CIRJE-F-575

Conditional Information Criteria for Selecting Variables in Linear Mixed Models

Muni S. Srivastava
University of Toronto

Tatsuya Kubokawa
University of Tokyo

July 2008; Revised in December 2008

Conditional Information Criteria for Selecting Variables in Linear Mixed Models

Muni S. Srivastava*and Tatsuya Kubokawa†
University of Toronto and University of Tokyo

December 24, 2008

Abstract

In this paper, we consider the problem of selecting the variables of the fixed effects in the linear mixed models where the random effects are present and the observation vectors have been obtained from many clusters. As the variable selection procedure, we here use the Akaike Information Criterion, AIC. In the context of the mixed linear models, two kinds of AIC have been proposed: marginal AIC and conditional AIC. In this paper, we derive three versions of conditional AIC depending upon different estimators of the regression coefficients and the random effects. Through the simulation studies, it is shown that the proposed conditional AIC's are superior to the marginal and conditional AIC's proposed in the literature in the sense of selecting the true model. Finally, the results are extended to the case when the random effects in all the clusters are of the same dimension but have a common unknown covariance matrix.

Key words and phrases: Akaike information criterion, analysis of variance, linear mixed model, nested error regression model, random effect, selection of variables.

1 Introduction

Consider the model in which the n_i-vector of response variables $\boldsymbol{y}_i$ in the i-th cluster is related by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{v}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, k, \tag{1}$$

where the k observation vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are independently distributed, $\boldsymbol{X}_i$ is a known $n_i \times p$ matrix, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ is a p-vector of unknown parameters, $\boldsymbol{Z}_i$ is an $n_i \times r_i$,

*Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, E-Mail: srivasta@utstat.toronto.edu

†Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

$n_i \geq r_i$, matrix of known covariables and $\boldsymbol{v}_i$ is an r_i-vector of random effects. The error n_i-vector $\boldsymbol{\epsilon}_i$ is distributed independently of the random effects vector $\boldsymbol{v}_i$, both are assumed to be normally distributed; $\boldsymbol{v}_i \sim \mathcal{N}_{r_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{G}_i)$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_i})$. It is assumed that $\sum_{i=1}^{k}(n_i - r_i) > p$, $\sum_{i=1}^{k} r_i \geq p$, and $\boldsymbol{Z}_i$ are of full rank r_i. Writing $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_k')'$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_k')'$, $\boldsymbol{Z} = \text{block diagonal}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k)$, $\boldsymbol{v} = (\boldsymbol{v}_1', \ldots, \boldsymbol{v}_k')'$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_k')'$, we can express the model in (1) in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{v}$ are independently distributed as $\mathcal{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$ and $\mathcal{N}_R(\boldsymbol{0}, \sigma^2 \boldsymbol{G})$ for

$$R = \sum_{i=1}^{k} r_i, \quad N = \sum_{i=1}^{k} n_i, \quad \boldsymbol{G} = \text{block diag}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_k).$$

The usual goal of the model (2) is to provide a good predicted value for a future observation. Often, it is achieved by reducing the dimension of the parameters, or equivalently by using fewer members of fixed variables than p, either by testing the hypothesis that some specified β_i's are zero or by model selection method such as Akaike Information Criterion (AIC) of Akaike (1973, 1974). Let $f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2)$ and $f(\boldsymbol{v}|\sigma^2)$ be the conditional density of $\boldsymbol{y}$ given $\boldsymbol{v}$ and the marginal density of $\boldsymbol{v}$, respectively. Then, the marginal density of $\boldsymbol{y}$ is written by $f_m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \int f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{v}|\sigma^2) \mathrm{d}\boldsymbol{v}$, which has

$$\mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Lambda}(\boldsymbol{G})) \quad \text{for} \quad \boldsymbol{\Lambda}(\boldsymbol{G}) = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'.$$

When $\boldsymbol{G}$ is known, the Akaike information based on the marginal density $f_m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ is

$$AI = -2 \int \int \{\log f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2)\} f_m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \mathrm{d}\boldsymbol{y}^* \mathrm{d}\boldsymbol{y},$$

where $\boldsymbol{y}^*$ is a future observation having the same distribution as $\boldsymbol{y}$ but independently distributed of $\boldsymbol{y}$, $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and σ^2 based on the observation $\boldsymbol{y}$ where the marginal distribution is given above. The expressions for the MLE $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are given in (5) and (6), respectively. Selection of the variables of the fixed effects $\boldsymbol{X}$ is based on the minimum value of an unbiased estimator of AI. In (7), an exact unbiased estimate of AI, which we denote by AIC is given.

Another interesting approach, proposed by Vaida and Blanchard (2005), is based on the so-called conditional Akaike information given by

$$cAI = -2 \int \int \int \log\{f(\boldsymbol{y}^*|\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}}, \hat{\sigma}^2)\} f(\boldsymbol{y}^*|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{v}|\sigma^2) \mathrm{d}\boldsymbol{y}^* \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{v}, \tag{3}$$

where $\widehat{\boldsymbol{v}}$ is the empirical Bayes estimator of $\boldsymbol{v}$ given in (8) and $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are given in (5) and (6), respectively. Vaida and Blanchard (2005) derived an unbiased estimator of cAI given by (9). We will denote this estimator by $cAIC$. Vaida and Blanchard have observed that

when $\boldsymbol{G}$ is known $cAIC$ takes the same value as DIC, the deviance information criterion proposed by Spiegelhalter, Best, Carlin and van der Linde (2002) for Bayesian inference.

In our simulation results, we, however, find that the performance of $cAIC$ is no better than the marginal $mAIC$. Thus, the performance of both of them is not good. This may be due to the fact that the estimate $\hat{\sigma}^2(\boldsymbol{G})$ ignores the existence of the random effects although it provides the largest degrees of freedom availale to estimate σ^2. But it also makes it necessary to obtain cAI marginally, that is averaging with respect to the density of $\boldsymbol{v}$. Thus, we consider estimating σ^2 from the conditional model, where

$$\boldsymbol{y}|\boldsymbol{v} \sim \mathcal{N}_N(\boldsymbol{W}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}),$$

where $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{v}')'$. Hence, we need to obtain the bias term only conditionally. That is, the conditional Akaike information we consider is

$$CAI(\boldsymbol{v}) = -2\int\int \log\{f(\boldsymbol{y}^*|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2)\} f(\boldsymbol{y}^*|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2) \mathrm{d}\boldsymbol{y}^* \mathrm{d}\boldsymbol{y}, \tag{4}$$

where $\hat{\sigma}_0 = (\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\gamma}}_0)'(\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\gamma}}_0)/N$ for the least squares estimator $\widehat{\boldsymbol{\gamma}}_0$ of $\boldsymbol{\gamma}$, and $\widetilde{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{\beta}}$ are linear estimators of $\boldsymbol{y}$ for $\boldsymbol{v}$ and $\boldsymbol{\beta}$. It is noted that $\hat{\sigma}_0^2$ is not affected by the random effect term $\boldsymbol{v}$ and $N\hat{\sigma}_0^2/\sigma^2$ has a chisquare distribution with $(N - p - R)$-degrees of freedom. The proposed new conditional AIC is an unbiased estimator of CAI given by

$$CAIC(\boldsymbol{G}) = -2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2) + 2\frac{N(\mathrm{tr}\,[\boldsymbol{W}\boldsymbol{L}] + 1)}{N - r_w - 2},$$

where $r_w = \mathrm{rank}\,(\boldsymbol{W})$ and $\boldsymbol{L}$ is a $(p + R) \times N$ matrix such that $(\widetilde{\boldsymbol{\beta}}', \widetilde{\boldsymbol{v}}')' = \boldsymbol{L}\boldsymbol{y}$. It is interesting to note that although this is obtained conditionally given $\boldsymbol{v}$, the bias term $-2N(\mathrm{tr}\,[\boldsymbol{W}\boldsymbol{L}] + 1)/(N - r_w - 2)$ does not depend on $\boldsymbol{v}$, and thus, it is also an unbiased estimator of cAI defined by Vaida and Blanchard (2005). Although $\hat{\sigma}_0^2$ is based on less degrees of freedom than $\hat{\sigma}^2$, we argue that it is the most appropriate estimate of σ^2 to use in $-2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}}, \hat{\sigma}_0^2)$ as it takes into account that not only $\boldsymbol{\beta}$ but $\boldsymbol{v}$ has also been estimated while $\hat{\sigma}^2$ does not take this into account; in fact in the conditional framework $\hat{\sigma}^2$ is a biased estimator of σ^2. Simulation experiments carried out in Section 4 show that the performance of $CAIC(\boldsymbol{G})$ is significantly better than $mAIC(\boldsymbol{G})$ and $cAIC(\boldsymbol{G})$ in the sense of selecting the true model. This method also facilitates to consider the case when $\boldsymbol{G}$ is a function of unknown parameters.

The organization of the paper is as follows. In Section 2, we introduce the concept of AIC and marginal and conditional AIC's. In Section 3, we derive three conditional AIC using three different kinds of estimators available for $(\boldsymbol{\beta}, \boldsymbol{v})$, assuming that $\boldsymbol{G}$ is known. The case of unknown $\boldsymbol{G}$ is considered in Section 4. We give simulation results in Section 5. The paper concludes in Section 6. The proof is given in the Appendix.

2 Marginal and Conditional AIC in the linear mixed model

2.1 Concept of AIC

We now introduce the marginal and conditional AIC's in the linear mixed model in the case of known $\boldsymbol{G}$. Before describing them, we first explain the concept of AIC briefly. The AIC is based on Kullback-Leibler distance. For a true density f and an approximating one g_ω, this distance is defined as

$$I(f, g_\omega) = E_{y^*} \log f(y^*) - E_f \log g_\omega(y^*),$$

where E_{y^*} denotes the expectation with repect to the true density $f(y^*)$. Let $\mathcal{G} = \{g_\omega : \omega \in \Omega\}$ be the class of approximating densities. If $f \in \mathcal{G}$, then there exist a $g_{\omega_0} \in \mathcal{G}$ such that $I(f, g_{\omega_0}) = 0$, otherwise $I(f, g_{\omega_0}) \geq 0$. Thus a g is chosen for which $I(f, g)$ is minimum. Usually ω is not known and estimated from the data y by $\hat{\omega} = \hat{\omega}(y)$. Thus, $I(f, g_\omega)$ is approximated by $I(f, g_{\hat{\omega}})$ and the quality of approximation is judged by

$$E_y f[I(f, g_{\hat{\omega}})] = E_{y^*}[\log f(y^*)] - E_y E_{y^*}[\log g_{\hat{\omega}(y)}(y^*)],$$

where E_y denotes the expectation with respect to the true density $f(y)$, which is independent of y^*. Akaike information is defined by

$$AI = -2E_y E_{y^*}[\log g_{\hat{\omega}(y)}(y^*)].$$

An unbiased estimator of AI is given by

$$AIC = -2 \log g_{\hat{\omega}(y)}(y) + \Delta,$$

where Δ is the bias

$$\Delta = E_y[-2 \log g_{\hat{\omega}(y)}(y)] - AI.$$

Akaike (1973, 1974) used an approximate value of the bias given by the number of free parameters. Thus, Akaike used the number of free parameters in place of Δ. It is noted that AIC is a criterion for selecting a good model in terms of minimizing the prediction error. It may be noted that the estimator $\hat{\omega}$ of ω need not be an MLE as any consistent estimator of ω may perform as good, see Konishi and Kitagawa (2007).

2.2 Marginal AIC

The marginal AIC in the linear mixed model is AIC based on the marginal distribution

$$\boldsymbol{y} \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Lambda}),$$

for $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{G}) = \boldsymbol{I}_N + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$, where the marginal density is given by

$$f_m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-N/2} |\boldsymbol{\Lambda}|^{-1/2} \exp\{-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Lambda}^{-1} (\boldsymbol{y} - X\boldsymbol{\beta})/(2\sigma^2)\}.$$

When $\boldsymbol{G}$ does not include unknown parameters, the maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and σ^2 are given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}} =&\widehat{\boldsymbol{\beta}}(\boldsymbol{G}) = (\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{y}, \qquad (5)\\ \hat{\sigma}^2 =&\hat{\sigma}^2(\boldsymbol{G}) = (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'\Lambda^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})/N\\ =&\boldsymbol{y}'\left\{\boldsymbol{\Lambda}^{-1}-\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{-}\boldsymbol{\Lambda}^{-1}\right\}\boldsymbol{y}, \qquad (6)\end{aligned}$$

where $\boldsymbol{A}^{+}$ denotes the Moore-Penrose inverse of $\boldsymbol{A}$ and satisfies (i) $\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A}=\boldsymbol{A}$, (ii) $\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+}=\boldsymbol{A}^{+}$, (iii) $(\boldsymbol{A}^{+}\boldsymbol{A})'=(\boldsymbol{A}^{+}\boldsymbol{A})$, (iv) $(\boldsymbol{A}\boldsymbol{A}^{+})'=\boldsymbol{A}\boldsymbol{A}^{+}$. Then, $-2\log f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}},\hat{\sigma}^2)$ is expressed as

$$-2\log f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}},\hat{\sigma}^2) = N[\log(2\pi\hat{\sigma}^2(\boldsymbol{G}))+1]+\log|\boldsymbol{\Lambda}|,$$

and the exact bias correction AIC based on the marginal likelihood is given by

$$mAIC(\boldsymbol{G}) = N[\log(2\pi\hat{\sigma}^2(\boldsymbol{G}))+1]+\log|\boldsymbol{\Lambda}|+2N(r_x+1)/(N-r_x-2), \qquad (7)$$

where $r_x=\mathrm{rank}\,(\boldsymbol{X})$. When $\boldsymbol{G}$ includes unknown parameters, we can use the criteria $AIC(\widehat{\boldsymbol{G}})$ and $mAIC(\widehat{\boldsymbol{G}})$ when a consistent estimator $\widehat{\boldsymbol{G}}$ of $\boldsymbol{G}$ is available.

2.3 Conditional AIC

The conditional AIC in the linear mixed model was proposed by Vaida and Blanchard (2005), who considered estimating the random effects $\boldsymbol{v}$ as well as $\boldsymbol{\beta}$ by the mixed model equation

$$\begin{pmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{v}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}'\\ \boldsymbol{Z}'\end{pmatrix}\boldsymbol{y},$$

which was given by Henderson (1950). Then the solution $\widehat{\boldsymbol{\beta}}$ is the generalized least squares estimator (GLS) given in (5), and $\widehat{\boldsymbol{v}}$ is given by

$$\widehat{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Lambda}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}), \qquad (8)$$

which can also be derived as an empirical Bayes estimator by considering the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$. Thus, using the estimator $\hat{\sigma}^2(\boldsymbol{G})$ defined in (6), they define the conditional Akaike information by (3). The conditional AIC, denoted by $cAIC(\boldsymbol{G})$, is given by

$$cAIC(\boldsymbol{G}) = -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}},\widehat{\boldsymbol{\beta}},\hat{\sigma}^2)-\Delta_c, \qquad (9)$$

where

$$\begin{aligned}\Delta_c =&E_{\boldsymbol{y}}[-2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}},\widehat{\boldsymbol{\beta}},\hat{\sigma}^2)]-cAI\\ =&-2\frac{N(N-r_x-1)}{(N-r_x)(N-r_x-2)}(\rho+1)+\frac{N(r_x+1)}{(N-r_x)(N-r_x-2)}.\end{aligned}$$

Here, ρ is defined by $\rho = \mathrm{tr}\,(\boldsymbol{H}_1)$, where

$$\boldsymbol{H}_1 = (\boldsymbol{X}, \boldsymbol{Z}) \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{pmatrix}.$$

which can be also expressed as

$$\begin{aligned} \boldsymbol{H}_1 =& \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{\Lambda}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Lambda}^{-1}\left\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\right\} \\ =& \boldsymbol{\Lambda}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{\Lambda}^{-1} + \boldsymbol{I} - \boldsymbol{\Lambda}^{-1}. \end{aligned} \tag{10}$$

3 Proposed conditional AIC

In our simulation results, we find that the performance of $cAIC(\boldsymbol{G})$ is only slightly better than the marginal $mAIC(\boldsymbol{G})$ in the sense of selecting true models. To improve the performance, we here propose another type of conditional AIC's. In this section, we assume that $\boldsymbol{G}$ is known, and we begin with rewriting the model (2) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon} \equiv \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$ is an $N \times (p+R)$ matrix and $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{v}')'$ is a $(p+R)$-dimensional vector. Given $\boldsymbol{v}$, the model $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ can be conditionally regarded as a usual linear regression model, and the conditional MLE of $\boldsymbol{\gamma}$ and σ^2 are given by

$$\widehat{\boldsymbol{\gamma}}_0 = (\boldsymbol{W}'\boldsymbol{W})^{+}\boldsymbol{W}'\boldsymbol{y}, \tag{11}$$

$$\hat{\sigma}_0^2 = (\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\gamma}}_0)'(\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\gamma}}_0)/N. \tag{12}$$

It can be shown that the MLE of $\boldsymbol{\gamma}$ can be expressed as

$$\begin{aligned} \widehat{\boldsymbol{\gamma}}_0 = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_0 \\ \widehat{\boldsymbol{v}}_0 \end{pmatrix} =& \begin{pmatrix} (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{+}\widetilde{\boldsymbol{X}}' \\ (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{I} - \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{+}\widetilde{\boldsymbol{X}}') \end{pmatrix} \boldsymbol{y} \\ \equiv& \boldsymbol{L}_0\boldsymbol{y}, \end{aligned}$$

where $\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{H}_z)\boldsymbol{X}$ for $\boldsymbol{H}_z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. It is noted that the MLE of $\boldsymbol{\gamma}$ is a linear function of $\boldsymbol{y}$. Thus, we may consider a general estimator of $\boldsymbol{\gamma}$ as a linear function of $\boldsymbol{y}$, namely,

$$\widetilde{\boldsymbol{\gamma}} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}} \\ \widetilde{\boldsymbol{v}} \end{pmatrix} = \boldsymbol{L}\boldsymbol{y},$$

where $\boldsymbol{L}$ is a $(p+R) \times N$ matrix which will be specified later. It is noted that $N\hat{\sigma}_0^2/\sigma^2$ has a chisquare distribution with $(N - r_w)$ degrees of freedom for $r_w = \mathrm{rank}\,(\boldsymbol{W})$. Thus, we shall obtain our $CAIC$ using $\widetilde{\boldsymbol{\gamma}}$ and $\hat{\sigma}_0^2$ as estimators of $\boldsymbol{\gamma}$ and σ^2, respectively.

The conditional Akaike information considered here is given in (4), which is different from cAI given in (3) in that $cAI = \int CAI(\boldsymbol{v}) f(\boldsymbol{v}|\sigma^2)\mathrm{d}\boldsymbol{v}$. Since $-2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2)$ is written as

$$-2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2) = N\log(2\pi\hat{\sigma}_0^2) + (\boldsymbol{y} - \boldsymbol{W}\widetilde{\boldsymbol{\gamma}})'(\boldsymbol{y} - \boldsymbol{W}\widetilde{\boldsymbol{\gamma}})/\hat{\sigma}_0^2.$$

We now consider another random vector $\boldsymbol{y}^*$ distributed independently of $\boldsymbol{y}$ but having the same distribution as $\boldsymbol{y}$. Hence,

$$\boldsymbol{y}^*|\boldsymbol{v} \sim \mathcal{N}_N(\boldsymbol{W}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}).$$

Thus,

$$\begin{aligned} E_{\boldsymbol{y}^*}[-2\log f(\boldsymbol{y}^*|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2)|\boldsymbol{v}] =& N\log(2\pi\hat{\sigma}_0^2) + E_{\boldsymbol{y}^*}[(\boldsymbol{y}^* - \boldsymbol{W}\widetilde{\boldsymbol{\gamma}})'(\boldsymbol{y}^* - \boldsymbol{W}\widetilde{\boldsymbol{\gamma}})/\hat{\sigma}_0^2|\boldsymbol{v}] \\ =& N\log(2\pi\hat{\sigma}_0^2) + N\frac{\sigma^2}{\hat{\sigma}_0^2} + \frac{(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{W}'\boldsymbol{W}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\hat{\sigma}_0^2}. \end{aligned}$$

Hence, the conditional Akaike information given $\boldsymbol{v}$ is given by

$$CAI(\boldsymbol{v}) = E_{\boldsymbol{y}}\left[N\log(2\pi\hat{\sigma}_0^2) + N\frac{\sigma^2}{\hat{\sigma}_0^2} + \frac{(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})'\boldsymbol{W}'\boldsymbol{W}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\hat{\sigma}_0^2}\Big|\boldsymbol{v}\right],$$

and the bias is expressed as

$$\begin{aligned} \Delta_C(\boldsymbol{v}) =& E_{\boldsymbol{y}}\left[-2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2)\Big|\boldsymbol{v}\right] - CAI(\boldsymbol{v}) \\ =& E_{\boldsymbol{y}}\left[\frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})}{\hat{\sigma}_0^2} - 2\frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'\boldsymbol{W}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\hat{\sigma}_0^2} - N\frac{\sigma^2}{\hat{\sigma}_0^2}\Big|\boldsymbol{v}\right]. \end{aligned} \tag{13}$$

It is here noted that $E_{\boldsymbol{y}}[\sigma^2/\hat{\sigma}_0^2|\boldsymbol{v}] = N/(N - r_w - 2)$ and $E_{\boldsymbol{y}}[(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})/\hat{\sigma}_0^2|\boldsymbol{v}] = N + Nr_w/(N - r_w - 2)$. Also for $\widetilde{\boldsymbol{\gamma}} = \boldsymbol{L}\boldsymbol{y}$, it is observed that

$$\begin{aligned} E_{\boldsymbol{y}}\left[\frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'\boldsymbol{W}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\hat{\sigma}_0^2}\Big|\boldsymbol{v}\right] =& E_{\boldsymbol{y}}\left[\frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'\boldsymbol{W}\boldsymbol{L}\boldsymbol{y}}{\hat{\sigma}_0^2}\Big|\boldsymbol{v}\right] \\ =& \frac{1}{2}E_{\boldsymbol{y}}\left[\frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})'(\boldsymbol{W}\boldsymbol{L} + \boldsymbol{L}'\boldsymbol{W}')(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})}{\hat{\sigma}_0^2}\Big|\boldsymbol{v}\right] \\ =& \frac{N}{2}E_{\boldsymbol{y}}\left[\frac{\boldsymbol{u}'(\boldsymbol{W}\boldsymbol{L} + \boldsymbol{L}'\boldsymbol{W}')\boldsymbol{u}}{\boldsymbol{u}'(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{u}}\Big|\boldsymbol{v}\right], \end{aligned}$$

where $\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma})/\sigma$ and $\boldsymbol{H}_w = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'$. Note that given $\boldsymbol{v}$, the conditional distribution of $\boldsymbol{u}$ has $\mathcal{N}_N(\boldsymbol{0}, \boldsymbol{I})$, so that $\boldsymbol{u}$ is independent of $\boldsymbol{v}$. Using Lemma A.1 in the Appendix, we can see that

$$\begin{aligned} \frac{N}{2}E\left[\frac{\boldsymbol{u}'(\boldsymbol{W}\boldsymbol{L} + \boldsymbol{L}'\boldsymbol{W}')\boldsymbol{u}}{\boldsymbol{u}'(\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{u}}\Big|\boldsymbol{v}\right] =& N\left\{\frac{\mathrm{tr}\,(\boldsymbol{W}\boldsymbol{L})}{N - r_w - 2} - \frac{2\mathrm{tr}\,[\boldsymbol{W}\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{H}_w)]}{(N - r_w)(N - r_w - 2)}\right\} \\ =& \frac{N\mathrm{tr}\,(\boldsymbol{W}\boldsymbol{L})}{N - r_w - 2}, \end{aligned}$$

since $\mathrm{tr}\,[\boldsymbol{W}\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{H}_w)] = 0$. Combining these evaluations gives the following expression for the bias term $\Delta_C(\boldsymbol{v})$:

$$\Delta_C(\boldsymbol{v}) = -2\frac{N(\mathrm{tr}\,[\boldsymbol{W}\boldsymbol{L}] + 1)}{N - r_w - 2}, \tag{14}$$

which yields the conditional AIC given by

$$CAIC(\boldsymbol{G}) = -2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}, \hat{\sigma}_0^2) + 2\frac{N(\mathrm{tr}\,[\boldsymbol{WL}]+1)}{N-r_w-2}. \tag{15}$$

It is interesting to note that the bias term $\Delta_C(\boldsymbol{v})$ does not depend on the given $\boldsymbol{v}$. Given $\boldsymbol{v}$, $CAIC$ is an unbiased estimator of CAI given in (4), which turns out to be an unbiased estimator of the conditional Akaike information cAI defined by Vaida and Blanchard (2005).

The matrix $\boldsymbol{L}$ given in (15) depends on the choice of $\widehat{\boldsymbol{\gamma}}$ as an estimator of $\boldsymbol{\gamma}$. We consider three kinds of estimators of $\boldsymbol{\gamma}$ as described below.

[1] Using maximum likelihood estimator $\widehat{\boldsymbol{\gamma}}_0$ for $\boldsymbol{\gamma}$. Since the maximum likelihood or lease squares estimator of $\boldsymbol{\gamma}$ is given in (11) as $\widehat{\boldsymbol{\gamma}}_0 = (\boldsymbol{W}'\boldsymbol{W})^+\boldsymbol{W}'\boldsymbol{y}$, the matrix $\boldsymbol{L}$ corresponds to $\boldsymbol{L} = (\boldsymbol{W}'\boldsymbol{W})^+\boldsymbol{W}'$. Hence, it is seen that $\mathrm{tr}\,[\boldsymbol{WL}] = \mathrm{tr}\,[(\boldsymbol{W}'\boldsymbol{W})^+(\boldsymbol{W}'\boldsymbol{W})] = \mathrm{rank}\,(\boldsymbol{W}) = r_w$, and we get the conditional AIC based on MLE $\widehat{\boldsymbol{\gamma}}_0 = (\widehat{\boldsymbol{\beta}}_0', \widehat{\boldsymbol{v}}_0')'$ as

$$\begin{aligned} CAIC_{ML}(\boldsymbol{G}) =& -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}_0, \widehat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2) + 2N(r_w+1)/(N-r_w-1) \\ =& N[\log(2\pi\hat{\sigma}_0^2)+1] + 2N(r_w+1)/(N-r_w-1). \end{aligned} \tag{16}$$

[2] Using empirical Bayes (EB) estimator of $\boldsymbol{v}$ and simple estimator of $\boldsymbol{\beta}$. Consider the case that the empirical Bayes estimator is used for $\boldsymbol{v}$ and the simple estimator $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}$ is utilized for $\boldsymbol{\beta}$. The empirical Bayes estimator of $\boldsymbol{v}$ is given in (8) as $\widehat{\boldsymbol{v}} = \boldsymbol{GZ}'\boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, and the matrix $\mathrm{tr}\,[\boldsymbol{WL}]$ corresponds to

$$\mathrm{tr}\,[\boldsymbol{WL}] = \mathrm{tr}\,\left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}' + \boldsymbol{ZGZ}'\boldsymbol{\Lambda}^{-1}\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^+\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\}\right].$$

Note that $\mathrm{tr}\,[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'] = \mathrm{tr}\,[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{\Lambda}^{-1}] = \mathrm{rank}\,(\boldsymbol{X})$. From the argument around (10), it is seen that $\mathrm{tr}\,[\boldsymbol{WL}] = \mathrm{tr}\,[\boldsymbol{H}_1] = \rho$. Hence, we get the conditional AIC based on the estimator $(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{v}}')'$ of $\boldsymbol{\gamma}$ as

$$\begin{aligned} CAIC_{SL}(\boldsymbol{G}) =& -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}}_1, \hat{\sigma}_0^2) + 2N(\rho+1)/(N-r_w-2) \\ =& N\log(2\pi\hat{\sigma}_0^2) + \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{Z}\widehat{\boldsymbol{v}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{Z}\widehat{\boldsymbol{v}})}{\hat{\sigma}_0^2} + \frac{2N(\rho+1)}{N-r_w-2}. \end{aligned} \tag{17}$$

[3] Using generalized least squares estimator of $\boldsymbol{\beta}$ and EB estimator of $\boldsymbol{v}$. Consider the generalized least squares estimator $\widehat{\boldsymbol{\beta}}$ and the EB estimator $\widehat{\boldsymbol{v}}$ given in (5) and (8), respectively. These are estimators treated by Vaida and Blanchard (2005), and it is seen that $\mathrm{tr}\,[\boldsymbol{WL}] = \rho$ since $\boldsymbol{WL} = \boldsymbol{H}_1$. Hence, we get the conditional AIC based on the estimator $(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{v}}')'$ of $\boldsymbol{\gamma}$ as

$$\begin{aligned} CAIC_{GL}(\boldsymbol{G}) =& -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{\beta}}, \hat{\sigma}_0^2) + 2N(\rho+1)/(N-r_w-2) \\ =& N\log(2\pi\hat{\sigma}_0^2) + \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{v}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{v}})}{\hat{\sigma}_0^2} + \frac{2N(\rho+1)}{N-r_w-2}. \end{aligned} \tag{18}$$

4 Extensions to the case of unknown variance components

We have explained the marginal and conditional AIC's under the assumption that $\boldsymbol{G}$ is known. In most applications, however, $\boldsymbol{G}$ depends on unknown parameters. In this section, we handle the case of $\boldsymbol{G}$ including unknown parameters. If a consistent estimator $\widehat{\boldsymbol{G}}$ is available for $\boldsymbol{G}$, then it can be substituted into $mAIC(\boldsymbol{G})$, $cAIC(\boldsymbol{G})$ and $CAIC(\boldsymbol{G})$ to get the marginal AIC $mAIC(\widehat{\boldsymbol{G}})$, the conditional AIC's $cAIC(\widehat{\boldsymbol{G}})$ and $CAIC(\widehat{\boldsymbol{G}})$, which will be suggested in this paper. In this case, the problem is how to estimate $\boldsymbol{G}$. The maximum likelihood method is an approach, but we need heavy computation as well as convergence of numerical iterations. Thus, in this section, we provide estimators of $\boldsymbol{G}$ in explicit forms in some specific models.

For the model given in (2), we begin with making the transformation

$$\begin{pmatrix} \tilde{\boldsymbol{y}}_{1i} \\ \tilde{\boldsymbol{y}}_{2i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_i \\ (\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i' \end{pmatrix} \boldsymbol{y}_i, \tag{19}$$

where $\boldsymbol{\Gamma}_i$ is an $(n_i - r_i) \times n_i$ matrix such that $\boldsymbol{\Gamma}_i\boldsymbol{Z}_i = \boldsymbol{0}$ and $\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i' = \boldsymbol{I}_{n_i - r_i}$. It may be noted that $\boldsymbol{I} - \boldsymbol{Z}_i(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i' = \boldsymbol{\Gamma}_i'\boldsymbol{\Gamma}_i$ and thus $\boldsymbol{\Gamma}_i$ can easily be computed. According to this transformation, let $\widetilde{\boldsymbol{X}}_{1i} = \boldsymbol{\Gamma}_i\boldsymbol{X}_i$, $\tilde{\boldsymbol{\epsilon}}_{1i} = \boldsymbol{\Gamma}_i\boldsymbol{\epsilon}_i$, $\widetilde{\boldsymbol{X}}_{2i} = (\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i'\boldsymbol{X}_i$ and $\tilde{\boldsymbol{\epsilon}}_{2i} = (\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i'\boldsymbol{\epsilon}_i$ for $i = 1, \ldots, k$. For $j = 1, 2$, let $\tilde{\boldsymbol{y}}_j = (\tilde{\boldsymbol{y}}_{j1}', \ldots, \tilde{\boldsymbol{y}}_{jk}')'$, $\widetilde{\boldsymbol{X}}_j = (\widetilde{\boldsymbol{X}}_{j1}', \ldots, \widetilde{\boldsymbol{X}}_{jk}')'$ and $\tilde{\boldsymbol{\epsilon}}_j = (\tilde{\boldsymbol{\epsilon}}_{j1}', \ldots, \tilde{\boldsymbol{\epsilon}}_{jk}')'$. Then, the model is decomposed as

$$\begin{aligned} \tilde{\boldsymbol{y}}_1 =& \widetilde{\boldsymbol{X}}_1\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_1, \\ \tilde{\boldsymbol{y}}_2 =& \widetilde{\boldsymbol{X}}_2\boldsymbol{\beta} + \boldsymbol{v} + \tilde{\boldsymbol{\epsilon}}_2, \end{aligned} \tag{20}$$

where $\tilde{\boldsymbol{y}}_1 : (N - R) \times 1$, $\tilde{\boldsymbol{y}}_2 : R \times 1$, $\widetilde{\boldsymbol{X}}_1 : (N - R) \times p$, $\widetilde{\boldsymbol{X}}_2 : R \times p$, $N = \sum_{i=1}^k n_i$ and $R = \sum_{i=1}^k r_i$. Let $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{y}}_1', \tilde{\boldsymbol{y}}_2')'$ and $\boldsymbol{W}_2 = \text{block diag}(\boldsymbol{G}_i + (\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}); i = 1, \ldots, k)$. Then, $\tilde{\boldsymbol{y}}_1$ and $\tilde{\boldsymbol{y}}_2$ are mutually independently distributed as

$$\begin{aligned} \tilde{\boldsymbol{y}}_1 \sim& \mathcal{N}_{N-R}(\widetilde{\boldsymbol{X}}_1\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{N-R}), \\ \tilde{\boldsymbol{y}}_2 \sim& \mathcal{N}_R(\widetilde{\boldsymbol{X}}_2\boldsymbol{\beta}, \sigma^2\boldsymbol{W}_2). \end{aligned}$$

From the marginal likelihood of $\tilde{\boldsymbol{y}}_1$, we get an unbiased estimator of σ^2 given by

$$\tilde{\sigma}_1^2 = \tilde{\boldsymbol{y}}_1' \left\{ \boldsymbol{I}_M - \widetilde{\boldsymbol{X}}_1(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}_1' \right\} \tilde{\boldsymbol{y}}_1 / (N - R - r_{(\widetilde{\boldsymbol{X}}_1)}), \tag{21}$$

where $r_{(\widetilde{\boldsymbol{X}}_1)} = \text{rank}\,(\widetilde{\boldsymbol{X}}_1)$. It can be seen that $\tilde{\sigma}_1^2$ is a consistent estimator of σ^2.

We now consider the estimation of $\boldsymbol{G}$. For this purpose, we handle two specific cases: (1) $\boldsymbol{G}_i = \psi\boldsymbol{D}_i$ for unknown scalar ψ and known matrix $\boldsymbol{D}_i$ and (2) $\boldsymbol{G}_i = \boldsymbol{\Psi}$ and $r_i = r$ for $i = 1, \ldots, k$.

[1] Case of $\boldsymbol{G}_i = \psi \boldsymbol{D}_i$ for unknown scalar ψ and known matrix $\boldsymbol{D}_i$. In this case, the marginal distribution of $\boldsymbol{y}$ has $\mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = \boldsymbol{I} + \psi \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$ for $\boldsymbol{D} = \text{blockdiag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k)$. Let $S = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{y}$ for $\boldsymbol{H}_x = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'$, and the expectation is written as

$$E[S] = \sigma^2 \text{tr}\,[(\boldsymbol{I} - \boldsymbol{H}_x)(\boldsymbol{I} + \psi \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}')] = \sigma^2\{N - r_x + \psi \text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}]\},$$

for $r_x = \text{rank}\,(\boldsymbol{X})$. Thus, we get an estimator of ψ given by

$$\widehat{\psi} = \{S/\tilde{\sigma}_1^2 - (N - r_x)\}/\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}],$$

for $\tilde{\sigma}_1^2$ given in (21).

For the consistency of $\widehat{\psi}$, we assume that $\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}] = O(k)$ for large k. From the consistency of $\tilde{\sigma}_1^2$ and the fact that $S/\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}] = O_p(1)$, it follows that $\widehat{\psi} - \psi = \{S/\sigma^2 - (N - r_x)\}/\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}] - \psi + O_p(k^{-1/2})$. It can be seen that $Var[\widehat{\psi}] = Var[S]/\{\sigma^4(\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}])^2\} + O(k^{-1})$ and that $Var[S] = \sigma^4 2\text{tr}\,[\{(\boldsymbol{I} - \boldsymbol{H}_x)(\boldsymbol{I} + \psi \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}')\}^2] = O(k)$. Hence, $Var[\widehat{\psi}] = O(k^{-1})$ under the assumption that $\text{tr}\,[\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{Z}\boldsymbol{D}] = O(k)$, which implies the consistency of $\widehat{\psi}$.

The estimator $\widehat{\psi}$ can take negative values with positive probability. Thus, we consider a truncated estimator

$$\widehat{\psi}^{TR} = \max\{\widehat{\psi}, R^{-2/3}\}, \tag{22}$$

which can be shown to be positive and consistent as $k \to \infty$.

[2] Case of $\boldsymbol{G}_i = \boldsymbol{\Psi}$, an unknown matrix, and $r_i = r$ for $i = 1, \ldots, k$. In this case, we recall the transformed model given in (20). Let $\boldsymbol{S}_i = (\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\widetilde{\boldsymbol{\beta}}_1)(\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\widetilde{\boldsymbol{\beta}}_1)'$ for $\widetilde{\boldsymbol{\beta}}_1 = (\widetilde{\boldsymbol{X}}'_1\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}'_1\boldsymbol{y}$. It is noted that $E[\boldsymbol{S}_i] = \sigma^2\{(\boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1} + \boldsymbol{\Psi} + \widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}'_1\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}'_{2i}\}$, which yields that $\sum_{i=1}^k E[\boldsymbol{S}_i] = \sigma^2\sum_{i=1}^k\{(\boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1} + \widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}'_1\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}'_{2i}\} + \sigma^2 k\boldsymbol{\Psi}$. Thus, $\boldsymbol{\Psi}$ can be estimated by

$$\widehat{\boldsymbol{\Psi}}^U = \frac{1}{k\tilde{\sigma}_1^2}\sum_{i=1}^k \boldsymbol{S}_i - \frac{1}{k}\sum_{i=1}^k \left\{(\boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1} + \widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}'_1\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}'_{2i}\right\}, \tag{23}$$

for $\tilde{\sigma}_1^2$ given in (21). It can be verified that $\widehat{\boldsymbol{\Psi}}^U$ is consistent. For the proof, see the Appendex.

For the covariance matrix $\textbf{Cov}\,(\boldsymbol{v}_i) = \sigma^2\boldsymbol{\Psi}$, we consider the two cases of $\boldsymbol{\Psi}$; $\boldsymbol{\Psi}$ is fully unknown, and $\boldsymbol{\Psi} = \text{diag}\,(\psi_1, \ldots, \psi_r)$. When $\boldsymbol{\Psi}$ is a fully unknown covariance matrix, $\widehat{\boldsymbol{\Psi}}$ should be estimated by a positive definite matrix. Let $\boldsymbol{P}$ be an orthogonal matrix such that $\widehat{\boldsymbol{\Psi}}^U = \boldsymbol{P}'\text{diag}\,(\omega_1, \ldots, \omega_m)\boldsymbol{P}$ where $\omega_1, \ldots, \omega_r$ are eigenvalues of $\widehat{\boldsymbol{\Psi}}^U$. Then, we can use the truncated estimator

$$\widehat{\boldsymbol{\Psi}}^{TR} = \boldsymbol{P}'\,\text{diag}\,\left(\max\{\omega_1, R^{-2/3}\}, \ldots, \max\{\omega_r, R^{-2/3}\}\right)\boldsymbol{P},$$

for $R = kr$. It can be shown that $\widehat{\boldsymbol{\Psi}}^{TR}$ is consistent with order $\widehat{\boldsymbol{\Psi}}^{TR} - \boldsymbol{\Psi} = O_p(k^{-1/2})$.

When $\boldsymbol{\Psi}$ has a covariance structure, we can use the structure to construct an appropriate estimator based on $\widehat{\boldsymbol{\Psi}}^{U}$. For instance, assume that $\boldsymbol{\Psi} = \mathrm{diag}\,(\psi_1, \ldots, \psi_r)$. This case implies that $\boldsymbol{Z}_i\boldsymbol{v}_i$ in (1) is expressed as $\boldsymbol{Z}_i\boldsymbol{v}_i = \boldsymbol{z}_{i1}v_{i1} + \cdots + \boldsymbol{z}_{ir}v_{ir}$ where $v_{ij} \sim \mathcal{N}(0, \sigma^2\psi_j)$ and $\boldsymbol{z}_{ij}$ is an $n_i \times 1$ vector for $j = 1, \ldots, r$. This model may be useful when several factors of random effects are considered in practical situations. Then, we can estimate each ψ_i by

$$\widehat{\psi}_i^{TR} = \max\{(\widehat{\boldsymbol{\Psi}}^{U})_{ii}, R^{-2/3}\}, \tag{24}$$

where $(\widehat{\boldsymbol{\Psi}}^{U})_{ii}$ denotes the (i, i) diagonal element of $\widehat{\boldsymbol{\Psi}}^{U}$. The resulting estimator of $\boldsymbol{\Psi}$ is given by $\widehat{\boldsymbol{\Psi}}^{TR} = \mathrm{diag}\,(\widehat{\psi}_1^{TR}, \ldots, \widehat{\psi}_r^{TR})$.

5 Simulation Studies

We now investigate the numerical performances of the marginal and the conditional AIC's derived in the previous sections through simulation and compare them in terms of the frequencies of selecting the true model.

The simulation experiments have been carried out for $k = 20$, $r_1 = \cdots = r_{20} = r = 1, 2, 3$ and $p = 7$. For the sample sizes n_i's, n_i's are generated as $n_i = 1 + Bin(8, 1/2)$ for $i = 1, \ldots, k$, where $Bin(8, 1/2)$ is a random variable distributed as a binomial distribution with mean 4 and success probability $1/2$. For the $N \times p$ matrix $\boldsymbol{X}$ of the regressor variables in the model (2), the row vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ for $\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ are generated as mutually independent random variables distributed as $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_x)$ where $\boldsymbol{\Sigma}_x = (1-\rho_x)\boldsymbol{I}_p + \rho_x\boldsymbol{J}_p$ for $\rho_x = 0.3$, where $\boldsymbol{J}_p = \boldsymbol{j}_p\boldsymbol{j}_p'$ for $\boldsymbol{j}_p = (1, \ldots, 1)'$, a p-vector of ones. For the $n \times r$ matrix $\boldsymbol{Z}_i$'s in the model (1), the row vectors in $\boldsymbol{Z}_i$ are generated as mutually independent random variables distributed as $\mathcal{N}_r(\boldsymbol{0}, \boldsymbol{\Sigma}_z)$ where $\boldsymbol{\Sigma}_z = (1 - \rho_z)\boldsymbol{I}_r + \rho_z\boldsymbol{J}_r$ for $\rho_z = 0.3$. In this experiment, we assume that the true model is given by

$$(p^*) \qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon},$$

where $1 \leq p^* \leq 7$, $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_{p^*}, 0, \ldots, 0)'$, and $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ are mutually independent random variables having $\boldsymbol{v} \sim \mathcal{N}_R(\boldsymbol{0}, \sigma^2\psi\boldsymbol{I}_R)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$; it may be noted that when $\mathbf{Cov}\,(\boldsymbol{v}) = \sigma^2\psi\boldsymbol{D}$, where $\boldsymbol{D}$ is known, we may assume without any loss of generality that $\boldsymbol{D} = \boldsymbol{I}_R$, as it can be absorbed with the matrix $\boldsymbol{Z}$ by defining $\boldsymbol{Z}* = \boldsymbol{Z}\boldsymbol{D}^{-1/2}$. Here, we handle the cases that $\sigma^2 = 1$ and $\psi = 0.01, 0.5, 1.0$. Also, β_ℓ for $1 \leq \ell \leq p^*$ is generated as a random variable distributed as $\beta_\ell = 2(-1)^{\ell+1}\{1 + U(0, 1)\}$ for a uniform random variable $U(0, 1)$ on the interval $(0, 1)$. Let (m) be the set $\{1, \ldots, m\}$, and we write the model using the first m regressor variables $\beta_1, \ldots, \beta_m$ by M_m or simply (m). Then, the full model is (7) and the true model is (p^*). As candidate models, we consider the nested subsets $(1), \ldots, (7)$ of $\{1, \ldots, 7\}$, namely,

$$(m) \qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}^{(m)} = (\beta_1, \ldots, \beta_m, 0, \ldots, 0)'$.

In the simulation experiments, 10 observations of the regressor variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are generated, and for each observation of $\boldsymbol{X}$ and $\boldsymbol{Z}$, 30 observations of the response variable $\boldsymbol{y}$ are generated from the true model (p^*) for $p^* = 2, 4, 6$. Thus, we have $10 \times 30 (= 300)$ total data sets. For each data set, we calculate the values of $mAIC$ given in (7), $cAIC$ given in (9) and $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$ given (16), (17) and (18), respectively, for the eight candidate models $(1), \ldots, (8)$, and we select the models minimizing the values of the information criteria. For each criterion and each candidate model (m), the number of selecting the model (m) is counted for 300 data set. We thus obtain the frequencies of the model (m) selected by the criteria by dividing the number by 300. These frequencies are reported in Table 1, where standard deviations in selecting the true model are less than 0.02. From the table, we can see that the proposed conditional AIC's $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$ are superior to $mAIC$ and $cAIC$ for most of the cases.

When the random effects $\boldsymbol{v}_i$ has the covariance matrix such that $\boldsymbol{\Psi} = \text{diag}\,(\psi_1, \ldots, \psi_r)$, we next investigate the similar performances of the five criteria $mAIC$, $cAIC$, $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$ with known and unknown $\boldsymbol{\Psi}$, where ψ_i is estimated by (24) in the unknown case of $\boldsymbol{\Psi}$. The frequencies of selecting the true model are reported in Table 2 for the balanced case of $n_1 = \cdots = n_{20} = 10$, $k = 20$ and $r = 2, 3$. This numerical results show that the conditional AIC's $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$ are better than $mAIC$ and $cAIC$. It is interesting to note that the performance of $CAIC_{ML}$ does not depend on whether $\boldsymbol{\Psi}$ is known or unknown, since $CAIC_{ML}$ does not include $\boldsymbol{\Psi}$ or its estimator. This means that $CAIC_{ML}$ can be used even if $\boldsymbol{\Psi}$ cannot be estimated appropriately, or n_i's are small.

6 Concluding Remarks

In this paper, we have considered linear mixed models. To select the fixed-effects variables, we have derived three conditional Akaike information criteria $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$, and have shown that these $CAIC$'s perform better than $mAIC$ as well as better than $cAIC$, proposed by Vaida and Blanchard (2005). We have also considered the case when $\mathbf{Cov}\,(\boldsymbol{v}_i) = \sigma^2\psi\boldsymbol{D}_i$ as well as when $\mathbf{Cov}\,(\boldsymbol{v}_i) = \sigma^2\boldsymbol{\Psi}$, but $r_i = r$, where ψ is a scalar unknown parameter, and $\boldsymbol{\Psi}$ is an unknown covariance matrix. The proposed $CAIC$'s perform better than $mAIC$ and $cAIC$. However, when n_i's are small, it is recommended to use $CAIC_{ML}$ when the matrix $\boldsymbol{\Psi}$ is completely unknown as it does not depend on the unknown parameters.

Table 1: Frequencies selected by the five criteria $mAIC$, $cAIC$, $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$, abbreviated by AIC, C_{VB}, C_{ML}, C_{EB} and C_{GL}, in 300 replications for the unbalanced case of n_i and $k = 20$: the dimension of a full model is $p = 7$ and the true model is $(p^*) = \{1, \ldots, p^*\}$

	known $\boldsymbol{\psi}$					unknown $\boldsymbol{\psi}$				
M_k	AIC	C_{VB}	C_{ML}	C_{EB}	C_{GL}	AIC	C_{VB}	C_{ML}	C_{EB}	C_{GL}
$p^* = 2$, $\psi = 0.01$, $r = 1$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	76.3	76.3	84.0	84.7	84.7	75.3	74.7	84.0	85.0	84.7
(3)	11.0	11.0	9.7	9.0	9.0	11.3	12.7	9.7	8.7	9.0
(4)	4.3	4.3	3.3	2.3	2.3	5.0	4.7	3.3	2.7	2.7
(5)	4.3	4.3	1.0	2.3	2.3	3.7	3.7	1.0	2.0	2.0
(6)	2.0	2.0	1.0	1.3	1.3	2.3	2.0	1.0	1.0	1.0
(7)	2.0	2.0	1.0	0.3	0.3	2.3	2.3	1.0	0.7	0.7
$p^* = 4$, $\psi = 0.5$, $r = 2$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(4)	76.3	75.0	94.3	98.3	96.3	77.0	72.0	94.3	98.7	97.3
(5)	12.7	12.0	3.3	1.7	2.7	12.0	11.7	3.3	1.3	2.0
(6)	8.0	9.7	2.0	0.0	1.0	8.0	9.3	2.0	0.0	0.7
(7)	3.0	3.3	0.3	0.0	0.0	3.0	7.0	0.3	0.0	0.0
$p^* = 6$, $\psi = 1.0$, $r = 3$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(4)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
(5)	0.0	0.0	0.0	0.0	0.03	0.0	0.0	0.0	9.7	0.0
(6)	88.0	87.3	98.7	96.0	99.7	87.0	77.3	98.7	89.0	100.0
(7)	12.0	12.7	1.3	0.0	0.3	13.0	22.7	1.3	0.3	0.0

Table 2: Frequencies selected by the five criteria $mAIC$, $cAIC$, $CAIC_{ML}$, $CAIC_{SL}$ and $CAIC_{GL}$, abbreviated by AIC, C_{VB}, C_{ML}, C_{EB} and C_{GL}, in 300 replications for known and unknown $\boldsymbol{\Psi}$ and the unbalanced case of $n_1 = \cdots = n_{20} = 10$ and $k = 20$: the dimension of a full model is $p = 7$ and the true model is $(p^*) = \{1, \ldots, p^*\}$

	known $\boldsymbol{\Psi}$					unknown $\boldsymbol{\Psi}$				
M_k	AIC	C_{VB}	C_{ML}	C_{EB}	C_{GL}	AIC	C_{VB}	C_{ML}	C_{EB}	C_{GL}
$\boldsymbol{p^* = 2}$, $\boldsymbol{(\psi_1, \psi_2) = (0.01, 1.0)}$, $\boldsymbol{r = 2}$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	78.0	74.7	87.7	94.7	86.3	75.7	78.0	87.7	95.0	89.7
(3)	10.7	12.0	8.0	3.7	8.3	12.3	10.3	8.0	2.3	5.7
(4)	3.7	4.0	1.3	0.7	1.3	3.0	4.0	1.3	0.7	1.7
(5)	3.0	4.3	1.3	0.3	2.0	4.0	3.0	1.3	1.0	2.0
(6)	2.7	2.3	1.0	0.3	1.0	2.3	2.3	1.0	0.7	0.7
(7)	2.0	2.7	0.7	0.3	1.0	2.7	2.3	0.7	0.3	0.3
$\boldsymbol{p^* = 4}$, $\boldsymbol{(\psi_1, \psi_2, \psi_3) = (0.01, 0.5, 1.0)}$, $\boldsymbol{r = 3}$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(4)	78.0	76.7	86.0	95.7	91.0	74.7	73.0	86.0	93.3	87.3
(5)	10.3	11.7	8.7	3.7	5.3	11.3	13.0	8.7	5.0	7.0.0
(6)	6.7	7.3	3.7	0.3	2.0	8.3	8.0	3.7	1.3	3.0
(7)	5.0	4.3	1.7	0.3	1.7	5.7	6.0	1.7	0.3	2.7
$\boldsymbol{p^* = 6}$, $\boldsymbol{(\psi_1, \psi_2, \psi_3) = (1.0, 1.0, 1.0)}$, $\boldsymbol{r = 3}$										
(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(4)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(5)	0.0	0.0	0.0	0.0	0.0	0.7	6.3	0.0	0.0	0.0
(6)	85.3	85.3	91.3	95.3	91.0	85.7	85.3	91.3	97.0	93.7
(7)	14.7	14.7	8.7	4.7	9.0	14.3	14.7	8.7	3.0	6.3

A Appendix

A.1 Expectation of a ratio of quadratic forms

We here evaluate the expected value of the ratio of two quadratic forms which have been used to derive the bias terms of the conditional AIC.

Lemma A.1 *Assume that* $\boldsymbol{A}$ *is an* $N\times N$ *symmetric matrix, and that* $\boldsymbol{H}$ *is an idempotent matrix of rank* q*. Then,*

$$E\left[\frac{\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}'(\boldsymbol{I}_N-\boldsymbol{H})\boldsymbol{u}}\right]=\frac{\mathrm{tr}\,\boldsymbol{A}}{N-q-2}-\frac{2\mathrm{tr}\,[A(\boldsymbol{I}_N-\boldsymbol{H})]}{(N-q)(N-q-2)},$$

where $\boldsymbol{u}\sim\mathcal{N}_N(\boldsymbol{0},\boldsymbol{I}_N)$.

Proof. Since $\boldsymbol{H}$ is an idempotent matrix of rank q, we can write $\boldsymbol{H}=\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2'$, $\boldsymbol{\Gamma}_2=N\times q$, and $\boldsymbol{\Gamma}_2'\boldsymbol{\Gamma}_2=\boldsymbol{I}_q$. Let $\boldsymbol{\Gamma}_1$ be $N\times(N-q)$ matrix such that $\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_2=\boldsymbol{0}$ and $\boldsymbol{\Gamma}_1'\boldsymbol{\Gamma}_1=\boldsymbol{I}_N-q$. Then, $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2)$ is an $N\times N$ orthogonal matrix and hence $\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1'+\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2'=\boldsymbol{I}_N$, and $\boldsymbol{I}_N-\boldsymbol{H}=\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1'$. Let $\boldsymbol{v}=\boldsymbol{\Gamma}'\boldsymbol{u}$, $\boldsymbol{v}_1=\boldsymbol{\Gamma}_1'\boldsymbol{u}$, $\boldsymbol{v}_2=\boldsymbol{\Gamma}_2'\boldsymbol{u}$ and

$$\boldsymbol{B}=\boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}'=\begin{pmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{12}' & \boldsymbol{B}_{22}\end{pmatrix}.$$

Then,

$$Q=\frac{\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}'(\boldsymbol{I}_N-\boldsymbol{H})\boldsymbol{u}}=\frac{\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v}}{\boldsymbol{v}_1'\boldsymbol{v}_1}=\frac{\boldsymbol{v}_1'\boldsymbol{B}_{11}\boldsymbol{v}_1+2\boldsymbol{v}_1'\boldsymbol{B}_{12}\boldsymbol{v}_2+\boldsymbol{v}_2'\boldsymbol{B}_{22}\boldsymbol{v}_2}{\boldsymbol{v}_1'\boldsymbol{v}_1}.$$

Noting that $\boldsymbol{v}\sim\mathcal{N}_N(\boldsymbol{0},\boldsymbol{I}_N)$, we can see that

$$\begin{aligned}E[Q]=&\frac{\mathrm{tr}\,\boldsymbol{B}_{11}}{N-q}+\frac{\mathrm{tr}\,\boldsymbol{B}_{22}}{N-q-2}=\frac{\mathrm{tr}\,\boldsymbol{A}\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1'}{N-q}+\frac{\mathrm{tr}\,\boldsymbol{A}\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2'}{N-q-2}\\ =&\frac{\mathrm{tr}\,\boldsymbol{A}(\boldsymbol{I}_N-\boldsymbol{H})}{N-q}+\frac{\mathrm{tr}\,\boldsymbol{A}\boldsymbol{H}}{N-q-2}\\ =&\frac{\mathrm{tr}\,\boldsymbol{A}}{N-q-2}-\frac{2\mathrm{tr}\,\boldsymbol{A}(\boldsymbol{I}_N-\boldsymbol{H})}{(N-q)(N-q-2)},\end{aligned}$$

which proves Lemma A.1. ■

A.2 Consistency of $\widehat{\boldsymbol{\Psi}}^U$

From the consistency of $\tilde{\sigma}_1^2$ and the fact that $\sum_{i=1}^k\boldsymbol{S}_i/k=O_p(1)$, it follows that $\widehat{\boldsymbol{\Psi}}^U=\boldsymbol{\Psi}^*/\sigma^2+O_p(k^{-1/2})$ for $\boldsymbol{\Psi}^*=\sum_{i=1}^k\boldsymbol{S}_i/k-\sigma^2\sum_{i=1}^k\{(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}+\widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^{+}\widetilde{\boldsymbol{X}}_{2i}'\}/k$. Note that $\boldsymbol{S}_i=(\tilde{\boldsymbol{y}}_{2i}-\widetilde{\boldsymbol{X}}_{2i}\widetilde{\boldsymbol{\beta}}_1)(\tilde{\boldsymbol{y}}_{2i}-\widetilde{\boldsymbol{X}}_{2i}\widetilde{\boldsymbol{\beta}}_1)'$ and that $\tilde{\boldsymbol{y}}_{2i}$ and $\widetilde{\boldsymbol{\beta}}_1$ are mutually independently

distributed as $\tilde{\boldsymbol{y}}_{2i} \sim \mathcal{N}(\widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ and $\widetilde{\boldsymbol{\beta}}_1 \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^+)$, where $\boldsymbol{\Sigma}_i = \sigma^2\{\boldsymbol{\Psi} + (\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\}$. Then, $\widehat{\boldsymbol{\Psi}}^* - \sigma^2\boldsymbol{\Psi}$ is expressed as

$$\begin{aligned}
\widehat{\boldsymbol{\Psi}}^* - \sigma^2\boldsymbol{\Psi} =& \frac{1}{k}\sum_{i=1}^{k}\Big\{(\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta})(\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta})' - \boldsymbol{\Sigma}_i\Big\} \\
&+ \frac{1}{k}\sum_{i=1}^{k}\Big\{\widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})'\widetilde{\boldsymbol{X}}_{2i}' - \sigma^2\widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}_{2i}'\Big\} \\
&- \frac{1}{k}\sum_{i=1}^{k}\Big\{(\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})'\widetilde{\boldsymbol{X}}_{2i}' + \widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})(\tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta})'\Big\} \\
=& \boldsymbol{K}_1 + \boldsymbol{K}_2 - \boldsymbol{K}_3. \quad \text{(say)}
\end{aligned}$$

It is easy to see that $E[\boldsymbol{K}_j] = 0$ for $j = 1, 2, 3$. For the consistency of $\boldsymbol{\Psi}^U$ or $\boldsymbol{\Psi}^*$, it is sufficient to show that $E[\mathrm{tr}\,[\boldsymbol{K}_j^2]] = O(k^{-1})$ for $j = 1, 2, 3$. Let $\boldsymbol{u}_i = \tilde{\boldsymbol{y}}_{2i} - \widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta}$, and it has $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$. Then,

$$E[\mathrm{tr}\,[\boldsymbol{K}_1^2]] = \frac{1}{k^2}E\Big[\mathrm{tr}\,\Big[(\sum_{i=1}^{k}\boldsymbol{u}_i\boldsymbol{u}_i')^2\Big]\Big] - \frac{1}{k^2}\mathrm{tr}\,\Big[(\sum_{i=1}^{k}\boldsymbol{\Sigma}_i)^2\Big]$$

and

$$\begin{aligned}
E\Big[\mathrm{tr}\,[\sum_{i=1}^{k}\boldsymbol{u}_i\boldsymbol{u}_i']^2\Big] =& E\Big[\sum_{i=1}^{k}(\boldsymbol{u}_i'\boldsymbol{u}_i)^2 + \sum_{i=1}^{k}\sum_{j\neq i}\mathrm{tr}\,[\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{u}_j\boldsymbol{u}_j']\Big] \\
=& \sum_{i=1}^{k}\Big\{2\mathrm{tr}\,[\boldsymbol{\Sigma}_i^2] + (\mathrm{tr}\,[\boldsymbol{\Sigma}_i])^2\Big\} + \sum_{i=1}^{k}\sum_{j\neq i}\mathrm{tr}\,\Big[\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j\Big] \\
=& 2\sum_{i=1}^{k}\mathrm{tr}\,[\boldsymbol{\Sigma}_i^2] + \mathrm{tr}\,\Big[(\sum_{i=1}^{k}\boldsymbol{\Sigma}_i)^2\Big].
\end{aligned}$$

Hence,

$$E[\mathrm{tr}\,[\boldsymbol{K}_1^2]] = 2\sum_{i=1}^{k}\mathrm{tr}\,[\boldsymbol{\Sigma}_i^2]/k^2,$$

which has the order $O(k^{-1})$. Similarly, we can see that

$$\begin{aligned}
E[\mathrm{tr}\,[\boldsymbol{K}_2^2]] =& 2\sigma^4\mathrm{tr}\,\Big[(\sum_{i=1}^{k}\widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}_{2i}')^2\Big]/k^2, \\
E[\mathrm{tr}\,[\boldsymbol{K}_3^2]] =& 4\sigma^2\sum_{i=1}^{k}\mathrm{tr}\,[\boldsymbol{\Sigma}_i]\mathrm{tr}\,[\widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}_{2i}']/k^2,
\end{aligned}$$

both of which have the order $O(k^{-1})$ under the assumption that $\mathbf{\Sigma}_i$ is bounded and $\sum_{i=1}^{k} \widetilde{\boldsymbol{X}}_{2i}(\widetilde{\boldsymbol{X}}'_1\widetilde{\boldsymbol{X}}_1)^+\widetilde{\boldsymbol{X}}'_{2i} = O(1)$. Thus, the consistency of $\widehat{\boldsymbol{\Psi}}^U$ is verified. ■

Acknowledgments. The research of the first author was supported by NSERC. The research of the second author was supported in part by a grant from the Ministry of Education, Japan, No. 19200020.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B.N. Petrov and Csaki, F, eds.), 267-281, Akademia Kiado, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Trans. Autom. Contr.*, **AC-19**, 716-723.

[3] Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.*, **21**, 309-310.

[4] Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling.* Springer.

[5] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. Royal Statist. Soc.*, **B 64**, 583-639.

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