Published in final edited form as:
Biometrika. 2008 ; 95(3): 773–778. doi:10.1093/biomet/asn023.

A Note on Conditional AIC for Linear Mixed-Effects Models

HUA LIANG,
Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A. hliang@bst.rochester.edu

HULIN WU, and
Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A. hwu@bst.rochester.edu

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. The conditional AIC is derived under the assumptions of the variance-covariance matrix or scaled variance-covariance matrix of random effects being known. We develop a general conditional AIC 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 method is 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 a more general conditional AIC.

This will allow Vaida and Blanchard's conditional AIC to be applied in a wide range. This note considers the case of known error variance. For the case of unknown error variance, a discussion can be found in Liang et al. (2006) which is available from the authors upon request.

2. GENERAL CONDITIONAL AIC FOR LME MODELS

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

$$\mathbf{y}_i=\mathbf{X}_i\beta+\mathbf{Z}_i\mathbf{b}_i+\varepsilon_i,\ i=1,\ldots,m, \tag{1}$$

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 for the fixed and random effects of full column rank, 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 an $n_i \times n_i$ 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 $\boldsymbol{\beta}$, σ^2 and the parameters in G. Model (1) can be written as

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

where $\mathbf{y}=(\mathbf{y}_1^{\mathrm{T}},\ldots,\mathbf{y}_m^{\mathrm{T}})^{\mathrm{T}}$ is an $N \times 1$ vector of observations, $\mathbf{X}=(\mathbf{X}_1^{\mathrm{T}},\ldots,\mathbf{X}_m^{\mathrm{T}})^{\mathrm{T}}$ is an $N \times p$ matrix of rank p, $\mathbf{Z}$ = diag($\mathbf{Z}_1$, …, $\mathbf{Z}_m$) is an $N \times r$ block-diagonal matrix of rank $r = mq$, $\mathbf{b}=(\mathbf{b}_1^{\mathrm{T}},\ldots,\mathbf{b}_m^{\mathrm{T}})^{\mathrm{T}}$, $\varepsilon=(\varepsilon_1^{\mathrm{T}},\ldots,\varepsilon_m^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{G}$ = diag(G, …, 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 $\mathbf{y}$ is $f(\mathbf{y} \mid \mathbf{u})$, 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

$$\begin{aligned}\mathrm{cAI}&=-2E_{f(\mathbf{y},\mathbf{u})}E_{f(\mathbf{y}^*|\mathbf{u})}\log g\{\mathbf{y}^*|\widehat{\theta}(\mathbf{y}),\widehat{\mathbf{b}}(\mathbf{y})\}\\&=\int -2\log g\left\{\mathbf{y}^*|\widehat{\theta}(\mathbf{y}),\widehat{\mathbf{b}}(\mathbf{y})\right\}f(\mathbf{y}^*|\mathbf{u})f(\mathbf{y},\mathbf{u})d\mathbf{y}^*d\mathbf{y}d\mathbf{u},\end{aligned} \tag{3}$$

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* $\mathbf{y}$, $\hat{\boldsymbol{\theta}}(\mathbf{y})$ *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 σ^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* σ^2 *is known, then an unbiased estimator of the cAI in* (*3*) *is given by*

$$\mathrm{cAIC}=-2\log g\{\mathbf{y}|\widehat{\theta}(\mathbf{y}),\widehat{\mathbf{b}}(\mathbf{y})\}+2\Phi_0(\mathbf{y}), \tag{4}$$

where $\Phi_0(\mathbf{y})=\sum_{i=1}^{N}\partial\widehat{y}_i/\partial y_i=\mathrm{tr}(\partial\widehat{\mathbf{y}}^{\mathrm{T}}/\partial\mathbf{y})$, 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.

From (4), it is seen that unlike for linear fixed-effects models, the penalty term generally depends on the observed data **y** for LME models. The calculation on the penalty function $\Phi_0(\mathbf{y})$ involves the first partial derivatives $\partial\hat{y}/\partial y_i$ ($i = 1, \ldots, N$) 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—Vaida and Blanchard (2005) developed a neat result (Theorem 1, p355) for the case of the known **G** when σ^2 is known. However, they claimed that no unbiased estimator for cAI such as (9) of their paper (see (5) below) exists for the unknown **G**. Our Theorem 1 provides an unbiased estimator of cAI for the unknown **G** when σ^2 is known.

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

$$\text{cAIC} = -2\log g\left\{\mathbf{y}|\widehat{\theta}(\mathbf{y}),\widehat{\mathbf{b}}(\mathbf{y})\right\} + 2\rho, \tag{5}$$

where $\rho = \text{tr}(\mathbf{H}_1)$, $\mathbf{H}_1$ is the "hat" matrix mapping the observed data vector **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 ρ, the penalty term when both σ^2 and **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 $\lambda_1, \ldots, \lambda_{r0}$ are the non-zero eigenvalues of the matrix $\mathbf{D}_0^{1/2}\mathbf{Z}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Z}\mathbf{D}_0^{1/2}$ with $\mathbf{D}_0 = \sigma^{-2}\mathbf{G}$ and $\mathrm{P}_{\mathrm{X}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. 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 ρ 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 **Z** in model (2) can be non-block diagonal). Further, when **G** is unknown, Vaida and Blanchard (2005) suggested to use the observed $\hat{\rho} = \text{tr}\{\mathbf{H}_1(\hat{\mathbf{G}})\}$, where $\hat{\mathbf{G}}$ is the maximum likelihood estimator of **G**. Observe that when **G** is unknown, we have $\hat{\mathbf{y}} = \mathbf{H}_1(\hat{\mathbf{G}})\mathbf{y}$. So from Theorem 2.1, the exact penalty term when **G** is unknown will be

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

where **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 $h_{ij}(\hat{\mathbf{G}})$ being the (i, j)-th element of the matrix $\mathbf{H}_1(\hat{\mathbf{G}})$ (here we write **H** as a function of $\hat{\mathbf{G}}$ but it may depend on **y** not only through $\hat{\mathbf{G}}$). The second term $\mathbf{1}^{\mathrm{T}}\mathbf{H}(\hat{\mathbf{G}})\mathbf{y}$ is the additional penalty due to the variability of estimating unknown **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 **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 $\boldsymbol{\mu}$ (= $\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Zu}$, where $\boldsymbol{\beta}_0$ is the true parameter for fixed effects) is not useful. This means that the traditional assumption that the candidate models include the true one can even be removed. As an example, if the data **y** come from a LME model mentioned in Vaida and Blanchard (2005): $\mathbf{y} = \mathbf{P}\boldsymbol{\alpha} + \mathbf{Qv} + \mathbf{e}$ with $\mathbf{v} \sim N(0, \mathbf{S})$, $\mathbf{e} \sim N(0, \sigma_0^2 \mathbf{I}_N)$, and **P** and **Q** containing covariates different from **X** and **Z**, then Theorem 1 and Corollary 1 still hold.

3. SIMULATION STUDY

In this section, we describe simulation results to study the behavior of the proposed method 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

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

where $\beta_0 = -2.78$, $\beta_1 = -0.186$, $t_j = 5j$, $(b_{0i}, b_{1i})^{\mathrm{T}}$ follows a normal distribution with mean of zero and variance-covariance matrix of $\begin{pmatrix} 0.0367 & -0.00126 \\ -0.00126 & 0.00279 \end{pmatrix}$, ε_{ij} are iid with $N(0, \sigma^2)$. In our simulation experiments, similarly to Vaida and Blanchard (2005), we consider $\sigma = 0.0705$, 0.141, and 0.282 and the following three scenarios (i) $j = 0, 1, \ldots, 5$, giving $n_i = 6$; (ii) $j = 0, 1, \ldots, 25$, giving $n_i = 26$; and (iii) $j = 0, 1, \ldots, 50$, giving $n_i = 51$. For each of the nine configurations, 500 independent sets of data are generated. We mainly compare the estimates of the bias correction (BC, which is defined as $\mathrm{cAI} = E_{f(\mathbf{y}, \mathbf{u})}\{-2 \log g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}, \hat{\mathbf{b}})\} + 2\mathrm{BC}$) based on our proposed method, $\Phi_0(\mathbf{y})$, and Vaida and Blanchard's (2005) method, $\hat{\rho}$ with the true BC values.

Table 1 summarizes the results of this small simulation study. The results obtained are in accord with the theory. The estimated values based on the proposed method and Vaida and Blanchard's (2005) method are both close to the BC values, and 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 true BC values than those based on the latter, showing that our method is promising.

4. CONCLUDING REMARKS

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

It is worthy of noting that the derivation of (A2) in the Appendix does not require the assumption that the candidate models include the true one. This means that when the error variance σ_0^2 under the true model is known, to derive a reasonable model selection criterion, this traditional assumption is not necessary. Further analysis shows that this conclusion is still true if σ_0^2 is the same as the error variance under the candidate model. Also, the assumption of the true model being included in the candidate model family is needed only in the derivation

of the estimator of σ_0^2 when it is unknown (c.f., Liang et al., 2006). Noting that σ_0^2 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 derivation in the model selection literature, we made use of the integration by part technique, which was used to obtain risk-unbiased estimators before (Stein, 1981; Lu 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 a referee for their constructive comments and suggestions. 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.

References

Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.; Csaki, F., editors. Second International Symposium on Information Theory. Budapest: Akademiai Kiado; 1973. p. 267-81.

Burnham, KP.; Anderson, DP. Model Selection and Inference: A Practical Information-Theoretical Approach. New York: Springer-Verlag; 1998.

Hodges JS, Sargent DJ. Counting degrees of freedom in hierarchical and other richly parameterized models. Biometrika 2001;88:367–79.

Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Statist Soc B 1998;60:271–93.

Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika 1989;76:297–307.

Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982;38:963–74. [PubMed: 7168798]

Liang, H.; Wu, HL.; Zou, GH. Technical report, Department of Biostatistics and Computational Biology, University of Rochester. 2006. General conditional AIC for linear mixed-effects models.

Lu KL, Berger JO. Estimation of normal means: frequentist estimation of loss. Ann Statist 1989;17:890–906.

Naik PA, Tsai CL. Single-index model selections. Biometrika 2001;61:821–32.

Ngo L, Brand R. Model selection in linear mixed effects models using SAS Proc Mixed. SUGI 2002;22

Pinheiro, JC.; Bates, DM. Mixed-Effects Models in S and S-PLUS. New York: Springer; 2000.

Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6:461–4.

Stein CM. Estimation of the mean of a multivariate normal distribution. Ann Statist 1981;9:1135–51.

Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. Biometrika 2005;92:351–70.

Verbeke, G.; Molenberghs, G. Linear Mixed Models for Longitudinal Data. New York: Springer; 2000.

Vonesh, EF.; Chinchilli, VM. Linear and Nonlinear Models for the Analysis of Repeated Measurements. New York: Marcel Dekker, Inc; 1996.

APPENDIX

Proof of Theorem 1

Denote $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{Zu}$, where $\boldsymbol{\beta}_0$ is the true parameter for fixed effects. Then it is readily seen that

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

Also, we have

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

Thus, after some calculations, we obtain

$$\begin{aligned}\text{cAI} - E_{f(\mathbf{y},\mathbf{u})}\{-2\log g(\mathbf{y}|\widehat{\theta},\widehat{\mathbf{b}})\} &= \tfrac{2}{\sigma^2}E_{f(\mathbf{y},\mathbf{u})}\{(\mathbf{y}-\mu)^{\mathrm{T}}\widehat{\mathbf{y}}\}\\ &= \tfrac{2}{\sigma^2}E_{p(\mathbf{u})}E_{f(\mathbf{y}|\mathbf{u})}\left\{\sum_{i=1}^{N}(y_i-\mu_i)\widehat{y}_i\right\},\end{aligned} \tag{A1}$$

where μ_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

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

providing each expectation on the right-hand side exists (see also Stein, 1981; and Lu and Berger, 1989). Therefore, (A1) becomes

$$\begin{aligned}\text{cAI} - E_{f(\mathbf{y},\mathbf{u})}\{-2\log g(\mathbf{y}|\widehat{\theta},\widehat{\mathbf{b}})\} &= 2E_{p(\mathbf{u})}E_{f(\mathbf{y}|\mathbf{u})}\left(\sum_{i=1}^{N}\tfrac{\partial\widehat{y}_i}{\partial y_i}\right)\\ &= 2E_{f(\mathbf{y},\mathbf{u})}\{\Phi_0(\mathbf{y})\}.\end{aligned} \tag{A2}$$

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 σ^2 and $\mathbf{G}$ are known, the fitted vector is

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

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

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

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

$$\Phi_0(\mathbf{y})=\mathrm{tr}\left(\frac{\partial \widehat{\mathbf{y}}^{\mathrm{T}}}{\partial \mathbf{y}}\right)=\mathrm{tr}(\mathbf{H_1})=\rho.$$

Table 1

Simulation study. Comparison of BC and its two estimates, $\hat{\rho}$ and $\Phi_0(\mathbf{y})$ based on 500 runs.

n_i	σ	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