Open Journal of Statistics, 2016, 6, 239-253
Published Online April 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.62021

Marginal Conceptual Predictive Statistic for Mixed Model Selection

Cheng Wenren[1], Junfeng Shang[2*], Juming Pan[2]

[1]Process Modeling Analytics Department, Bristol-Myers Squibb, New York, NY, USA
[2]Bowling Green State University, Bowling Green, OH, USA
Email: cwenren@gmail.com, *jshang@bgsu.edu, panj@bgsu.edu

Received 9 March 2016; accepted 23 April 2016; published 26 April 2016

Abstract

We focus on the development of model selection criteria in linear mixed models. In particular, we propose the model selection criteria following the Mallows' Conceptual Predictive Statistic (C_p) [1] [2] in linear mixed models. When correlation exists between the observations in data, the normal Gauss discrepancy in univariate case is not appropriate to measure the distance between the true model and a candidate model. Instead, we define a marginal Gauss discrepancy which takes the correlation into account in the mixed models. The model selection criterion, marginal C_p, called MC_p, serves as an asymptotically unbiased estimator of the expected marginal Gauss discrepancy. An improvement of MC_p, called IMC_p, is then derived and proved to be a more accurate estimator of the expected marginal Gauss discrepancy than MC_p. The performance of the proposed criteria is investigated in a simulation study. The simulation results show that in small samples, the proposed criteria outperform the Akaike Information Criteria (AIC) [3] [4] and Bayesian Information Criterion (BIC) [5] in selecting the correct model; in large samples, their performance is competitive. Further, the proposed criteria perform significantly better for highly correlated response data than for weakly correlated data.

Keywords

Mixed Model Selection, Marginal C_p, Improved Marginal C_p, Marginal Gauss Discrepancy, Linear Mixed Model

1. Introduction

With the development in data science over the past decades, people become more aware of the complexity of

*Corresponding author.

How to cite this paper: Wenren, C., Shang, J.F. and Pan, J.M. (2016) Marginal Conceptual Predictive Statistic for Mixed Model Selection. *Open Journal of Statistics*, **6**, 239-253. http://dx.doi.org/10.4236/ojs.2016.62021

data in real life. Univariate linear regression models with independent identically distributed (i.i.d.) Gaussian errors cannot achieve good fitness for some types of data, especially for the data with observations that are correlated. For instance, in longitudinal data, observations are usually recorded from the same individual over time. It is reasonable to assume that correlation exists among the observations from the same individual and linear mixed models are therefore appropriately utilized for modeling such data.

Since linear mixed models are extensively used, mixed model selection plays an important role in statistical literature. The aim of mixed model selection is to choose the most appropriate model from a candidate pool in the mixed model setting. To facilitate this task, a variety of model selection criteria are employed to implement the selection process.

In linear mixed models, a number of criteria have been developed to characterize model selection. The most widely used criteria are the information criteria such as the AIC [3] [4] and the BIC [5]. Sugiura [6] proposed a marginal AIC (mAIC) which involved the number of random effects parameters into the penalty term. Shang and Cavanagh [7] employed the bootstrap method to estimate the penalty term of mAIC for proposing two variants of AIC. For longitudinal data, a special case of linear mixed models, Azari, Li and Tsai [8] proposed a corrected Akaike Information Criterion (AICc). In the justification of AICc, the paper mainly handled the challenge initiated by the correlation matrix under certain conditions for the mixed models. Vaida and Blanchard [9] redefined the Akaike information based on the best linear unbiased predictor (BLUP) [10]-[12] for the random effects in the mixed models, and proposed a conditional AIC (cAIC). Dimova *et al*. [13] derived a series of variants of the Akaike Information Criterion in small samples for linear mixed models.

Another information criterion, BIC, can be considered as a Bayesian alternative to AIC. In linear mixed models, BIC is converted from marginal AIC by replacing the constant 2 in the penalty by $\log(N)$, where N is the sample size (mBIC) [14]. Jones [15] proposed a measure of the effective sample size to replace the sample size in the penalty term of BIC, leading to a new criterion BIC_J.

We note that the BIC-type information criteria are derived using Bayesian approaches. Different from that, the AIC-type information selection criteria are justified from the frequentist perspective and based upon the information discrepancy. However, little research has relied on other discrepancy to propose criteria including Mallows' C_p [1] [2] in linear mixed models. In fact, because of dissimilar derivation, each selection criterion has its own advantages, and no unique selection criterion can cover all the benefits for model selection. To further develop the selection criteria in the mixed modeling setting, we aim to justify the C_p-type ones relying on the Gauss discrepancy.

Mallows' C_p [1] [2] in linear regression models targets to estimate the Gauss discrepancy between the true model and a candidate model. It serves as an asymptotically unbiased estimator of the expected Gauss discrepancy. Fujikoshi and Satoh [16] identified C_p in multivariate linear regression. Davies *et al.* [17] presented the estimation optimality of C_p in linear regression models. Cavanaugh *et al*. [18] provided an alternate version of C_p. The Gauss discrepancy is an L2 norm measuring the distance between the true model and a candidate model in linear models. To select the most appropriate model among competing fitted models, the candidate model leading to the smallest value of C_p is chosen. However, since the covariance matrix of linear mixed models poses the challenge for the justification of selection criteria, C_p statistic in linear mixed models has not been identified.

This paper extends the justification of C_p from linear models to linear mixed models. We first define a marginal Gauss discrepancy reflecting the correlation for measuring the distance between the true model and a candidate model. We utilize the assumption that under certain conditions, the estimator of the correlation matrix for the candidate model is consistent to that for the true correlation matrix. The marginal C_p, abbreviated as MC_p. MC_p serves as an asymptotically unbiased estimator of the expected marginal Gauss discrepancy between the true model and a candidate model. An improvement of MC_p, abbreviated as IMC_p, is also proposed and proved. We then justify IMC_p as an asymptotically more precisely unbiased estimator of the expected marginal Gauss discrepancy. We examine the performance of the proposed criteria in a simulation study where we utilize various correlation structures and different sample sizes.

The paper is organized as follows: Section 2 presents the notation and defines the marginal Gauss discrepancy in the setting of linear mixed models. In Section 3, we provide the derivations of the model selection criteria MC_p and IMC_p. Section 4 presents a simulation study to demonstrate the effectiveness of the proposed criteria. Section 5 concludes.

2. Marginal Gauss Discrepancy

In this section, we will introduce the true model, also called the generating model, and the candidate model in the setting of linear mixed models, then define the marginal Gauss discrepancy.

Suppose that the generating model for the data is given by

$$y = X_o\beta_o + Zb_o + \varepsilon_o, \tag{2.1}$$

where y denotes an $N \times 1$ response vector, X_o is an $N \times p_o$ design matrix of full column rank, β_o is a $p_o \times 1$ unknown vector for fixed effects. Z is an $N \times mr$ known matrix of full column rank and b_o is an $mr \times 1$ unknown vector for random effects, where m is the number of cases, the sample size, and r is the dimension of the random effects for each case. Here, $b_o \sim N\left(0, \sigma_o^2 G_o\right)$, $\varepsilon_o \sim N\left(0, \sigma_o^2 I_N\right)$, and b_o and ε_o are mutually independent and G_o is a positive definite matrix and σ_o^2 is a scalar.

We fit the data with a candidate model of the form

$$y = X\beta + Zb + \varepsilon, \tag{2.2}$$

where X is an $N \times p$ design matrix of full column rank, β is a $p \times 1$ unknown vector, $b \sim N\left(0, \sigma^2 G\right)$, $\varepsilon \sim N\left(0, \sigma^2 I_N\right)$, and b and ε are mutually independent. The design matrix of the random effects Z and the random effects b are the same as those in the generating model. The matrix G is a positive definite matrix with the q unknown parameters in it.

Since the random part of the model (*i.e.* Zb) is not subject to selection, it is easier to use the marginal form in [19] of linear mixed models. Let $\zeta_o = Zb_o + \varepsilon_o$, then the generating model (2.1) can be written as

$$\begin{aligned} y &= X_o\beta_o + \zeta_o, \\ \zeta_o &\sim N\left(0, \sigma_o^2\Sigma_o\right), \end{aligned} \tag{2.3}$$

where the scaled variance $\Sigma_o = ZG_oZ^{\mathrm{T}} + I_N$.

For the candidate model (2.2), let $\zeta = Zb + \varepsilon$, we have

$$\begin{aligned} y &= X\beta + \zeta, \\ \zeta &\sim N\left(0, \sigma^2\Sigma\right), \end{aligned} \tag{2.4}$$

where the scaled variance $\Sigma = ZGZ^{\mathrm{T}} + I_N$. Therefore, the Σ is a nonsingular positive definite matrix.

In models (2.3) and (2.4), the terms ζ_o and ζ are the combinations of the random effects and errors in the model, respectively. Since they are both assumed to have mean zero, the parameters scaled variances Σ_o and Σ contain all the information of the random effects and errors, including the correlation structures.

We measure the distance between the true model and a candidate model by defining the marginal Gauss discrepancy based on the marginal forms of models (2.3) and (2.4). The true model is assumed to be included in the pool of candidate models. Let θ_o and θ denote the vectors of parameters $\left(\beta_o^{\mathrm{T}}, \sigma_o, \Sigma_o\right)^{\mathrm{T}}$ and $\left(\beta^{\mathrm{T}}, \sigma, \Sigma\right)^{\mathrm{T}}$, respectively. The marginal Gauss discrepancy between the true model and a candidate model is defined as

$$d^G\left(\theta, \theta_o\right) = E_o\left\{\left(y - X\beta\right)^{\mathrm{T}}\Sigma^{-1}\left(y - X\beta\right)\right\},$$

where E_o denotes the expectation with respect to the true model. Note that the marginal Gauss discrepancy contains a weight of inverse scaled variance Σ^{-1} into the L_2 norm. Therefore, the correlation between observations is involved when we use the marginal Gauss discrepancy to measure the distance between the true model and a candidate model.

Now let $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}, \hat{\Sigma}\right)^{\mathrm{T}}$ denote an estimate of θ. For instance, $\hat{\theta}$ could be the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (REML). However, in this paper, the MLE is utilized. The marginal Gauss discrepancy between the true model and the fitted candidate model is defined as

$$d^G\left(\hat{\theta}, \theta_o\right) = d^G\left(\theta, \theta_o\right)\Big|_{\theta=\hat{\theta}},$$

which can be therefore expressed as

$$
\begin{aligned}
& d^{G}\left(\hat{\theta},\theta_{o}\right) \\
&= E_{o}\left\{\left(y-X\beta\right)^{\mathrm{T}}\Sigma^{-1}\left(y-X\beta\right)\right\}\Big|_{\theta=\hat{\theta}} \\
&= E_{o}\left\{\left(y-X_{o}\beta_{o}+X_{o}\beta_{o}-X\beta\right)^{\mathrm{T}}\Sigma^{-1}\left(y-X_{o}\beta_{o}+X_{o}\beta_{o}-X\beta\right)\right\}\Big|_{\theta=\hat{\theta}} \\
&= E_{o}\left\{\left(y-X_{o}\beta_{o}\right)^{\mathrm{T}}\Sigma^{-1}\left(y-X_{o}\beta_{o}\right)\right\}\Big|_{\theta=\hat{\theta}}+E_{o}\left\{\left(X_{o}\beta_{o}-X\beta\right)^{\mathrm{T}}\Sigma^{-1}\left(X_{o}\beta_{o}-X\beta\right)\right\}\Big|_{\theta=\hat{\theta}} \\
&= \sigma_{o}^{2}tr\left(\hat{\Sigma}^{-1}\Sigma_{o}\right)+\left(X_{o}\beta_{o}-X\hat{\beta}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(X_{o}\beta_{o}-X\hat{\beta}\right).
\end{aligned}
\tag{2.5}
$$

We define a transformed marginal Gauss discrepancy between the true generating model and the fitted candidate model as a linear function of the marginal Gauss discrepancy (2.5) as

$$
d_{C_{p}}\left(\hat{\theta},\theta_{o}\right)=\frac{1}{\sigma_{o}^{2}}d^{G}\left(\hat{\theta},\theta_{o}\right)-N \tag{2.6}
$$

Taking the expectation of the transformed marginal Gauss discrepancy (2.6), we obtain the expected transformed marginal Gauss discrepancy as

$$
\begin{aligned}
\Delta_{C_{p}}\left(\theta_{o}\right) &= E_{o}\left\{d_{C_{p}}\left(\hat{\theta},\theta_{o}\right)\right\} \\
&= E_{o}\left\{tr\left(\hat{\Sigma}^{-1}\Sigma_{o}\right)\right\}+\frac{E_{o}\left\{\left(X_{o}\beta_{o}-X\hat{\beta}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(X_{o}\beta_{o}-X\hat{\beta}\right)\right\}}{\sigma_{o}^{2}}-N.
\end{aligned}
\tag{2.7}
$$

To serve as a model selection criterion based on the expected transformed marginal Gauss discrepancy in Equation (2.7), an unbiased estimator or an asymptotically unbiased estimator will be proposed. To simplifying the procedure, we will first abbreviate this discrepancy in Equation (2.7).

From expression (2.7), the expectation part in the numerator can be written as

$$
E_{o}\left\{\left(X_{o}\beta_{o}-\hat{H}y\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(X_{o}\beta_{o}-\hat{H}y\right)\right\}, \tag{2.8}
$$

where $\hat{H}=X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)X^{\mathrm{T}}\hat{\Sigma}^{-1}$ is a projection matrix such that $X\hat{\beta}=\hat{H}y$. To explore a further expression of (2.8), we need to know the properties of $\hat{H}$.

Theorem 1. *For every* $\hat{\Sigma}$, *the matrix* $\hat{H}=X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}$ *satisfies the following properties*:

1) $\hat{H}$ *is idempotent.*
2) $tr\left(\hat{H}\right)=p$ *and* $tr\left(I_{N}-\hat{H}\right)=N-p$.

The proof is given in the Appendix.

Corollary 1. *Following Theorem* 1, *we have*:

1) $\hat{H}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{H}=\hat{\Sigma}^{-1}\hat{H}$.
2) $\hat{H}^{\mathrm{T}}\hat{\Sigma}^{-1}\left(\hat{H}-I\right)=0$.
3) $\left(\hat{H}-I\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(\hat{H}-I\right)=\left(I-\hat{H}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)=\hat{\Sigma}^{-1}\left(I-\hat{H}\right)$.

The proof of Corollary 1 can be easily completed following Theorem 1.

By Corollary 1, expression (2.8) can be written as

$$
\begin{aligned}
& E_{o}\left\{\left(\hat{H}y-X_{o}\beta_{o}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(\hat{H}y-X_{o}\beta_{o}\right)\right\} \\
&= E_{o}\left\{\left(\left(\hat{H}y-\hat{H}X_{o}\beta_{o}\right)+\left(\hat{H}X_{o}\beta_{o}-X_{o}\beta_{o}\right)\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(\left(\hat{H}y-\hat{H}X_{o}\beta_{o}\right)+\left(\hat{H}X_{o}\beta_{o}-X_{o}\beta_{o}\right)\right)\right\} \\
&= E_{o}\left\{\left(y-X_{o}\beta_{o}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{H}\left(y-X_{o}\beta_{o}\right)\right\}+E_{o}\left\{\left(X_{o}\beta_{o}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(\hat{H}-I\right)\left(X_{o}\beta_{o}\right)\right\} \\
&= E_{o}\left\{\zeta_{o}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{H}\zeta_{o}\right\}+E_{o}\left\{\left(X_{o}\beta_{o}\right)^{T}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\left(X_{o}\beta_{o}\right)\right\}.
\end{aligned}
\tag{2.9}
$$

Note that the scaled variance Σ is a function of the q unknown parameter vector of variance components γ, *i.e.*, $\Sigma = \Sigma(\gamma)$. Azari, Li and Tsai [8] noted that under the assumption that the set of candidate models includes the true model, it is reasonable to assume that the MLE $\hat{\gamma}$ is a consistent estimator of γ_o. Therefore, we can approximate $\hat{\Sigma}$ by Σ, *i.e.*, $\hat{\Sigma} = \Sigma_o + o(1)$. In what follows, we will make use of this approximation.

First, since $E_o\{\zeta_o\} = 0$ and $var_o\{\zeta_o\} = \sigma_o^2\Sigma_o$, using the approximation $\hat{\Sigma} = \Sigma_o + o(1)$ and Theorem 1, we have the first term of (2.9) as

$$\begin{aligned} E_o\left\{\zeta_o^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{H}\zeta_o\right\} &= \sigma_o^2 tr\left(\hat{\Sigma}^{-1}\hat{H}\Sigma_o\right) \\ &\approx \sigma_o^2 tr\left(\hat{H}\right) = p\sigma_o^2. \end{aligned} \tag{2.10}$$

Second, using the approximation $\hat{\Sigma} = \Sigma_o + o(1)$ again, the first term of Equation (2.7) can be simplified as

$$E_o\left\{tr\left(\hat{\Sigma}^{-1}\Sigma_o\right)\right\} \approx N. \tag{2.11}$$

Using expressions (2.9), (2.10), and (2.11), $\Delta_{C_p}(\theta_o)$ in (2.7) can be therefore approximated as

$$\begin{aligned} \Delta_{C_p}(\theta_o) &\approx \frac{p\sigma_o^2 + E_o\left\{(X_o\beta_o)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)(X_o\beta_o)\right\}}{\sigma_o^2} \\ &= p + \frac{E_o\left\{(X_o\beta_o)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)(X_o\beta_o)\right\}}{\sigma_o^2}. \end{aligned} \tag{2.12}$$

Following Mallows' interpretation, $\Delta_{C_p}(\theta_o)$ in (2.12) can be expressed as

$$\Delta_{C_p}(\theta_o) \approx V_P + \frac{B_p}{\sigma_o^2},$$

where V_P and B_p are respectively "variance" and "bias" contributions given by

$$V_P = p$$

and

$$B_p = E_o\left\{(X_o\beta_o)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)(X_o\beta_o)\right\}.$$

We comment that increasing the number of the parameters of the fixed effects p will decrease the bias B_p for the fitted model, yet will increase the variance V_P at the same time. The marginal Gauss discrepancy can therefore be considered as a bias-variance trade-off. Since a smaller value of the discrepancy indicates a smaller distance between the true model and a candidate model, the size of the Gauss discrepancy can really reflect how a fitted model is close to the true model.

3. Derivations of Marginal C_p and Improved Marginal C_p

3.1. Marginal C_p

In this section, model selection criteria based on $\Delta_{C_p}(\theta_o)$ are developed by finding a statistic that has an expectation which equals to or asymptotically equals to the expected transformed marginal Gauss discrepancy.

We start with the expectation of the sum of squared errors SS_{Res} from a candidate model. In linear mixed models, the sum of squared errors SS_{Res} can be written as

$$SS_{Res} = \left(y - X\hat{\beta}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(y - X\hat{\beta}\right).$$

By Theorem 1 and Corollary 1, the expectation of the "scaled sum of squared error" $\frac{SS_{Res}}{\sigma_o^2}$ can be expressed by

$$E_o\left\{\frac{SS_{Res}}{\sigma_o^2}\right\}=E_o\left\{\frac{\left(y-X\hat{\beta}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(y-X\hat{\beta}\right)}{\sigma_o^2}\right\}$$
$$=E_o\left\{\frac{\left(y-\hat{H}y\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(y-\hat{H}y\right)}{\sigma_o^2}\right\}$$
$$=E_o\left\{\frac{y^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)y}{\sigma_o^2}\right\},$$

and then we have

$$E_o\left\{\frac{SS_{Res}}{\sigma_o^2}\right\}=\frac{E_o\left\{\left(y-X_o\beta_o+X_o\beta_o\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\left(y-X_o\beta_o+X_o\beta_o\right)\right\}}{\sigma_o^2}$$
$$=\frac{E_o\left\{\left(y-X_o\beta_o\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\left(y-X_o\beta_o\right)\right\}}{\sigma_o^2}+\frac{E_o\left\{\left(X_o\beta_o\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\left(X_o\beta_o\right)\right\}}{\sigma_o^2} \quad (3.1)$$
$$=\frac{E_o\left\{\zeta_o^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\zeta_o\right\}}{\sigma_o^2}+\frac{E_o\left\{\left(X_o\beta_o\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\left(X_o\beta_o\right)\right\}}{\sigma_o^2}.$$

Similar to the derivation of Equation (2.11), the numerator of first term of Equation (3.1) is expressed as

$$E_o\left\{\zeta_o^{\mathrm{T}}\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\zeta_o\right\}=\sigma_o^2 tr\left(\hat{\Sigma}^{-1}\left(I-\hat{H}\right)\Sigma_o\right)$$
$$\approx\sigma_o^2 tr\left(I-\hat{H}\right)=\left(N-p\right)\sigma_o^2. \quad (3.2)$$

Then, by Equations (3.1) and (3.2), it is straightforward to construct a function $T=\frac{SS_{Res}}{\sigma_o^2}+2p-N$, which is a linear combination of $\frac{SS_{Res}}{\sigma_o^2}$. It can be shown that the function T has the expectation

$$E_o\{T\}=E_o\left\{\frac{SS_{Res}}{\sigma_o^2}+2p-N\right\}=\frac{1}{\sigma_o^2}E_o\left\{SS_{Res}\right\}+2p-N$$
$$=N-p+\frac{E_o\left\{\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}\left(I-\hat{H}\right)\left(X_o\beta_o\right)\right\}}{\sigma_o^2}+2p-N$$
$$=p+\frac{E_o\left\{\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}\left(I-H\right)\left(X_o\beta_o\right)\right\}}{\sigma_o^2}=V_P+\frac{B_p}{\sigma_o^2}\approx\Delta_{C_p}\left(\theta_o\right).$$

Note that the function T is not a statistic since the parameter σ_o^2 is unknown. Here, we would like to use an estimator $\hat{\sigma}^2$ to replace σ_o^2 in the function. Let X_* denote the design matrix for the largest model in the candidate pool with $rank\left(X_*\right)=p_*$. We assume that $C\left(X\right)\subseteq C\left(X_*\right)$. Let SS_{Res}^* represent the sum of squared errors for the corresponding fitted model and is written as

$$SS_{Res}^*=\left(y-X_*\hat{\beta}_*\right)^{\mathrm{T}}\hat{\Sigma}_*^{-1}\left(y-X_*\hat{\beta}_*\right),$$

where $\hat{\beta}_*$ and $\hat{\Sigma}_*^{-1}$ are the MLEs for parameters β_* and Σ_*^{-1} in the largest candidate model respectively. The estimator $\hat{\Sigma}_*^{-1}$ cannot be expressed in a closed form and is calculated by computational algorithm where the iterations are needed.

For the estimator of σ_o^2, we use the mean squared error of the largest candidate model

$$\hat{\sigma}^2 = \frac{SS^*_{Res}}{N - p_*}, \tag{3.3}$$

which is an asymptotically unbiased estimator for σ_o^2, yet it is biased. In the justification of this estimator, using the approximation $\hat{\Sigma}_* = \Sigma_o + o(1)$, we can represent $\hat{\beta}_*$ in terms of Σ_o, then the expected value of SS^*_{Res} can be easily calculated as $(N - p_*)\sigma_o^2$, *i.e.*, asymptotically we can have $E_o(\hat{\sigma}^2) = \sigma_o^2$. Serving as an asymptotically unbiased estimator of σ_o^2, the $\hat{\sigma}^2$ in Equation (3.3) for the largest candidate model is preferred to estimate σ_o^2.

MC_p is then obtained as

$$MC_p = \frac{SS_{Res}}{\hat{\sigma}^2} + 2p - N = \frac{(N - p_*)SS_{Res}}{SS^*_{Res}} + 2p - N. \tag{3.4}$$

Note that MC_p is biased for $\Delta_{C_p}(\theta_o)$. However, under the assumption that the true model is included in the pool of candidate models, MC_p serves as an asymptotically unbiased estimator of the discrepancy in expression (2.7). The proof is nontrivial, yet the simulations (not presented here) can show that as the samples size increases, the curves of the average values for MC_p and the discrepancy $\Delta_{C_p}(\theta_o)$, along with IMC_p, which will be introduced in the following subsection, collectively get merged, indicating that MC_p and IMC_p are all asymptotically unbiased estimators of the discrepancy $\Delta_{C_p}(\theta_o)$.

3.2. Improved Marginal C_p

To improve the performance of the MC_p statistic in linear mixed models, we wish to propose an improved marginal C_p, called IMC_p, which is expected to be a more accurate or less biased estimator of the expected transformed marginal Gauss discrepancy than MCp. IMC_p is proposed as

$$IMC_p = \frac{(N - p_* - 2)SS_{Res}}{SS^*_{Res}} + 2p - N + 2, \tag{3.5}$$

where SS_{Res} and SS^*_{Res} are the sum of squared errors from the candidate fitted model and the largest fitted model, respectively. Note that IMC_p provides us an asymptotically unbiased estimator of $\Delta_{C_p}(\theta_o)$, *i.e.*, $E_o\{IMC_p\} \cong \Delta_{C_p}(\theta_o)$, and it will be shown in what follows.

To evaluate the expectation of IMC_p, we first need to calculate the ratio of the sum of squared errors $\frac{SS_{Res}}{SS^*_{Res}}$ between the candidate model and the largest candidate model in the pool. By Corollary 1, we have

$$\begin{aligned}\frac{SS_{Res}}{SS^*_{Res}} &= \frac{(y - X\hat{\beta})^{\mathrm{T}}\hat{\Sigma}^{-1}(y - X\hat{\beta})}{(y - X_*\hat{\beta}_*)^{\mathrm{T}}\hat{\Sigma}_*^{-1}(y - X_*\hat{\beta}_*)} = \frac{(y - \hat{H}y)^{\mathrm{T}}\hat{\Sigma}^{-1}(y - \hat{H}y)}{(y - \hat{H}_*y)^{\mathrm{T}}\hat{\Sigma}_*^{-1}(y - \hat{H}_*y)} \\ &= \frac{y^{\mathrm{T}}(I - \hat{H})^{\mathrm{T}}\hat{\Sigma}^{-1}(I - \hat{H})y}{y^{\mathrm{T}}(I - \hat{H}_*)^{\mathrm{T}}\hat{\Sigma}_*^{-1}(I - \hat{H}_*)y} = \frac{y^{\mathrm{T}}\hat{\Sigma}^{-1}(I - \hat{H})y}{y^{\mathrm{T}}\hat{\Sigma}_*^{-1}(I - \hat{H}_*)y}.\end{aligned}$$

By using the approximation $\hat{\Sigma} = \Sigma_o + o(1)$ for all $\hat{\Sigma}$, we approximate $\hat{H}$ and $\hat{H}_*$ by H and H_*, respectively, and $H = X(X^{\mathrm{T}}\Sigma_o^{-1}X)^{-1}X^{\mathrm{T}}\Sigma_o^{-1}$ and $H_* = X_*(X_*^{\mathrm{T}}\Sigma_o^{-1}X_*)^{-1}X_*^{\mathrm{T}}\Sigma_o^{-1}$. Then, the ratio $\frac{SS_{Res}}{SS^*_{Res}}$ can be written as

$$\begin{aligned}\frac{SS_{Res}}{SS^*_{Res}} &= \frac{y^{\mathrm{T}}\hat{\Sigma}^{-1}(I - H)y}{y^{\mathrm{T}}\hat{\Sigma}_*^{-1}(I - H_*)y} \approx \frac{y^{\mathrm{T}}\Sigma_o^{-1}(I - H)y}{y^{\mathrm{T}}\Sigma_o^{-1}(I - H_*)y} \\ &= \frac{y^{\mathrm{T}}\Sigma_o^{-1}(I - H_* + H_* - H)y}{y^{\mathrm{T}}\Sigma_o^{-1}(I - H_*)y} = 1 + \frac{y^{\mathrm{T}}\Sigma_o^{-1}(H_* - H)y}{y^{\mathrm{T}}\Sigma_o^{-1}(I - H_*)y}.\end{aligned} \tag{3.6}$$

To continue the proof, we will use the following theorem and corollary.

Theorem 2. *If* $C(X)\subseteq C(X_*)$, *then for any* $N\times N$ *matrix K, we have* $C\left(K^{\mathrm{T}}X\right)\subseteq C\left(K^{\mathrm{T}}X_*\right)$.
The proof of Theorem 2 is presented in the Appendix.
Corollary 2. *Following Theorem* 2, *we can obtain following results*:
1) $\Sigma_o^{-1}HH_* = \Sigma_o^{-1}H_*H = \Sigma_o^{-1}H$.
2) $\Sigma_o^{-1}H_*X = \Sigma_o^{-1}X$.
The proof of Corollary 2 is included in the Appendix.
By Theorem 1 and Corollary 2, we have

$$\Sigma_o^{-1}(H_*-H)\Sigma_o\Sigma_o^{-1}(I-H_*)=0,$$

such that the quadratic forms $y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y$ and $y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y$ are independent. It follows that the expectation of $\frac{SS_{Res}}{SS_{Res}^*}$ in (3.6) can be written as

$$\begin{aligned}E_o\left\{\frac{SS_{Res}}{SS_{Res}^*}\right\} &\approx 1+E_o\left\{\frac{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\}\\ &=1+E_o\left\{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y\right\}E_o\left\{\frac{1}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\}.\end{aligned}\tag{3.7}$$

For the term $E_o\left\{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y\right\}$ in (3.7), since $y\sim N\left(X_o\beta_o,\sigma_o^2\Sigma_o\right)$, we have

$$\begin{aligned}&E_o\left\{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y\right\}\\ &=\sigma_o^2 tr\left(\Sigma_o^{-1}(H_*-H)\Sigma_o\right)+\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)\left(X_o\beta_o\right)\\ &=\sigma_o^2 tr(H_*-H)+\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)\left(X_o\beta_o\right)\\ &=\sigma_o^2(p_*-p)+\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)\left(X_o\beta_o\right)\\ &=\sigma_o^2(p_*-p)+\left(X_o\beta_o\right)^{\mathrm{T}}\Sigma_o^{-1}(I-H)\left(X_o\beta_o\right).\end{aligned}\tag{3.8}$$

For the term $E_o\left\{\frac{1}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\}$ in (3.7), we can prove that

$$\frac{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}{\sigma_o^2}\sim\chi^2_{rank(I-H_*)}.$$

Note that $rank(I-H_*)=N-p_*$. To justify the distribution of $\frac{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}{\sigma_o^2}$, we have

$$\frac{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}{\sigma_o^2}=y^{\mathrm{T}}Ay,$$

where $A=\frac{\Sigma_o^{-1}(I-H_*)}{\sigma_o^2}$. For the distribution of *y*, we know that $y\sim N\left(X_o\beta_o,\sigma_o^2\Sigma_o\right)$. We calculate that $\sigma_o^2\Sigma_o A=I-H_*$, and by Theorem 1, the matrix $I-H_*$ is idempotent. Therefore, we have $y^{\mathrm{T}}Ay\sim\chi^2_{\nu,\frac{\lambda}{2}}$, where $\nu=rank(I-H_*)=N-p_*$ and by Corollary 2, we can calculate λ as

$$\lambda=\left(X_o\beta_o\right)^{\mathrm{T}}AX_o\beta_o=\left(X_o\beta_o\right)^{\mathrm{T}}\frac{\Sigma_o^{-1}(I-H_*)}{\sigma_o^2}X_o\beta_o=0.$$

Now, its inverse $\frac{\sigma_o^2}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}$ follows an inverse Chi-square distribution, *i.e.*,

$\frac{\sigma_o^2}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y} \sim I\chi^2_{rank(I-H_*)}$, with the expectation as

$$E_o\left\{\frac{\sigma_o^2}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\} = \frac{1}{N-p_*-2}. \tag{3.9}$$

Using the results of (3.8) and (3.9), we have the expectation of $E_o\left\{\frac{SS_{Res}}{SS^*_{Res}}\right\}$ in (3.7) as

$$\begin{aligned} E_o\left\{\frac{SS_{Res}}{SS^*_{Res}}\right\} &\approx 1 + E_o\left\{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y\right\} E_o\left\{\frac{1}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\} \\ &= 1 + E_o\left\{y^{\mathrm{T}}\Sigma_o^{-1}(H_*-H)y\right\} E_o\left\{\frac{\sigma_o^2}{y^{\mathrm{T}}\Sigma_o^{-1}(I-H_*)y}\right\}\frac{1}{\sigma_o^2} \\ &= 1 + \frac{1}{\sigma_o^2}\left(\sigma_o^2(p_*-p) + (X_o\beta_o)^{\mathrm{T}}\Sigma_o^{-1}(I-H)(X_o\beta_o)\right)\left(\frac{1}{N-p_*-2}\right) \\ &= 1 + \frac{p_*-p}{N-p_*-2} + \frac{(X_o\beta_o)^{\mathrm{T}}\Sigma_o^{-1}(I-H)(X_o\beta_o)}{\sigma_o^2(N-p_*-2)} \\ &= \frac{N-p-2}{N-p_*-2} + \frac{(X_o\beta_o)^{\mathrm{T}}\Sigma_o^{-1}(I-H)(X_o\beta_o)}{\sigma_o^2(N-p_*-2)}. \end{aligned} \tag{3.10}$$

We recall that the criterion $\mathrm{IMC_p}$ in (3.5) is defined as

$$\mathrm{IMC_p} = (N-p_*-2)\frac{SS_{Res}}{SS^*_{Res}} + 2p - N + 2.$$

By the result of (3.10) and the approximation $\hat{\Sigma} = \Sigma_o + o(1)$ again , we have the expectation of $\mathrm{IMC_p}$ as

$$\begin{aligned} E_o\left\{\mathrm{IMC_p}\right\} &= (N-p_*-2)E_o\left\{\frac{SS_{Res}}{SS^*_{Res}}\right\} + 2p - N - 2 \\ &\approx (N-p_*-2)\left\{\frac{N-p-2}{N-p_*-2} + \frac{(X_o\beta_o)^{\mathrm{T}}\Sigma_o^{-1}(I-H)(X_o\beta_o)}{\sigma_o^2(N-p_*-2)}\right\} + 2p - N + 2 \\ &\approx p + \frac{E_o\left\{(X_o\beta_o)^{\mathrm{T}}\hat{\Sigma}^{-1}(I-\hat{H})(X_o\beta_o)\right\}}{\sigma_o^2} \approx \Delta_{C_p}(\theta_o). \end{aligned}$$

Hence, $\mathrm{IMC_p}$ is an asymptotically unbiased estimator of the expected overall transformed Gauss discrepancy $\Delta_{C_p}(\theta_o)$ in Equation (2.7). The advantage of $\mathrm{IMC_p}$ is that it avoids the bias of using $\frac{1}{\hat{\sigma}^2}$ to estimate $\frac{1}{\sigma_o^2}$ to derive the criterion comparing to the derivation of $\mathrm{MC_p}$.

We comment that the proposed $\mathrm{MC_p}$ and $\mathrm{IMC_p}$ are justified based upon the assumption that the true model is contained in the candidate models. Hence, we can calculate the $\mathrm{MC_p}$ and $\mathrm{IMC_p}$ values for the correctly and over-fitted candidate models. However, the proposed criteria are also can be utilized for the underspecified models except that the values will be quite large and not behave well.

4. Simulation Study

In this simulation study, we investigate the ability of $\mathrm{MC_p}$ in (3.4) and $\mathrm{IMC_p}$ in (3.5) to determine the correct set of fixed effects for the simulated data in different models.

4.1. Presentation of Simulations

Consider a setting in which data are generated by the model of the following form

$$y_{ij} = X_{ij}^{\mathrm{T}}\beta + b_i + \varepsilon_{ij},\ i = 1,\cdots,m,\ j = 1,\cdots,n,$$

where the random effects $b_1,\cdots,b_m$ are uncorrelated with mean 0 and variance τ^2, the errors ε_{ij} are independent with each other with mean 0 and variance σ^2. It follows that the correlation between any two observations from the same case is $\frac{\tau^2}{\tau^2+\sigma^2}$, whereas the observations from different cases are uncorrelated. Let ϕ denote the proportion between the variance of the random effects and the variance of the errors, *i.e.* $\phi = \frac{\tau^2}{\sigma^2}$. We can obtain that the correlation between the observations from the same case equals $\frac{\phi}{1+\phi}$, which is an increasing function of ϕ. Therefore, a higher ϕ implies a higher correlation between the observations in the same case.

For convenience, the generating model can also be expressed by

$$y = X\beta + Zb + \varepsilon,$$

where β are unknown coefficients of the fixed effects. It is assumed that the random effects $b \sim N\left(0,\sigma^2 G\right)$ with $G = \phi I_m$, and $r = 1$. We set $Z_i = j_{n_i}$ for $j_{n_i} = (1,\cdots,1)^{\mathrm{T}}$, an n_i-vector of ones, and $n_1 = \cdots = n_m = n = N/m$. We also assume that the error term $\varepsilon \sim N\left(0,\sigma^2 I_N\right)$, and is independent of the random effects *b*.

Since the random part of the model (*i.e. Zb*) is not subject to selection, we would like to express the model by its marginal form. Let $\zeta_{ij} = b_i + \varepsilon_{ij}$, we have

$$y_{ij} = X_{ij}^{\mathrm{T}}\beta + \zeta_{ij},$$

which can also be expressed by the general form as

$$y = X\beta + \zeta,\ \zeta \sim N\left(0,\sigma^2\Sigma\right), \tag{4.1}$$

where $\zeta = Zb + \varepsilon$, $\Sigma = \frac{\tau^2}{\sigma^2}ZZ^{\mathrm{T}} + I_N$ is a scaled covariance matrix. Equivalently, the term ζ has the following exchangeable correlation structure: $Var(\zeta) = (\phi+1)\sigma^2\left\{\left(1-\frac{\phi}{1+\phi}\right)I + \frac{\phi}{1+\phi}J\right\}$, where $\phi = \frac{\tau^2}{\sigma^2}$, I is the identity matrix and J is the matrix of 1's.

In this simulation study, we generate the design matrix X with $rank(X)$ of 5. The first column of X is 1 and the other four columns of X are generated randomly from uniform distributions but are fixed throughout the simulations. Therefore, the number of fixed effects including the intercept in the largest model is $p_* = 5$. We assume that the candidate vectors of covariates, $X_1,\cdots,X_5$ from which the columns of X are to be selected, then there are $2^{p_*-1} = 16$ candidate models in the candidate pool. Here, we will illustrate the behavior of model selection criteria by choosing three generating models:

1) Model 1: $y_{ij} = \beta_o + \beta_3 X_{ij3} + b_i + \varepsilon_{ij}$, $\beta_o = 2, \beta_3 = -3$;
2) Model 2: $y_{ij} = \beta_o + \beta_3 X_{ij3} + \beta_4 X_{ij4} + b_i + \varepsilon_{ij}$, $\beta_o = 2, \beta_3 = -3, \beta_4 = 4$;
3) Model 3: $y_{ij} = \beta_o + \beta_1 X_{ij1} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + b_i + \varepsilon_{ij}$, $\beta_o = 2, \beta_2 = 2, \beta_2 = -3, \beta_4 = 4$.

These three models correspond to the three βs: $(2,0,0,-3,0)'$, $(2,0,0,-3,4)'$ and $(2,0,2,-3,4)'$ in model (4.1) with the number of fixed effects p_o equals 2, 3, 4, respectively. Again, the MLEs are used for estimation in the simulations.

Furthermore, we consider the case where the correlated errors have varying degrees of exchangeable structure. The variance component of error term σ^2 is taken to be 1, and four values in an increasing order of τ^2 are considered: 3, 6, 9, corresponding to three values of ϕ: 3, 6, 9, respectively. We take the number of clusters (*m*) to be 5, 10 and 20, the number of repetitions in a cluster to be fixed at *n* = 5. We employ a total of 100 realizations for each model.

4.2. Results

4.2.1. Model 1: $\beta=(2,0,0,-3,0)'$

Table 1 presents the performance of the two versions of marginal C_p (MC_p and IMC_p), mAIC and mBIC, under model 1 with the true fixed effects parameter $\beta=(2,0,0,-3,0)'$, and corresponding to p_o = 2. The correct model selection rate for each criterion is listed. We observe that corresponding to each ϕ, the IMC_p outperforms the MC_p, and both outperform mAIC and mBIC in selecting the correct model for small samples. With the increasing of the ratio ϕ, we can observe the better performance in selecting the correct model from our proposed criteria.

4.2.2. Model 2: $\beta=(2,0,0,-3,4)'$

We evaluate the proposed criteria for model 2 in the same manner as for model 1. **Table 2** presents the performance of MC_p and IMC_p, mAIC and mBIC under model 2, where the true fixed effects parameter is $\beta=(2,0,0,-3,4)'$ and p_o = 3. The only change on model 2 from model 1 is that we add one more fixed effect variable X_5 and set the coefficient of that variable $\beta_5=4$. In **Table 2**, the simulation results of model 2 are similar to those of model 1. With the increasing of the ratio ϕ, we can have the better performance from our proposed criteria MCp and IMCp, indicating that the proposed MCp and IMCp can effectively fulfill the mission of model selection in the mixed models. We can also observe and conclude that IMC_p has improved the performance of MC_p for model selection in small samples. With the increasing of m, the performance of IMC_p and MC_p becomes closer. Comparing to the correct selection rates in model 1, all model selection criteria behave better in model 2.

4.2.3. Model 3: $\beta=(2,0,2,-3,4)'$

As in the first two models, we evaluate the performance of model selection criteria by the rates in correctly selecting the true model. The results are presented in **Table 3**. Model 3 is identical to model 2 with the exception that we add one more significant fixed effect variable X_2 with the coefficient $\beta_2=2$.

The simulation results of model 3 are similar to those of models 1 - 2. Considering the rates in choosing the correct model, we can find the trend of dramatic improvement of all criteria on model 3 over those on models 1 and 2, implying that the proposed MC_p and IMC_p essentially and effectively implement model selection when the fixed-effects are significant. In moderately large (m = 20) sample sizes, compared to that of mAIC and mBIC, MC_p and IMC_p have comparative performance in selecting the correct model.

Table 1. Correct selection rate in model 1.

Sample size	Criterion	correlation parameter		
		ϕ = 3	ϕ = 6	ϕ = 9
m = 5	MC_p	0.78	0.86	0.85
	IMC_p	0.85	0.92	0.88
	mAIC	0.55	0.48	0.53
	mBIC	0.81	0.75	0.69
m = 10	MC_p	0.76	0.88	0.89
	IMC_p	0.77	0.89	0.90
	mAIC	0.62	0.52	0.53
	mBIC	0.86	0.82	0.80
m = 20	MC_p	0.81	0.88	0.93
	IMC_p	0.82	0.89	0.93
	mAIC	0.57	0.61	0.59
	mBIC	0.86	0.93	0.91

Table 2. Correct selection rate in model 2.

Sample size	Criterion	Correlation parameter		
		$\phi = 3$	$\phi = 6$	$\phi = 9$
$m = 5$	MC_p	0.81	0.90	0.93
	IMC_p	0.82	0.92	0.93
	mAIC	0.63	0.66	0.62
	mBIC	0.76	0.83	0.74
$m = 10$	MC_p	0.81	0.88	0.94
	IMC_p	0.83	0.88	0.94
	mAIC	0.62	0.65	0.69
	mBIC	0.85	0.85	0.83
$m = 20$	MC_p	0.87	0.94	0.91
	IMC_p	0.88	0.94	0.91
	mAIC	0.74	0.70	0.63
	mBIC	0.92	0.93	0.88

Table 3. Correct selection rate in model 3.

Sample size	Criterion	Correlation parameter		
		$\phi = 3$	$\phi = 6$	$\phi = 9$
$m = 5$	MC_p	0.92	0.87	0.92
	IMC_p	0.93	0.89	0.92
	mAIC	0.81	0.72	0.77
	mBIC	0.93	0.85	0.87
$m = 10$	MC_p	0.93	0.93	0.96
	IMC_p	0.93	0.96	0.96
	mAIC	0.83	0.83	0.87
	mBIC	0.93	0.94	0.94
$m = 20$	MC_p	0.92	0.96	0.97
	IMC_p	0.93	0.96	0.98
	mAIC	0.84	0.85	0.77
	mBIC	0.97	0.96	0.96

5. Concluding Remarks

The simulation results illustrate that the proposed criteria MC_p and IMC_p outperform mAIC and mBIC when the observations are highly correlated in small samples. The results also show that with the increasing of the ratio ϕ between the variance for the random effects and that for errors, the MC_p and IMC_p perform better. Since a larger ϕ implies a higher correlation between the observations, we can conclude that with the correlation between observations increases, a better performance from the proposed criteria MC_p and IMC_p would be observed. Since the model with a small ϕ which close to 0 is similar to a linear regression model with independent errors, our proposed criteria are not advantageous to be applied in such case.

The simulation results show that the proposed criteria MC_p and IMC_p significantly outperform mAIC and mBIC when the sample size is small. As the sample size increases, the performance of the proposed criteria becomes comparable to that of mAIC and mBIC. Therefore, MC_p and IMC_p are highly recommended in small samples in the setting of linear mixed models.

Our research (not shown in this paper) also shows that both proposed criteria behave best when the maximum likelihood estimation (MLE) is employed, comparing to those when the restricted maximum likelihood estimation or least squares estimation are used. The research on MC_p and IMC_p under REML estimation needs to be further developed in the future.

In the simulation study, by the comparison among models 1, 2 and 3, we see that when the true model includes more significant fixed effect covariates, the proposed criteria perform better in selecting the correct model. This fact indicates that the models with more significant variables (larger βs) are more identifiable by the proposed criteria than the models with variables which are not quite significant.

Comparing the performance between MC_p and IMC_p, we find that when the sample size is small, IMC_p obtains a higher correct selection rate than MC_p, which demonstrates that IMC_p improves the performance of MC_p in selecting the most appropriate model. However, when the sample size becomes larger, the performance of MC_p and IMC_p is quite identical.

Regarding the consistency of a model selection criterion, it means that as the sample size increases, the model selection will select the true model with probability 1. Note that MC_p, IMC_p, and mAIC are not consistent, whereas mBIC is consistent as expected since its penalty term $\log(N)$ prevents the overfitting in large samples. As the simulation study demonstrates, we can address again that the proposed criteria MC_p and IMC_p validate their advantages in small samples, although they are originally justified with large sample approximations, which is similar to quite a few other model selection criteria. The details for the consistency of model selection criteria in linear mixed models can also see Jiang and Rao [20].

References

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

[2] Mallows, C.L. (1995) More Comments on C_p. *Technometrics*, **37**, 362-372.

[3] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., *International Symposium on Information Theory*, 267-281.

[4] Akaike, H. (1974) A New Look at the Model Selection Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

[5] Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464. http://dx.doi.org/10.1214/aos/1176344136

[6] Sugiura, N. (1978) Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Communications in Statistics—Theory and Methods A*, **7**, 13-26. http://dx.doi.org/10.1080/03610927808827599

[7] Shang, J. and Cavanaugh, J.E. (2008) Bootstrap Variants of the Akaike Information Criterion for Mixed Model Selection. *Computational Statistics & Data Analysis*, **52**, 2004-2021. http://dx.doi.org/10.1016/j.csda.2007.06.019

[8] Azari, R., Li, L. and Tsai, C. (2006) Longitudinal Data Model Selection. Applied Times Series Analysis, Academic Press, New York, 1-23. http://dx.doi.org/10.1016/j.csda.2005.05.009

[9] Vaida, F. and Blanchard, S. (2005) Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, **92**, 351-370. http://dx.doi.org/10.1093/biomet/92.2.351

[10] Henderson, C.R. (1950) Estimation of Genetic Parameters. *Annals of Mathematical Statistics*, **21**, 309-310.

[11] Harville, D.A. (1990) BLUP (Best Linear Unbiased Prediction) and beyond. In: Gianola, D. and Hammond, K., Eds., *Advances in Staitstical Methods for Genetic Improvement of Livestock*, Springer, New York, 239-276. http://dx.doi.org/10.1007/978-3-642-74487-7_12

[12] Robinson, G.K. (1991) That BLUP Is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**, 15-32. http://dx.doi.org/10.1214/ss/1177011926

[13] Dimova, R.B., Mariantihi, M. and Talal, A.H. (2011) Information Methods for Model Selection in Linear Mixed Effects Models with Application to HCV Data. *Computational Statistics & Data Analysis*, **55**, 2677-2697. http://dx.doi.org/10.1016/j.csda.2010.10.031

[14] Müller, S., Scealy, J.L. and Welsh, A.H. (2013) Model Selection in Linear Mixed Models. *Statistical Science*, **28**, 135-167. http://dx.doi.org/10.1214/12-STS410

[15] Jones, R.H. (2011) Bayesian Information Criterion for Longitudinal and Clustered Data. *Statistics in Medicine*, **30**, 3050-3056. http://dx.doi.org/10.1002/sim.4323

[16] Fujikoshi, Y. and Satoh, K. (1997) Modified AIC and C_p in Multivariate Linear Regression. *Biometrika*, **84**, 707-716. http://dx.doi.org/10.1093/biomet/84.3.707

[17] Davies, S.L., Neath, A.A. and Cavanaugh, J.E. (2006) Estimation Optimality of Corrected AIC and Modified C_p in Linear Regression. *International Statistical Review*, **74**, 161-168. http://dx.doi.org/10.1111/j.1751-5823.2006.tb00167.x

[18] Cavanaugh, J., Neath, A.A. and Davies, S.L. (2010) An Alternate Version of the Conceptual Predictive Statistic Based on a Symmetrized Discrepancy Measure. *Journal of Statistical Planning and Inference*, **140**, 3389-3398. http://dx.doi.org/10.1016/j.jspi.2010.05.002

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

[20] Jiang, J. and Rao, J.S. (2003) Consistent Procedures for Mixed Linear Model Selection. *Sankhya*, **65**, 23-42.

Appendix

Proof of Theorem 1. 1) To prove that $\hat{H}$ is idempotent, we calculate

$$\hat{H}\hat{H} = X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1} = X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1} = \hat{H}.$$

Thus, we prove that $\hat{H}$ is idempotent.

2) By the properties of trace, we have

$$tr\left(\hat{H}\right) = tr\left(X\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}\right) = tr\left(\left(X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-1}X\right) = tr\left(I_p\right) = p.$$

Therefore, we have

$$tr\left(I_N - \hat{H}\right) = tr\left(I_N\right) - tr\left(\hat{H}\right) = N - p.$$

Thus, Theorem 1 is proved. □

Proof of Theorem 2. Let $y \in C\left(K^{\mathrm{T}}X\right)$. We need to show that $y \in C\left(K^{\mathrm{T}}X_*\right)$.

Since $y \in C\left(K^{\mathrm{T}}X\right)$, there exists a $p \times 1$ vector β_1 such that $y = K^{\mathrm{T}}X\beta_1$.

By $C(X) \subseteq C(X_*)$, there also exists a $p \times 1$ vector β_2 such that $X\beta_1 = X_*\beta_2$, which makes $y = K^{\mathrm{T}}X\beta_1 = K^{\mathrm{T}}X_*\beta_2$.

So we have $y \in C\left(K^{\mathrm{T}}X_*\right)$. □

Proof of Corollary 2. 1) Since Σ_o is positive definite, there exists an $N \times N$ matrix V with $rank(V) = N$, such that $\Sigma_o = VV^{\mathrm{T}}$. It follows that $\Sigma_o^{-1} = \left(VV^{\mathrm{T}}\right)^{-1} = \left(V^{\mathrm{T}}\right)^{-1}(V)^{-1} = \left(V^{-1}\right)^{\mathrm{T}}(V)^{-1}$.

Let $K = \left(V^{-1}\right)^{\mathrm{T}}$, we can have $\Sigma_o^{-1} = KK^{\mathrm{T}}$. Then, we arrive at

$$\begin{aligned}\Sigma_o^{-1}HH_* &= \Sigma_o^{-1}X\left(X^{\mathrm{T}}\Sigma_o^{-1}X\right)X^{\mathrm{T}}\Sigma_o^{-1}X_*\left(X_*^{\mathrm{T}}\Sigma_o^{-1}X_*\right)X_*^{\mathrm{T}}\Sigma_o^{-1}\\ &= KK^{\mathrm{T}}X\left(X^{\mathrm{T}}KK^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}KK^{\mathrm{T}}X_*\left(X_*^{\mathrm{T}}KK^{\mathrm{T}}X_*\right)^{-1}X_*^{\mathrm{T}}KK^{\mathrm{T}}\\ &= K\left(K^{\mathrm{T}}X\right)\left(\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X\right)\right)^{-1}\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\left(\left(K^{\mathrm{T}}X_*\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\right)^{-1}\left(K^{\mathrm{T}}X_*\right)K^{\mathrm{T}}.\end{aligned}$$

Now, let

$$\bar{H} = \left(K^{\mathrm{T}}X\right)\left(\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X\right)\right)^{-1}\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}$$

and

$$\bar{H}_* = \left(K^{\mathrm{T}}X_*\right)\left(\left(K^{\mathrm{T}}X_*\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\right)^{-1}\left(K^{\mathrm{T}}X_*\right).$$

Since $C(X) \subseteq C(X_*)$, by Theorem 2, we have $C\left(K^{\mathrm{T}}X\right) \subseteq C\left(K^{\mathrm{T}}X_*\right)$, so that we can have $\bar{H}\bar{H}^* = \bar{H}$, which leads to

$$\begin{aligned}\Sigma_o^{-1}HH_* &= K\left(K^{\mathrm{T}}X\right)\left(\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X\right)\right)^{-1}\left(K^{\mathrm{T}}X\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\left(\left(K^{\mathrm{T}}X_*\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\right)^{-1}\left(K^{\mathrm{T}}X_*\right)K^{\mathrm{T}}\\ &= K\bar{H}\bar{H}_*K^{\mathrm{T}} = K\bar{H}K^{\mathrm{T}} = KK^{\mathrm{T}}X\left(X^{\mathrm{T}}KK^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}KK^{\mathrm{T}}\\ &= \Sigma_o^{-1}X\left(X^{\mathrm{T}}\Sigma_o^{-1}X\right)^{-1}X^{\mathrm{T}}\Sigma_o^{-1} = \Sigma_o^{-1}H.\end{aligned}$$

The first part of Corollary 2 is therefore proved.

2) Following the first part proof of Corollary 2, since $C\left(K^{\mathrm{T}}X\right) \subseteq C\left(K^{\mathrm{T}}X_*\right)$, we have $H_1^*\left(K^{\mathrm{T}}X\right) = K^{\mathrm{T}}X$. Then, we can conclude that

$$\begin{aligned}\Sigma_o^{-1}H_*X &= K\left(\left(K^{\mathrm{T}}X_*\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X_*\right)\right)^{-1}\left(K^{\mathrm{T}}X_*\right)^{\mathrm{T}}\left(K^{\mathrm{T}}X\right)\\ &= KH_1^*\left(K^{\mathrm{T}}X\right) = KK^{\mathrm{T}}X = \Sigma_o^{-1}X.\end{aligned}$$

Therefore, the proof for the second part of Corollary 2 is completed. □