

**2nd Revision of #06.338**

**A Note on Conditional AIC for Linear Mixed-Effects  
Models**

Hua Liang, Hulin Wu, and Guohua Zou

Jun 22, 2007

**Correspondence:**

Hua Liang, Ph.D.  
Department of Biostatistics  
and Computational Biology  
University of Rochester Medical Center  
601 Elmwood Avenue, Box 630  
Rochester, NY 14642

**Email:** [hliang@bst.rochester.edu](mailto:hliang@bst.rochester.edu)

# A Note on Conditional AIC for Linear Mixed-Effects Models

By HUA LIANG, HULIN WU

*Department of Biostatistics and Computational Biology, University of Rochester Medical Center,  
Rochester, New York 14642, U.S.A.*

hliang@bst.rochester.edu hwu@bst.rochester.edu

AND GUOHUA ZOU

*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080,  
China*

Guohua.Zou@urmc.rochester.edu

## SUMMARY

The conventional model selection criterion AIC has been applied to choose candidate models in mixed-effects models by the consideration of marginal likelihood. Vaida and Blanchard (2005) demonstrated that such a marginal AIC and its small sample correction are inappropriate when the research focus is on clusters. Correspondingly, these authors suggested to use conditional AIC. Their conditional AIC is derived under the assumption of the variance-covariance matrix or scaled variance-covariance matrix of random effects being known. We develop two general conditional AICs but without these strong assumptions. This allows Vaida and Blanchard's conditional AIC to be applied in a wide range. Simulation studies show that the proposed methods are promising.

*Some key words:* Akaike information criterion, conditional likelihood, Kullback-Leibler information, longitudinal data, marginal likelihood, profile likelihood.

## 1 INTRODUCTION

Linear mixed-effects (LME) models (Laird and Ware, 1982), as a powerful tool for the analysis of longitudinal data, have been paid more and more attentions because they can incorporate within-cluster and between-cluster variations into consideration. Statistical estimation and inference for LME models have widely been studied and applied in literature (Vonesh and Chinchilli, 1996; Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000). A fundamental question in LME models, model selection, seems to be disregarded, however. Traditional selection criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978) for cross-sectional data have been parallelly applied for the selection of LME models without justification (Pinheiro and Bates, 2000; Ngo and Brand, 2002). This deficiency was recently noticed by Vaida and Blanchard (2005). These authors explicitly elucidated that, when the researchers' focus is on clusters instead of population,

the traditional AIC and its small sample correction  $AIC_C$  are not appropriate, and suggested the conditional Akaike information and the corresponding model selection criterion: conditional AIC. However, in deriving the conditional AIC, they required that the variance-covariance matrix of random effects should be known when the variance of the measurement error term is known, or the scaled variance-covariance matrix of random effects should be known when the variance of the measurement error term is unknown. These requirements may limit the use of the conditional AIC. The objective of this note is to remove Vaida and Blanchard's assumptions and to propose two more general conditional AICs. This will allow Vaida and Blanchard's conditional AIC to be applied in a wide range.

## 2 GENERAL CONDITIONAL AIC FOR LME MODELS

Assume the data from  $m$  clusters to be modelled by the following LME model:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, m, \quad (1)$$

where  $\mathbf{y}_i$  is an  $n_i \times 1$  vector of observations for cluster  $i$ ,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of fixed effects,  $\mathbf{b}_i$  is a  $q \times 1$  vector of random effects for cluster  $i$ ,  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are the  $n_i \times p$  and  $n_i \times q$  design matrices of full column rank for the fixed and random effects, respectively, and  $\boldsymbol{\varepsilon}_i$  is the disturbance. We assume that  $\mathbf{b}_i$  and  $\boldsymbol{\varepsilon}_i$  are independently and normally distributed with mean of zero and variance-covariance matrices of  $G$  and  $\sigma^2\mathbf{I}_{n_i}$ , respectively, where  $\mathbf{I}_{n_i}$  is the  $n_i \times n_i$  identity matrix. Let  $N = \sum_{i=1}^m n_i$  be the total number of observations, and let  $\boldsymbol{\theta}$  be the vector of parameters in the model, including  $\boldsymbol{\beta}$ ,  $\sigma^2$  and the parameters in  $G$ . Model (1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \mathbf{b} \sim N(0, \mathbf{G}), \quad (2)$$

where  $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_m^T)^T$  is an  $N \times 1$  vector of observations,  $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_m^T)^T$  is an  $N \times p$  matrix of rank  $p$ ,  $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$  is an  $N \times r$  block-diagonal matrix of rank  $r = mq$ ,  $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_m^T)^T$ ,  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \dots, \boldsymbol{\varepsilon}_m^T)^T$ , and  $\mathbf{G} = \text{diag}(G, \dots, G)$  is a  $r \times r$  block-diagonal matrix. Denote the joint density function of  $\mathbf{y}$  and  $\mathbf{b}$  under model (2) by  $g(\mathbf{y}, \mathbf{b} \mid \boldsymbol{\theta})$ . Thus, given  $\mathbf{b}$ , the conditional likelihood is  $g(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{b})$  and the marginal likelihood is  $g(\mathbf{y} \mid \boldsymbol{\theta}) = \int g(\mathbf{y}, \mathbf{b} \mid \boldsymbol{\theta})d\mathbf{b}$ .

Let the true conditional distribution of  $\mathbf{y}$  be  $f(\mathbf{y} \mid \mathbf{u})$ , where  $\mathbf{u}$  is the true random effects vector with distribution  $p(\mathbf{u})$ , and  $f(\mathbf{y}, \mathbf{u})$  be the joint density of  $\mathbf{y}$  and  $\mathbf{u}$ . Then Vaida and Blanchard (2005) defined the conditional Akaike information as follows.

DEFINITION 1 *The conditional Akaike information is defined to be*

$$\begin{aligned} \text{cAI} &= -2E_{f(\mathbf{y}, \mathbf{u})}E_{f(\mathbf{y}^* | \mathbf{u})} \log g\{\mathbf{y}^* \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} \\ &= \int -2 \log g\{\mathbf{y}^* \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} f(\mathbf{y}^* \mid \mathbf{u}) f(\mathbf{y}, \mathbf{u}) d\mathbf{y}^* d\mathbf{y} d\mathbf{u}, \end{aligned} \quad (3)$$

where  $\mathbf{y}^*$  is the prediction dataset which is independent of  $\mathbf{y}$  conditional on  $\mathbf{u}$  and from the same distribution  $f(\cdot \mid \mathbf{u})$  as  $\mathbf{y}$ , and  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  and  $\hat{\mathbf{b}}(\mathbf{y})$  are the estimators of  $\boldsymbol{\theta}$  and  $\mathbf{b}$ , respectively.

The following theorem derives an unbiased estimator of cAI when the variance  $\sigma^2$  is known. The proof is given in the Appendix. Let  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  and  $\hat{\mathbf{b}}(\mathbf{y})$  be the maximum likelihood and the empirical Bayes estimators of  $\boldsymbol{\theta}$  and  $\mathbf{b}$ , respectively.

THEOREM 1 *Assume that the data  $\mathbf{y}$  have true density  $f(\mathbf{y} \mid \mathbf{u}) = g(\mathbf{y} \mid \boldsymbol{\theta}_0, \mathbf{u})$  for some  $\boldsymbol{\theta}_0$  and some random effect  $\mathbf{u}$  with distribution  $p(\mathbf{u})$ . Let the data be modelled by (2) with densities denoted by  $g(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{b})$  and  $p(\mathbf{b})$ . If  $\sigma^2$  is known, then an unbiased estimator of the cAI in (3) is given by*

$$\text{cAIC} = -2 \log g\{\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} + 2\Phi_0(\mathbf{y}), \quad (4)$$

where  $\Phi_0(\mathbf{y}) = \sum_{i=1}^N \partial \hat{y}_i / \partial y_i = \text{tr}(\partial \hat{\mathbf{y}}^T / \partial \mathbf{y})$ , and  $y_i$  and  $\hat{y}_i$  are the  $i$ -th components of  $\mathbf{y}$  and the fitted vector  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$ , respectively.

It is interesting to observe that the expectation of  $\Phi_0(\mathbf{y})$  in (4), conditional on  $\mathbf{u}$ , is just the generalized degrees of freedom defined by Ye (1998) for LME models. From (4), it is seen that unlike for linear fixed-effects models, the penalty term generally depends on the observed data  $\mathbf{y}$  for LME models. The calculation on the penalty function  $\Phi_0(\mathbf{y})$  involves the first partial derivatives  $\partial \hat{y}_i / \partial y_i$  ( $i = 1, \dots, N$ ) which can be directly calculated or numerically approximated by  $\{\hat{y}_i(\mathbf{y} + h\mathbf{e}_i) - \hat{y}_i(\mathbf{y})\} / h$ , where  $h$  is a small number and  $\mathbf{e}_i$  is the  $N \times 1$  vector with the  $i$ -th component of one and other components of zero.

*Remark 1* Assuming that  $\sigma^2$  is known, Vaida and Blanchard (2005) developed a neat result (Theorem 1, p355) for the case of the known  $\mathbf{G}$ . Our Theorem 1 generalizes their result and provides an unbiased estimator of cAI for the unknown  $\mathbf{G}$ .

*Corollary 1* (Vaida and Blanchard 2005) Under the assumptions of Theorem 1, further assume that  $\mathbf{G}$  is known. Then an unbiased estimator of the cAI is

$$\text{cAIC} = -2 \log g\{\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} + 2\rho, \quad (5)$$

where  $\rho = \text{tr}(\mathbf{H}_1)$  is the “effective degrees of freedom” of Hodges and Sargent (2001),  $\mathbf{H}_1$  is the “hat” matrix mapping the observed data vector  $\mathbf{y}$  into the fitted vector  $\hat{\mathbf{y}}$ , that is,  $\hat{\mathbf{y}} = \mathbf{H}_1\mathbf{y}$ .

**Proof:** See the Appendix.

An intuitive explanation on  $\rho$ , the penalty term when both  $\sigma^2$  and  $\mathbf{G}$  are known, can be provided as follows: From the definition of  $\mathbf{H}_1$  (see the proof of Corollary 1), it can be shown that

$$\rho = p + \sum_{i=1}^{r_0} \frac{\lambda_i}{1 + \lambda_i},$$

where  $\lambda_1, \dots, \lambda_{r_0}$  are the non-zero eigenvalues of the matrix  $\mathbf{D}_0^{1/2}\mathbf{Z}^\top(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}\mathbf{D}_0^{1/2}$  with  $\mathbf{D}_0 = \sigma^{-2}\mathbf{G}$  and  $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ . Note that in the scenario of Corollary 1, only  $\boldsymbol{\beta}$  is unknown. So the first term on the right-hand side of the above formula is the total number of parameters in LME model. Thus, unlike for the usual linear fixed-effects model, the penalty term is not only the number of unknown parameters for LME model. The second term in the expression of  $\rho$  is the extra penalty due to random effects. Also, observe that this term is smaller than the number of random effects,  $r$ , showing that the extra penalty is not the number of random effects terms, although these random effects may be independent (note that the covariate matrix  $\mathbf{Z}$  in model (2) can be non-block diagonal). Further, when  $\mathbf{G}$  is unknown, Vaida and Blanchard (2005) suggested to use the observed  $\hat{\rho} = \text{tr}\{\mathbf{H}_1(\widehat{\mathbf{G}})\}$ , where  $\widehat{\mathbf{G}}$  is the maximum likelihood estimator of  $\mathbf{G}$ . Observe that when  $\mathbf{G}$  is unknown, we have  $\hat{\mathbf{y}} = \mathbf{H}_1(\widehat{\mathbf{G}})\mathbf{y}$ . So from Theorem 1, the exact penalty term when  $\mathbf{G}$  is unknown will be

$$\Phi_0(\mathbf{y}) = \hat{\rho} + \mathbf{1}^\top \mathbf{H}(\widehat{\mathbf{G}})\mathbf{y},$$

where  $\mathbf{1}$  is the  $N \times 1$  vector of ones, and

$$\mathbf{H}(\widehat{\mathbf{G}}) = \begin{pmatrix} \frac{\partial h_{11}(\widehat{\mathbf{G}})}{\partial y_1} & \cdots & \frac{\partial h_{1N}(\widehat{\mathbf{G}})}{\partial y_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial h_{N1}(\widehat{\mathbf{G}})}{\partial y_N} & \cdots & \frac{\partial h_{NN}(\widehat{\mathbf{G}})}{\partial y_N} \end{pmatrix}$$

with  $h_{ij}(\widehat{\mathbf{G}})$  being the  $(i, j)$ -th element of the matrix  $\mathbf{H}_1(\widehat{\mathbf{G}})$  (here we write  $\mathbf{H}$  as a function of  $\widehat{\mathbf{G}}$  but it may depend on  $\mathbf{y}$  not only through  $\widehat{\mathbf{G}}$ ). The second term  $\mathbf{1}^T \mathbf{H}(\widehat{\mathbf{G}}) \mathbf{y}$  is the additional penalty due to the variability of estimating unknown  $\mathbf{G}$ .

*Remark 2* In Theorem 1 and Corollary 1, the assumption of  $f(\mathbf{y} \mid \mathbf{u}) = g(\mathbf{y} \mid \boldsymbol{\theta}_0, \mathbf{u})$  means that the true model is included in the candidate model family. This is a *traditional assumption* in deriving model selection criterion (see, for example, Akaike, 1973; Hurvich and Tsai, 1989; Burnham and Anderson, 1998; and Hurvich et al., 1998). The further assumption of  $\mathbf{G}$  being known in Corollary 1 implies that the covariate matrices for random effects under the true and candidate models are in fact exactly the same. The removal of this further assumption shows that the covariate matrices for random effects under the true and candidate models can be different. Further, in the proof of Theorem 1, the expression of  $\boldsymbol{\mu}$  ( $= \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{u}$ , where  $\boldsymbol{\beta}_0$  is the true parameter for fixed effects) is not used. This means that the traditional assumption that the candidate models include the true one can even be removed. As an example, if the data  $\mathbf{y}$  come from a LME model mentioned in Vaida and Blanchard (2005):  $\mathbf{y} = \mathbf{P}\boldsymbol{\alpha} + \mathbf{Q}\mathbf{v} + \mathbf{e}$  with  $\mathbf{v} \sim N(0, \mathbf{S})$ ,  $\mathbf{e} \sim N(0, \sigma_0^2 \mathbf{I}_N)$ , and  $\mathbf{P}$  and  $\mathbf{Q}$  containing covariates different from  $\mathbf{X}$  and  $\mathbf{Z}$ , then Theorem 1 and Corollary 1 still hold.

In practice,  $\sigma^2$  is generally unknown and needs to be estimated. In this case, the maximum likelihood or restricted maximum likelihood method (Davidian and Giltinan, 1995) is an alternative. Let  $\hat{\sigma}^2$  be an estimator of  $\sigma^2$ . We have the following result, whose proof is postponed to the Appendix.

**THEOREM 2** Under the set-up of Theorem 1, assume that  $\sigma^2$  is unknown. Then an approximate unbiased estimator of the cAI in (3) is given by

$$\text{cAIC} = -2 \log g\{\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} + 2\hat{\sigma}_0^2/\hat{\sigma}^2 \cdot \Phi_0(\mathbf{y}) + \Phi_1(\mathbf{y}), \quad (6)$$

where  $\hat{\sigma}_0^2$  is an estimator of the true variance of error term  $\sigma_0^2$ , and

$$\begin{aligned}\Phi_1(\mathbf{y}) &= 2\hat{\sigma}_0^2 \sum_{i=1}^N \left\{ (\hat{y}_i - y_i) \cdot \frac{\partial(\hat{\sigma}^{-2})}{\partial y_i} \right\} + \hat{\sigma}_0^4 \sum_{i=1}^N \frac{\partial^2(\hat{\sigma}^{-2})}{\partial y_i^2} \\ &= 2\hat{\sigma}_0^2 (\hat{\mathbf{y}} - \mathbf{y})^\top \frac{\partial(\hat{\sigma}^{-2})}{\partial \mathbf{y}} + \hat{\sigma}_0^4 \text{tr} \left\{ \frac{\partial^2(\hat{\sigma}^{-2})}{\partial \mathbf{y} \partial \mathbf{y}^\top} \right\}.\end{aligned}$$

Let us explain the penalty term. When  $\sigma^2$  is known,  $\Phi_1(\mathbf{y})$  will be zero and the corresponding penalty term will be  $\Phi_0(\mathbf{y})$  as in Theorem 1. So  $\Phi_1(\mathbf{y})$  is in fact an extra penalty due to the variability of estimating unknown  $\sigma^2$ .

In Theorem 2, to obtain a feasible model selection criterion, we need to input the estimator of  $\sigma_0^2, \hat{\sigma}_0^2$ . Note that  $\hat{\sigma}_0^2$  is generally different from the estimator of the variance of candidate model  $\sigma^2, \hat{\sigma}^2$ , although they can be the same for some special cases such as linear fixed-effects models and LME models with the known scaled variance-covariance matrix  $\sigma^{-2}\mathbf{G}$ . So we provide an estimator here. Consider the conventional estimators of  $\beta, \mathbf{b}, \mathbf{G}$  and  $\sigma^2$ . For given  $\mathbf{G}$  and  $\sigma^2$ , it is well known that the MLEs of  $\beta$  and  $\mathbf{b}$  are given by  $\hat{\beta} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}$ , and  $\hat{\mathbf{b}} = \mathbf{G} \mathbf{Z}^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta})$ , respectively, where  $\Sigma = \sigma^2 \mathbf{I}_N + \mathbf{Z} \mathbf{G} \mathbf{Z}^\top$  (see, for example, Laird and Ware, 1982). The  $\sigma^2$  and unknown parameters in  $\mathbf{G}$  can be estimated using the maximum likelihood or restricted maximum likelihood method (Davidian and Giltinan, 1995), and  $\hat{\beta}$  and  $\hat{\mathbf{b}}$  can be obtained by replacing the unknown parameters in  $\mathbf{G}$  and  $\sigma^2$  by their estimates. Denote

$$\mathbf{B} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1}\}^\top \{\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1}\}.$$

Then an estimator of  $\sigma_0^2$  can be taken as

$$\hat{\sigma}_0^2 = \frac{1}{\text{tr}(\mathbf{B})} (\mathbf{y} - \mathbf{X} \hat{\beta})^\top (\mathbf{y} - \mathbf{X} \hat{\beta}). \quad (7)$$

It can be readily shown that  $\hat{\sigma}_0^2$  is an unbiased estimator of  $\sigma_0^2$  under the assumptions of the candidate models including the true one and  $\mathbf{Z} \in \mathbf{R}(\mathbf{X})$ , where  $\mathbf{R}(\mathbf{X})$  is the space spanned by the matrix  $\mathbf{X}$  (Of course, these assumptions only serve as the purpose of derivation of the model selection criterion).

To apply our criterion to choose the variables in LME model, we have to calculate the function  $\Phi_1(\mathbf{y})$  which depends only on the data but involves the first and second partial derivatives of the estimator with respect to  $\mathbf{y}$ . This can be done in a similar way to calculating  $\Phi_0(\mathbf{y})$ .

Theorem 2 provides an approximately unbiased estimator of cAI for the unknown  $\mathbf{G}$  when  $\sigma^2$  is unknown. If we further assume that the scaled variance-covariance matrix  $\sigma^{-2}\mathbf{G}$  is known as in Vaida and Blanchard (2005), then we have

*Corollary 2* Under the set-up of Theorem 2, assume further that  $\mathbf{D}_0 = \sigma^{-2}\mathbf{G}$  is known. Then an approximately unbiased estimator of the cAI in (3) is

$$\text{cAIC} = -2 \log g\{\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} + 2\rho \left(1 + \frac{1}{N}\right) - 2 + 4 \left(1 + \frac{2}{N}\right) \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)^2\mathbf{y}}{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)\mathbf{y}}. \quad (8)$$

**Proof:** See the Appendix.

Unlike Corollary 1, it can be seen from (8) that our model selection criterion in Corollary 2 is different from that of Vaida and Blanchard (2005) which is

$$\text{cAIC}_{\text{VB}} = -2 \log g\{\mathbf{y} \mid \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} + 2K,$$

where  $K$  is given by

$$K = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)}. \quad (9)$$

But they are indeed close. Clearly, from formula (8), for large  $N$ , we have

$$\text{cAIC} \approx \text{cAIC}_{\text{VB}} + 4 \cdot \left\{ \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)^2\mathbf{y}}{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)\mathbf{y}} - 1 \right\}.$$

The second term is the increasing quantity ( $< 0$ ) resulting from  $\sigma_0^2$  being replaced by its estimator

$$\hat{\sigma}_0^2 = \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)\mathbf{y}}{N}$$

in (A.6) directly.

*Remark 3* Like Theorem 1, in the proof of (A.6), the assumption of  $f(\mathbf{y} \mid \mathbf{u}) = g(\mathbf{y} \mid \boldsymbol{\theta}_0, \mathbf{u})$ , that is, the true model is included in the candidate model family, is not necessary. However, to derive the estimator of  $\sigma_0^2$  as in Theorem 2 and Corollary 2, the traditional assumption is generally needed. This also means that the assumption of the true model being included in the candidate model family is needed only in the derivation of the estimator of  $\sigma_0^2$ .

## 3 SIMULATION STUDY

In this section, we present simulation results to study the behavior of the proposed methods under small and moderate sample sizes. To make a comparison, we generate data from the framework that Vaida and Blanchard (2005) used, that is, the data are generated from the model

$$y_{ij} = (\beta_0 + \beta_1 t_j) + (b_{0i} + b_{1i} t_j) + \varepsilon_{ij}, \quad i = 1, \dots, m = 10, \quad j = 1, \dots, n_i,$$

where  $\beta_0 = -2.78$ ,  $\beta_1 = -0.186$ ,  $t_j = 5j$ ,  $(b_{0i}, b_{1i})^T$  follows a normal distribution with the mean of zero and variance-covariance matrix of  $\begin{pmatrix} 0.0367 & -0.00126 \\ -0.00126 & 0.00279 \end{pmatrix}$ ,  $\varepsilon_{ij}$  are iid with the distribution of  $N(0, \sigma^2)$ . In our simulation experiments, similar to Vaida and Blanchard (2005), we consider  $\sigma = 0.0705, 0.141$ , and  $0.282$  and the following three scenarios: (i)  $j = 0, 1, \dots, 5$ , giving  $n_i = 6$ ; (ii)  $j = 0, 1, \dots, 25$ , giving  $n_i = 26$ ; and (iii)  $j = 0, 1, \dots, 50$ , giving  $n_i = 51$ . For each of the nine configurations, 500 independent sets of data are generated. Assuming the variance-covariance matrix  $G$  is unstructured, we mainly compare the estimates of the bias correction (BC), which is defined as  $\text{cAI} = E_{f(\mathbf{y}, \mathbf{u})} \{-2 \log g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} + 2\text{BC}$ . Specifically, we compare  $\Phi_0(\mathbf{y})$  in Theorem 1 and  $\hat{\rho}$  given by Vaida and Blanchard (2005) with the true BC values when  $\sigma^2$  is known, and  $\Phi(\mathbf{y}) \equiv \hat{\sigma}_0^2 / \hat{\sigma}^2 \cdot \Phi_0(\mathbf{y}) + \Phi_1(\mathbf{y}) / 2$  in Theorem 2,  $K$  with  $\rho$  replaced by  $\hat{\rho}$  (see formula (9)) and  $K^*$  which is  $K$  with  $\rho$  replaced by  $\Phi_0(\mathbf{y})$  with the true BC values when  $\sigma^2$  is unknown.

Tables 1 and 2 summarize the results of this small simulation study for known and unknown  $\sigma^2$ , respectively. The results obtained are in accord with the theory. Table 1 shows that when  $\sigma^2$  is known, the estimated values based on the proposed method and Vaida and Blanchard's (2005) method are both close to the BC values especially for the small values of error variance. Generally, the larger the sample size, the closer. However, it is worthy of emphasizing that the estimated values based on the former are consistently closer to the BC values than those based on the latter, especially for the large values of error variance or small values of  $n_i$ . From Table 2, the similar conclusions are observed when  $\sigma^2$  is unknown. All of the three estimates  $K$ ,  $K^*$  and  $\Phi(y)$  are close to the BC values. Among these three estimates,  $K$  and  $\Phi(y)$  perform slightly better and these two estimates are comparable. These observations show that our methods are promising.

Table 1: Simulation study. Comparison of BC and its two estimates,  $\hat{\rho}$  and  $\Phi_0(\mathbf{y})$  based on 500 runs for known  $\sigma^2$ .

$n_i$	$\sigma$	BC	$\hat{\rho}$	$\Phi_0(\mathbf{y})$
6	0.0705	19.549	19.994	19.38
26	0.0705	19.875	19.999	19.837
51	0.0705	19.926	19.999	19.891
6	0.141	17.638	19.731	18.253
26	0.141	19.339	19.976	19.355
51	0.141	19.547	19.986	19.597
6	0.282	15.818	16.944	15.436
26	0.282	17.832	19.265	17.927
51	0.282	18.723	19.763	18.648

#### 4 CONCLUDING REMARKS

This note removed the assumption on the variance-covariance matrix or scaled variance-covariance matrix of random effects being known in the conditional AIC of Vaida and Blanchard (2005) and developed two more general conditional AICs. This would substantially enlarge the use of the conditional AIC in LME model selection.

It is worthy of noting that the derivations of (A.2) and (A.6) in the Appendix do not require the assumption that the candidate models include the true one. This means that when the error variance  $\sigma^2$  is known, to derive a reasonable model selection criterion, this traditional assumption is not necessary. Further analysis shows that this conclusion is still true even the error variances under the true and candidate models are unknown but the same. Also, the assumption of the true model being included in the candidate model family is needed only in the derivation of the estimator of the true error variance. Noting that the error variance is a nuisance parameter, this explains in part why the commonly used AIC and  $AIC_C$  in fixed-effects models often perform well even the candidate model family does not include the true model, although these selection criteria were derived under the above traditional assumption.

Different from the derivations in the model selection literature, we made use of the integration by part technique, which has been used to obtain risk-unbiased estimators before (Stein, 1981; Lu

Table 2: Simulation study. Comparison of BC and its three estimates,  $K$  given in Vaida and Blanchard (2005),  $K^*$ , and  $\Phi(y)$  based on 500 runs for unknown  $\sigma^2$ .

$n_i$	$\sigma$	BC	$K$	$K^*$	$\Phi(y)$
6	0.0705	21.557	21.124	21.515	20.697
26	0.0705	21.121	21.256	21.092	20.936
51	0.0705	21.220	20.130	21.021	20.967
6	0.141	20.049	19.843	20.328	20.562
26	0.141	20.109	20.013	21.605	20.919
51	0.141	20.435	20.118	21.725	20.954
6	0.282	17.669	16.904	17.363	17.813
26	0.282	18.494	18.360	19.162	18.375
51	0.282	19.354	19.187	19.770	19.492

and Berger, 1989), to derive the selection criterion for LME models. It can be seen that our method can also be applied to obtain marginal AIC based on the marginal likelihood and overall AIC based on the joint likelihood for LME models, and  $AIC_C$  for nonparametric regression models (Hurvich et al., 1998) and single-index models (Naik and Tsai, 2001) etc. Further, the principle of this note may be extended to generalized mixed-effects models, and applied to select smoothing parameters in the semiparametric regression. These topics warrant our future researches.

#### ACKNOWLEDGEMENTS

The authors thank the Editor and one referee for their constructive comments and suggestions which greatly improved the original manuscript. Liang and Zou's research was partially supported by two grants from the National Institute of Allergy and Infectious Diseases. Wu's research was partially supported by three grants from the National Institute of Allergy and Infectious Diseases. Zou's research was also partially supported by one grant from the NSF of China.

APPENDIX

*Proof of Theorem 1.* Denote  $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{u}$ , where  $\boldsymbol{\beta}_0$  is the true parameter for fixed effects.

Then it is readily seen that

$$\begin{aligned} \text{cAI} &= -2E_{f(\mathbf{y},\mathbf{u})}E_{f(\mathbf{y}^*|\mathbf{u})} \log g\{\mathbf{y}^* | \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\} \\ &= E_{f(\mathbf{y},\mathbf{u})} \left\{ N \log(2\pi\sigma^2) + N + \frac{1}{\sigma^2}(\hat{\mathbf{y}} - \boldsymbol{\mu})^\top(\hat{\mathbf{y}} - \boldsymbol{\mu}) \right\}. \end{aligned}$$

Also, we have

$$E_{f(\mathbf{y},\mathbf{u})}\{-2 \log g(\mathbf{y} | \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} = E_{f(\mathbf{y},\mathbf{u})} \left\{ N \log(2\pi\sigma^2) + \frac{1}{\sigma^2}(\mathbf{y} - \hat{\mathbf{y}})^\top(\mathbf{y} - \hat{\mathbf{y}}) \right\}.$$

Thus, after some calculations, we obtain

$$\begin{aligned} \text{cAI} - E_{f(\mathbf{y},\mathbf{u})}\{-2 \log g(\mathbf{y} | \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} &= \frac{2}{\sigma^2}E_{f(\mathbf{y},\mathbf{u})}\{(\mathbf{y} - \boldsymbol{\mu})^\top \hat{\mathbf{y}}\} \\ &= \frac{2}{\sigma^2}E_{p(\mathbf{u})}E_{f(\mathbf{y}|\mathbf{u})} \left\{ \sum_{i=1}^N (y_i - \mu_i) \hat{y}_i \right\}, \quad (\text{A.1}) \end{aligned}$$

where  $\mu_i$  is the  $i$ -th component of  $\boldsymbol{\mu}$ .

Note that under the true model, for given  $\mathbf{u}$ ,  $\mathbf{y}$  follows a normal distribution with the mean  $\boldsymbol{\mu}$  and variance-covariance matrix  $\sigma^2\mathbf{I}_N$ . Assuming that  $\hat{y}_i$  is a continuous function with piecewise continuous partial derivatives with respect to  $\mathbf{y}$ , it can be shown from the integration by part that

$$E_{f(\mathbf{y}|\mathbf{u})} \left\{ \sum_{i=1}^N (y_i - \mu_i) \hat{y}_i \right\} = \sigma^2 E_{f(\mathbf{y}|\mathbf{u})} \left( \sum_{i=1}^N \frac{\partial \hat{y}_i}{\partial y_i} \right),$$

providing each expectation on the right-hand side exists (see also Stein, 1981; and Lu and Berger, 1989). Therefore, (A.1) becomes

$$\begin{aligned} \text{cAI} - E_{f(\mathbf{y},\mathbf{u})}\{-2 \log g(\mathbf{y} | \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} &= 2E_{p(\mathbf{u})}E_{f(\mathbf{y}|\mathbf{u})} \left( \sum_{i=1}^N \frac{\partial \hat{y}_i}{\partial y_i} \right) \\ &= 2E_{f(\mathbf{y},\mathbf{u})}\{\Phi_0(\mathbf{y})\}. \quad (\text{A.2}) \end{aligned}$$

Thus, an unbiased estimator of the cAI is given by cAIC in (4) and this completes the proof of Theorem 1.

*Proof of Corollary 1.* From Hodges and Sargent (2001) or Vaida and Blanchard (2005), when  $\sigma^2$  and  $\mathbf{G}$  are known, the fitted vector is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} = \mathbf{H}_1\mathbf{y},$$

where  $\mathbf{H}_1 = (\mathbf{X} \mathbf{Z})(\mathbf{M}^\top \mathbf{M})^{-1}(\mathbf{X} \mathbf{Z})^\top$  with

$$\mathbf{M} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{O} & -\Delta \end{pmatrix}$$

and  $\Delta$  being some  $r \times r$  matrix such that  $\sigma^{-2}\mathbf{G} = (\Delta^\top \Delta)^{-1}$ . Thus,

$$\Phi_0(\mathbf{y}) = \text{tr} \left( \frac{\partial \hat{\mathbf{y}}^\top}{\partial \mathbf{y}} \right) = \text{tr}(\mathbf{H}_1) = \rho.$$

*Proof of Theorem 2.* Similar to the proof of Theorem 1, we have

$$\text{cAI} = E_{f(\mathbf{y}, \mathbf{u})} \left\{ N \log(2\pi\hat{\sigma}^2) + \frac{N\sigma_0^2}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^2} (\boldsymbol{\mu} - \hat{\mathbf{y}})^\top (\boldsymbol{\mu} - \hat{\mathbf{y}}) \right\},$$

and

$$E_{f(\mathbf{y}, \mathbf{u})} \{-2 \log g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} = E_{f(\mathbf{y}, \mathbf{u})} \left\{ N \log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) \right\}.$$

So, after some calculations, we have

$$\begin{aligned} \text{cAI} &= E_{f(\mathbf{y}, \mathbf{u})} \{-2 \log g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} \\ &= E_{f(\mathbf{y}, \mathbf{u})} \left[ \frac{N\sigma_0^2}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^2} \{-2(\mathbf{y} - \boldsymbol{\mu})^\top (\mathbf{y} - \hat{\mathbf{y}}) + (\mathbf{y} - \boldsymbol{\mu})^\top (\mathbf{y} - \boldsymbol{\mu})\} \right] \\ &= E_{p(\mathbf{u})} E_{f(\mathbf{y}|\mathbf{u})} \left\{ \frac{N\sigma_0^2}{\hat{\sigma}^2} + 2 \sum_{i=1}^N (y_i - \mu_i) \cdot \frac{\hat{y}_i - y_i}{\hat{\sigma}^2} + \sum_{i=1}^N (y_i - \mu_i)^2 \cdot \frac{1}{\hat{\sigma}^2} \right\}. \end{aligned} \quad (\text{A.3})$$

Assuming that  $(\hat{y}_i - y_i)/\hat{\sigma}^2$  is a continuous function with piecewise continuous partial derivatives with respect to  $\mathbf{y}$ , it can be shown from the integration by part that

$$E_{f(\mathbf{y}|\mathbf{u})} \left\{ \sum_{i=1}^N (y_i - \mu_i) \cdot \frac{\hat{y}_i - y_i}{\hat{\sigma}^2} \right\} = \sigma_0^2 E_{f(\mathbf{y}|\mathbf{u})} \left\{ \sum_{i=1}^N \frac{\partial}{\partial y_i} \left( \frac{\hat{y}_i - y_i}{\hat{\sigma}^2} \right) \right\}, \quad (\text{A.4})$$

providing each expectation on the right-hand side exists.

Similarly, assuming that  $\partial(\hat{\sigma}^{-2})/\partial y_i (i = 1, \dots, N)$  are continuous functions with piecewise continuous partial derivatives and the corresponding expectations exist, we have

$$E_{f(\mathbf{y}|\mathbf{u})} \left\{ \sum_{i=1}^N (y_i - \mu_i)^2 \cdot \frac{1}{\hat{\sigma}^2} \right\} = E_{f(\mathbf{y}|\mathbf{u})} \left[ \sum_{i=1}^N \left\{ \frac{\sigma_0^2}{\hat{\sigma}^2} + \sigma_0^4 \cdot \frac{\partial^2(\hat{\sigma}^{-2})}{\partial y_i^2} \right\} \right]. \quad (\text{A.5})$$

Substituting (A.4) and (A.5) in (A.3), we obtain

$$\begin{aligned} \text{cAI} &= E_{f(\mathbf{y}, \mathbf{u})} \{-2 \log g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} \\ &= E_{f(\mathbf{y}, \mathbf{u})} \left[ 2\sigma_0^2 \sum_{i=1}^N \left\{ \frac{\partial \hat{y}_i}{\partial y_i} \cdot \frac{1}{\hat{\sigma}^2} + (\hat{y}_i - y_i) \cdot \frac{\partial(\hat{\sigma}^{-2})}{\partial y_i} \right\} + \sigma_0^4 \sum_{i=1}^N \frac{\partial^2(\hat{\sigma}^{-2})}{\partial y_i^2} \right] \\ &= E_{f(\mathbf{y}, \mathbf{u})} \left[ \frac{2\sigma_0^2}{\hat{\sigma}^2} \Phi_0(\mathbf{y}) + 2\sigma_0^2 \sum_{i=1}^N \left\{ (\hat{y}_i - y_i) \cdot \frac{\partial(\hat{\sigma}^{-2})}{\partial y_i} \right\} + \sigma_0^4 \sum_{i=1}^N \frac{\partial^2(\hat{\sigma}^{-2})}{\partial y_i^2} \right]. \end{aligned} \quad (\text{A.6})$$

From this, we see that Theorem 2 is true.

*Proof of Corollary 2.* From Hodges and Sargent (2001) or Vaida and Blanchard (2005), it can be seen that both  $\sigma_0^2$  in the true model and  $\sigma^2$  in the candidate model can be estimated by

$$\hat{\sigma}_0^2 = \hat{\sigma}^2 = \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)\mathbf{y}}{N}.$$

After some algebras, it can be shown that

$$(\hat{\mathbf{y}} - \mathbf{y})^\top \frac{\partial(\hat{\sigma}^{-2})}{\partial \mathbf{y}} = \frac{2}{N} \cdot \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)^2 \mathbf{y}}{\hat{\sigma}^4},$$

and

$$\text{tr} \left\{ \frac{\partial^2(\hat{\sigma}^{-2})}{\partial \mathbf{y} \partial \mathbf{y}^\top} \right\} = -\frac{2}{N} \frac{N - \rho}{\hat{\sigma}^4} + \frac{8}{N^2} \cdot \frac{\mathbf{y}^\top(\mathbf{I} - \mathbf{H}_1)^2 \mathbf{y}}{\hat{\sigma}^6}.$$

Substituting these expressions in  $\Phi_1(y)$  of (6), we obtain (8).

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. Petrov and F. Csaki, pp. 267-81. Budapest: Akademiai Kiado.
- Burnham, K. P. & Anderson, D. P. (1998). *Model Selection and Inference: A Practical Information-Theoretical Approach*. New York: Springer-Verlag.
- Hodges, J. S. & Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367-79.
- Hurvich, C. M., Simonoff, J. S. & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* **B 60**, 271-93.
- Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.
- Laird, N. M. & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-74.
- Lu, K. L. & Berger, J. O. (1989). Estimation of normal means: frequentist estimation of loss. *Ann. Statist.* **17**, 890-906.
- Naik, P. A. & Tsai, C. L. (2001). Single-index model selections. *Biometrika* **61**, 821-32.

- Ngo, L. & Brand, R. (2002). Model selection in linear mixed effects models using SAS Proc Mixed. SUGI 22.
- Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*, New York: Springer.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-51.
- Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-70.
- Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- Vonesh, E. F. & Chinchilli, V. M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker, Inc.
- Ye, J. M. (1998). On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* **93**, 120-31.