Adaptive Model Selection in Linear Mixed Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Bo Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Xiaotong Shen, Adviser

August 2009

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Xiaotong Shen, for his guidance, encouragement and patience through the development of this dissertation. His profound understanding of numerous areas has been a great resource for me and has made my Ph.D. research such a pleasant journey. He is not only an academic advisor, but also a mentor, a friend and inspiration. Professor Xiaotong Shen, being passionate, energetic and intellectual, has set up a solid professional model that I will be pursuing.

I am thankful to Professor Glen Meeden for his serving as my defense committee chair and reviewing my thesis. My thanks also go to Professor Galin Jones and Professor Xianghua Luo for their time and effort for reviewing my thesis, and for their valuable suggestion regarding my research. I am grateful to the faculty, staff and students in the School of Statistics for making my five-year study at the University of Minnesota such a wonderful experience.

Abstract

Linear mixed models are commonly used models in the analysis of correlated data, in which the observed data are grouped according to one or more clustering factors. The selection of covariates, the variance structure and the correlation structure is crucial to the accuracy of both estimation and prediction in linear mixed models. Information criteria such as Akaike's information criterion, Bayesian information criterion, and the risk inflation criterion are mostly applied to select linear mixed models. Most information criteria penalize an increase in the size of a model through a fixed penalization parameter. In this dissertation, we firstly derive the generalized degrees of freedom for linear mixed models. A resampling technique, data perturbation, is employed to estimate the generalized degrees of freedom of linear mixed models. Further, based upon the generalized degrees of freedom of linear mixed models, we develop an adaptive model selection procedure with a data-adaptive model complexity penalty for selecting linear mixed models. The asymptotic optimality of the adaptive model selection procedure in linear mixed models is shown over a class of information criteria. The performance

of the adaptive model selection procedure in linear mixed models is studied by numerical simulations. Simulation results show that the adaptive model selection procedure outperforms information criteria such as Akaike's information criterion and Bayesian information criterion in selecting covariates, the variance structure and the correlation structure in linear mixed models. Finally, an application to diabetic retinopathy is examined.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Model selection is one key step in statistical modeling. A fundamental question, given a data set, is to choose the approximately "best" model from a class of competing candidate models. The issue becomes more complex in the case of linear mixed models, where selecting the "best" model means not only to select the best mean structure but also the most optimal variance-covariance structure. For this purpose, a suitable model selection criterion is needed. This dissertation focuses on model selection in linear mixed models. Shen and Ye (2002), Shen, Huang and Ye (2004), and Shen and Huang (2006) proposed a novel model selection procedure, called adaptive model selection, derived from information criteria and based on the generalized degrees of freedom. Here the adaptive model selection is discussed in linear mixed models. The discussion starts with introducing linear mixed models and reviewing a class of information criteria for

selecting linear mixed models.

1.1 Linear Mixed Models

1.1.1 Introduction to Linear Mixed Models

Many common statistical models can be expressed as models that incorporate two parts: one part is fixed effects, which are parameters associated with the entire population or certain repeatable levels of experimental factors; another part is random effects, which are associated with experimental subjects or units randomly drawn or observed from the population. The models with both fixed effects and random effects are called mixed models, or mixed-effects models, or random effect models, or sometimes variance components models. In this dissertation, we will call them *mixed models*.

Mixed models are primarily used to describe relationships between a one-dimensional or multidimensional response variable and some possibly related covariates in the observed data that are grouped according to one or more clustering factors. Such data include longitudinal data (or panel data in econometrics), repeated measurement data, multilevel/hierarchical data, and block design data. In such data, observations that are grouped into one cluster are correlated for a reason. Therefore, this type of data is called *correlated data* or *grouped data*. By associating common random effects to observations sharing the same level of a classification factor or same investigated subject, mixed models flexibly represent the covariance structure induced by the grouping of correlated

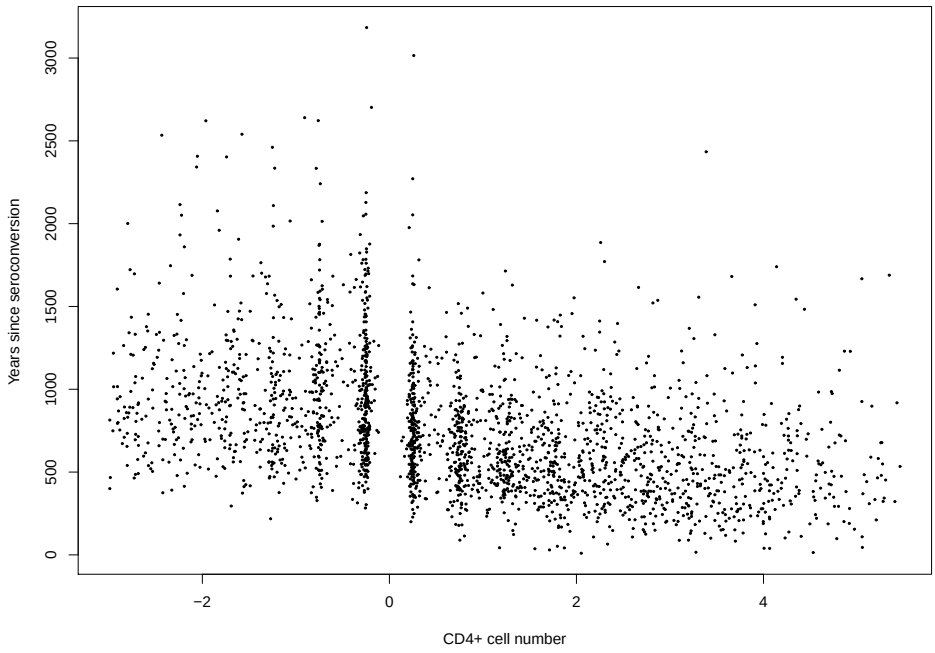

Figure 1.1: Scatterplot showing relationship between time since HIV patients' seroconversio due to infection with the HIV virus and patients' CD4+ cell numbers.

data. For example, Figure 1.1 displays 2376 values of CD4+ cell number plotted against time since seroconversion (seroconversion is the time when human immune deficiency virus (HIV) becomes detectable) for 369 infected men enrolled in the Multicenter Acquired Immune Deficiency Syndrome (AIDS) Cohort Study (Kaslow et al., 1987). In the data set, repeated measurements for some patients are connected to accentuate the longitudinal nature of the study. The main objective of analyzing the data is to capture the time course of CD4+ cell depletion, which is caused by the HIV infection. Analyzing the data will help to clarify the interaction of HIV with the immune system and can

assist when counseling infected men. Diggle et al. (2002) suggested that linear mixed models was one of the options for modeling the progression of mean CD4+ counts as a function of time since seroconversion.

For a single level of grouping in correlated data, the *linear mixed models* specify the n_i-dimensional response vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_j})^T$ for the ith subject, $i = 1, 2, \cdots, m$, with total number of observation $\sum_{i=1}^{m} n_i$, as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \cdots, m, \tag{1.1}$$

$$\boldsymbol{b}_i \sim N(0, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{\Lambda}_i),$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects, the $\boldsymbol{b}_i$'s are the q-dimensional vector of random effects, $\boldsymbol{X}_i$'s are $n_i \times p$ fixed-effects regressor matrices, $\boldsymbol{Z}_i$'s are $n_i \times q$ random-effects regressor matrices, and $\boldsymbol{\epsilon}_i$ are the n_i-dimensional within-group error vectors, $\boldsymbol{\Psi}$ is a positive-definite symmetric matrix denoting the variance-covariance structure between random effects $\boldsymbol{b}_i$, $\boldsymbol{\Lambda}_i$ are positive-definite matrices denoting the variance-covariance structure of heteroscedastic and dependent within-group errors. The random effects $\boldsymbol{b}_i$'s and the within-group errors $\boldsymbol{\epsilon}_i$'s are assumed to be independent for different groups and to be independent of each other with the same group. In linear mixed models, the Gaussian continuous response is assumed to be a linear function of covariates with regression coefficients that vary over individuals, which reflects natural heterogeneity

due to unmeasured factors.

Linear mixed models with flexible variance-covariance structure of within-group errors provide a flexible and powerful tool for the analysis of correlated data, which arise in many areas such as agriculture, biology, economics, manufacturing, and geophysics. Usually, when linear mixed models are fitted to data, a number of covariates are involved and there are also more than one possible variance-covariance structures for random effects and within-group errors. Therefore, model selection is required prior to model fitting. This dissertation is devoted to model selection for linear mixed models.

1.1.2 Linear Mixed Models in Longitudinal Data Analysis

A *longitudinal study* is an observational study that involves repeated observations of the same items over periods of time. Longitudinal studies are applied in a variety of fields. In medicine, longitudinal studies are used to uncover predictors of certain diseases. In advertising, longitudinal studies are used to identify the changes that advertising has produced in the attitudes and behaviors of those within the target audience who have seen the advertising campaign. Longitudinal studies are also used in psychology to study developmental trends across the life span, and in sociology to study life events throughout lifetimes or generations. Data collected from a longitudinal study is *longitudinal data*. In the analysis of longitudinal data, linear mixed models are widely used (Diggle et al., 2002). Repeated measurements in longitudinal data are observed

on subjects across *time*. In addition to the *time variable*, other covariates are typically observed, which in turn requires more accurate selection for seeking the best model from a large pool of candidate models. In the analysis of longitudinal data by linear mixed models, model selection also involves both covariate selection and the variance-covariance structure selection of the random effects and the within-group errors.

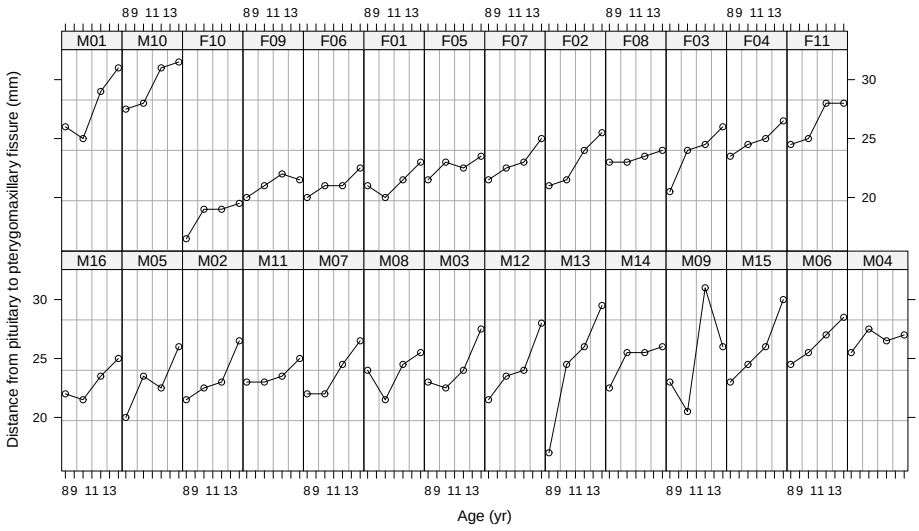

Figure 1.2: Distance from the pituitary to the pterygomaxillary fissure versus age for a sample of 16 boys (subjects M01 to M16) and 11 girls (subjects F01 to F11).

Linear mixed models (1.1) are generally specified. However, in the analysis of longitudinal data, the linear mixed models either with only random intercepts or with random intercepts and random slopes for time variable are frequently considered. For example,

a set of longitudinal data collected by orthodontists from x-rays of children's skulls is a set of measurements of the distance from the pituitary gland to the pterygomaxillary fissure taken every two years from eight years of age until fourteen years of age on a sample of 27 children of 16 males and 11 females (Potthoff and Roy, 1964, and Pinheiro and Bates, 2000), in which *age* is time-varying. The distances from the pituitary gland to the pterygomaxillary fissure of each subject vary linearly with either arbitrary intercept plus common slope or arbitrary intercept plus arbitrary slope of *age* (see Figure 1.2). In this study, besides the time variable, there are also some clinical variables, such as gender, race, physical exercise frequency, athletic condition, etc.. For the jth observation of the ith subject, let t_{ij} be the time variable *age*, let Y_{ij} be the distances from the pituitary gland to the pterygomaxillary fissure, and $\boldsymbol{x} = (x_{ij1}, x_{ij2}, \cdots, x_{ijq})^T$ represents the clinical variables, a linear mixed model with random intercept and random slope is appropriate for the data:

$$Y_{ij} = \beta_0 + b_{i0} + t_{ij}\beta_0 + t_{ij}b_{i1} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}, \tag{1.2}$$

with $b_{i0} \overset{iid}{\sim} N(0, \sigma_0^2)$, $b_{i1} \overset{iid}{\sim} N(0, \sigma_1^2)$ and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$; $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n_i$. The clinical variables are included in the study, but may or may not have influence on the distances from the pituitary gland to the pterygomaxillary fissure. Therefore, it is of importance to do covariate selection among the clinical variables when we fit a linear mixed model to the data.

1.2 Model Selection in Linear Mixed Models

Model selection is seen almost everywhere in statistical modeling. There is no exemption for modeling correlated data with linear mixed models. Let $\{M_\gamma, \gamma \in \Gamma\}$ be a class of candidate linear mixed models for grouped observations, where Γ is an index set. The true model may or may not be included in the class of candidate models $\{M_\gamma, \gamma \in \Gamma\}$. Given the data, a model is selected from $\{M_\gamma, \gamma \in \Gamma\}$ through a suitable model selection criterion, to better describe the underlying data. There are mainly three aspects of model selection in linear mixed models: (1) selecting appropriate covariates with nonzero regression coefficients in the mean structure of linear mixed models; (2) selecting the appropriate variance structure for modeling heteroscedasticity of within-group errors and random effect components of linear mixed models; (3) selecting the appropriate correlation structure for modeling dependence of within-group errors and random effect components of linear mixed models.

1.2.1 Information Criteria for Model Selection

Many model selection procedures for linear mixed models have been proposed in the literature. The most popular procedures are *information criteria*. They can be applied in selecting in linear mixed models. Most information criteria choose the optimal model $\widehat{M}$ among candidate models $\{M_\gamma, \gamma \in \Gamma\}$ by minimizing a model selection criterion of

the form

$$-2\ell_{M_\gamma} + \lambda(n, k_{M_\gamma}) \cdot k_{M_\gamma}, \tag{1.3}$$

where n denotes the number of observations, k_{M_γ} is the total number of independent parameters in a candidate model M_γ, ℓ_{M_γ} is the maximum log-likelihood given model M_γ. In (1.3), the total number of parameters k_{M_γ} in the candidate model M_γ is multiplied by $\lambda(n, k_{M_\gamma})$, a function of n and k_{M_γ} called a penalization parameter, to penalize an increase size of the candidate model. Information criteria, which were motivated by various principles, differ only in the value of penalization parameter $\lambda(n, k_{M_\gamma})$ in (1.3): Akaike's information criterion (AIC), (Akaike, 1973) used the expected Kullback-Leibler information with $\lambda(n, k_{M_\gamma}) = 1$; Hurvich and Tsai (1989) derived a bias-corrected version of AIC, called AICc, with $\lambda(n, k_{M_\gamma}) = nk_{M_\gamma}/(n - k_{M_\gamma} - 1)$, by estimating the expected Kullback-Leibler information directly in a regression model where a second order bias adjustment was made; Bayesian information criterion (BIC), (Schwarz, 1978) used an asymptotic Bayes factor and advocated $\lambda(n, k_{M_\gamma}) = \log(n)/2$; the risk inflation criterion (RIC), (George and Foster, 1994) was based on the minimax principle, and adjusted the penalization parameter to be $\lambda(n, k_{M_\gamma}) = \log(p)$, where p is the number of available covariates; and the covariance inflation criterion (CIC), (Tibshirani and Knight, 1999) with $\lambda = 2\sum_{l=1}^{k_{M_\gamma}} \log(n/l)/k_{M_\gamma}$ in (1.3) adjusts the training error by the average covariance of the predictions and responses when the prediction rule is applied to permuted versions of the data set, and many others.

1.2.2 Adaptive Model Selection

In information criterion (1.3), the penalization parameter $\lambda(n, k_{M_\gamma})$ penalizes an increase in the size of a model as a fixed penalization parameter in the sense that it is pre-determined by the sample size n and the number of independent parameters k_{M_γ}, and it is not adaptive. The model selection procedures with the form of (1.3) are hereby referred as "nonadaptive" selection procedures. Shen and Ye (1998) have showed in linear models that these nonadaptive model selection procedures perform well only in one type of situation: some of them, such as the BIC, with a large penalty has been proven to be effective when the number of parameters k_{M_γ} in the true model is small, but yield large selection bias when k_{M_γ} in the true model is large; and the AIC does just the opposite. Shen, Huang and Ye (2004) confirmed this disadvantages of nonadaptive model selection procedures in logistic regression and poisson regression.

The need for an adaptive model selection procedure that reduces the selection bias and essentially performs well uniformly across a variety of situations is compelling. As the solution to nonadaptive problem of information criteria with form (1.3), Shen and Ye (2002) proposed the generalized degrees of freedom and used it to derive a data-adaptive model selection procedure for linear models. Shen, Huang and Ye (2004) proposed a similar procedure for exponential family distributions with no or only one dispersion parameter. In their work, the adaptive model selection procedures with data-adaptive penalization parameter $\hat{\lambda}$ were theoretically and numerically proven to be optimal. Shen

and Huang (2006) generalized the concept of the generalized degrees of freedom and the adaptive model selection criteria of Shen, Huang and Ye (2004) from the Kullback-Leibler loss to a more general class of loss functions.

The present work discusses the generalized degrees of freedom in linear mixed models, and develops an adaptive model selection procedure for linear mixed models. The proposed adaptive selection procedure overcomes the shortcoming of fixed penalty model selection procedures that are widely used in selection among linear mixed models. It leads to large penalization parameter in (1.3) when the true model has a parsimonious representation, and leads to small penalization parameter when the true model is large. It approximates the best performance of nonadaptive alternatives that can be written in the form of (1.3). In summary, an adaptive model selection procedure offers an uniformly better solution than nonadaptive ones in linear mixed models.

1.3 Outline of the Dissertation

This dissertation is organized as follows. Chapter 2 derives the generalized degrees of freedom for linear mixed models via the Kullback-Leibler loss, and discusses the data perturbation estimation for the generalized degrees of freedom. Chapter 3 introduces adaptive model selection criteria for linear mixed models, involving an adaptive penalty estimated through the generalized degrees of freedom of linear mixed models. The asymptotic optimality of the proposed model selection criteria is also showed in Chapter

3. Chapter 4 applies the adaptive model selection procedure to covariate selection, variance structure selection, and correlation structure selection in linear mixed models. The simulations suggest that the proposed model selection procedure performs well against the AIC, the BIC, and the RIC across a variety of different settings. Chapter 5 presents an application of the methodology to the study of diabetic retinopathy. Chapter 6 discusses the methodology. The appendices contain technical discussions and proofs.

Chapter 2

Generalized Degrees of Freedoms of Linear Mixed Models: Concept and Estimation

This chapter discusses the concept of the generalized degrees of freedom in linear mixed-effect models and its data perturbation estimation process. Section 2.1 reviews the concept of the degrees of freedom and the effective degrees of freedom in linear regression models (Weisberg, 2005), as well as the generalized degrees of freedom proposed by Ye (1998), and Shen and Huang (2006). Then we derived the concept of generalized degrees of freedom for linear mixed-effect models in Section 2.1.3. Section 2.2 introduces parametric and nonparametric data perturbation techniques. Finally, the parametric

data perturbation is applied in Section 2.2.2, to develop the estimators of the generalized degrees of freedom in linear mixed models.

2.1 Generalized Degrees of Freedom of Linear Mixed Models

2.1.1 Degrees of Freedom and Effective Degrees of Freedom of Linear Models

In the theory of linear regression, the concept of *the degrees of freedom* (Weisberg, 2005) plays a crucial role. The degrees of freedom is the dimension of the domain that a random vector belongs to, or essentially the number of free components, i.e., how many components need to be known before a random vector is fully determined. But, the degrees of freedom is most often used in the context of linear regression and analysis of variance where certain random vectors are constrained to lie in linear subspaces, and the degrees of freedom is the dimension of the subspace. Consider a linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I})$, $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ is the response vector, $\boldsymbol{\beta}$ is a $p\times 1$ coefficient vecotr, $\boldsymbol{X}_{n\times p}$ is the design matrix with p covariates, and $\boldsymbol{I}$ is the identity matrix. Here, the least squares estimation of the mean of response vector $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ is $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y} =$

$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. In least squares estimation, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called *hat matrix*. In the classical theory, the degrees of freedom was used for the estimation of the residual variance σ^2. That is, the unbiased estimate of σ^2 is $(\boldsymbol{Y}-\hat{\boldsymbol{\mu}})^T(\boldsymbol{Y}-\hat{\boldsymbol{\mu}})/(n-p)$, where $n-p$ is the *degrees of freedom of the residuals*. On the other hand, the cost of the *degrees of freedom of fit*, or the degrees of freedom in regression, is equal to the quantity p, the number of covariates. The degrees of freedom in regression p, as a model complexity measurement, is involved in various model selection criteria such as the AIC, generalized cross-validation, and the RIC. For example, to measure how well the linear model (2.1) captures the underlying structure, we can consider the quadratic loss $(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})$. The unbiased estimator of quadratic loss $(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})$ is

$$(\boldsymbol{Y}-\hat{\boldsymbol{\mu}})^T(\boldsymbol{Y}-\hat{\boldsymbol{\mu}}) - n\sigma^2 + 2p\sigma^2,$$

which is referred to as *Akaike's information criterion* (AIC). Different linear models can then be compared based on their AIC values. The linear models with smaller AIC values are preferred. The quantity $(\boldsymbol{Y}-\hat{\boldsymbol{\mu}})^T(\boldsymbol{Y}-\hat{\boldsymbol{\mu}})$ in the AIC is the goodness-of-fit measurement, which decreases when the size of model increases. The degrees of freedom in regression p, which equals the number of covariates in the AIC, penalizes the increasing of the model size in the AIC.

Many regression methods, including ridge regression, linear smoothers and smoothing splines are not based on least squares methods. As a consequence, the degrees of

freedom in terms of dimensionality can be applied. However, the *effective degrees of freedom* (Weisberg, 2005) of fit was defined in various ways to perform goodness-of-fit tests, cross-validation and other inferential procedures which are still linear in the observations and the fitted values of the regression can be expressed in the form of $\hat{\boldsymbol{\mu}} = \boldsymbol{HY}$. Depending on a procedure, appropriate definitions of the effective degrees of freedom of fit include Trace($\boldsymbol{H}$), Trace($\boldsymbol{H}^T\boldsymbol{H}$), or even Trace($\boldsymbol{H}^T\boldsymbol{H}$)^2/Trace($\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H}$). In the case of linear regression, all these definitions reduce to the usual degrees of freedom of fit. There are corresponding definitions of the *effective degrees of freedom of the residuals*, with $\boldsymbol{H}$ replaced by $\boldsymbol{I} - \boldsymbol{H}$. For instance, the k-nearest neighbor smoother (Hastie, Tibshirani and Friedman, 2001), which is the average of the k nearest measured values to a given point. At each of the n measured points, the weight of the original value on a linear combination that makes up the predicted value is just $1/k$. Thus, the trace of the hat matrix is n/k. So, the smoother costs n/k effective degrees of freedom of fit.

2.1.2 Generalized Degrees of Freedom

Since model selection procedures are studied here, the degrees of freedom which measures a model's complexity is of our interest. The generalized degrees of freedom is generalized from the concepts of degrees of freedom of fit and effective degrees of freedom of fit.

First of all, the degrees of freedom and the effective degrees of freedom in linear model (2.1) with $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y}$ can be generalized to the response vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ without the assumption of linear model. For $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, a modeling procedure M is defined as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ that produces the vector of fitted values $\hat{\boldsymbol{\mu}}_M = (\hat{\mu}_{1,M}, \hat{\mu}_{2,M}, \cdots, \hat{\mu}_{n,M})^T$ from $\boldsymbol{Y}$. To emphasize the dependence of $\hat{\boldsymbol{\mu}}_M$ on $\boldsymbol{Y}$, we use the notation $\hat{\boldsymbol{\mu}}_M = \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}) = (\hat{\mu}_{1,M}(\boldsymbol{Y}), \hat{\mu}_{2,M}(\boldsymbol{Y}), \cdots, \hat{\mu}_{n,M}(\boldsymbol{Y}))^T$ here. Note that the alternative expression of the degrees of freedom in least squares method is

$$p = \text{Trace}(\boldsymbol{H}) = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_{i,M}(\boldsymbol{Y})}{\partial Y_i}, \tag{2.2}$$

where $\hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. It motives the definition of the generalized degrees of freedom, since the degrees of freedom in (2.2) is equal to the sum of the sensitivities of the fitted values $\hat{\mu}_i$ with respect to the response values Y_i. For the response vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and any modeling procedure M, Ye (1998) defined the *generalized degrees of freedom* in linear models as

$$\begin{aligned} GDF(M) = & \sum_{i=1}^{n} \frac{\partial E\hat{\mu}_{i,M}(\boldsymbol{Y})}{\partial Y_i} \\ = & \sum_{i=1}^{n} \frac{E\hat{\mu}_{i,M}(\boldsymbol{Y})(Y_i - \mu_i)}{\sigma^2} = \sum_{i=1}^{n} \frac{\text{cov}(\hat{\mu}_{i,M}(\boldsymbol{Y}), Y_i)}{\sigma^2}. \end{aligned} \tag{2.3}$$

In (2.3), the generalized degrees of freedom is defined to be the sum of the average

sensitivities of fitted value $\hat{\mu}_{i,\,M}(\boldsymbol{Y})$ to a small change in Y_i for a more general modeling procedure. It measures the flexibility of the procedure and allows us to think about modeling procedures in terms of their complexity and tendency to overfitting. Therefore, the generalized degrees of freedom in (2.3) will depend on the known true model and the modeling procedure. It is not necessary for the generalized degrees of freedom to be the number of parameters like in linear regression.

Secondly, the definition of the generalized degrees of freedom for a general modeling procedure is given in Shen and Huang (2006) as follows. For unknown parameter vector $\boldsymbol{\mu}$, a modeling procedure M, defined as a mapping from $\mathbb{R}^n$ to the parameter space of $\boldsymbol{\mu}$, will produce an estimator $\hat{\boldsymbol{\mu}}_M(\boldsymbol{Y})$ from observations $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^T$. Each component of $\boldsymbol{Y}$ can be either one-dimensional as in linear regression or logistic regression, or multidimensional as in longitudinal data where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n)^T$ is used so that each $\boldsymbol{Y}_i$ denotes the observations from one subject. A loss function $Q(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}))$ is introduced as the measurement of accuracy of the modeling procedure M. Loss $Q(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}))$ can not be observed. So, to assess the current modeling procedure, it is necessary to seek the estimator of loss $Q(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}))$. If there exists a goodness-of-fit measure $G(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}))$ such that

$$E[Q(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_M) - G(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y}))] = E\bigg[\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{cov}(g_{ij}(\hat{\boldsymbol{\mu}}_M(\boldsymbol{Y})), f_j(Y_i))\bigg],$$

where $n, m \in \mathbb{N}$, g_{ij}'s and f_j's are known functions, and κ is a penalization term of

overfitting, a class of loss estimator

$$G(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y})) + \kappa \tag{2.4}$$

can be considered. As shown in Shen and Huang (2006), the optimal κ that minimizes the L_2-distance $E[Q(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_M) - (G(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_M(\boldsymbol{Y})) + \kappa)]^2$ is

$$\kappa = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{cov}(g_{ij}(\hat{\boldsymbol{\mu}}_M(\boldsymbol{Y})), f_j(Y_i)), \tag{2.5}$$

which is defined as the *generalized degrees of freedom* of current modeling procedure M.

2.1.3 Generalized Degrees of Freedom of Linear Mixed Models

The generalized degrees of freedom of linear mixed models, as a framework of performing adaptive model selection procedures, has not been explicitly discussed yet. In this section, the generalized degrees of freedom of linear mixed models is derived. And its estimation will be discussed in subsequent sections.

In (1.1), a linear mixed model expresses an n_i-dimensional response vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_j})^T$ for the ith subject, $i = 1, 2, \cdots, m$, with a total number of observations $\sum_{i=1}^{m} n_i$ as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \cdots, m, \tag{2.6}$$

$$\boldsymbol{b}_i \sim N(0, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \boldsymbol{\Lambda}_i).$$

Let $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\alpha}_i) = \boldsymbol{Z}_i \boldsymbol{\Psi} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{\Lambda}_i$ with $\boldsymbol{\alpha}_i$ denoting the parameters introduced by the random effects variance-covariance components and within-group variance-covariance components, then

$$\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad i = 1, 2, \cdots, m. \tag{2.7}$$

The independent Gaussian formulation (2.7), instead of hierarchical model (1.1), simplifies the log-likelihood in selection criterion (1.3). Note that the estimator of each $\boldsymbol{\Sigma}_i$ relies on $\boldsymbol{\alpha}_i$, i.e. $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Sigma}}_i(\widehat{\boldsymbol{\alpha}}_i)$. Based upon the formulations (2.6) and (2.7), the Kullback-Leibler loss is a natural choice for measuring accuracy of estimation from the perspective of the maximum likelihood method. The Kullback-Leibler loss measures the deviation of the estimated likelihood from the true likelihood while pretending the truth is known. The performance of each pair of estimates $(\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ of $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for subject i is evaluated by its closeness to $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ in terms of the individual Kullback-Leibler loss of $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ versus $(\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$:

$$KL(p(\boldsymbol{y}_i|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)) = \int p(\boldsymbol{y}_i|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \log \frac{p(\boldsymbol{y}_i|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{p(\boldsymbol{y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i))} d\boldsymbol{y}_i. \tag{2.8}$$

This yields the Kullback-Leibler loss for all independent subjects:

$$KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \sum_{i=1}^{m} KL(p(\boldsymbol{y}_i|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)),$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_n)^T$ and $\boldsymbol{\Sigma} = \text{Diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_m)$, and so do their estimators. Direct calculations (see Appendix A) yield

$$\begin{aligned} & KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \\ = \; & \frac{1}{m}\left[-\log \sum_{i=1}^{m} p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i \right. \\ & \left. \frac{1}{2}\sum_{i=1}^{m} \boldsymbol{Y}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i + \frac{1}{2}\boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i + \frac{1}{2}\text{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1}) \right]. \end{aligned} \tag{2.9}$$

The Kullback-Leibler loss $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ compares different estimations of linear mixed models. If $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ were known, then we could select the optimal estimator or model by minimizing $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ with respect to all candidates. However, in general, $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ is not known. Consequently, $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ needs to be estimated.

Motivated from (2.4), we consider a class of loss estimators of the form

$$-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \kappa\,. \tag{2.10}$$

Members of this class penalize an increase in the size of a model used in estimation, with κ controlling the degree of penalization. Clearly, different choices of κ yield different model selection criteria; for instance, $\kappa = k_{M_\gamma}$ gives the AIC. For the optimal estimation of $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$, we choose to minimize the criterion

$$E\left[KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) - \left(-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \kappa\right)\right]^2,$$

which is the expected L_2 distance between $KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ and the class (2.10) of loss estimators. Minimizing this with respect to κ, we obtain (see Appendix A) the optimal κ as

$$GDF(M) = \sum_{i=1}^{m} h_i(M)$$

with

$$\begin{aligned} h_i(M) &= E\left[(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i\right] - \frac{1}{2} E\left[\boldsymbol{Y}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i\right] \\ &\quad + \frac{1}{2} E\left[\boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i\right] + \frac{1}{2} E\left[\ \mathrm{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1})\ \right], \end{aligned} \tag{2.11}$$

for model M which yields estimates $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$. The optimal quantity $GDF(M)$, which measures the degrees of freedom cost in model selection, is thereby defined as the *generalized degrees of freedom* for general linear mixed modeling procedure M. The $GDF(M)$ of linear mixed models (2.11) is close to the example with heteroscedastic response as in Shen and Huang (2006), but can be distinguished from it because each subject in

linear mixed models yields dissimilar variance-covariance structure.

In linear mixed models, the $GDF(M)$ sums up four parts from each subject. Because $GDF(M)$ depends on unknown parameters, we will construct the data perturbation estimator $\widehat{GDF}(M)$ of $GDF(M)$ in next section, in order to help evaluate the performance of linear mixed modeling procedures M.

As mentioned in Shen and Huang (2006), model assessment is a process of assessing the accuracy of a modeling procedure in estimation and prediction. Moreover, linear mixed models can be optimally assessed based on the Kullback-Leibler loss and the generalized degrees of freedom of linear mixed models. In fact, the estimator of the Kullback-Leibler loss of a linear mixed model M

$$-\sum_{i=1}^{m} \log p\left(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_{i,\,M}, \widehat{\boldsymbol{\Sigma}}_{i,\,M}\right) + \widehat{GDF}(M)$$

can be used to assess linear mixed model M.

2.2 Data Perturbation, Estimation of Generalized Degrees of Freedom

2.2.1 Data Perturbation

In this section, a resampling technique, *data perturbation*, is outlined for the estimation of the proposed generalized degrees of freedom of linear mixed models.

The ideas of data perturbation can be traced back to Breiman (1992) and Ye (1998) for the normal models. Data perturbation was firstly proposed in Shen, Huang and Ye (2004). Shen and Huang (2006) further developed data perturbation for a general class of losses to handle estimators of any form and a general distribution. Data perturbation assesses the sensitivity of estimated parameter through pseudo response vector

$$\boldsymbol{Y}_i^* = (Y_{i1}^*, Y_{i2}^*, \cdots, Y_{in_i}^*) = \boldsymbol{Y}_i + \tau(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{Y}_i) \tag{2.12}$$

generated from $\boldsymbol{Y}_i$, $i = 1, \cdots, n$, with the size of perturbation controlled by $\tau \in (0, 1]$ and perturbed response vector $\widetilde{\boldsymbol{Y}}_i = (Y_{i1}^*, Y_{i2}^*, \cdots, Y_{i1}^*)$. Data perturbation retains the support range of the conditional distribution of $\boldsymbol{Y}_i^*$ given $\boldsymbol{Y}_i$ to be the same as that of the distribution of $\boldsymbol{Y}_i$. The distribution of perturbed response vector $\widetilde{\boldsymbol{Y}}_i$ is specified as follows. To generate $\widetilde{\boldsymbol{Y}}_i$, there are two situations: parametric data perturbation and nonparametric data perturbation, corresponding to situations with or without specified likelihood. Parametric data perturbation samples $\widetilde{\boldsymbol{Y}}_i$ from the distribution of $\boldsymbol{Y}_i$, $p(\boldsymbol{y}_i|\boldsymbol{\mu}_i)$, with $\boldsymbol{\mu}_i = E\boldsymbol{Y}_i$ replaced by $\boldsymbol{Y}_i$, whereas nonparametric data perturbation deals with the case of unknown distribution, where $\widetilde{\boldsymbol{Y}}_i$ is sampled from an estimated distribution with the same support of $\boldsymbol{Y}_i$. Data perturbation is applicable to any sampling distribution of $\widetilde{\boldsymbol{Y}}_i$ as long as it satisfies that $E(\widetilde{\boldsymbol{Y}}_i|\boldsymbol{Y}_i) = \boldsymbol{Y}_i$, $i = 1, \cdots, n$.

For (1.1), parametric data perturbation is a reasonable choice because model (1.1) owns a explicit parametric expression. Recall that the reparameterized expression of

linear mixed models (1.1) is $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2, \cdots, m$, where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\boldsymbol{\alpha}_i) = \boldsymbol{Z}_i \boldsymbol{\Psi} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{\Lambda}_i$. Again, the $\boldsymbol{\alpha}_i$'s are denoting the parameters introduced by the random effects variance-covariance component $\boldsymbol{\Psi}$ and within-group variance-covariance components $\boldsymbol{\Lambda}_i$'s. In order to generate pseudo response vector in linear mixed models, the $\boldsymbol{\alpha}_i$'s should be estimated first to derive $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Sigma}}_i(\widehat{\boldsymbol{\alpha}}_i)$, the estimation of $\boldsymbol{\Sigma}_i$. With the $\widehat{\boldsymbol{\Sigma}}_i$'s, $\widetilde{\boldsymbol{Y}}_i$ can be sampled from $N(\boldsymbol{Y}_i, \widehat{\boldsymbol{\Sigma}}_i)$, and the pseudo response vector $\boldsymbol{Y}_i^*$ is generated by (2.12), $i = 1, 2, \cdots, m$.

2.2.2 Estimation of Generalized Degrees of Freedom of Linear Mixed Models

We now provide some heuristics for our proposed estimator. Rewrite $GDF(M)$ as

$$\begin{aligned} GDF(M) &= \sum_{i=1}^{m} E(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i - \frac{1}{2} \sum_{i=1}^{m} E \Big[\boldsymbol{Y}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i - \boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i \\ &\qquad - \text{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1}) \Big] \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y}) \widehat{\mu}_{ij}(\boldsymbol{Y}), Y_{ik}) - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y}), Y_{ij} Y_{ik}) \\ &\stackrel{set}{=} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} GDF_{ijk}^{(1)} - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} GDF_{ijk}^{(2)}, \end{aligned}$$

in which $\widehat{\sigma}_{ijk}$ is the jkth element of $\widehat{\boldsymbol{\Sigma}}_i^{-1}$ and $\widehat{\mu}_{ij}$ is the jth element of $\widehat{\boldsymbol{\mu}}_i$; $i = 1, 2, \cdots, m$, and $j, k = 1, 2, \cdots, n_i$. With perturbed $\boldsymbol{Y}_i^*$, let E^*, var*, and cov* denote the conditional mean, variance, and covariance, respectively, given $\boldsymbol{Y}_i$. Following Shen and Huang (2006), we can estimate the generalized degrees of freedom $GDF(M)$ of a linear mixed model M. To estimate $\text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y}), Y_{ik})$, for any combination of i, j, and k, note that it equals $E\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y})(Y_{ik} - EY_{ik}) = E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y})\text{var}(Y_{ik}) = \frac{E^*\text{var}^*Y_{ik}^*}{\text{var}^*Y_{ik}^*}E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y})\text{var}(Y_{ik}) = \tau^{-2}(E^*\text{var}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y})$. Then we can approximate $(E^*\text{var}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y})$ by $E^*\frac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*)\widehat{\mu}_{ij}(\boldsymbol{Y}^*)\text{var}^*(Y_{ik}^*) = \text{cov}^*(\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*)\widehat{\mu}_{ij}(\boldsymbol{Y}^*), Y_{ik}^*)$. Therefore, $\text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y})\widehat{\mu}_{ij}(\boldsymbol{Y}), Y_{ik})$ is eventually estimated by

$$\widehat{GDF}_{ijk}^{(1)} = \tau^{-2}\text{cov}^*(\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*)\widehat{\mu}_{ij}(\boldsymbol{Y}^*), Y_{ik}^*)$$

through perturbed data $\boldsymbol{Y}^*$. Similarly, to estimate $\text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y}), Y_{ij}Y_{ik})$, for any combination of i, j, and k, note that $\text{cov}(\widehat{\sigma}_{ijk}(\boldsymbol{Y}), Y_{ij}Y_{ik}) = \frac{E^*\text{var}^*(Y_{ij}^*Y_{ik}^*)}{\text{var}^*(Y_{ij}^*Y_{ik}^*)}E\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\text{var}($ $Y_{ij}Y_{ik}) = \frac{\text{var}(Y_{ij}Y_{ik})}{\text{var}^*(Y_{ij}^*Y_{ik}^*)}E^*\text{var}^*(Y_{ij}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\text{var}(Y_{ij}Y_{ik})$. Then we can approximate $E^*\text{var}^*(Y_{ij}^*Y_{ik}^*)E\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk}(\boldsymbol{Y})\text{var}(Y_{ij}Y_{ik})$ by $E^*\text{var}^*(Y_{ij}^*Y_{ik}^*)\frac{\partial}{\partial Y_{ij}^*Y_{ik}^*}\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*)\text{var}(Y_{ij}^*$ $Y_{ik}^*) = \text{cov}^*(\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*), Y_{ij}^*Y_{ik}^*)$. Denote $\pi_{ijk}^2 = \text{var}(Y_{ij}Y_{ik}) = \mu_{ik}^2\sigma_{ijj} + \mu_{ij}^2\sigma_{ikk} + 2\mu_{ik}\mu_{ij}\sigma_{ijk}$ $+ \sigma_{ijj}\sigma_{ikk} + \sigma_{ijk}^2$. A full model with all possible covariates can be fitted to estimate all the essential unknown parameters in π_i^{-2}. Suppose $\widehat{\boldsymbol{\nu}}_i$ and $\widehat{\boldsymbol{\Delta}}_i$ are estimates, by the full model, of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Denote $\widehat{\delta}_{ijk}$ to be the jkth element of $\widehat{\boldsymbol{\Delta}}_i$ and $\widehat{\nu}_{ij}$ to be the jth

element of $\widehat{\boldsymbol{\nu}}_i$. Then $\hat{\pi}^2_{ijk} = \widehat{\nu}^2_{ik}\widehat{\delta}_{ijj} + \widehat{\nu}^2_{ij}\widehat{\delta}_{ikk} + 2\widehat{\nu}_{ik}\widehat{\nu}_{ij}\widehat{\delta}_{ijk} + \widehat{\delta}_{ijj}\widehat{\delta}_{ikk} + \widehat{\delta}^2_{ijk}$ Therefore

$$\widehat{GDF}^{(2)}_{ijk} = \hat{\pi}^2_{ijk}\mathrm{cov}^*(\widehat{\sigma}_{ijk}(\boldsymbol{Y}^*), Y^*_{ij}Y^*_{ik})/\mathrm{var}^* Y^*_{ij}Y^*_{ik}.$$

Now an estimated generalized degrees of freedom of linear mixed models is

$$\widehat{GDF}(M) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left[\widehat{GDF}^{(1)}_{ijk} - \frac{1}{2}\widehat{GDF}^{(2)}_{ijk}\right] \tag{2.13}$$

involves both summations and expectations, which can be computed via a numerical approximation. For the problems considered in linear mixed models, we use a Monte Carlo numerical approximation for our implementation as suggested in Shen and Huang (2006). We sample $\boldsymbol{Y}^{*d} = (\boldsymbol{Y}^{*d}_1, \boldsymbol{Y}^{*d}_2, \cdots, \boldsymbol{Y}^{*d}_m)$, $d = 1, 2, \cdots, D$, independently from the distribution of $\boldsymbol{Y}^* = (\boldsymbol{Y}^*_1, \boldsymbol{Y}^*_2, \cdots, \boldsymbol{Y}^*_m)$ as described earlier in Section 3.1. Note that $\boldsymbol{Y}^{*d}_i$ follows the conditional distribution of $\boldsymbol{Y}^*_i$ given $\boldsymbol{Y}_i$, $i = 1, 2, \cdots, n$ and $d = 1, 2, \cdots, D$. Then $\widehat{GDF}(M)$ is approximated by

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widetilde{GDF}^{(1)}_{ijk} - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widetilde{GDF}^{(2)}_{ijk},$$

in which

$$\widetilde{GDF}^{(1)}_{ijk} = \tau^{-2}D^{-1}\left[\widehat{\sigma}_{ijk}(\boldsymbol{Y}^{*d})\widehat{\mu}_{ij}(\boldsymbol{Y}^{*d}) - D^{-1}\sum_{d=1}^{D}\widehat{\sigma}_{ijk}(\boldsymbol{Y}^{*d})\widehat{\mu}_{ij}(\boldsymbol{Y}^{*d})\right]\times$$

$$\left[Y_{ik}^{*d} - D^{-1}\sum_{d=1}^{D} Y_{ik}^{*d}\right],$$

and

$$\widetilde{GDF}_{ijk}^{(2)} = \hat{\pi}_{ijk}^2 D^{-1}\left[\widehat{\sigma}_{ijk}(\boldsymbol{Y}^{*d}) - D^{-1}\sum_{d=1}^{D}\widehat{\sigma}_{ijk}(\boldsymbol{Y}^{*d})\right]\left[Y_{ij}^{*d}Y_{ik}^{*d} - D^{-1}\sum_{d=1}^{D} Y_{ij}^{*d}Y_{ik}^{*d}\right]$$

$$\Big/ D^{-1}\left[Y_{ij}^{*d}Y_{ik}^{*d} - D^{-1}\sum_{d=1}^{D} Y_{ij}^{*d}Y_{ik}^{*d}\right].$$

The D is chosen to be sufficiently large to ensure the precision of a Monte Carlo approximation. For the problems that we consider, it is recommended that D be at least n and τ be 0.5 for small and moderate sample sizes. Of course, D may be smaller than n, as in model selection when the size of candidate models is small. We will choose D to be n in our simulations.

Chapter 3

Adaptive Model Selection in Linear Mixed Models: Procedures and Theories

In Chapter 1, the existing model selection procedures for selecting linear mixed models are reviewed. The drawbacks of non-adaptive model selection procedures are revealed. This chapter discusses adaptive model selection procedure for selecting linear mixed models. Furthermore, we discuss the theoretical properties of the adaptive model selection procedure in linear mixed models. We show that adaptive model selection procedure is asymptotically optimal among the other model selection procedures in the form of (3.1). The proofs of the theorems are defer to the appendices.

3.1 Adaptive Selection Procedure in Linear Mixed Models

In this section, we propose a data-adaptive model selection procedure for selecting linear mixed models. In Chapter 2, the generalized degrees of freedom of linear mixed models was introduced. Based on the generalized degrees of freedom of linear mixed models and its estimation, data-adaptive model selection can be performed as follows.

Now consider a class of the model selection criteria

$$-\sum_{i=1}^{m} \log p\left(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_{i,\,M}, \widehat{\boldsymbol{\Sigma}}_{i,\,M}\right) + \lambda \cdot k_M \ , \qquad \lambda \in (0, \infty). \tag{3.1}$$

It is well known that a fixed penalty λ performs well in some situations and poorly in others. To achieve the goal of adaptive selection, we choose the optimal λ by selecting the optimal model selection procedure from (3.1) indexed by $\lambda \in (0, \infty)$. For each fixed $\lambda \in (0, \infty)$, let $\widehat{M}_\lambda$ be the optimal model selected by minimizing (3.1), and let the corresponding estimate be $(\widehat{\boldsymbol{\mu}}_{i,\,\widehat{M}_\lambda}, \widehat{\boldsymbol{\Sigma}}_{i,\,\widehat{M}_\lambda})$, for $i = 1, 2, \cdots, m$. Also, denote the estimated generalized degrees of freedom of $\widehat{M}_\lambda$ by $\widehat{GDF}(\widehat{M}_\lambda)$. We are now ready to introduce $\widehat{\lambda}$. The optimal $\hat{\lambda}$ is obtained by minimizing the estimated loss of linear mixed model $\widehat{M}_\lambda$

$$-\sum_{i=1}^{m} \log p\left(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_{i,\,\widehat{M}_\lambda}, \widehat{\boldsymbol{\Sigma}}_{i,\,\widehat{M}_\lambda}\right) + \widehat{GDF}(\widehat{M}_\lambda) \tag{3.2}$$

with respect to $\lambda \in (0, \infty)$. Inserting $\widehat{\lambda}$ into (3.1) yields our adaptive model selection procedure. The adaptive selection procedure chooses the optimal model $\widehat{M}_{\widehat{\lambda}}$ from

$\{M_\gamma, \gamma \in \Gamma\}$ by minimizing a model selection criterion of the form

$$-\sum_{i=1}^{m} \log p\left(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_{i,\,M}, \widehat{\boldsymbol{\Sigma}}_{i,\,M}\right) + \hat{\lambda} \cdot k_M \ . \tag{3.3}$$

The $\widehat{\lambda}$ may depend on the data, and so does our data-adaptive model selection procedure. The adaptive penalty $\widehat{\lambda}$ estimates the ideal optimal penalization parameter over the class (3.1), whose value becomes large or small depending on whether the size of the true model is small or the true model has a parsimonious representation or not. Therefore, $\widehat{\lambda}$ allows us to approximate the best performance of a class of model selection criteria (3.1), including AIC, BIC, and RIC.

3.2 Optimality of Adaptive Selection Procedure in Linear Mixed Models

This section discusses the proposed data-adaptive model selection procedure for linear mixed models theoretically, based on the properties of data perturbation. Particularly, the asymptotic optimality of $\hat{\lambda}$ which minimizes (3.2) is established in Theorem 1 and Theorem 2; that is, $\hat{\lambda}$ approximates the best ideal performance assuming that the knowledge about the true parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in linear mixed models were known in advance.

Theorem 1. Assume that: (1) (Integrability) For some $\delta > 0$ and $\lambda \in (0, \infty)$,

$$E \sup_{\tau \in (0,\delta)} |\widehat{GDF}(\widehat{M}(\lambda))| < \infty.$$

(2) (Positive Kullback-Leibler loss)

$$\inf_{\lambda \in (0,\infty)} |KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})| > 0.$$

(3) (Fineness of variance function estimation) For any i, j, and k,

$$\lim_{m,n_i \to \infty} \lim_{\tau \to 0^+} E\left[(\widehat{\mathrm{var}}(Y_{ij}^* Y_{ik}^*) - \mathrm{var}(Y_{ij}^* Y_{ik}^*)) \frac{\mathrm{cov}^*(\widehat{\sigma}_{ijk,\hat{\lambda}}(\boldsymbol{Y}^*), Y_{ij}^* Y_{ik}^*)}{\mathrm{var}(Y_{ij}^* Y_{ik}^*)} \right] = 0.$$

Let $\hat{\lambda}$ be the minimizer of (3.2), then

$$\lim_{m,n_i \to \infty} \lim_{\tau \to 0^+} \frac{E(KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})}))}{\inf_{\lambda \in (0,\infty)} E(KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)}))} = 1. \tag{3.4}$$

The proof of Theorem 1 is deferred to Appendix B.

Theorem 2. Under the assumptions (1)–(3) in Theorem 1 and (4) (Loss and risk)

$$\lim_{m,n_i \to \infty} \sup_{\lambda \in (0,\infty)} \left| \frac{KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})}{E(KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)}))} - 1 \right| = 0,$$

then

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\frac{KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})})}{\inf_{\lambda\in\Lambda} KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})}=1. \tag{3.5}$$

The proof of Theorem 2 is deferred to Appendix C.

Theorems 1 and 2 yield the asymptotic optimality of the proposed adaptive model selection procedure in linear mixed models. The estimated penalization parameter $\hat{\lambda}$ by the adaptive selection procedure is optimal in terms that the linear mixed model $\widehat{M}(\lambda)$ selected by minimizing (3.3) with a data-adaptive $\hat{\lambda}$ asymptotically achieves the minimal loss among all models selected by procedures with form (3.1).

Chapter 4

Adaptive Model Selection in Linear Mixed Models: Simulations Studies

This chapter numerically illustrates some aspects of adaptive model selection procedure in linear mixed models. It reinforces our view on various key issues in selecting linear mixed models: (1) selecting covariates in the mean structure; (2) selecting variance structures for modeling heteroscedasticity of within-group errors and random effect components; (3) selecting correlation structures for modeling dependence of within-group errors and random effect components. The first three sections of this chapter deal with the three key issues, respectively. In the final section, the sensitivity study of perturba-

tion size of the proposed adaptive model selection procedure in linear mixed models is studied.

4.1 Selection of Covariates of Linear Mixed Models

In this section, the advantages of the adaptive model selection procedure for selecting covariates in linear mixed models are demonstrated numerically. We compare the proposed procedure with three nonadaptive model selection procedures: the AIC, the BIC, and the RIC.

The simulation example, modified from the setup similar to the examples in George and Foster (2000), Shen and Ye (2002), and Shen and Huang (2004), are conducted with correlated and independent covariates. Consider a linear mixed model with random intercept and random slope

$$Y_{ij} = \alpha_0 + b_{i0} + t_{ij}\alpha_1 + t_{ij}b_{i1} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}, \tag{4.1}$$

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right), \qquad \epsilon_{ij} \sim N(0, \sigma^2),$$

for simulated longitudinal data with $m = 50$ subjects and $n_i = 5$ observed response values for each subject at time point $t_{ij} = j - 1$, $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$. In

(4.1), the response Y_{ij} depends on the time variable t_{ij} and time-independent covariates $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijq})^T$. The parameters α_0 and α_1 are the regression intercept and slope of interest. Here $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_q)^T$ is a vector of nuisance parameters that may or may not have impact on responses. In (4.1), $(b_{i0}, b_{i1})^T$, $i = 1, 2, \cdots, m$, are random effects variance components, and $\epsilon_{ij} \sim N(0, \sigma^2)$, $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$, are within-group variance components.

We perform simulations under different scenarios, including different sizes of models with both correlated and independent time-independent covariates. A random sample

$$\{(Y_{ij}, \boldsymbol{x}_{ij}) : i = 1, 2, \cdots, m, \text{ and, } j = 1, 2, \cdots, n_i\}$$

is generated according to (4.1), with $m = 50$ and $n_i = 5$ for any $i = 1, 2, \cdots, m$, where Y_{ij} follows (4.1) with $q = 40$, $\alpha_0 = 1$, $\alpha_1 = 1$, $t_{ij} = j - 1$ for $\forall\ i$, j, $\sigma_{11}^2 = \sigma_{22}^2 = \sigma^2 = 0.5$; $\boldsymbol{x}_{ij}$ follows $N(\mathbf{0}, \boldsymbol{\Phi})$ with a covariance matrix $\boldsymbol{\Phi}_{50\times 50}$ whose diagonal elements are 1 and whose $l_1 l_2$th element is $\rho^{|l_1 - l_2|}$ for $l_1 \neq l_2$. $\boldsymbol{\beta}$ is designed to consist of 5 replications of its first ten elements $\beta_1, \beta_2, \cdots, \beta_{10}$, that is

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_{10}, \beta_1, \beta_2, \cdots, \beta_{10}, \beta_1, \beta_2, \cdots, \beta_{10}, \beta_1, \beta_2, \cdots, \beta_{10})^T.$$

Nine situations are now considered when $c = 0, 1, 2, \cdots, 8$. For each c, the cth choice of $(\beta_1, \beta_2, \cdots, \beta_{10})^T$ comprises the first c values that are equal to a constant B_c and

0's otherwise; for instance, when $c = 2$, $(\beta_1, \beta_2, \cdots, \beta_{10})^T = (B_2, B_2, 0, \cdots, 0)^T$. The values of B_c's are chosen such that $\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} / (\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} + 100) = 3/4$, in which $\boldsymbol{X}$ is the design matrix of $\boldsymbol{x}_{ij}$'s.

We examine four procedures: the AIC, the BIC, the RIC and the adaptive selection procedure for $\rho = 0$, 0.5 and -0.5. To make a fair comparison, the performance of the procedures is evaluated by the Kullback-Leibler loss (2.9) for selected model $\widehat{M}$. Because of simulation study, where the truth is known each time, the Kullback-Leibler distance can be easily calculated for selected model $\widehat{M}$.

For each $c = 0, 1, 2, \cdots, 8$ and each of $\rho = 0$, 0.5, -0.5, the averages of the Kullback-Leibler loss as well as the corresponding standard errors are computed over 100 simulation replications and are reported in Tables 4.1 — 4.3 for $\rho = 0$, 0.5, and -0.5 respectively. Four competing procedures, the AIC, the BIC, the RIC, and our adaptive model selection procedure, are examined. The Kullback-Leibler loss is calculated assuming that the true model would have been known. This provides a baseline for comparison. The plots of the averages of the Kullback-Leibler loss as a function of c are displayed in Figures 4.1 — 4.3 for $\rho = 0$, 0.5, and -0.5 respectively. The average estimated numbers of selected variables are given to indicate the quality of estimating the size of the true model.

As suggested by Tables 4.1 — 4.3 and Figures 4.1 — 4.3, the adaptive selection procedure with a data-adaptive $\widehat{\lambda}$ performs well in selection covariates in linear mixed

	Adaptive Selection	AIC
$c = 0$	0.0731(0.00135)[0.46]	0.3884(0.00312)[6.34]
$c = 1$	0.1306(0.00181)[5.14]	0.4291(0.00328)[9.76]
$c = 2$	0.2201(0.00235)[10.23]	0.4569(0.00338)[15.04]
$c = 3$	0.2900(0.00269)[15.98]	0.4898(0.00350)[19.45]
$c = 4$	0.3821(0.00309)[22.44]	0.5252(0.00362)[24.39]
$c = 5$	0.4623(0.00340)[29.96]	0.5642(0.00376)[29.09]
$c = 6$	0.5509(0.00371)[35.15]	0.5955(0.00386)[32.29]
$c = 7$	0.6146(0.00392)[40.18]	0.6220(0.00394)[35.37]
$c = 8$	0.6694(0.00409)[45.41]	0.6537(0.00404)[38.18]
	BIC	RIC
$c = 0$	0.1043(0.00161)[1.41]	0.0536(0.00116)[0.38]
$c = 1$	0.1739(0.00209)[5.94]	0.1141(0.00169)[5.01]
$c = 2$	0.2314(0.00241)[10.84]	0.1861(0.00216)[9.89]
$c = 3$	0.3007(0.00274)[15.91]	0.2590(0.00254)[15.27]
$c = 4$	0.3644(0.00302)[20.99]	0.3369(0.00290)[20.97]
$c = 5$	0.4406(0.00332)[24.70]	0.4328(0.00329)[25.68]
$c = 6$	0.5101(0.00357)[27.16]	0.5527(0.00372)[28.10]
$c = 7$	0.6168(0.00393)[28.15]	1.0161(0.00504)[27.13]
$c = 8$	0.7765(0.00441)[29.01]	1.5471(0.00622)[29.92]

Table 4.1: Averages of the Kullback-Leibler loss based on 100 simulation replications and corresponding standard errors (in parentheses) for the simulations in Section 4.1 with $\rho = -0.5$, as well as averages of the number of covariates selected (in brackets).

	Adaptive Selection	AIC
$c = 0$	0.0788(0.00140)[0.45]	0.4116(0.00321)[6.34]
$c = 1$	0.1616(0.00201)[5.13]	0.4402(0.00332)[9.41]
$c = 2$	0.2057(0.00227)[10.93]	0.4686(0.00342)[14.01]
$c = 3$	0.2840(0.00266)[15.80]	0.5027(0.00355)[18.31]
$c = 4$	0.3754(0.00306)[21.03]	0.5263(0.00363)[21.26]
$c = 5$	0.5218(0.00361)[25.48]	0.5599(0.00374)[22.59]
$c = 6$	0.6576(0.00405)[27.31]	0.6441(0.00401)[23.07]
$c = 7$	0.6785(0.00412)[30.03]	0.6594(0.00406)[23.71]
$c = 8$	0.7075(0.00421)[35.79]	0.7337(0.00428)[24.70]
	BIC	RIC
$c = 0$	0.1184(0.00172)[1.41]	0.0574(0.00120)[0.31]
$c = 1$	0.1907(0.00218)[5.84]	0.1319(0.00182)[5.07]
$c = 2$	0.2377(0.00244)[10.88]	0.1786(0.00211)[10.39]
$c = 3$	0.3000(0.00274)[14.59]	0.2531(0.00252)[14.94]
$c = 4$	0.3619(0.00301)[15.96]	0.3768(0.00307)[16.54]
$c = 5$	0.4797(0.00346)[16.58]	0.6564(0.00405)[16.15]
$c = 6$	0.7496(0.00433)[17.42]	1.2084(0.00550)[17.20]
$c = 7$	1.0065(0.00502)[17.95]	1.6317(0.00639)[17.92]
$c = 8$	1.3596(0.00583)[18.40]	1.8472(0.00715)[17.21]

Table 4.2: Averages of the Kullback-Leibler loss based on 100 simulation replications and corresponding standard errors (in parentheses) for the simulations in Section 4.1 with $\rho = 0$, as well as averages of the number of covariates selected (in brackets).

	Adaptive Selection	AIC
$c = 0$	0.0768(0.00139)[0.34]	0.3916(0.00313)[6.77]
$c = 1$	0.1623(0.00201)[5.17]	0.4310(0.00328)[9.90]
$c = 2$	0.3912(0.00313)[10.37]	0.4787(0.00346)[14.73]
$c = 3$	0.5691(0.00377)[15.74]	0.5457(0.00369)[19.47]
$c = 4$	0.7126(0.00422)[22.35]	0.6431(0.00401)[23.94]
$c = 5$	0.7592(0.00436)[29.84]	0.7296(0.00427)[28.49]
$c = 6$	0.7690(0.00438)[35.92]	0.7898(0.00444)[31.63]
$c = 7$	0.7647(0.00437)[40.20]	0.8460(0.00460)[33.66]
$c = 8$	0.7585(0.00435)[44.62]	0.8956(0.00473)[34.95]
	BIC	RIC
$c = 0$	0.1175(0.00171)[1.43]	0.0484(0.00110)[0.25]
$c = 1$	0.1737(0.00208)[6.06]	0.1155(0.00170)[5, 07]
$c = 2$	0.3443(0.00293)[11.19]	0.4300(0.00328)[10.86]
$c = 3$	0.5870(0.00383)[15.84]	0.7175(0.00424)[14.38]
$c = 4$	0.7884(0.00444)[20.63]	0.9424(0.00485)[19.51]
$c = 5$	0.9430(0.00486)[23.93]	1.0894(0.00522)[20.70]
$c = 6$	1.0481(0.00512)[25.70]	1.2452(0.00558)[23.12]
$c = 7$	1.1197(0.00529)[26.20]	1.3639(0.00584)[24.10]
$c = 8$	1.2179(0.00552)[26.51]	1.4594(0.00604)[23.55]

Table 4.3: Averages of the Kullback-Leibler loss based on 100 simulation replications and corresponding standard errors (in parentheses) for the simulations in Section 4.1 with $\rho = 0.5$, as well as averages of the number of covariates selected (in brackets).

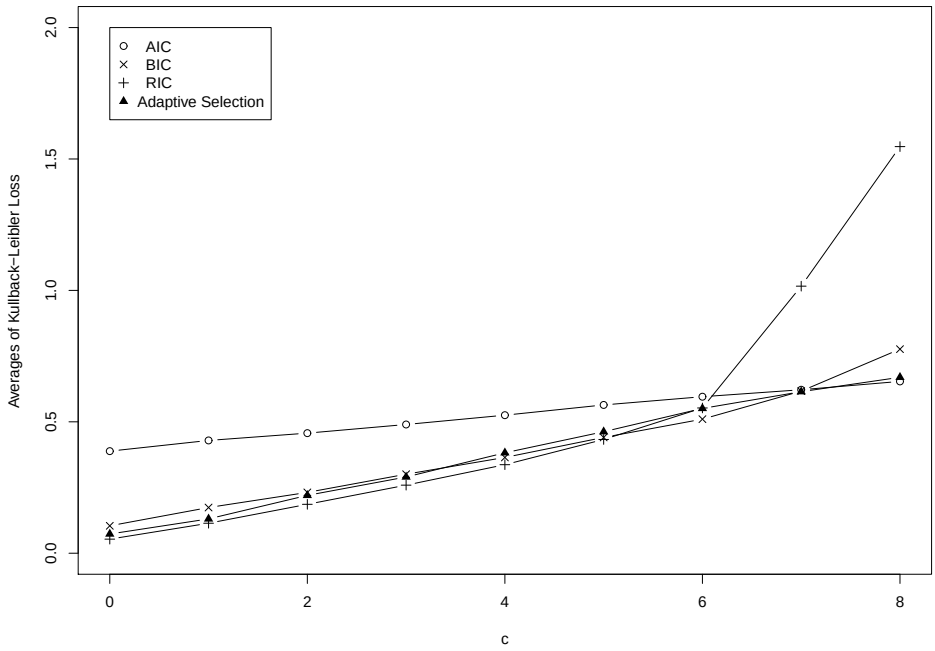

Figure 4.1: Averages of the Kullback-Leibler loss of AIC, BIC, RIC, and adaptive model selection procedure for the simulations in Section 4.1 based on 100 simulation replications with $\rho = -0.5$.

models, across all the nine situations ($c = 0, 1, 2, \cdots, 8$) with independent covariates ($\rho = 0$) and correlated covariates ($\rho = 0.5$ and $\rho = -0.5$). The proposed procedure yields the competitive performance in most of situations. Even though, in a few cases, the adaptive model selection procedure doesn't not yield the best performance, it is still substantially close to the best performance. It is evident from Figures 1 — 3 that a nonadaptive penalty, as defined in (1.3) such as the AIC, the BIC and the RIC, cannot perform well for both large and small c simultaneously. The BIC and the RIC yield

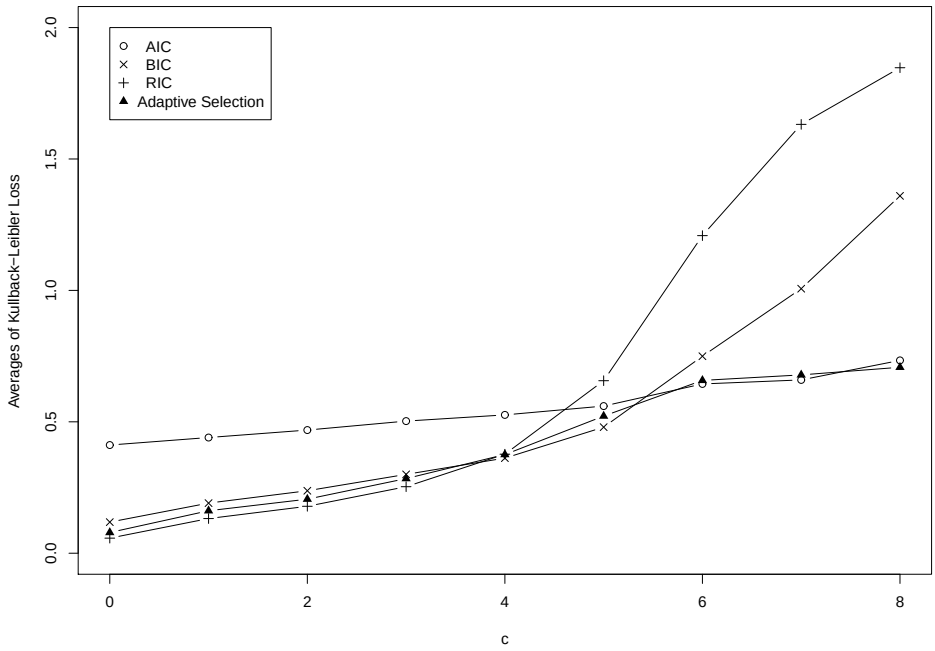

Figure 4.2: Averages of the Kullback-Leibler loss of AIC, BIC, RIC, and adaptive model selection procedure for the simulations in Section 4.1 based on 100 simulation replications with $\rho = 0$.

better performance than the AIC for small c and yield worse performance for large c. The AIC does just the opposite. Moreover, as c increases, the Kullback-Leibler loss of the proposed adaptive model selection procedure and the AIC stabilize, whereas the Kullback-Leibler loss of the BIC and the RIC increases dramatically. The simulations show that the proposed model selection procedure with a data-adaptive penalization parameter $\widehat{\lambda}$ performs well for both large and small c, which constitutes a basis for the adaptive model selection and provides a solution to the problem of a nonadaptive

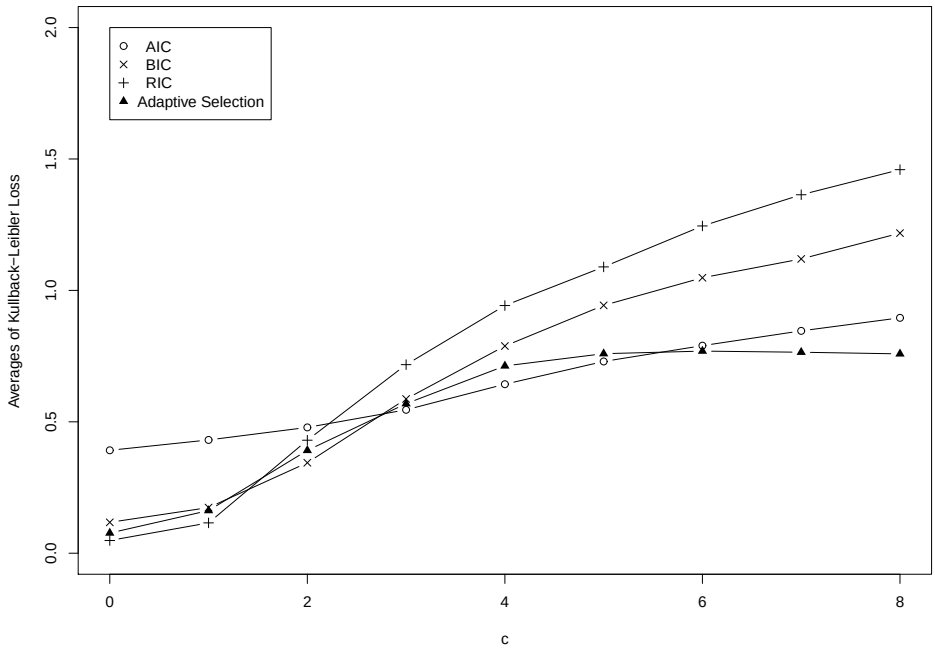

Figure 4.3: Averages of the Kullback-Leibler loss of AIC, BIC, RIC, and adaptive model selection procedure for the simulations in Section 4.1 based on 100 simulation replications with $\rho = 0.5$.

penalty.

4.2 Selection of the Variance Structure of Linear Mixed Model

In this section, we perform simulations to show how the proposed adaptive model selection procedure is beneficial to the variance structure selection of linear mixed models. We will show the advantages of the adaptive model selection procedure in selecting

correlation structure of linear mixed models in subsequent section.

As described in Chapter 1, a linear mixed model can expressed as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \cdots, m,$$

$$\boldsymbol{b}_i \sim N(0, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{\Lambda}_i), \tag{4.2}$$

where the $\boldsymbol{\Lambda}_i$ are positive-definite matrices parameterized by a fixed, generally small set of parameters. As before, the within-group errors $\boldsymbol{\epsilon}_i$ are assumed to be independent for different i and independent of the random effect $\boldsymbol{b}_i$. The σ^2 is factored out of the $\boldsymbol{\Lambda}_i$ for computational convenience. The flexibility of the specification of $\boldsymbol{\Lambda}_i$ allows the linear mixed models (4.2) to capture heteroscedasticity and correlation within-group errors simultaneously. There are many applications involving grouped data for which the within-group errors are heteroscedastic (i.e., have unequal variances) or are correlated or are both heteroscedastic and correlated. As in Pinheiro and Bates (2000), the $\boldsymbol{\Lambda}_i$ in (4.2) can be reparameterized, so that one part of $\boldsymbol{\Lambda}_i$ models correlation between within-group errors and another part of it models heteroscedasticity of within-group errors. The $\boldsymbol{\Lambda}_i$ matrices in (4.2) can be decomposed into a product of three matrices $\boldsymbol{\Lambda}_i = \boldsymbol{V}_i\boldsymbol{C}_i\boldsymbol{V}_i$, where $\boldsymbol{V}_i$ is a diagonal matrix and $\boldsymbol{C}_i$ is a positive-definite correlation matrix with all diagonal elements equal to one. The matrix $\boldsymbol{V}_i$ is not uniquely defined in decomposition $\boldsymbol{\Lambda}_i = \boldsymbol{V}_i\boldsymbol{C}_i\boldsymbol{V}_i$. We require that the diagonal elements of $\boldsymbol{V}_i$ must be

positive to ensure uniqueness of decomposition of $\boldsymbol{V}_i$. It can be easily conclude that $\text{var}(\epsilon_{ij}) = \sigma^2[\boldsymbol{V}_i]^2_{jj}$, and $\text{cor}(\epsilon_{ij}, \epsilon_{ik}) = [\boldsymbol{C}_i]_{jk}$. The $\boldsymbol{V}_i$ describes the variance of the within-group errors $\boldsymbol{\epsilon}_i$, and therefore it is called the *variance structure component* of $\boldsymbol{\epsilon}_i$ or $\boldsymbol{\Lambda}_i$. The $\boldsymbol{C}_i$ describes the correlation of the within-group errors $\boldsymbol{\epsilon}_i$, and it is called the *correlation structure component* of $\boldsymbol{\epsilon}_i$ or $\boldsymbol{\Lambda}_i$.

Variance functions are used to model the variance structure of the within-group errors using covariates. Following Pinheiro and Bates (2000), we define that the general *variance function* model for the within-group errors in linear mixed models (4.2) as

$$\text{var}(\epsilon_{ij}|\boldsymbol{b}_i) = \sigma^2 g^2(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}), \qquad i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, n_i, \tag{4.3}$$

where $\mu_{ij} = E[y_{ij}|\boldsymbol{b}_i]$, $\boldsymbol{v}_{ij}$ is a vector of variance covariates, $\boldsymbol{\delta}$ is a vector of variance parameters and $g(\cdot)$ is the variance function assumed to be continuous in $\boldsymbol{\delta}$. The variance function (4.3) is flexible and intuitive, because it allows the within-group variance to depend on the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}_i$ through the expected value μ_{ij}.

Statistical simulations are performed to examine the performance of the adaptive model selection procedure in the selecting the variance structure of linear mixed models. Consider a linear mixed model

$$Y_{ij} = \alpha_0 + b_{i0} + t_{ij}\alpha_t + t_{ij}b_{i1} + \sum_{l=1}^{p} x_{ij\,l}\beta_l + \epsilon_{ij}, \qquad i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, n_i, \tag{4.4}$$

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right),$$

with variance structure

$$\text{var}(\epsilon_{ij} | \, \boldsymbol{b}_i) = \sigma^2 \sum_{l=1}^{r} x_{ijl}^2, \qquad i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, n_i,$$

and correlation structure

$$\text{cor}(\epsilon_{ij}, \epsilon_{ij'}) = 0, \qquad i = 1, 2, \cdots, m, \quad j, j' = 1, 2, \cdots, n_i,$$

where $m = 20$, $n_i = 10$, $p = 10$, and $\beta_1 = \cdots = \beta_{10} = 1$, $\alpha_0 = \alpha_0 = 1$, $\sigma_0^2 = \sigma_1^2 = \sigma^2 = 0.5$, $t_{ij} = j - 1$ for $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$, and $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijp})^T$ follows $N(\boldsymbol{0}, \boldsymbol{I}_{10\times 10})$.

In this simulation study, a random sample

$$\{(Y_{ij}, \boldsymbol{x}_{ij}) : i = 1, 2, \cdots, m, \text{ and, } j = 1, 2, \cdots, n_i\}$$

is generated according to (4.4). Ten variance structures $r = 1, 2, \cdots, 10$ are examined, with the prior knowledge that all the covariates are active in the mean function. The goal of the variance structure selection is to decide which of the 10 covariates are as-

	Adaptive Selection	AIC	BIC
$r = 1$	0.2478(0.00253)[1.49]	0.4789(0.00364)[4.74]	0.2750(0.00339)[1.88]
$r = 2$	0.2462(0.00394)[2.30]	0.4850(0.00427)[5.05]	0.3067(0.00403)[2.31]
$r = 3$	0.2606(0.00278)[3.36]	0.4960(0.00504)[5.60]	0.2971(0.00297)[3.32]
$r = 4$	0.3179(0.00281)[4.37]	0.5287(0.00475)[6.35]	0.3381(0.00446)[4.47]
$r = 5$	0.2964(0.00405)[5.05]	0.5036(0.00457)[6.56]	0.3555(0.00327)[5.21]
$r = 6$	0.3235(0.00319)[6.25]	0.5047(0.00419)[7.04]	0.3521(0.00474)[5.79]
$r = 7$	0.3536(0.00319)[7.12]	0.5284(0.00429)[7.58]	0.3817(0.00364)[6.67]
$r = 8$	0.3772(0.00374)[7.81]	0.5403(0.00423)[8.18]	0.3726(0.00455)[7.83]
$r = 9$	0.3997(0.00311)[8.66]	0.5662(0.00370)[8.55]	0.3844(0.00372)[8.64]
$r = 10$	0.3934(0.00500)[9.58]	0.5418(0.00539)[9.09]	0.4043(0.00469)[9.18]

Table 4.4: Averages of the Kullback-Leibler loss based on 100 simulation replications and corresponding standard errors (in parentheses) for the simulations in Section 4.2, as well as averages of the number of variance covariates selected (in brackets).

sociated with the variance of within-group error ϵ_{ij}. We will select the best variance function from the candidates with the form $\text{var}(\epsilon_{ij}|\boldsymbol{b}_i) = \sigma^2 \sum_{l=1}^{10} \delta_l x_{ij\,l}^2$, $\delta_l = 0$ or 1. For implementation, we apply exhaust search for all possible candidate variance structures in $\text{var}(\epsilon_{ij}|\boldsymbol{b}_i) = \sigma^2 \sum_{l=1}^{10} \delta_l x_{ij\,l}^2$ with $\delta_l = 0$ or 1, $l = 1, 2, \cdots, 10$. We compare three procedures: the AIC, the BIC, and the adaptive model selection procedure. For adaptive model selection procedure $\hat{\lambda}$ is obtained by minimizing (3.2). The performance of the procedures is evaluated by the Kullback-Leibler loss (2.9) for selected models.

Table 4.4 displays the averages of the Kullback-Leibler loss of 100 replications of above simulations, as well as the corresponding standard errors, and the averages of covariates in the variance structure of selected models. As suggested by Tables 4.4, the adaptive selection procedure with a data-adaptive $\widehat{\lambda}$ yields the best performance in eight situations of ten, and perform close to the best in the other two situations. It is evident from Table 4.4 that a nonadaptive penalty, as defined in (1.3) such as the AIC and the BIC, cannot perform well for both large and small r simultaneously. The AIC performs the worst among three procedures, giving the largest Kullback-Leibler loss in all ten situations. The simulations show that the proposed model selection procedure with a data-adaptive penalization parameter $\widehat{\lambda}$ is able to perform well in selecting the variance structure of linear mixed models.

4.3 Selection of the Correlation Structure of Linear Mixed Model

In linear mixed models (4.2), correlation structures are used to model dependence among the within-group errors. We assume that the within-group errors ϵ_{ij} in (4.2) are associated with position vector $\boldsymbol{p}_{ij}$. For time series models, the $\boldsymbol{p}_{ij}$ are typically integer scalars, while for spatial models they are generally two-dimensional coordinate vectors. The correlation structures considered here are assumed to be *isotropic*, i.e., the correlation between two within-group errors ϵ_{ij} and $\epsilon_{ij'}$ is assumed to depend on the

corresponding position vectors $\boldsymbol{p}_{ij}$ and $\boldsymbol{p}_{ij'}$ only through some distance between them, say $d(\boldsymbol{p}_{ij}, \boldsymbol{p}_{ij'})$, and not on the particular values they assume. The general within-group isotropic *correlation structure* for grouping is modeled as

$$\text{cor}(\epsilon_{ij}, \epsilon_{ij'}) = h[d(\boldsymbol{p}_{ij}, \boldsymbol{p}_{ij'}), \boldsymbol{\rho}], \qquad i = 1, 2, \cdots, m, \quad j, j' = 1, 2, \cdots, n_i,$$

where $\boldsymbol{\rho}$ is a vector of correlation parameters and $h(\cdot)$ is a correlation function taking values between -1 and 1 assumed continuous in $\boldsymbol{\rho}$ and $h(0, \boldsymbol{\rho}) = 1$, that is, if two observations have identical position vectors, they have correlation 1.

One type of classic isotropic correlation structure models is *mixed autoregressive-moving average* models, or *ARMA* models (Box et al., 1994). They are obtained by combining together an autoregressive model and a moving average model. The within-group errors of $ARMA(r_1, r_2)$ models are expressed as

$$\epsilon_t = \sum_{i=1}^{r_1} \phi_i \epsilon_{t-i} + \sum_{j=1}^{r_2} \theta_j a_{t-j} + a_t.$$

There are $r_1 + r_2$ correlation parameter $\boldsymbol{\rho}$ in an $ARMA(r_1, r_2)$ model, corresponding to the combination of the r_1 autoregressive parameters $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_{r_1})$ and the r_2 moving average parameters $\boldsymbol{\theta} = \theta_1, \cdots, \theta_{r_2}$. The correlation function for an $ARMA(r_1, r_2)$

model can be obtained using the recursive relations

$$h(k,\boldsymbol{\rho}) = \begin{cases} \sum_{i=1}^{r_1} \phi_i h(|k-i|,\boldsymbol{\rho}) + \sum_{j=1}^{r_2} \theta_j g(|k-j|,\boldsymbol{\rho}), & k = 1,2,\cdots,r_2, \\ \\ 0, & k = r_2+1, r_2+2, \cdots, \end{cases}$$

where $g(k,\boldsymbol{\rho}) = E(\epsilon_{t-k}a_t)/\mathrm{var}(\epsilon_t)$.

Simulations are performed as follows, to demonstrate the advantages of the adaptive model selection procedure in selecting linear mixed models with $ARMA(r_1, r_2)$ correlation structures. Consider a linear mixed model

$$Y_{ij} = \alpha_0 + b_{i0} + t_{ij}\alpha_t + t_{ij}b_{i1} + \epsilon_{ij}, \qquad i = 1,2,\cdots,m, \quad j = 1,2,\cdots,n_i, \tag{4.5}$$

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} \right),$$

with variance-covariance structure being $ARMA(r_1, r_2)$

$$\epsilon_{ij} = \sum_{l_1=1}^{r_1} \phi_{l_1}\epsilon_{i,\,j-l_1} + \sum_{l_2=1}^{r_2} \theta_{l_2} a_{i,\,j-l_2} + a_{ij},$$

where $m = 5$, $n_i = 50$, $\alpha_0 = \alpha_0 = 1$, $\sigma_0^2 = \sigma_1^2 = 0.5$, $t_{ij} = j-1$ for $i = 1,2,\cdots,m$ and $j = 1,2,\cdots,n_i$, $\phi_{l_1} = \theta_{l_2} = 0.5$, and a_{ij} follows $N(0, 0.5)$.

We perform simulations under the different $ARMA(r_1, r_2)$ correlation structures of within-group errors of (4.5). A random sample

$$\{(Y_{ij}, \boldsymbol{x}_{ij}) : i = 1, 2, \cdots, m, \text{ and, } j = 1, 2, \cdots, n_i\}$$

is generated according to (4.5). In this simulation study, two-way combinations of $r_1 = 0, 1, \cdots, 5$ and $r_2 = 5$ are examined. We suppose that $r_1 \leq 10$ and $r_2 \leq 10$. The goal of the correlation structure selection is to decide parameters r_1 and r_2. For implementation, we apply exhaust search for all possible $ARMA(r_1, r_2)$ candidate models for $r_1 = 0, 1, \cdots, 10$ and $r_2 = 0, 1, \cdots, 10$. We compare three procedures: the AIC, the BIC, and the adaptive model selection. For the adaptive model selection, $\hat{\lambda}$ is obtained by minimizing (3.2). The performance of all three procedures is evaluated by the Kullback-Leibler loss (2.9) for selected model.

Table 4.5 displays the averages of the Kullback-Leibler loss of 100 replications of above simulations, as well as the corresponding standard errors of Kullback-Leibler distances, and the averages of parameters r_1 and r_2 of selected models. As suggested by Tables 5, the adaptive selection procedure with a data-adaptive $\widehat{\lambda}$ performs well in selecting the variance structure of linear mixed models, across all the five situations of combinations of $r_1 = 0, 1, \cdots, 5$ and $r_2 = 5$. The proposed procedure yields the best performance in all situations. It is evident from Table 4.5 that a nonadaptive penalty, as defined in (1.3) such as in the AIC and the BIC, cannot perform well for both large

True Structure	Adaptive Selection	AIC	BIC
$ARMA(0,5)$	0.1867(0.00475) $(0.46, 4.77)$	0.4693(0.00576) $(3.40, 5.76)$	0.2008(0.00245) $(0.58, 6.28)$
$ARMA(1,5)$	0.1770(0.00299) $(1.29, 5.10)$	0.4816(0.00466) $(4.33, 5.57)$	0.2496(0.00446) $(1.33, 5.40)$
$ARMA(2,5)$	0.1832(0.00343) $(2.09, 4.86)$	0.4928(0.00552) $(5.29, 5.88)$	0.2406(0.00262) $(2.02, 4.86)$
$ARMA(3,5)$	0.2324(0.00446) $(2.95, 5.18)$	0.4752(0.00407) $(4.20, 8.08)$	0.2577(0.00255) $(2.75, 4.82)$
$ARMA(4,5)$	0.2027(0.00300) $(3.72, 5.14)$	0.4674(0.00387) $(4.22, 5.63)$	0.2976(0.00430) $(3.43, 5.90)$
$ARMA(5,5)$	0.2216(0.00459) $(4.55, 4.92)$	0.4734(0.00416) $(6.21, 6.21)$	0.3255(0.00291) $(4.17, 2.87)$

Table 4.5: Averages of the Kullback-Leibler loss based on 100 simulation replications and corresponding standard errors (in parentheses) for the simulations in Section 4.3, as well as averages of selected r_1 and r_2 (underneath).

and small r_1 simultaneously. The AIC performs the worst among three procedures, giving the largest Kullback-Leibler distances in all five situations. Simulation results show that the proposed model selection procedure with the data-adaptive penalization parameter $\widehat{\lambda}$ is able to perform well in selecting the variance structure of linear mixed models.

4.4 Sensitivity Study of Perturbation Size in Data Perturbation

The estimator of the generalized degrees of freedom in linear mixed models in (2.11) depends on perturbation size τ. As suggested by Theorems 1 and 2, the perturbation size should be chosen as small as possible, yielding the optimality of adaptive model selection procedure in linear mixed models. But, in practice, very small τ may not be preferable. A possible solution is to use another model assessment criterion, such as cross-validation to estimate the optimal τ from data, which removes the dependency of data perturbation estimator on τ. However, it will be too computationally intensive for selection problems in linear mixed models. Shen et al. (2004) and Shen and Huang (2006) suggested to use $\tau = 0.5$. In all previous simulation studies, $\tau = 0.5$ is also applied to abtain the estimation of generalized degrees of freedom in linear mixed models.

In this section, we investigate the sensitivity of the performance of the proposed adaptive selection procedure of linear mixed model to the perturbation size τ via a

	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
$\rho = 0$, $c = 8$	0.7112	0.7288	0.7075	0.7100	0.7419
$r = 10$	0.4095	0.3977	0.3934	0.3981	0.3220
$ARMA(5,5)$	0.2660	0.2450	0.2216	0.2935	0.2565
	AIC	BIC			
$\rho = 0$, $c = 8$	0.7337	1.3596			
$r = 10$	0.5418	0.4043			
$ARMA(5,5)$	0.4734	0.3255			

Table 4.6: Sensitivity study of perturbation size of specific model selection settings.

small simulation study. The simulation study is conducted in three particular settings of previous simulations (see Table 4.6), with $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$. The three particular settings are (1) $\rho = -0.5$ and $p = 50$ in the simulation example of covariate selection in Section 4.1, (2) $r = 10$ in the simulation example of variance structure selection in Section 4.2, and (3) $ARMA(5,5)$ in the simulation example of correlation structure selection in Section 4.3. The results are shown in Table 4.6. Evidently, the Kullback-Leibler loss of the adaptive model selection procedure in linear mixed models hardly varies as a function of size τ in all of the situations. Therefore, $\tau = 0.5$ is a reliable choice for the adaptive model selection procedure in linear mixed models.

Chapter 5

Adaptive Model Selection in Linear Mixed Models: Applications

In this chapter, we apply the proposed adaptive model selection procedure in linear mixed models to the data from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984).

The Wisconsin Epidemiologic Study of Diabetic Retinopathy began in 1979. It was initially funded by the National Eye Institute. One purpose of the Wisconsin Epidemiologic Study of Diabetic Retinopathy was to describe the frequency and incidence of complications associated with diabetes especially eye complications such as diabetic

retinopathy and visual loss, kidney complications such as diabetic nephropathy, and amputations. Another purpose was to identify risk factors such as poor glycemic control, smoking, and high blood pressure, which may contribute to the development of these complications. In addition, another purpose of the Wisconsin Epidemiologic Study of Diabetic Retinopathy was to examine health care delivery for people with diabetes. The Wisconsin Epidemiologic Study of Diabetic Retinopathy examinations were done in a large 40-foot mobile examining van in an eleven-county area in southern Wisconsin and involved all persons (720 individuals) with younger-onset Type 1 diabetes and older-onset persons mostly with Type 2 diabetes who were first examined from 1980 to 1982. The examinations were done near participants' residences. The van provided standardized conditions to examine participants and minimized participants' travel time. The examination involved refraction and measurement of best corrected visual acuity, examination of the front and back of the eye, measurement of the pressure in the eye, fundus photography, blood tests for glycosylated hemoglobin (a measure of recent blood sugar control) and random blood sugar, and urine protein analysis. The fundus photographs were masked for anonymity and then graded by trained graders. This provided objective information about the presence and severity of diabetic retinopathy (damage of the retinal blood vessels from diabetes) and macular edema (swelling of the center of the retina) in each eye.

The available data set from the Wisconsin Epidemiologic Study of Diabetic Retino-

pathy involves the response, which is the severity of diabetic retinopathy, and 13 potential risk factors, which are age at diagnosis of diabetes (years), duration of diabetes(years), glycosylated hemoglobin level, systolic blood pressure, diastolic blood pressure, body mass index, pulse rate (beats/30 seconds), gender (male = 0, female = 1), proteinuria(absent = 0, present = 1), doses of insulin per day (0, 1, 2), residence (urban = 0, rural = 1), refractive error of eye, intraocular pressure of eye. The goal is to determine the risk factors for diabetic retinopathy. The linear mixed model with random intercept is fitted to the data.

To identify the risk factors of diabetic retinopathy and better estimate the influence of the risk factors on the severity of diabetic retinopathy, we performed covariate selection by the AIC, the BIC, and the proposed adaptive model selection procedure. The selected models, along with the estimated coefficients of selected covariates, are reported in Table 5.1. Duration of diabetes, diastolic blood pressure and body mass index are identified as risk factors by all three procedures. The AIC, with smaller penalization parameters than the BIC, selects one more covariate, glycosylated hemoglobin level, as the risk factor of diabetic retinopathy. The risk factors that the adaptive model selection procedure selects, however, include not only five factors chosen by the AIC and the BIC but also include age at diagnosis of diabetes and proteinuria, indicating that adding age at diagnosis of diabetes and proteinuria into the model may improve its performance. From Table 5.1, we can see that proteinuria and age at diagnosis of diabetes

are important risk factors of diabetic retinopathy, and adding intraocular pressure or refractive error of eye into the linear mixed model may not improve its performance.

Covariates	Adaptive Selection	AIC	BIC
Intercept	1.391	−2.064	−2.469
age at diagnosis of diabetes (years)	0.016	—	—
gender (male, female)	—	—	—
doses of insulin per day (1, 2, 3)	—	—	—
residence (urban, rural)	—	—	—
duration of diabetes (years)	0.054	0.090	0.101
glycosylated hemoglobin level	1.259	1.403	—
systolic blood pressure	—	—	—
diastolic blood pressure	0.051	0.077	0.087
body mass index	0.108	0.213	0.271
pulse rate (beats/30 seconds)	—	—	—
proteinuria (absent, present)	0.212	—	—
refractive error	—	—	—
intraocular pressure	—	—	—

Table 5.1: Estimated coefficients of the selected models by AIC, BIC, and adaptive model selection procedure in Wisconsin Epidemiologic Study of Diabetic Retinopathy.

Chapter 6

Conclusion

This dissertation firstly develops the generalized degrees of freedom for linear mixed models and derives data perturbation estimation procedure of the generalized degrees of freedom of linear mixed models. As an excellent model complexity measurement of linear mixed models, the generalized degrees of freedom constructs the foundation where the adaptive model selection procedure is built upon. Most importantly, this dissertation proposes an adaptive model selection procedure for selecting linear mixed models. We theoretically show that the adaptive model selection procedure approximates the best performance of nonadaptive alternatives within the class (1.3) in selecting linear mixed models. The theorems ensure the remarkable asymptotic behavior of adaptive model selection. We also examine the performance of proposed model selection procedure in all three aspects, including covariate selection, variance structure selection and correlation

structure selection, and conclude it performs well in terms of Kullback-Leibler loss against the AIC, the BIC, and other information criterion with form (1.3). As seen from the simulation and theoretical results, the adaptive model selection procedure has advantages over their nonadaptive counterparts in the setting of linear mixed models. These results suggest that the adaptive model selection procedure should also perform well in large-sample situations.

Bibliography

[1] Akaike, H. (1973), "Information Theory and the Maximum Likelihood Principle," in *International Symposium on Information Theory*, eds. V. Petrov and F. Csáki, Budapest: Akademiai Kiádo.

[2] Breiman, L. (1992), "The Little Bootstrap and OtherMethods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

[3] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd Edition, Holden-Day, San Francisco.

[4] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, 2nd Edition, Oxford, U.K.: Oxford University Press.

[5] George, E. I., and Foster, D. P. (2000), " Calibration and Emprirical Bayes Variable Selection ", *Biometrika*, 87, 731–747.

[6] George, E. I., and Foster, D. P. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

[7] Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

[8] Hurvich, C. M., and Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

[9] Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., and DeMets, D. L. (1984). "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years," *Archives of Ophthalmology*, 102, 520–526.

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

[11] Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag.

[12] Potthoff, R. F. and Roy, S. N. (1964), "A generalized multivariate analysis of variance model useful especially for growth curve problems," Biometrika, 51, 313–326.

[13] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

[14] Shen, X., and Huang, H.(2006), "Optimal Model Assessment, Selection, and Combination," *Journal of the American Statistical Association*, 102, 554–568.

[15] Shen, X., Huang, H. and and Ye, J. (2004), "Adaptive Model Selection and Assessment for Exponential Family," *Techinometrics*, 46, 306–317.

[16] Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221.

[17] Tibshirani, R., and Knight, K. (1999), "The Covariance In. ation Criterion for Model Selection," *Journal of the Royal Statistical Society, Ser. B*, 61, 529–546.

[18] Weisberg, S. (2005). *Applied Linear Regression*, 3rd Edition, Wiley/Interscience, New York.

[19] Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131.

Appendix A

Generalized Degrees of Freedom of Linear Mixed Models

In this section, we sketch the technical details of deriving generalized degrees of freedom of linear mixed models.

In (1.1), a linear mixed model expresses response vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_j})^T$ as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, 2, \cdots, m, \tag{A.1}$$

$$\boldsymbol{b}_i \sim N(0, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2\boldsymbol{\Lambda}_i).$$

It is reparameterized as $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2, \cdots, m.$ in (2.7). Then, the individual Kullback-Leibler loss that measures the deviation of the estimated likelihood

$p(\boldsymbol{y}_i|\,\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$ from the true likelihood $p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is

$$
\begin{aligned}
& K(p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\,\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)) \\
= & \int p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \log \frac{p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{p(\boldsymbol{y}_i|\,\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i))} d\boldsymbol{y}_i \\
= & \log |\boldsymbol{\Sigma}_i|^{-1/2} - \log |\widehat{\boldsymbol{\Sigma}}_i|^{-1/2} + \frac{1}{2}\widehat{\boldsymbol{\mu}}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i - \frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \\
& + \frac{1}{2}\boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i + \frac{1}{2}\mathrm{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1}) - \frac{1}{2} \int p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \boldsymbol{y}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i d\boldsymbol{y}_i.
\end{aligned}
$$

The constant terms related to only $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are independent of $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$. Hence for the purpose of comparison, they can be dropped from Kullback-Leibler loss $KL(p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\,\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i))$. Now,

$$
\begin{aligned}
& KL(p(\boldsymbol{y}_i|\,\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\,\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)) \\
= & \log |\widehat{\boldsymbol{\Sigma}}_i|^{-1/2} + \frac{1}{2}\widehat{\boldsymbol{\mu}}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i + \frac{1}{2}\boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i \qquad \text{(A.2)} \\
& + \frac{1}{2}\mathrm{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1}),
\end{aligned}
$$

where (A.2) is defined as the individual comparative Kullback-Leibler loss of the ith

subject. This yields the comparative Kullback-Leibler loss of all independent subjects:

$$\begin{aligned} KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) = & \sum_{i=1}^{m} K(p(\boldsymbol{y}_i|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), p(\boldsymbol{y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)) \\ = & -\log \sum_{i=1}^{m} p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\mu}}_i \\ & -\frac{1}{2} \sum_{i=1}^{m} \boldsymbol{Y}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i + \frac{1}{2} \boldsymbol{\mu}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\mu}_i + \frac{1}{2} \mathrm{Trace}(\boldsymbol{\Sigma}_i \widehat{\boldsymbol{\Sigma}}_i^{-1}). \end{aligned}$$

Consider a class of loss estimators of the form

$$-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \kappa \,. \tag{A.3}$$

In order to access the optimal estimation of $KL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$, we choose to minimize the criterion

$$E\left[KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) - \left(-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \kappa \right) \right]^2,$$

which is the expected L_2 distance between $KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ and the class (A.3) of loss estimators. Note that

$$E\left[KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) - \left(-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i|\, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) + \kappa \right) \right]^2$$

$$
\begin{aligned}
= \quad & E\left[\sum_{i=1}^{m}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\widehat{\boldsymbol{\mu}}_i-\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{Y}_i^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{Y}_i+\frac{1}{2}\boldsymbol{\mu}_i^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\mu}_i\right. \\
& \left.+\frac{1}{2}\mathrm{Trace}(\boldsymbol{\Sigma}_i\widehat{\boldsymbol{\Sigma}}_i^{-1})-\kappa\right]^2.
\end{aligned}
$$

Therefore, we obtain the optimal κ by minimizing this with respect to κ as

$$
\begin{aligned}
\kappa = \quad & E\left[\sum_{i=1}^{m}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\widehat{\boldsymbol{\mu}}_i\right]-\frac{1}{2}E\left[\sum_{i=1}^{m}\boldsymbol{Y}_i^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{Y}_i\right] \\
& +\frac{1}{2}\left[E\boldsymbol{\mu}_i^T\widehat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\mu}_i\right]+\frac{1}{2}E\left[\ \mathrm{Trace}(\boldsymbol{\Sigma}_i\widehat{\boldsymbol{\Sigma}}_i^{-1})\ \right],
\end{aligned}
\tag{A.4}
$$

which is desired generalized degrees of freedom in Section 2.1.

Appendix B

Proof of Theorem 1

Before we show the proof of Theorem 1, a lemma is presented. The proof of the lemma is straightforward by the definition of the generalized degrees of freedom and the Kullback-Leilber loss of linear mixed models.

Lemma 1 (Unbiased estimator of loss) For any $\lambda \in (0, \infty)$,

$$E\left[KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})\right] = E\left[-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \, \widehat{\boldsymbol{\mu}}_{i,\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{i,\widehat{M}(\lambda)}) + GDF(\widehat{M}(\lambda))\right].$$

Now we present the proof of Theorem 1 as follows: in (2.13), we proposed data perturbation estimator $\widehat{GDF}(\widehat{M}_\lambda)$ of the generalized degrees of freedom $\widehat{GDF}(\widehat{M}_\lambda)$ of linear mixed model $\widehat{M}_\lambda$ for any $\lambda \in (0, \infty)$. By (2.13), and assumptions (1), (3) in

Theorem 1, we have that, for any $\lambda \in (0, \infty)$,

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+} E\left[\widehat{GDF}(\widehat{M}_\lambda)\right]$$

$$= \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\tau^{-2}\mathrm{cov}^*(\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y}^*), Y_{ik}^*))\right]$$

$$+ \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\frac{\widehat{\mathrm{var}}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}\mathrm{cov}^*(\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*), Y_{ij}^*Y_{ik}^*)\right]$$

$$= \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\frac{\mathrm{var}(Y_{ik})}{\mathrm{var}^*(Y_{ik}^*)}E^*\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y}^*)(Y_{ik}^* - E^*Y_{ik}^*)\right]$$

$$+ \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\frac{\mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}\mathrm{cov}^*(\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*), Y_{ij}^*Y_{ik}^*)\right]$$

$$+ \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\frac{\widehat{\mathrm{var}}(Y_{ij}Y_{ik}) - \mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}\mathrm{cov}^*(\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*), Y_{ij}^*Y_{ik}^*)\right]$$

$$= \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\lim_{\tau\to 0^+} E\left[\frac{\mathrm{var}(Y_{ik})}{\mathrm{var}^*(Y_{ik}^*)}E^*\frac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y}^*))\mathrm{var}^*(Y_{ik}^*)\right]$$

$$+ \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\lim_{\tau\to 0^+} E\left[\frac{\mathrm{var}(Y_{ij}Y_{ik})}{\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)}E^*\frac{\partial}{\partial Y_{ij}^*Y_{ik}^*}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\mathrm{var}^*(Y_{ij}^*Y_{ik}^*)\right]$$

$$
\begin{aligned}
&= \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\mathrm{var}(Y_{ik})\left(E^*\mathrm{var}(Y_{ik}^*)\frac{\partial}{\partial Y_{ik}^*}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y}^*)\right)\right] \\
&\quad + \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\mathrm{var}(Y_{ij}Y_{ik})\left(E^*\mathrm{var}(Y_{ij}^*Y_{ik}^*)\frac{\partial}{\partial Y_{ij}^*Y_{ik}^*}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y}^*)\right)\right] \\
&= \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\mathrm{var}(Y_{ik})\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y})\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y})\right] \\
&\quad + \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\left[\mathrm{var}(Y_{ij}Y_{ik})\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y})\right] \\
&= \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\frac{\partial}{\partial Y_{ik}}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y})\widehat{\mu}_{ij,\lambda}(\boldsymbol{Y})(Y_{ik}-EY_{ik}) \\
&\quad + \lim_{m,n_i\to\infty}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E\frac{\partial}{\partial Y_{ij}Y_{ik}}\widehat{\sigma}_{ijk,\lambda}(\boldsymbol{Y})(Y_{ij}Y_{ik}-EY_{ij}Y_{ik}) \\
&= \lim_{m,n_i\to\infty} GDF(\widehat{M}_\lambda).
\end{aligned}
$$

That is, for any $\lambda \in (0,\infty)$,

$$
\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+} E\left[\widehat{GDF}(\widehat{M}_\lambda)\right] = \lim_{m,n_i\to\infty} GDF(\widehat{M}_\lambda). \tag{B.1}
$$

Suppose λ_{opt} minimizes $KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})$ with respect to $\lambda \in (0, \infty)$. By Lemma 1 and the definition of λ_{opt}, we have

$$\lim_{m,n_i\to\infty} \lim_{\tau\to 0^+} E(KL(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})}, \widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})}))$$

$$= \lim_{m,n_i\to\infty} \lim_{\tau\to 0^+} E\left[-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \widehat{\boldsymbol{\mu}}_{i,\widehat{M}(\hat{\lambda})}, \widehat{\boldsymbol{\Sigma}}_{i,\widehat{M}(\hat{\lambda})}) + GDF(\widehat{M}(\hat{\lambda}))\right]$$

$$= \lim_{m,n_i\to\infty} \lim_{\tau\to 0^+} E\Bigg[-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \widehat{\boldsymbol{\mu}}_{i,\widehat{M}(\hat{\lambda})}, \widehat{\boldsymbol{\Sigma}}_{i,\widehat{M}(\hat{\lambda})}) + \widehat{GDF}(\widehat{M}(\hat{\lambda}))$$

$$-\widehat{GDF}(\widehat{M}(\hat{\lambda})) + GDF(\widehat{M}(\hat{\lambda}))\Bigg]$$

$$\leq \lim_{m,n_i\to\infty} \lim_{\tau\to 0^+} E\Bigg[-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \widehat{\boldsymbol{\mu}}_{i,\widehat{M}(\lambda_{opt})}, \widehat{\boldsymbol{\Sigma}}_{i,\widehat{M}(\lambda_{opt})}) + \widehat{GDF}(\widehat{M}(\lambda_{opt}))$$

$$-\widehat{GDF}(\widehat{M}(\hat{\lambda})) + GDF(\widehat{M}(\hat{\lambda}))\Bigg]$$

$$= \lim_{m,n_i\to\infty} \lim_{\tau\to 0^+} E\Bigg[-\sum_{i=1}^{m} \log p(\boldsymbol{Y}_i | \widehat{\boldsymbol{\mu}}_{i,\widehat{M}(\lambda_{opt})}, \widehat{\boldsymbol{\Sigma}}_{i,\widehat{M}(\lambda_{opt})}) + GDF(\widehat{M}(\lambda_{opt}))$$

$$-GDF(\widehat{M}(\lambda_{opt})) + \widehat{GDF}(\widehat{M}(\lambda_{opt})) - \widehat{GDF}(\widehat{M}(\hat{\lambda})) + GDF(\widehat{M}(\hat{\lambda}))\Bigg]$$

$$= \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+} E(KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda_{opt})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda_{opt})})) + \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}(\widehat{EGDF}(\widehat{M}(\lambda_{opt}))$$

$$-GDF(\widehat{M}(\lambda_{opt}))) + \lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}(\widehat{EGDF}(\widehat{M}(\hat{\lambda})) - GDF(\widehat{M}(\hat{\lambda})))$$

By (B.1), $\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}(\widehat{EGDF}(\widehat{M}(\lambda_{opt})) - GDF(\widehat{M}(\lambda_{opt}))) = 0$, and $\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}(E\widehat{GDF}(\widehat{M}(\hat{\lambda})) - GDF(\widehat{M}(\hat{\lambda})) = 0$. Therefore

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+} E(KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})})) =$$

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+} E(KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda_{opt})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda_{opt})})),$$

which implies

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\frac{E(KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})}))}{\inf_{\lambda\in(0,\infty)} E(KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)}))} = 1. \tag{B.2}$$

This completes the proof.

Appendix C

Proof of Theorem 2

With the proof of Theorem 1 and assumption (4) in Theorem 1, it can be further concluded that

$$\lim_{m,n_i\to\infty}\lim_{\tau\to 0^+}\frac{KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\hat{\lambda})},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\hat{\lambda})})}{\inf_{\lambda\in\Lambda} KL(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{\boldsymbol{\mu}}_{\widehat{M}(\lambda)},\widehat{\boldsymbol{\Sigma}}_{\widehat{M}(\lambda)})}=1. \tag{C.1}$$