hat{\boldsymbol{\theta}}$ (Searle et al. (1992)), given the event $\mathrm{E} = [\boldsymbol{X}_{\mathrm{FL}} = \boldsymbol{X}_0]$, $\forall \epsilon > 0$, $\exists M_\epsilon$ and m_ϵ such that when $m > m_\epsilon$,

$$\mathrm{P}[|\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| > M_\epsilon | \mathrm{E}] < \epsilon.$$

Since Fan and Li (2001) proved the oracle property of $\hat{\boldsymbol{\beta}}$ for their method, which implies $\mathrm{P}(\mathrm{E}) \to 1$ and $\mathrm{P}(\overline{\mathrm{E}}) \to 0$.

Now $\forall \epsilon > 0$, $\exists M_\epsilon$ and m_ϵ such that when $m > m_\epsilon$,

$$\begin{aligned}
& \mathrm{P}[|\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| > M_\epsilon] \\
= \; & \mathrm{P}[|\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| > M_\epsilon | \mathrm{E}]\mathrm{P}(\mathrm{E}) + \mathrm{P}[|\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| > M_\epsilon | \overline{\mathrm{E}}]\mathrm{P}(\overline{\mathrm{E}}) \\
< \; & \epsilon \times 1 + 1 \times \epsilon = 2\epsilon,
\end{aligned}$$

which implies $\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 = O_P(m^{-\frac{1}{2}})$.

Next, we prove Theorem 1 which describes the consistence of $\hat{\boldsymbol{\beta}}$.

Theorem 1, Let $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ be independently distributed each with the normal density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{\beta})$ which satisfies regularity conditions (1) and (2). Assume $\boldsymbol{Y}_1|\boldsymbol{X}_1$,...,$\boldsymbol{Y}_m|\boldsymbol{X}_m$ follow the linear mixed model (2.1). Let $\boldsymbol{X}_{\mathrm{FL}}$ be the covariates selected using Fan and Li's (2001) method and $\hat{\boldsymbol{\theta}}_m$ be the REML estimate based on $\boldsymbol{X}_{\mathrm{FL}}$. If $h_m = \max\{|p''_{\lambda_m}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m)$ such that $\hat{\boldsymbol{\beta}}$ is $\sqrt{m}$ consistent to $\boldsymbol{\beta}_0$.

Proof: Let $\alpha_m = m^{-\frac{1}{2}} + a_m$, where $a_m = \max\{|p'_{\lambda_m}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$. We want to show that for any given $\epsilon \geq 0$, there exists a large constant C such that

$$P\{\sup_{||u||=C} Q(\boldsymbol{\beta}_0 + \alpha_n u, \hat{\boldsymbol{\theta}}_m) < Q(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}_m)\} \geq 1 - \epsilon. \tag{3.5}$$

Note that

$$\begin{aligned}
& Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m) - Q(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}_m) \\
= \;& \ell(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m) - \ell(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}_m) - m[\sum_{j=1}^{s} p_{\lambda_m}(|\beta_j|) - \sum_{j=1}^{s} p_{\lambda_m}(|\beta_{j0}|)] \\
= \;& \ell(\boldsymbol{\beta}_0 + \alpha_m u, \hat{\boldsymbol{\theta}}_m) - \ell(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}_m) - m[\sum_{j=1}^{s} p_{\lambda_m}(|\beta_{j0} + \alpha_m u_j|) \\
& - \sum_{j=1}^{s} p_{\lambda_m}(|\beta_{j0}|)],
\end{aligned} \tag{3.6}$$

where s is the number of nonzero components in $\boldsymbol{\beta}$, or the dimension of $\boldsymbol{\beta}_{10}$ and q is the dimension of $\boldsymbol{\theta}$. In particular, we have

$$\begin{aligned}
& \ell(\boldsymbol{\beta}_0 + \alpha_m u, \hat{\boldsymbol{\theta}}_m) - \ell(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}_m) \\
= \;& \alpha_m u^T \boldsymbol{X}^T \boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0) + \frac{1}{2} u^T [-\boldsymbol{X}^T \boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)\boldsymbol{X}] u \alpha_m^2.
\end{aligned} \tag{3.7}$$

Now

$$
\begin{aligned}
& \frac{1}{m}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0) \\
= & \frac{1}{m}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)+\frac{1}{m}\sum_{l=1}^{q}\boldsymbol{X}^T\frac{\partial\boldsymbol{V}^{-1}(\boldsymbol{\theta}_{l0})}{\partial\boldsymbol{\theta}_l}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_{ml}-\boldsymbol{\theta}_{l0}) \\
& +\frac{1}{2m}\sum_{l=1}^{q}\sum_{k=1}^{q}\boldsymbol{X}^T\frac{\partial^2\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_l\partial\boldsymbol{\theta}_k}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_{ml}-\boldsymbol{\theta}_{l0})(\hat{\boldsymbol{\theta}}_{mk}-\boldsymbol{\theta}_{k0})O_P(1). \quad (3.8)
\end{aligned}
$$

Since $\mathrm{E}[\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)]=0$, $\mathrm{Var}[\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)]=\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}_i$ and regularity condition 1, it follows from the multivariate central limit theorem (Rao (1973)) that

$$
\sqrt{m}\frac{1}{m}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)\xrightarrow{D}N(0,\boldsymbol{A}(\boldsymbol{\theta}_0)), \quad (3.9)
$$

where $\boldsymbol{A}(\boldsymbol{\theta}_0)=\lim\limits_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}_i$ and $\boldsymbol{A}(\boldsymbol{\theta}_0)$ exists and is positive definite. Equivalently, (3.9) implies $m^{-\frac{1}{2}}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)=O_P(1)$, or,

$$
\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)=O_P(m^{\frac{1}{2}}).
$$

Thus by the consistency of $\boldsymbol{\theta}_m$ given in lemma 1,

$$
\sum_{l=1}^{q}\boldsymbol{X}^T\frac{\partial\boldsymbol{V}^{-1}(\boldsymbol{\theta}_{l0})}{\partial\boldsymbol{\theta}_l}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_{ml}-\boldsymbol{\theta}_{l0})=O_p(m^{\frac{1}{2}})O_P(m^{-\frac{1}{2}})=O_P(1).
$$

In addition, the regularity condition (2) implies $\frac{1}{m}\sum_{l=1}^{q}\sum_{k=1}^{q}\boldsymbol{X}^T\frac{\partial^2\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_l\partial\boldsymbol{\theta}_k}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)=O_P(1)$, we can simplify $\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)$ in a form as

$$
\begin{aligned}
& \boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0) \\
= & O_P(m^{\frac{1}{2}})+O_P(1)+\frac{1}{2}O_P(m)O_P(m^{-\frac{1}{2}})O_P(m^{-\frac{1}{2}})O_P(1) \\
= & O_P(m^{\frac{1}{2}})+O_P(1)+\frac{1}{2}O_P(1). \quad (3.10)
\end{aligned}
$$

Similarly, for $-\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)\boldsymbol{X}$ in the second term of (3.7), we have

$$
\begin{aligned}
& -\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)\boldsymbol{X} \\
= \quad & -\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X} - \boldsymbol{X}^T\sum_{l=1}^{q}\frac{\partial\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)}{\partial\theta_l}\boldsymbol{X}(\hat{\theta}_{ml}-\theta_{0l})O_P(1) \\
= \quad & -m\boldsymbol{A}(\boldsymbol{\theta}_0) - O_P(m)O_P(m^{-\frac{1}{2}})O_P(1) \\
= \quad & -m\boldsymbol{A}(\boldsymbol{\theta}_0) - O_P(m^{\frac{1}{2}}). \qquad (3.11)
\end{aligned}
$$

By (3.10) and (3.11), (3.7) is equivalent to

$$
\begin{aligned}
& \alpha_m u^T\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0) + \frac{1}{2}u^T[-\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)\boldsymbol{X}]u\alpha_m^2 \\
= \quad & u^T O_P(m^{\frac{1}{2}}\alpha_m) + \alpha_m u O_P(1) + \frac{1}{2}\alpha_m u O_P(1) \\
& -\frac{1}{2}u^T\boldsymbol{A}(\boldsymbol{\theta}_0)um\alpha_m^2 - \frac{1}{2}u^T O_P(m^{\frac{1}{2}}\alpha_m^2)u \qquad (3.12)
\end{aligned}
$$

as $\alpha_m = m^{-\frac{1}{2}} + a_m$ and $a_m = \max\{|p'_{\lambda m}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$, $O_P(m^{\frac{1}{2}}\alpha_m) = O_P(m\alpha_m^2)$, therefore, when C is chosen sufficient large, $-\frac{1}{2}u^T\boldsymbol{A}(\boldsymbol{\theta}_0)um\alpha_m^2$ dominates all other terms in (3.12).

Fan and Li (2001) indicated that the penalty terms $-m[\sum_{j=1}^{s}p_{\lambda_m}(|\beta_{j0}+\alpha_m u_j|) - \sum_{j=1}^{s}p_{\lambda_m}(|\beta_{j0}|)]$ in (3.6) are bounded by $\sqrt{s}m\alpha_m a_m||u|| + m\alpha_m^2 h_m||u||^2$, as $m\alpha_m a_m = O_P(m\alpha_m^2)$, the penalty terms are dominated by $-\frac{1}{2}u^T\boldsymbol{A}(\boldsymbol{\theta}_0)um\alpha_m^2$ when C is chosen sufficient large and $h_m \to 0$. As a result, (3.5) holds when C is sufficiently large. Thus theorem 1 is proved.

3.4 Proof of Sparsity

It is clear from Theorem 1 that by choosing a proper λ_m, there exists a root-m consistent penalized likelihood estimator. We now show that this estimator possesses the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma 2: Let $\boldsymbol{Y}_1|\boldsymbol{X}_1,...,\boldsymbol{Y}_m|\boldsymbol{X}_m$ be independently distributed each with the conditional normal density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\boldsymbol{\beta},\boldsymbol{\theta})$ which satisfies regularity conditions (1) and (2). Assume $\boldsymbol{Y}_1|\boldsymbol{X}_1,...,\boldsymbol{Y}_m|\boldsymbol{X}_m$ follow the linear mixed model (2.1). If the penalty function satisfies the condition $\liminf_{m\to\infty}\liminf_{\boldsymbol{\beta}\to 0^+}\lambda_m^{-1}p'_{\lambda_m}(\boldsymbol{\beta}) > 0$ when $\lambda_m \to 0$ and $\sqrt{m}\lambda_m \to \infty$ as $m \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_P(m^{-\frac{1}{2}})$ and any constant C,

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ 0\end{pmatrix}, \hat{\boldsymbol{\theta}}_m\right\} = \max_{||\boldsymbol{\beta}_2||\le Cm^{-\frac{1}{2}}} Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}, \hat{\boldsymbol{\theta}}_m\right\}.$$

proof: Note

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j} &= \frac{\partial \ell(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j} - mp'_{\lambda_m}(|\beta_j|)\mathrm{sgn}(\beta_j)\\ &= \boldsymbol{X}_j^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}_j\boldsymbol{\beta}_{j0}) - mp'_{\lambda_m}(|\beta_j|)\mathrm{sign}(\beta_j). \qquad (3.13)\end{aligned}$$

By (3.10), (3.13) is equivalent to

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j} &= O_P(m^{\frac{1}{2}}) + O_P(1) + \frac{1}{2}O_P(1) - mp'_{\lambda_m}(|\beta_j|)\mathrm{sign}(\beta_j)\\ &= m\lambda_m\{-\lambda_m^{-1}p'_{\lambda_m}(|\beta_j|)\mathrm{sign}(\beta_j) + \frac{m^{-\frac{1}{2}}}{\lambda_m}O_P(1) + m^{-\frac{1}{2}}\frac{m^{-\frac{1}{2}}}{\lambda_m}O_P(1)\\ &\quad + \frac{1}{2}m^{-\frac{1}{2}}\frac{m^{-\frac{1}{2}}}{\lambda_m}O_P(1)\}\end{aligned}$$

Since $\liminf_{m\to\infty}\liminf_{\boldsymbol{\beta}\to 0^+}\lambda_m^{-1}p'_{\lambda_m}(\boldsymbol{\beta}) > 0$ and $\frac{m^{-\frac{1}{2}}}{\lambda_m} \to 0$, the sign of the derivatives is determined by that of β_j, thus with probability tending to 1 as $m \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(m^{-\frac{1}{2}})$ and for some small $\epsilon_m = Cm^{-\frac{1}{2}}$ and $j = s+1, ..., d$, we have

$$\frac{\partial Q(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j} < 0 \qquad \text{for } \ 0 < \beta_j < \epsilon_m$$

and

$$\frac{\partial Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_m)}{\partial \beta_j} > 0 \qquad \text{for} \quad -\epsilon_m < \beta_j < 0$$

therefore,

$$Q\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ 0 \end{pmatrix}, \hat{\boldsymbol{\theta}}_m\right\} = \max_{||\boldsymbol{\beta}_2|| \leq Cm^{-\frac{1}{2}}} Q\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \hat{\boldsymbol{\theta}}_m\right\}.$$

3.5 Proof of Oracle Property

Theorem 2: Let $\boldsymbol{Y}_1|\boldsymbol{X}_1,...,\boldsymbol{Y}_m|\boldsymbol{X}_m$ be independently distributed each with the conditional normal density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i;\boldsymbol{\beta},\boldsymbol{\theta})$ which satisfies regularity conditions. Assume $\boldsymbol{Y}_1|\boldsymbol{X}_1,...,\boldsymbol{Y}_m|\boldsymbol{X}_m$ follow the linear mixed model (2.1). Let $\boldsymbol{h} = (p'_{\lambda_m}(|\beta_{10}|)\text{sign}(\beta_{10}), ..., p'_{\lambda_m}(|\beta_{s0}|))\text{sign}(\beta_{s0}))^T$, if the penalty function satisfies $\liminf_{m\to\infty} \liminf_{\boldsymbol{\beta}\to 0^+} \lambda_m^{-1} p'_{\lambda_m}(\boldsymbol{\beta}) > 0$ when $\lambda_m \to 0$ and $\sqrt{m}\lambda_m \to \infty$ as $m \to \infty$, then with probability tending to 1, the root m consistent local maximizers $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Theorem 1 must satisfy:

(i) (**Sparsity**) $\hat{\boldsymbol{\beta}}_2 = 0$;

(ii) (**Asymptotic normality**)

$$\sqrt{m}(\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m) + \boldsymbol{\Sigma})\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\boldsymbol{I}_1(\boldsymbol{\theta}_0) + \boldsymbol{\Sigma})^{-1}\boldsymbol{h}\} \xrightarrow{D} N(0, \boldsymbol{A}_1(\boldsymbol{\theta}_0)),$$

where $\boldsymbol{I}_1(\boldsymbol{\theta}) = \frac{1}{m}\boldsymbol{X}_{\text{imp}}^T \boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{\text{imp}}$, $\boldsymbol{A}_1(\boldsymbol{\theta}) = \lim_{m\to\infty} \frac{1}{m}\boldsymbol{X}_{\text{imp}}^T \boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{\text{imp}}$ and $\boldsymbol{X}_{\text{imp}}$ includes the important covariate only.

proof: Lemma 2 shows sparsity. We now prove asymptotic normality, by Theorem 1 and Lemma 2, there is $\hat{\boldsymbol{\beta}}_1$ which is a root-m consistent local maximizer of

$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\0\end{pmatrix},\hat{\boldsymbol{\theta}}_m\right\}$ and satisfies the estimating equation:

$$\frac{\partial Q(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\0\end{pmatrix}}=0,\qquad \text{for}\qquad j=1,...,s.$$

Since

$$\begin{aligned}0 &= \frac{\partial Q(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\0\end{pmatrix}}\\ &= \frac{\partial\ell(\beta_{j0},\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j}+\sum_{l=1}^{s}\frac{\partial^2\ell(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}}_m)}{\partial\beta_j\partial\beta_l}(\hat{\beta}_l-\beta_{l0})\\ &\quad -m\{p'_{\lambda_m}(|\beta_{j0}|)\mathrm{sign}(\beta_{j0})+[p''_{\lambda_m}(|\beta_{j0}|)+o_P(1)](\hat{\beta}_j-\beta_{j0})\}),\end{aligned}\tag{3.14}$$

according to (3.9) and slutsky's theorem,

$$\sqrt{m}\frac{1}{m}\frac{\partial\ell(\boldsymbol{\beta}_0,\hat{\boldsymbol{\theta}}_m)}{\partial\boldsymbol{\beta}_1}=m^{-\frac{1}{2}}\boldsymbol{X}_{\text{imp}}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{Y}-\boldsymbol{X}_{\text{imp}}\boldsymbol{\beta}_{10})\xrightarrow{D}N(0,\boldsymbol{A}_1(\boldsymbol{\theta}_0)),\tag{3.15}$$

where $\boldsymbol{A}_1(\boldsymbol{\theta}_0)=\lim_{m\to\infty}\frac{1}{m}\boldsymbol{X}_{\text{imp}}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}_{\text{imp}}$. Regularity condition (1) implies $\boldsymbol{A}_1(\boldsymbol{\theta}_0)$ is finite and positive definite. Let $h_j=p'_{\lambda_m}(|\beta_{j0}|)\mathrm{sign}(\beta_{j0})$ and $\Sigma_{jj}=p''_{\lambda_m}(|\beta_{j0}|)$, by (3.14), (3.15) and slutsky's theorem, we have

$$\sqrt{m}(\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m)+\boldsymbol{\Sigma})\{\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+[\boldsymbol{I}_1(\boldsymbol{\theta}_0)+\boldsymbol{\Sigma}]^{-1}\boldsymbol{h}\}\xrightarrow{D}N(0,\boldsymbol{A}_1(\boldsymbol{\theta}_0)),$$

where $\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m)=\frac{1}{m}\boldsymbol{X}_{\text{imp}}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\theta}}_m)\boldsymbol{X}_{\text{imp}}$, $\boldsymbol{A}_1(\boldsymbol{\theta}_0)=\lim_{m\to\infty}\frac{1}{m}\boldsymbol{X}_{\text{imp}}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}_{\text{imp}}$ and $\boldsymbol{X}_{\text{imp}}$ includes important covariate only. As a consequence, the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_1)=\frac{1}{m}[\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m)+\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1)]^{-1}\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m)[\boldsymbol{I}_1(\hat{\boldsymbol{\theta}}_m)+\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_1)]^{-1}.\tag{3.16}$$

3.6 Discussion

For the SCAD penalty function, if $\lambda_m \to 0$, then $a_m = 0$. By Theorem 2, when $\sqrt{m}\lambda_m \to \infty$, the penalized likelihood estimator $\boldsymbol{\beta}$ from the simple procedure possesses the oracle property and perform as well as the estimate for $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = 0$.

This chapter demonstrates how the rates of convergence for the penalized likelihood estimator $\boldsymbol{\beta}$ from the simple procedure depends on the number of subjects. We have shown that $\boldsymbol{\beta}$ performs as well as the oracle procedure in terms of selecting the correct model, under certain conditions. In other words, when the true parameters have some zero components, they are estimated as 0 with probability tending to one, and the non-zero components are estimated as well as if the correct submodel were known.

It is much more complicated to study the theoretical properties of $\hat{\boldsymbol{\beta}}$ obtained from the iterative procedure because of the more sophisticated algorithm. Through simulation study in Chapter 4, We will present a detailed investigation of the performance for $\hat{\boldsymbol{\beta}}$ obtained from the iterative procedure.

Chapter 4

Simulation Studies

4.1 Introduction

We have presented the SCAD penalized likelihood variable selection method for linear mixed models in longitudinal data in Chapter 2. We have shown that $\hat{\boldsymbol{\beta}}$ possesses the oracle property under certain conditions. In this chapter, we study the finite sampling performance of the proposed procedures and compared with Fan and Li's (2001) procedure which ignores the correlation for longitudinal data.

4.2 Simulation Design

We carry out simulation studies for the linear mixed model (2.1) in the form of

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \epsilon_{ij},$$

where $i = 1, ..., m$, m represents the total number of subjects, $j = 1, ..., n_i$, n_i represents the total number of responses measured for subject i. We generate n_i from $Bin(7, 0.7)$, so that on average, each subject has about five observations.

We add the intercept $\beta_0 = 2$ in all simulation examples, but we do not penalize β_0. We generate 9 independent variables from a multivariate normal distribution. To simulate practical longitudinal data, we assume X_1, X_4, X_5, X_6, X_9 are one time covariates, X_2, X_3, X_7, X_8 are time varying covariates. To make the example more challenging, we make the covariates correlated: the correlation among four one time variables has an AR(1) structure with $\rho = 0.5$. Within each subject, the correlation among n_i observations has an AR(1) correlation structure with $\rho = 0.5$.

The random errors ϵ_{ij} independently follow a normal distribution $N(0, \sigma_\epsilon^2)$. We consider two types of random effect for $\boldsymbol{b}_i$: (1) the random intercept only, (2) the random intercept and random slope.

For the random intercept only case, $b_i \sim N(0, \sigma_{b0}^2)$, Let $\boldsymbol{Z}_i$ be the design matrix for b_i, then $\boldsymbol{Z}_i$ is the constant vector $\mathbf{1}_{n_i}$ and $\boldsymbol{b} = (b_1, b_2, ..., b_m)^T$,

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \mathbf{0} & ... & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_2 & ... & \mathbf{0} & ... & \mathbf{0} \\ . & . & ... & . & ... & . \\ \mathbf{0} & \mathbf{0} & ... & \boldsymbol{Z}_i & ... & \mathbf{0} \\ . & . & ... & . & ... & . \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} & ... & \boldsymbol{Z}_m \end{pmatrix}_{n \times m}.$$

For the random intercept and random slope case, $\boldsymbol{b}_i$ follows bivariate normal $\boldsymbol{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b01} \\ \sigma_{b01} & \sigma_{b1}^2 \end{bmatrix}\right)$. Let $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2})$ be the design matrix for the random effects $\boldsymbol{b}_i$, where $\boldsymbol{Z}_{i1}$ is the constant vector $\mathbf{1}_{n_i}$, $\boldsymbol{Z}_{i2} = (\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, ..., \frac{n_i}{5})^T$,

$\boldsymbol{b} = (\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, ..., \boldsymbol{b}_m^T)^T$ and

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \mathbf{0} & ... & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_2 & ... & \mathbf{0} & ... & \mathbf{0} \\ . & . & ... & . & ... & . \\ \mathbf{0} & \mathbf{0} & ... & \boldsymbol{Z}_i & ... & \mathbf{0} \\ . & . & ... & . & ... & . \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} & ... & \boldsymbol{Z}_m \end{pmatrix}_{n \times 2m} .$$

For simplicity, we set true values $\sigma^2 = 2$, the nonzero coefficients are clustered in the first four variables from X_1 to X_4.

We investigate the performance of our procedures and Fan and Li's (2001) method under two settings: $\boldsymbol{\beta} = (2, 2, 1.5, 1, 0.5, 0, 0, 0, 0, 0)^T$ and $\boldsymbol{\beta} = (2, 1, 1, 1, 1, 0, 0, 0, 0, 0)^T$. Figure 4.1 shows the values of nonzero regression coefficients for two different settings. Also two different forms of random effect $\boldsymbol{b}_i$, one is with random intercept only and $\boldsymbol{b}_i = b_{i0} \sim N(0, \sigma_0^2 = 0.8^2)$, another is with random intercept and random slope, $\boldsymbol{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b01} \\ \sigma_{b01} & \sigma_{b1}^2 \end{bmatrix} = \begin{bmatrix} 0.8^2 & 0.12 \\ 0.12 & 0.6^2 \end{bmatrix}\right)$. Totally we have $2 \times 2 \times 2 = 8$ scenarios in the design of the simulation with sample size $m = 100, 200$. For each simulation, $S = 100$ sets of data at each of the scenarios are generated.

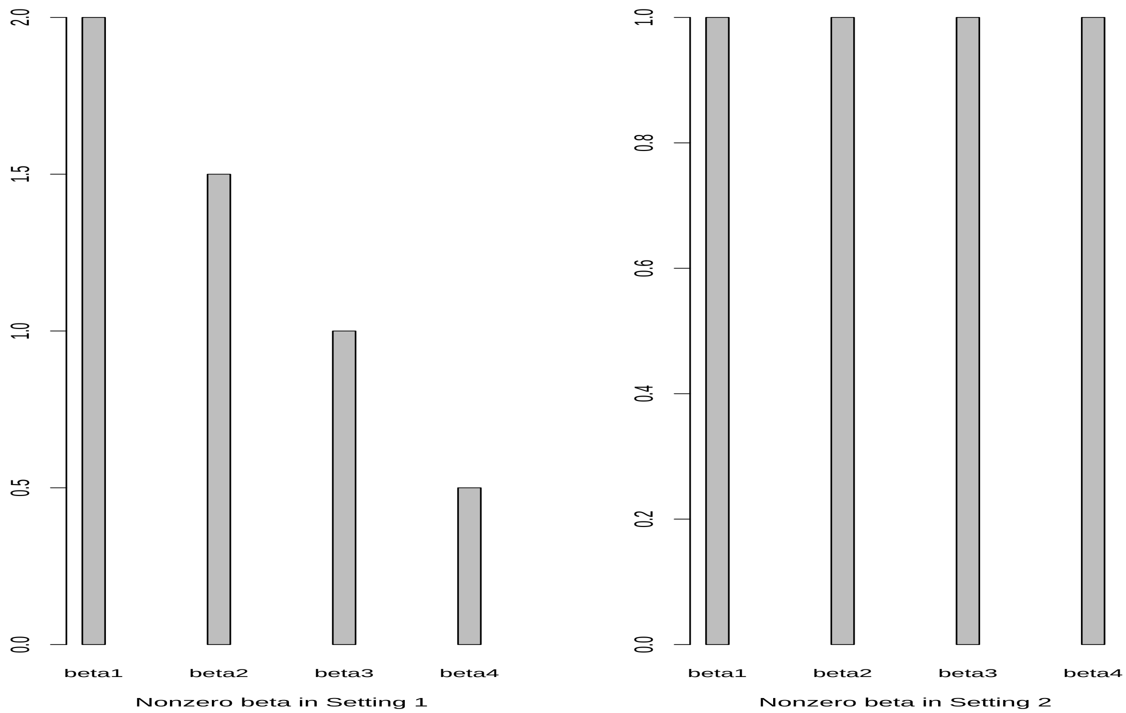

Figure 4.1: Values of nonzero regression coefficients $\beta_1, \beta_2, \beta_3, \beta_4$.

4.3 Data Analysis

For the purpose of this simulation, we select significant variables and estimate parameters for each data set under each scenario by three ways: (i), Fan and Li's (2001) approach, denoted by FL, which uses standard SCAD by ignoring the correlation in the longitudinal data; (ii), Simple procedure, denoted by SMP; (iii), Iterative procedure, denoted by ITER. For Fan and Li's (2001) procedure, we use GCV as a criterion to select λ, and for our procedures, we apply GCV score, REML value and $\mathrm{BIC}_{\mathrm{scad}}$.

To evaluate the variable selection methods in terms of variable selection and model fitting, we report five different measures: (1) the median of model errors (ME, see (1.2)) over 100 simulated data sets, (2) the median of relative model errors (MRME) over 100 data, or the model error of the proposed procedures is compared to that of the weighted least squares estimator for the true model, (3) the average number of selected true 0 coefficients (Correct 0), (4) the average number of coefficients erroneously set to 0 (Incorrect 0), and (5) the average model size (Average Size).

To examine the performance of the parameter estimates, we report (1), the Monte Carlo sample mean of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ over 100 data sets, denoted by $\beta_1, \beta_2, \beta_3, \beta_4$; (2), the median absolute deviation divided by 0.6745, denoted by SD in tables, is regarded as the estimate of true standard error; (3), the median of the 100 estimated SD's, denoted by $\mathrm{SD_m}$.

4.4 Simulation Results

4.4.1 Model Selection

Based on the simulation design, the true model includes the subset X_1 X_2 X_3 X_4 and the intercept X_0, since we do not add penalty on β_0 and β_0 is always included in all the models, the true model size is 4, the number of true zero coefficients is 5 and the number of true nonzero coefficients is 4.

Table 4.1, 4.2 and 4.3 present the model selection results under the first setting $(\beta_1, \beta_2, \beta_3, \beta_4)^T = (2, 1.5, 1, 0.5)^T$ and $\boldsymbol{b} = (b_0, b_1)^T$. Figure 4.2 plots the average model error for FL, SMP and ITER. The advantage of estimating the variance components in the random effects is appreciated via the facts: it is obvious that MEs and MRMEs for SMP and ITER are smaller than those for FL; reflected in the average model size, SMP and ITER are closer to the truth than FL; the average numbers of true 0 coefficient for FL is further away from the truth five than that for SMP and ITER. These findings indicate that SMP and ITER outperform FL when data are correlated.

Comparing SMP to ITER, the latter performs better in terms of the evaluation values such as ME, MRME, Correct 0 and Average Size. It is reasonable since iteratively estimating $\boldsymbol{\theta}$ may improve the accuracy of the estimates.

Moreover, when the subject size m increases from 100 to 200, both the ME and MRME become smaller, the average model size and the average number of true 0 coefficients are closer to the true sizes.

The results for all other scenarios follow the similar pattern, more tables are given in Appendix.

Table 4.1: Selection Results in Setting 1 with Random Intercept and Random Slope Using GCV

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.50	124.8	3.85	0.02	5.15
100	SMP	0.41	79.6	4.41	0.02	4.59
100	ITER	0.38	74.1	4.57	0.02	4.43
200	FL	0.24	98.0	3.98	0.01	5.02
200	SMP	0.16	52.1	4.56	0.01	4.44
200	ITER	0.15	42.7	4.67	0.01	4.33

Table 4.2: Selection Results in Setting 1 with Random Intercept and Random Slope Using REML

size(m)	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	SMP	0.36	78.1	4.55	0.02	4.45
100	ITER	0.32	66.1	4.59	0.02	4.41
200	SMP	0.15	61.4	4.68	0	4.32
200	ITER	0.15	60.0	4.70	0	4.30

Table 4.3: Selection Results in Setting 1 with Random Intercept and Random Slope Using $\mathrm{BIC}_{\mathrm{scad}}$

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.37	106.2	4.75	0.02	4.25
100	SMP	0.31	59.0	4.79	0.02	4.21
100	ITER	0.23	33.2	4.79	0.02	4.21
200	FL	0.19	100.3	4.81	0.01	4.19
200	SMP	0.15	58.3	4.83	0	4.17
200	ITER	0.13	32.1	4.83	0	4.17

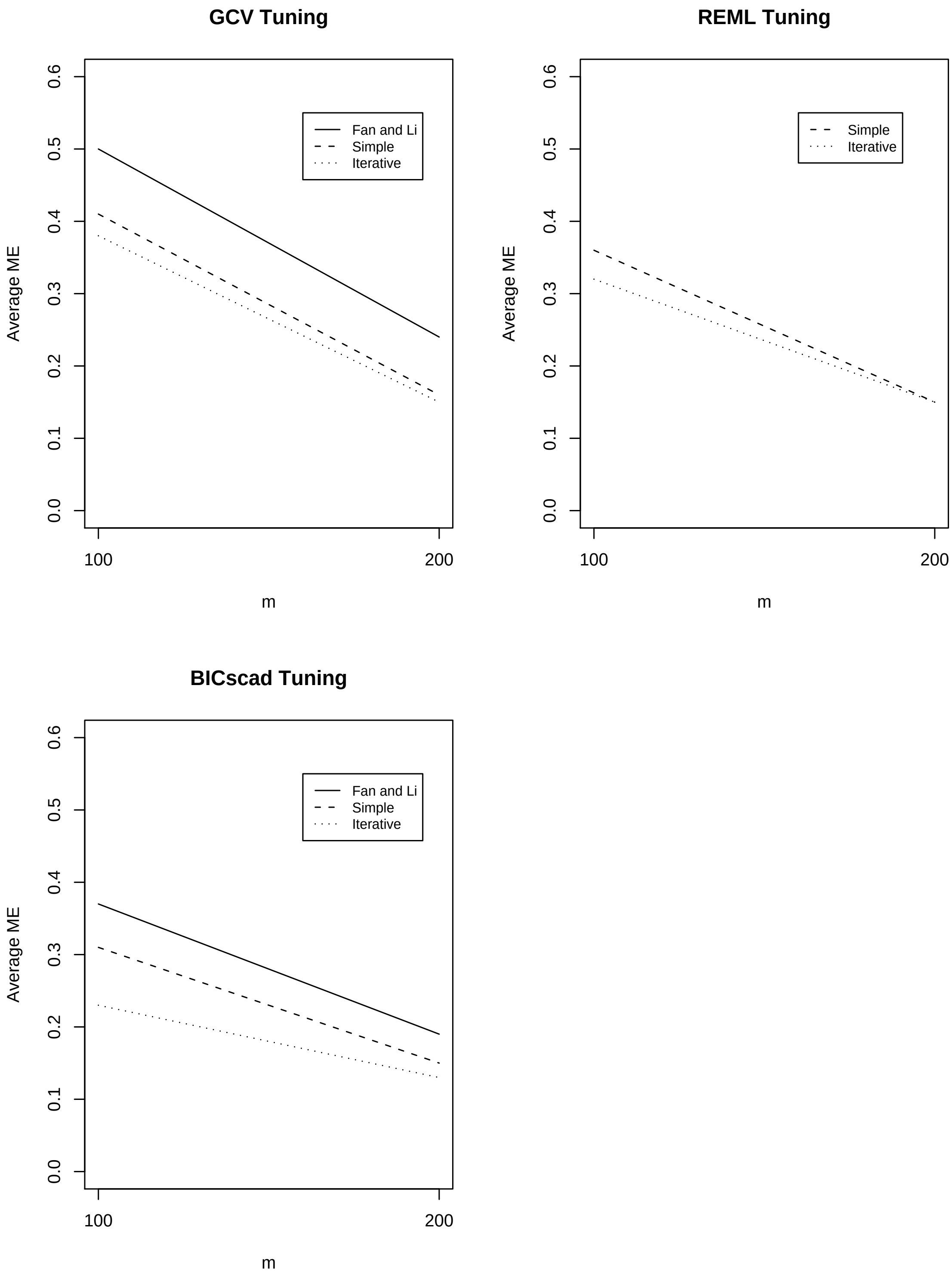

Figure 4.2: Average Model Error for FL, SMP and ITER.

4.4.2 Parameter Estimation

As stated previously, the proposed procedures select variables and estimate parameters simultaneously. We are interested in investigating how good the parameter estimates are, and how well the proposed sandwich formula (2.33) performs.

Results for $\hat{\boldsymbol{\beta}}$

Simulation results indicate the estimates for the intercept β_0 are always unbiased. Since we are more interested in the estimates for the regression coefficients of the important variables, $\hat{\beta}_0$ is not reported. Table (4.4) summarizes the estimates for ($\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$) under the first setting $(\beta_1, \beta_2, \beta_3, \beta_4) = (2, 1.5, 1, 0.5)$ and $\boldsymbol{b} = (b_0, b_1)^T$ with λ tuned by GCV score. The relative bias for (β_1, β_2, β_3) are small (less than 2%), while for β_4, the relative bias is relatively large. The weak signal of X_4 with $\beta_4 = 0.5$ may bring relative bias of $\hat{\beta}_4$. Reflected in Table (4.5) under $(\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 1, 1)^T$, stronger signal of variable X_4 results in less bias to $\hat{\beta}_4$. Noted in both (4.4) and (4.5), the median absolute deviation divided by 0.6745 (SD) is very close to the corresponding $\mathrm{SD_m}$, the median of the 100 estimated SD, this suggests that the sandwich formula (2.33) performs well.

Under the same scenario for Table (4.4), referring to the observation for REML tuning in Table (4.6), REML tuning produces negligible bias for important coefficients including β_4. Also, the small difference between SD and $\mathrm{SD_m}$ again indicates the performance of the sandwich formula is very good.

Table 4.4: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept and Random Slope Using GCV, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.99(0.16)	0.11	1.51(0.12)	0.10
100	SMP	1.99(0.15)	0.15	1.51(0.11)	0.11
100	ITER	1.99(0.15)	0.14	1.51(0.11)	0.11
200	FL	2.00(0.10)	0.08	1.49(0.10)	0.09
200	SMP	2.01(0.10)	0.11	1.50(0.08)	0.08
200	ITER	2.01(0.10)	0.10	1.50(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.00(0.13)	0.11	0.51(0.18)	0.11
100	SMP	0.99(0.12)	0.11	0.50(0.16)	0.15
100	ITER	0.99(0.12)	0.11	0.50(0.16)	0.14
200	FL	1.00(0.08)	0.06	0.51(0.12)	0.11
200	SMP	1.00(0.08)	0.08	0.51(0.11)	0.11
200	ITER	1.00(0.08)	0.08	0.51(0.11)	0.10

Table 4.5: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept and Random Slope Using GCV, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	0.99(0.14)	0.09	1.00(0.13)	0.10
100	SMP	1.00(0.13)	0.15	1.00(0.11)	0.11
100	ITER	1.00(0.13)	0.14	1.00(0.10)	0.11
200	FL	0.99(0.10)	0.07	1.00(0.10)	0.07
200	SMP	0.99(0.09)	0.10	1.00(0.08)	0.08
200	ITER	1.00(0.09)	0.10	1.00(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.02(0.12)	0.09	1.00(0.15)	0.09
100	SMP	1.02(0.10)	0.11	1.01(0.14)	0.14
100	ITER	1.02(0.10)	0.11	1.01(0.14)	0.14
200	FL	1.01(0.11)	0.09	1.00(0.12)	0.07
200	SMP	1.01(0.08)	0.08	1.00(0.10)	0.10
200	ITER	1.01(0.07)	0.08	1.00(0.10)	0.10

Table 4.6: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept and Random Slope Using REML, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	SMP	2.01(0.14)	0.14	1.51(0.10)	0.11
100	ITER	2.01(0.14)	0.13	1.51(0.10)	0.11
200	SMP	2.00(0.08)	0.10	1.50(0.08)	0.08
200	ITER	2.00(0.08)	0.09	1.50(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	SMP	1.01(0.13)	0.11	0.51(0.14)	0.14
100	ITER	1.01(0.13)	0.11	0.51(0.13)	0.13
200	IMP	1.00(0.08)	0.08	0.51(0.09)	0.10
200	ITER	1.01(0.08)	0.08	0.51(0.09)	0.09

Tables (4.4) (4.5) (4.6) further demonstrate that the median of the 100 estimated SD of $\hat{\boldsymbol{\beta}}$ from SMP and ITER is smaller than that from FL under the same setting, which shows $\hat{\boldsymbol{\beta}}$ from our method performs better than that from Fan and Li's (2001) when data are longitudinal.

Results For $\hat{\boldsymbol{\theta}}$

Based on the simulation design, the true values for variance components are $\begin{bmatrix} \sigma_{b0}^2 & \sigma_{b01} \\ \sigma_{b01} & \sigma_{b1}^2 \end{bmatrix} = \begin{bmatrix} 0.8^2 & 0.12 \\ 0.12 & 0.6^2 \end{bmatrix}$ and $\sigma_e^2 = 2$. Tables (4.7) and (4.8) summarize the information related to $\hat{\boldsymbol{\theta}}$ estimated by GCV score and REML value respectively. Compared to $\hat{\boldsymbol{\beta}}$, there exists considerable bias for $\hat{\boldsymbol{\theta}}$, especially for the small values of σ_{01} and σ_{b1}^2. The overall performance of $\boldsymbol{\theta}$ is good.

Table 4.7: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept and Random Slope Using GCV, the true value of $(\sigma^2_{b0}, \sigma_{b01}, \sigma^2_{b1}, \sigma^2_e) = (0.64, 0.12, 0.36, 2)$, $\sigma^2_{b0}, \sigma_{b01}, \sigma^2_{b1}, \sigma^2_e$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma^2_{b0}(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.67(0.25)	0.23	0.16(0.21)	0.20
100	ITER	0.66(0.21)	0.24	0.12(0.24)	0.22
200	SMP	0.67(0.21)	0.21	0.15(0.20)	0.21
200	ITER	0.67(0.21)	0.24	0.12(0.21)	0.23
size(m)	Procedure	$\sigma^2_{b1}(SD)$	SD_m	$\sigma^2_e(SD)$	SD_m
100	SMP	0.46(0.22)	0.23	1.97(0.16)	0.17
100	ITER	0.40(0.26)	0.24	1.97(0.16)	0.18
200	SMP	0.42(0.23)	0.22	1.99(0.11)	0.13
200	ITER	0.41(0.24)	0.21	1.99(0.11)	0.13

Table 4.8: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept and Random Slope Using REML, the true value of $(\sigma^2_{b0}, \sigma_{b01}, \sigma^2_{b1}, \sigma^2_e) = (0.64, 0.12, 0.36, 2)$, $\sigma^2_{b0}, \sigma_{b01}, \sigma^2_{b1}, \sigma^2_e$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma^2_{b0}(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.67(0.25)	0.26	0.09(0.20)	0.23
100	ITER	0.68(0.25)	0.23	0.09(0.21)	0.24
200	SMP	0.65(0.19)	0.20	0.11(0.19)	0.20
200	ITER	0.65(0.19)	0.21	0.11(0.20)	0.22
size(m)	Procedure	$\sigma^2_{b1}(SD)$	SD_m	$\sigma^2_e(SD)$	SD_m
100	SMP	0.46(0.20)	0.21	2.00(0.15)	0.16
100	ITER	0.46(0.19)	0.21	2.00(0.15)	0.17
200	SMP	0.40(0.20)	0.21	2.00(0.10)	0.12
200	ITER	0.41(0.20)	0.19	2.00(0.11)	0.12

4.5 Discussion

We investigate the factors which may affect the selection results for the proposed procedures.

4.5.1 Number of Subjects

From Tables (4.1) to (4.8), when sample size m increases from 100 to 200, the sizes of the selected models are always more close to those of underlying true models, and the parameter estimates perform better. This is consistent with the theoretical results, Fan and Li (2001) proves that the penalized likelihood estimator for SCAD is root-n consistent for independent data. We also show that, for longitudinal normal data, $\hat{\boldsymbol{\beta}}$ from SMP possesses oracle property and is root-m consistent. Although it is challenging to prove the theoretical properties for ITER estimates, the simulation results indicate ITER estimates should also be consistent.

4.5.2 Variance Structure for Random Effects

From Tables (4.2) and (4.9), the significant variable selection results for random intercept models are comparable to those for the model with random slope and random intercept. However, the performance of $\hat{\boldsymbol{\theta}}$ from our procedures is affected by the complexity of the random effect structure. Comparing Tables (4.10) to (4.6), Tables (4.11) to (4.8), the performance of $\hat{\boldsymbol{\beta}}$ is similar while $\hat{\boldsymbol{\theta}}$ from the former has less bias and smaller standard deviation than $\hat{\boldsymbol{\theta}}$ from the latter has. It is reasonable because it is more difficult for model fitting when the variance structure is more complicated.

Table 4.9: Selection Results in Setting 1 with Random Intercept Using REML

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	SMP	0.28	62.5	4.53	0	4.47
100	ITER	0.27	52.2	4.54	0	4.46
100	SMP	0.15	66.7	4.64	0	4.28
100	ITER	0.13	64.2	4.77	0	4.22

Table 4.10: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept Using REML, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	SMP	2.00(0.12)	0.12	1.51(0.09)	0.10
100	ITER	2.00(0.12)	0.12	1.51(0.09)	0.11
200	SMP	2.00(0.10)	0.09	1.50(0.08)	0.08
200	ITER	2.00(0.10)	0.09	1.50(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	SMP	1.01(0.11)	0.11	0.49(0.11)	0.12
100	ITER	1.01(0.11)	0.11	0.49(0.14)	0.12
200	SMP	0.99(0.07)	0.08	0.50(0.10)	0.09
200	ITER	0.99(0.07)	0.08	0.50(0.10)	0.09

Table 4.11: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept Using REML, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.64(0.14)	0.16	2.00(0.14)	0.14
100	ITER	0.64(0.14)	0.15	2.00(0.14)	0.14
200	SMP	0.64(0.11)	0.11	2.01(0.10)	0.10
200	ITER	0.64(0.11)	0.11	2.01(0.10)	0.10

4.5.3 Tuning Approach

From Tables (4.1) (4.2) (4.3), it can be seen that tuning method plays an important role in selecting important variables. REML performs slightly better than GCV and BIC_{scad} performs the best among three.

4.5.4 AIC/BIC Forward Selection

To investigate the performance of the standard information criteria variable selection approaches, we use forward selection with AIC/BIC to select significant covariates based on ML and REML. Table (4.12) gives the results. The selection results are the same for two criterion: AIC and BIC. Compare Table (4.12) to Tables (4.1) and (4.2), our method works better than the traditional AIC/BIC forward selection method. Furthermore, we can not detect the trend of selection via AIC/BIC when subject size m increases from 100 to 200.

Table 4.12: Selection Results in Setting 1 with Random Intercept and Random Slope Using AIC/BIC by Forward Selection

size(m)#	Criterion	Correct 0	Incorrect 0	Average Size
100	AIC	4.52	0.02	4.48
100	BIC	4.49	0.02	4.51
200	AIC	4.67	0.01	4.33
200	BIC	4.68	0.01	4.32

4.6 Summary

In summary, by estimating the variance components, the proposed simple procedure and iterative procedure outperform Fan and Li's (2001) approach for longitudinal normal data: the selected models are more close to the truth and the parameter estimates are more accurate. In addition, our procedures can estimate the variance components in the random effect distribution while Fan and Li's (2001) can not. Thus the proposed procedures are promising, especially in linear mixed models.

We are aware that some factors affecting the performance of the proposed procedures, the forms of the random effects, the number of the subjects, the total number of the sample size and the tuning criterion.

The procedures have been implemented using SAS. More simulation results are listed in Appendix.

Chapter 5

Real Data Analysis

5.1 MACS DATA

The Multi-center AIDS Cohort (MACS) Study includes the HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. Kaslow et al. (1987) described the study design, methods, and medical implications in detail. This data has been widely studied, such as Fan and Li (2004), Fan and Zhang (2000), Huang et al. (2002).

It is well known that human immune-dificiency virus (HIV) destroys CD4 cells (or T-lymphocytes, a vital component of the immune system). As a result, the number or percentage of the CD4 cells in the blood of a human body decreases when a patient is infected with HIV. In general, researchers are interested in investigating the association of CD4 cell counts with the candidate covariates.

In this study, subjects were required to visit clinical sites and to be taken measurements semiannually. The response variable is the CD4 cell percentage of a subject at different time points after HIV infection. Because of the missing visits and different starting time for HIV infection, individuals have different number of repeated measurements and distinct measurement times. This data structure is typical for longitudinal data. Figure 5.1 shows the CD4 counts over years since the study for the 283 men.

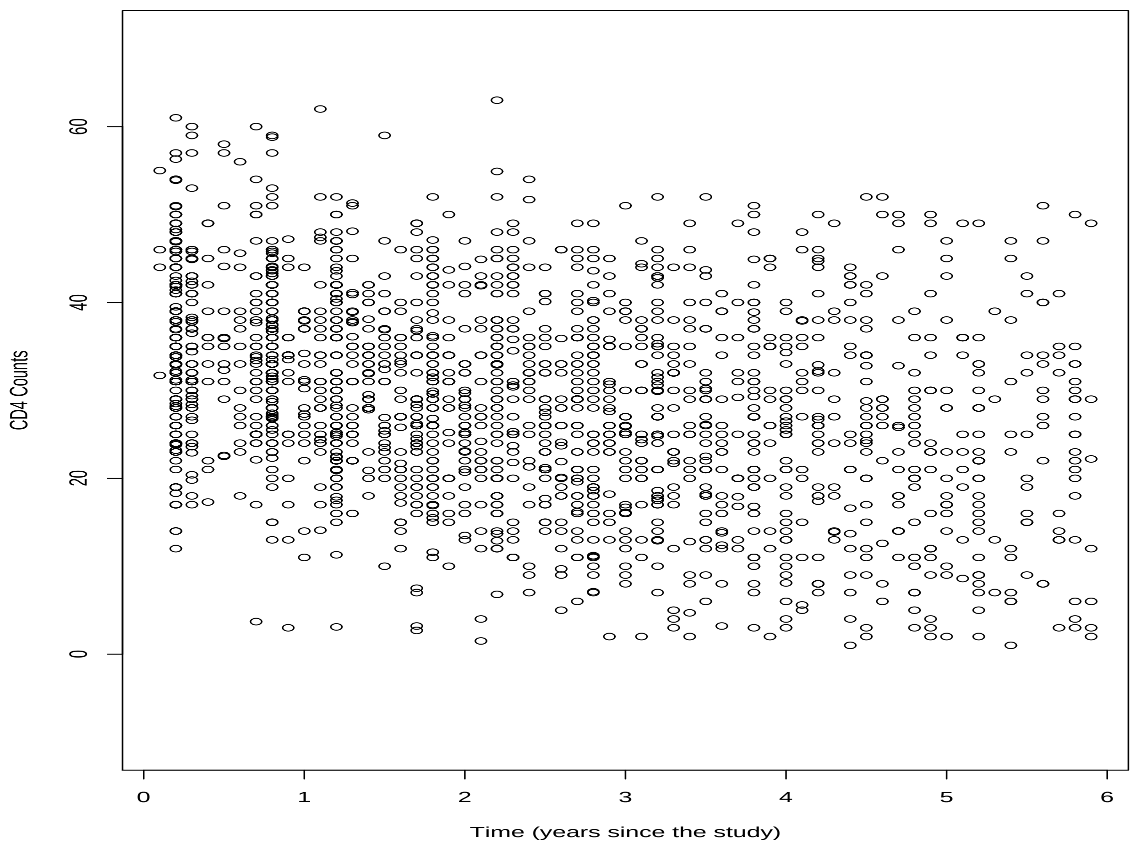

Figure 5.1: CD4 Counts Over Time for Multi-center AIDS Cohort Study.

5.2 Data Analysis Results

We apply the proposed SCAD penalized likelihood method on the Multi-center AIDS Cohort Study. The covariates include: (1) time since the study; (2), a binary variable smoke takes 1 or 0 according to whether a subject is currently smoke or not; (3) the baseline age of the subject when the first measurement was taken; (4), the baseline CD4, i.e., the CD4 cell percentage level before HIV infection.

We are interested in investigating the effects of the baseline CD4, the age, the visit time and smoking status on the CD4 percentage after HIV infection. In addition, we take any pairwise interaction and any quadratic effect from these 4 covariate as predictor variables.

We consider random intercept only, then variance components contain σ_{b0}^2 for the random intercept b_{0i} and σ_e^2 for the measurement error. We further assume the full model is as follows,

$$\begin{aligned} \log CD4_{ij} &= b_{0i} + \beta_1 + \beta_2 \log \text{baselineCD4}_i + \beta_3 \text{age}_i + \beta_4 \text{time}_{ij} + \beta_5 \text{smoke}_i \\ &\quad + \beta_6 \log \text{baselineCD4}_i{}^2 + \beta_7 \text{age}_i^2 + \beta_8 (\text{time}_{ij})^2 + \beta_9 \log \text{baselineCD4}_i \times \text{age}_i \\ &\quad + \beta_{10} \log \text{baselineCD4}_i \times \text{time}_{ij} + \beta_{11} \log \text{baselineCD4}_i \times \text{smoke}_i + \\ &\quad \beta_{12} \text{age}_i \times \text{time}_{ij} + \beta_{13} \text{age}_i \times \text{smoke}_i + \beta_{14} \text{time}_{ij} \times \text{smoke}_i + \epsilon_{ij}, \end{aligned} \tag{5.1}$$

where $i = 1, ..., 283$, $j = 1, ..., n_i$ and n_i is the total number of measurements for subject i, in MACS data, the range of n_i is between 1 and 14 and the average number of observations per subject is 6, the random intercept b_{0i} follows normal $N(0, \sigma_{b0}^2)$ and is independent of ϵ_{ij}, the random errors ϵ_{ij} independently follows a normal distribution

$N(0, \sigma_e^2)$.

Table 5.1: Selection Results for MACS Data Using GCV with Random Intercept

Procedure	Time	Age×Time	σ_{b0}^2	σ_e^2
SMP	-0.02(0.03)	-0.001(0.001)	0.13	0.11
ITER	-0.02(0.03)	-0.001(0.001)	0.14	0.14

Table 5.2: Selection Results for MACS Data Using REML with Random Intercept

Procedure	Time	Age×Time	σ_{b0}^2	σ_e^2
SMP	-0.08(0.03)	-0.001(0.001)	0.13	0.11
ITER	-0.08(0.03)	-0.001(0.001)	0.14	0.11

Table 5.3: Selection Results for MACS Data Using BIC_{scad} with Random Intercept

Procedure	Time	Age×Time	σ_{b0}^2	σ_e^2
SMP	-0.03(0.03)	-0.001(0.001)	0.13	0.12
ITER	-0.03(0.03)	-0.001(0.001)	0.14	0.11

Tables (5.1), (5.2) and (5.3) show results from fitting model (5.1) by the proposed SCAD penalized likelihood method using the tuning criterion GCV, REML and BIC_{SCAD} respectively. All tuning approaches select the optimal model with the tuning parameter $\lambda = 0.125$. Variables "Time" and "Age×Time" are selected by both the simple and the iterative procedure under GCV, REML and BIC_{SCAD}. This result suggests that, the variation of CD4 counts is associated with factors time, baseline age and their interactions. All other factors including baseline CD4, the smoking status, certain pairwise interaction and quadratic effects are not significant and not selected. The time by age interaction is interesting, it appears that, the rate of CD4 counts change over time is lower for the younger subjects than that for the older ones. Overall, the regression results indicate the average of $\log \mathrm{CD4}$ decreases over years at rate 0.02 per year by GCV, 0.08 by REML and 0.03 by $\mathrm{BIC}_{\mathrm{SCAD}}$. The estimated variance component for the among-subject random effect is around 0.13, which is slightly larger than the estimated variance for the within-subject error.

Chapter 6

Discussion

In this study, we have presented a detailed investigation of one new method for variable selection in linear mixed models. We propose two procedures, the simple procedure and the iterative procedure. When estimating variance components, the simple procedure updates the variance components once while the iterative procedure update them iteratively. We maximize the penalized likelihood with SCAD penalty on the fixed effect coefficients $\boldsymbol{\beta}$ to select important variables and estimate mean structure, and use the REML approach to estimate variance components $\boldsymbol{\theta}$.

We have derived two criteria for selecting the tuning parameter λ: the GCV score and the REML score. Under certain common regularity conditions, we have demonstrated that the regression coefficient estimator from the simple procedure possesses oracle properties. Simulation studies indicate the iterative procedure estimator possesses similar property. Rates of convergence of this procedure is derived. Supported by the statistical theory, estimators from our procedures have good statistical properties such as consistency and asymptotic normality. Simulation results show that our procedures are effective and the standard errors are estimated with good accuracy.

We demonstrate that standard information criteria may be used to choosing the tuning parameter λ, the combination of SCAD penalized likelihood with information criteria has worked well to select significant covariates.

Compared with the classical best subset variable selection methods, the proposed procedures are more efficient. Our method is faster to implement when the number of candidate variables are large. For a data set with 200 subjects and 5 longitudinal observations for each subject, the computing time for 10 candidate covariates using simple and iterative procedure are 25 seconds and 40 seconds respectively, including estimation and calculation of standard errors. However, since we need to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ iteratively, good initial values and good convergency rules are critical for our procedures. The selection of tuning parameters λ and a depends on the grid search which is influenced by the grid settings, and is also determined by the tuning criteria. The algorithm includes some high dimensional matrix operation when the number of observations are large, some approximation needs to be applied.

For the method studied, normality of within-subject longitudinal data measurement errors, which is often reasonable, is assumed, so the analyst needs to explore the relevance of this assumption.

Future investigation includes extending the proposed procedures to the semiparametric regression models when the assumptions on the parametric forms are relaxed, and the generalized linear mixed models when the observed longitudinal data are non-normal. A theoretical basis for oracle property under these extended assumptions may exist, which may need further research.

Appendix A

Tables of Simulation Results

Table A.1: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept and Random Slope Using BIC_{scad}, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.99(0.15)	0.11	1.51(0.12)	0.10
100	SMP	1.98(0.14)	0.14	1.51(0.11)	0.10
100	ITER	1.99(0.14)	0.14	1.50(0.11)	0.11
200	FL	2.01(0.10)	0.08	1.49(0.09)	0.07
200	SMP	2.00(0.09)	0.08	1.50(0.08)	0.08
200	ITER	2.00(0.09)	0.08	1.50(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.01 (0.13)	0.10	0.51(0.16)	0.12
100	SMP	1.00 (0.11)	0.10	0.51(0.15)	0.14
100	ITER	1.00 (0.10)	0.09	0.51(0.15)	0.14
200	FL	1.00(0.08)	0.06	0.51(0.13)	0.11
200	SMP	0.99(0.07)	0.07	0.51(0.11)	0.10
200	ITER	1.00(0.07)	0.07	0.51(0.11)	0.12

Table A.2: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept and Random Slope Using BIC_{scad}, the true value of $(\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2) = (0.64, 0.12, 0.36, 2)$, $\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.67(0.23)	0.23	0.15(0.23)	0.24
100	ITER	0.67(0.22)	0.23	0.15(0.23)	0.25
200	SMP	0.66(0.19)	0.17	0.14 (0.22)	0.19
200	ITER	0.67(0.18)	0.17	0.14(0.21)	0.18
size(m)	Procedure	$\sigma_{b1}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.46(0.29)	0.30	2.01(0.14)	0.13
100	ITER	0.46(0.29)	0.31	2.01(0.14)	0.14
200	SMP	0.40(0.25)	0.23	2.00(0.10)	0.09
200	ITER	0.39(0.24)	0.23	2.00(0.10)	0.10

Table A.3: Selection Results in Setting 2 with Random Intercept and Random Slope Using GCV

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.47	114.2	3.98	0	5.02
100	SMP	0.39	69.1	4.44	0	4.56
100	ITER	0.27	29.2	4.69	0	4.31
200	FL	0.28	139.6	4.03	0	4.97
200	SMP	0.23	94.9	4.51	0	4.49
200	ITER	0.16	58.3	4.80	0	4.20

Table A.4: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept and Random Slope Using GCV, the true value of $(\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2) = (0.64, 0.12, 0.36, 2)$, $\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.69(0.32)	0.35	0.10(0.29)	0.33
100	ITER	0.68(0.31)	0.33	0.09(0.27)	0.29
200	SMP	0.68(0.22)	0.23	0.10(0.23)	0.26
200	ITER	0.67(0.22)	0.24	0.10(0.23)	0.27
size(m)	Procedure	$\sigma_{b1}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.46(0.37)	0.34	1.98(0.17)	0.18
100	ITER	0.45(0.38)	0.40	1.98(0.17)	0.18
200	SMP	0.45(0.31)	0.31	2.01(0.11)	0.13
200	ITER	0.40(0.30)	0.30	2.01(0.11)	0.12

Table A.5: Selection Results in Setting 2 with Random Intercept and Random Slope Using REML

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	SMP	0.38	64.5	4.62	0	4.38
100	ITER	0.35	61.6	4.76	0	4.24
200	SMP	0.16	63.9	4.72	0	4.28
200	ITER	0.15	54.6	4.88	0	4.12

Table A.6: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept and Random Slope Using REML, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	SMP	0.99(0.15)	0.15	1.00(0.13)	0.11
100	ITER	0.99(0.15)	0.14	1.00(0.13)	0.11
200	SMP	1.00(0.10)	0.10	0.99(0.10)	0.09
200	ITER	1.00(0.10)	0.10	0.99(0.10)	0.09
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	SMP	0.97(0.13)	0.11	1.00(0.14)	0.15
100	ITER	0.98(0.13)	0.11	0.99(0.14)	0.14
200	SMP	1.01(0.08)	0.08	0.99(0.10)	0.10
200	ITER	1.00(0.08)	0.08	0.99(0.10)	0.10

Table A.7: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept and Random Slope Using REML, the true value of $(\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2) = (0.64, 0.12, 0.36, 2)$, $\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.68(0.29)	0.30	0.10(0.28)	0.29
100	ITER	0.68(0.30)	0.31	0.10(0.28)	0.30
200	SMP	0.67(0.23)	0.25	0.14(0.26)	0.28
200	ITER	0.66(0.23)	0.23	0.14(0.25)	0.27
size(m)	Procedure	$\sigma_{b1}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.43(0.30)	0.32	1.97(0.14)	0.17
100	ITER	0.43(0.30)	0.31	1.97(0.13)	0.15
200	SMP	0.35(0.27)	0.29	2.00(0.10)	0.11
200	ITER	0.35(0.27)	0.28	2.00(0.10)	0.11

Table A.8: Selection Results in Setting 2 with Random Intercept and Random Slope Using BIC_{scad}

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.40	82.1	4.73	0	4.27
100	SMP	0.38	65.2	4.74	0	4.26
100	ITER	0.30	40.1	4.74	0	4.26
200	FL	0.26	108.5	4.89	0	4.11
200	SMP	0.20	35.4	4.91	0	4.09
200	ITER	0.19	29	4.92	0	4.08

Table A.9: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept and Random Slope Using BIC_{scad}, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.01(0.16)	0.12	1.01(0.14)	0.13
100	SMP	0.99(0.15)	0.14	1.00(0.13)	0.12
100	ITER	1.00(0.15)	0.14	1.00(0.13)	0.12
200	FL	1.00(0.11)	0.10	1.00 (0.10)	0.08
200	SMP	1.00(0.10)	0.09	1.00 (0.09)	0.09
200	ITER	1.00(0.10)	0.10	1.00 (0.09)	0.09
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	0.98(0.16)	0.12	1.01(0.14)	0.11
100	SMP	0.99(0.13)	0.12	1.01(0.14)	0.13
100	ITER	1.00(0.13)	0.12	1.01(0.14)	0.13
200	FL	1.00(0.09)	0.07	1.00(0.10)	0.08
200	SMP	1.01(0.08)	0.07	1.00(0.08)	0.08
200	ITER	1.01(0.08)	0.08	1.00(0.08)	0.08

Table A.10: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept and Random Slope Using BIC_{scad}, the true value of $(\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2) = (0.64, 0.12, 0.36, 2)$, $\sigma_{b0}^2, \sigma_{b01}, \sigma_{b1}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_{b01}(SD)$	SD_m
100	SMP	0.68(0.29)	0.28	0.14(0.28)	0.29
100	ITER	0.68(0.28)	0.28	0.14(0.28)	0.28
200	SMP	0.66(0.21)	0.20	0.10(0.25)	0.24
200	ITER	0.66(0.21)	0.20	0.10(0.25)	0.24
size(m)	Procedure	$\sigma_{b1}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.39(0.31)	0.29	1.99(0.15)	0.16
100	ITER	0.39(0.31)	0.30	1.99(0.14)	0.15
200	SMP	0.40(0.29)	0.28	2.00(0.11)	0.10
200	ITER	0.40(0.29)	0.29	2.00(0.11)	0.10

Table A.11: Selection Results in Setting 1 with Random Intercept Using GCV

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.39	149.2	3.86	0	5.14
100	SMP	0.36	82.3	4.43	0	4.57
100	ITER	0.30	63.7	4.53	0	4.47
200	FL	0.21	122.5	4.01	0	4.99
200	SMP	0.19	79.2	4.62	0	4.38
200	ITER	0.14	37.3	4.75	0	4.25

Table A.12: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept Using GCV, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.99(0.12)	0.09	1.51(0.12)	0.10
100	SMP	1.99(0.11)	0.12	1.51(0.11)	0.11
100	ITER	1.99(0.11)	0.12	1.51(0.11)	0.11
200	FL	1.99(0.08)	0.06	1.49(0.08)	0.05
200	SMP	1.99(0.08)	0.09	1.50(0.07)	0.08
200	ITER	1.99(0.08)	0.09	1.50(0.07)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	0.97(0.12)	0.10	0.53(0.12)	0.09
100	SMP	0.98(0.11)	0.11	0.53(0.12)	0.12
100	ITER	0.98(0.11)	0.11	0.53(0.12)	0.12
200	FL	1.00(0.09)	0.06	0.51(0.11)	0.07
200	SMP	1.00(0.08)	0.08	0.51(0.11)	0.09
200	ITER	1.00(0.08)	0.07	0.51(0.11)	0.09

Table A.13: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept Using GCV, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.63(0.18)	0.16	2.01(0.15)	0.14
100	ITER	0.63(0.17)	0.15	2.01(0.15)	0.14
200	SMP	0.64(0.10)	0.11	2.00(0.10)	0.10
200	ITER	0.64(0.10)	0.10	2.00(0.10)	0.10

Table A.14: Selection Results in Setting 1 with Random Intercept Using BIC_{scad}

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.37	91.3	4.55	0	4.45
100	SMP	0.27	61.5	4.61	0	4.39
100	ITER	0.25	43.2	4.62	0	4.38
200	FL	0.20	96.4	4.79	0	4.21
200	SMP	0.16	50.1	4.81	0	4.19
200	ITER	0.15	23.4	4.82	0	4.18

Table A.15: Estimation of $\boldsymbol{\beta}$ in Setting 1 with Random Intercept Using BIC_{scad}, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.99(0.13)	0.11	1.50(0.12)	0.10
100	SMP	1.98(0.12)	0.11	1.49(0.11)	0.11
100	ITER	1.99(0.12)	0.11	1.49(0.11)	0.11
200	FL	1.99(0.08)	0.07	1.50(0.08)	0.07
200	SMP	1.99(0.08)	0.08	1.50(0.07)	0.07
200	ITER	1.99(0.08)	0.08	1.50(0.07)	0.07
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.00(0.12)	0.10	0.51(0.12)	0.10
100	SMP	1.00(0.11)	0.11	0.51(0.12)	0.11
100	ITER	1.00(0.11)	0.10	0.51(0.12)	0.11
100	FL	1.01(0.09)	0.08	0.51(0.10)	0.08
200	SMP	1.00(0.08)	0.07	0.51(0.10)	0.09
200	ITER	1.00(0.08)	0.07	0.51(0.10)	0.10

Table A.16: Estimation of $\boldsymbol{\theta}$ in Setting 1 with Random Intercept Using BIC_{scad}, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.63(0.17)	0.16	1.99(0.14)	0.13
100	ITER	0.63(0.17)	0.16	1.99(0.14)	0.13
200	SMP	0.64(0.10)	0.11	2.00(0.09)	0.09
200	ITER	0.64(0.10)	0.10	2.00(0.09)	0.09

Table A.17: Selection Results in Setting 2 with Random Intercept Using GCV

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.50	102.6	3.99	0	5.01
100	SMP	0.43	87.2	4.46	0	4.54
100	ITER	0.28	69.6	4.73	0	4.27
200	FL	0.19	73.0	4.03	0	4.97
200	SMP	0.16	67.5	4.53	0	4.47
200	ITER	0.11	34.1	4.85	0	4.15

Table A.18: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept Using GCV, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	0.99(0.13)	0.09	1.02(0.12)	0.08
100	SMP	0.99(0.13)	0.12	1.02(0.11)	0.11
100	ITER	0.99(0.13)	0.12	1.02(0.11)	0.11
200	FL	0.99(0.09)	0.06	1.00(0.09)	0.06
200	SMP	0.99(0.09)	0.08	1.00(0.07)	0.08
200	ITER	0.99(0.09)	0.08	1.00(0.07)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.01(0.13)	0.08	1.01(0.15)	0.09
100	SMP	1.00(0.12)	0.11	1.01(0.15)	0.13
100	ITER	1.00(0.12)	0.11	1.01(0.14)	0.12
200	FL	1.00(0.08)	0.06	1.00(0.10)	0.07
200	SMP	1.00(0.08)	0.08	1.00(0.10)	0.08
200	ITER	1.00(0.08)	0.08	1.01(0.93)	0.08

Table A.19: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept Using GCV, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.63(0.14)	0.15	2.00(0.15)	0.14
100	ITER	0.63(0.14)	0.15	2.00(0.15)	0.14
200	SMP	0.65(0.11)	0.11	2.01(0.12)	0.10
200	ITER	0.65(0.10)	0.11	2.01(0.12)	0.10

Table A.20: Selection Results in Setting 2 with Random Intercept Using REML

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	SMP	0.38	64.5	4.62	0	4.38
100	ITER	0.35	61.6	4.76	0	4.24
200	SMP	0.16	63.9	4.72	0	4.28
200	ITER	0.15	54.6	4.88	0	4.12

Table A.21: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept Using REML, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	SMP	1.00(0.13)	0.12	0.99(0.12)	0.11
100	ITER	1.00(0.13)	0.12	0.99(0.13)	0.11
200	SMP	0.99(0.08)	0.08	0.99(0.09)	0.07
200	ITER	0.99(0.08)	0.08	0.99(0.08)	0.07
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	SMP	1.00(0.12)	0.11	1.01(0.11)	0.12
100	ITER	1.00(0.12)	0.11	1.01(0.11)	0.12
200	SMP	0.99(0.11)	0.09	1.00(0.08)	0.08
200	ITER	0.99(0.11)	0.08	1.00(0.08)	0.08

Table A.22: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept Using REML, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.64(0.14)	0.15	2.01(0.14)	0.14
100	ITER	0.63(0.13)	0.14	2.01(0.14)	0.14
200	SMP	0.64(0.11)	0.11	1.99(0.09)	0.09
200	ITER	0.64(0.10)	0.10	1.99(0.09)	0.09

Table A.23: Selection Results in Setting 2 with Random Intercept Using BIC_{scad}

size(m)#	Procedure	ME	$MRME(\%)$	Correct 0	Incorrect 0	average size
100	FL	0.31	102.3	4.64	0	4.36
100	SMP	0.24	59.1	4.66	0	4.34
100	ITER	0.22	52.1	4.67	0	4.33
200	FL	0.18	69.3	4.89	0	4.11
200	SMP	0.14	48.6	4.89	0	4.11
200	ITER	0.14	47.7	4.89	0	4.11

Table A.24: Estimation of $\boldsymbol{\beta}$ in Setting 2 with Random Intercept Using BIC_{scad}, β_1-β_4, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\beta_1(SD)$	SD_m	$\beta_2(SD)$	SD_m
100	FL	1.01(0.13)	0.10	0.99(0.15)	0.12
100	SMP	1.01(0.12)	0.12	1.00(0.13)	0.13
100	ITER	1.00(0.11)	0.12	1.00(0.13)	0.13
200	FL	1.01(0.11)	0.10	1.00(0.09)	0.07
200	SMP	0.99(0.10)	0.10	1.00(0.08)	0.09
200	ITER	0.99(0.10)	0.09	1.00(0.08)	0.08
size(m)	Procedure	$\beta_3(SD)$	SD_m	$\beta_4(SD)$	SD_m
100	FL	1.00(0.11)	0.09	1.01(0.12)	0.09
100	SMP	1.01(0.10)	0.11	0.99(0.11)	0.10
100	ITER	1.00(0.10)	0.10	0.99(0.11)	0.10
100	FL	1.00(0.09)	0.07	1.00(0.09)	0.07
200	SMP	1.00(0.08)	0.08	1.00(0.08)	0.08
200	ITER	1.00(0.08)	0.08	1.00(0.08)	0.07

Table A.25: Estimation of $\boldsymbol{\theta}$ in Setting 2 with Random Intercept Using BIC_{scad}, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, the true value of $(\sigma_{b0}^2, \sigma_e^2) = (0.64, 2)$, $\sigma_{b0}^2, \sigma_e^2$, the Monte Carlo sample mean; SD, the Monte Carlo median absolute deviation divided by 0.6745; SD_m, the median of the 100 estimated SD's.

size(m)	Procedure	$\sigma_{b0}^2(SD)$	SD_m	$\sigma_e^2(SD)$	SD_m
100	SMP	0.65(0.13)	0.12	1.01(0.13)	0.13
100	ITER	0.64(0.13)	0.12	1.01(0.13)	0.13
200	SMP	0.64(0.11)	0.10	1.00(0.09)	0.09
200	ITER	0.64(0.11)	0.10	1.00(0.08)	0.08

Bibliography

Akaike (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255-265.

Berger, J. O. and Pericchi, L. R. (1996a), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109-122.

—— (1996b), "The Intrinsic Bayes Factor for Linear Models," *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 25-44.

—— (2001), "Objective Bayesian Methods for Model Selection: Introduction and Com- parison (with discussion)," *Institute of Mathematical Statistics Lecture Notes - Mono- graph Serires (Lahiri, P., ed.)*, vol. 38, 135-207.

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-xed Prediction Error," *Journal of the American Statistical Association,* 87, 738-754.

Burman, P. (1989), “A Comparative Study of Ordinary Cross-Validation, V-fold Cross-Validation and the Repeated Learning-testing Methods,” *Biometrika*, 76, 503-514.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), “Multiple Shrinkage and Subset Selection in Wavelets,” *Biometrika*, 85, 391-402.

Donoho, D. L. and Johnstone, I. M. (1994), “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, 81 425-455.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), “Least Angle Regression,” *The Annals of Statistics*, 32, No. 2, 407-499.

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

—— (2004), “New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis,” *Journal of American Statistical Association*, 99, 710-723.

Fan, J., and Zhang, J. (2000), “Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data,” *Journal of the Royal Statistical Society, Ser. B*, 62, 303-322.

Frank, I. E. and Friedman, J. H. (1993), “A statistical view of some chemometric regression tools (with discussion).” *Technometrics* 35 109-148.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881-889.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.

Hoerl, A. E. and Kennard, R. W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

—— (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 69-82.

Huang, J. Z.,Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111-128.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987), "The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310-318.

Knight, K. and Fu, W. J. (2000), "Asymptotics for Lasso-type estimators," *The Annal of Statistics*, 28 1356-1378.

Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression (with discussion) ," *Journal of the Royal Statistical Society, Series B*, 30, 31-66.

Luo, X., Stefanski, L. A., and Boos, D. D. (2005), "Tuning Variable Selection Procedures," to appear in *Technometrics*.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661-675.

Miller, A. J. (2002), *Subset Selection in Regression*, London: Chapman & Hall.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression (with discussion)," *Journal of the American Statistical Association*, 83, 1023-1036.

Patterson, H.D. and Thompson, R. (1971), "Recovery of inter-block information when block sizes are unequal," *Biometrika*, 58, 545-554.

Picard, R. R. and R. D. Cook (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*. 79, 575-583.

Thompson, M. L. (1978), "Selection of Variables in Multiple Regression," *International Statistical Review*, 46, 1-19 and 129-146.

Rao, C. R. (1973), "Linear Statistical Inference and Its Applications," 2nd ed., Wiley, New York.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267-288. —— (1997), "The Lasso method for variable selection in the Cox model," *Statistics in Medicine* 16 385-395.

Wahba, G. (1990), "Spline Models for Observational Data," Philadelphia: Society for

Industrial and Applied Mathematics.

Wang, Y. (1998), "Smoothing spline models with correlated random errors," *Journal of the American Statistical Association*, 93, 341-348.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2005), "Controlling Variable Selection by the Addition of Pseudo Variables," to appear in *Journal of the American Statistical Association.*

Zhang, P. (1993), "Model Selection via Multifold Cross-Validation," *The Annal of Statistics*, 21, 299-313.