Communications for Statistical Applications and Methods
2015, Vol. 22, No. 5, 507–518

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.5.507
Print ISSN 2287-7843 / Online ISSN 2383-4757

A Note on Performance of Conditional Akaike Information Criteria in Linear Mixed Models

Yonghee Lee[1,a]

[a]Department of Statistics, University of Seoul, Korea

Abstract

It is not easy to select a linear mixed model since the main interest for model building could be different and the number of parameters in the model could not be clearly defined. In this paper, performance of conditional Akaike Information Criteria and its bias-corrected version are compared with marginal Bayesian and Akaike Information Criteria through a simulation study. The results from the simulation study indicate that bias-corrected conditional Akaike Information Criteria shows promising performance when candidate models exclude large models containing the true model, but bias-corrected one prefers over-parametrized models more intensively when a set of candidate models increases. Marginal Bayesian and Akaike Information Criteria also have some difficulty to select the true model when the design for random effects is nested.

Keywords: linear mixed models, variance components, selection, Akaike Information Criteria, Bayesian Information Criteria, conditional distribution, maximum likelihood estimation

1. Introduction

Linear mixed models are useful statistical models in which various dependency among responses can be incorporated using the framework of linear models. Dependency or correlation among observations can be implemented by introducing random effects and observations with the same random effect that are correlated so that linear mixed models can analyze data in which various levels of dependency occur. For example, in a longitudinal study, responses from the same subject are supposed to be correlated for genetics traits of siblings from the same parent that have strong dependency. Hence, linear mixed models are widely used for data analysis in many research fields such as medicine, epidemiology, psychometrics and genetics.

If a researcher is interested in a specification of the regression mean of responses rather than covariance structure, linear mixed models can be considered as usual general linear models in which covariance of responses can be modeled as a positive definite matrix with a small number of nuisance parameters. Traditional methods to select regression variables in general linear models can be used without modification. For example, Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) could be used for model selection in linear mixed models. Linear mixed models can give specific structure of dependency among responses; therefore, different specification of random effects leads to different covariance structure of observations. It is important to select random effects to determine covariance structure as well as select fixed effects to specify mean structure of responses. Linear

[1] Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 130-743, Korea.
E-mail: ylee@uos.ac.kr

Published 30 September 2015 / journal homepage: http://csam.or.kr

mixed models specify covariance structure of responses and prediction of a response within a group or level which is explained by the same random effect. In such cases, it is important to select a model for random effects as much as selecting fixed effects. This creates difficulty in developing statistical methods to select fixed and random effects simultaneously, since the objective of data analysis using linear mixed models could have many folds.

Recently, modifications of traditional model selection methods using information criteria are developed to focus on different objectives of linear mixed models in analyzing complex data. Conditional Akaike Information Criteria (cAIC) is a major modification and its objective is to provide model selection methods for cluster level focus. The prediction at the group level is conditional on the corresponding random effect; therefore, cAIC is using conditional likelihood with predicted random effects instead of marginal likelihood. Also cAIC includes random effects when enumerating the number of parameters with certain weights as well as the number of unknown fixed parameters. Vaida and Blanchard (2005) proposed the conditional model selection formation in terms of cAIC with effective degrees of freedom that account for certain shrinkage weight in the random effects under assumption of known variance parameters for random effects. Liang *et al.* (2008) proposed a bias-corrected cAIC that accounts for uncertainty in the estimation of the variance parameters, but its implementation could be computationally intensive since it uses numerical differentiation. Greven and Kneib (2010) developed and explicit formulation of bias-corrected cAIC by Liang *et al.* (2008) so that the corrected cAIC good be obtained without numerical differentiation.

Theoretical properties of cAIC and corrected versions are derived by Greven and Kneib (2010), but it is limited since it does not show properties for consistency and efficiency in the selection procedures. Many researchers investigate performance of model selection methods for linear mixed models by using simulation studies since it is complicated to derive theoretical properties of cAIC in linear mixed models. But, as indicated in Müller *et al.* (2013), simulation studies are not comprehensive since most of them use very simple models such as random intercept model. The most comprehensive simulation studies are provided by Dimova *et al.* (2011), but their studies do not include bias-corrected cAIC and its focus is the selection of fixed effects under simple structure for random effects. It is important to investigate the performance of bias-corrected cAIC in comparison with other criteria especially when structure for random effects is complicated and the main interest is selecting random effects rather than fixed effects.

In this paper, performance of bias-corrected cAIC is investigated under the nested design for random effects as well as multivariate random effect for longitudinal data in comparison with other selection methods based on information criteria. The nested design and longitudinal data are common in real data analysis and the selection of the structure of random effects is crucial in fields such as genetics and paediatrics. Hence, the focus on model selection in this paper is to select the structure of random effects in linear mixed models rather than selecting fixed effects. There are two methods for estimation of variance components, one is ordinary maximum likelihood estimation and another is restricted (or residual) maximum likelihood estimation. In this paper, I consider only maximum likelihood estimation for simplicity of presentation since the performance between two methods in model selection is similar.

In Sections 2 and 3 brief review for linear mixed models and model selection methods based on information criteria are provided respectively and Section 4 provides simulation studies. Finally Section 5 includes some comments on simulation studies and the conclusion. Programs for computation for all methods in this paper are written by `JULIA` in `http://julialang.org/` and all programs can be found in `https://github.com/iloveall data/SelectMixedModels`.

2. Linear Mixed Models

Linear mixed models is defined by

$$y = X\beta + Zb + \epsilon, \tag{2.1}$$

where y is $n \times 1$ response vector, X is $n \times p$ design matrix for fixed effect with parameter β which is $p \times 1$ vector. The known matrix Z is $n \times q$ design matrix for $q \times 1$ vector b for unknown random effects. Sometimes Z and b can be partitioned into $Z = \text{diag}(Z_1, Z_2, \ldots, Z_m)$ and $b = (b_1^t, b_2^t, \ldots, b_m^t)$ according to grouping levels. Also $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^t$ is $n \times 1$ vector for random errors.

The distribution of random error ϵ_i follows normal distribution with mean 0 and variance $\sigma^2 > 0$ and all random errors are assumed independent. The distribution of random effects follows multivariate normal distribution $N(0, \sigma^2\Sigma)$ where relative covariance matrix Σ is determined by unknown parameters for variance components $\tau = (\tau_1, \tau_2, \ldots, \tau_r)^t$ and relative covariance matrix $\Sigma = \Sigma(\tau)$ is assumed to be non-negative definite since a variance component is allowed to be zero (i.e., $\tau_j = 0$ for some j's). I will define parameter vector θ by $\theta = (\beta^t, \tau^t, \sigma^2)^t$ which includes all unknown parameters in linear mixed model (2.1). Then, mean and covariance V of response vector y is defined by

$$E(y) = X\beta \quad \text{and} \quad \text{Var}(y) = \sigma^2\left(I + Z\Sigma Z^t\right) := V = \sigma^2 V_*,$$

where $V_* = (I + Z\Sigma Z^t)$ is relative covariance of y to the error variance σ^2.

Now the marginal log likelihood function $\ell(\theta|y)$ for linear mixed models (2.1) is given by

$$-2\ell(\theta|y) = \log|V| + (y - X\beta)^t V^{-1}(y - X\beta) = n\log\sigma^2 + \log|V_*| + \frac{1}{\sigma^2}(y - X\beta)^t V_*^{-1}(y - X\beta). \tag{2.2}$$

The estimator of β and σ^2 based on maximum likelihood estimation when V_* is assumed to be known are given by

$$\hat\beta = \left(X^t V_*^{-1} X\right)^{-1} X^t V_*^{-1} y \quad \text{and} \quad \hat\sigma^2 = \frac{1}{n}\left(y - X\hat\beta\right)^t V_*^{-1}\left(y - X\hat\beta\right). \tag{2.3}$$

In practice, relative covariance V_* is replaced with estimated one $\hat V_*$. If the maximum likelihood estimators are plugged in the marginal log likelihood function in (2.2), it leads to the profile likelihood function $\ell(\tau|y, \hat\beta, \hat\sigma^2)$ for variance components τ and it is given by, up to constant,

$$-2\ell\left(\tau|y, \hat\beta, \hat\sigma^2\right) = \log|V_*| + n\log\left(y - X\hat\beta\right)^t V_*^{-1}\left(y - X\hat\beta\right). \tag{2.4}$$

The variance components τ can be estimated by maximizing the profile likelihood function (2.4) in terms of parameter τ and it is commonly done using an algorithm for nonlinear optimization.

The random effect b is an unobservable quantity; however, its prediction is commonly main interest. The best linear unbiased prediction (BLUP) under known V_* is frequently used to predict random effects in the model (2.1) and it is given by

$$\hat b = \Sigma Z V_*^{-1}\left(y - X\hat\beta\right) \tag{2.5}$$

and unknown parameters in V_* (so Σ) are replaced by estimated ones in practice.

Commonly, if interest of research is prediction of new response from the same level or group specified in the model, prediction could be based on the linear model conditional on random effects such that

$$y|b \sim N\left(X\beta + Zb, \sigma^2 I\right)$$

which leads to the conditional distribution y given b and it is defined by

$$-2\log f(y|b) = n\log\sigma^2 + \frac{1}{\sigma^2}(y - X\beta - Zb)^t(y - X\beta - Zb). \tag{2.6}$$

This conditional distribution $f(y|b)$ has an important role in defining cAIC which is proposed for prediction of new observation when new observation is assumed to share the same random effect specified in the given model. All details in theory and its application of linear mixed models can be found in Searle *et al.* (1992) and McCulloch and Searle (2001).

3. Review of Model Selection Methods based on Akaike Information Criteria

Model selection methods for linear models based on Information Criteria are well-developed and most frequently used methods are BIC and AIC. AIC for given parametric statistical model $f(y|\theta)$ is defined as

$$\text{AIC} = -2\ell\left(\hat{\theta}\middle|y\right) + 2\rho,$$

where $\ell(\theta|y)$ is the log likelihood function based on the distribution $f(y|\theta)$ and ρ is number of unknown parameters θ in the model specified by $f(y|\theta)$. Also BIC is defined by

$$\text{BIC} = -2\ell\left(\hat{\theta}\middle|y\right) + (\log n)\rho.$$

Main difference between AIC and BIC is the leading constant in the penalty term, which is 2 and $\log n$ respectively. The penalty term in BIC increases as number of responses n increases; therefore, performance of BIC is different from AIC and BIC is known to be consistent in sense that BIC prefer to sparse model among candidate models which includes true model. However, AIC is efficient and not consistent in the sense that a selected model by AIC gives the most less predictive mean squared error among candidate models.

The definition for BIC and AIC for linear model can be extended to linear mixed models in (2.1) easily. When BIC and AIC are defined for linear mixed models, the marginal likelihood function in (2.2) should be used and the number of unknown parameters in $\theta = (\beta^t, \tau^t, \sigma^2)^t$ is $p + r + 1$. Hence, marginal BIC (mBIC) and marginal AIC (mAIC) for linear mixed models can be defined by

$$\text{mAIC} = -2\ell\left(\hat{\theta}\middle|y\right) + 2(p + r + 1) \quad \text{and} \quad \text{mBIC} = -2\ell\left(\hat{\theta}\middle|y\right) + (\log n)(p + r + 1). \tag{3.1}$$

Note that the term 'marginal' is used since mBIC and mAIC are based on the marginal likelihood instead of conditional distribution given random effects.

Vaida and Blanchard (2005) proposed the conditional model selection formation based on the conditional distribution of responses given by random effects as defined in (2.6). The proposed methods commonly called as cAIC and the number of parameters in penalty terms is implemented as the effective degree of freedom accounting for certain shrinkage weight in the random effect under the assumption of known variance parameters for random effects. Vaida and Blanchard (2005) proposed conditional Akaike Information such that

$$\text{cAI} = -2E_{g(y,u)}E_{g(y^*|u)}\log f\left(y^*\middle|\hat{\theta}(y), \hat{b}(y)\right), \tag{3.2}$$

where $g(y, u)$ is the true joint distribution of responses y and random effects u. $f(y|\theta, b)$ is the conditional distribution of y given b with unknown parameter θ defined in (2.6). Note that conditional

Akaike Information (3.2) is different from the marginal Akaike Information which should be estimated by marginal AIC in (3.1). The marginal Akaike Information is defined by

$$\text{mAI} = -2E_{g(y)}E_{g(y^*)} \log f\left(y^* \middle| \hat{\theta}(y)\right)$$

and its estimator mAIC in (3.1) is based on the marginal likelihood in (2.2).

Assuming the error variance σ^2 and variance components τ are known, the cAIC is derived as

$$\text{cAIC} = -2\log f\left(y^* \middle| \hat{\theta}(y), \hat{b}(y)\right) + 2\rho, \tag{3.3}$$

where ρ is the effective degrees of freedom whose value lies between number of fixed effects β and the total number of fixed effects and random effects. The value of ρ is defined by

$$\rho = \text{trace}(H_1) = \text{tr}\left\{\begin{pmatrix} X & Z \end{pmatrix}\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + \Sigma^{-1}\end{pmatrix}^{-1}\begin{pmatrix} X^t \\ Z^t\end{pmatrix}\right\} = \text{tr}\left\{\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ + \Sigma^{-1}\end{pmatrix}^{-1}\begin{pmatrix} X^tX & X^tZ \\ Z^tX & Z^tZ\end{pmatrix}\right\}, \tag{3.4}$$

where H_1 is the pseudo projection matrix for $\hat{y} = X\hat{\beta} + Z\hat{b} = H_1y$.

The formula for ρ in (3.4) is developed under the assumption that the error variance and variance components are known; therefore, several adjustments have been developed in cases of unknown error variance and variance components. Vaida and Blanchard (2005) propose adjustment for ρ when the error variance σ^2 is unknown but variance components τ are known and it gives

$$\text{cAIC} = -2\log f\left(y^* \middle| \hat{\theta}(y), \hat{b}(y)\right) + 2K, \tag{3.5}$$

where K is given by

$$K = \frac{n(n-p-1)}{(n-p)(n-p-2)}(\rho + 1) + \frac{n(p+1)}{(n-p)(n-p-2)}$$

and p is number of column in X, the number of fixed effects. Further, when variance components τ are unknown, Vaida and Blanchard (2005) proposed that ρ can be replaced with estimated version $\hat{\rho} = \rho(\hat{\Sigma})$.

Liang *et al.* (2008) provided an alternative presentation for cAIC by using Stein's formula (Stein, 1981) and when the error variance σ^2 is known but variance components τ are unknown, cAIC is derived as

$$\text{cAIC} = -2\log f\left(y^* \middle| \hat{\theta}(y), \hat{b}(y)\right) + 2\Phi_0, \quad \text{where } \Phi_0 = \sum_{i=1}^{n}\frac{\partial \hat{y}_i}{\partial y_i} = \text{trace}\left(\frac{\partial \hat{y}^t}{\partial y}\right). \tag{3.6}$$

Note that when both the error variance and variance components are known, $\Phi_0 = \rho$ in (3.4). Liang *et al.* (2008) also showed that Φ_0 can be calculated in terms of ρ and H_1 in (3.4) such that

$$\Phi_0 = \hat{\rho} + 1^tD\left(\hat{\Sigma}\right)y, \tag{3.7}$$

where $\hat{\rho} = \rho(\hat{\Sigma})$, 1 is the vector of ones and $D(\hat{\Sigma}) = \{d_{ij}\}$ is $n \times n$ matrix with its $(i, j)^{th}$ element d_{ij} is defined by partial derivative of $\hat{H}_1 = H_1(\hat{\Sigma}) = \{h_{ij}\}$ in (3.4) with respect to y such that

$$d_{ij} = \frac{\partial h_{ij}}{\partial y_i}.$$

Even though their forms are simple, the penalty terms in (3.6) and (3.7) require numerical differentiation at least n times so that it is not practical when number of responses n becomes large. Note that Liang *et al.* (2008) proposed cAIC when both the error variance and variance components are known, but its evaluation requires a second order numerical differentiation so that it is not easy to calculate.

Greven and Kneib (2010) give explicit formula for penalty term of cAIC in (3.6) without differentiation and its form is given by

$$\Phi_0 = \hat{\rho} + \sum_{i=1}^{s} e_i^t \hat{B}^{-1} \hat{G} \hat{A} \hat{W}_i \hat{A} y, \tag{3.8}$$

where $s \leq r$ is the number of non-zero estimates in variance components $\tau = (\tau_1, \tau_2, \ldots, \tau_r)^t$, $W_i = \partial V_* / \partial \tau_i$, $A = V_*^{-1} - V_*^{-1} X (X^t V_*^{-1} X)^{-1} X^t V_*^{-1}$, and G is $s \times n$ matrix whose ith row is given by $2\{(y^tAy)y^tAW_iA - (y^tAW_iAy)y^tA\}$ and $B = \{b_{ij}\}$ is $s \times s$ the negative definite matrix with its $(i, j)^{th}$ element b_{ij} is defined by

$$b_{ij} = \left(y^tAy\right)^2 \frac{\text{trace}\left(U_{ij}V^{-1} - W_iV^{-1}W_jV^{-1}\right)}{n} - \left(y^tAW_iAy\right)\left(y^tAW_jAy\right) - \left[y^t\left(AU_{ij}A - 2AW_iAW_jA\right)y\right]\left(y^tAy\right),$$

where $U_{ij} = \partial^2 V / \partial \tau_i \partial \tau_j$. Note that the nonzero estimates $(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_s)$ are plugged in B, G, A, W_i, U_{ij} for evaluation of (3.8).

Even though its formulation in (3.8) looks complicated, the derivatives W_i and U_{ij} can be easily calculated in practice since most common form for $Z^t \Sigma Z$ in V_* is diagonal such that $Z^t \Sigma Z = \sum_{k=1}^{r} \tau_k Z_k^t \Sigma_k Z_k$ where Σ_k's are known and $U_{ij} = 0$ for this case.

4. Simulation Study

4.1. Design for models and parameters

A simulation study is considered to evaluate the performance various methods based on Akaike Information Criteria. The true linear mixed model for simulation study has the following model formulation:

$$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij}) t_{ijk} + \epsilon_{ijk}, \tag{4.1}$$

for $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$. In the true model (4.1), a_i is the random effect for the first group level and it is assumed that $a_i \sim N(0, \sigma_a^2)$. The second group level, which is nested within the first level i, consists of a linear trend with fixed effects $\beta_0 + \beta_1 t$ for population level and the bivariate random effects $(b_{0ij}, b_{1ij})^t$ for individual effects.The bivariate random effects follow bivariate normal distribution such that

$$\begin{pmatrix} b_{0ij} \\ b_{1ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b0}^2 & \phi\sigma_{b0}\sigma_{b1} \\ \phi\sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 \end{pmatrix} \right), \tag{4.2}$$

where ϕ is correlation between b_{0ij} and b_{1ij}. Finally independent random errors ϵ_{ijk}'s follow the normal distribution $N(0, \sigma^2)$ and it is assumed that all a_i's, $(b_{0ij}, b_{1ij})^t$'s, ϵ_{ijk}'s are independent.

In terms of parameter specification for linear mixed models (2.1), we can consider $\beta = (\beta_0, \beta_1)^t$ and $(\tau_1, \tau_2, \tau_3, \tau_4) = (\sigma_a^2, \sigma_{b0}^2, \sigma_{b1}^2, \phi\sigma_{b0}\sigma_{b1})$. Note that parameter space for variance components τ_i is $[0, \infty)$, which includes boundary point 0, for $i = 1, 2, 3$ and the variance component τ_4 leads to zero when either τ_2 or τ_3 is estimated by zero.

Table 1: Specification of parameters in variance components in the true model (4.1) for simulation

Case	σ_a	$(\sigma_{b0}, \sigma_{b1})$	ϕ
1	0.5	(1.0, 1.0)	0.5
2	2.0	(1.0, 1.0)	0.5
3	4.0	(1.0, 1.0)	0.5
4	1.0	(2.0, 0.5)	0.5
5	1.0	(4.0, 0.25)	0.5
6	1.0	(4.0, 0.125)	0.5
7	1.0	(0.5, 0.5)	0.5
8	1.0	(2.0, 2.0)	0.5
9	1.0	(4.0, 4.0)	0.5
10	1.0	(1.0, 1.0)	0.2
11	1.0	(1.0, 1.0)	0.5
12	1.0	(1.0, 1.0)	0.8

For example, in the model (4.1), the first level could be a school whose effect is a_i and the second level could be a student within the school. Then, a student has responses along several time points and responses in time follows a simple linear trend. Hence the fixed effects (β_0,β_1) represent population linear model and the random effects (b_{0ij}, b_{1ij}) represent a student's individual effect on the linear model.

For simulation study, the fixed parameters (β_0,β) = $(3, 2)$ and the error variance $\sigma = 1.0$ for ϵ_{ijk} are fixed for all cases. Then, 12 combinations of parameters for variance components are considered (Table 1). The first three cases in Table 1 investigate the effects of variance size for the first level and the next 6 cases evaluate the effects of different variance components for the second level. The last three cases are considered to see the effects of different correlations for bivariate random effects in the second level. The number of parameters r for unknown variance components in Table 1 is the dimension of $\tau = (\tau_1, \tau_2, \ldots, \tau_r)^t$ which specifies covariance $\Sigma = \Sigma(\tau)$ of random effects b. Some of the maximum likelihood estimates for τ could be zero and the number of positive estimates s for τ could therefore be less than or equal to the number of specified parameters in the model ($s \leq r$).

Regarding sample size in the simulation study, we consider two schemes for the model (4.1). The first scheme uses $I = 10, J = 2, K = 5$ which lead to the total number of responses is $n = 10 \times 2 \times 5 = 100$ and the second scheme uses $I = 20, J = 4, K = 5$ which lead to the total number of responses is $n = 400$. The time point t_{ijk} in the model (4.1) has five time points such that $t_{ijk} = k$ in all cases.

Settle candidate models for given data are considered in practice and the best is selected based on the applied model selection criteria. Hence, we consider ten different models in this simulation to mimic a real situation (Table 2). All models include the same fixed effects term $\beta_0 + \beta_1 t_{ijk}$ in the true model (4.1) since main interest of simulation study to evaluate the performance of the selection methods for structure of random effects. The simplest model includes just random effect a_i for the first level. Then, the complexity of structure for random effects increases by adding more random effect as well as the correlation structure among random effects. The second order quadratic term $(\beta_2 + b_{2ij})t_{ijk}^2$ is also included in some models for more complexity. The true model which generates data is at the seventh model and the most complex model including the true model (4.1) is the last model in Table 2.

We consider ten candidate models for simulation; however, the number of candidate models sequentially increased in the simulation study. First, the simplest six models (from Model 1 to 6 in Table 2) are considered candidate models in order to select the best one based on the selection method so that this design represents a situation where candidate models do not include the true model. The numbers of candidate models increase by adding one more model which is more complex sequen-

Table 2: Ten candidate models for simulation

No.	Model specification	Correlation for mulitivar. random effects	Number of parameters for variance comp. (r)
1	$y_{ijk} = a_i + \beta_0 + \beta_1 t_{ijk} + \epsilon_{ijk}$		1
2	$y_{ijk} = (\beta_0 + b_{0ij}) + \beta_1 t_{ijk} + \epsilon_{ijk}$		1
3	$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + \beta_1 t_{ijk} + \epsilon_{ijk}$		2
4	$y_{ijk} = (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij} are independent)	2
5	$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij} are independent)	3
6	$y_{ijk} = (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij} are correlated)	3
7	$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij}) + \epsilon_{ijk}$	(b_{0ij}, b_{1ij} are correlated)	4(true model)
8	$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + (\beta_2 + b_{2ij})t^2_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij}, b_{2ij} are independent)	4
9	$y_{ijk} = (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + (\beta_2 + b_{2ij})t^2_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij}, b_{2ij} are correlated)	6
10	$y_{ijk} = a_i + (\beta_0 + b_{0ij}) + (\beta_1 + b_{1ij})t_{ijk} + (\beta_2 + b_{2ij})t^2_{ijk} + \epsilon_{ijk}$	(b_{0ij}, b_{1ij}, b_{2ij} are correlated)	7

Table 3: Relative frequencies of selected models by four methods when sample size $n = 100$

Selected model	Candidate models																			
	1–6				1–7				1–8				1–9				1–10			
	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.18	0.06	0.17	0.02	0.17	0.05	0.13	0.02	0.17	0.05	0.06	0.01	0.17	0.05	0.03	0.00	0.17	0.05	0.02	0.00
5	0.09	0.11	0.66	0.22	0.03	0.02	0.28	0.04	0.03	0.02	0.11	0.02	0.03	0.02	0.05	0.00	0.03	0.02	0.04	0.00
6	0.71	0.82	0.17	0.76	0.64	0.61	0.03	0.19	0.64	0.60	0.03	0.17	0.64	0.57	0.01	0.02	0.64	0.56	0.01	0.01
7(T)	x	x	x	x	0.14	0.31	0.55	0.75	0.14	0.31	0.50	0.67	0.14	0.31	0.34	0.22	0.14	0.29	0.17	0.07
8	x	x	x	x	x	x	x	x	0.00	0.00	0.30	0.14	0.00	0.00	0.09	0.02	0.00	0.00	0.06	0.00
9	x	x	x	x	x	x	x	x	x	x	x	x	0.00	0.05	0.49	0.74	0.00	0.04	0.16	0.38
10	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.00	0.02	0.54	0.54

Candidate models correspond to numbers considered for selection listed in Table 2.
† BIC = Bayesian Information Criteria; AIC = Akaike Information Criteria; mBIC = marginal BIC; mAIC = marginal AIC; cAIC = conditional AIC; bAIC = bias-corrected conditional AIC.

tially. As soon as Model 7 in Table 2 is included in the sequential design, it considers the true model in candidates. The final design considers all 10 candidate models with the most complex model that include the true model.

4.2. Simulation and its results

Four selection methods are considered: mBIC and mAIC in (3.1), cAIC in (3.5), bAIC in (3.8) in the simulation study for to compare the performance of model selection methods for linear mixed models. Note that bias-corrected cAIC in (3.8) by Greven and Kneib (2010) is an explicit version of cAIC in (3.7) which requires numerical differentiation.

I use 1000 iterations to generate responses from the true model (4.1) for each specification for parameters in variance components listed in Table 1 under two different sample size $n = 100$ and $n = 400$. For each iteration, four selection methods are applied to the models in Table 2 with generated data; subsequently, the best one is selected based on each selection method.

To assess the overall performance of four methods, the relative frequencies of selected models under all 12 combinations of parameter settings are listed under five designs in which the number of candidate models are increasingly sequential. Tables 3 and 4 list the results of the simulation study for sample size $n = 100$ and $n = 400$, respectively.

For the design where the true model is not included (candidates Models 1–6 in Tables 3 and 4), the Model 6 (the model with correlated random effects for the second level and without the random effect for the first level) is mostly (71–87%) preferred by mBIC, mAIC and bAIC, but cAIC prefer Model 5

Table 4: Relative frequencies of selected models by four methods when sample size $n = 400$

Selected model	Candidate models																			
	1–6				1–7				1–8				1–9				1–10			
	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.07	0.02	0.14	0.01	0.06	0.02	0.12	0.01	0.06	0.02	0.06	0.00	0.06	0.01	0.03	0.00	0.06	0.01	0.03	0.00
5	0.10	0.11	0.76	0.20	0.01	0.01	0.31	0.02	0.01	0.00	0.12	0.01	0.01	0.00	0.07	0.00	0.01	0.00	0.05	0.00
6	0.83	0.87	0.10	0.79	0.69	0.51	0.01	0.13	0.69	0.51	0.01	0.12	0.69	0.48	0.01	0.01	0.69	0.48	0.01	0.00
7(T)	x	x	x	x	0.24	0.47	0.56	0.84	0.24	0.47	0.52	0.78	0.24	0.46	0.40	0.27	0.24	0.44	0.24	0.04
8	x	x	x	x	x	x	x	x	0.00	0.00	0.29	0.09	0.00	0.00	0.11	0.01	0.00	0.00	0.07	0.00
9	x	x	x	x	x	x	x	x	x	x	x	x	0.00	0.04	0.37	0.71	0.00	0.03	0.13	0.32
10	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.00	0.03	0.46	0.64

Candidate models correspond to numbers considered for selection listed in Table 2.

† BIC = Bayesian Information Criteria; AIC = Akaike Information Criteria; mBIC = marginal BIC; mAIC = marginal AIC; cAIC = conditional AIC; bAIC = bias-corrected conditional AIC.

(the model with the random effect for the first level and with independent random effects for the second level). As soon as the true model is included in candidate models (candidates Models 1–7 in Tables 3 and 4), bAIC selects the true model mostly (75%, 84%) and cAIC selects the true model about 55% and still select the model with independent random effects about 30%. However, mBIC still selects the model without random effect for the first level mostly about 65% and mAIC shows similar behavior to mBIC and selects the complex model than mBIC. After an unnecessary quadratic term t^2_{ijk} with independent random effects is included in candidate models (candidates Models 1–8 in Tables 3 and 4), performance of mBIC, mAIC, and bAIC do not change, but cAIC starts preferring the model with quadratic term to simple models. Then, after unnecessary quadratic term t^2_{ijk} with correlated random effects but without random effect for the first level is included (candidates Models 1–9 in Tables 3 and 4), both cAIC and bAIC start selecting more complex models than the true models, but mBIC and mAIC show similar behavior before including unnecessary terms. Finally the most complex model including the true model is considered (candidates Models 1–10 in Tables 3 and 4), both cAIC and bAIC start selecting the most complex models and the behavior of mBIC and mAIC does not change either.

Performance of four selection methods (shown in Tables 3 and 4) can be summarized as

- mBIC prefers sparse models more than the true model mostly and it especially fails to catch the first level effect in the nested design.
- mAIC selects the true model about 50% mostly and it also fails to catch the first level effect in the nested design.
- cAIC selects the true model less than 60% mostly but starts preferring the complex model to the true model intensively.
- Bias-corrected cAIC prefers the true model more than 70% when candidate models do not include the larger model than the true model. However, it starts preferring over-parametrized model to the true model intensively when candidate models include larger models.
- The performance of all methods to select the true model improves as sample size increases; however, the described pattern of selection does not change very much.

The most interesting behavior in this simulation study is that mBIC and mAIC fail to select models that include the random effect a_i for the first level in the true model (4.1). The simulation study in this

Table 5: Relative frequencies of selected models by four methods when only variance component for the first level is changed

Selected model	n = 100															
	σ_a = 0.5				σ_a = 1.0				σ_a = 2.0				σ_a = 4.0			
	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.05	0.01	0.07	0.00	0.22	0.05	0.04	0.01	0.19	0.02	0.01	0.00	0.05	0.01	0.00	0.00
5	0.00	0.00	0.11	0.00	0.02	0.02	0.12	0.02	0.05	0.02	0.08	0.01	0.17	0.05	0.04	0.02
6	0.94	0.89	0.08	0.31	0.70	0.67	0.02	0.14	0.35	0.22	0.00	0.04	0.02	0.01	0.00	0.00
7(T)	0.01	0.10	0.48	0.62	0.06	0.25	0.45	0.66	0.41	0.74	0.71	0.81	0.75	0.93	0.79	0.83
8	0.00	0.00	0.26	0.07	0.00	0.00	0.37	0.17	0.00	0.00	0.19	0.14	0.00	0.01	0.17	0.15
Selected model	n = 400															
	σ_a = 0.5				σ_a = 1.0				σ_a = 2.0				σ_a = 4.0			
	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.09	0.00	0.00	0.00	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.13	0.00	0.00	0.00	0.12	0.00	0.01	0.00	0.06	0.00	0.03	0.00	0.03	0.00
6	0.98	0.86	0.06	0.30	0.84	0.52	0.01	0.10	0.15	0.03	0.00	0.01	0.00	0.00	0.00	0.00
7(T)	0.02	0.14	0.47	0.68	0.16	0.48	0.57	0.85	0.82	0.97	0.77	0.92	0.97	1.00	0.85	0.92
8	0.00	0.00	0.25	0.02	0.00	0.00	0.26	0.05	0.00	0.00	0.17	0.07	0.00	0.00	0.12	0.08

Only cases 1, 2 and 3 are considered for specification of parameters in Table 1;
Candidate models are from number 1 to 8 as listed in Table 2.
[†] BIC = Bayesian Information Criteria; AIC = Akaike Information Criteria; mBIC = marginal BIC; mAIC = marginal AIC; cAIC = conditional AIC; bAIC = bias-corrected conditional AIC.

paper considers a nested design for the first time among recent research papers. For more information, the performance of selection methods is investigated for cases where the variance component σ_a changes only when others are fixed. Table 5 shows relative frequencies to select models among four methods when σ_a increases. The results in Table 5 show that mBIC and mAIC select the model without the random effect a_i for the first level more often than the true model when sample size and value of σ_a are relatively small. The performance of mBIC and mAIC is comparable to cAIC and bias-corrected cAIC when sample size and value of σ_a increases; consequently, simulation results indicate that more attention is needed when using mBIC and mAIC under nested designs.

4.3. Simulation using real data

A small simulation study using real data is also considered to verify results from simulation study with artificial data. I use data from a pharmacokinetics dataset from a Cadralazine study in Vaida and Blanchard (2005), which consists of 6 repeated measures from each 10 subjects. Original responses are plaza drug concentrations from 30mg of Cadralazine. The analyzed responses are log transformation of concentrations subtract by the log of dose amount. In this simulation study, the estimates of parameters from fitting the linear mixed effect model (no. 6 and no. 7 in Table 2) are considered true parameters. For estimation stability, the number of subjects changes from 10 to 20 and the number of groups is set to 10. There is no other grouping factor in original data; however, random effects for artificial grouping factor are added to responses with various scales relative to the error variance. We compare four selection methods, as similar to the simulation study with artificial data under the first 8 candidate models.

Table 6 shows similar patterns to results by the simulation study in the previous section. mBIC and mAIC fails to catch the grouping factor even though its variance σ_a^2 increases and also dependency between random effects b_0 and b_1 are ignored for all cases. cAIC and bAIC prefer to over-parametrize models in most cases.

Table 6: Relative frequencies of selected models by four methods for simulation using parameters from real data (n = 240)

	Relative value of σ_b to σ_ϵ																			
Model	σ_a = 0.0				σ_a = (0.5)σ_ϵ				σ_a = (1.0)σ_ϵ				σ_a = (2.0)σ_ϵ				σ_a = (3.0)σ_ϵ			
	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC	mBIC	mAIC	cAIC	bAIC
1	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.89	0.74	0.08	0.04	0.93	0.74	0.08	0.00	0.92	0.76	0.11	0.05	0.90	0.76	0.11	0.08	0.94	0.74	0.09	0.05
5	0.00	0.00	0.11	0.04	0.00	0.00	0.12	0.06	0.00	0.01	0.08	0.02	0.00	0.00	0.13	0.05	0.00	0.00	0.10	0.06
6	0.11	0.26	0.31	0.27	0.05	0.22	0.36	0.34	0.07	0.22	0.32	0.24	0.09	0.20	0.31	0.25	0.06	0.21	0.30	0.18
7	0.00	0.00	0.12	0.33	0.00	0.00	0.05	0.26	0.00	0.00	0.01	0.20	0.01	0.02	0.09	0.22	0.00	0.02	0.13	0.27
8	0.00	0.00	0.38	0.32	0.01	0.04	0.39	0.34	0.00	0.01	0.48	0.49	0.00	0.02	0.36	0.40	0.00	0.03	0.38	0.44

Candidate models are from number 1 to 8 as listed in Table 2.

† BIC = Bayesian Information Criteria; AIC = Akaike Information Criteria; mBIC = marginal BIC; mAIC = marginal AIC; cAIC = conditional AIC; bAIC = bias-corrected conditional AIC.

5. Conclusions

In linear mixed models, it is not easy to select models since the main interest of model building could be different and the number of parameters in the models is unclear. Therefore, many different methods to select linear mixed models are developed under different assumptions and circumstances. This development can interfere with research and it is very difficult to choose an appropriate selection method for the given purpose and interest of research. It also represents another obstacle for choosing a model selection method in the linear mixed model because most of newly proposed methods have computational complexity.

We compared four methods for model selection: mBIC, mAIC, cAIC and bias-corrected cAIC. Bias-corrected cAIC shows promising performance when candidate models do not included larger models that contain the true model and it is improved from the original cAIC as indicated by its developer. But still bias-corrected cAIC prefers complex model intensively than other methods based on the marginal distribution of linear mixed models. This fact is consistent with theoretical properties of mAIC. It also turns out that mBIC and mAIC have some difficulty in selecting random effects when the design is nested so that more careful attention should be needed applying selection methods based on marginal distributions.

The main findings in this paper provide several new research topics such that performance of model selection methods for linear mixed models should be investigated under more complex nested designs. It is also worthwhile to develop model selection method based on the conditional distribution to prevent intensive preference to the over-parametrized model in cAIC.

References

Dimova, R. B., Markatou, 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.

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

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

McCulloch, C. E. and Searle, S. R. (2001). *Generalized Linear Mixed Models*, John Wiley & Sons, New York.

Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model selection in linear mixed models, *Statistical*

Science, **28**, 135–167.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution, *Annals of Statistics*, **9**, 1135–1151.

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

Received August 17, 2015; Revised September 3, 2015; Accepted September 5, 2015