MODIFIED BIC FOR MODEL SELECTION IN LINEAR MIXED MODELS

THI HANG LAI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

August 2022

Abstract

Linear mixed effects models are widely used in applications to analyze clustered and longitudinal data. Model selection in linear mixed models is more challenging than that of linear models as the parameter vector in a linear mixed model includes both fixed effects and variance components parameters. When selecting the variance components of the random effects, the variance of the random effects must be non-negative and therefore, parameters may lie on the boundary of the parameter space. In this dissertation, we propose a modified BIC for model selection with linear mixed effects models that can solve the case when the variance components are on the boundary of the parameter space. We first derive a modified BIC to choose random effects assuming that the random effects are independent. Then, we propose a modified BIC to choose random effects when random effects are assumed to be correlated. Lastly, we propose a modified BIC to choose both fixed effects and random effects simultaneously. Through the simulation results, we found that the modified BIC performs well and performs better than the regular BIC in most cases.

The modified BIC is also applied to a real data set to choose the most appropriate linear mixed model.

Acknowledgements

I would like to express my greatest appreciation to Professor Xin Gao, my supervisor, for her excellent guidance, caring, and encouragement. Her patience and support helped me overcome many challenges to complete this dissertation.

I am also very appreciative of Professor Yuehua (Amy) Wu and Professor Dong Liang for being members of my supervisory committee and for their great support and advise. I much appreciate all your valuable comments which have helped me improve my dissertation. Additionally, I would like to thank Professor Gene Cheung and Professor Chen Xu for taking time out of their summer to review this dissertation and take part in the dissertation examining committee.

I would like to extend my thanks to the professors and staff in the Department of Mathematics and Statistics at York University for their teaching, support, and for giving me a chance to pursue my dream of obtaining a PhD. My thanks to my fellow friends there for their support and encouragement.

I deeply thank my employers and colleagues for their continuous support.

Finally, I would like to thank my parents, my family, my husband and my children for their constant support and encouragement.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Introduction and Objective

With the development of technology in recent times, more complex, and large datasets have become available. Statisticians and researchers are also developing different statistical models to extract valuable information from data to aid decision-making processes. Classical multiple linear regression models can be used to model the relationship between variables. However, one of the assumptions of linear regression is that the errors are independent. Therefore, when the observations are correlated as with longitudinal data, clustered data, and hierarchical data, linear regression models are no longer appropriate. A more powerful class of models used to model correlated data are mixed-effects models, which have been used in many fields of applications.

Correlation between observations may appear when data is collected hierarchically, for example, students may be sampled from the same school, and schools may

be sampled within the same district. Consequently, students in the same school have the same teachers and school environment and therefore, observations are not independent of one another. Observations may be taken from members of the same family, where each family is considered a group or a cluster. As the observations are dependent, we can consider this clustered data. Another type of correlated data pertains to when observations from the same subjects are collected over time, such as repeated blood pressure measurements over a patient's treatment period. This is an example of longitudinal data. Patients (or subjects) may vary in the number and date of the collected measurements. Since observations are recorded from the same individual over time, it is reasonable to assume that subject-specific correlations exist in the trend of the response variable over time. We wish to model the pattern of the response variable over time within subjects, and the variation in the time trends between subjects.

Linear mixed-effects models are used to model correlated data, accounting for the variability within and between clusters in clustered data, or the variability within and between repeated measurements in longitudinal data. Model selection is an important procedure in statistical analysis, allowing the most appropriate model to be chosen from a set of potential candidate models. A desired model is one that can adequately fit the data and not too complex in order to improve two important aspects:

interpretability and predictability. In linear mixed models, identifying significant random effects that should be included in mixed effects model is a challenging step in model selection, since it involves conducting a hypothesis test for whether or not the variance components of random effects are equal to zero. For example, we want to test $H_0 : \sigma^2 = 0$ against $H_1 : \sigma^2 > 0$, where the parameter space of σ^2 is $[0, \infty)$. Under the null hypothesis, the testing value of variance component parameter lies on the boundary of the parameter space. This violates one of the classical regularity conditions that the true value of the parameter must be an interior point of the parameter space. Therefore, classical hypothesis tests such as the likelihood ratio test, the score test, and the Wald tests are no longer appropriate. We call this a boundary issue. When the boundary issue happens, the asymptotic null distribution of the likelihood ratio test statistic is not a chi-square distribution. Chernoff (1954), Self and Liang (1987), Stram and Lee (1994), Azadbakhsh et al. (2021), and Baey et al. (2019) pointed out that, under some conditions on the parameter space and the likelihood functions, the null asymptotic distribution of the likelihood ratio test statistic is a mixture of chi-square distributions. For instance, the asymptotic null distribution of the likelihood ration test statistic for testing $H_0 : \sigma^2 = 0$ against $H_1 : \sigma^2 > 0$ is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, not χ_1^2 ((Stram and Lee, 1994)). Shapiro (1985) provided the expressions to calculate the exact weights used in the mixture of chi-square dis-

tributions for some special cases. However, in general, determining the exact weights used in the mixture of chi-squared distributions is difficult when the number of the variance components being tested under the null hypothesis is large as the weights are not available in a tractable form (Baey et al. (2019)).

There are a number of information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), that have been developed for model selection with linear mixed models by Vaida and Blanchard (2005), Pauler (1998), Jones (2011), and Delattre and Poursat (2020). Other methods, including shrinkage and permutation methods for identifying the important fixed effects and random effects variance components, have been considered as in Ibrahim et al. (2011), Bondell et al. (2010), Peng and Lu (2012), and Drikvandi et al. (2012).

The boundary issue has impacted the BIC. If we use regular BIC in linear mixed models, that is, we treat this case as if there were no constraints on the model's parameter vector, then the penalty term of the regular BIC would include all number of components of the parameter vector. Therefore, the regular BIC would overestimate the number of degrees of freedom of the linear mixed model (called model complexity in the thesis) and would not take into account the fact that variances components are constrained and bounded below by 0. Consequently, the regular BIC tends to choose under-fitted linear mixed models.

Several versions of modified BIC have been proposed for model selection in linear mixed models as in (Jones, 2011; Pauler, 1998; Pauler et al., 1999). However, to our knowledge, none of the current BIC can directly deal with the boundary issue.

Therefore, the main objective of this dissertation is to introduce a modified BIC for model selection when the true value of the variance components parameters lie on the boundary of the parameter space, allowing the most appropriate model to be chosen from a set of candidate linear mixed models.

Here is the general idea on how our proposed method solves the boundary problem. From the previous literature, we know that the asymptotic null distribution of the likelihood ratio test statistic of testing the nullity of several variances is a chi-bar square distribution (Baey et al. (2019)). Based on this theoretical result, we take the average of the chi-bar square distribution and include this average in the complexity of the model. When random effects are correlated, the tricky problem is in calculating the weights of the chi-bar square distribution. The weights depend on a cone C^* that contains the set of positive definite matrices. Describing the set of positive definite matrices explicitly using constraints on the components of random effects covariance matrix, D, is almost impossible. Thus, calculating the weights of the chi-bar square distribution for this case is not an easy task and has not been addressed in the literature. Our solution to this problem is: we bound cone C^* by a

bigger cone. The bigger cone has a much simpler structure and allows us to calculate the weights of the chi-bar distribution.

1.2 Outline of the dissertation

This thesis consists of two projects. The first project summarizes the "Modified BIC for Model Selection in Linear Mixed Models" and the second project is an application of "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques".

In Chapter 1 of this dissertation, we provide an introduction to this dissertation. After introducing the objective of the dissertation research, we review the linear mixed models, previous research on model selection with linear mixed models, and the hypothesis tests on random effects variances.

In Chapter 2, we introduce a modified BIC to choose random effects components in a linear mixed model when random effects are assumed to be independent and prove the model selection consistency of the modified BIC.

In Chapter 3, we introduce a modified BIC to choose random effects components in a linear mixed model when random effects are assumed to be correlated.

In Chapter 4, we propose a modified BIC to choose both fixed effects and random effects components in a linear mixed model where the random effects are assumed

to be independent or correlated.

In Chapter 5, we present our second project which is "Predictive Models for Diabetes Mellitus using Machine Learning Techniques". In this project, we propose predictive models utilizing Gradient Boosting Machine and Logistic Regression techniques to predict the probability of patients having Diabetes Mellitus based on their demographic information and laboratory results collected from their visits to medical facilities.

In Chapter 6, we summarize the contributions of this thesis based on the proposed modified BIC for linear mixed models that can deal with the boundary problem. A discussion on potential future work follows.

1.3 Literature Review

In this section, we provide a brief overview of linear mixed models, hypothesis testing under non-standard conditions, and review several model selection methods from the literature, which are applied to linear mixed models based on information criteria such as the AIC and BIC.

1.3.1 Linear Mixed Models

Consider the linear mixed model introduced in Laird and Ware (1982):

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{1.1}$$

for $i = 1, \ldots, N$, where $\boldsymbol{y}_i$ denotes the n_i-dimensional vector of response measurements for cluster i with $i = 1, \ldots, N$; $\boldsymbol{\beta}$ is a $p \times 1$ fixed effect parameter vector; $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of covariates for the fixed effects; $\boldsymbol{Z}_i$ is a $n_i \times q$ matrix of covariates for the random effects; and $\boldsymbol{b}_i$ denotes the random effects vector of the ith cluster; $\boldsymbol{b}_i$ is assumed to follow a multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a $q \times q$ covariance matrix. $\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}$ are assumed to be independent. Fixed effects are used to model the population mean, while random effects are used to model between-cluster variation in the response. $\boldsymbol{\epsilon}_i$ is the vector of random errors and is assumed to follow a multivariate normal distribution, $N(0, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ denotes the $n_i \times n_i$ identity matrix. $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are assumed to be pairwise independent for $i = 1, \ldots, N$. The marginal distribution of $\boldsymbol{y}_i$ is $N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$ where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_{n_i}$.

Model (1.1) has broad application in many fields such as medical studies and educational studies. For example, we look at the data set "Orthodont" in the nlme package. The dental data set was introduced by Potthoff and Roy (1964), where dental measurements were made on 11 girls and 16 boys at ages 8, 10, 12 and 14. The response variable was the distance, in millimeters, from the center of pituitary

to the pterygomaxillary fissure. There are 27 subjects in the data set with the following variables: **Distance** is a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). These distances are measured on x-ray images of the skull. **Age** is a numeric vector of ages of the subject (in years). **Subject** is an ordered factor indicating the subject on which the measurement was made. **Sex** is a factor with levels Male and Female. The objective is to study the change in an orthodontic measurement over time for young boys and girls. This is an example of longitudinal data. Each boy or girl can be considered as a subject or cluster. There are 27 independent subjects or clusters. Each subject was measured 4 times. Thus, the measurements within each subject are correlated. We can use model (1.1) to analyze this data set. Fixed effects can be used to model the relationship between Distance and Age for boys and girls. The regression coefficients, $\boldsymbol{\beta}$, measures the population average intercept and the population average slope meanwhile the random effects $\boldsymbol{b_i} = (b_{i0}, b_{i1})$ are the effects in intercept and slope associated with subject i. The subject intercept measures the deviation of that subject intercept from the population average intercept. We assume that $\boldsymbol{b_1}, \ldots, \boldsymbol{b_{27}}$ are independent and each has a bivariate normal distribution $N(\mathbf{0}, \boldsymbol{D})$. The covariance matrix $\boldsymbol{D}$ of random effects captures the variation between subjects.

We can also write model (1.1) in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{y} = (\boldsymbol{y_1}^T, \ldots, \boldsymbol{y_N}^T)^T$ is an $n \times 1$ vector; $n = \sum_{i=1}^{N} n_i$; and $\boldsymbol{X} = (\boldsymbol{X_1}^T, \ldots, \boldsymbol{X_N}^T)^T$ is an $n \times p$ matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector; $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_N})$ is an $n \times Nq$ matrix; $\boldsymbol{b} = (\boldsymbol{b_1}^T, \ldots, \boldsymbol{b_N}^T)^T$ is an $Nq \times 1$ vector; and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_N^T)^T$ is an $n \times 1$ vector. Let $\boldsymbol{\tau}$ be the vector of distinct variance and covariance components in matrix $\boldsymbol{D}$. Let $\boldsymbol{\eta}$ be $(\boldsymbol{\tau}^T, \sigma_{\boldsymbol{\epsilon}}^2)^T$. Then the vector of parameters for this model is $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$. We assume that the response vectors $\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}$ from N clusters are independent random observations. Given a clustered data set, we wish to choose a linear mixed model that yields fits the data well and is also a parsimonious model.

For the linear mixed model (1.1), the marginal log-likelihood function is:

$$l_N(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \left[-\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{V}_i| - \frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right]. \tag{1.3}$$

If the random effects components are independent, then $\boldsymbol{D}$ is a diagonal matrix. If the random effects components are correlated, $\boldsymbol{D}$ is a full matrix. For our linear mixed model (1.1), the information matrix is a block diagonal matrix (Searle, 1970).

$$\boldsymbol{I}_N(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\eta} \\ \boldsymbol{I}_{\beta\eta}^T & \boldsymbol{I}_{\eta\eta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\eta\eta} \end{bmatrix}. \tag{1.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$; $\boldsymbol{\eta}$ is the vector containing all different variances and covariances in $\boldsymbol{D}$ and $\sigma_{\boldsymbol{\epsilon}}^2$ and $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n = \text{diag}(\boldsymbol{V_1}, \boldsymbol{V_2}, \ldots, \boldsymbol{V_N})$ with

$\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_{n_i}$; $n = \sum_{i=1}^{N} n_i$ and $\boldsymbol{G} = \text{diag}(\boldsymbol{D}, \ldots, \boldsymbol{D})$. The st-th element of $\boldsymbol{I}_{\eta\eta}$ is $\frac{1}{2}tr(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{\eta}_s}\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{\eta}_t})$ where $tr(\boldsymbol{A})$ denotes the trace of matrix $\boldsymbol{A}$, for any matrix $\boldsymbol{A}$. When $\boldsymbol{V}$ is unknown, $\boldsymbol{V}$ can be approximated by $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$ and the partial derivatives are evaluated at $\hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the full maximum likelihood estimator or restricted maximum likelihood estimator of $\boldsymbol{\theta}$.

The parameters in linear mixed models are usually estimated using the maximum likelihood (ML) or restricted maximum likelihood (REML) method. First, for fixed $\boldsymbol{\eta}$, the log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{y})$ is maximized over $\boldsymbol{\beta}$ by the generalized least squares estimator:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{1.5}$$

Then, replace $\boldsymbol{\beta}$ in $l(\boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{y})$ by $\hat{\boldsymbol{\beta}}(\boldsymbol{\eta})$ to obtain the profile log-likelihood function, $l_p(\boldsymbol{\eta}) = l(\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}), \boldsymbol{\eta})$, which is a function of $\boldsymbol{\eta}$, and

$$l_p(\boldsymbol{\eta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\{\log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\}. \tag{1.6}$$

Maximizing $l_p(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ gives the maximum likelihood estimator $\hat{\boldsymbol{\eta}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}_{ML})$, which is the maximum likelihood estimator of $\boldsymbol{\beta}$. The maximum likelihood estimator $\hat{\boldsymbol{\eta}}_{ML}$ has been found to be biased downwards because when the variance components are estimated the maximum likelihood method does not take into account the degrees of freedom lost by estimating the fixed effects (Harville,

1974). Therefore, the restricted maximum likelihood estimator of $\boldsymbol{\eta}$ is often used to produce unbiased estimates of variance and covariance parameters.

The restricted maximum log-likelihood, $l_R(\boldsymbol{\eta})$, is obtained by integrating out the fixed effect parameter $\boldsymbol{\beta}$ from the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{y})$ which is the joint distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, and then taking the natural logarithm. The following derivation is from Czado (2017). We have

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{y} \\ &= (\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y}) + \boldsymbol{y}^T [\boldsymbol{V}^{-1} - \boldsymbol{B}^T \boldsymbol{A}\boldsymbol{B}] \boldsymbol{y},\end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}$ and $\boldsymbol{B} = \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}$.

Therefore,

$$\begin{aligned}&\int L(\boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{y}) d\boldsymbol{\beta} \\ &\quad = \int (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\{(\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y}) + \boldsymbol{y}^T (\boldsymbol{V}^{-1} - \boldsymbol{B}^T \boldsymbol{A}\boldsymbol{B}) \boldsymbol{y}\}\} d\boldsymbol{\beta} \\ &\quad = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \boldsymbol{y}^T (\boldsymbol{V}^{-1} - \boldsymbol{B}^T \boldsymbol{A}\boldsymbol{B}) \boldsymbol{y}\} \int \exp\{-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{y})\} d\boldsymbol{\beta} \\ &\quad = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \boldsymbol{y}^T (\boldsymbol{V}^{-1} - \boldsymbol{B}^T \boldsymbol{A}\boldsymbol{B}) \boldsymbol{y}\} (2\pi)^{\frac{p}{2}} |\boldsymbol{A}|^{-\frac{1}{2}} \\ &\quad = (2\pi)^{-\frac{n-p}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \boldsymbol{y}^T \boldsymbol{P}\boldsymbol{y}\} |\boldsymbol{A}|^{-\frac{1}{2}},\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}$. We also have

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\hat{\boldsymbol{\beta}} - 2\boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{y} = \boldsymbol{y}^T \boldsymbol{P}\boldsymbol{y},$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$. Hence, the restricted maximum log-likelihood function is

$$l_R(\boldsymbol{\eta}) = -\frac{1}{2}\{\log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\} - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{n-p}{2}\log 2\pi.$$

Thus,

$$l_R(\boldsymbol{\eta}) = l_p(\boldsymbol{\eta}) - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| + C, \tag{1.7}$$

where C is a constant term which does not depend on parameters.

Maximizing $l_R(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ gives the restricted maximum likelihood estimator $\hat{\boldsymbol{\eta}}_R$ and the restricted maximum likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}_R)$.

The conditional log-likelihood of $\boldsymbol{y}$ given b is

$$l(\boldsymbol{\theta}; \boldsymbol{y} \mid \boldsymbol{b}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2_{\boldsymbol{\epsilon}} - \frac{1}{2\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}). \tag{1.8}$$

1.3.2 Review of Model Selection for Linear Mixed Models

In this subsection we will provide a brief review of previous work on model selection for linear mixed models. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) play an important role in model selection. AIC is an asymptotically unbiased estimator of Akaike Information (AI) which is based on

Kullback-Leibler discrepancy between the true model and the approximating model. The Akaike information criterion or AIC (Akaike, 1974) is:

$$AIC = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + 2p, \tag{1.9}$$

where $l(\hat{\boldsymbol{\theta}}; \boldsymbol{y})$ is the value of the log-likelihood function evaluated at $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimator of $\boldsymbol{\theta}$; and p is the number of parameters in the model. Out of the set of candidate models, the model with the minimum AIC is chosen.

The Bayesian information criterion (BIC) provides a large-sample estimator of a transformation for the Bayesian posterior probability associated with the approximating model. The Bayesian information criterion or BIC (Schwarz, 1978) is:

$$BIC = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + p\log(n), \tag{1.10}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, p is the number of parameters in the model, and n is the number of observations. This criterion includes two terms: the first term measures the discrepancy of the model fitting and the second term is the penalty for the model complexity. The BIC is used to compare and select models, where the selected model is the one that minimizes the BIC. Under some conditions, Schwarz (1978) proved that the BIC is consistent. That is, if the set of candidate models contains the true model with parameter $\boldsymbol{\theta}_T$, then as n becomes large, with probability approaching 1, the BIC will select the model of

lowest dimension, containing $\boldsymbol{\theta}_T$. In multiple regression models, Rao and Wu (1989) introduced a modified criterion with a flexible penalty function and demonstrated its consistency property without making any distributional assumptions on the error term of a multiple regression model.

There are also a number of information criteria developed for linear mixed models in the literature. Vaida and Blanchard (2005) introduced the marginal AIC (based on the marginal likelihood),

$$mAIC = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + 2a_n(p+q), \tag{1.11}$$

where $a_n = 1$ for the asymptotic form or $a_n = \frac{n}{n-p-q-1}$ for the finite sample form, p is the number of parameters of $\boldsymbol{\beta}$, and q is the number of variance and covariance parameters of $\boldsymbol{\tau}$. The author also introduced the conditional AIC (based on the conditional log-likelihood function) of the form,

$$cAIC = -2l(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{b}}) + 2(\rho+1), \tag{1.12}$$

with $\hat{\boldsymbol{b}}$ is the empirical best linear unbiased predictor (EBLUP) of $\boldsymbol{b}$ and $\rho = \text{Tr}(\boldsymbol{H}_1)$ or the trace of the "hat" matrix, $\boldsymbol{H}_1$, such that $\hat{\boldsymbol{y}} = \boldsymbol{H}_1\boldsymbol{y}$. The authors recommend to use $\hat{\rho}$ when the variance and covariance components of random effects are unknown, arguing that the difference between $\hat{\rho}$ and ρ are negligible asymptotically. Liang et al. (2008) propose a corrected version of cAIC that takes into account the estimation of

the variance and covariance components. However, evaluating the bias correction is computationally expensive. Greven and Kneib (2010) develop an analytic version of the corrected cAIC and the bias correction of the corrected cAIC can be calculated. Their method is implemented in the cAIC4 package in R (Säfken et al., 2018). We will use this criterion in chapters 3 and 4.

The formula for the BIC used by the lmer function in the lme4 package in R for linear mixed models is

$$BIC = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + (p + q + 1)\log(n), \tag{1.13}$$

where p is the number of fixed effects parameters; $q + 1$ is the number of distinct parameters in the random effects covariance matrix and the error term variance parameter; and n is the number of observations. Pauler (1998) developed a BIC to select fixed effects parameters, $\boldsymbol{\beta}$, in independent cluster models, making an adjustment to effective sample size.

Pauler et al. (1999) addressed the boundary problems in the variance component model by assuming that the parameter space $\boldsymbol{\Theta}$ can be extended to an open set $\boldsymbol{\Theta}^*$ such that the boundary of $\boldsymbol{\Theta}$ is in the interior of $\boldsymbol{\Theta}^*$. Jones (2011) developed an effective sample size based on the correlation matrix corresponding to the covariance

matrix of the response variable and proposed a BIC of the following form:

$$BIC_J = -2l(\hat{\boldsymbol{\theta}}) + (p + q + 1)\log\{n_e\}. \tag{1.14}$$

where $n_e = \sum_{i=1}^{N}\{\mathbf{1}_{n_i}^T \boldsymbol{C}_i^{-1}\mathbf{1}_{n_i}\}$ is the effective sample size. $\boldsymbol{C}_i$ is the correlation matrix corresponding to $\boldsymbol{V}_i$ which is the covariance matrix of the response variable, $\boldsymbol{y}_i$, for cluster i.

Some other methods have been developed recently to overcome the limitations of mixed-effects models in choosing the random effects. Chen and Dunson (2003) used a hierarchical Bayesian model to identify random effects with zero variance. They reparameterized the mixed model by decomposing the covariance matrix of random effects, $\boldsymbol{D}$ as,

$$\boldsymbol{D} = \boldsymbol{\Lambda\Gamma\Gamma}^T\boldsymbol{\Lambda}, \tag{1.15}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with elements proportional to the standard deviations of the random effects, and Γ is a lower triangular matrix that relates to the correlations among the random effects. The parameters in either $\boldsymbol{\Lambda}$ or $\boldsymbol{\Gamma}$ are considered regression coefficients in a normal linear model. Therefore, the authors used normal priors for these parameters in the Gibbs sampling algorithm in order to compute the posterior distributions. Then, any random effects with zero variance will be identified and dropped out of the model. Saville and Herring (2008) developed a test based on Bayes

factors, however, this test relies on the subjective choice of the prior distribution of parameters. Drikvandi et al. (2012) proposed a permutation test to test any subset of the variance components.

The shrinkage approach to model selection in the linear mixed model is also proposed as in Ibrahim et al. (2011), Bondell et al. (2010) and Peng and Lu (2012) for choosing both fixed effects and random effects. The authors consider model selection for the independent cluster model (1.1). This approach does both model selection and model estimation by maximizing a penalized likelihood function. Ibrahim et al. (2011) proposed joint selection of fixed and random effects by maximizing the following penalized likelihood function when p and q are fixed:

$$\ell(\boldsymbol{\theta}) - N\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|) - N\sum_{k=1}^{q}\phi_{\lambda_{p+k}}(\|\boldsymbol{\gamma}_k\|)$$

where $l(\boldsymbol{\theta})$ is the marginal likelihood function as defined in (1.3); $\boldsymbol{\gamma}_k$ contains all nonzero components of row k of $\boldsymbol{\Gamma}$. Here $\boldsymbol{\Gamma}$ is a $q \times q$ lower triangular matrix and is a factor in the Cholesky parametrization of the random effects covariance matrix $\boldsymbol{D}$. Matrix $\boldsymbol{D}$ is written as $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. The penalty functions can be SCAD (smoothly clipped absolute deviation) function or ALASSO (Adaptive least absolute shrinkage and selection operator) function. The ALASSO penalty functions are defined as:

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

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

for fixed effect and random effects, respectively. Here, $\hat{\beta}_j$ and $\hat{\boldsymbol{\gamma}}_k$ are the unpenalized maximum likelihood estimators.

The tuning parameters, $\boldsymbol{\lambda_j}$ for $j = 1, \ldots, p$ and $\boldsymbol{\lambda_{p+k}}$ for $k = 1, \ldots, q$ are for fixed effects and random effects, respectively. The author considered two tuning constants which are defined as:

$$\lambda_j = \lambda^{(1)}, \qquad j = 1, 2, ..., p$$

$$\lambda_{p+k} = \lambda^{(2)} \sqrt{k}, \qquad k = 1, 2, ..., q.$$

Compared to the Information Criterion approach, one of the advantages of the shrinkage method is that we do not need to consider all possible models. Therefore, the shrinkage method is very helpful when the number of fixed effects and/or random effects is large. However, this shrinkage approach also has problems with boundary issues in linear mixed model. As pointed out in Müller et al. (2013), this shrinkage approach uses the unpenalized maximum likelihood or restricted maximum likelihood estimates as the weights in the ALASSO penalty. When the boundary issue happens, some of the maximum likelihood estimates of variance parameters could be exactly on the zero boundary. In this case, the ALASSO weight is infinity and the optimization

procedure may fail to converge.

Please see Müller et al. (2013) for a thorough review of model selection criteria in linear mixed models. We will propose a modified BIC for model selection in linear mixed models in the following chapters.

1.3.3 Review of Hypothesis Testing with Boundary Problem

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be n independent observations with common density function, $f(\boldsymbol{y}; \boldsymbol{\theta})$, where the parameter vector, $\boldsymbol{\theta}$, takes values in a parameter space, $\boldsymbol{\Theta}$, which is a subset of $\mathbb{R}^m$. Denote by $\boldsymbol{\theta}^*$ the true value of the parameter vector. Consider a hypothesis test, $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ vs. $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, where $\boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1 = \varnothing$, $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta} \subset R^m$. The log-likelihood ratio test statistic is

$$\lambda_n = -2 \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \log L(\boldsymbol{\theta}; \boldsymbol{y}) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log L(\boldsymbol{\theta}; \boldsymbol{y}) \right),$$

where $L(\boldsymbol{\theta}; \boldsymbol{y})$ denotes the likelihood function and $L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta})$. The likelihood ratio statistic can also be written as,

$$\lambda_n = -2 \left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} l(\boldsymbol{\theta}; \boldsymbol{y}) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l(\boldsymbol{\theta}; \boldsymbol{y}) \right),$$

where $l(\boldsymbol{\theta}; \boldsymbol{y})$ is the log-likelihood function.

When $\boldsymbol{\theta}^*$ is an interior point of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_0$ is a r-dimensional subspace of a m-dimensional linear space $\boldsymbol{\Theta}$ with $r < m$, under classical regularity conditions, the null distribution of the likelihood ratio test statistic λ_n is χ^2_{m-r} as $n \to \infty$ where $m - r$ is the difference between the dimensions of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_0$. However, when the true value of the parameter lies on the boundary of the parameter space, the distribution of the likelihood ratio test statistic may no longer follow a chi-square distribution.

It has been proven in some literature that the limiting distribution of the like-

lihood ratio test statistic in the non-standard condition is a mixture of chi-square distributions. Generalizing Chernoff (1954)'s work on the case when $\boldsymbol{\theta}^*$ lies on the boundary of $\boldsymbol{\Theta}$, Self and Liang (1987) showed that, under modified regularity conditions, the likelihood ratio test statistic

$$\lambda_n = -2\left(\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}\log L(\boldsymbol{\theta};\boldsymbol{y}) - \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\log L(\boldsymbol{\theta};\boldsymbol{y})\right),$$

asymptotically has the same distribution as

$$\inf_{\boldsymbol{\theta}\in C_{\boldsymbol{\Theta}_0}-\boldsymbol{\theta}^*}(\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{I}(\boldsymbol{\theta}^*)(\boldsymbol{z}-\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in C_{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*}(\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{I}(\boldsymbol{\theta}^*)(\boldsymbol{z}-\boldsymbol{\theta}), \tag{1.16}$$

where $C_{\boldsymbol{\Theta}_0}$ is the cone approximating the set $\boldsymbol{\Theta}_0$ and $C_{\boldsymbol{\Theta}}$ is the cone approximating the set $\boldsymbol{\Theta}$ at the point $\boldsymbol{\theta}^*$. $C_{\boldsymbol{\Theta}_0}-\boldsymbol{\theta}^*$ and $C_{\boldsymbol{\Theta}_0}-\boldsymbol{\theta}^*$ are translated cones of $C_{\boldsymbol{\Theta}_0}$ and $C_{\boldsymbol{\Theta}}$, respectively, such that their vertices are at the origin. Here $\boldsymbol{z}\sim N_m(\boldsymbol{0},\boldsymbol{I}^{-1}(\boldsymbol{\theta}^*))$ and $\boldsymbol{I}(\boldsymbol{\theta}^*)$ is the Fisher information matrix at $\boldsymbol{\theta}^*$.

Silvapulle and Sen (2005) considered the case when observations are independent but not necessarily identically distributed. Let $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N$ be N independent observations and $\boldsymbol{Y}_i$ has the probability density function $f_i(\boldsymbol{y}_i;\boldsymbol{\theta})$, for $i=1,\ldots,N$. Let $l_N(\boldsymbol{\theta})$ be the log-likelihood function, $\sum_{i=1}^N\log f_i(\boldsymbol{y}_i;\boldsymbol{\theta})$. The parameter vector, $\boldsymbol{\theta}$, takes values in a parameter space, $\boldsymbol{\Theta}$, which is a subset of $\mathbb{R}^m$. Suppose that, for all $\boldsymbol{\theta}$, $N^{-\frac{1}{2}}l_N'(\boldsymbol{\theta})\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\}\xrightarrow{a.s.}\boldsymbol{\nu}(\boldsymbol{\theta})$ for some positive definite matrix $\boldsymbol{\nu}(\boldsymbol{\theta})$ and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$.

Silvapulle and Sen (2005) showed that, under some regularity conditions, the null distribution of the likelihood ratio test statistic

$$\lambda_N = -2\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} l_N(\boldsymbol{\theta};\boldsymbol{y}) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l_N(\boldsymbol{\theta};\boldsymbol{y})\right),$$

is asymptotically the same as the distribution of

$$\inf_{\boldsymbol{\theta}\in T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)} (\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{\nu}(\boldsymbol{\theta}^*)(\boldsymbol{z}-\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)} (\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{\nu}(\boldsymbol{\theta}^*)(\boldsymbol{z}-\boldsymbol{\theta}), \tag{1.17}$$

where $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is the tangent cone of $\boldsymbol{\Theta}_0$ at $\boldsymbol{\theta}^*$, $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ is the tangent cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$; and $\boldsymbol{z} \sim N_m(\boldsymbol{0}, \boldsymbol{\nu}^{-1}(\boldsymbol{\theta}^*))$.

The *tangent cone* of a set A at a point $\boldsymbol{\theta}_0$ in A is denoted by $T_A(\boldsymbol{\theta}_0)$ and is defined in Silvapulle and Sen (2005, Section 4.7.1) as follows:

$$T_A(\boldsymbol{\theta}_0) = \{\boldsymbol{w} : \exists\, t_n \downarrow 0,\ \exists\, \boldsymbol{\theta}_n \in A \text{ such that } \boldsymbol{\theta}_n \to \boldsymbol{\theta}_0 \text{ and } t_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \to \boldsymbol{w}\}.$$

More generally, $T_A(\boldsymbol{\theta}_0)$ is the set of all tangents to the set A at $\boldsymbol{\theta}_0$.

Based on the work in Shapiro (1985), Silvapulle and Sen (2005) present the definition of $\bar{\chi}^2$ and the following results: Let $\mathcal{C} \subset \mathbb{R}^m$ be a closed convex cone and let $\boldsymbol{Z} \sim N_m(\boldsymbol{0}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a positive definite matrix. $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C})$ is a random variable which has the same distribution as $\left[\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})\right]$. So, we write

$$\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) = \boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta}). \tag{1.18}$$

Also, let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^m$ and $\boldsymbol{V}$ be a $m \times m$ positive definite matrix. Then the distribution of $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C})$ is given by,

$$Pr\left(\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \leq c\right) = \sum_{i=0}^{m} w_i\left(m, \boldsymbol{V}, \mathcal{C}\right) Pr(\chi_i^2 \leq c), \tag{1.19}$$

where $w_i\left(m, \boldsymbol{V}, \mathcal{C}\right), i = 0, \ldots, m,$ are some non-negative numbers and $\sum_{i=0}^{m} w_i\left(m, \boldsymbol{V}, \mathcal{C}\right) = 1$. If $\boldsymbol{V} = \boldsymbol{I}$ which is an identity matrix and $\mathcal{C} = \mathbb{R}_+^m$, the nonnegative orthant, then

$$Pr\left(\bar{\chi}^2(\boldsymbol{I}, \mathbb{R}_+^m) \leq c\right) = \sum_{i=0}^{m} 2^{-m} \binom{m}{i} Pr(\chi_i^2 \leq c), \tag{1.20}$$

where $\binom{m}{i}$ is m choose i and χ_i^2 is a chi-squared distribution with i degrees of freedom.

When $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace contained in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ and $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ is a closed convex cone, according to Silvapulle and Sen (2005, Theorem 3.7.1), the distribution of

$$\inf_{\boldsymbol{\theta} \in T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)} (\boldsymbol{z} - \boldsymbol{\theta})^T \boldsymbol{\nu}(\boldsymbol{\theta}^*)(\boldsymbol{z} - \boldsymbol{\theta}) - \inf_{\boldsymbol{\theta} \in T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)} (\boldsymbol{z} - \boldsymbol{\theta})^T \boldsymbol{\nu}(\boldsymbol{\theta}^*)(\boldsymbol{z} - \boldsymbol{\theta})$$

is

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*), \tag{1.21}$$

where $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp}$; $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp}$ is the orthogonal complement of $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ in the sense that $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \{\boldsymbol{y} : \text{ such that } \boldsymbol{x}^T \boldsymbol{\nu}(\boldsymbol{\theta}^*) \boldsymbol{y} = 0 \text{ for all } \boldsymbol{x} \in T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)\}$, and $\boldsymbol{z} \sim N_m(\boldsymbol{0}, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1})$.

Baey et al. (2019) used the results from both Self and Liang (1987) and Silvapulle and Sen (2005) for testing the random effects variance components in linear mixed models and generalized linear mixed models.

1.3.4 Review of Testing Random Effects Variances

In model (1.1), the random effects vector $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \ldots, b_{iq})^T$ is assumed to follow a multivariate normal distribution, $N_q(\boldsymbol{0}, \boldsymbol{D})$. If the random effects components are uncorrelated, then $\boldsymbol{D}$ is a diagonal matrix. If the random effects components are correlated, then $\boldsymbol{D}$ is a full matrix. Let $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ denote the marginal log-likelihood function of the linear mixed model (1.1). $l_N(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \log f_i(\boldsymbol{y}_i; \boldsymbol{\theta})$ where $f_i(\boldsymbol{y}_i; \boldsymbol{\theta})$ is the density function of $\boldsymbol{y}_i$, and,

$$l_N(\boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \quad (1.22)$$

where $\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N)$ and $n = \sum_{i=1}^{N} n_i$.

Stram and Lee (1994) used Self and Liang (1987)'s results in testing for the nonzero variance components of random effects in a linear mixed model using likelihood ratio test statistic, given by,

$$\lambda_N = -2\left(\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} l_N(\boldsymbol{\theta}; \boldsymbol{y}) - \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l_N(\boldsymbol{\theta}; \boldsymbol{y})\right). \quad (1.23)$$

The authors assumed that the true value of $\sigma^2_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\beta}$ lie in the interior of the admissible region of these parameters. The authors, thus, could restrict the geometry

of $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}$ to deal with the variance components parameters in matrix $\boldsymbol{D}$. Let $\boldsymbol{\tau}$ be the vector of distinct variance and covariance components in matrix $\boldsymbol{D}$ and $\dim(\boldsymbol{\tau}) = q$.

Case 1: $q = 1$ and $\boldsymbol{D} = [\sigma_1^2]$. In this hypothesis test, the null hypothesis is, $H_0 : \sigma_1^2 = 0$ and the alternative hypothesis is $H_1 : \sigma_1^2 > 0$. The parameter is $\boldsymbol{\theta} = (\sigma_1^2)$. The parameter spaces are $\boldsymbol{\Theta}_0 = \{0\}$ and $\boldsymbol{\Theta} = [0, \infty)$. The approximating cones at the vertex $\{0\}$ are $A_{\boldsymbol{\Theta}_0}(0) = \{0\}$ and $A_{\boldsymbol{\Theta}}(0) =[0, \infty)$, respectively. Under the null hypothesis, the limiting distribution of the log-likelihood ratio test statistic, λ_N, is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ as N goes to infinity. Please refer to the definition of a approximating cone in the Appendix section 7.

Case 2: $q = 2$ and $\boldsymbol{D} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$ where $\sigma_{12} = \sigma_{21}$. Consider the hypothesis test, $H_0 : \boldsymbol{D} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}$ with positive σ_1^2 against H_1: $\boldsymbol{D}$ is positive semi-definite. The parameter is $\boldsymbol{\theta} = (\sigma_1^2, \sigma_{12}, \sigma_2^2)$. The parameter spaces are $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} : \sigma_1^2 > 0; \sigma_{12} = 0; \sigma_2^2 = 0\}$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta} : \sigma_1^2 > 0; \sigma_1^2\sigma_2^2 - (\sigma_{12})^2 \geq 0; \sigma_2^2 \geq 0\}$. The author argued that since σ_1^2 is assumed to be positive under both H_0 and H_1, the only relevant constraint is $\sigma_1^2\sigma_2^2 - (\sigma_{12})^2 \geq 0$. The approximating cones at the vertex $\{(0,\ 0,\ 0)\}$ are $A_{\boldsymbol{\Theta}_0}(0, 0, 0) = (0, \infty) \times \{0\} \times \{0\}$ and $A_{\boldsymbol{\Theta}}(0, 0, 0) = (0, \infty) \times \{R\} \times [0, \infty)$, respectively. And, under the null hypothesis, the limiting distribution of the log-likelihood ratio test statistic, λ_N, is $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$.

Silvapulle and Sen (2005) also discussed about *Case 2* above in their example 4.8.3. The authors pointed out that, under the null hypothesis, the the limiting distribution of the log-likelihood ratio test statistic, λ_N, is

$$\lambda_N \xrightarrow{d} \|\boldsymbol{U} - T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)\|^2 - \|\boldsymbol{U} - T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)\|^2,$$

where $\boldsymbol{U} \sim N_m(\boldsymbol{0}, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1})$, m is the dimension of θ,

$$\text{and } \|\boldsymbol{U} - T_{\Omega}(\boldsymbol{\theta}^*)\|^2 = \min_{\boldsymbol{\theta} \in T_{\Omega}(\boldsymbol{\theta}^*)} \{(\boldsymbol{U} - \boldsymbol{\theta})^T \boldsymbol{\nu}(\boldsymbol{\theta}^*)(\boldsymbol{U} - \boldsymbol{\theta})\}$$

where $\Omega \in \{\boldsymbol{\Theta}, \boldsymbol{\Theta}_0\}$.

The authors also pointed out that there are two scenarios for case 2 above.

Scenario 1: The null hypothesis is assumed and $\boldsymbol{\theta}^* = (0, 0, 0)^T$. Since $\boldsymbol{\Theta}$ is a closed convex cone with vertex at origin, $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \boldsymbol{\Theta}$. The authors noted that since neither $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ nor $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear space, the asymptotic null distribution of λ_N may not be a chi-bar square distribution.

Scenario 2: The null hypothesis is assumed and $\boldsymbol{\theta}^* = (a, 0, 0)^T$ where a is a positive value. In this case, $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \in \mathbb{R}^3 : \sigma_{12} = 0; \sigma_2^2 = 0\}$ and $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \in \mathbb{R}^3 : \sigma_2^2 \geq 0\}$. Since $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ is a closed convex cone and $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, the asymptotic null distribution of λ_N is a chi-bar square distribution.

Based on Silvapulle and Sen (2005)'s results (Equation 1.21), without loss of generality, Baey et al. (2019) tested the nullity of the last r variances and corresponding

covariances of matrix $\boldsymbol{D}$ using likelihood ratio test statistic, assuming that the variances that are not being tested are strictly positive. Assume that matrix $\boldsymbol{D}$ is written as $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}$ where the size of $\boldsymbol{D}_{11}$ is $(q-r) \times (q-r)$ and the size of $\boldsymbol{D}_{22}$ is $r \times r$. Under conditions as in 7.0.1.1, Baey et al. (2019) obtained the following results: When $\boldsymbol{D}$ is a diagonal matrix, $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & 0 \\ 0 & \boldsymbol{D}_{22} \end{bmatrix}$ where $\boldsymbol{D}_{11} = \text{diag}(d_1, \ldots, d_{q-r})$ and $\boldsymbol{D}_{22} = \text{diag}(d_{q-r+1}, \ldots, d_q)$. The parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_{\boldsymbol{\epsilon}}^2)^T \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ with $\boldsymbol{\tau} = (d_1, \ldots, d_q)$. Consider the hypothesis test, $H_0 : \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & 0 \\ 0 & 0 \end{bmatrix}$ with positive-definite matrix $\boldsymbol{D}_{11}$ versus $H_1 :$ $\boldsymbol{D}$ is positive definite. Let $\boldsymbol{\theta}^*$ be the true value of the parameter $\boldsymbol{\theta}$. Assume that the null hypothesis holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, then the asymptotic null distribution of the log-likelihood ratio test statistic is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ (see Equation 1.18), which is a mixture of chi-squared distributions with degree of freedom ranging from 0 to r.

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=0}^{r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{1.24}$$

where $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = 0, \ldots, m,$ are some non-negative numbers and $\sum_{i=0}^{m} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right) = 1$; χ_i^2 is a chi-square distribution with i degrees of freedom; and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$. $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp}$ where $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp}$ is the orthogonal complement of $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ in $\mathbb{R}^m$.

When $\boldsymbol{D}$ is a full matrix, the number of distinct variances and covariances is

$q(q+1)/2$. Consider the hypothesis test, $H_0 : \boldsymbol{D} = \left[\begin{smallmatrix} \boldsymbol{D}_{11} & 0 \\ 0 & 0 \end{smallmatrix}\right]$ with positive-definite matrix $\boldsymbol{D}_{11}$ versus $H_1 : \boldsymbol{D}$ is a positive definite matrix. Let $\boldsymbol{\theta}^*$ be the true value of the parameter. Assume that the null hypothesis holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, then, Baey et al. (2019) pointed out that the null asymptotic distribution of the log-likelihood test statistic is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$, which is a mixture of chi-squared distributions with degree of freedom ranging from $r(q-r)$ to $r(q-r)+r(r+1)/2$.

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{1.25}$$

where $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = r(q-r), \ldots, r(q-r)+r(r+1)/2$, are some non-negative numbers and $\sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right) = 1$; χ_i^2 is a chi-square distribution with i degrees of freedom; and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

We will use these results in our work to develop a modified BIC in the following chapters.

1.3.5 Notations

The following list contains general notations that will be used in the dissertation.

- $\boldsymbol{b}_i$ denotes the random effects vector of the ith cluster; $\boldsymbol{b}_i$ is assumed to follow a multivariate normal distribution $N_q(\mathbf{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a $q \times q$ covariance

matrix. $\boldsymbol{\tau}$ denotes the vector of distinct variance and covariance components in matrix $\boldsymbol{D}$.

- I_{n_i} denotes the $n_i \times n_i$ identity matrix. N is the number of clusters; n_i is the number of observations in cluster i.

- The vector of parameters, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_\epsilon)^T \in \Theta \subset \mathbb{R}^m$.

- Denote by $\boldsymbol{\theta}^*$ the true value of the parameter vector.

- $T_{\Theta_0}(\boldsymbol{\theta}^*)$ is the tangent cone of Θ_0 at $\boldsymbol{\theta}^*$ and $T_{\Theta}(\boldsymbol{\theta}^*)$ is the tangent cone of Θ at $\boldsymbol{\theta}^*$.

- $\nu(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l'_N(\boldsymbol{\theta}) \xrightarrow{d} N(0, \nu(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \nu(\boldsymbol{\theta})$.

- Also, let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. Then the distribution of $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C})$ is given by,

$$Pr\left(\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \leq c\right) = \sum_{i=0}^{p} w_i\left(p, \boldsymbol{V}, \mathcal{C}\right) Pr(\chi_i^2 \leq c), \tag{1.26}$$

  where $w_i\left(p, \boldsymbol{V}, \mathcal{C}\right), i = 0, \ldots, p$, are some non-negative numbers and $\sum_{i=0}^{p} w_i\left(p, \boldsymbol{V}, \mathcal{C}\right) = 1$.

- $\xrightarrow{a.s.}$ means convergence almost surely as N goes to ∞.

- $\xrightarrow{d}$ means convergence in distribution as N goes to ∞.

- $\mathbb{S}_+^r$ denotes the set of symmetric positive semi-definite matrices of size $r \times r$.

  $\mathbb{S}^r$ is the set of symmetric matrices of size $r \times r$.

- $\mathbb{R}_+^r$ denotes the positive orthant in $\mathbb{R}^r$.

- $\boldsymbol{A}^T$ denotes the transposition of matrix $\boldsymbol{A}$, for any matrix $\boldsymbol{A}$.

2 Modified BIC for Linear Mixed Models with Independent Random Effects

In this chapter, we provide a brief theoretical background and propose a modified Bayesian information criterion (BIC) for choosing random effects in linear mixed models with independent random effects.

Model selection in linear mixed models includes selection of both regression parameters $\boldsymbol{\beta}$ (fixed effects) and variance components of random effects. In this chapter, we start with the selection of random effects for a special case when the random effects are uncorrelated. In chapter 3 we will select random effects for a general case when random effects are correlated. In chapter 4, we will propose a modified BIC to select both fixed effects and random effects.

Selecting random effects in linear mixed models is important because by selecting the most appropriate random effect model, we can determine the underlying correlation structures of the observations. Furthermore, it helps to interpret the data

analysis results. Fixed effects describe the population level effect, whereas random effects model the individual variability away from the population mean levels. For example, for the data set "Orthodont" in the nlme package that is introduced in chapter 1, beside modelling the pattern of the response variable over time within subjects, it is also important to investigate the variation in the time trends between subjects. This can be done using random effects in linear mixed models. When considering a BIC for choosing random effects in linear mixed models, the challenge is in finding the asymptotic null distribution of the likelihood ratio test statistic when the testing values of variance components are on the boundary of the parameter space. We will tackle this problem in our work.

2.1 Background

Assume that the covariance matrix $\boldsymbol{D}$ of random effects in the linear mixed model (1.1) is a $q \times q$ diagonal matrix. That is, the random effects are uncorrelated. Without loss of generality, we want to test the nullity of the last r variance components, $r < q$. Matrix $\boldsymbol{D}$ can be written as $\boldsymbol{D} = \left[\begin{smallmatrix} \boldsymbol{D}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{22} \end{smallmatrix}\right]$ where $\boldsymbol{D}_{11} = \text{diag}(d_1, \ldots, d_{q-r})$ and $\boldsymbol{D}_{22} = \text{diag}(d_{q-r+1}, \ldots, d_q)$. The parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_{\boldsymbol{\epsilon}})^T \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ with $\boldsymbol{\tau} = (d_1, \ldots, d_q)^T$. Consider the hypothesis test, $H_0 : \boldsymbol{D} = \left[\begin{smallmatrix} \boldsymbol{D}_{11} & 0 \\ 0 & 0 \end{smallmatrix}\right]$ with positive-definite matrix $\boldsymbol{D}_{11}$ versus $H_1 :$ $\boldsymbol{D}$ is positive definite. Let $\boldsymbol{\theta}^*$ be the true value of

the parameter. The parameter spaces under the null and alternative hypotheses and their corresponding tangent cones at $\boldsymbol{\theta}^*$ are:

$$\begin{aligned}
\boldsymbol{\Theta}_0 &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; d_1 > 0, \ldots, d_{q-r} > 0, \\
&\qquad d_{q-r+1} = 0, \ldots, d_q = 0, \sigma^2_{\boldsymbol{\epsilon}} \geq 0\}. \\
T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) &= \{\mathbb{R}^p \times \mathbb{R}^{q-r} \times \{0\}^r \times \mathbb{R}\}. \\
\boldsymbol{\Theta} &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; d_1 > 0, \ldots, d_{q-r} > 0, \\
&\qquad d_{q-r+1} \geq 0, \ldots, d_q \geq 0, \sigma^2_{\boldsymbol{\epsilon}} \geq 0\}. \\
T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) &= \mathbb{R}^p \times \mathbb{R}^{q-r} \times \mathbb{R}^r_+ \times \mathbb{R}.
\end{aligned}$$

In this case, $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$. Therefore, $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^\perp = \{0\}^p \times \{0\}^{q-r} \times \mathbb{R}^r_+ \times \{0\}$ (Baey et al., 2019; Proposition 7.1). C^* is contained in a linear subspace of dimension r. Thus, $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = 0$ for $i = r+1, \ldots, m$ (Shapiro, 1985, 1988; proof of Corollary 1 in Baey et al., 2019). Assuming that the null hypothesis holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, Baey et al. (2019) pointed out that the asymptotic null distribution of the log-likelihood ratio test statistic is a mixture of chi-squared distributions with degree of freedom ranging from 0 to r (Corollary 7.1).

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=0}^{r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi^2_i, \tag{2.1}$$

where $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = 0, \ldots, r$, are some non-negative numbers and $\sum_{i=0}^{r} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right) = 1$, χ_i^2 is a chi-square distribution with i degrees of freedom, $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$, and m is the dimension of $\boldsymbol{\theta}$. We will use this result in our work to develop a modified BIC in the next section.

2.2 Derivation of the modified BIC

In this section, we introduce a modified BIC for choosing random effects in linear mixed models with independent random effects.

In the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_{\boldsymbol{\epsilon}}^2)^T$, $\boldsymbol{\tau}$ is the parameter of interest; $\boldsymbol{\beta}$ and $\sigma_{\boldsymbol{\epsilon}}^2$ are considered as nuisance parameters. Let $\boldsymbol{\lambda} = (\boldsymbol{\beta}^T, \sigma_{\boldsymbol{\epsilon}})^T$ and let p be the number of parameters of $\boldsymbol{\beta}$ and q is the number of parameters of $\boldsymbol{\tau}$. We first consider the linear mixed model (1.1) with the covariance matrix for random effects $\boldsymbol{b}_i$ being diagonal.

Let $\mathcal{M} = \{M_k : k \geq 1\}$ be a countable set of possible candidate linear mixed models. Let $\boldsymbol{\theta}_k$ denote the vector of parameters of model M_k and let d_k be the complexity of model M_k. Assume that d_k can be calculated and $d_k < d_l$ if $M_k \subset M_l$. Let M_T be the model that generates the data (called the true model) with parameter $\boldsymbol{\theta}_T$ and the true value of $\boldsymbol{\theta}_T$ is $\boldsymbol{\theta}_{T,0}$. Any model M_k that is more complex than the true

model is called an over-fitting model. That is, $M_T \subset M_k$ or $\boldsymbol{\theta}_T \subset \boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_T \neq \boldsymbol{\theta}_k$. Let $\mathcal{M}^+$ be the set of all over-fitting models. An under-fitting model M_k is a model such that $\boldsymbol{\theta}_T$'s components are not a subset of its parameter vector's components. That is, $\boldsymbol{\theta}_T \subsetneq \boldsymbol{\theta}_k$. Let $\mathcal{M}^-$ be the set of all under-fitting models. Assume that model M_k has parameter vector $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_k}^T, {\boldsymbol{\tau}_k}^T, \sigma^2_{\boldsymbol{\epsilon},k})^T$ where $\boldsymbol{\beta}_k$ is the vector of fixed effects parameters which includes the population regression coefficients; $\boldsymbol{\tau}_k$ contains the distinct variance and covariance elements of matrix $\boldsymbol{D}$; and $\sigma_{\boldsymbol{\epsilon},k}$ is the parameter for the variance of the random error vector $\boldsymbol{\epsilon}$. For a general covariance matrix case, model M_k is uniquely defined by its non-zero parameters in $\boldsymbol{\beta}$ and non-zero variance components on the diagonal of matrix $\boldsymbol{D}$. If $d_{ii} = 0$, then all elements on row i and column i of this matrix are set to 0.

In this chapter, we consider the case when $\boldsymbol{D}$ is a diagonal matrix. Therefore, $\boldsymbol{\tau}$ contains all variances on the diagonal of matrix $\boldsymbol{D}$. Let $\boldsymbol{\tau} = (d_1, \ldots, d_q)^T$. Assume that we test the model M_k (with $\boldsymbol{\tau}_k = (d_1, \ldots, d_k)$) against model M_1 (with $\boldsymbol{\tau}_1 = (d_1, 0, \ldots, 0)$). Here k is the number of variances of random effects in model M_k and $k \geq 2$. In this case, $m = \dim(\boldsymbol{\theta}_k) = p + k + 1$; $q = k$; and $r = k - 1$. Thus, based on (2.1), the null limiting distribution of the likelihood ratio test statistic is:

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=0}^{k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi^2_i, \tag{2.2}$$

where $C^* = \{0\}^p \times \{0\} \times \mathbb{R}_+^{k-1} \times \{0\}$; $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = 0, \ldots, k-1$, are

some non-negative numbers and $\sum_{i=0}^{k-1} w_i\,(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = 1$; matrix $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l'_N(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$, and m is the dimension of $\boldsymbol{\theta}$.

For example, M_2 is a model with 2 variance components. $\boldsymbol{\tau} = (d_1, d_2)^T$ and $\boldsymbol{D} = \left[\begin{smallmatrix} d_1 & 0 \\ 0 & d_2 \end{smallmatrix}\right]$. We want to test $H_0 : d_2 = 0$ vs. $H_1 : d_2 > 0$, assuming that d_1 is strictly positive. In this example, $m = p + 2 + 1; q = 2; k = 2; r = 1$, and,

$$\begin{aligned}
\boldsymbol{\Theta} &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; \boldsymbol{\tau} \in \mathbb{R}^2_+; \sigma^2_{\boldsymbol{\epsilon}} \geq 0\} \\
&= \mathbb{R}^p \times \mathbb{R}^2_+ \times \mathbb{R}_+ \\
\boldsymbol{\Theta}_0 &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; \boldsymbol{\tau} = (d_1, 0)^T \text{ and } d_1 > 0; \sigma^2_{\boldsymbol{\epsilon}} \geq 0\} \\
&= \{\mathbb{R}^p \times \mathbb{R}_+ \times \{0\} \times \mathbb{R}_+\}.
\end{aligned}$$

The corresponding tangent cones to $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$ are:

$$\begin{aligned}
T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) &= \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \\
T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) &= \mathbb{R}^p \times \mathbb{R} \times \{0\} \times \mathbb{R} \\
T^{\perp}_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) &= \{0\}^p \times \{0\} \times \mathbb{R} \times \{0\}.
\end{aligned}$$

Therefore, $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \{0\}^p \times \{0\} \times \mathbb{R}_+ \times \{0\}$. Thus, the null

asymptotic distribution of the log-likelihood ratio test statistic is

$$\begin{aligned}\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) &= \sum_{i=0}^{1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2 \\ &= w_0(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_0^2 + w_1(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_1^2 \\ &= \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2.\end{aligned}$$

This is consistent with Stram and Lee (1994)'s results that the distribution of the likelihood ratio test statistic for this case is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

If we compare model M_1 with model M_0 where modelM_0 contains the fixed effects only, then the likelihood ratio test statistic is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Stram and Lee (1994)).

We now take the expectation of the chi-bar distribution in equation (2.2) and include it in the complexity of model M_k.

$$E[\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)] = \sum_{i=0}^{k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)i.$$

Also, the expectation of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ is $\frac{1}{2} \times 0 + \frac{1}{2} \times 1$ or 0.5.

We propose the following modified BIC:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{2.3}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k; $n = \sum_{i=1}^{N} n_i$ and $d_k = p + 1.5 + \sum_{i=0}^{k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)i$ for $k > 1$; $d_k = p + 1.5$ for $k = 1$; and $d_k = p + 1$ for $k = 0$.

The first term, $-2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$, measures the goodness-of-fit for model M_k and the second term, $d_k \log(n)$, is the penalty for model complexity, which makes sure that the model selected is parsimonious.

The rationale of choosing the complexity d_k for model M_k when $k > 1$ is as follows: p is the number of fixed effects parameters, 1 is for $\sigma_{\boldsymbol{\epsilon}}$ parameter, 0.5 for the assumed random effect in the model (such as random intercept), and the rest of d_k is the expectation of the chi-bar distribution in equation (2.2). When $k = 1$, $d_1 = p + 1.5$ is the complexity of model M_1 which is the model with fixed effects and only one random effect (such as random intercept). When $k = 0$, $d_0 = p + 1$ is the complexity of model M_0 which is the model with fixed effects and no random effects. In this case, d_0 is exactly the regular BIC for multiple regression models.

A Bayesian information criterion has been proposed for model selection in linear mixed models by some authors, such as Pauler (1998), Pauler et al. (1999), Jones (2011), as summarized in Chapter 1 of our introduction. Pauler (1998) uses Bayes factors to derive a Schwarz criterion to select the regression parameter $\boldsymbol{\beta}$ in independent cluster models. However, it is difficult to obtain simple approximations to the Bayes factor (Müller et al. (2013)). Pauler et al. (1999) addressed the boundary problems in the variance component model by assuming that the parameter space $\boldsymbol{\Theta}$ can be extended to an open set $\boldsymbol{\Theta}^*$ such that the boundary of $\boldsymbol{\Theta}$ is in the interior

of $\boldsymbol{\Theta}^*$. Jones (2011) proposed to replace the total number of observations, n, in the penalty term of BIC by an effective sample size that is calculated based on the covariance matrix of the response variable. The penalty term in the formula for the BIC used in the "lmer" function implemented in the current R package 'lme4' (Bates et al., 2015) for linear mixed models is $(p+q+1)\log(n)$, where $(p+q+1)$ is the total number of parameters in the model. In this version, the boundary issue is totally ignored. In our work, we have considered the boundary issue when obtaining the asymptotic null distribution of the likelihood ratio test statistic and used its expected value in the penalty term of our modified BIC.

We are now going to prove the consistency of the proposed BIC.

2.2.0.1 Assumptions for Theorem 1 and Theorem 2

(C1). The observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ from different clusters are independent random vectors. All the assumptions of the linear mixed model (1.1) are satisfied.

(C2). Let $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ be the log-likelihood function of the linear mixed model (1.1). Denote by $\boldsymbol{\Theta}$ the parameter space of the model parameter vector, $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}^*$ be the true value of the parameter vector. Denote the vector of first partial derivatives of $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ with respected to $\boldsymbol{\theta}$ by $l'_N(\boldsymbol{\theta})$ and denote the matrix of second partial derivatives of $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ with respected to $\boldsymbol{\theta}$ by $l''_N(\boldsymbol{\theta})$. Directional

derivatives are used when $\boldsymbol{\theta}$ is on the boundary of $\boldsymbol{\Theta}$. (i) Assume that, for all $\boldsymbol{\theta}$, the first three partial derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$ exist almost everywhere. (ii) Also, assume that N^{-1} times the absolute value of the third derivative of $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ is bounded by a function of $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ whose expectation exists and finite on the intersection of neighborhoods of $\boldsymbol{\theta}^*$ and $\boldsymbol{\Theta}$.

(C3). $n_1, \ldots, n_N$ are uniformly bounded. That is, there exists a constant $K > 0$ such that $n_i \leq K$ for $i = 1, \ldots, N$.

(C4). Let $\boldsymbol{\theta}_T$ be the parameter vector of the true model M_T and let $\boldsymbol{\theta}_{T,0}$ denote the true value of $\boldsymbol{\theta}_T$.

(i) For any under-fitting model, M_k, with model parameter $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k$, assume that $E_{T,0}\left[\log \frac{f(\boldsymbol{y};\boldsymbol{\theta}_{T,0})}{f(\boldsymbol{y};\boldsymbol{\theta}_k)}\right]$ exists and there exists a unique pseudo true, $\boldsymbol{\theta}_{k,0}$, such as $\boldsymbol{\theta}_{k,0} = \underset{\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k}{\arg\min} E_{T,0}\left[\log \frac{f_i(\boldsymbol{y};\boldsymbol{\theta}_{T,0})}{f_i(\boldsymbol{y};\boldsymbol{\theta}_k)}\right]$ for all i.

(ii) For all $\boldsymbol{\theta}$, $\frac{1}{N}(l(\boldsymbol{\theta}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}; \boldsymbol{Y})]) \xrightarrow{p} 0$.

(iii) For any two nested models, $M_k \subset M_l$, $-2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_l; \boldsymbol{Y})\right)$ is bounded by an integrable function, $M(\boldsymbol{Y})$, and $E[M(\boldsymbol{Y})] < \infty$.

Assumption $(C1)$ is to ensure that the assumptions in the linear mixed model (1.1) are satisfied. Assumption $(C2)$ is to ensure that conditions $(C1), (C2)(i)(ii)(iv)$ and

$(C3)$ in Baey et al. (2019)'s work are fulfilled so that the asymptotic null distribution of the likelihood statistic is a chi-bar square distribution and Baey et al. (2019)'s results in their Proposition 1, Corollary 1, and Theorem 2 can apply. Assumptions $(C3), (C4)$ are used in the proof of consistency of the proposed BIC.

Theorem 2.1 *Theorem 1: Consistency of the modified BIC. Assume that the assumptions* $(C1) - (C4)$ *are satisfied, then*

$$\lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1 \ \textit{for all}\ M_k \in M^+$$

$$\textit{and} \qquad \lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1 \ \textit{for all}\ M_k \in M^-.$$

Theorem 1 says that as the sample size goes to infinity, the proposed BIC would correctly identify the true model with probability tending to 1.

Given a set of candidate models, we calculate the proposed BIC value for each model. Then, the selected model is the one that minimizes the proposed BIC. Please see a proof of Theorem 1 in the subsection below.

2.2.1 Proof to Theorem 1

Proof.

We first compare the proposed BIC of the true model to an underfitted model. We prove that the proposed BIC of the true model is less than the proposed BIC of any underfitted model with probability tending to 1 as n goes to ∞. We then show that the proposed BIC of the true model is also less than that of any overfitted model with probability tending to 1. Combining two parts, we get the result that the proposed BIC of the true model is less than that of any other model with probability tending to 1 as n goes to ∞ or the proposed BIC would correctly identify the true model with probability tending to 1.

In the following, we use $l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$ instead of $l_N(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$ for the convenience of exposition.

Case 1: For any under-fitting model, $M_k \in M^-$, we want to prove that $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$.

We have that

$$
\begin{aligned}
BIC^*(M_k) &= -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \\
BIC^*(M_T) &= -2l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) + d_T \log(n), \\
BIC^*(M_k) - BIC^*(M_T) &= -2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) + (d_k - d_T)\log(n). \\
-2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) &= -2\Big(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - [l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y})] \\
&\quad - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) + l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\Big) \\
&= -2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right) + 2\left[l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y})\right] \\
&\quad + 2\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right] - 2E_{T,0}\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{Y})\right] \\
&\quad + 2E_{T,0}\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{Y})\right].
\end{aligned}
$$

We have that $l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) = o_p(1)$ and $l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) = o_p(1)$ because $\hat{\boldsymbol{\theta}}_k \xrightarrow{p} \boldsymbol{\theta}_{k,0}$ and $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_{T,0}$ (as shown in the proof of Theorem 2 in Baey et al. (2019)) and function $l(\boldsymbol{\theta}; \boldsymbol{y})$ is continuous with respect to $\boldsymbol{\theta}$. Also, under assumption $C4(ii)$, $\frac{1}{N}(l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y})]) \xrightarrow{p} 0$ and $\frac{1}{N}(l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{k,0}; \boldsymbol{Y})]) \xrightarrow{p} 0$. Thus,

$$
\frac{1}{N}(l(\boldsymbol{\theta}_{T,0}; ; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{Y})]) \xrightarrow{p} 0,
$$

and therefore, $l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{Y})] = o_p(N)$.

The last term can be evaluated as,

$$E_{T,0}[l(\boldsymbol{\theta}_{T,0};\boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0};\boldsymbol{Y})] = \sum_{i=1}^{N} E_{T,0}[\log f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0}) - \log f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})]$$
$$= \sum_{i=1}^{N} E_{T,0}\left[\log \frac{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})}{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})}\right] = O_p(N).$$

This is because $E_{T,0}\left[\log \frac{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})}{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})}\right]$ is the Kullback-Leibler distance between $f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})$ and $f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})$; and is positive and finite by assumption $C4(i)$.

Assume that the cluster sample sizes, $n_1, \ldots, n_N$ are uniformly bounded (assumption $C3$), then $O_p(N)$ dominates $(d_k - d_T)\log(n)$ as $N \to \infty$. Thus, $BIC^*(M_k) - BIC^*(M_T) > 0$. And, $\lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1$ for all $M_k \in M^-$.

Case 2: For any over-fitting model, $M_k \in M^+$, we also prove that $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$. Without loss of generality, assume that $\boldsymbol{\theta}_T = ({\boldsymbol{\beta}_T}^T, {\boldsymbol{\psi}_T}^T, \underline{0}, \sigma^2_{\boldsymbol{\epsilon},T})^T$ and $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_k}^T, {\boldsymbol{\psi}_{k,1}}^T, {\boldsymbol{\psi}_{k,2}}^T, \sigma^2_{\boldsymbol{\epsilon},k})^T$ where $\boldsymbol{\psi}_T$ has the same dimension as $\boldsymbol{\psi}_{k,1}$ and $\underline{0}$ has the same dimension as ${\boldsymbol{\psi}_{k,2}}^T$. Let r be the dimension of $\boldsymbol{\psi}_{k,2}; \dim(\boldsymbol{\psi}_{k,2}) = r$ and all elements of $\underline{0}$ are 0. We have that

$$BIC^*(M_k) - BIC^*(M_T) = -2\left(l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})\right) + (d_k - d_T)\log(n). \qquad (2.4)$$

Then $-2(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}))$ is the likelihood ratio test statistic of the following

hypothesis test,

$$H_0 : \boldsymbol{\psi}_{k,1} \geq \mathbf{0}, \boldsymbol{\psi}_{k,2} = \mathbf{0}$$

$$H_1 : \boldsymbol{\psi}_{k,1} \geq \mathbf{0}, \boldsymbol{\psi}_{k,2} > \mathbf{0}.$$

According to Baey et al. (2019), under H_0, the asymptotic distribution of $-2(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}))$ is

$$\sum_{i=0}^{r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2.$$

Therefore, $-2(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})) = O_p(1)$, according to van der Vaart (2000, Theorem 2.4). We also have that,

$$\begin{aligned}
2(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})) &= 2\left(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y}) - \left[l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y})\right]\right) \\
&= -2\left(l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) \\
&\quad - \left[-2\left(l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})\right)\right]. \\
\Rightarrow E\left[2\left(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y})\right)\right] &= E\left[-2\left(l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{Y})\right)\right] \\
&\quad - E\left[-2\left(l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y})\right)\right] = d_T - d_k,
\end{aligned}$$

where $l(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y})$ is the maximum log-likelihood of the simplest model; that is the model with only the random intercept. Therefore,

$$E\left[-2\left(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y})\right)\right] = d_k - d_T.$$

On the other hand, $-2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})-l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y})\right)$ asymptotically has the distribution which is a mixture of the chi-square distributions. Therefore, $E\left[-2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{Y})-l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{Y})\right)\right]$ must be positive and, hence, $d_k - d_T > 0$. Thus, $BIC^*(M_k) - BIC^*(M_T) \to \infty$ as $n \to \infty$ and $\lim_{n\to\infty} P\left(BIC^*(M_k) - BIC^*(M_T) > 0\right) = 1$ for $M_k \in M^+$. Combining the two cases, we have proved that the proposed BIC of the true model is less than that of any other model almost surely as n goes to ∞ or the proposed BIC would correctly identify the true model with probability tending to 1. This completes the proof of Theorem 1. ■

2.3 Simulation

In this section, we evaluate the performance of the proposed BIC*. We compare the performance of the proposed BIC* to the regular BIC. For each candidate model, we compute BIC* and regular BIC; then for each method we choose the model with minimum value of BIC* and regular BIC, respectively.

Following the methods used in Gao and Song (2010) and Chen and Chen (2012), the criteria we used to evaluate and compare the proposed BIC to the regular BIC are (1) Positive Selection Rate (PSR), (2) False Discovery Rate (FDR), and (3) Correction Rate (CR).

For each chosen model, the positive selection rate (PSR) is the ratio of the number

of predictors that are correctly identified as significant in the chosen model to the number of predictors that are truly significant in the data-generating model, for which we take the average of the PSRs over all chosen models. The false discovery rate (FDR) is the ratio of the number of predictors that are incorrectly identified as significant in the chosen model to the number of predictors that are identified as significant in the chosen model, for which we take the average of the FDRs over all chosen models. The correction rate (CR) is the proportion of the times the true data-generating model is selected in all chosen models. For example, in the data-generating model, $\boldsymbol{X}_1, \boldsymbol{X}_3$ are significant predictors while $\boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5$ are not significant predictors. Therefore, the selected model would include $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_4, \boldsymbol{X}_5$ as significant predictors, yielding $PSR = 1/2$ and $FDR = 3/4$.

2.3.1 Simulation Settings

Our data is generated from linear mixed model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$.

Scenario 1: With total number of observations $n = 500$ and number of clusters, $N = 100$. X is a $n \times p$ matrix with $n = 500$; $p = 2$; the first column of $\boldsymbol{X}$ includes all 1's. The second column is $\boldsymbol{X}_1$ which is generated from the standard normal distribution. The vector of fixed effects, $\boldsymbol{\beta} = (1, 2)^T$. Matrix $\boldsymbol{Z}$ contains the first two columns $\boldsymbol{z}_0$, $\boldsymbol{z}_1$ which are the same as two columns of matrix $\boldsymbol{X}$ and two more

columns $\boldsymbol{z}_2$, $\boldsymbol{z}_3$ both are generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, are generated from multivariate normal distribution $N_q(\mathbf{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ is a 4×4 diagonal matrix and $\boldsymbol{D} = \text{diag}(\sigma_0^2, \ldots, \sigma_3^2)$. To measure the ability to detect the significance of variance components parameter of the proposed BIC^*, we created different scenarios for different sizes of σ_1^2, σ_2^2, σ_3^2 as shown below. $\boldsymbol{\epsilon}$ is generated from a multivariate normal distribution, $N(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n)$ with $\sigma_{\boldsymbol{\epsilon}}^2 = 1$.

The random intercept, b_0, has the standard deviation of $\sigma_0 = 5$. Random effects, b_1, b_2, and b_3, have standard deviations σ_1, σ_2, σ_3, respectively. We consider 2 cases.

Case 1: σ_1 is a sequence of values from 0 to 0.5 incrementing by 0.05; σ_2 is a sequence of values from 0 to 1 incrementing by 0.1; σ_3 is a sequence of values from 0 to 2 incrementing by 0.2.

Case 2: σ_1 is a sequence of values from 0 to 0.5 incrementing by 0.05; σ_2 is a sequence of values from 0 to 0.4 incrementing by 0.04; σ_3 is a sequence of values from 0 to 0.6 incrementing by 0.06. Repeat all the above cases with $n = 1000$ ($N = 200$), and $n = 250$ ($N = 50$).

Scenario 2: With total number of observations $n = 500$ and number of clusters, $N = 100$. $\boldsymbol{X}$ is a $n \times p$ matrix with $n = 500$; $p = 3$; the first column of $\boldsymbol{X}$ includes all 1's. The last two columns of matrix $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are generated from the standard normal distributions. The vector of fixed effects, $\boldsymbol{\beta} = (1, 2, 3)^T$. Matrix $\boldsymbol{Z}$

contains the first three columns $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ which are the same as three columns of matrix $\boldsymbol{X}$ and three more columns $\boldsymbol{z}_3$, $\boldsymbol{z}_4$, $\boldsymbol{z}_5$ which are generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, are generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ is a 6×6 diagonal matrix and $\boldsymbol{D} = \text{diag}(\sigma_0^2, \ldots, \sigma_5^2)$. To measure the ability to detect the significance of variance components parameter of the BIC^*, we created different scenarios for different sizes of σ_1^2, σ_2^2, σ_3^2 as shown below; $\sigma_4^2 = 0$ and $\sigma_5^2 = 0$. $\boldsymbol{\epsilon}$ is generated from a multivariate normal distribution, $N(\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n)$ with $\sigma_{\boldsymbol{\epsilon}}^2 = 1$.

The random intercept, b_0, has the standard deviation of $\sigma_0 = 5$. Random effects, b_1, b_2, b_3, b_4, and b_5, have standard deviations σ_1, σ_2, σ_3, σ_4, σ_5, respectively. σ_1 is a sequence of values from 0 to 0.5 incrementing by 0.05; σ_2 is a sequence of values from 0 to 1 incrementing by 0.1; σ_3 is a sequence of values from 0 to 2 incrementing by 0.2; $\sigma_4 = 0$ and $\sigma_5 = 0$. Repeat all the above cases with $n = 1000$ ($N = 200$), and $n = 250$ ($N = 50$).

Scenario 3: With total number of observations $n = 500$ and number of clusters, $N = 100$. $\boldsymbol{X}$ is a $n \times p$ matrix with $n = 500$; $p = 3$; the first column of $\boldsymbol{X}$ includes all 1's. The last two columns of matrix $\boldsymbol{X}_1$ are generated from the standard normal distributions. The vector of fixed effects, $\boldsymbol{\beta} = (1, 2, 3)^T$. Matrix $\boldsymbol{Z}$ contains the first three columns $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ which are the same as three columns of matrix

$\boldsymbol{X}$ and eight more columns $\boldsymbol{z}_3, \ldots, \boldsymbol{z}_{10}$ which are generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, are generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ is a 11×11 diagonal matrix and $\boldsymbol{D} = \text{diag}(\sigma_0^2, \ldots, \sigma_{10}^2)$, where $\sigma_1^2 = 0.16$, $\sigma_2^2 = 0.64$, $\sigma_3^2 = 1$, $\sigma_4^2 = 1.44$ and $\sigma_5^2, \ldots, \sigma_{10}^2$ are all 0. The error term, $\boldsymbol{\epsilon}$, is generated from a multivariate normal distribution, $N(\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n)$ with $\sigma_{\boldsymbol{\epsilon}}^2 = 1$.

The random intercept, b_0, has the standard deviation of $\sigma_0 = 5$. Random effects, $b_1, \ldots, b_{10}$, have standard deviations, $\sigma_1, \ldots, \sigma_{10}$, respectively. Repeat this simulation set up with $n = 1000$ ($N = 200$), and $n = 250$ ($N = 50$).

2.3.2 Simulation Procedure

In scenario 1, for each set of values of σ_1^2, σ_2^2, σ_3^2, $B = 1001$ simulations are run using parallel programming with 7 processors to yield $1001/7 = 143$ simulation rounds. In each simulation, all possible candidate models, M_k, are fitted, having the same fixed effect covariates (including $\boldsymbol{X}_1$ and the intercept); meanwhile the covariates for random effects part vary in the power set of $\{1, 2, 3\}$. The proposed BIC^* and regular BIC are calculated for each model. The chi-bar-square weights, $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$, are calculated using function "con-weights-boot" in the R package "restriktor" Vanbrabant et al. (2019). The matrix $\boldsymbol{\nu}(\boldsymbol{\theta}^*)$ is approximated by

$N^{-1}\{I_N(\hat{\boldsymbol{\theta}}_k)\}$ where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k and $I_N(\boldsymbol{\theta})$ is the Fisher information matrix as in 1.4, and the cone C^* is as defined in 2.2. Linear mixed models are run using the "lmer" function implemented in the R package 'lme4' (Bates et al., 2015). The regular BIC is given by

$$BIC(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + (p + k + 1)\log(n). \tag{2.5}$$

For the proposed BIC^* criterion, we choose the model with minimum proposed BIC^*. For the regular BIC criterion, we also choose the model with minimum regular BIC. For each selection criterion, we have a set of 1001 models obtained from 1001 simulations. The correction rate (CR) was calculated for each criterion.

In scenario 2, for each set of values of $\sigma_1^2, \ldots, \sigma_5^2$, $B = 1001$ simulations are run, where for each simulation, all possible candidate models is fitted having the same fixed effect covariates (including $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and the intercept); meanwhile the covariates for random effects vary in the power set of $\{1, \ldots, 5\}$. The proposed BIC^* and regular BIC are calculated for each model. Subsequently, one model with minimum proposed BIC is selected and one model with minimum regular BIC is selected. From 1001 simulations, we obtained 1001 models under each selection criterion. We calculate the means and standard deviations of Positive Selection Rate and False Discovery Rate. We also calculate the correction rate for each criterion.

In scenario 3, with the given set of values of $\sigma_1^2, \ldots, \sigma_{10}^2$, $B = 1001$ simulations are

ran. In each simulation, all possible candidate models is fitted with the same fixed effect covariates (including $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and the intercept); meanwhile the covariates for random effects vary in the power set of $\{1, \ldots, 10\}$. For each candidate model, the proposed BIC^* and regular BIC are calculated. Subsequently, the models with the minimum proposed BIC and minimum regular BIC are selected, resulting in 1001 models obtained from 1001 simulations for each selection criterion. We calculate the means and standard deviations of Positive Selection Rate and False Discovery Rate; and calculate the Correction Rate for each criterion. All simulations are performed by using R version 4.0.2 (Team, 2000).

2.3.3 Simulation Results

Scenario 1: Table 2.1 summarizes the results of Scenario 1 Case 1 and Table 2.2 summarizes the results of Scenario 1 Case 2. In both cases, we observe that the correction rate for the proposed BIC^* is greater than that of the regular BIC. Furthermore, the difference in the correction rate between these two methods is bigger when the values of σ_1^2, σ_2^2, σ_3^2 are smaller.

σ_1	σ_2	σ_3	Correction Rate	
			Proposed BIC	Regular BIC
0.00	0.00	0.00	0.00	0.00
0.05	0.10	0.20	0.00	0.00
0.10	0.20	0.40	0.01	0.00
0.15	0.30	0.60	0.04	0.01
0.20	0.40	0.80	0.11	0.02
0.25	0.50	1.00	0.24	0.08
0.30	0.60	1.20	0.36	0.18
0.35	0.70	1.40	0.54	0.32
0.40	0.80	1.60	0.67	0.47
0.45	0.90	1.80	0.78	0.60
0.50	1.00	2.00	0.87	0.72

"Correction Rate" reports the proportion of times the selected model is the true data-generating model.

Table 2.1: Comparison of the Proposed BIC and Regular BIC methods in terms of Correction Rate for the simulation in case 1 of Scenario 1 with n = 500 and N = 100.

σ_1	σ_2	σ_3	Correction Rate	
			Proposed BIC	Regular BIC
0.00	0.00	0.00	0.00	0.00
0.05	0.04	0.06	0.00	0.00
0.10	0.08	0.12	0.00	0.00
0.15	0.12	0.18	0.00	0.00
0.20	0.16	0.24	0.01	0.00
0.25	0.20	0.30	0.03	0.01
0.30	0.24	0.36	0.14	0.03
0.35	0.28	0.42	0.29	0.10
0.40	0.32	0.48	0.49	0.24
0.45	0.36	0.54	0.67	0.39
0.50	0.40	0.60	0.80	0.58

"Correction Rate" reports the proportion of times the selected model is the true data-generating model.

Table 2.2: Comparison of the Proposed BIC and Regular BIC methods in terms of Correction Rate for the simulation in case 2 of Scenario 1 with n = 500 and N = 100.

Scenario 2: Table 2.3 summarizes the results of Scenario 2. The simulation results suggest that the values of Positive Selection Rate (PSR) for the proposed BIC^* are higher than the regular BIC when the values of variance components are close to 0. That is, the ability to choose the significant variance components is higher for the

proposed BIC^* than the regular BIC. Almost all of the False Discovery Rate (FDR) values are within 5 percent in both cases.

			Proposed BIC			Regular BIC		
σ_1	σ_2	σ_3	PSR (SD)	FDR (SD)	Correction Rate	PSR (SD)	FDR (SD)	Correction Rate
0.00	0.00	0.00	0.029 (0.010)	0.067 (0.062)	0.00	0.004 (0.001)	0.009 (0.008)	0.00
0.05	0.10	0.20	0.118 (0.031)	0.060 (0.053)	0.00	0.037 (0.011)	0.013 (0.013)	0.00
0.10	0.20	0.40	0.392 (0.034)	0.033 (0.016)	0.01	0.292 (0.027)	0.007 (0.005)	0.00
0.15	0.30	0.60	0.537 (0.034)	0.027 (0.011)	0.03	0.429 (0.026)	0.008 (0.004)	0.01
0.20	0.40	0.80	0.657 (0.032)	0.026 (0.009)	0.12	0.568 (0.032)	0.007 (0.003)	0.04
0.25	0.50	1.00	0.734 (0.028)	0.016 (0.005)	0.23	0.660 (0.025)	0.003 (0.001)	0.10
0.30	0.60	1.20	0.796 (0.028)	0.019 (0.006)	0.37	0.719 (0.021)	0.004 (0.001)	0.18
0.35	0.70	1.40	0.854 (0.027)	0.018 (0.005)	0.53	0.777 (0.025)	0.004 (0.001)	0.33
0.40	0.80	1.60	0.902 (0.023)	0.013 (0.004)	0.67	0.830 (0.028)	0.004 (0.001)	0.48
0.45	0.90	1.80	0.945 (0.015)	0.015 (0.004)	0.80	0.888 (0.025)	0.004 (0.001)	0.66
0.50	1.00	2.00	0.960 (0.012)	0.016 (0.004)	0.83	0.916 (0.021)	0.003 (0.001)	0.74

PSR is Positive Selection Rate and FDR is False Discovery Rate, both are averaged over 1001 simulations. All values in brackets are sample standard deviations.

Table 2.3: Comparison of the Proposed BIC and Regular BIC methods in terms of the Positive Selection Rate, the False Discovery Rate, and Correction Rate for different values of σ_1, σ_2, σ_3, $\sigma_4 = 0$, and $\sigma_5 = 0$ in Scenario 2 with n = 500 and N = 100.

As the values of variance components increase, the PSR increases. From the results obtained, we also see that the ability to choose the true model also becomes larger as the values of variance components increase. We also noted that the standard deviations are small for all cases. This means that the estimated PSR and FDR are very consistent.

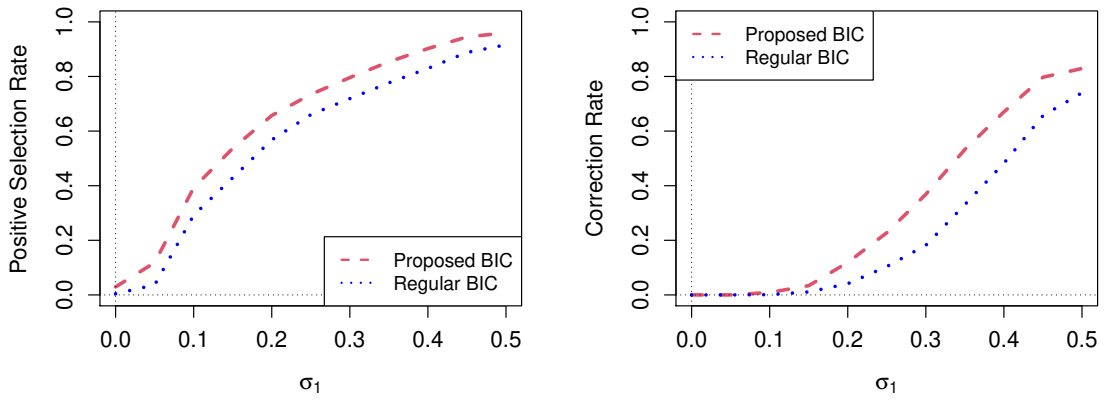

Figure 2.1: Comparison of the Proposed BIC and Regular BIC methods in terms of the Positive Selection Rate and Correction Rate for different values of σ_1, σ_2, σ_3, $n = 500(N = 100)$

Figure 2.1 shows the comparison of the Proposed BIC and Regular BIC methods in terms of the Positive Selection Rate and Correction Rate for different values of σ_1, σ_2, σ_3 when $n = 500$ and $(N = 100)$ in scenario 2.

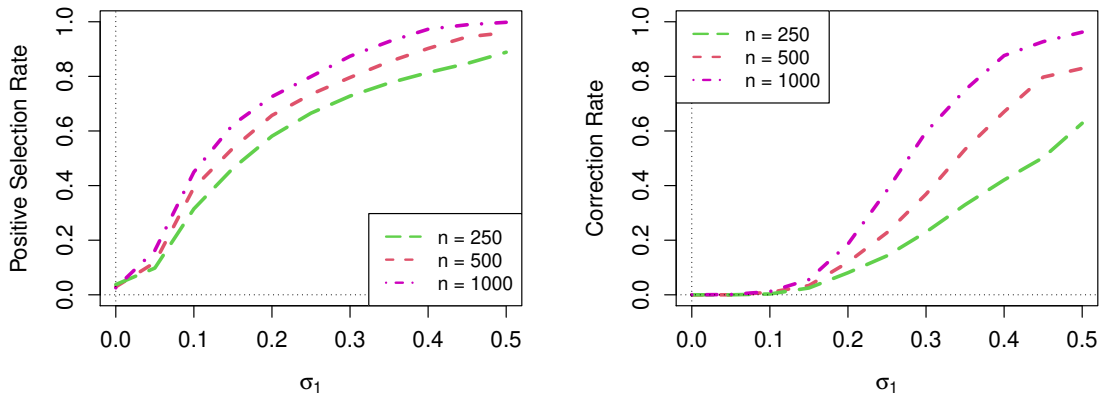

Figure 2.2: Comparison of the Positive Selection Rate and Correction Rate for $n = 250(N = 50)$, $n = 500(N = 100)$, and $n = 1000(N = 200)$

Figure 2.2 shows the comparison of the Positive Selection Rate (PSR) and correction rates for Scenario 2 when $n = 250$, 500, and 1000 with $N = 50$, 100, 200, respectively. Given the same set of values of $\sigma_1^2, \ldots, \sigma_5^2$, we observe that the positive selection rate increases as the number of clusters N increases.

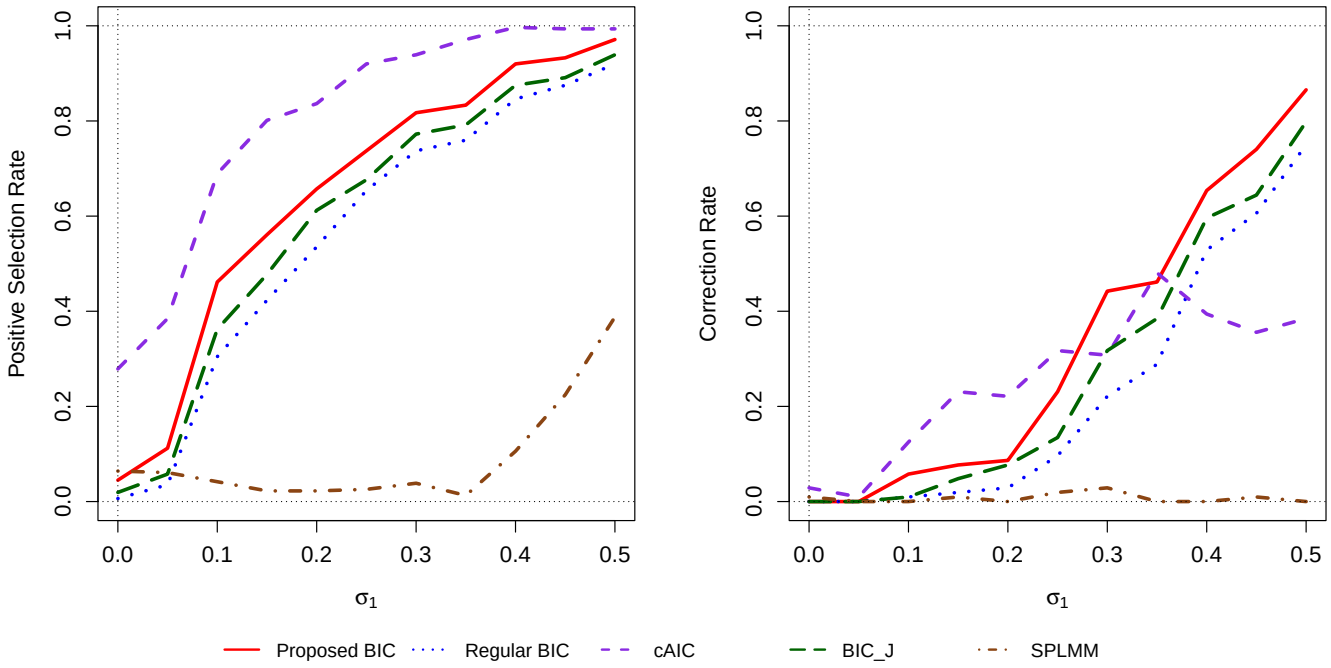

Figure 2.3: Comparison of the Positive Selection Rate and Correction Rate for $n = 500(N = 100)$ with different competing methods for different values of σ_1, σ_2, σ_3 with $\sigma_4 = 0$ and $\sigma_5 = 0$.

We ran 104 simulations with 3 more competing methods: "cAIC", "BIC_J", and "Splmm" using the same setting as in scenario 2. cAIC is the conditional AIC as introduced in (1.12). "BIC_J" is a modified BIC for linear mixed models as introduced in (Jones (2011)). "Splmm" (Simultaneous Penalized Linear Mixed Effects Models) is a method for choosing both the fixed effects and random effects for variable selection using penalized likelihood function. The R-package "*Splmm*" contains functions that fit linear mixed models for high-dimensional data ($p > n$) with penalty functions for

both the fixed effects and random effects. This method is based on the results in (Yang and Wu (2022)). Figure 2.3 shows that the modified BIC performs better than the regular BIC, "BIC_J", and "$Splmm$" in this scenario in terms of the positive selection rate and correction rate. The ability to choose correct variables is higher for cAIC than the modified BIC. However, the correction rates for cAIC are not always higher than that of the modified BIC. The "$Splmm$" method does not seem to work well in this scenario. This may be because the method works better for the case when the number of parameters is much higher than the number of observations.

Scenario 3: Table 2.4 summarizes the results of Scenario 3. We see that in all cases for the sample sizes $n = 500$, 1000, 250, the mean PSR and the correction rates are higher for the proposed BIC, meanwhile the FDR are kept around 5%.

(n, N)	Method	Average PSR (SD)	Average FDR (SD)	Correction Rate
(250, 50)	Proposed BIC	0.825 (0.015)	0.063 (0.014)	0.21
	Regular BIC	0.771 (0.014)	0.017 (0.004)	0.15
(500, 100)	Proposed BIC	0.883 (0.016)	0.051 (0.010)	0.40
	Regular BIC	0.832 (0.015)	0.009 (0.002)	0.32
(1000, 200)	Proposed BIC	0.959 (0.009)	0.041 (0.008)	0.67
	Regular BIC	0.916 (0.014)	0.005 (0.001)	0.65

"Correction Rate" reports the proportion of times the selected model is the true data-generating model

Table 2.4: Comparison of the Proposed BIC and Regular BIC methods in terms of the Positive Selection Rate and Correction Rate for $n = 250$, $n = 500$, and $n = 1000$ in Scenario 3.

The computational time for the proposed BIC: It took about 0.07 second to calculate the complexity, d_k and the proposed BIC, using our personal laptop. The simulation time depends on the number of simulations; number of variances of random effects; and number of possible models for each simulation. For example, in scenario 2, there are 11 sets of different values of variances. For each set of variances, we ran 1001 simulations. Each simulation has 32 possible models, and the number of random effects is 6, including the random intercept. The simulation time was 6.458593 hours.

Steps to compute the model complexity, d_k:

- Run a linear mixed model using 'lmer' function in the R-package 'lme4' and get the covariance matrix of the MLE estimator of model parameters.
- Get the R matrix from the constraints $R \times \theta \geq 0$ that define cone C^*.
- Obtain the weights using the "con-weights-boot" function in the 'restriktor' package in R.
- Calculate the expected value of the chi-bar square distribution and obtain the penalty term to get modified BIC.

2.4 Real-Data Application

In this section, we apply the proposed BIC to a real data set. We work with a data set which is a subset of 120 schools of dataset "hsfull" from package 'spida2' in R (Monette et al., 2019). This dataset is originally from the 1982 "High School and Beyond" (HSB) survey data set in Raudenbush and Bryk's text on hierarchical linear models, Raudenbush and Bryk (2002). The data includes mathematics achievement test scores of 5307 students from 50 Catholic and 70 Public high schools, with the number of students in each school ranging from 19 to 66 students.

The variables included in the analysis are school identification number, mathematics achievement score (Y), socioeconomic status ($\boldsymbol{X}_1$), sex (female (0) or male (1); $\boldsymbol{X}_2$), visible minority status (yes (1) or no (0); $\boldsymbol{X}_3$), school sector (catholic (0) or public (1); $\boldsymbol{X}_4$). Variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ are group-centered. The objective is to study the relationship between students' mathematics achievement score and socioeconomic status, sex, visible minority status in public and catholic schools; and whether this relationship varies across schools within each sector.

We fit linear mixed models to the data set, where all models have the same fixed effects which include $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$, while the random effects include the random intercept and a subset of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$, where $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$ are the socioeconomic status, sex, and visible minority variables which are group-centered.

It may be too restrictive to fit linear mixed models with uncorrelated random effects to this dataset because when we ran a full linear mixed model with all possible random effects, we obtained the estimated correlation between random slope of $\boldsymbol{z}_1$ and random slope of $\boldsymbol{z}_3$ which is -0.87. However, it is still desirable to assume uncorrelated random effects and compare the scenario with the scenario obtained in chapter 3, where the random effects are assume to be correlated.

We calculate the proposed BIC^* and the regular BIC for each model. We also use cAIC as presented in Säfken et al. (2018), Greven and Kneib (2010), and Vaida and Blanchard (2005).

Model	Proposed BIC	Regular BIC	cAIC
Random Intercept (RI)	34379.33	34379.83	34176.92
RI, $\boldsymbol{z}_1$	34380.07	34384.86	34172.44
RI, $\boldsymbol{z}_2$	34383.27	34388.06	34179.66
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$	34384.10	34393.11	34175.22
RI, $\boldsymbol{z}_3$	34377.73	34382.52	34167.84
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_3$	34379.13	34388.33	34165.13
RI, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	34381.57	34390.70	34170.34
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	34383.25	34396.52	34167.60

All values are rounded to two decimal places.

Table 2.5: Results of the proposed BIC, regular BIC, and cAIC for all models considered in the Real-Data Application section.

Table 2.5 provides the results of the proposed BIC, regular BIC, and cAIC for all models considered. The optimal model we obtain using the proposed BIC^* is the model with random intercept and random slope of $\boldsymbol{z}_3$; the proposed BIC^* is 34377.73. The optimal model we obtain using the regular BIC is the model with random intercept only. The regular BIC of this model is 34379.83. The optimal model we obtain using the cAIC is the model with random intercept, random slopes of $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$. The cAIC of the optimal model is 34165.13. We rerun the final model chosen using the proposed BIC. All fixed effects are highly significant with p-value < 0.001. The estimates of fixed effects coefficients, $\hat{\boldsymbol{\beta}}$, are $(14.4712, 1.8621, 1.1786, -2.9348, -3.0057)^T$. The estimated variance matrix of the random effects is $\text{diag}(\hat{\sigma}_0^2, \hat{\sigma}_3^2)$ where $\hat{\sigma}_0 = 2.514$ and $\hat{\sigma}_3 = 1.509$, and the estimated standard deviation of the error term is 5.971. We observe that the estimated random effects variances are quite small compared to the estimated variance of the error term. The estimated variance of the random intercept term is just about 18% of the variance of the random error term and the estimated variance of random slope term for $\boldsymbol{z}_3$ is just about 6.4% of the variance of the random error term. Based on the optimal model chosen by the proposed BIC, there is a significant linear relationship between students' mathematics achievement score and socioeconomic status, sex, visible minority status in public and catholic schools. On average, students in catholic schools, being male, non-minority, and had high social

economic status tend to have higher mathematics achievement score. Furthermore, the mean mathematics achievement score and minority gap effect significantly vary across schools within each sector.

2.5 Discussion

In this chapter, we have proposed a modified BIC for for choosing random effects in linear mixed models with uncorrelated random effects. Through the simulation results and the application results, we see that the proposed BIC^* performs quite well in selecting the optimal model, which can capture the amount of information contained in the data. Compared to the regular BIC, the proposed BIC^* performs better when the values of the random effects variances are small.

In chapter 3, we will present the case when the random effects are assumed to be correlated. That is, the covariance matrix of random effects is a full matrix.

3 Modified BIC for Linear Mixed Models with Correlated Random Effects

In this chapter, we propose a BIC model criteria for choosing random effects in linear mixed models with correlated random effects. In model (1.1), the vector of random effects, $\boldsymbol{b_i}$, is assumed to follow a multivariate normal distribution. It is quite natural that its covariance matrix can be a full matrix. For example, the random effects for subject i is $\boldsymbol{b}_i = (b_{i0}, b_{i1})$, where b_{i0} is the subject's random intercept and b_{i1} is the subject's random slope. There may be the case that subjects with higher random intercept also have higher random slope. That is, there is a positive correlation between subjects' random intercepts and random slopes. And this correlation structure should be captured by the model. Therefore, we will consider the case with correlated random effects in this chapter.

3.1 Background

When random effects are correlated, the covariance matrix $\boldsymbol{D}$ of random effects in the linear mixed model (1.1) is a full matrix. Assume that matrix $\boldsymbol{D}$ is written as $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}$ where the size of $\boldsymbol{D}_{11}$ is $(q-r) \times (q-r)$ and the size of $\boldsymbol{D}_{22}$ is $r \times r$. When $\boldsymbol{D}$ is a full matrix, the number of distinct variances and covariances is $q(q+1)/2$. Consider the hypothesis test, $H_0 : \boldsymbol{D}_{11} > \boldsymbol{0}, \boldsymbol{D}_{12} = \boldsymbol{0}, \boldsymbol{D}_{22} = \boldsymbol{0}$ versus H_1 : $\boldsymbol{D} > \boldsymbol{0}$. That is, $\boldsymbol{D}$ is a positive definite matrix. Let $\boldsymbol{\theta}^*$ be the true value of the parameter. The parameter space under the null hypothesis is:

$$\begin{aligned} \boldsymbol{\Theta}_0 &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; \boldsymbol{D}_{11} > \boldsymbol{0}; \boldsymbol{D}_{12} = \boldsymbol{0}, \boldsymbol{D}_{22} = \boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2 \geq 0\} \\ &= \{\mathbb{R}^p \times \mathbb{S}_+^{q-r} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}_+\}, \end{aligned}$$

where $\mathbb{S}_+^{q-r}$ is the set of symmetric positive semi-definite matrices of size $(q-r) \times (q-r)$.

Assume that the null hypothesis is true and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$. Applying the result from (Baey et al., 2019; Proposition 7.1), we obtain the tangent cone to $\boldsymbol{\Theta}_0$ at $\boldsymbol{\theta}^*$:

$$\begin{aligned} T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) &= \{\mathbb{R}^p \times \mathbb{S}^{q-r} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}\} \\ &= \{\mathbb{R}^p \times \mathbb{R}^{(q-r)(q-r+1)/2} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}\}, \end{aligned}$$

where $\mathbb{S}^{(q-r)}$ is the set of symmetric matrices of size $(q-r) \times (q-r)$. Also, the

parameter space under the alternative hypothesis is:

$$\begin{aligned}\boldsymbol{\Theta} &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; D \in \mathbb{S}_+^q, \sigma_{\boldsymbol{\epsilon}}^2 \geq 0\} \\ &= \{\mathbb{R}^p \times \mathbb{S}_+^q \times \mathbb{R}_+\}.\end{aligned}$$

According to Baey et al. (2019)'s proof of Proposition 1, the tangent cone to $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$ is:

$$T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \mathbb{R}^p \times \mathbb{R}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}_+^r \times \mathbb{R},$$

where $\mathbb{S}_+^r$ is the set of symmetric positive semi-definite matrices of size $r \times r$. Since $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, the null asymptotic distribution of the likelihood ratio test statistic for the above hypothesis test is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ (see Equation 1.18), where $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \{0\}^p \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}_+^r \times \{0\}$ (Baey et al., 2019; Proposition 7.1).

When $\boldsymbol{D}$ is a full matrix, under conditions $B1$ to $B5$ as in (7.0.1.1), Baey et al. (2019) pointed out that the asymptotic null distribution of the log-likelihood ratio test statistic is:

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{3.1}$$

where $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*), i = r(q-r), \ldots, r(q-r)+r(r+1)/2$, are some non-negative numbers and $\sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right) = 1$, m is the dimension of $\boldsymbol{\theta}$, χ_i^2 is

a chi-square distribution with i degrees of freedom, and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l'_N(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$. $\mathbb{S}_+^r$ is the set of symmetric positive semi-definite matrices of size $r \times r$ (we note that it is too complex to define $\mathbb{S}_+^r$ using equality and inequality constraints on variance and covariance components of matrix $\boldsymbol{D}$). Since $\mathbb{S}_+^r \subset \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r$, in our work, we will approximate C^* by $C = \{0\}^p \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r \times \{0\}$. And, thus, $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ is approximated by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=r(q-r)+r(r-1)/2}^{r(q-r)+r(r-1)/2+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \tag{3.2}$$

where $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C), i = r(q-r) + r(r-1)/2, \ldots, r(q-r) + r(r-1)/2 + r$, are some non-negative numbers and $\sum_{i=r(q-r)+r(r-1)/2}^{r(q-r)+r(r-1)/2+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = 1$, χ_i^2 is a chi-square distribution with i degrees of freedom, and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l'_N(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$. This is because $C = \{0\}^p \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r \times \{0\}$ which contains a linear space of dimension $r(q-r) + r(r-1)/2$ and is included in a linear space of dimension $r(q-r)+r(r-1)/2+r$. Therefore, the weights $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)$ are zero for $i = 0, \ldots, r(q-r)+r(r-1)/2-1$ and for $i = r(q-r)+r(r-1)/2+r+1, \ldots, m$ (7.2).

3.2 Derivation of Modified BIC for Correlated Random Effects

In this section, we introduce a modified BIC for selecting linear mixed models with correlated random effects. We still focus on only selecting random effects.

In the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma^2_{\boldsymbol{\epsilon}})^T$, $\boldsymbol{\tau}$ is the parameter of interest; $\boldsymbol{\beta}$ and $\sigma^2_{\boldsymbol{\epsilon}}$ are considered as nuisance parameters. We now consider the linear mixed model (1.1) with the covariance matrix for random effects $\boldsymbol{b}_i$ is a full matrix. Therefore, $\boldsymbol{\tau}$ contains all distinct variances and covariances of matrix $\boldsymbol{D}$.

Assume that model M_k has parameter vector $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_k}^T, {\boldsymbol{\tau}_k}^T, \sigma^2_{\boldsymbol{\epsilon},k})^T$ where $\boldsymbol{\beta}_k$ represents the parameter vector of fixed effects; $\boldsymbol{\tau}_k$ contains distinct variances and covariances of random effect covariance matrix $\boldsymbol{D}_k$ and $\sigma_{\boldsymbol{\epsilon},k}$ is the parameter from the variance of the random error term $\boldsymbol{\epsilon}$. Let p be the number of parameters of $\boldsymbol{\beta}_k$ and q_k is the number of parameters of $\boldsymbol{\tau}_k$. Model M_k is uniquely defined by its non-zero parameters in $\boldsymbol{\beta}$ and non-zero variance components on the diagonal of matrix $\boldsymbol{D}_k$. If $d_{ii} = 0$, then all elements on row i and column i of this matrix are set to 0.

Assume that we test the model M_k against model M_1, where M_1 contains only one random effect which is random intercept and M_k contains k random effects including random intercept. Assume that the two models contain the same fixed effects part. In this case, $m = \dim(\boldsymbol{\theta}_k) = p + q_k + 1$, $r = k - 1$, $q = k$, and $q - r = 1$. Thus,

$r(q-r)+r(r-1)/2 = k(k-1)/2$ and $r(q-r)+r(r-1)/2+r = (k-1)(k+2)/2$. Thus, based on (3.2), the asymptotic null distribution of the log-likelihood ratio test statistic is

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=k(k-1)/2}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \tag{3.3}$$

where $C = \{0\}^p \times \{0\} \times \mathbb{R}^{k(k-1)/2} \times \mathbb{R}_+^{k-1} \times \{0\}$; m is the dimension of $\boldsymbol{\theta}$, $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C), i = k(k-1)/2, \ldots, (k-1)(k+2)/2$, are some non-negative numbers and $\sum_{i=k(k-1)/2}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = 1$, χ_i^2 is a chi-square distribution with i degrees of freedom, and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

For example, M_3 is a model with 3 random effects. $\boldsymbol{\tau} = (d_{11}, d_{12}, d_{13}, d_{22}, d_{23}, d_{33})$ and $\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{12} & d_{22} & d_{23} \\ d_{13} & d_{23} & d_{33} \end{bmatrix}$. We want to test $H_0 : d_{11} > 0$; $d_{12} = 0$, $d_{13} = 0$, $d_{22} = 0$, $d_{23} = 0$, $d_{33} = 0$ vs. $H_1 :$ $\boldsymbol{D}$ is positive definite. In this example, $q_k = 6, m = p + 6 + 1, q = k = 3, r = 2$. Thus, $k(k-1)/2 = 3$ and $(k-1)(k+2)/2 = 5$. Therefore, the null asymptotic distribution of the log-likelihood ratio test statistic is approximated by:

$$\begin{aligned}\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) &= \sum_{i=3}^{5} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2 \\ &= w_3(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_3^2 + w_4(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_4^2 + w_5(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_5^2,\end{aligned}$$

where $C = \{0\}^p \times \{0\} \times \mathbb{R}^3 \times \mathbb{R}_+^2 \times \{0\}$.

From (3.3), let $c_k = E(\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=k(k-1)/2}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)i$.

We propose the following modified BIC:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{3.4}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k; $n = \sum_{i=1}^{N} n_i$ and $d_k = p + 1.5 + c_k$ for $k > 1$; $d_k = p + 1.5$ for $k = 1$; and $d_k = p + 1$ for $k = 0$. The first term, $-2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$, measures the goodness-of-fit for model M_k and the second term, $d_k \log(n)$, is the penalty for model complexity, which makes sure that the model selected is parsimonious. The selected model is the one that minimizes the proposed BIC.

The proof of this proposed BIC's consistency is similar to the proof of consistency presented in Chapter 2.

We will compare the performance of this proposed BIC to the regular BIC. The regular BIC for the case when matrix $\boldsymbol{D}$ is a full matrix is:

$$BIC(M_k) = -2l(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + (p + k(k+1)/2 + 1) \log n, \tag{3.5}$$

where p is the number of fixed effects parameters; $k(k+1)/2 + 1$ is the number of distinct parameters in the random effects covariance matrix and the error term variance parameter; and n is the number of observations.

3.3 Simulation

In this section, we evaluate the performance of the proposed BIC*. We compare the performance of the proposed BIC* to the regular BIC. For each candidate model, we compute BIC* and regular BIC; then for each method we choose the model with minimum value of BIC* and regular BIC, respectively.

3.3.1 Simulation Set up

Our data is generated from linear mixed model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$. The total number of observations is $n = 1000$ and number of clusters is $N = 100$. $\boldsymbol{X}$ is a $n \times p$ matrix with $n = 1000$; $p = 3$; the first column of $\boldsymbol{X}$ includes all 1's. The last two columns, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are generated from the standard normal distribution. The vector of fixed effects, $\boldsymbol{\beta} = (1, 2, 3)^T$. Matrix $\boldsymbol{Z}$ contains the first three columns $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ which are the same as three columns of matrix $\boldsymbol{X}$ and three more columns $\boldsymbol{z}_3$, $\boldsymbol{z}_4$, $\boldsymbol{z}_5$ are generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, are generated from a multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a 6×6 full matrix. The correlation matrix between the random effects in the data-generating model is:

$$R = \begin{bmatrix} 1 & 0.7 & 0.6 & 0.5 \\ 0.7 & 1 & 0.4 & 0.3 \\ 0.6 & 0.4 & 1 & 0.5 \\ 0.5 & 0.3 & 0.5 & 1 \end{bmatrix}.$$

To measure the ability to detect the significance of variance components parameter of the proposed BIC^*, we created different cases for different sizes of σ_0^2, σ_1^2, σ_2^2, σ_3^2 as shown below. σ_4^2, σ_5^2 and covariances corresponding to random effects of $\boldsymbol{z}_4$ and $\boldsymbol{z}_5$ are all 0. $\boldsymbol{\epsilon}$ is generated from a multivariate normal distribution, $N(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n)$ with $\sigma_{\boldsymbol{\epsilon}}^2 = 1$.

Case 1: The standard deviations of random effects are $\sigma_0 = 5$, $\sigma_1 = 1.5$, $\sigma_2 = 1$, $\sigma_3 = 0.5$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 2: The standard deviations of random effects are $\sigma_0 = 5$, $\sigma_1 = 1.0$, $\sigma_2 = 0.8$, $\sigma_3 = 0.4$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 3: The standard deviations of random effects are $\sigma_0 = 2$, $\sigma_1 = 0.8$, $\sigma_2 = 0.5$, $\sigma_3 = 0.4$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 4: The standard deviations of random effects are $\sigma_0 = 2$, $\sigma_1 = 0.8$, $\sigma_2 = 0.5$, $\sigma_3 = 0.3$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 5: The standard deviations of random effects are $\sigma_0 = 2$, $\sigma_1 = 0.5$, $\sigma_2 = 0.4$, $\sigma_3 = 0.2$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 6: In this case, we keep the standard deviations of random effects the same as the ones in case 4. However, we increase the correlations by 0.1 for each non-zero correlation in the correlation matrix to see how this affects the correction rates. The correlation matrix between the random effects is:

$$R_1 = \begin{bmatrix} 1 & 0.8 & 0.7 & 0.6 \\ 0.8 & 1 & 0.5 & 0.4 \\ 0.7 & 0.5 & 1 & 0.6 \\ 0.6 & 0.4 & 0.6 & 1 \end{bmatrix}.$$

3.3.2 Simulation Procedure

In each case presented above, $B = 1001$ simulations are run. In each simulation, all possible candidate models is fitted, with the same fixed effect covariates (including the intercept, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$); meanwhile the covariates for random effects part vary in the power set of $\{1, 2, 3, 4, 5\}$ and also include random intercept. The proposed BIC^*, regular BIC, and cAIC are calculated for each model. Models with the minimum proposed BIC, minimum regular BIC and minimum cAIC are selected, obtaining 1001 models from 1001 simulations for each selection criterion. We calculate the means and standard deviations of Positive Selection Rate and False Discovery Rate; and the correction rate for each criterion.

3.3.3 Simulation results

Table 3.1 shows the comparison of the Proposed BIC, Regular BIC, and cAIC methods in terms of the Positive Selection Rate, the False Discovery Rate, and Correction Rate for case 1 to case 6. In all cases, the correction rate for the proposed BIC is greater than that of the regular BIC. The difference in the correction rate between these two methods is bigger when the values of σ_1^2, σ_2^2, σ_3^2 are smaller. And in most cases, the two methods seem to perform better than the cAIC method. In case 5, the values of variances are small except the variance for the random intercept and we observed that cAIC performs better than the proposed BIC and regular BIC in this case. As we observed in figure 2.3 of chapter 2, cAIC works better than the proposed BIC in terms of the positive selection rate. However, cAIC doesn't work better than the proposed BIC in terms of correction rate. For some sets of values of variances, cAIC performs better but for other sets of values, cAIC doesn't perform better. The correction rate of cAIC is not high. This may be because cAIC is not a consistent criterion for model selection.

	Proposed BIC			Regular BIC			cAIC		
Case	PSR (SD)	FDR (SD)	CR	PSR (SD)	FDR (SD)	CR	PSR (SD)	FDR (SD)	CR
1	1 (0.000)	0.0 (0.000)	1.00	0.999 (0.0002)	0.0 (0.000)	0.998	0.9987 (0.0004)	0.1190 (0.0197)	0.5644
2	0.99 (0.0032)	0.0007 (0.0002)	0.967	0.9837 (0.0052)	0.0007 (0.0002)	0.9481	0.9950 (0.0025)	0.1110 (0.0212)	0.6054
3	0.9937 (0.0021)	0.0 (0.000)	0.981	0.989 (0.0035)	0.0 (0.000)	0.967	0.9990 (0.0003)	0.0871 (0.0170)	0.6783
4	0.8911 (0.0244)	0.0003 (0.0001)	0.6733	0.8541 (0.0273)	0.0 (0.000	0.5624	0.9933 (0.0026)	0.0935 (0.0188)	0.6603
5	0.7106 (0.0132)	0.0003 (0.0001)	0.1339	0.6893 (0.0084)	0.0003 (0.0001)	0.0739	0.9314 (0.0184)	0.0940 (0.0206)	0.5355
6	0.9204(0.0202)	0.0 (0.000)	0.7612	0.8901(0.0246)	0.0 (0.000)	0.6703	0.995(0.0016)	0.0904(0.0189)	0.6763

PSR is Positive Selection Rate, FDR is False Discovery Rate, CR is Correction Rate, and SD is Standard Deviation. All are averaged over 1001 simulations. All values in brackets are sample standard deviations.

Table 3.1: Comparison of the Proposed BIC, Regular BIC, and cAIC methods in terms of the Positive Selection Rate, the False Discovery Rate, and Correction Rate for different values of σ_0, σ_1, σ_2, σ_3, $\sigma_4 = 0$, and $\sigma_5 = 0$ with correlated random effects.

3.4 Real-Data Application

In this section, we apply the proposed BIC to the subset of 120 schools of dataset "hsfull" which is the same dataset as in the previous chapter. Data are on mathematics achievement test scores from 50 Catholic and 70 Public high schools. That is, $N = 120$. The number of students in each school ranges from 19 to 66 students. Total number of students is 5307. That is, $n = 5307$. Variables included in the analysis are school identification number, mathematics achievement score $(; \boldsymbol{y})$, socioeconomic status $(\boldsymbol{X}_1)$, sex (female or male; $\boldsymbol{X}_2$), visible minority status (yes/no; $\boldsymbol{X}_3$), school sector (catholic or public; $\boldsymbol{X}_4$). The candidate variables for random effects part are $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$, being the socioeconomic status, sex, and visible minority variables which are group-centered. Assume that the fixed effects includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$, we wish to find significant random effects components. We can use the proposed method to analyze this data because in this dataset, students are nested within schools. There are 120 schools in this dataset and each school is considered as a cluster, thus, there are 120 independent clusters. The mathematics achievement test scores from the students within the same school are correlated because these students share the same school environment and teachers. We can also consider this data as hierarchical data with two levels: the school level and the student level.

We first fit a linear mixed model which includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ and random intercept and compared this model to a linear model which includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ to test whether the random intercept effect is significant. We found that the random intercept is very significant with p-value < 0.00001. We then fit linear mixed models with correlated random effects to the data set. All the models have the same fixed effects part which includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$, meanwhile the random effects part includes the random intercept and a subset of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$. We calculate the proposed BIC and the regular BIC for each model. The optimal model we obtain using the proposed BIC is the model with random intercept; the proposed BIC is 34379.83. The optimal model we obtain using the regular BIC is also the model with random intercept only. The BIC of this model is 34379.83. The optimal model we obtain using the cAIC is the model with random intercept, random slopes of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$. The cAIC of the optimal model is 34166.25.

Table 3.2 shows of the proposed BIC, regular BIC, and cAIC for all models with correlated random effects considered in the Real-Data Application section.

Model	Proposed BIC	Regular BIC	cAIC
Random Intercept (RI)	34379.33	34379.83	34176.92
RI, $\boldsymbol{z}_1$	34380.61	34385.4	34167.38
RI, $\boldsymbol{z}_2$	34391.38	34396.17	34181.23
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$	34401.4	34410.2	34174.07
RI, $\boldsymbol{z}_2$	34384.58	34389.37	34169.18
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$	34391.03	34400.38	34167.38
RI, $\boldsymbol{z}_2$, $\boldsymbol{z}_2$	34405.59	34414.63	34169.18
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_2$	34420.4	34433.63	34166.25

All values are rounded to two decimal places.

Table 3.2: Results of the proposed BIC, regular BIC, and cAIC for all models with correlated random effects considered for the subset of the "hsfull" dataset.

Combine the results obtained for independent random effects in 2.5 and correlated random effects in 3.2, we see that the model that includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, and independent random effects which are random intercept and $\boldsymbol{z}_3$ is selected using the proposed BIC method. The model that includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ and random intercept is selected using the regular BIC method. The model that includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, and independent random effects which are random intercept, $\boldsymbol{z}_1$, and $\boldsymbol{z}_3$ is selected using the cAIC method. Table 3.3 shows the optimal model chosen by each method when the random effects are assumed to be independent and when the

random effects are correlated.

	Proposed BIC		Regular BIC		cAIC	
Case	Optimal model	Proposed BIC	Optimal model	Regular BIC	Optimal model	cAIC
Correlated Random Effects	RI	34379.33	RI	34379.83	RI, $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$	34166.25
Independent Random Effects	RI, $\boldsymbol{z}_3$	34377.73	RI	34379.83	RI, $\boldsymbol{z}_1, \boldsymbol{z}_3$	34165.13

RI means Random Intercept.

Table 3.3: Compare the optimal model chosen by each method for correlated random effects and independent random effects.

In general, if we have a reason to justify that the random effects are independent, then we choose the optimal model chosen by this scenario. Otherwise, we choose the optimal model chosen by the scenario in which the random effects are assumed to be correlated. In our application, the model that includes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, and independent random effects which are random intercept and $\boldsymbol{z}_3$ should be selected because the estimated correlation between random intercept and random slope of $\boldsymbol{z}_3$ is -0.34 in the full model will all possible random effects and, thus, it is reasonable to assume that the random intercept and random slope of $\boldsymbol{z}_3$ are uncorrelated.

We also look at the data set "Orthodont" in the nlme package, which is introduced in chapter 1. The dental data set introduced by Potthoff and Roy (1964), where dental measurements were made on 11 girls and 16 boys at ages 8, 10, 12 and 14. The response variable was the distance, in millimeters, from the center of pituitary to the pterygomaxillary fissure. There are 27 subjects in the data set with the following variables: Distance is a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). Age is a numeric vector of ages of the subject (in years). Subject is an ordered factor indicating the subject on which the measurement was made. Sex is a factor with levels Male and Female The objective is to study the change in an orthodontic measurement over time for young boys and girls.

We first check if the random intercept is significant by comparing models with and

without random intercept. We found the random intercept is also very significant with p-value < 0.00001. We then fit all possible linear mixed models to the data set. All models contain the same fixed effects part which is Sex, Age, and the interaction between Sex and Age.

Table 3.4 shows the proposed BIC, regular BIC, and cAIC values of all models considered for the "Orthodont" dataset. Based on the results, all three criteria (the proposed BIC, regular BIC, and cAIC) choose the model with fixed effects as Sex, Age, the interaction between Sex and Age, and the random intercept. This result is also consistent with the results from Baey et al. (2019) and Potthoff and Roy (1964).

Model	Proposed BIC	Regular BIC	cAIC
Random Intercept (RI)	461.35	461.85	405.47
Age	464.31	464.81	409.97
RI, Age (uncorrelated)	467.77	465.92	405.60
RI, Age (correlated)	467.20	470.04	405.51

All values are rounded to two decimal places.

Table 3.4: Results of the proposed BIC, regular BIC, and cAIC for all models considered for the "Orthodont" dataset

3.5 Discussion

In this chapter, we have proposed a modified BIC for choosing random effects in linear mixed models with correlated random effects. Based on the simulation results, we see that the proposed BIC performs well in selecting the random effects in all cases. The proposed BIC performs better than the regular BIC and cAIC methods in all cases. The performance of the proposed BIC and regular BIC is better when the magnitude of the variance or correlation values are larger. In Chapter 4, we will propose a modified BIC for choosing both fixed effects and random effects in linear mixed models. We also look at two scenarios: when the random effects are assumed to be independent and when the random effects are assumed to be correlated.

4 Modified BIC for selecting both Fixed Effects and Random Effects in Linear Mixed Models

In this chapter, we propose a modified BIC to select both fixed effects and random effects for linear mixed models. We also divide the situations into two cases: when the random effects are independent, that is, the covariance matrix, $\boldsymbol{D}$, of random effects is diagonal and when the random effects are correlated, that is, the covariance matrix, $\boldsymbol{D}$, is a full matrix.

Very often in practice when we consider fitting a linear mixed model to a dataset, we want to know which variables should be included for fixed effects and which variables should be included for random effects from a set of candidate variables. For example, from the set of candidate variables in the subset of 120 schools of dataset "hsfull" used in chapter 3, which variable(s) should be chosen for fixed effects and which one(s) should be chosen for random effects. In this case we need to choose both regression coefficients and random effects variance components.

4.1 Modified BIC for Selecting Both Fixed Effects and Random Effects when Random Effects are Independent

When the covariance matrix, $\boldsymbol{D}$, of random effects in the linear mixed model (1.1) is a diagonal matrix, $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_{q-r}, d_{q-r+1}, \ldots, d_q)$. The fixed effects parameter in the linear mixed model (1.1) is $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$. Without the loss of generality, assume that we want to test the nullity of s components of $\boldsymbol{\beta}$ which are $\beta_1, \ldots, \beta_s$ and the nullity of the last r variances components of matrix $\boldsymbol{D}$ which are $d_{q-r+1}, \ldots, d_q$.

Consider the hypothesis test, $H_0 : \beta_1 = 0, \ldots, \beta_s = 0; d_{q-r+1} = 0, \ldots, d_q = 0$ versus $H_1 : \beta_1 \neq 0, \ldots, \beta_s \neq 0; d_{q-r+1} > 0, \ldots, d_q > 0$, with the variances that are not tested $(d_1, \ldots, d_{q-r})$ are positive. Let $\boldsymbol{\theta}^*$ be the true value of the parameter. Assuming that the null hypothesis is true and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, we obtain the tangent cones to the parameter spaces under the null and alternative hypotheses, respectively.

$$\boldsymbol{\Theta}_0 = \{\{0\}^s \times \mathbb{R}^{p-s} \times \{0\}^r \times \mathbb{R}^{q-r} \times \mathbb{R}_+\},$$

$$T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) = \{\{0\}^s \times \mathbb{R}^{p-s} \times \mathbb{R}^{q-r} \times \{0\}^r \times \mathbb{R}\},$$

$$\boldsymbol{\Theta} = \{\mathbb{R}^p \times \mathbb{R}_+^{q-r} \times \mathbb{R}_+^r \times \mathbb{R}\},$$

$$T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \mathbb{R}^p \times \mathbb{R}^{q-r} \times \mathbb{R}_+^r \times \mathbb{R}.$$

Since $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is also a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, Baey et al. (2019) pointed out

that the null asymptotic distribution of $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ is a mixture of chi-squared distributions with degree of freedom ranging from s to $s+r$.

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=s}^{s+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{4.1}$$

where $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \mathbb{R}^s \times \{0\}^{p-s} \times \{0\}^{q-r} \times \mathbb{R}_+^r \times \{0\}$; $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = s, \ldots, s+r$, are some non-negative numbers and $\sum_{i=s}^{s+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = 1$, χ_i^2 is a chi-square distribution with i degrees of freedom, and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

For example, if we test the nullity of one regression coefficient and the nullity of one variance component, $r = 1$ and $s = 1$, then

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=1}^{2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2 \tag{4.2}$$

$$= \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2, \tag{4.3}$$

where $C^* = \mathbb{R} \times \{0\} \times \{0\} \times \mathbb{R}_+ \times \{0\}$. This result is also obtained in Baey et al. (2019)'s paper.

In the model selection, we assume that the smallest model (called model M_1) contains only the y-intercept for fixed effects and random intercept for random effects. Model M_k contains $(p_k + 1)$ fixed effects and the covariance matrix, $\boldsymbol{D}_k$, of random effects is of order $k \times k$. If random effects are assumed to be independent, then the number of random effects variances components is $q_k = k$. When we test model

M_k against model M_1, we are testing the nullity of $s = p_k$ regression coefficients and $r = k - 1$ random effects variance components. Therefore, based on (4.1), the asymptotic null distribution of the log-likelihood ratio test statistic is

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=p_k}^{p_k+k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{4.4}$$

where $C^* = \mathbb{R}^{p_k} \times \{0\} \times \{0\} \times \mathbb{R}_+^{k-1} \times \{0\}$; $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right), i = p_k, \ldots, p_k+k-1$, are some non-negative numbers and $\sum_{i=p_k}^{p_k+k-1} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\right) = 1$, m is the dimension of $\boldsymbol{\theta}$.

Let u_k be the expectation of $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$, then $u_k = \sum_{i=p_k}^{p_k+k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)i$.

We propose a modified BIC for this case as:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{4.5}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k; $n = \sum_{i=1}^{N} n_i$ and $d_k = 2.5 + u_k$ for $k > 1$; $d_k = p_k + 2.5$ for $k = 1$; and $d_k = p_k + 2$ for $k = 0$. Here, in the formula $d_k = 2.5 + u_k$ for $k > 1$, we add 2.5 to u_k to account for the degrees of freedom of fixed effect intercept (1 degree of freedom), random intercept (0.5 degree of freedom), and the variance component of the error term, $\boldsymbol{\epsilon}$ (1 degree of freedom).

We will later compare the performance of this proposed BIC to the regular BIC. The corresponding regular BIC is:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + (p_k + 2 + k) \log(n). \tag{4.6}$$

The difference between the regular BIC and the proposed BIC is in the model complexity measurement. The model complexity in the regular BIC is the number of components in the model parameter vector. The complexity in the proposed BIC is the total of number of fixed effects, residual variance, and the expected value of the chi-bar square distribution. Since the sum of the weights in the chi-bar square distribution is 1, the expected value of the chi-bar square distribution, c_k, is always less than p_k+k-1 for k greater than 1. Therefore, $c_k+2.5$ is always less than p_k+2+k. Since the regular BIC does not consider the boundary issue, the complexity in the regular BIC does not actually reflect the model's degrees of freedom in this case. The complexity in the proposed BIC considers the boundary issue and is a corrected number of degrees of freedom for the model.

To show assess the magnitude of difference between the complexity of model M_k using proposed BIC and regular BIC in an example, we run a simulation in which model M_k has 6 fixed effects including the y-intercept and 4 random effects including the random intercept. The true model contains 4 fixed effects including the fixed effect intercept with $\beta = (1, 2, 3, 1, 0, 0)$ and 2 random effects including the random intercept with standard deviations of 2 and 0.8 for random intercept and random slope of $\boldsymbol{z}_1$, respectively. We ran all possible models. All models contain the fixed effect intercept and random intercept. Table 4.1 shows the results of the complexity

for the proposed BIC and regular BIC of a model with $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, and $\boldsymbol{X}_5$ as fixed effects and random effects as shown in the rows of the table.

Model	Proposed, d_k	Regular
Random Intercept (RI)	7.5	8
RI, $\boldsymbol{z}_1$	8	9
RI, $\boldsymbol{z}_2$	8	9
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$	8.53	10
RI, $\boldsymbol{z}_3$	8	9
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_3$	8.54	10
RI, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	8.52	10
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	9.07	11

Table 4.1: Comparing the complexity, d_k of the proposed BIC and regular BIC when random effects are independent

4.2 Modified BIC for Selecting Both Fixed Effects and Random Effects when Random Effects are Correlated

When random effects in the linear mixed model (1.1) are correlated, their covariance matrix, $\boldsymbol{D}$, is a full matrix. Matrix $\boldsymbol{D}$ can be written as $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}$ where the size of $\boldsymbol{D}_{11}$ is $(q-r)\times(q-r)$ and the size of $\boldsymbol{D}_{22}$ is $r \times r$. The number of distinct variance and covariance components in $\boldsymbol{D}$ is $q(q+1)/2$.

Consider the hypothesis test, $H_0 : \beta_1 = 0, \ldots, \beta_s = 0, \boldsymbol{D}_{11} > \mathbf{0}, \boldsymbol{D}_{12} = \mathbf{0}, \boldsymbol{D}_{22} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} > \mathbf{0}$. That is, $\boldsymbol{D}$ is a positive definite matrix. Let $\boldsymbol{\theta}^*$ be the true value of the parameter vector. Assume that the null hypothesis holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, then the parameter spaced under the null hypothesis and its tangent cone at $\boldsymbol{\theta}^*$ are:

$$\begin{aligned}
\boldsymbol{\Theta}_0 &= \{\{0\}^s \times \mathbb{R}^{p-s} \times \mathbb{S}_+^{q-r} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}_+\}, \\
T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*) &= \{\{0\}^s \times \mathbb{R}^{p-s} \times \mathbb{S}^{q-r} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}\}, \\
&= \{\{0\}^s \times \mathbb{R}^{p-s} \times \mathbb{R}^{(q-r)(q-r+1)/2} \times \{0\}^{r(q-r)} \times \{0\}^{r(r+1)} \times \mathbb{R}\},
\end{aligned}$$

where $\mathbb{S}_+^{q-r}$ is the set of symmetric positive semi-definite matrices of size $(q-r) \times (q-r)$. Also, the parameter space under the alternative hypothesis is:

$$\begin{aligned}
\boldsymbol{\Theta} &= \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{\beta} \in \mathbb{R}^p; D \in \mathbb{S}_+^q, \sigma_{\boldsymbol{\epsilon}}^2 \geq 0\} \\
&= \{\mathbb{R}^p \times \mathbb{S}_+^q \times \mathbb{R}_+\}.
\end{aligned}$$

The tangent cone to $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$ is:

$$T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \mathbb{R}^p \times \mathbb{R}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}_+^r \times \mathbb{R}.$$

where $\mathbb{S}_+^r$ is the set of symmetric positive semi-definite matrices of size $r \times r$. Since $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, the null asymptotic distribution of the likelihood ratio test statistic for the above hypothesis test is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ where

$C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \mathbb{R}^s \times \{0\}^{p-s} \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}_+^r \times \{0\}$. As in chapter 3, it is challenging to define $\mathbb{S}_+^r$ using equality and inequality constraints. Since $\mathbb{S}_+^r \subset \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r$, we will approximate C^* by $C = \mathbb{R}^s \times \{0\}^{p-s} \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r \times \{0\}$. And $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ is approximated by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=s+r(q-r)+r(r-1)/2}^{s+r(q-r)+r(r-1)/2+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2. \tag{4.7}$$

where $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C), i = s+r(q-r)+r(r-1)/2, \ldots, s+r(q-r)+r(r-1)/2+r$, are some non-negative numbers and $\sum_{i=s+r(q-r)+r(r-1)/2}^{s+r(q-r)+r(r-1)/2+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = 1$, χ_i^2 is a chi-square distribution with i degrees of freedom, and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$. In the model selection, we assume that model M_1 contains only the y-intercept for fixed effects and random intercept for random effects. Model M_k contains $(p_k + 1)$ fixed effects and the covariance matrix, $\boldsymbol{D}_k$, of random effects is of order $k \times k$. When random effects are assumed to be correlated, the number of distinct random effects variance and covariance components is $q_k = k(k+1)/2$. When we test model M_k against model M_1, applying (4.7) with $s = p_k$ and $r = k-1$, then $s+r(q-r)+r(r-1)/2 = p_k+k(k-1)/2$ and $s+r(q-r)+r(r-1)/2+r = p_k+(k-1)(k+2)/2$. Thus, the asymptotic null distribution of the log-likelihood ratio test statistic is approximated

by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=p_k+k(k-1)/2}^{p_k+(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \tag{4.8}$$

where $C = \mathbb{R}^{p_k} \times \{0\} \times \{0\} \times \mathbb{R}^{k(k-1)/2} \times \mathbb{R}_+^{k-1} \times \{0\}$.

Also, let h_k be the expectation of $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)$, then,

$$h_k = \sum_{i=p_k+k(k-1)/2}^{p_k+(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)i.$$

Our proposed modified BIC for this case is:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{4.9}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k; $n = \sum_{i=1}^N n_i$ and $d_k = 2.5 + h_k$ for $k > 1$; $d_k = p_k + 2.5$ for $k = 1$; and $d_k = p_k + 2$ for $k = 0$.

The corresponding regular BIC that we will use in our simulation is:

$$BIC(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + (p_k + 2 + k(k+1)/2) \log(n). \tag{4.10}$$

We also use the same setting for correlated random effects with correlation of 0.7 between random intercept and random slope of $\boldsymbol{z}_1$ in the true model. Table 4.2 shows the results of the complexity for the proposed BIC and regular BIC of a model

with $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$, and $\boldsymbol{X}_5$ as fixed effects and random effects as shown in the rows of the table.

Model	Proposed, d_k	Regular
Random Intercept (RI)	7.5	8
RI, $\boldsymbol{z}_1$	9	10
RI, $\boldsymbol{z}_2$	9	10
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$	11.52	13
RI, $\boldsymbol{z}_3$	9	10
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_3$	11.49	13
RI, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	11.50	13
RI, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$	15.08	17

Table 4.2: Comparing the complexity, d_k of the proposed BIC and regular BIC when random effects are correlated.

Given the same model, we calculate the difference between the complexity of model M_k using the regular BIC and the complexity using the proposed BIC. Then we take the average over the models with the same number of random effects. The average differences between the complexity of the regular BIC and the complexity of the proposed BIC are 0.5 , 1, 1.494, and 1.968 for $k = 1$, 2, 3, and 4, respectively, where k is the number of random effects in the model M_k, including the random intercept.

Table 4.3 presents the average difference between the complexity of the regular BIC and the complexity of the proposed BIC, where k is the number of random effects in the model M_k, including the random intercept.

Random Effects	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Independent	0.5	1	1.49	1.97
Correlated	0.5	1	1.47	1.89

Table 4.3: Average difference in Model Complexity between Proposed BIC and Regular BIC

We notice that the difference between the complexity of the regular BIC and the complexity of the proposed BIC is larger when the number of random effects in model M_k is bigger.

Theorem 4.1 *Theorem 2: Consistency of the modified BIC*

Assume that the assumptions $(C1) - (C4)$ *are satisfied and* $BIC^*(M_k)$ *is defined as in (4.9), then*

$$\lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1 \ \textit{for all}\ M_k \in M^+,$$

$$\textit{and} \qquad \lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1 \ \textit{for all}\ M_k \in M^-.$$

Please see a proof of Theorem 2 in the section below.

4.2.1 Proof to Theorem 2

Proof. **Case 1**: For any over-fitting model, $M_k \in M^+$, we also prove that $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$. Assume that model M_k contains p_k fixed effects and q_k random effects and the true model M_T contains p_T fixed effects and q_T random effects. Let $s = p_k - p_T$ and $r = q_k - q_T$ with $s \geq 0$, $q \geq 0$, and $s + r > 0$. Without loss of generality, assume that the covariance matrix of random effects in model M_k is $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}^T & \boldsymbol{D}_{22} \end{bmatrix}$ where $\boldsymbol{D}_{11}$ is the covariance matrix of random effects of the true model M_T. The size of $\boldsymbol{D}_{11}$ is $q_T \times q_T$ and the size of $\boldsymbol{D}_{22}$ is $r \times r$. Let $\boldsymbol{\theta}_T = (\underline{\mathbf{0}_{\boldsymbol{\beta}}}, {\boldsymbol{\beta}_T}^T, {\boldsymbol{\psi}_T}^T, \underline{\mathbf{0}}, \sigma^2_{\boldsymbol{\epsilon},T})^T$ and $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_{k,1}}^T, {\boldsymbol{\beta}_{k,2}}^T, {\boldsymbol{\psi}_{k,1}}^T, {\boldsymbol{\psi}_{k,2}}^T, \sigma^2_{\boldsymbol{\epsilon},k})^T$ where $\underline{\mathbf{0}_{\boldsymbol{\beta}}}$ has the same dimension as $\boldsymbol{\beta}_{k,1}$; and $\boldsymbol{\beta}_T$ has the same dimension as $\boldsymbol{\beta}_{k,2}$; $\boldsymbol{\psi}_T$ has the same dimension as $\boldsymbol{\psi}_{k,1}$ and $\underline{\mathbf{0}}$ has the same dimension as $\boldsymbol{\psi}_{k,2}$. All elements of $\underline{\mathbf{0}_{\boldsymbol{\beta}}}$ and $\underline{\mathbf{0}}$ are 0. We have that

$$BIC^*(M_k) - BIC^*(M_T) = -2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) + (d_k - d_T)\log(n). \quad (4.11)$$

Then $-2(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}))$ is the likelihood ratio test statistic of the following hypothesis test,

$$H_0 : {\boldsymbol{\beta}_{k,1}}^T = \mathbf{0}; \boldsymbol{D}_{11} > \mathbf{0}; \boldsymbol{D}_{12} = \mathbf{0}, \boldsymbol{D}_{22} = \mathbf{0},$$

$$H_1 : \boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{D} > \mathbf{0}.$$

As presented in the Background section of this chapter, under H_0, the asymptotic

distribution of $-2(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}))$ is approximated by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=s}^{s+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \tag{4.12}$$

where $C = \mathbb{R}^s \times \{0\}^{p_T} \times \{0\}^{q_T} \times \mathbb{R}_+^r \times \{0\}$; $w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C\right), i = s, \ldots, s+r$, are some non-negative numbers and $\sum_{i=s}^{s+r} w_i\left(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C\right) = 1$; χ_i^2 is a chi-square distribution with i degrees of freedom and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

Therefore, $-2(l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})) = O_p(1)$. We also have that,

$$\begin{aligned}
2(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y})) &= 2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{y}) - \left[l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{y})\right]\right) \\
&= -2\left(l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})\right) \\
&\quad - \left[-2\left(l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y})\right)\right]. \\
\Rightarrow E\left[2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{Y})\right)\right] &= E\left[-2\left(l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{Y})\right)\right] \\
&\quad - E\left[-2\left(l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{Y})\right)\right] = d_T - d_k,
\end{aligned}$$

where $l(\hat{\boldsymbol{\theta}}_1;\boldsymbol{y})$ is the maximum log-likelihood of the simplest model; that is, the model with only the intercept for fixed effects and random intercept for random effects. Therefore,

$$E\left[-2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{Y})\right)\right] = d_k - d_T.$$

On the other hand, $-2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y})\right)$ asymptotically has the distribution which is a mixture of the chi-square distributions. Therefore, $E\left[-2\left(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{Y})\right)\right]$

must be positive and, therefore, $d_k - d_T > 0$. Thus, $BIC^*(M_k) - BIC^*(M_T) \to \infty$ as $n \to \infty$ and $\lim_{n\to\infty} P\left(BIC^*(M_k) - BIC^*(M_T) > 0\right) = 1$ for $M_k \in M^+$.

Case 2: For any under-fitting model, $M_k \in M^-$, we want to prove that $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$. The proof for this part is similar to the proof for case 1 in theorem 1. We include the proof here for completeness. We have that

$$\begin{aligned}
BIC^*(M_k) &= -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \\
BIC^*(M_T) &= -2l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) + d_T \log(n), \\
BIC^*(M_k) - BIC^*(M_T) &= -2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) + (d_k - d_T)\log(n). \\
-2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\right) &= -2\Big(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - [l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y})] \\
&\quad - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) + l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\Big) \\
&= -2\left(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right) + 2\left[l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y})\right] \\
&\quad + 2\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right] - 2E\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{Y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right] \\
&\quad + 2E\left[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})\right].
\end{aligned}$$

We have that $l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) = o_p(1)$ and $l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) = o_p(1)$ because $\hat{\boldsymbol{\theta}}_k \xrightarrow{p} \boldsymbol{\theta}_{k,0}$ and $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_{T,0}$ and function $l(\boldsymbol{\theta}; \boldsymbol{y})$ is continuous with respect to $\boldsymbol{\theta}$. Also, under assumption $C4(ii)$, $\frac{1}{N}(l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y})]) \xrightarrow{p} 0$ and $\frac{1}{N}(l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - E_{T,0}[l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})]) \xrightarrow{p} 0$. Thus,

$$\frac{1}{N}(l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y}) - E[l(\boldsymbol{\theta}_{T,0}; \boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0}; \boldsymbol{y})]) \xrightarrow{p} 0.$$

Therefore, $l(\boldsymbol{\theta}_{T,0};\boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0};\boldsymbol{y}) - E[l(\boldsymbol{\theta}_{T,0};\boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0};\boldsymbol{y})] = o_p(N)$. The last term can be evaluated as,

$$\begin{aligned} E[l(\boldsymbol{\theta}_{T,0};\boldsymbol{y}) - l(\boldsymbol{\theta}_{k,0};\boldsymbol{y})] &= \sum_{i=1}^{N} E[\log f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0}) - \log f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})] \\ &= \sum_{i=1}^{N} E\left[\log \frac{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})}{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})}\right] = O_p(N). \end{aligned}$$

This is because $E\left[\log \frac{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})}{f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})}\right]$ is the Kullback-Leibler distance between $f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{T,0})$ and $f_i(\boldsymbol{Y}_i;\boldsymbol{\theta}_{k,0})$; and is positive and finite by assumption $C4$. Assume that the cluster sample sizes, $n_1,\ldots,n_N$ are uniformly bounded, then $O_p(N)$ dominates $(d_k - d_T)\log(n)$ as $N \to \infty$. Thus, $BIC^*(M_k) - BIC^*(M_T) > 0$. And, $\lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1$ for all $M_k \in M^-$. This completes the proof of Theorem 2. ■

4.3 Simulation

4.3.1 Simulation Set up

We generated data from linear mixed model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ with total number of observations $n = 1000$ and number of clusters, $N = 100$. $\boldsymbol{X}$ is a $n \times p$ matrix with $n = 1000$; $p = 6$; the first column of $\boldsymbol{X}$ includes all 1's. The last five columns, $\boldsymbol{X}_1$ to $\boldsymbol{X}_5$, are generated from the standard normal distribution. The vector of fixed effects, $\boldsymbol{\beta} = (1,2,3,1,0,0)^T$. Matrix $\boldsymbol{Z}$ contains the first three columns $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ which are the same as three columns of matrix $\boldsymbol{X}$ and three more columns $\boldsymbol{z}_3$, $\boldsymbol{z}_4$,

$\boldsymbol{z}_5$ are generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, are generated from multivariate normal distribution $N_q(\mathbf{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ is a 6×6 full matrix. The correlation matrix between the random effects in the data-generating model is:

$$R = \begin{bmatrix} 1 & 0.7 & 0.6 & 0.5 \\ 0.7 & 1 & 0.4 & 0.3 \\ 0.6 & 0.4 & 1 & 0.5 \\ 0.5 & 0.3 & 0.5 & 1 \end{bmatrix}.$$

To measure the ability to detect the significance of fixed effects and variance components parameter of the proposed BIC^*, we explore two different cases for different sizes of σ_0^2, σ_1^2, σ_2^2, σ_3^2 as shown below. σ_4^2, σ_5^2 and covariances corresponding to random effects of $\boldsymbol{z}_4$ and $\boldsymbol{z}_5$ are all 0. $\boldsymbol{\epsilon}$ is generated from a multivariate normal distribution, $N_n(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_n)$ with $\sigma_{\boldsymbol{\epsilon}}^2 = 1$.

Case 1: The standard deviations of random effects are $\sigma_0 = 5$, $\sigma_1 = 1.5$, $\sigma_2 = 1$, $\sigma_3 = 0.5$, $\sigma_4 = 0$, $\sigma_5 = 0$.

Case 2: The standard deviations of random effects are $\sigma_0 = 2$, $\sigma_1 = 0.8$, $\sigma_2 = 0.5$, $\sigma_3 = 0.3$, $\sigma_4 = 0$, $\sigma_5 = 0$.

4.3.2 Simulation Procedure

For each case above, we run $B = 1001$ simulations. In each simulation, all possible candidate models are run. All models contains the y-intercept for fixed effect and random intercept for random effects. The covariates for fixed effects part vary in the power set of $\{1, 2, 3, 4, 5\}$ for $\boldsymbol{X}_1$ to $\boldsymbol{X}_5$ and the covariates for random effects part vary in the power set of $\{1, 2, 3, 4, 5\}$ for $\boldsymbol{z}_1$ to $\boldsymbol{z}_5$. We also include the models that includes only y-intercept for fixed effect with varying random effects and the models that includes random intercept only with varying fixed effects. The proposed BIC^* and regular BIC were calculated for each model. Then the model with minimum proposed BIC is selected and the model with minimum regular BIC is selected. For each selection criterion, we have 1001 models obtained from 1001 simulations. We then calculate the means and standard deviations of Positive Selection Rate and False Discovery Rate; and the correction rate for each criterion.

We also run simulations for the case when random effects are assumed to be uncorrelated and the variances of random effects are the same as the values in case 1 and case 2.

4.3.3 Simulation results

Table 4.4 shows the comparison of the Proposed BIC, Regular BIC, and cAIC methods in terms of Fixed Effects Correction Rate, Random Effects Correction Rate, and Both Effects Correction Rate for both case 1 and case 2 when random effects are assumed to be correlated.

	Proposed BIC			Regular BIC			cAIC		
Case	FE-CR	RE-CR	Both-CR	FE-CR	RE-CR	Both-CR	FE-CR	RE-CR	Both-CR
1	0.983	0.999	0.982	0.982	0.997	0.979	0.3147	0.3177	0.1708
2	0.975	0.6673	0.6503	0.979	0.5684	0.5554	0.3377	0.3746	0.2128

FE-CR is the correction rate for fixed effects variables; RE-CR is the correction rate for random effects variables; and Both-CR is the correction rate of selecting the true model. All the rates are calculated over 1001 simulations.

Table 4.4: Comparison of the Proposed BIC,Regular BIC, and cAIC methods in terms of the Correction Rate for Fixed Effects, Random Effects, and Both for different values of σ_0, σ_1, σ_2 with σ_3, $\sigma_4 = 0$, $\sigma_5 = 0$ and correlated random effects.

Based on the simulation results for the situation when random effects are assumed correlated in table 4.4, we see that the proposed BIC method performs better than the regular BIC and the cAIC methods in terms of the correction rate for selecting the fixed effects, the correction rate for selecting the random effects and also for selecting both fixed effects and random effects simultaneously. We also see that when the values of the variances for random effects are smaller, the correction rates are lower for all methods. However, the performance of the proposed method is still much better than the other two methods.

When random effects are assumed uncorrelated, based on the simulation results in table 4.5, we see that the proposed BIC and regular BIC still perform well and better than the cAIC method. The proposed BIC method performs better than the regular BIC in case 2 but doesn't performs better than the regular BIC in case 1. This may be because the penalty term of the regular BIC is calculated using the exact chi-square distribution and the calculation of the penalty term is without any error. However, for the proposed BIC, the weights of the chi-bar distribution are approximated. Therefore, the penalty term is approximated only.

	Proposed BIC			Regular BIC			cAIC		
Case	FE-CR	RE-CR	Both-CR	FE-CR	RE-CR	Both-CR	FE-CR	RE-CR	Both-CR
1	0.985	0.9481	0.9351	0.986	0.994	0.981	0.5105	0.4865	0.2667
2	0.980	0.8661	0.8511	0.981	0.7672	0.7532	0.5664	0.5385	0.3017

FE-CR is the correction rate for fixed effects variables; RE-CR is the correction rate for random effects variables; and Both-CR is the correction rate of selecting the true model. All the rates are calculated over 1001 simulations.

Table 4.5: Comparison of the Proposed BIC, Regular BIC, and cAIC methods in terms of Fixed Effects Correction Rate, Random Effects Correction Rate, and Both Effects Correction Rate for different values of σ_0, σ_1, σ_2, σ_3 with $\sigma_4 = 0$, and $\sigma_5 = 0$ and independent random effects.

From the simulation results, we notice that when the values of the variances for random effects are smaller, the correction rates are lower for the proposed and regular BIC methods. However, the correction rates in case 2 are better than case 1 for the cAIC method.

4.4 Real-Data Application

In this section, we again apply the proposed BIC, regular BIC, and cAIC methods to the subset of 120 schools of dataset "hsfull" as used in the previous chapters. Data are on math achievement test scores from 50 Catholic and 70 Public high schools. The number of students in each school ranges from 19 to 66 students. Total number of students is 5307.

Variables are school identification number, math achievement score ($\boldsymbol{Y}$), socioeconomic status ($\boldsymbol{X}_1$), sex (female or male; $\boldsymbol{X}_2$), visible minority status (yes/no; $\boldsymbol{X}_3$), school sector (catholic or public; $\boldsymbol{X}_4$). We wish to study the relationship between students' math achievement score and socioeconomic status, sex, visible minority status in public and catholic schools; and whether this relationship varies across schools within each sector.

The candidate variables in the fixed effects part are $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ which are group-centered. The candidate variables in the random effects part are $\boldsymbol{z}_1$, $\boldsymbol{z}_2$,

$\boldsymbol{z}_3$ which are the same as $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$.

We first fit a linear mixed model which includes only y-intercept and random intercept. Then, we fit the models with only y-intercept for fixed effects and all possible combinations of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$ with random intercept for random effects. Next, we fit models with all possible combinations of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ for fixed effects and only random intercept for random effects. Lastly, for each combination of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ for fixed effects, we fit models with all possible combinations of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{z}_3$ with random intercept for random effects. For each model, we record the values of proposed BIC, regular BIC, and cAIC. There are $(2^4) * (2^3)$ or $16 * 8 = 128$ values for each method. Now, for each method, we choose the model with minimum value of the corresponding criterion. We apply this procedure for both cases when random effects are assumed to be correlated and uncorrelated.

When random effects are assumed to be correlated, the optimal model we obtain using the proposed BIC is the model with all $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ and random intercept; the proposed BIC is 34379.83. The optimal model we obtain using the regular BIC is also the model with $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ and random intercept only. The regular BIC of this model is also 34379.83. The cAIC yields the optimal model which contains $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ with random intercept, and random slopes of $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$. The cAIC of the optimal model is 34166.25.These results are consistent with

the results obtained in the in the Real-data Application section of chapter 3.

When random effects are assumed to be uncorrelated, the optimal model we obtain using the proposed BIC is the model with all $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$, random intercept, and random slopes of $\boldsymbol{z}_3$; the proposed BIC value is 34378.23. The optimal model we obtain using the regular BIC is the model with $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ and random intercept only. The regular BIC of this model is 34379.83. The cAIC yields the optimal model which contains $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ with random intercept, and random slopes of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$. The cAIC of the optimal model is 34165.13. These results are also consistent with the results obtained in the Real-data Application section of chapter 2. Based on the results presented above, we would choose the model with all $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ for fixed effects; and random intercept, and random slope of $\boldsymbol{z}_3$ for random effects assuming that random effects are uncorrelated. There is a significant relationship between students' math achievement score and socioeconomic status, sex, visible minority status in public and catholic schools; and the school mean math achievement score and minority gap effect vary across schools within each sector.

5 Predictive Models for Diabetes Mellitus using Machine Learning Techniques

5.1 Abstract

Background: Diabetes Mellitus is an increasingly prevalent chronic disease characterized by the body's inability to metabolize glucose. The objective of this study was to build an effective predictive model with high sensitivity and selectivity to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results during their visits to medical facilities. **Methods**: Using the most recent records of 13309 Canadian patients aged between 18 and 90 years, along with their laboratory information (age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein), we built predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques. The area under the receiver operat-

ing characteristic curve (AROC) was used to evaluate the discriminatory capability of these models. We used the adjusted threshold method and the class weight method to improve sensitivity – the proportion of Diabetes Mellitus patients correctly predicted by the model. We also compared these models to other learning machine techniques such as Decision Tree and Random Forest.

Results: The AROC for the proposed GBM model is 84.7% with a sensitivity of 71.6% and the AROC for the proposed Logistic Regression model is 84.0% with a sensitivity of 73.4%. The GBM and Logistic Regression models perform better than the Random Forest and Decision Tree models.

Conclusions: The ability of our models to predict patients with Diabetes using some commonly used lab results is high with satisfactory sensitivity. These models can be built into an online computer program to help physicians in predicting patients with future occurrence of diabetes and providing necessary preventive interventions. The model is developed and validated on the Canadian population which is more specific and powerful to apply on Canadian patients than existing models developed from US or other populations. Fasting blood glucose, body mass index, high-density lipoprotein, and triglycerides were the most important predictors in these models.

5.2 Background

Diabetes Mellitus (DM) is an increasingly prevalent chronic disease characterized by the body's inability to metabolize glucose. Finding the disease at the early stage helps reduce medical costs and the risk of patients having more complicated health problems. Wilson et al. (2007) developed the Framingham Diabetes Risk Scoring Model (FDRSM) to predict the risk for developing DM in middle-aged American adults (45 to 64 years of age) using Logistic Regression. The risk factors considered in this simple clinical model are parental history of DM, obesity, high blood pressure, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose. The number of subjects in the sample was 3140 and the area under the receiver operating characteristic curve (AROC) was reported to be 85.0%. The performance of this algorithm was evaluated in a Canadian population by Mashayekhi et al. (2015) using the same predictors as Wilson et al. (2007) with the exception of parental history of DM. The number of subjects in the sample was 4403 and the reported AROC was 78.6%. Data mining techniques have been widely used in DM studies to explore the risk factors for DM as in Iyer et al. (2015), Ioannis et al. (2017),Kahn et al. (2009), and Lindström and Tuomilehto (2003). Machine learning methods, such as logistic regression, artificial neural network, and decision tree were used by Meng et al. (2013) to predict DM and pre-diabetes. The data

included 735 patients who had DM or pre-diabetes and 752 who are healthy from Guangzhou, China. The accuracy was reported to be 77.87% using a decision tree model; 76.13% using a logistic regression model; and 73.23% using the Artificial Neural Network (ANN) procedure. Other machine learning methods, such as Random Forest, Support Vector Machines (SVM), k-nearest Neighbors (KNN), and the naïve Bayes have also been used as in Ioannis et al. (2017), Jayalakshmi and Santhakumaran (2010), Kahn et al. (2009), and Meng et al. (2013). Sisodia and Sisodia (2018) recently used three classification algorithms: Naïve Bayes, Decision Tree, and SVM, to detect DM. Their results showed that Naïve Bayes algorithm works better than the other two algorithms.

In this chapter, we present predictive models using Gradient Boosting Machine and Logistic Regression techniques to predict the probability of patients having DM based on their demographic information and laboratory results from their visits to medical facilities. We also compare these methods with other widely used machine learning techniques such as Rpart and Random Forest. The MLR (Machine Learning in R) package in R Bischl et al. (2016) was used to develop all the models.

5.3 Methods

The data used in this research were obtained from CPCSSN (www.cpcssn.ca). The case definition for diabetes is described in Williamson et al. (2014). "Diabetes includes diabetes mellitus type 1 and type 2, controlled or uncontrolled, and excludes gestational diabetes, chemically induced (secondary) diabetes, neonatal diabetes, polycystic ovarian syndrome, hyperglycemia, prediabetes, or similar states or conditions" (page 4 in Williamson et al. (2014)). The dataset was generated as follows: 1) Every blood pressure reading (over 6 million) were pulled into a table for all patients over the age of 17 along with the patient ID, their age on the date of the exam and their sex. 2) For each blood pressure reading, we joined the following records that were closest in time, within a specific time period, based on the type of measurement: BMI $\pm$ 1 year, LDL $\pm$ 1 year, HDL $\pm$ 1 year, triglyceride (TG) $\pm$ 1 year, Fasting blood sugar (FBS) $\pm$ 1 month, HbA1c $\pm$ 3 months. 3) We removed records with missing data in any one of the columns. This left approximately 880000 records, of which approximately 255000 records were from patients who have diabetes. 4) Patients on insulin, who might have Type 1 diabetes, and patient on corticosteroids, which can affect blood sugar levels, were removed from the dataset, leaving 811000 records with 235000 from patients with DM. 5) We then curated a dataset for records of patients that preceded the onset of DM and identified those

patients for whom there were at least 10 visits worth of data. For patients who had not developed DM, we removed the last year of records before the end of the database to minimize the impact of patients who might be on the verge of becoming diabetic. There are 215544 records pertaining to patient visits in the dataset. The outcome variable is Diabetes Mellitus which is encoded a binary variable, with category 0 indicating patients with no DM and category 1 indicating patients with DM. The predictors of interest are: Sex, Age (Age at examination date), BMI (Body Mass Index), TG (Triglycerides), FBS (Fasting Blood Sugar), sBP (Systolic Blood Pressure), HDL (High Density Lipoprotein), and LDL (Low Density Lipoprotein). Since a patient may have multiple records representing their multiple visits to medical facilities, we took each patient's last visit to obtain a dataset with 13317 patients. In the exploratory data analysis step, we found some extreme values in BMI and TG, and thereafter, excluded these values to obtain a final analysis dataset with 13309 patients. About 20.9% of the patients in this sample have DM. 40% of the patients are male and about 60% are female. The age of the patients in this dataset ranges from 18 to 90 years with a median of around 64 years. Age is also encoded as a categorical variable represented by the four categories: Young, Middle-Aged, Senior, and Elderly. About 44.6% of patients are middle-aged, between 40 and 64 years old; 47.8% are senior, between 65 and 84; 4.8% are elderly who are older than 85; and

2.9% are younger than 40 years old. Body mass index was calculated by dividing the patient's weight (in kilograms) by the patient's height (in meters) squared. The body mass index ranges from 11.2 to 70 with a median of 28.9. The distributions of BMI, FBS, HDL and TG are all right-skewed.

Group	BMI	FBS	HDL	TG	LDL	sBP	Age
DM	31.16	6.10	1.20	1.56	2.71	130	64.00
No DM	28.32	5.20	1.40	1.24	2.74	130	66.00

All values are rounded to two decimal places.

Table 5.1: Comparing the median of continuous variables between DM and No DM groups.

Table 5.1 shows that the medians of BMI, FBS, and TG of the group of patients with DM are higher than those of the group of patients with no DM; the median HDL is higher for the group of patients with no DM meanwhile the median LDL, median sBP, and the median Age are similar. The correlation matrix of the continuous variables (Age, BMI, TG, FBS, sBP, HDL, LDL) shows no remarkable correlation among the variables, except for a moderate negative correlation of -0.39 between HDL and TG.

Gradient Boosting Machine is a powerful machine-learning technique that has

shown considerable success in a wide range of practical applications (Natekin and Knoll (2013)). In this research study, we used Logistic Regression and Gradient Boosting Machine techniques in the MLR package in R to build predictive models. We then compared these methods to two other modern machine-learning techniques which are Decision Tree Rpart and Random Forest.

Procedure: We first created a training dataset by randomly choosing 80% of all patients in the dataset and created a test dataset with the remaining 20% of patients. The training dataset has 10647 patients and the test dataset has 2662 patients. We used the training dataset to train the model and used the test dataset to evaluate how well the model performs based on an unseen dataset. Using the training dataset and the 10-fold cross-validation method, we tuned the model hyperparameters to obtain the set of optimal hyperparameters that yields the highest area under the receiver operating characteristic curve (AROC). Since the dataset is imbalanced with only 20.9% of the patients in the DM group, we used different misclassification costs to find the optimal threshold (or the cut off value) for the DM class (i.e., Diabetes Mellitus = 1). In the tuning threshold approach, we set up a matrix of misclassification costs in which the diagonal elements are zero and the ratio of the cost of a false negative to the cost of a false positive is 3 to 1. We validated the model with the

optimal hyperparameters using a 10-fold cross validation. In this step, we measured both AROC values and the misclassification costs. We tuned the threshold for the positive class (Diabetes = 1) by choosing the threshold that yields the lowest expected misclassification cost. We obtained our final model by fitting the model with the optimal set of hyperparameters on the entire training dataset. Finally, using the optimal threshold we evaluated the performance of the final model on the test dataset. Sensitivity was calculated by dividing the model-predicted number of DM patients by the observed number of DM patients. Specificity was calculated by dividing the model-predicted number of No DM patients by the observed number of No DM patients. The misclassification rate is the number of incorrectly classified patients divided by the total number of patients.

We next overview the tuning process for each model. The training data set was used in the tuning process.

1. *GBM model*

    - **The search space**: We create a search space by defining a parameter grid of hyperparameters as follows: the number of trees (n.trees) is an integer ranging from 200 to 600; the depth of tree (interaction.depth) is from 2 to 6; the minimum number of observations in the terminal nodes (n.minobsinnode) is from 30 to 80; and learning rate (shrinkage) is from

0.01 to 0.3.

- **Tuning method**: We perform random search on the parameter space with the maximum number of iterations of 100.
- **Evaluation method**: We use 10-fold cross validation and use AROC as the performance measure.
- **The tuning results are as follows**: the number of iterations (n.trees) is 257; the interaction depth (interaction.depth) is 2; the minimum number of observations in the terminal nodes (n.minobsinnode) is 75; the shrinkage rate (shrinkage) is 0.126. The average AROC is 83.6%.

2. *Logistic Regression model*: There are no hyperparameters for the Logistic Regression models so we do not use the tuning process for Logistic Regression models.

3. *Random Forest model*:

   - **Parameter grid**: the number of trees to grow (ntree) is from 80 to 500; the number of variables should be selected at a node split (mtry) is an integer ranging from 3 to 6; the number of observations at terminal nodes (nodesize) is from 20 to 50.
   - **Tuning method**: We perform random search on the parameter space

with the number of iterations of 100.

- **Evaluation method**: We use 10-fold cross validation and use AROC as the performance measure.
- **The tuning results are as follows**: the number of trees to grow (ntree) is 407; the number of variables should be selected at a node split (mtry) is 3; the number of observations at terminal nodes (nodesize) is 22. The average AROC is 82.9%.

4. *Decision Tree model*

- **Parameter grid**: the smallest number of observations in the parent node that could be split further (minsplit) is from 30 to 50; the smallest number of observations that are allowed in a terminal node (minbucket) is an integer ranging from 10 to 50; depth of tree (maxdepth) can be 8, 12, 16, or 30; the complexity parameter (cp) is from 0.001 to 0.2.
- **Tuning method**: We perform random search on the parameter space with the maximum number of iterations of 100.
- **Evaluation method**: We use 10-fold cross validation and use AROC as the performance measure.
- **The tuning results are as follows**: The smallest number of observa-

tions in the parent node that could be split further (minsplit) is 41; the smallest number of observations that are allowed in a terminal node (minbucket) is 19; depth of tree (maxdepth) is 8; the complexity parameter (cp) is from 0.001. The average AROC is 76.7%.

5.4 Results

The optimal set of hyperparameters we obtained for this GBM model is as follows: the number of iterations (n.trees) is 257; the interaction depth (interaction.depth) is 2; the minimum number of observations in the terminal nodes (n.minobsinnode) is 75; the shrinkage rate (shrinkage) is 0.126. Since the outcome variable is a binary variable, we used the Bernoulli loss function and tree-based learners in this GBM model. Using the cross-validation method to validate this model, we obtained AROC values ranging from 81.6% to 85.0% with an average AROC of 83.6%, indicating a high reliability of the method. The optimal threshold for the DM class using the misclassification cost matrix method is 0.24. We also used the train/test split method to validate this model and obtained similar results with average AROC of 83.3%.

When testing the model on the test dataset we obtained the following results: the AROC is 84.7%; the misclassification rate is 18.9%; the sensitivity is 71.6% and the specificity is 83.7%. We observed that there is a trade off between the sensitivity

and the misclassification rate. Using a default threshold of 0.5, the misclassification rate for the GBM model was 15%; the sensitivity was low at 48.3%; the specificity was 95.2%; and the AROC remained the same at 84.7%. For our Logistic Regression model, the AROC was 84.0%; the misclassification rate was 19.6%; the sensitivity was 73.4% and the specificity was 82.3%. The optimal threshold was estimated to be 0.24 and Age was treated as a categorical variable in this model. We validated this model using the cross-validation method and obtained AROC values ranging from 80.6% to 85.7% with an average AROC of 83.2%. Fasting blood glucose, high-density lipoprotein, body mass index, and triglycerides were very significant predictors in this model ($P < 0.0001$). Interestingly, based on this sample data, we found that age was also a significant factor (Table 5.2); elderly and senior patients significantly have lower chance of having DM than the middle-aged patients, given that all other factors are kept the same. Checking the model assumptions, we found no severe collinearity; all variables had a variance inflation factor (VIF) values less than 1.5. Variables FBS, SBP, TG, and BMI were all strongly linearly associated with the DM outcome on the logit scale. With respect to standardized residuals, there were 9 outliers ranging from 3.1 to 3.4. Since the number of potential influential observations was not large, all patients were kept in the dataset.

Variables	Estimated Coefficient	OR	95% CI for OR	P Value
Intercept	-11.816	–	–	< 0.0001
Age				
Middle-Aged (40-64)	(Reference)	1.000	–	–
Elderly (85-90)	-0.829	0.436	(0.31, 0.61)	< 0.0001
Senior (65-84)	-0.127	0.881	(0.78, 0.99)	0.036
Young (¡ 40)	0.238	1.269	(0.90, 1.79)	0.170
Male	-0.250	0.779	(0.69, 0.88)	< 0.0001
FBS	1.963	7.122	(6.45, 7.87)	< 0.0001
BMI	0.023	1.024	(1.01, 1.03)	< 0.0001
HDL	-0.894	0.409	(0.34, 0.49)	< 0.0001
TG	0.158	1.171	(1.09, 1.26)	< 0.0001
sBP	-0.001	0.999	(0.96, 1.00)	0.560
LDL	-0.011	0.990	(0.93, 1.05)	0.740

OR = Odds Ratio. All values are rounded.

Table 5.2: Predictors associated with the Logistic Regression Model.

Figure 5.1 shows the Information Gain measure from the potential predictors. Based on the information gain criterion which measures the amount of information gained by each predictor, we also found that fasting blood glucose is the most important predictor, followed by high-density lipoprotein, body mass index, and triglycerides; then age, sex, blood pressure, and low-density lipoprotein (Figure 5.1).

To compare the performance of the obtained Logistic Regression and GBM models with other machine-learning techniques, we used the same training dataset, test dataset, and procedure on the Rpart and Random Forest techniques. The AROC values from the models are presented in Table 5.3.

Model	Area Under the ROC Curve, AROC
GBM	84.7%
Logistic Regression	84.0%
Random Forest	83.4%
Rpart	78.2%

All values are rounded to two decimal places.

Table 5.3: Comparing the AROC values with other machine-learning techniques.

The results in Table 5.3 show that the GBM model performs the best based on highest AROC value, followed by the Logistic Regression model and the Random Forest model. The Rpart model gives the lowest AROC value at 78.2%. Figure 5.2 illustrates the Receiver Operating Curves (ROC) curves of the Rpart, Random Forest, Logistic Regression, and GBM models.

Our models can be implemented in practice. For the Logistic Regression model, we outline an algorithm for estimating the risk of DM. sBP and LDL were excluded

from this model as their contributions were not statistically significant.

Algorithm: Step 1: Calculate the estimated linear predictor, η. Step 2: Calculate the risk of having DM risk $= \exp(\eta)/(1+\exp(\eta))$. Step 3: If the risk is 0.24 or more, then the patient has a high chance of having DM.

Age from 40 to 64	Male	$\eta = -12.066 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
	Female	$\eta = -11.816 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
Age from 65 to 84	Male	$\eta = -12.193 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
	Female	$\eta = -11.943 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
Age from 85 to 90	Male	$\eta = -12.895 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
	Female	$\eta = -12.645 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
Age from 18 to 39	Male	$\eta = -11.828 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$
	Female	$\eta = -11.578 + 1.963FBS + 0.023BMI - 0.894HDL + 0.158TG$

All coefficients are kept to three decimal places.

Table 5.4: Calculation of η.

For the GBM model, it is more difficult to display the equations explicitly. However, it is feasible to set up an online real-time DM risk predictor program so that a patients' risk of developing DM can be reported when the patient's predictor values are entered. The trained GBM model can be saved in the Predictive Model Markup Language (PMML) format, which is an XML-based format, using the pack-

age r2pmml in R. Thereafter, the model can be deployed to make predictions using a Java platform (Scoruby and Goscore packages) or the Yellowfin platform.

To compare the performance of the four models, we conducted 10-fold cross validation on the whole dataset with the following steps:

1. Divide data set into 10 parts. Use 9 parts as training data set and the last part as the testing data set.

2. Train the four 4 models on the training data set.

3. Measure AROC for each model based on the testing data set

4. Repeat for all 10 folds.

Shuffle the whole data set and repeat the above procedure 2 more times.

Based on 30 values of AROC obtained for each model (with age is treated as a continuous variable), we are able to estimate the mean of their AROC values.

Model	Mean of AROC value
GBM	83.9%
Logistic Regression	83.5%
Random Forest	83.0%
Rpart	77.1%

All values are rounded to two decimal places.

Table 5.5: Mean of AROC for the four models from the cross-validation results.

We also created a box plot to compare the AROC values of the four models. Figure 5.3 presents a Box plot to compare the AROC values of the four models in the cross-validation results. The box plot shows that the medians of AROC values for GBM, Logistic Regression and Random Forest are quite close to each other and they are all greater than that of the Rpart model.

Due to the independence and normality assumptions of the t-test, it may not be safe to use the paired t-test for testing equality between the mean AROC values for any two models based on the AROC values we obtained. Therefore, to estimate the consistency of the predictive power for each model, we used the DeLong et al. (1988) to find the standard deviation and the 95% confidence interval for the AROC value of each model. We also used the DeLong method to compare the AROC values of two correlated ROC curves. For each pair, we wanted to test the equality of AROCs of two ROC curves and whether the AROC value of the first mode is significantly greater than that of the second model. The DeLong method is a nonparametric method that was implemented in pROC package in R Robin et al. (2011). The obtained results are presented in Tables 5.6 and 5.7.

Table 5.6 presents the estimated AROC values, Standard Deviations, and 95% Confidence Intervals of the AROC values for the four models using the DeLong method. The standard deviations are small and the confidence intervals are not

wide. This indicates that the values of AROC of the four models are consistent.

Model	AROC value	Standard deviation	95% CI
GBM	84.5%	0.97%	(82.6%, 86.4%)
Logistic Regression	84.1%	1.01%	(82.1%, 86.1%)
Random Forest	83.2%	1.05%	(81.1%, 85.2%)
Rpart	78.1%	1.10%	(76.0%, 80.3%)

All values are rounded to two decimal places.

Table 5.6: Mean of AROC for the four models from the cross-validation results.

Table 5.7 shows the paired one-sided DeLong test to compare the AROC values of the four models.

Test Name	z-statistic	p-value
GBM vs. Logistic Regression	1.392	0.081
GBM vs. Random Forest	3.885	5.13e-05
GBM vs. Rpart	8.914	2.20e-16
Logistic Regression vs. Random Forest	2.038	0.021
Logistic Regression vs. Rpart	8.006	5.95e-16
Random Forest vs. Rpart	7.028	1.05e-12

All values are rounded to two decimal places.

Table 5.7: Paired one-sided DeLong test to compare the AROC values of the four models.

The results in Table 5.7 show that the AROC value of the GBM model is significantly greater than that of Random Forest, and Rpart models ($P < 0.001$), but not significantly greater than that of Logistic Regression model ($P > 0.05$). The Logistic Regression model also has an AROC value greater than that of Random Forest and of Rpart. The AROC of Random Forest model is significantly greater than that of Rpart model, as well. We also noted that the comparison of the tests are statistically significant but this relative performance may be restricted to the specific population and data we are dealing with.

To see how our models work on a different data set, we used Pima Indians Dataset

which is publicly available Patel, Ashish (2018). All patients in this data set are females at least 21 years old of Pima Indian heritage. There are 768 observations with 9 variables as followings: Pregnant, number of times pregnant; Glucose, plasma glucose concentration (glucose tolerance test); BP, diastolic blood pressure (mm/Hg); Thickness (triceps skin fold thickness (mm)); Insulin (2-Hour serum insulin (mu U/ml); BMI (body mass index (weight in kg/(height in m) squared)); Pedigree (diabetes pedigree function); Age (Age of the patients in years); Diabetes (binary variable with 1 for Diabetes and 0 for No Diabetes). When working on this data set, we noticed that there are many rows with missing data and the missing values in Glucose, BP, Thickness, and BMI are labeled as 0. For example, about 48.7% of Insulin values are missing. For purpose of validating our methods, we chose not to impute the data but excluded all rows with missing values. There are 392 observations left in the working data set in which 130 patients with diabetes and 262 without diabetes. We applied our methods on this dataset to predict whether or not a patient has diabetes. We also divided the PIMA data set into the training data set 80% of the observations) and the testing data set 20% of the observations). We trained the four models on the training data set and validate the models on the testing data set. On the testing data set, we obtained the AROC of 84.7% for GBM model, 88.0% for Logistic Regression Model, 87.1% for Random Forest Model, and

77.0% for Rpart model.

We also conducted 10-fold cross-validation and repeated the procedure for two more times. Table 5.8 presents the results based on the 30 AROC values from the cross-validation results conducted on the PIMA Indian data set.

Model	Mean of AROC value
GBM	85.1%
Logistic Regression	84.6%
Random Forest	85.5%
Rpart	80.5%

All values are rounded to two decimal places.

Table 5.8: Comparing the AROC values of the four models using PIMA Indian data set.

The results we obtained for this data set are quite consistent with what we observed in our main data set. Based on these results, GBM, Logistic Regression, and Random Forest are comparable and they all give higher mean AROC than that of the Rpart model on the testing data set. We also created a box plot to compare the sampling distributions of the AROC values for the four models. Figure 5.4 shows a Box plot of AROC values for the Rpart, Random Forest, Logistic Regression, and

GBM models applied to PIMA Indian data set. The box plot shows that the variability in the AROC values of GBM, Logistic Regression, and Random Forest are quite the same and less than that of the Rpart model.

5.5 Discussion

In this research study, we used the Logistic Regression and GBM machine learning techniques to build a model to predict the probability that a patient develops DM based on their personal information and recent laboratory results. We also compared these models to other machine learning models to see that the Logistic Regression and GBM models perform best and give highest AROC values. During the analysis, we also used the class weight method for our imbalanced dataset. We first tuned the class weight for the DM class to find the optimal class weight that minimized the average classification cost. We found that the optimal class weight for the GBM model is 3 and the optimal class weight for the Logistic Regression is 3.5. These optimal class weights are then incorporated into the model during the training process. We obtained similar results for GBM, Logistic Regression, and Random Forest model. However, the Decision Tree Rpart model gives a higher AROC at 81.8% compared to 78.2% when the threshold adjustment method was used. We also applied a natural logarithmic transformation on the continuous variables, however, this did not

improve AROC and sensitivity. Compared to the simple clinical model presented by Wilson et al. (2007), the AROC value from our GBM model was very similar. The AROC value of our Logistic Regression model was lower, given the fact that the parental history of the disease was not available in our sample data. We also note that the characteristics of the sample data used in this study were not the same as the ones used by Wilson et al. (2007). For example, the age of the patients in our dataset ranges from 18 to 90, while the patients studied by Wilson et al. (2007) ranges from 45 to 64. Schmid et al. (2011) conducted a study on Swiss patients to compare different score systems used to estimate the risk of developing type 2 diabetes such as the 9-year risk score from Balkau et al. (2008), the Finnish Diabetes Risk Score (FINDRISC) Lindström and Tuomilehto (2003), the prevalent undiagnosed diabetes risk score from Griffin et al. (2000), 10-year-risk scores from Kahn et al. (2009), 8-year risk score from Wilson et al. (2007), and the risk score from the Swiss Diabetes Association. Their results indicated that the risk for developing type 2 diabetes varies considerably among the scoring systems studied. They also recommended that different risk-scoring systems should be validated for each population considered to adequately prevent type 2 diabetes. These scoring systems all include the parental history of diabetes factor and the AROC values reported in these scoring systems range from 71% to 86%. Mashayekhi et al. (2015) had previously applied

Wilson's simple clinical model to the Canadian population. Comparing our results to the results reported by Mashayekhi et al. (2015), the AROC values suggest that our GBM and Logistic Regression models perform better with respect to predictive ability. Using the same continuous predictors from the simple clinical model with the exception of parental history of diabetes, we also obtained an AROC of 83.8% for the Logistic Regression model on the test dataset.

5.6 Conclusion

The main contribution of our research study was proposing two predictive models using machine-learning techniques, Gradient Boosting Machine and Logistic Regression, in order to identify patients with high risk of developing DM. We applied both the classical statistical model and modern learning-machine techniques to our sample dataset. We dealt with the issue of imbalanced data using the adjusted-threshold method and class weight method. The ability to detect patients with DM using our models is high with fair sensitivity. These predictive models are developed and validated on Canadian population reflecting the risk patterns of DM among Canadian patients. These models can be set up in a computer program online to help physicians in assessing Canadian patients' risk of developing Diabetes Mellitus.

Abbreviations: DM (Diabetes Mellitus), BMI (Body Mass Index), TG (Triglyc-

erides), FBS (Fasting Blood Sugar), sBP (systolic Blood Pressure), HDL (High Density Lipoprotein), LDL (Low Density Lipoprotein), AROC (Area under the Receiver Operating Characteristics curve), GBM (Gradient Boosting Machine).

5.7 Figures

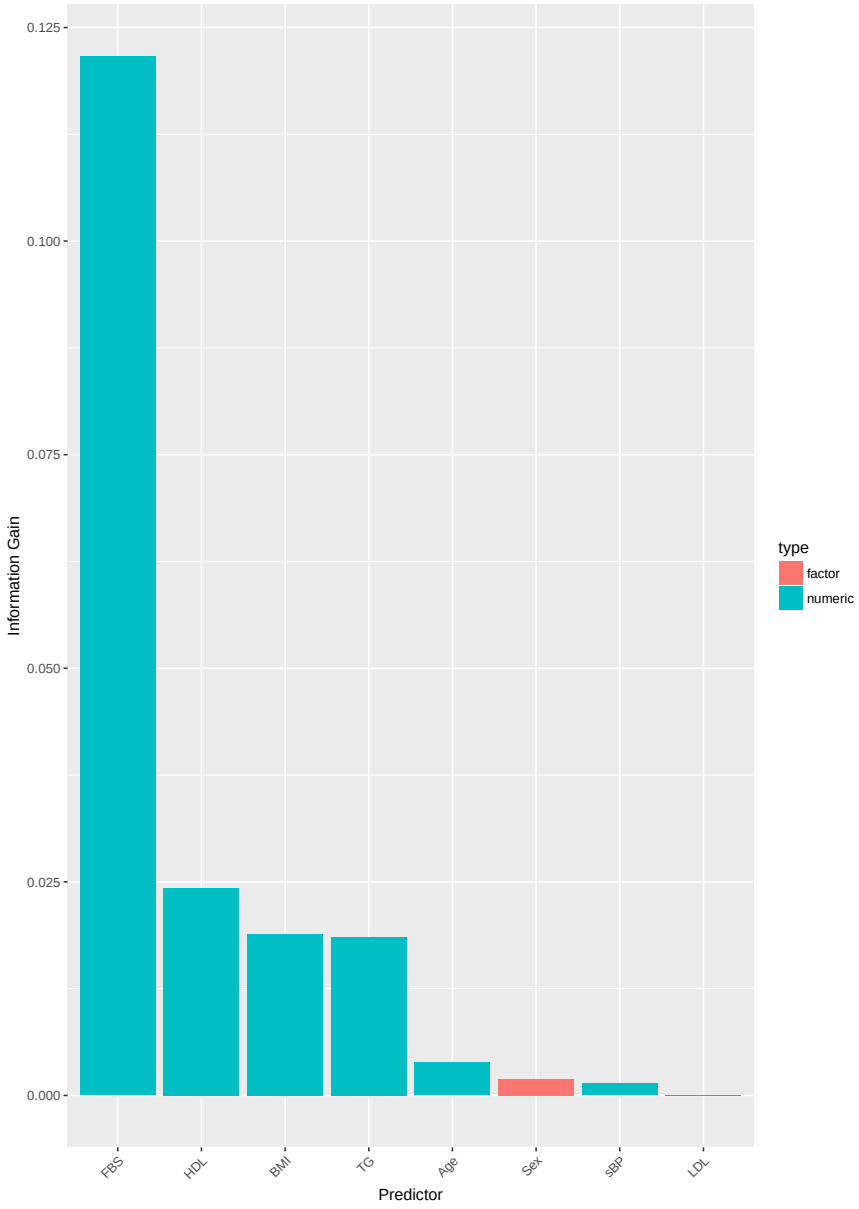

Figure 5.1: Information Gain measure from the potential predictors

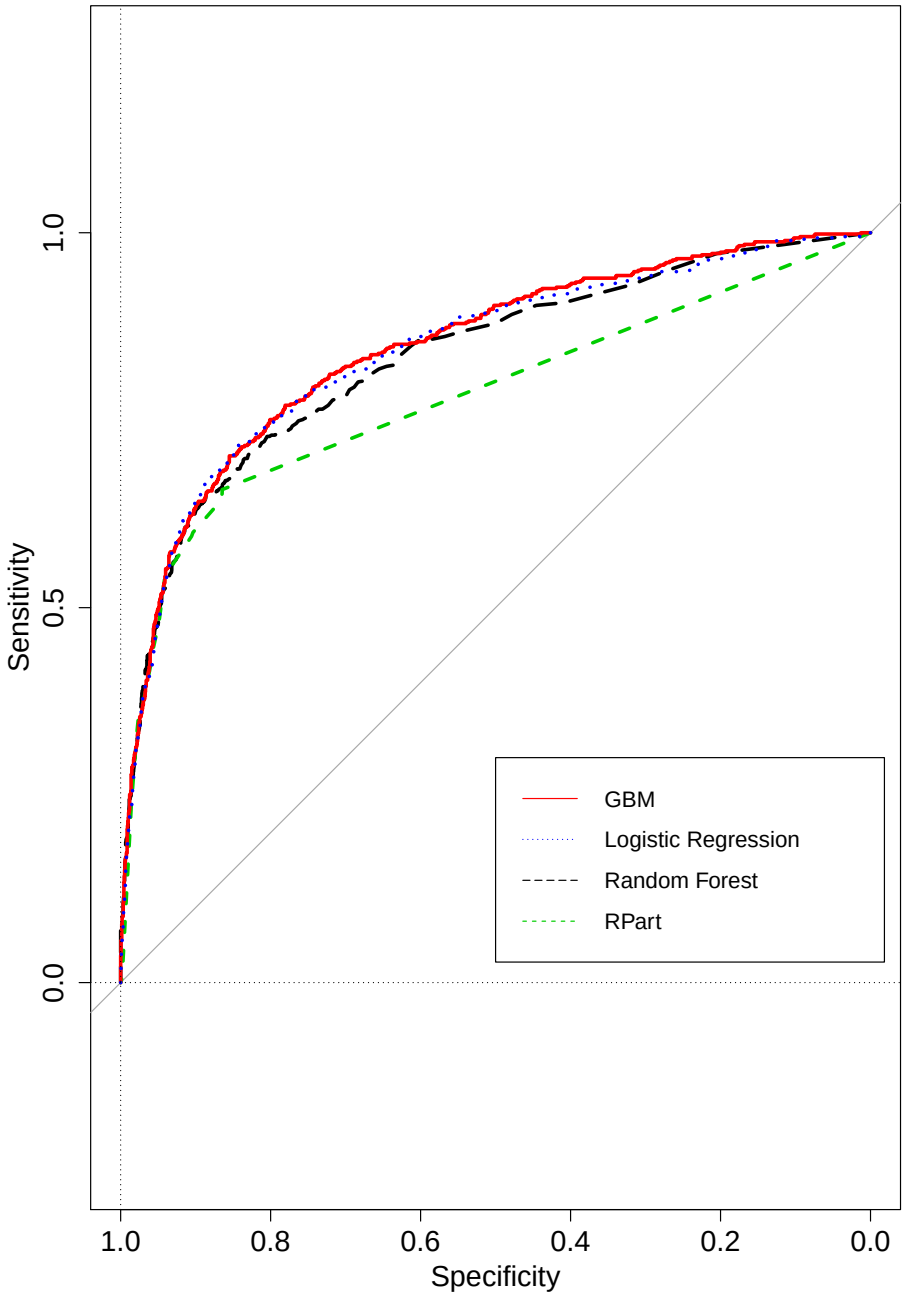

Figure 5.2: Receiver Operating Curves for the Rpart, Random Forest, Logistic Regression, and GBM models

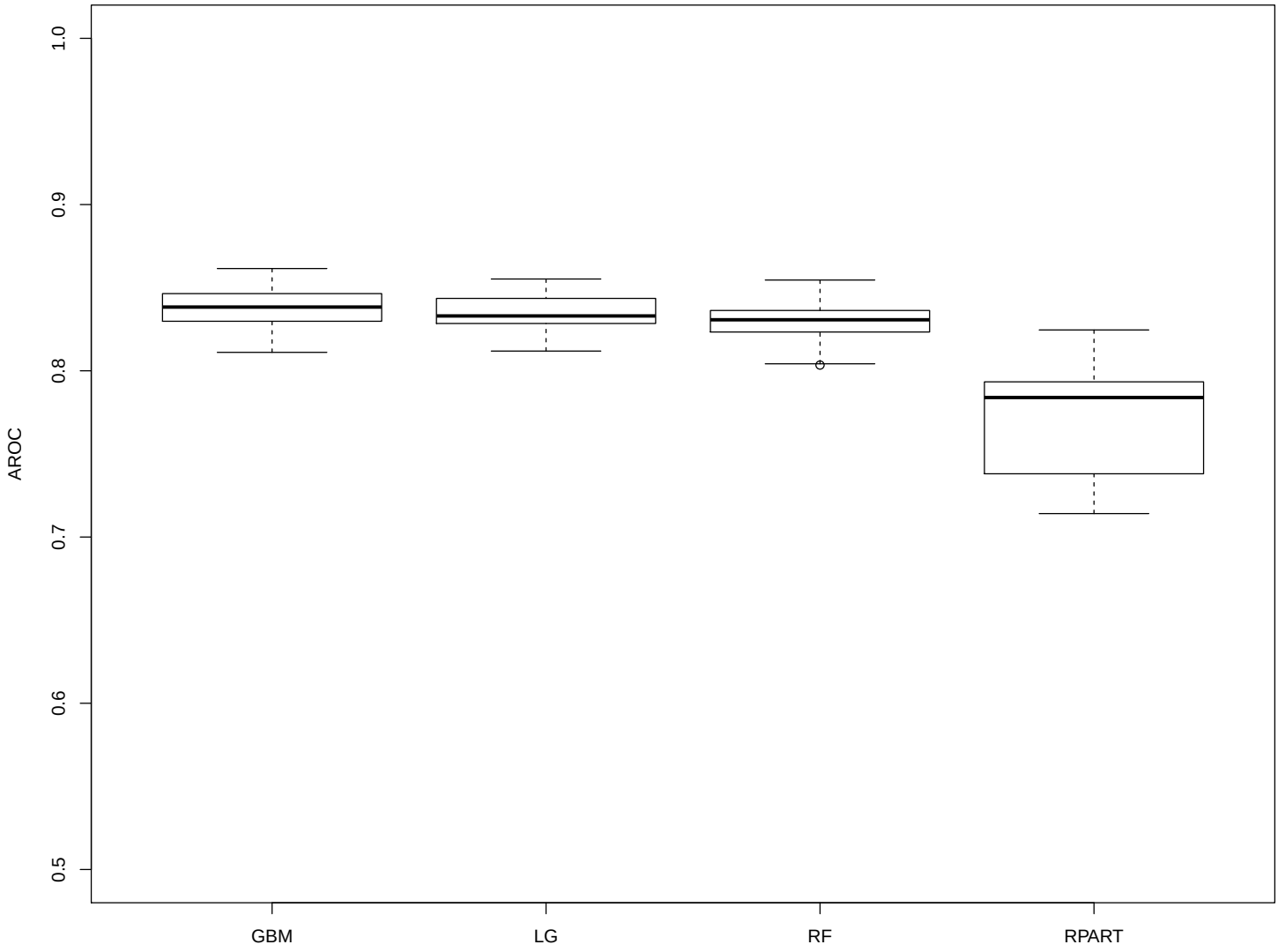

Figure 5.3: Box plot to compare the AROC values of the four models in the cross-validation results.

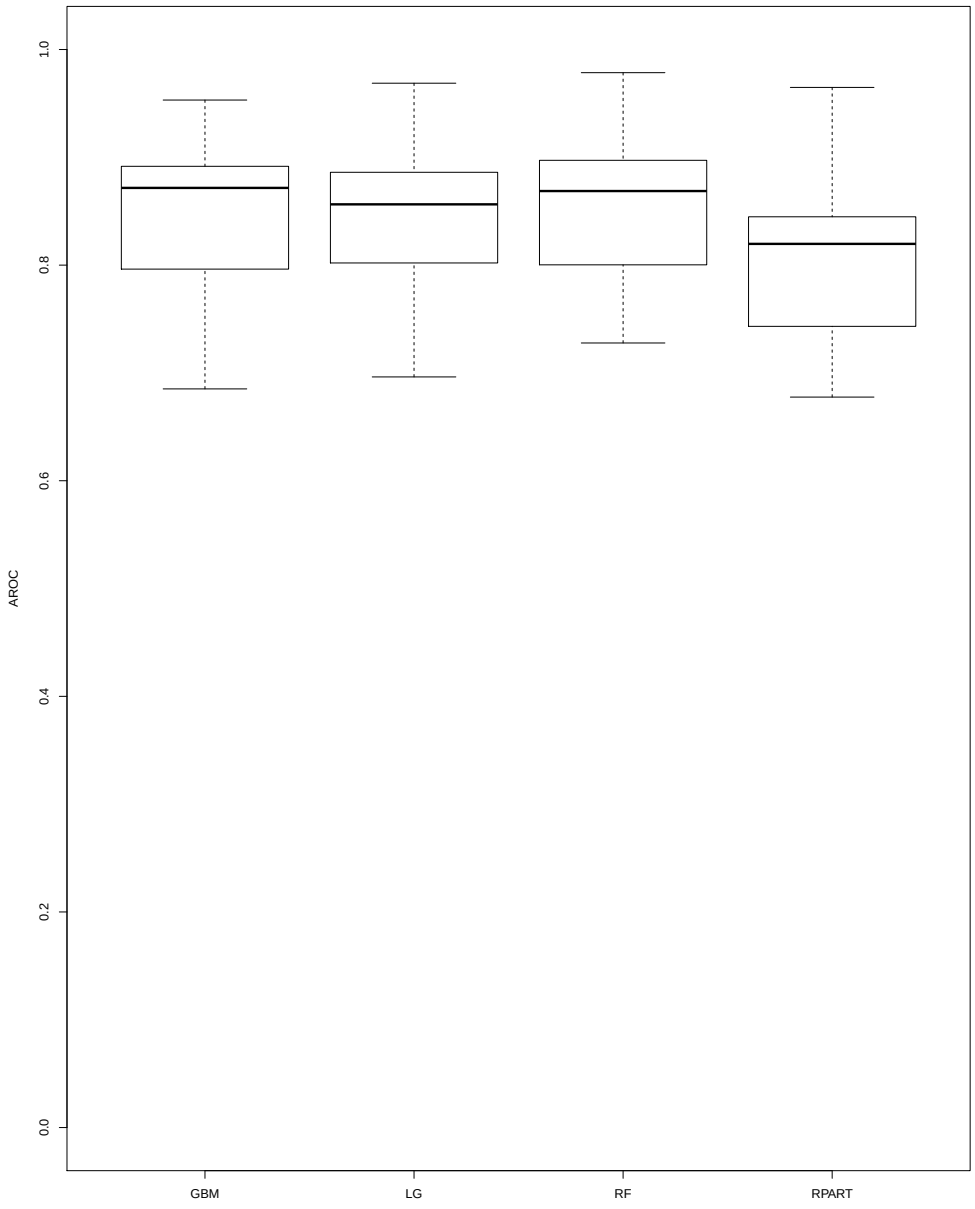

Figure 5.4: Box plot of AROC values for the Rpart, Random Forest, Logistic Regression, and GBM models applied to PIMA Indian data set

Our second project was published in the BMC Indocrine Disorder journal in 2019. (Lai et al. (2019))

6 Conclusions and Future Work

In this chapter, we summarize the conclusions and contributions of this dissertation and discuss potential future work.

6.1 Conclusions

In this dissertation, we have introduced a modified BIC for linear mixed models that can directly deal with the boundary issue of variance components. First, we focused on selecting random effects variance components and proposed a model selection criterion when the random effects are assumed to be independent (the covariance matrix of random effects is a diagonal matrix). Second, we proposed a criterion for choosing random effects variance components when the random effects are assumed to be correlated. Instead of working with a complex tangent cone to the alternative parameter space, we approximated the tangent cone using a bigger but simpler cone. This allowed us to obtain the weights of the chi-bar square dis-

tribution. Lastly, we presented a model selection criterion for choosing both fixed effects and random effects simultaneously in both cases: when random effects are assumed to be independent and when they are correlated. We have also proven the consistency of the modified BIC.

Based on the simulation studies, the modified BIC performs quite well in terms of the correction rate. The ability to select the data-generating model of the modified BIC is better when the size of random effects variance component or the size of correlation component is bigger. Compared to the regular BIC, the modified BIC gives higher correction rates, especially, when the variances of random effects are small. Based on the correction rate, the modified BIC and performs better than the regular BIC in most cases.

Furthermore, we also present predictive models using Gradient Boosting Machine and Logistic Regression techniques to predict the probability of patients having Diabetes Mellitus based on their demographic information and laboratory results from their visits to medical facilities. The ability of our models to predict patients with Diabetes is high with satisfactory sensitivity.

One limitation of the modified BIC is that when choosing the optimal model, the proposed method looks at all possible models. Since the number of possible models increases exponentially as the number of fixed effects and random effects increases,

the model selection process may be increasingly computationally intensive.

6.2 Future Work

There are some potential future directions that we can consider. The first direction is to extend the modified BIC to the case when the covariance matrix of the random error $\boldsymbol{\epsilon}$ is a general covariance matrix, $\boldsymbol{R}$. In this work we assume that the vector of random errors $\boldsymbol{\epsilon}_i$ follows a multivariate normal distribution, $N(0, \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ denotes the $n_i \times n_i$ identity matrix.

We can also investigate how the performance of the proposed BIC is affected if the model is misspecified. For example, the normal distribution of random effects does not hold.

We also look at the case if the first term in the proposed BIC can be replaced by some other model fitting measurements just as quasi-likelihood.

"Missingness" is also common in clustered data and longitudinal data. We can investigate how missingness affects the performance of the proposed BIC.

The second direction is to extend the modified BIC for model selection on generalized linear mixed models or non-linear mixed models. This future direction is feasible because the theoretical results obtained in Baey et al. (2019) paper are for non-linear mixed models as well.

And the final direction is to extend the modified BIC to the case when the number of random effects, q_n, becomes very large. That is, $q_n \to \infty$ as $n \to \infty$. How can we handle the case when the number of variances of random effects is large. Since the number of candidate model increases exponentially with the number of model parameters, it may be not possible to compare all possible models. Also, when the number of variances of random effects is large, the size of the covariance matrix will be large and this may cause a computational issue. However, we may combine the proposed BIC with some selection procedure such as shrinkage methods or fence methods as introduced in Müller et al. (2013). We can first use a fence method to reduce the number of candidate models. Then we can use the proposed BIC method to do model selection.

Bibliography

Akaike, H. (1974). A new look at the model selection identification. *In IEEE Trans. Automat. Control AC*, 19:716–723.

Azadbakhsh, M., Gao, X., and Jankowski, H. (2021). Composite likelihood ratio testing under nonstandard conditions using tangent cones. *Stat*, 10(1):e375.

Baey, C., Cournède, P.-H., and Kuhn, E. (2019). Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 135:107–122.

Balkau, B., Lange, C., and Fezeu, L., e. a. (2008). Predicting diabetes: clinical, biological, and genetic approaches: data from the epidemiological study on the insulin resistance syndrome (desir). *Diabetes Care*, 31:2056 – 2061.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., and Richter, J., e. a. (2016). mlr: Machine learning in r. journal of machine learning research. *Journal of Machine Learning Research*, 17(170):1 – 5.

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

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22(2):555–574.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25:573–578.

Czado, C. (2017). Lecture 10: Linear mixed models (linear models with random effects).

Delattre, M. and Poursat, M.-A. (2020). An iterative algorithm for joint covariate and random effect selection in mixed effects models. *The International Journal of Biostatistics*, 16.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

Drikvandi, R., Verbeke, G., Khodadadi, A., and Nia, V. P. (2012). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1):144–159.

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, 97(4):773–789.

Griffin, S., Little, P., Hales, C., Kinmonth, A., and Wareham, N. (2000). Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. *Diabetes Metab Res*, 16:164 – 171.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

Ioannis, K., Olga, T., Athanasios, S., and Nicos, M., e. a. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104 – 116.

Iyer, A., Jeyalatha, S., and Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 5(1):1 – 14.

Jayalakshmi, T. and Santhakumaran, A. (2010). A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. *International Conference on Data Storage and Data Engineering, India*, pages 159 – 163.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25):3050–3056.

Kahn, H., Cheng, Y., Thompson, T., Imperatore, G., and Gregg, E. (2009). Two risk-scoring systems for predicting incident diabetes mellitus in u.s. adults age 45 to 64 years. *Ann Intern Med*, 150:741 – 751.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorder*, 19(101).

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional aic for linear mixed-effects models. *Biometrika*, 95(3):773–778.

Lindström, J. and Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26:725 –731.

Mashayekhi, M., Prescod, F., Shah, B., Dong, L., Keshavjee, K., and Guergachi, A. (2015). Performance of the framingham diabetes risk scoring model in canadian electronic medical records. *Canadian Journal of Diabetes.*, 39(30):152 – 156.

Meng, X., Huang, Y. X., Rao, D., Zhang, Q., and Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, 29(2):93–99.

Monette, G., Fox, J., Friendly, M., Krause, H., and Zhu, F. (2019). *spida2: Collection of tools developed for the Summer Programme in Data Analysis 2000-2012*. R package version 0.2.1.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot*, 7(21).

Patel, Ashish (2018). Pima indians diabetes dataset missing value handling using imputation. `http://www.github.com/ashishpatel26/Pima-Indians-Diabetes-Dataset-Missing-Value-Imputation`.

Pauler, D. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika*, 85(1):13–27.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association*, 94(448):1242–1253.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Multivariate Anal*, 109:109–129.

Potthoff, R. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3–4):313–326.

Rao, R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2):369–374.

Raudenbush, S. and Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE Publications.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze andcompare roc curves. *BMC Bioinformatics*, 12:77–77.

Säfken, B., Rügamer, D., Kneib, T., and Greven, S. (2018). Conditional model selection in mixed-effects models with caic4. *arXiv preprint arXiv:1803.05664*.

Saville, B. R. and Herring, A. H. (2008). Testing random effects in the linear mixed model using approximate bayes factors. *Biometrics*, 65(2):369–376.

Schmid, R., Vollenweider, P., Waeber, G., and Marques-Vidal, P. (2011). Estimating the risk of developing type 2 diabetes: a comparison of several risk scores: the cohorte lausannoise study. *Diabetes Care*, 34:1863 –1868.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrika*, 26(3):505–524.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in

multivariate analysis. *International Statistical Review / Revue Internationale de Statistique*, 56(1):49–62.

Silvapulle, M. J. and Sen, P. K. (2005). *Constrained statistical inference: Order,inequality, and shape constraints*. John Wiley & Sons.

Sisodia, D. and Sisodia, D. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, 132:1578 –1585.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1179.

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2):351–370.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

Vanbrabant, L., Rosseel, Y., and Dacko, A. (2019). con_weights_boot: function for computing the chi-bar-square weights based on Monte Carlo simulation. `https://www.rdocumentation.org/packages/restriktor/versions/0.2-250/topics/con{_}weights_boot/`.

Williamson, T., Green, M., Birtwhistle, R., Khan, S., Garies, S., Wong, S., Natarajan, N., Manca, D., and Drummond, N. (2014). Validating the 8 cpcssn case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med.*, 12(4):367 –372.

Wilson, P., Meigs, J., Sullivan, L., Fox, C., and Nathan, D.M., e. a. (2007). Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study. *Arch. Intern. Med.*, 167:1068 –1074.

Yang, L. and Wu, T. (2022). Model-based clustering of high-dimensional longitudinal data via regularization. *Biometrics*, 109.

7 Appendix

7.0.1 Some definitions

Definition 1 *Definition of an approximating cone (Chernoff, 1954). Let* $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ *and* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$*. The set* $\boldsymbol{\Theta}$ *is said to be approximated by a cone* A *at* $\boldsymbol{\theta}_0$ *if* $d(\boldsymbol{y}, A) = o(||\boldsymbol{y} - \boldsymbol{\theta}_0||)$*, for all* $\boldsymbol{y} \in \boldsymbol{\Theta}$*, and* $d(\boldsymbol{x}, \boldsymbol{\Theta}) = o(||\boldsymbol{x} - \boldsymbol{\theta}||)$*, for all* $\boldsymbol{x} \in A$ *where* $d(\boldsymbol{x}, \boldsymbol{\Omega}) = \inf_{\boldsymbol{y} \in \boldsymbol{\Omega}} ||\boldsymbol{x} - \boldsymbol{y}||$*, which is the distance between point* $\boldsymbol{x}$ *and its projection onto any space* $\boldsymbol{\Omega}$*. In this case* A *is called the approximating cone of* $\boldsymbol{\Theta}$ *at* $\boldsymbol{\theta}_0$ *and* $\boldsymbol{\Theta}$ *is said to be Chernoff-regular at* $\boldsymbol{\theta}_0$*.*

Definition 2 *Definition of a tangent cone (Silvapulle and Sen, 2005). A tangent cone* $T_A(\boldsymbol{\theta}_0)$ *of a set* $\boldsymbol{\Theta}$ *at a point* $\boldsymbol{\theta}_0$ *in* $\boldsymbol{\Theta}$ *is the set of limits of sequences* $t_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$*, where* t_n *are positive real numbers,* $t_n \to 0$ *and* $\boldsymbol{\theta}_n$ *in* $\boldsymbol{\Theta}$ *converge to* $\boldsymbol{\theta}_0$*.*

Definition 3 *Definition of chi-bar square distribution Silvapulle and Sen (2005). Let* $\mathcal{C} \subset \mathbb{R}^m$ *be a closed convex cone and let* $\boldsymbol{Z} \sim N_m(\boldsymbol{0}, \boldsymbol{V})$*, where* $\boldsymbol{V}$ *is a positive definite matrix.* $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C})$ *is a random variable which has the same distribution as*

$\left[\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})\right]$. *So, we write*

$$\bar{\chi}^2(\boldsymbol{V},\mathcal{C}) = \boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})$$

where $w_i\left(m,\boldsymbol{V},\mathcal{C}\right), i=0,\ldots,m$, *are some non-negative numbers and* $\sum_{i=0}^{m} w_i\left(m,\boldsymbol{V},\mathcal{C}\right) = 1$.

Definition 4 *Results from Baey et al. (2019).*

1. *Denote by* $\boldsymbol{D}$ *the covariance matrix of the vector of random effects in model (1.1). Let* $r \in \{1,\ldots,p\}$. *We consider general test hypotheses of the following form, to test the nullity of* r *variances and of the corresponding covariances in matrix* D*:*

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \text{ against } H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{7.1}$$

where $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta} \subset R^m$. *Up to permutations of rows and columns of the covariance matrix* $\boldsymbol{D}$, *we can assume that we are testing the nullity of the last* r *variances. We write* $\boldsymbol{D}$ *in blocks as follows:* $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{12}{}^T & \boldsymbol{D}_{22} \end{bmatrix}$ *with* $\boldsymbol{D}_{11}$ *a* $(q-r)\times(q-r)$ *matrix,* $\boldsymbol{D}_{12}$ *a* $(q-r)\times r$ *matrix, and* $\boldsymbol{D}_{22}$ *a* $r\times r$ *matrix; and where* $\boldsymbol{A}^T$ *denotes the transposition of matrix* $\boldsymbol{A}$, *for any matrix* $\boldsymbol{A}$.

The spaces associated to the null and alternative hypotheses are then:

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}\in\mathbb{R}^m/\boldsymbol{\beta}\in\mathbb{R}^p; \boldsymbol{D}_{11}\in\mathbb{S}_+^{q-r}; \boldsymbol{D}_{12}=\boldsymbol{0}, \boldsymbol{D}_{22}=\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2\geq 0\}$$

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}\in\mathbb{R}^m/\boldsymbol{\beta}\in\mathbb{R}^p; D\in\mathbb{S}_+^q, \sigma_{\boldsymbol{\epsilon}}^2\geq 0\},$$

where $\mathbb{S}_+^{q-r}$ *is the set of symmetric positive semi-definite matrices of size* $(q-r)\times(q-r)$.

2. *Let us denote by* $\boldsymbol{y}$ *the joint vector of a* N*-sample* $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N)$ *and* $L_N(\boldsymbol{\theta};\boldsymbol{y})$ *the joint likelihood. We then define the likelihood ratio test statistics by:*

$$\lambda_N = -2\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} l_N(\boldsymbol{\theta};\boldsymbol{y}) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} l_N(\boldsymbol{\theta};\boldsymbol{y})\right),$$

3. *Let* $\boldsymbol{\theta}^*$ *be the true value of the parameter and* $\boldsymbol{\nu}(\boldsymbol{\theta})$ *be some positive definite matrix such that* $N^{-\frac{1}{2}}l'_N(\boldsymbol{\theta}) \xrightarrow{d} N_m(\mathbf{0},\boldsymbol{\nu}(\boldsymbol{\theta}))$ *and* $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

7.0.1.1 Baey, Cournède, and Kuhn (2019)'s assumptions

(B1). The function $L_N(\boldsymbol{\theta};\boldsymbol{y})$, the joint likelihood of a N-sample $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N)$, is injective in $\boldsymbol{\theta}$. That is, the model is identifiable.

(B2). The first three derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$ exist and are bounded by a function whose expectation exists. Denote the first derivatives of $l_N(\boldsymbol{\theta};\boldsymbol{y})$ with respected to $\boldsymbol{\theta}$ by $l'_N(\boldsymbol{\theta})$ and denote the second derivatives of $l_N(\boldsymbol{\theta};\boldsymbol{y})$ with respected to $\boldsymbol{\theta}$ by $l''_N(\boldsymbol{\theta})$.

(B3). For all $\boldsymbol{\theta}$, there exists some positive definite matrix $\boldsymbol{\nu}(\boldsymbol{\theta})$ such that $N^{-\frac{1}{2}}l'_N(\boldsymbol{\theta}) \xrightarrow{d} N(0,\boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$; and $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a continuous function with respect to $\boldsymbol{\theta}$.

(B4). Denote by $\boldsymbol{\theta}^*$ the true value of the parameters. The value $\boldsymbol{\theta}^*$ is in $\boldsymbol{\Theta}_0$ and is of the form $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{D}^*, \sigma_{\boldsymbol{\epsilon}}^{2*})^T$ with $\boldsymbol{D}^* = \left[\begin{smallmatrix} \boldsymbol{D}_1^* & 0 \\ 0 & 0 \end{smallmatrix}\right]$ where $\boldsymbol{D}_1^*$ is positive definite and $\sigma_{\boldsymbol{\epsilon}}^{2*}$ is positive.

(B5). The maximum likelihood estimates on $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}$ are consistent.

Theorem 7.1 *(modified from Theorem 2 in Baey, Cournède, and Kuhn (2019)). Assume that conditions (B1) to (B5) (from 7.0.1.1) are fulfilled. Consider the test defined in (7.1). Then:*

$$\lambda_N \underset{N\to\infty}{\longrightarrow} \bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*) \cap T(\boldsymbol{\Theta}_0, \boldsymbol{\theta}^*)^{\perp}),$$

where $T(\boldsymbol{\Theta}_0, \boldsymbol{\theta})$ is the tangent cone to $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}$, and $S^{\perp}$ is the orthogonal complement of S, for any subset S of $\mathbb{R}^m$.

Proposition 7.1 *(modified from Proposition 1 in Baey, Cournède, and Kuhn (2019)).*

(i) Assume that $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} \in \mathbb{S}_+^q, \boldsymbol{D}$ diagonal$, \sigma_{\boldsymbol{\epsilon}}^2 \in \mathbb{R}_+\}$. Then,

$$T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*) \cap T(\boldsymbol{\Theta}_0, \boldsymbol{\theta}^*)^{\perp} = \{0\}^p \times \{0\}^{q-r} \times \mathbb{R}_+^r \times \{0\}.$$

(ii) Assume that $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} \in \mathbb{S}_+^q, \boldsymbol{D}$ full$, \sigma_{\boldsymbol{\epsilon}}^2 \in \mathbb{R}_+\}$. Then,

$$T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*) \cap T(\boldsymbol{\Theta}_0, \boldsymbol{\theta}^*)^{\perp} = \{0\}^p \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}_+^r \times \{0\}.$$

Corollary 7.1 *(modified from Corollary 1 in Baey, Cournède, and Kuhn (2019)).*

(i) Assume that $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} \in \mathbb{S}_+^q, \boldsymbol{D}$ *diagonal*, $\sigma_{\boldsymbol{\epsilon}}^2 \in \mathbb{R}_+\}$. *Then the distribution of the random variable* $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*) \cap T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*)^{\perp})$ *is a mixture of* $(r+1)$ *chi-square distributions with degrees of freedom between* 0 *and* r.

(ii) Assume that $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} \in \mathbb{S}_+^q, \boldsymbol{D}$ *full*, $\sigma_{\boldsymbol{\epsilon}}^2 \in \mathbb{R}_+\}$. *Then the distribution of the random variable* $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*) \cap T(\boldsymbol{\Theta}, \boldsymbol{\theta}^*)^T)$ *is a mixture of* $(r(r+1)/2+1)$ *chi-square distributions with degrees of freedom between* $r(q-r)$ *and* $r(q-r)+r(r+1)/2$.

Corollary 7.2 *(Shapiro, 1985, 1988). Let* V *be a positive-definite matrix and* $\mathcal{C}$ *a closed convex cone of* $\mathbb{R}^m$. *The polar cone of cone* $\mathcal{C}$ *is denoted by* $\mathcal{C}^0$ *and* $\mathcal{C}^0 = \{x \in \mathbb{R}^m \mid x^T y \leq 0, \ \forall \ y \in \mathcal{C}\}$. *Some properties for the weights of the chi-bar-square distribution* $\bar{\chi}^2(V, \mathcal{C})$.

1. *For* $0 \leq i \leq m$, $w_i(m, V, C) = w_{m-i}(m, V, \mathcal{C}^0)$,

2. *If* $\mathcal{C}$ *is included in a linear space of dimension* $(m-k)$, *for* $1 \leq k \leq m$, *then the first* k *weights* $\{w_i(m, V, \mathcal{C}^0), i = 0, ..., k-1\}$ *are zero,*

3. *If* $\mathcal{C}$ *contains a linear space of dimension* l, *for* $1 \leq l \leq m$, *then the last* l *weights* $\{w_i(m, V, \mathcal{C}^0), i = m-l+1, ..., m\}$ *are zero.*