mathematics

MDPI

Article

Modified BIC Criterion for Model Selection in Linear Mixed Models

Hang Lai [1,*] and Xin Gao [2]

[1] Business Program, University of Guelph-Humber, Toronto, ON M9W 5L7, Canada
[2] Department of Mathematics & Statistics, Faculty of Science, York University, Toronto, ON M3J 1P3, Canada; xingao@yorku.ca
* Correspondence: hang.lai@guelphhumber.ca

Abstract: Linear mixed-effects models are widely used in applications to analyze clustered, hierarchical, 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 component parameters. When selecting the variance components of the random effects, the variance of the random effects must be non-negative and the parameters may lie on the boundary of the parameter space. Therefore, classical model selection methods cannot be directly used to handle this situation. In this article, 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. Through the simulation results, we found that the modified BIC performed better than the regular BIC in most cases for linear mixed models. The modified BIC was also applied to a real dataset to choose the most-appropriate model.

Keywords: linear mixed models; BIC; model selection; chi-bar-squared distribution; complex data; statistical modeling

MSC: 62J05

check for updates

Citation: Lai, H.; Gao, X. Modified BIC Criterion for Model Selection in Linear Mixed Models. *Mathematics* **2023**, *11*, 2130. https://doi.org/10.3390/math11092130

Academic Editor: Dan Vilenchik

Received: 5 April 2023
Revised: 25 April 2023
Accepted: 28 April 2023
Published: 2 May 2023

1. Introduction

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. Recently, Sheng et al. [1] compared the linear models with linear mixed-effects models and showed that estimators from the latter are more advantageous in terms of both efficiency and unbiasedness. This shows the importance of applying linear mixed-effects models in longitudinal settings.

The correlation between observations may appear when data are 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, the 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 observations from the same subjects collected over time, such as repeated blood pressure measurements over a patient's treatment period—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 parsimonious and can adequately fit the data in order to improve two important aspects: interpretability and predictability. In linear mixed models, identifying significant random effects is a challenging step in model selection, as 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 the 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, score, and Wald tests are no longer appropriate. We refer to this violation as the boundary issue. (Please see a graphical example of the boundary issue in Appendix A.3). When the boundary issue occurs, the asymptotic null distribution of the likelihood ratio test statistic does not follow a chi-squared distribution. Chernoff [2], Self and Liang [3], Stram and Lee [4], Azadbakhsh et al. [5], and Baey et al. [6] pointed out that, under some conditions on the parameter space and the likelihood functions, the asymptotic null distribution of the likelihood ratio test statistic is a mixture of chi-squared distributions. For instance, the asymptotic null distribution of the likelihood ratio 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 [4]. The distribution of a chi-bar-squared random variable depends on its mixing weights (Appendix A.2). Dykstra [7] discussed conditions on the weight distribution to ensure asymptotic normality for chi-bar-squared distributions. Shapiro [8] provided expressions to calculate the exact weights used in the mixture of chi-squared distributions for some special cases. However, in general, determining the exact weights used in the mixture of chi-squared distributions is challenging 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. [6]).

There are a number of information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), that were developed for model selection with linear mixed models by Vaida and Blanchard [9], Pauler [10], Jones [11], and Delattre and Poursat [12]. Other methods for identifying important fixed effects and random effects variance components, including shrinkage and permutation methods, were considered in Ibrahim et al. [13], Bondell et al. [14], Peng and Lu [15], and Drikvandi et al. [16].

The BIC is susceptible to the boundary issue. If we use the 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 the components of the parameter vector. Therefore, the regular BIC would overestimate the number of degrees of freedom of the linear mixed model (which we refer to as model complexity for this article) 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 the modified BIC have been proposed for model selection in linear mixed models [10,11,17]. However, to our knowledge, none of the current BICs can directly deal with the boundary issue.

The main objective of this article was to introduce a modified BIC for model selection when the true values 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-squared distribution (Baey et al. [6]). Based on this theoretical result, we took the average of the chi-bar-squared distribution and included this average in the complexity of the model. When random effects are correlated, calculating the weights of the chi-bar-squared distribution is not straightforward, as 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 the random effects covariance matrix is almost impossible. Thus, calculating the weights of the chi-bar-squared distribution for this case is not an easy task and has not been addressed in the literature. Our solution to this problem is to place a bound on cone C^* with a bigger cone. The bigger cone has a much simpler structure and allowed us to calculate the weights of the chi-bar-squared distribution. The rest of the paper is arranged as follows. In Section 2, we develop the methodology. The simulation and application are provided in Sections 3 and 4. We conclude with a brief discussion in Section 5.

2. Methodology

2.1. Model Setup and Definitions

Consider the linear mixed model introduced in Laird and Ware [18]:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

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 an $n_i \times p$ matrix of covariates for the fixed effects; $\boldsymbol{Z}_i$ is an $n_i \times q$ matrix of covariates for the random effects; $\boldsymbol{b}_i$ denotes the random effects vector of the i-th 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. $\boldsymbol{b}_{\boldsymbol{1}}, \ldots, \boldsymbol{b}_{\boldsymbol{N}}$ were 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. The vector of random errors $\boldsymbol{\epsilon}_i$ was assumed to follow a multivariate normal distribution, $N(\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ denotes the $n_i \times n_i$ identity matrix. It was assumed that $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are 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 $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}_{n_i}$.

Let $\boldsymbol{\tau}$ denote the vector of distinct variance and covariance components in matrix $\boldsymbol{D}$, and let $\boldsymbol{\eta} = (\boldsymbol{\tau}^T, \sigma_{\boldsymbol{\epsilon}}^2)^T$. The vector of parameters for this model is $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$. We assumed that the response vectors $\boldsymbol{y}_{\boldsymbol{1}}, \ldots, \boldsymbol{y}_{\boldsymbol{N}}$ from N clusters are independent random observations. Given a clustered dataset, we wish to choose a linear mixed model that fits the data well and is also a parsimonious model.

Definition 1 (Definition of an approximating cone [2])**.** *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 [19])**.** *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 and $t_n \to 0$ and $\boldsymbol{\theta}_n$ in $\boldsymbol{\Theta}$ converge to $\boldsymbol{\theta}_0$.*

Definition 3 (Definition of chi-bar-squared distribution [19])**.** *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 vari-*

able, which has the same distribution as $\Big[\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\mathbf{Z}-\boldsymbol{\theta})^T\mathbf{V}^{-1}(\mathbf{Z}-\boldsymbol{\theta})\Big]$. *Therefore, we write*

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

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

2.2. Proposed Methods

In this section, we introduce a modified BIC for linear mixed model selection. In linear mixed models, model selection includes the selection of the regression parameters $\boldsymbol{\beta}$ (fixed effects) and variance components of random effects. We first derived 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. We also considered two cases for when the covariance matrix for random effects b_i are diagonal and full matrices.

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 $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_k}^T, {\boldsymbol{\tau}_k}^T, \sigma^2_{\epsilon,k})^T$, 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 $\mathbf{D}$; $\sigma^2_{\epsilon,k}$ is the parameter for the variance of the random error vector $\boldsymbol{\epsilon_k}$. For a general covariance matrix, model M_k is uniquely defined by its non-zero parameters in $\boldsymbol{\beta}$ and non-zero variance components on the diagonal of matrix $\mathbf{D}$. If $d_{ii} = 0$, then all elements on row i and column i of this matrix are set to 0.

2.2.1. Modified BIC for Choosing Random Effects Assuming That the Random Effects Are Independent

In this section, we considered the case where the covariance matrix of random effects, D, is a diagonal matrix. Here, $\boldsymbol{\tau}$ is a vector of variances on the diagonal of matrix D.

Lemma 1. *When* $\mathbf{D}$ *is a diagonal matrix, under assumptions* (C1) − (C4) (*Appendix* A.1), *assume that we wish to test the model* M_k (*with* $\boldsymbol{\tau}_k = (d_1,\dots,d_k)$) *against model* M_1 (*with* $\boldsymbol{\tau}_1 = (d_1, 0,\dots,0)$) *and both models have the same fixed effects part, then the null limiting distribution of the likelihood ratio test is*

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

where $C^* = \{0\}^p \times \{0\} \times \mathbb{R}_+^{k-1} \times \{0\}$; $w_i\big(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\big), i = 0,\dots,k-1$, *are some non-negative numbers and* $\sum_{i=0}^{k-1} w_i\big(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*\big) = 1$; *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* $\boldsymbol{\theta}$. $\boldsymbol{\theta}^*$ *denotes the true value of the parameter* $\boldsymbol{\theta}$; $l_N(\boldsymbol{\theta}; y)$ *denotes the marginal log-likelihood function of the linear mixed model* (1).

Proof. We applied the results from Baey et al. [6] on testing the nullity of r variance components of the $q \times q$ diagonal covariance matrix, $\mathbf{D}$, using the likelihood ratio test statistic, assuming that the variances that are not being tested are strictly positive. Without loss of generality, assume that matrix $\boldsymbol{D}$ can be written as $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{22} \end{bmatrix}$, 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 $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_\epsilon)^T \in \Theta \subset \mathbb{R}^m$ with $\boldsymbol{\tau} = (d_1, \ldots, d_q)^T$. Consider the hypothesis test, $H_0 : \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$ with positive definite matrix $\boldsymbol{D}_{11}$ versus $H_1 : \boldsymbol{D}$ is positive definite.

The parameter spaces under and their corresponding tangent cones are

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

In this case, $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$. Therefore, $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^\perp = \{0\}^p \times \{0\}^{q-r} \times \mathbb{R}_+^r \times \{0\}$. C^* is contained in a linear subspace of dimension r. Thus, $w_i(m, v(\boldsymbol{\theta}^*)^{-1}, C^*) = 0$ for $i = r+1, \ldots, m$. Assume that the null hypothesis holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$,. Baey et al. [6] pointed out that the asymptotic null distribution of the log-likelihood ratio test statistic is a mixture of chi-squared distributions with the degree of freedom ranging from 0 to r, denoted by

$$\bar{\chi}^2(v(\boldsymbol{\theta}^*)^{-1}, C^*) \quad = \quad \sum_{i=0}^{r} w_i(m, v(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{3}$$

where χ_i^2 is a chi-squared distribution with i degrees of freedom and $v(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l_N'(\boldsymbol{\theta}) \xrightarrow{d} N(0, v(\boldsymbol{\theta}))$ and $N^{-1}\{-l_N''(\boldsymbol{\theta})\} \xrightarrow{a.s.} v(\boldsymbol{\theta})$.

We applied this result to our case with $m = p + k + 1$; $q = k$; and $r = k - 1$. Thus, based on (3), the null limiting distribution of the likelihood ratio test statistic is

$$\bar{\chi}^2(\boldsymbol{v}(\boldsymbol{\theta}^*)^{-1}, C^*) \quad = \quad \sum_{i=0}^{k-1} w_i(m, \boldsymbol{v}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{4}$$

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

We now take the expectation of the chi-bar-squared distribution in Equation (2) and include it in the complexity of model M_k.

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

We propose the following modified BIC:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{5}$$

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 $d_k = p + 1.5 + \sum_{i=0}^{k-1} w_i(m, \boldsymbol{v}(\boldsymbol{\theta}^*)^{-1}, C^*)i$ for $k > 1$; $d_k = p + 1.5$ for $k = 1$; $d_k = p + 1$ for $k = 0$. The first term, $-2l(\hat{\theta}_k; y)$, measures the goodness-of-fit for model M_k, and the second term, $d_k \log(n)$, is the penalty for the 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 the σ_ϵ parameter and 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-squared distribution in Equation (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. For example, M_3 is a model with three independent random effects. $\boldsymbol{D} = \text{diag}(\sigma_0^2, \sigma_1^2, \sigma_2^2)$, and $\boldsymbol{\tau} = (\sigma_0^2, \sigma_1^2, \sigma_2^2)$. We want to test $H_0 : \sigma_0^2 > 0; \sigma_1^2 = 0, \sigma_2^2 = 0$, vs. $H_1 : \sigma_0^2 > 0; \sigma_1^2 > 0, \sigma_2^2 > 0$. In this example, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_\epsilon^2)^T; m = p + 3 + 1, k = 3, r = 2$. Therefore, the asymptotic null distribution of the log-likelihood ratio test statistic is

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

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

Cone C^* can be written as $C^* = \{\boldsymbol{\theta} \in \mathbb{R}^m / \boldsymbol{R}\boldsymbol{\theta} \geq 0\}$, where $\boldsymbol{R} = \left(\boldsymbol{0}_{p+1} | \boldsymbol{I}_2 | \boldsymbol{0}\right)$ is a $2 \times m$ matrix and $\boldsymbol{I}_2$ is an identity matrix of order two. The chi-bar-squared weights are $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = w_i(r, \boldsymbol{R}\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}^T, \mathbb{R}_+^2)$ under Proposition 3.6.1 of [19]. The matrix, $\boldsymbol{\nu}(\boldsymbol{\theta}^*)$, is approximated by $\boldsymbol{\Gamma} = N^{-1}\{I_N(\hat{\boldsymbol{\theta}}_k)\}$, where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator of $\boldsymbol{\theta}_k$ in model M_k and $\boldsymbol{I}_N(\boldsymbol{\theta})$ is the Fisher information matrix. The chi-bar-squared weights, $w_i(r, \boldsymbol{R}\boldsymbol{\Gamma}^{-1}\boldsymbol{R}^T, \mathbb{R}_+^2)$, can be calculated using function "con-weights-boot" in the R package "restriktor" of Vanbrabant et al. [20]. In this example, we assumed that $\boldsymbol{R}\boldsymbol{\Gamma}^{-1}\boldsymbol{R}^T = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$. Then, we obtained the weights $w_0 = 0.334$, $w_1 = 0.503$, and $w_2 = 0.163$. Thus,

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) \quad = \quad 0.334\chi_0^2 + 0.503\chi_1^2 + 0.163\chi_2^2,$$

We note that, in theory, w_1 must be 0.5. However, in our simulation, $w_1 = 0.503$. The expectation of this chi-bar-squared distribution is $0.334(0) + 0.503(1) + 0.163(2) = 0.829$. The complexity of this model is $d_3 = p + 1.5 + 0.829$.

Theorem 1. *Assume that Assumptions (C1)–(C4) in Appendix A.1 are satisfied, then*

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

Proof. We used $l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})$ 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

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

We also 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 $\hat{\boldsymbol{\theta}}_k \overset{p}{\to} \boldsymbol{\theta}_{k,0}$ and $\hat{\boldsymbol{\theta}}_T \overset{p}{\to} \boldsymbol{\theta}_{T,0}$ (as shown in the proof of Theorem 2 in Baey et al. [6]) and function $l(\boldsymbol{\theta};\boldsymbol{y})$ is continuous with respect to $\boldsymbol{\theta}$. Furthermore, 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

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

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 for all $M_k \in M^-$,

$$\lim_{n\to\infty} P(BIC^*(M_T) < BIC^*(M_k)) = 1$$

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 $\boldsymbol{\theta}_T = (\boldsymbol{\beta}_T{}^T, \boldsymbol{\psi}_T{}^T, \underline{0}, \sigma^2_{\epsilon,T})^T$ and $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k{}^T, \boldsymbol{\psi}_{k,1}{}^T, \boldsymbol{\psi}_{k,2}{}^T, \sigma^2_{\epsilon,k})^T$, 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

$$BIC^*(M_k) - BIC^*(M_T) = -2\big(l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})\big) + (d_k - d_T)\log(n). \quad (6)$$

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:

$$\begin{aligned} H_0 &: \boldsymbol{\psi}_{k,1} \geq \mathbf{0}, \boldsymbol{\psi}_{k,2} = \mathbf{0} \\ H_1 &: \boldsymbol{\psi}_{k,1} \geq \mathbf{0}, \boldsymbol{\psi}_{k,2} > \mathbf{0}. \end{aligned}$$

According to Baey et al. [6], under H_0, the asymptotic distribution of $-2(l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}))$ is

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

Therefore, $-2(l(\hat{\boldsymbol{\theta}}_k;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T;\boldsymbol{y})) = O_p(1)$, according to Theorem 2.4 of [21]. We also have that

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

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 follows a mixture of the chi-squared 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

$$\lim_{n \to \infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$$

for $M_k \in M^+$. This completes the proof of Theorem 1. □

Given a set of candidate models, we calculated the proposed BIC value for each model. Then, the selected model is the one that minimizes the proposed BIC.

2.2.2. Modified BIC for Choosing Random Effects Assuming That the Random Effects Are Correlated

In this section, we introduce a modified BIC for selecting linear mixed models with correlated random effects. We still focused on only selecting random effects. In the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \sigma_\epsilon^2)^T$, $\boldsymbol{\tau}$ is the parameter of interest; $\boldsymbol{\beta}$ and σ_ϵ^2 are considered as nuisance parameters. We now considered that the linear mixed model (1) with the covariance matrix for random effects $\boldsymbol{b}_i$ is a full matrix. Therefore, vector $\boldsymbol{\tau}$ contains all distinct variances and covariances of matrix $\boldsymbol{D}$.

Lemma 2. *When $\boldsymbol{D}$ is a full matrix and under Assumptions (C1)–(C4) in Appendix A.1, assume that we test the model M_k against model M_1, where M_1 contains only one random effect, which is a random intercept, M_k contains k random effects including a random intercept, and both models have the same fixed effects part, then the null limiting distribution of the likelihood ratio test is*

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) \quad = \quad \sum_{i=k-1}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{7}$$

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

Proof. When $\boldsymbol{D}$ is a full matrix, the number of distinct variances and covariances is $q(q+1)/2$. We also applied the results from Baey et al. [6] on testing the nullity of r variance components of the $q \times q$ covariance matrix, $\boldsymbol{D}$, when this matrix 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 $r \times r$. Consider the hypothesis test: $H_0 : \boldsymbol{D}_{11} > \boldsymbol{0}, \boldsymbol{D}_{12} = \boldsymbol{0}, \boldsymbol{D}_{22} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{D}$ is a positive definite matrix.

The parameter space under the null hypothesis is

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

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 holds and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, then applying the results of Baey et al. [6], 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 $\mathbb{S}^{(q-r)\times(q-r)}$ is the set of symmetric matrices of size $(q-r)\times(q-r)$. Furthermore,

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

According to the results of Baey et al. [6], 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 asymptotic null distribution of the likelihood ratio test statistic for the above hypothesis test is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$, where $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^\perp = \{0\}^p \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}^r_+ \times \{0\}$.

When $\boldsymbol{D}$ is a full matrix, under the null hypothesis, Baey et al. [6] pointed out that the asymptotic null 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 the degree of freedom ranging from $r(q-r)$ to $r(q-r) + r(r+1)/2$.

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi^2_i, \tag{8}$$

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 $\sum_{i=r(q-r)}^{r(q-r)+r(r+1)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = 1$; χ^2_i is a chi-squared distribution with i degrees of freedom; $\boldsymbol{\nu}(\boldsymbol{\theta})$ is a positive definite matrix such that $N^{-\frac{1}{2}} l'_N(\boldsymbol{\theta}) \xrightarrow{d} N_m(\boldsymbol{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''_N(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

Assume that model M_k has parameter vector $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k{}^T, \boldsymbol{\tau}_k{}^T, \sigma^2_{\epsilon,k})^T$, where $\boldsymbol{\beta}_k$ represents the parameter vector of the fixed effects; $\boldsymbol{\tau}_k$ contains distinct variances and covariances of the random effect covariance matrix $\boldsymbol{D}_k$, and $\sigma^2_{\epsilon,k}$ is the variance of the random error term $\boldsymbol{\epsilon_k}$. Let p be the number of parameters of $\boldsymbol{\beta}_k$ and q_k be the number of parameters of $\boldsymbol{\tau}_k$. Assume that we tested the model M_k against model M_1, where M_1 contains only one random effect, which is a random intercept, and M_k contains k random effects including a 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, $r(q-r) = k-1$ and $r(q-r) + r(r+1)/2 = (k-1)(k+2)/2$. Therefore, based on (8), the asymptotic null distribution of the log-likelihood ratio test statistic is

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) = \sum_{i=k-1}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi^2_i, \tag{9}$$

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

We note that it is too complex to define $\mathbb{S}^{k-1}_+$ using equality and inequality constraints on the variance and covariance components of matrix $\boldsymbol{D}_k$. Since $\mathbb{S}^{k-1}_+ \subset \mathbb{R}^{(k-1)(k-2)/2} \times$

$\mathbb{R}_{+}^{k-1}$, in our work, we approximated C^* by $C = \{0\}^p \times \{0\} \times \mathbb{R}^{k-1} \times \mathbb{R}^{(k-1)(k-2)/2} \times \mathbb{R}_{+}^{k-1} \times \{0\}$. Thus, $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ is approximated by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) \quad = \sum_{i=k(k-1)/2}^{(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \tag{10}$$

where $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C), i = k(k-1)/2, \ldots, (k-1)(k+2)/2$, 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-squared 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 $C = \{0\}^p \times \{0\} \times \mathbb{R}^{k-1} \times \mathbb{R}^{(k-1)(k-2)/2} \times \mathbb{R}_{+}^{k-1} \times \{0\}$ contains a linear space of dimension $(k-1) + (k-1)(k-2)/2$ and is included in a linear space of dimension $(k-1) + (k-1)(k-2))/2 + (k-1)$. Therefore, the weights $w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)$ are zero for $i = 0, \ldots, (k-1) + (k-1)(k-2)/2 - 1$ and for $i = (k-1) + (k-1)(k-2))/2 + (k-1) + 1, \ldots, m$ [8]. From (10), 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:

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{11}$$

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 $k = 0$.

2.2.3. Modified BIC for Selecting Both Fixed Effects and Random Effects in Linear Mixed Models

In this section, we propose a modified BIC to select both fixed effects and random effects for linear mixed models. We also divided 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.

Scenario 1: modified BIC for selecting both fixed effects and random effects when random effects are independent.

In the model selection, we assumed that the smallest model (called model M_1) contains only the intercept term for fixed effects and a 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 variance components is $q_k = k$.

Lemma 3. *When $\boldsymbol{D}$ is a diagonal matrix, under Assumptions (C1)–(C4) in Appendix A.1, assume that we tested model M_k against model M_1, then the asymptotic null distribution of the log-likelihood test statistic is*

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) \quad = \sum_{i=p_k}^{p_k+k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{12}$$

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

Proof. When $\boldsymbol{D}$ is a diagonal matrix, $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_{q-r}, d_{q-r+1}, \ldots, d_q)$. The fixed effects parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$. Without loss of generality, assume that we wanted to test the nullity of the s components of $\boldsymbol{\beta}$, which are $\beta_1, \ldots, \beta_s$, and the nullity of the last r variance 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 $H_1 : \beta_1 \neq 0, \ldots, \beta_s \neq 0; d_{q-r+1} > 0, \ldots, d_q > 0$, assuming that the variances that are not tested $(d_1, \ldots, d_{q-r})$ are positive. 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 spaces under the null and alternative hypotheses and their tangent cones at $\boldsymbol{\theta}^*$ are

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

Since $T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)$ is also a linear subspace in $T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, Baey et al. [6] pointed out that the asymptotic null distribution of $\bar{\chi}^2(\nu(\boldsymbol{\theta}^*)^{-1}, C^*)$ is a mixture of chi-squared distributions with the degree of freedom ranging from s to $s + r$.

$$\bar{\chi}^2((\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)) \;=\; \sum_{i=s}^{s+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{13}$$

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

When we test model M_k against model M_1, we are testing the nullity of the $s = p_k$ regression coefficients and $r = k - 1$ random effects variance components. Therefore, based on Equation (13), the asymptotic null distribution of the log-likelihood test statistic is

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*) \;=\; \sum_{i=p_k}^{p_k+k-1} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)\chi_i^2, \tag{14}$$

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

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

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \tag{15}$$

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 $k = 1$; $d_k = p_k + 2$ for $k = 0$. Here, in the formula $d_k = 2.5 + u_k$ for $k > 1$, we added 2.5 to u_k to account for the degrees of freedom of a fixed effect intercept (1 degree of freedom), a random intercept (0.5 degree of freedom), and the variance component of the error term, $\boldsymbol{\epsilon}$ (1 degree of freedom).

Scenario 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) 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, $H_0 : \beta_1 = 0, \ldots, \beta_s = 0, \boldsymbol{D}_{11} > \boldsymbol{0}, \boldsymbol{D}_{12} = \boldsymbol{0}, \boldsymbol{D}_{22} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{D} > \boldsymbol{0}$. 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 spaces 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 $(q-r) \times (q-r)$. Furthermore, the parameter space under the alternative hypothesis is

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

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 asymptotic null 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 Lemma 2, 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 approximated C^* by $C = \mathbb{R}^s \times \{0\}^{p-s} \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{R}^{r(r-1)/2} \times \mathbb{R}_+^r \times \{0\}$. $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$ is approximated by

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=s+r(q-r)+r(r-1)/2}^{s+r(q-r)+r(r-1)/2+r} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2. \quad (16)$$

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-squared distribution with i degrees of freedom; $\boldsymbol{\nu}(\boldsymbol{\theta})$ 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 assumed that model M_1 contains only the intercept term for fixed effects and a 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 tested model M_k against model M_1, applying (16) 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

$$\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C) = \sum_{i=p_k+k(k-1)/2}^{p_k+(k-1)(k+2)/2} w_i(m, \boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C)\chi_i^2, \quad (17)$$

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

Furthermore, 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

$$BIC^*(M_k) = -2l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) + d_k \log(n), \quad (18)$$

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 $k = 0$.

Theorem 2. *Assume that Assumptions $(C1) - (C4)$ Appendix A.1 are satisfied and $BIC^*(M_k)$ is defined as in (18), then*

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

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 $s \geq 0$, $r \geq 0$, 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 $\boldsymbol{\theta}_T = (\underline{\boldsymbol{0}_\beta}, {\boldsymbol{\beta}_T}^T, {\boldsymbol{\psi}_T}^T, \underline{\boldsymbol{0}}, \sigma^2_{\epsilon,T})^T$ and $\boldsymbol{\theta}_k = ({\boldsymbol{\beta}_{k,1}}^T, {\boldsymbol{\beta}_{k,2}}^T, {\boldsymbol{\psi}_{k,1}}^T, {\boldsymbol{\psi}_{k,2}}^T, \sigma^2_{\epsilon,k})^T$, where $\underline{\boldsymbol{0}_\beta}$ has the same dimension as $\boldsymbol{\beta}_{k,1}$; $\boldsymbol{\beta}_T$ has the same dimension as $\boldsymbol{\beta}_{k,2}$; $\boldsymbol{\psi}_T$ has the same dimension as $\boldsymbol{\psi}_{k,1}$; $\underline{\boldsymbol{0}}$ has the same dimension as $\boldsymbol{\psi}_{k,2}$. All elements of $\underline{\boldsymbol{0}_\beta}$ and $\underline{\boldsymbol{0}}$ are 0. We have that

$$BIC^*(M_k) - BIC^*(M_T) = -2\big(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y})\big) + (d_k - d_T)\log(n). \tag{19}$$

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:

$$\begin{aligned} &H_0 : {\boldsymbol{\beta}_{k,1}}^T = \boldsymbol{0}; \boldsymbol{D}_{11} > \boldsymbol{0}; \boldsymbol{D}_{12} = \boldsymbol{0}, \boldsymbol{D}_{22} = \boldsymbol{0}, \\ &H_1 : \boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{D} > \boldsymbol{0}. \end{aligned}$$

As in Lemma 2, under H_0, the asymptotic distribution of $-2(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y}))$ is $\bar{\chi}^2(\boldsymbol{\nu}(\boldsymbol{\theta}^*)^{-1}, C^*)$, where $C^* = T_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \cap T_{\boldsymbol{\Theta}_0}(\boldsymbol{\theta}^*)^{\perp} = \mathbb{R}^s \times \{0\}^{p-s} \times \{0\}^{(q-r)(q-r+1)/2} \times \mathbb{R}^{r(q-r)} \times \mathbb{S}^r_+ \times \{0\}$ with $s = p_k - p_T$, $r = q_k - q_T$, $p = p_k + 1$, and $q = q_k$; $\boldsymbol{\nu}(\boldsymbol{\theta})$ is some positive definite matrix such that $N^{-\frac{1}{2}} l'(\boldsymbol{\theta}) \xrightarrow{d} N_m(\boldsymbol{0}, \boldsymbol{\nu}(\boldsymbol{\theta}))$ and $N^{-1}\{-l''(\boldsymbol{\theta})\} \xrightarrow{a.s.} \boldsymbol{\nu}(\boldsymbol{\theta})$.

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

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

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 a random intercept for random effects. Therefore,

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

On the other hand, $-2\big(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{y})\big)$ asymptotically follows a mixture of the chi-squared distributions. Therefore, $E\big[-2\big(l(\hat{\boldsymbol{\theta}}_T; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y})\big)\big]$ must be positive and, therefore, $d_k - d_T > 0$. Thus, $BIC^*(M_k) - BIC^*(M_T) \to \infty$ as $n \to \infty$ and $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*\ (M_T) > 0) = 1$ for $M_k \in M^+$.

Case 2: For any under-fitting model, $M_k \in M^-$, we want to prove that $\lim_{n\to\infty} P(BIC^*(M_k) - BIC^*(M_T) > 0) = 1$. Using similar arguments as in the proof for Case 1 in Theorem 1, we obtain this result. □

3. Simulation

In this section, we evaluated the performance of the proposed BIC*. We compared the performance of the proposed BIC* to the regular BIC. For each candidate model, we computed the BIC* and regular BIC; then for each method, we chose the model with the minimum value of the BIC* and the regular BIC, respectively. All models were run using function "lmer" in the R package lme4 [22]. The chi-bar-squared weights were calculated using function "con-weights-boot" in the R package "restriktor" [20]. Following the methods used in Gao and Song [23] and Chen and Chen [24], the criteria we used to evaluate and compare the proposed BIC to the regular BIC were (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. Then, we took the average of the PSR 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. Then, we took the average of the FDR 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 each selection criterion, we had 1001 models obtained from 1001 simulations. We then calculated the means and standard deviations of the positive selection rate and false discovery rate and the correction rate for each criterion.

3.1. Simulation Setup

Our data were generated from the linear mixed model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$. For all simulation, ϵ was generated from a multivariate normal distribution, $N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$ with $\sigma_\epsilon^2 = 1$.

3.1.1. Setup A: Choose Random Effects Assuming That the Random Effects Are Independent

Scenario 1: With total number of observations $n = 500$ and number of clusters $N = 100$, $\boldsymbol{X}$ is an $n \times p$ matrix with $p = 2$; the first column of $\boldsymbol{X}$ includes all ones. The second column is $\boldsymbol{X}_1$, which was 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 generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, were generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ a 4×4 diagonal matrix and $\boldsymbol{D} = \text{diag}(\sigma_0^2, \ldots, \sigma_3^2)$. The random intercept, b_{i0}, had a standard deviation of $\sigma_0 = 5$. Random effects components, b_{i1}, b_{i2}, and b_{i3}, had standard deviations σ_1, σ_2, and σ_3, respectively. To measure the ability to detect the significance of the variance component parameters of the proposed BIC^*, we considered different sizes of σ_1^2, σ_2^2, and σ_3^2. 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 1 incrementing by 0.1; σ_3 is a sequence of values from 0 to 2 incrementing by 0.2.

Scenario 2: With the total number of observations $n = 500$ and number of clusters $N = 100$, $\boldsymbol{X}$ is an $n \times p$ matrix with $n = 500$; $p = 3$; the first column of $\boldsymbol{X}$ includes all ones. The last two columns of matrix $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ were 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 $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, and $\boldsymbol{z}_2$, which are the same as three columns of matrix $\boldsymbol{X}$ and three more columns $\boldsymbol{z}_3$, $\boldsymbol{z}_4$, and $\boldsymbol{z}_5$, which were generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, were generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ 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 the variance component parameters of the BIC^*, we also considered different sizes of σ_1^2, σ_2^2, and σ_3^2 as in Scenario 1 with $\sigma_4^2 = 0$ and $\sigma_5^2 = 0$. We then repeated this setup with $n = 1000$ ($N = 200$) and $n = 250$ ($N = 50$).

Scenario 3: The setup was similar to the one in Scenario 2. However, matrix $\boldsymbol{Z}$ contains the first three columns z_0, z_1, and z_2, which are the same as the three columns of matrix $\boldsymbol{X}$, and eight more columns $z_3, \ldots, z_{10}$, which were generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, were generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ 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. We also repeated this simulation setup with $n = 1000$ ($N = 200$) and $n = 250$ ($N = 50$).

3.1.2. Setup B: Choose Random Effects Assuming That the Random Effects Are Correlated

In this set up, the total number of observations is $n = 1000$ and the number of clusters is $N = 100$. Matrix $\boldsymbol{X}$ and the vector of fixed effects, $\boldsymbol{\beta}$, were generated the same as in Setup A Scenario 2. Matrix $\boldsymbol{Z}$ contains the first three columns z_0, z_1, and z_2, which are the same as three columns of matrix $\boldsymbol{X}$, and three more columns z_3, z_4, and z_5 were generated from the standard normal distributions. Random effects, $\boldsymbol{b}_i$, were generated from multivariate normal distribution $N_q(\boldsymbol{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ a 6×6 matrix. The correlation matrix between the random effects components, b_{i0}, b_{i1}, b_{i2}, and b_{i3}, 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 component parameters of the proposed BIC^*, we created different cases for different sizes of σ_0^2, σ_1^2, σ_2^2, and σ_3^2, as shown below. σ_4^2, σ_5^2, and the covariances of random effects b_{i4} and b_{i5} corresponding to z_4 and z_5 are all 0.

Case 1: The standard deviations of the random effects were $\sigma_0 = 5$, $\sigma_1 = 1.0$, $\sigma_2 = 0.8$, $\sigma_3 = 0.4$, $\sigma_4 = 0$, and $\sigma_5 = 0$.

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

Case 3: The standard deviations of the random effects were $\sigma_0 = 2$, $\sigma_1 = 0.5$, $\sigma_2 = 0.4$, $\sigma_3 = 0.2$, $\sigma_4 = 0$, and $\sigma_5 = 0$.

Case 4: In this case, we kept the standard deviations of the random effects the same as the ones in Case 2. However, we increased 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

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

3.1.3. Setup C: Choose Both Fixed Effects and Random Effects Assuming That the Random Effects Are Correlated

With the total number of observations $n = 1000$ and number of clusters $N = 100$, $\boldsymbol{X}$ is an $n \times p$ matrix with $p = 6$; the first column of $\boldsymbol{X}$ includes all ones. The last five columns, $\boldsymbol{X}_1$ to $\boldsymbol{X}_5$, weer 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 z_0, z_1, and z_2, which are the same as three columns of matrix $\boldsymbol{X}$, and three more columns z_3, z_4, and z_5 were generated from the standard normal distributions. The correlation matrix between the random effects components, b_{i0}, b_{i1}, b_{i2}, and b_{i3}, 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 the fixed effects and variance component parameters of the proposed BIC^*, we explored two different cases for different sizes of σ_0^2, σ_1^2, σ_2^2, and σ_3^2, as shown below. The σ_4^2, σ_5^2, and covariances corresponding to the random effects of $\boldsymbol{z}_4$ and $\boldsymbol{z}_5$ are all 0.

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

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

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

3.2. Simulation Procedure

3.2.1. For Setup A

In all scenarios, for each set of values of σ_1^2, σ_2^2, and σ_3^2, $B = 1001$ simulations were run. In each simulation, all possible candidate models were run. All these models had the same fixed effect covariates (including X_1 and the intercept); meanwhile, the covariates for random effects part varied in the power set of $\{1, 2, 3\}$. The proposed BIC^* and regular BIC were calculated for each model. Then, one model with the minimum proposed BIC was selected and one model with the minimum regular BIC. Now, for each selection criterion, we had 1001 models obtained from 1001 simulations. We calculated the correction rate (CR) for each criterion.

In Scenario 2, for each set of values of $\sigma_1^2, \ldots, \sigma_5^2$, $B = 1001$ simulations were run. In each simulation, all possible candidate models were run. All these models had the same fixed effect covariates (including $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and the intercept); meanwhile, the covariates for random effects varied in the power set of $\{1, \ldots, 5\}$. The proposed BIC^* and regular BIC were calculated for each model. Then, one model with the minimum proposed BIC was selected, and one model with the minimum regular BIC was selected. We calculated the means and standard deviations of the positive selection rate and false discovery rate. We also calculated 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 were run. In each simulation, all possible candidate models were run. All these models had the same fixed effect covariates; meanwhile, the covariates for the random effects varied in the power set of $\{1, \ldots, 10\}$. We calculated the means and standard deviations of the positive selection rate and false discovery rate and calculated the correction rate for each criterion. All simulations were performed by using R Version 4.0.2 [25].

3.2.2. For Setup B

In each case presented above, $B = 1001$ simulations were run. In each simulation, all possible candidate models were run. All these models had the same fixed effect covariates (including the intercept, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$); meanwhile, the covariates for the random effects part varied in the power set of $\{1, 2, 3, 4, 5\}$ and also included a random intercept. The proposed BIC^*, regular BIC, and cAIC were calculated for each model. Greven and Kneib [26] developed an analytic version of the corrected cAIC, and their method was implemented in the cAIC4 package in R [27]. Then, one model with the minimum proposed BIC was selected; one model with the minimum regular BIC was selected; one model with the minimum cAIC was selected. We calculated the means and standard deviations of the positive selection rate and false discovery rate and the correction rate for each criterion.

3.2.3. For Setup C

For each case above, we ran $B = 1001$ simulations. In each simulation, all possible candidate models were run. All models contained the intercept term for the fixed effect and a random intercept for the random effects. The covariates for the fixed effects part varied 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 varied in the power set of $\{1,2,3,4,5\}$ for z_1 to z_5. We also included the models that included only the intercept term for the fixed effect with varying random effects and the models that included a random intercept only with varying fixed effects. The proposed BIC^* and regular BIC were calculated for each model. Then, the model with the minimum proposed BIC was selected, and the model with the minimum regular BIC was selected.

3.3. Simulation Results

Scenario 1: Table 1 summarizes the results of Scenario 1. We observed that the correction rate for the proposed BIC^* was greater than that of the regular BIC. Furthermore, the correction rates of the two methods were higher when the values of σ_1^2, σ_2^2, and σ_3^2 were bigger.

Table 1. Comparison of the proposed BIC and regular BIC methods in terms of correction rate for the simulation in 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	0.10	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.

Scenario 2: Table 2 summarizes the results of Scenario 2. The simulation results suggested that the values of the positive selection rate (PSR) for the proposed BIC^* were higher than the regular BIC when the values of the variance components were close to 0. That is, the ability to choose the significant variance components was higher for the proposed BIC^* than the regular BIC. Almost all of the false discovery rate (FDR) values were within 5 percent in all cases. We also observed that the proposed BIC approach had a higher FDR and corresponding SD as compared to the regular BIC approach. For some very low values of the sigma values, the FDR values of the proposed BIC were greater than 5 percent. The possible reason behind this is because the calculation of the penalty term of the regular BIC uses an exact chi-squared distribution, meanwhile the penalty term of proposed BIC uses the approximated weights of the chi-bar-square distribution.

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

Figure 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, and σ_3 when $n = 500$ and $(N = 100)$ in Scenario 2.

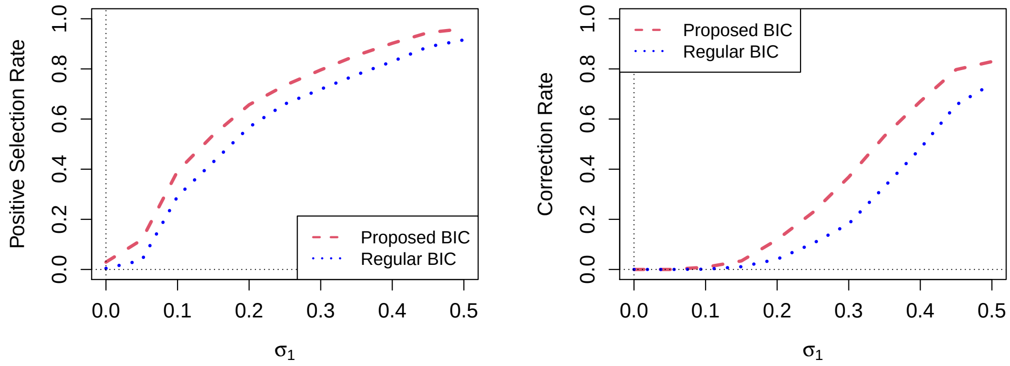

Figure 1. Comparison of the proposed BIC and regular BIC methods in terms of the positive selection rate and correction rate for different values of σ_1, σ_2, and σ_3, $n = 500$ $(N = 100)$. In this simulation setup, for each value of σ_1 on the horizontal axis, the value of σ_2 is $2 * \sigma_1$ and the value of σ_3 is $4 * \sigma_1$; $\sigma_4 = 0$ and $\sigma_5 = 0$.

Table 2. 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 $N = 100$.

			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.

Figure 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, and 200, respectively. Given the same set of values of $\sigma_1^2, \ldots, \sigma_5^2$, we observed that the positive sensitivity rate increased as the number of clusters N increased.

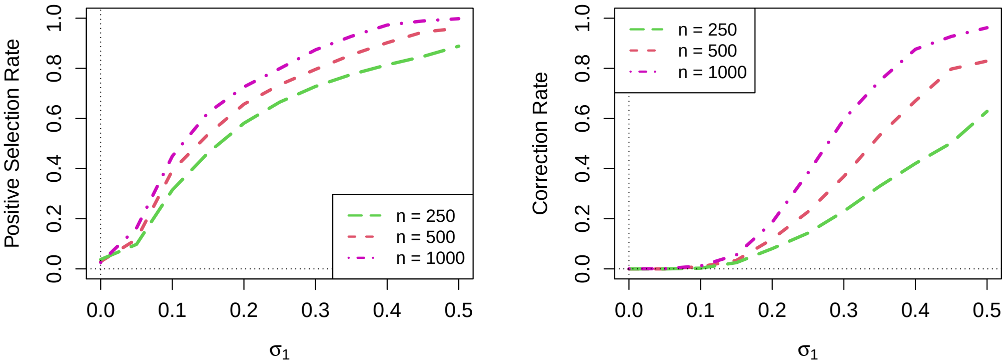

Figure 2. Comparison of the positive selection rate and correction rate for $n = 250$ $(N = 50)$, $n = 500$ $(N = 100)$, and $n = 1000$ $(N = 200)$. For each value of σ_1 on the horizontal axis, the value of σ_2 is $2 * \sigma_1$ and the value of σ_3 is $4 * \sigma_1$; $\sigma_4 = 0$ and $\sigma_5 = 0$.

We also ran 104 simulations with three more competing methods: "cAIC", "BIC_J", and "Splmm", using the same setting as in Scenario 2. The cAIC is the corrected conditional AIC as implemented in the cAIC4 package in R [27]. The "BIC_J" is a modified BIC for linear mixed models as introduced in (Jones [11]). "Splmm" (simultaneous penalized linear mixed-effects models) is a method for choosing both the fixed effects and random effects for variable selection using the penalized likelihood function. This method is based on the results in (Yang and Wu [28]) and was implemented in the R-package "$Splmm$". Figure 3 shows that the modified BIC performed 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 was higher for the cAIC than the modified BIC. However, the correction rates for the cAIC were not always higher than that of the modified BIC. The "$Splmm$" method did 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.

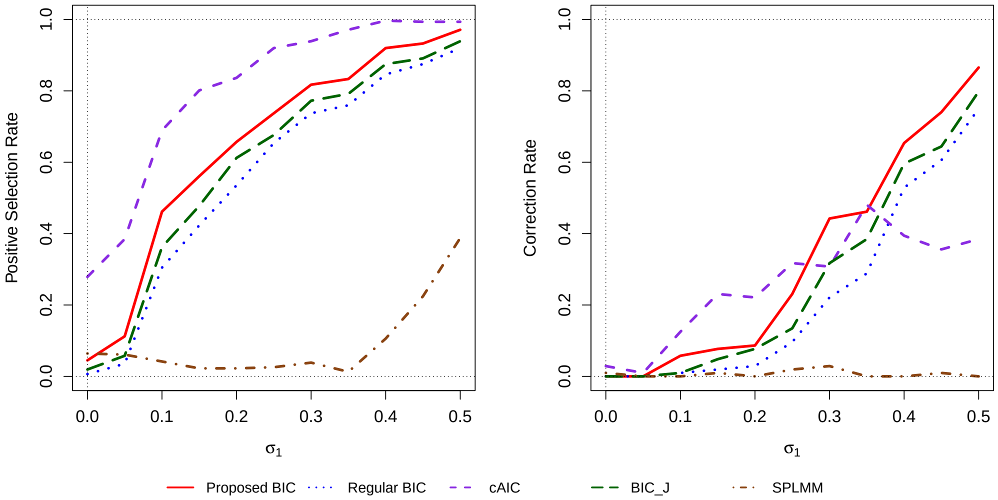

Figure 3. Comparison of the positive selection rate and correction rate for $n = 500$ $(N = 100)$ with different competing methods for different values of σ_1, σ_2, and σ_3. For each value of σ_1 on the horizontal axis, the value of σ_2 is $2 * \sigma_1$ and the value of σ_3 is $4 * \sigma_1$; $\sigma_4 = 0$ and $\sigma_5 = 0$.

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

Table 3. Comparison of the proposed BIC and regular BIC methods in terms of the positive sensitivity rate and correction rate for $n = 250$, $n = 500$, and $n = 1000$ in Scenario 3.

(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 4 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 4. In all cases, the correction rate for the proposed BIC was greater than that of the regular BIC. The difference in the correction rate between these two methods was bigger when the values of σ_1^2, σ_2^2, and σ_3^2 were smaller. In most cases, the two methods seemed to perform better than the cAIC method.

Table 4. 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.

	Proposed BIC			Regular BIC			cAIC		
Case	PSR (SD)	FDR (SD)	Correction Rate	PSR (SD)	FDR (SD)	Correction Rate	PSR (SD)	FDR (SD)	Correction Rate
1	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
2	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
3	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
4	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, and FDR is false discovery rate; both are averaged over 1001 simulations. All values in brackets are sample standard deviations.

Table 5 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 were assumed to be correlated.

Based on the simulation results for the situation when random effects were assumed correlated in Table 5, we saw that the proposed BIC method performed 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 saw that, when the values of the variances for random effects were smaller, the correction rates were lower for all methods. However, the performance of the proposed method was still much better than the other two methods.

Table 5. 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, and σ_2 with σ_3, $\sigma_4 = 0$, $\sigma_5 = 0$, and correlated random effects.

	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; Both-CR is the correction rate of selecting the true model. All the rates are calculated over 1001 simulations.

When random effects were assumed uncorrelated, based on the simulation results in Table 6, we saw that the proposed BIC and regular BIC still performed well and better than the cAIC method. The proposed BIC method performed better than the regular BIC in

Case 2, but did not perform better than the regular BIC in Case 1. This may be because the penalty term of the regular BIC was calculated using the exact chi-squared distribution and the calculation of the penalty term was without any error. However, for the proposed BIC, the weights of the chi-bar-squared distribution were approximated. Therefore, the penalty term was approximated only. From the simulation results, we noticed that when the values of the variances for random effects were smaller, the correction rates were lower for the proposed and regular BIC methods. However, the correction rates in Case 2 were better than Case 1 for the cAIC method.

Table 6. 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, and σ_3 with $\sigma_4 = 0$, $\sigma_5 = 0$, and independent random effects.

	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; Both-CR is the correction rate of selecting the true model. All the rates are calculated over 1001 simulations.

Comparing the computational complexity, the proposed method requires Monte Carlo simulations to estimate the weights so that the penalty parameter can be computed. This is more computational intensive than the regular BIC. In our simulation, for one dataset with a given model, for Setup A, it took about 0.11 to 0.42 s for the proposed BIC method and about 0.04 to 0.07 s for the regular BIC. For Setup B, it took about 0.19 s for the proposed BIC, 0.10 s for the regular BIC, and 0.22 s for the cAIC. For Setup C with independent random effects, it took about 0.11 s for the proposed BIC, 0.05 s for the regular BIC, and 0.09 s for the cAIC. For Setup C with correlated random effects, it took about 0.19 s for the proposed BIC, 0.10 s for the regular BIC, and 0.21 s for the cAIC. We noted that the model with correlated random effects took longer than the one with independent random effects. Furthermore, the computational time of the proposed method was longer than that of the regular BIC, but quite close to that of the cAIC method. The OS and CPU system specifications that we used to run our methods were Windows 10, CPU: Intel Core $i7 - 8550U$ with 4 cores, 8 threads. The memory requirements of our methods are 8 GB RAM.

4. Real-Data Application

In this section, we applied the proposed BIC to a real dataset. We worked with a dataset that is a subset of 120 schools of dataset "hsfull" from package "spida2" in R, which was developed by Monette et al. [29]. This dataset was originally from the 1982 "High School and Beyond" (HSB) survey dataset in Raudenbush and Bryk's text on hierarchical linear models (Raudenbush and Bryk [30]). The data include the 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 were school identification number, mathematics achievement score ($\boldsymbol{Y}$), socioeconomic status ($\boldsymbol{X}_1$), sex (female (0) or male (1); $\boldsymbol{X}_2$), visible minority status (yes (1) or no (0); $\boldsymbol{X}_3$), and school sector (Catholic (0) or public (1); $\boldsymbol{X}_4$). Variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ are group-centered. The objective was to study the relationship between students' mathematics achievement score and socioeconomic status, sex, and 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 were $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$, which are group-centered. The candidate variables in the random effects part were z_1, z_2, and z_3, which are the same as $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$.

We first fit a linear mixed model that included only the intercept term for fixed effects and a random intercept. Then, we fit the models with only the intercept term for fixed effects and all possible combinations of z_1, z_2, and z_3 with a random intercept for random effects. Next, we fit the models with all possible combinations of X_1, X_2, X_3, and X_4 for fixed effects and only a random intercept for random effects. Lastly, for each combination of X_1, X_2, X_3, and X_4 for fixed effects, we fit the models with all possible combinations of z_1, z_2, and z_3 with a random intercept for random effects. For each model, we recorded the values of the proposed BIC, regular BIC, and cAIC. There were 128 values for each method. Now, for each method, we chose the model with the minimum value of the corresponding criterion. We applied this procedure for both cases when random effects were assumed to be correlated and uncorrelated.

When random effects were assumed to be correlated, the optimal model we obtained using the proposed BIC was the model with all X_1, X_2, X_3, and X_4 and a random intercept; the proposed BIC was 34,379.83. The optimal model we obtained using the regular BIC was also the model with X_1, X_2, X_3, and X_4 and a random intercept only. The regular BIC of this model was also 34,379.83. The cAIC yielded the optimal model, which contained X_1, X_2, X_3, and X_4 with a random intercept and random slopes of z_1 and z_3. The cAIC of the optimal model was 34,166.25.

When random effects were assumed to be uncorrelated, the optimal model we obtained using the proposed BIC was the model with all X_1, X_2, X_3, and X_4, a random intercept, and random slopes of z_3; the proposed BIC value was 34,378.23. The optimal model we obtained using the regular BIC was the model with X_1, X_2, X_3, and X_4 and a random intercept only. The regular BIC of this model was 34,379.83. The cAIC yielded the optimal model, which contained X_1, X_2, X_3, and X_4 with a random intercept and random slopes of z_1, z_2, and z_3. The cAIC of the optimal model was 34,165.13.

Table 7 shows the proposed BIC, regular BIC, and cAIC for all models that contained X_1, X_2, X_3, and X_4 with correlated random effects considered.

Table 7. Results of the proposed BIC, regular BIC, and cAIC for all models with correlated random effects considered for the subset of the "hsfull" dataset.

Model	Proposed BIC	Regular BIC	cAIC
Random Intercept (RI)	34,379.33	34,379.83	34,176.92
RI, z_1	34,380.61	34,385.4	34,167.38
RI, z_2	34,391.38	34,396.17	34,181.23
RI, z_1, z_2	34,401.4	34,410.2	34,174.07
RI, z_2	34,384.58	34,389.37	34,169.18
RI, z_1, z_2	34,391.03	34,400.38	34167.38
RI, z_2, z_2	34,405.59	34,414.63	34169.18
RI, z_1, z_2, z_2	34,420.4	34,433.63	34,166.25

All values are rounded to two decimal places.

Table 8 shows the optimal model chosen by each method when the random effects were assumed to be independent and when the random effects were correlated. All X_1, X_2, X_3, and X_4 were included in the models.

Table 8. Comparison of the optimal model chosen by each method for correlated random effects and independent random effects.

	Proposed BIC		Regular BIC		cAIC	
Case	**Optimal Model**	**Proposed BIC**	**Optimal Model**	**Regular BIC**	**Optimal Model**	**cAIC**
Correlated Random Effects	RI	34,379.33	RI	34,379.83	RI, z_1, z_2, z_3	34,166.25
Independent Random Effects	RI, z_3	34,377.73	RI	34,379.83	RI, z_1, z_3	34,165.13

RI means random intercept.

Based on the results presented above, we would choose the model with all X_1, X_2, X_3, and X_4 for fixed effects and a random intercept and a random slope of z_3 for random effects assuming that random effects are uncorrelated. There was a significant relationship between students' math achievement score and socioeconomic status, sex, and visible minority status in public and Catholic schools, and the school mean math achievement score and minority gap effect varied across the schools within each sector.

5. Discussion

In this article, we 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 were assumed to be independent (the covariance matrix of random effects was a diagonal matrix). Second, we proposed a criterion for choosing random effects variance components when the random effects were 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-squared distribution. Lastly, we presented a model selection criterion for choosing both fixed effects and random effects simultaneously in both cases: when random effects were assumed to be independent and when they were correlated. We also proved the consistency of the modified BIC.

Based on the simulation studies, the modified BIC performed quite well in terms of the correction rate. The ability to select the data-generating model of the modified BIC was better when the size of the random effects variance component or the size of correlation component was bigger. Compared to the regular BIC, the modified BIC gave higher correction rates, especially when the variances of random effects were small. Based on the correction rate, the modified BIC and performed better than the regular BIC in most cases. Furthermore, there was significant improvement in the positive selection rate in most of the simulation scenarios.

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 increased exponentially as the number of fixed effects and random effects increased, the model selection process may be increasingly computationally intensive. We may combine the proposed BIC with some selection procedure such as shrinkage methods or fence methods as introduced in Müller et al. [31] to reduce the number of candidate models. Then, we can use the proposed BIC method to perform model selection.

Author Contributions: Writing, original draft, H.L.; writing, reviewing and editing, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by an NSERC grant held by Gao.

Data Availability Statement: The dataset used in the Application Section is a subset of the dataset "hsfull" from the package "spida2" in R, using the R command: data(hsfull, package = 'spida2'). The package is available at https://github.com/gmonette/spida2, accessed on 25 January 2021.

Acknowledgments: The authors would like to thank the Editor and reviewers for their helpful comments and suggestions, which have helped us to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Assumptions for Lemmas 1–3 and Theorems 1 and 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) are satisfied.

(C2). Let $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ be the log-likelihood function of the linear mixed model (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 respect to $\boldsymbol{\theta}$ by $l_N^{'}(\boldsymbol{\theta})$, and denote the matrix of the second partial derivatives of $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ with respect 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) Furthermore, assume that N^{-1} times the absolute value of the third derivative of $l_N(\boldsymbol{\theta}; \boldsymbol{y})$ is bounded as a function of $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$, whose expectation exists, and finite on the intersection of the neighborhoods of $\boldsymbol{\theta}^*$ and $\boldsymbol{\Theta}$.

(C3). Assume that $n_1, \ldots, n_N$ 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, $M_k \subset M_l$, $-2\big(l(\hat{\boldsymbol{\theta}}_k; \boldsymbol{Y}) - l(\hat{\boldsymbol{\theta}}_l; \boldsymbol{Y})\big)$ is bounded by an integrable function, $M(\boldsymbol{Y})$, and $E[M(\boldsymbol{Y})] < \infty$.

Appendix A.2. Graphs of Chi-Bar-Squared Distributions

In this section, we created graphs of some density functions of different chi-bar-squared distributions.

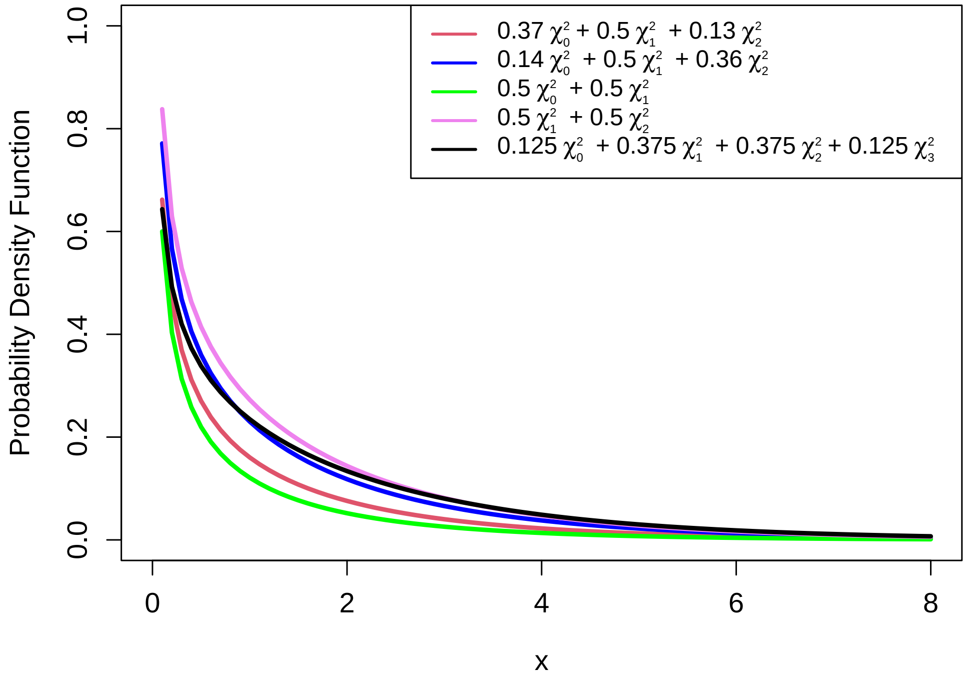

Figure A1. Chi-bar-squared distributions.

The graphs show that the distribution of a chi-bar-squared distribution depends on it mixing weights.

Appendix A.3. Graphical Example of the Boundary Issue

In this example, we tested $H_0 : \sigma_1^2 > 0, \sigma_2^2 = 0$ against $H_1 : \sigma_1^2 > 0, \sigma_2^2 > 0$. The parameter space under the null hypothesis was the set of all points of the form $(a, 0)$ with $a > 0$, illustrated by the blue interval along the axis of σ_1^2. Under the alternative hypothesis, the parameter space was the set of all points (a, b) with $a > 0$ and $b \geq 0$ and is illustrated by the shaded orange region on the graph. Under the null hypothesis, the testing value of the parameter vector lies on the boundary of the parameter space.

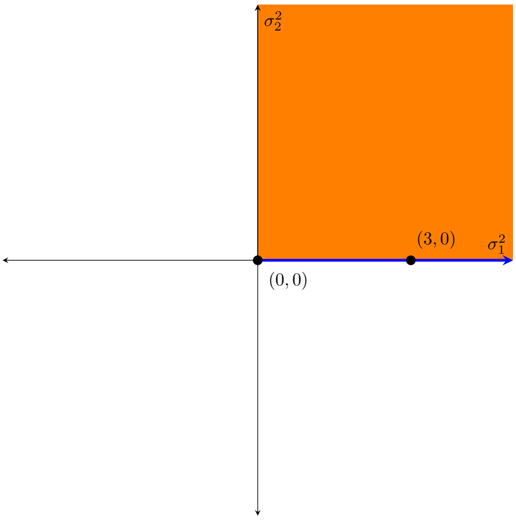

Figure A2. Graphical example of boundary issue.

References

1. Sheng, Y.; Yang, C.; Curhan, S.; Curhan, G.; Wang, M. Analytical methods for correlated data arising from multicenter hearing studies. *Stat. Med.* **2022**, *41*, 5335–5348. [CrossRef] [PubMed]
2. Chernoff, H. On the distribution of the likelihood ratio. *Ann. Math. Stat.* **1954**, *25*, 573–578. [CrossRef]
3. Self, S.G.; Liang, K.Y. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *J. Am. Stat. Assoc.* **1987**, *82*, 605–610. [CrossRef]
4. Stram, D.O.; Lee, J.W. Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics* **1994**, *50*, 1171–1179. [CrossRef] [PubMed]
5. Azadbakhsh, M.; Gao, X.; Jankowski, H. Composite likelihood ratio testing under nonstandard conditions using tangent cones. *Stat* **2021**, *10*, e375. [CrossRef]
6. Baey, C.; Cournède, P.H.; Kuhn, E. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed-effects models. *Comput. Stat. Data Anal.* **2019**, *135*, 107–122. [CrossRef]
7. Dykstra, R. Asymptotic normality for chi-bar-squared distributions. *Can. J. Stat.* **1991**, *19*, 297–306. [CrossRef]
8. Shapiro, A. Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures Under Inequality Constraints. *Biometrika* **1985**, *72*, 133–144. [CrossRef]
9. Vaida, F.; Blanchard, S. Conditional Akaike information for mixed-effects models. *Biometrika* **2005**, *92*, 351–370. [CrossRef]
10. Pauler, D. The Schwarz criterion and related methods for normal linear models. *Biometrika* **1998**, *85*, 13–27. [CrossRef]
11. Jones, R.H. Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* **2011**, *30*, 3050–3056. [CrossRef] [PubMed]
12. Delattre, M.; Poursat, M.A. An iterative algorithm for joint covariate and random effect selection in mixed-effects models. *Int. J. Biostat.* **2020**, *16*, 1–12. [CrossRef] [PubMed]
13. Ibrahim, J.G.; Zhu, H.; Garcia, R.I.; GUO, R. Fixed and random effects selection in mixed-effects models. *Biometrics* **2011**, *67*, 495–503. [CrossRef] [PubMed]
14. Bondell, H.D.; Krishna, A.; Ghosh, S.K. Joint variable selection for fixed and random effects in linear mixed effects models. *Biometrics* **2010**, *66*, 1069–1077. [CrossRef]
15. Peng, H.; Lu, Y. Model selection in linear mixed effect models. *Multivar. Anal.* **2012**, *109*, 109–129. [CrossRef]
16. Drikvandi, R.; Verbeke, G.; Khodadadi, A.; Nia, V.P. Testing multiple variance components in linear mixed-effects models. *Biostatistics* **2012**, *14*, 144–159. [CrossRef]
17. Pauler, D.K.; Wakefield, J.C.; Kass, R.E. Bayes Factors and Approximations for Variance Component Models. *J. Am. Stat. Assoc.* **1999**, *94*, 1242–1253. [CrossRef]
18. Laird, N.M.; Ware, J.H. Random-Effects Models for Longitudinal Data. *Biometrics* **1982**, *38*, 963–974. . [CrossRef]
19. Silvapulle, M.J.; Sen, P.K. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*; John Wiley & Sons: Hoboken, NJ, USA, 2005. [CrossRef]
20. Vanbrabant, L.; Rosseel, Y.; Dacko, A. con_weights_boot: Function for Computing the Chi-Bar-Square Weights Based on Monte Carlo Simulation. 2019. Available online: https://www.rdocumentation.org/packages/restriktor/versions/0.2-250/topics/con_weights_boot/ (accessed on 12 August 2020).
21. van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000.
22. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]

23. Gao, X.; Song, P.X.K. Composite Likelihood Bayesian Information Criteria for Model Selection in High-Dimensional Data. *J. Am. Stat. Assoc.* **2010**, *105*, 1531–1540. [CrossRef]
24. Chen, J.; Chen, Z. Extended BIC for small-n-large-P sparse GLM. *Stat. Sin.* **2012**, *22*, 555–574. [CrossRef]
25. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: https://www.R-project.org/ (accessed on 29 December 2021).
26. Greven, S.; Kneib, T. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **2010**, *97*, 773–789. [CrossRef]
27. Säfken, B.; Rügamer, D.; Kneib, T.; Greven, S. Conditional model selection in mixed-effects models with cAIC4. *arXiv* **2018**, arXiv:1803.05664.
28. Yang, L.; Wu, T. Model-based clustering of high-dimensional longitudinal data via regularization. *Biometrics* **2022**, 1–14. [CrossRef]
29. Monette, G.; Fox, J.; Friendly, M.; Krause, H.; Zhu, F. spida2: Collection of Tools Developed for the Summer Programme in Data Analysis 2000–2012. R Package Version 0.2.1. 2019. Available online: https://github.com/gmonette/spida2 (accessed on 30 April 2020).
30. Raudenbush, S.; Bryk, A. *Hierarchical Linear Models: Applications and Data Analysis Methods*; SAGE Publications: Thousand Oaks, CA, USA, 2002.
31. Müller, S.; Scealy, J.L.; Welsh, A.H. Model Selection in Linear Mixed Models. *Stat. Sci.* **2013**, *28*, 135–167. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.