CIRJE-F-895

Modfiied Conditional AIC in Linear Mixed Models

Yuki Kawakubo
Graduate School of Economics, University of Tokyo

Tatsuya Kubokawa
University of Tokyo

July 2013

Modified Conditional AIC in Linear Mixed Models

Yuki Kawakubo[*]and Tatsuya Kubokawa[†]

University of Tokyo

July 13, 2013

Abstract

In linear mixed models, the conditional Akaike Information Criterion (cAIC) is a procedure for variable selection in light of the prediction of specific clusters or random effects. This is useful in problems involving prediction of random effects such as small area estimation, and much attention has been received since suggested by Vaida and Blanchard (2005). A weak point of cAIC is that it is derived as an unbiased estimator of conditional Akaike information (cAI) in the overspecified case, namely in the case that candidate models include the true model. This results in larger biases in the underspecified case that the true model is not included in candidate models. In this paper, we derive the modified cAIC (McAIC) to cover both the underspecified and overspecified cases, and investigate properties of McAIC. It is numerically shown that McAIC has less biases and less prediction errors than cAIC.

Key words and phrases: Asymptotically unbiased estimator, Akaike information criterion, conditional AIC, Kullback-Leibler information, linear mixed model, small area estimation, variable selection.

1 Introduction

Linear mixed models (LMM) and empirical best linear unbiased predictors (EBLUP) have been studied for a long time in the literature from both theoretical and applied aspects. The problem of selecting explanatory variables in LMM is important since one needs to select significant variables in order to give a good prediction. One of the conventional procedures for variable selection is the Akaike Information Criterion (AIC) proposed by Akaike (1973, 1974) based on the marginal likelihood, which integrates out the likelihood with respect to random effects in LMM. However, AIC is not appropriate when one is interested in prediction of specific clusters or random effects. One of such problems is the

[*]Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: y.k.5.58.2010@gmail.com

[†]Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

small area estimation, and the predictor induced from LMM has been used and studied actively and extensively in estimation of means of small geographical areas. Then it is important to select variables in terms of minimizing the prediction errors for the focus on specific random effects. An appropriate method in this direction is the conditional Akaike information criterion (cAIC) suggested by Vaida and Blanchard (2005). The cAIC is related to estimation of the expected Kullback-Leibler information based on the conditional density given random effects, and cAIC is derived as an (asymptotically) unbiased estimator of the conditional Akaike information (cAI) when the true model is included in candidate models. Since Vaida and Blanchard (2005), the cAIC and the relevant criteria have been studied by Liang, Wu and Zou (2008), Greven and Kneib (2010), Srivastava and Kubokawa (2010), Donohue, Overholser, Xu and Vaida (2011), Kubokawa (2011), Kubokawa and Nagashima (2012) and others.

A critical point of AIC, cAIC and Mallows' C_p is that those procedures are derived when a candidate model includes the true model. This assumption is called the overspecified case. On the other hand, the underspecified case means that a candidate model does not include the true model. Thus, we have the following questions:

(I) Is cAIC appropriate as an estimator of cAI in the underspecified case ?

(II) Can one extend cAIC to a procedure useful for both the under- and over-specified cases ?

For the query (I), it is noted that the cAIC is not an asymptotically unbiased estimator of cAI in the underspecified case. In fact, cAIC has large biases in the underspecified case as illustrated in Tables 1 and 2. Thus, the drawback of cAIC gives a motivation for addressing the query (II).

In this paper, we derive an asymptotically unbiased estimator of cAI in both under- and over-specified cases. This procedure is here called the modified conditional AIC (McAIC). The setup of linear mixed models and the concept of cAIC is explained in Section 2. The problem of variable selection which we consider in this paper is also described. In Section 3, we derive the McAIC as an asymptotically unbiased estimator of cAI in both under- and over-specified cases. This approach was used by Fujikoshi and Satoh (1997) to modify AIC and Mallows' C_p in multivariate linear regression models. The performance of McAIC is investigated numerically by simulation in Section 4, and it is shown that McAIC and the corresponding model averaging procedure are better than cAIC in terms of the prediction error. In Section 5, we apply the McAIC to estimate small area land prices. All the proofs are given in the Appendix.

2 Setup of Models and Conditional AIC

2.1 Linear mixed models and conditional AIC

Consider a linear mixed model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \tag{2.1}$$

where $\boldsymbol{y}_i$ is an n_i-variate vector of observations from i-th small area (or cluster), $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ matrices of covariates, respectively, $\boldsymbol{\beta}$ is a p-variate vector of unknown regression coefficients, $\boldsymbol{b}_i$ is a q-variate vector of random effects, and $\boldsymbol{\epsilon}_i$ is an n_i-variate vector of random errors. Here, $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{G}_i), \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{n_i})$, and $\boldsymbol{b}_i$'s and $\boldsymbol{\epsilon}_i$'s are mutually independent, where $\boldsymbol{G}_i$ is a $q \times q$ covariance matrix, and σ^2 is an unknown variance. Let $N = \sum_{i=1}^m n_i$ be the total number of observations.

The model (2.1) is rewritten in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1^t, \ldots, \boldsymbol{X}_m^t)^t$ and $\boldsymbol{Z} = \mathrm{diag}\,(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$ are $N \times p$ and $N \times r$ matrices, respectively, for $r = mq$, $\boldsymbol{b} = (\boldsymbol{b}_1^t, \ldots, \boldsymbol{b}_m^t)^t$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^t, \ldots, \boldsymbol{\epsilon}_m^t)^t$. It is seen that $\boldsymbol{b} \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{G})$ for $\boldsymbol{G} = \mathrm{diag}\,(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_m)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$.

Let $\boldsymbol{\theta}$ be the collection of unknown parameters in model (2.2). The conditional density function of $\boldsymbol{y}$ given $(\boldsymbol{b}, \boldsymbol{\theta})$ is denoted by $f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\theta})$, and the density of $\boldsymbol{b}$ is $\pi(\boldsymbol{b}|\boldsymbol{G})$. The marginal density function of $\boldsymbol{y}$ is written as

$$f_\pi(\boldsymbol{y}|\boldsymbol{\theta}) = \int f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\theta})\pi(\boldsymbol{b}|\boldsymbol{G})\mathrm{d}\boldsymbol{b}.$$

Using these notations, we consider to predict the distribution for the focus on specific clusters or random effects. To measure the prediction error of some candidate model $\hat{f}$, Vaida and Blanchard (2005) used the expected Kullback-Leibler information based on the conditional density, given by

$$\iint \left[\int \log \left\{ \frac{f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\theta})}{\hat{f}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y}))} \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\theta})\mathrm{d}\tilde{\boldsymbol{y}} \right] f(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b}, \tag{2.3}$$

where $\tilde{\boldsymbol{y}}$ is a future sample vector independent of $\boldsymbol{y}$ given $\boldsymbol{b}$, $\hat{\boldsymbol{b}}(\boldsymbol{y})$ is the empirical Bayes estimator of $\boldsymbol{b}$ and $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ is some estimator of $\boldsymbol{\theta}$. Since the numerator of (2.3) is irrelevant to the model $\hat{f}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y}))$, it is sufficient to consider

$$cAI = -2 \iiint \log\{\hat{f}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y}))\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\theta}) f(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\theta})\mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b}, \tag{2.4}$$

which is called the conditional Akaike Information (cAI). When cAI is estimated by $-2\log \hat{f}(\boldsymbol{y}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y}))$, the bias is denoted by

$$\Delta_{cAI} = cAI - E\left[-2\log \hat{f}(\boldsymbol{y}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y})) \right].$$

Then the conditional AIC (cAIC) is defined by the bias corrected unbiased estimator of cAI, given by

$$cAIC = -2\log \hat{f}(\boldsymbol{y}|\hat{\boldsymbol{b}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y})) + \widehat{\Delta_{cAI}},$$

where $\widehat{\Delta_{cAI}}$ is an (asymptotically) unbiased estimator of Δ_{cAI}, and is called a bias correction term or a penalty term.

2.2 Setup for the problem of variable selection

In this paper, we focus on the problem of selecting explanatory variables in linear mixed model in the following setup.

First, we assume the same setup of the true model as in Vaida and Blanchard (2005). Let $N \times p_\omega$ matrix $\boldsymbol{X}$ consist of all the explanatory variables, and the true model be given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ in the same framework of (2.2), where $\boldsymbol{\beta}^*$ is a vector of the true regression coefficients with p^* non-zero components and $p_\omega - p^*$ zero components, $\boldsymbol{b} \sim \mathcal{N}_r(\boldsymbol{0}, \sigma^2\boldsymbol{G}_0)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ for a common parameter σ^2 and a known matrix $\boldsymbol{G}_0$. Thus the marginal density of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma^2\boldsymbol{\Sigma}_0), \tag{2.5}$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{Z}\boldsymbol{G}_0\boldsymbol{Z}^t + \boldsymbol{I}_N$.

In the setup of the true model, the assumption of the covariance matrix $\sigma^2\boldsymbol{G}_0$, used by Vaida and Blanchard (2005), seems restrictive. However, it is not very restrictive as long as we handle the problem of selecting only explanatory variables. When $\boldsymbol{G}_0$ includes unknown parameters $\boldsymbol{\psi}$, namely, $\boldsymbol{G}_0 = \boldsymbol{G}(\boldsymbol{\psi})$, we can use the variable selection procedure by replacing $\boldsymbol{\psi}$ with an estimator $\widehat{\boldsymbol{\psi}}$. More explanations about it will be given in Remark 3.1. If the selection of both explanatory variables and random effects were treated, the setup of $\sigma^2\boldsymbol{G}_0$ would be inappropriate.

Second, our objective is to select a good model from the collection of candidate models $\{M_j\}$ for $j = 1, \ldots, F$. Model M_F denotes the full (biggest) model including all the explanatory variables, given by

$$M_F : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}_\omega + \boldsymbol{\epsilon}_\omega, \tag{2.6}$$

where $\boldsymbol{\beta}$ is a p_ω-variate vector, $\boldsymbol{b}_\omega \sim \mathcal{N}_r(\boldsymbol{0}, \sigma_\omega^2\boldsymbol{G}_0)$ and $\boldsymbol{\epsilon}_\omega \sim \mathcal{N}_N(\boldsymbol{0}, \sigma_\omega^2\boldsymbol{I}_N)$. Model M_j is expressed as

$$M_j : \boldsymbol{y} = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b}_j + \boldsymbol{\epsilon}_j, \tag{2.7}$$

where $\boldsymbol{X}_j$ consists of p_j columns of $\boldsymbol{X}$, $\boldsymbol{\beta}_j$ is a p_j-variate vector corresponding to $\boldsymbol{X}_j$, $\boldsymbol{b}_j$ is a q-variate vector of random effects, and $\boldsymbol{\epsilon}_j$ is an N-variate vector of random errors. It is here assumed that $\boldsymbol{b}_j \sim \mathcal{N}_r(\boldsymbol{0}, \sigma_j^2\boldsymbol{G}_0)$ and $\boldsymbol{\epsilon}_j \sim \mathcal{N}_N(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}_N)$.

Third, we assume that the collection of candidate models includes both underspecified and overspecified models, and that the full model includes the true model. Here, a candidate model M_j is overspecified if $\boldsymbol{X}\boldsymbol{\beta}^* \in \mathcal{R}[\boldsymbol{X}_j]$, which means that $\boldsymbol{X}\boldsymbol{\beta}^*$ is in the column space of $\boldsymbol{X}_j$. If $\boldsymbol{X}\boldsymbol{\beta}^* \notin \mathcal{R}[\boldsymbol{X}_j]$, M_j is called an underspecified model. This definition is the same as in Fujikoshi and Satoh (1997), who modified AIC and Mallows' C_p for underspecification in multivariate linear regression. We shall modify cAIC by Vaida and Blanchard (2005) so that we can use a modified procedure in both the overspecified and underspecified cases.

3 Modification of cAIC

3.1 Evaluation of the bias of cAIC

We begin by deriving the bias of cAIC for model M_j in (2.7). The unknown parameters are $\boldsymbol{\beta}_j$ and σ_j^2 and their maximum likelihood (ML) estimators are

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j =&(\boldsymbol{X}_j^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{y},\\ \hat{\sigma}_j^2 =&\frac{1}{N}(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j).\end{aligned}$$

The bias is given by

$$\Delta_{cAI} = cAI - E[-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}_j(\boldsymbol{y}))],$$

where $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}_j(\boldsymbol{y}))$ is

$$N\log(2\pi\hat{\sigma}_j^2)+(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j-\boldsymbol{Z}\hat{\boldsymbol{b}}_j)^t(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j-\boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2. \tag{3.1}$$

As shown in the Appendix, the bias can be expressed as

$$\Delta_{cAI} = E\Big[\frac{1}{\hat{\sigma}_j^2}\big\{(2N-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])\sigma^2-\boldsymbol{u}^t\boldsymbol{\Sigma}_0^{-2}\boldsymbol{u}+2\boldsymbol{u}^t\boldsymbol{\Sigma}_0^{-2}(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j-\boldsymbol{X}\boldsymbol{\beta}^*)\big\}\Big], \tag{3.2}$$

for $\boldsymbol{u}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*$.

It is important to note that the distribution of $\hat{\sigma}_j^2$ under the underspecified case is different from that under the overspecified case. Thus, we need to clarify the distribution of $\hat{\sigma}_j^2$. To this end, $N\hat{\sigma}_j^2$ is decomposed as

$$\begin{aligned}N\hat{\sigma}_j^2 =&\Big\{\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{y}-\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{M}_j\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{y})\Big\}^t\Big\{\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{y}-\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{M}_j\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{y})\Big\}\\ =&\sigma^2\left\{\boldsymbol{z}^t(\boldsymbol{I}_N-\boldsymbol{M}_\omega)\boldsymbol{z}+\boldsymbol{z}^t(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{z}\right\}\\ =&\sigma^2(W_0+W_1),\end{aligned}$$

where $W_0=\boldsymbol{z}^t(\boldsymbol{I}_N-\boldsymbol{M}_\omega)\boldsymbol{z}$, $W_1=\boldsymbol{z}^t(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{z}$,

$$\begin{aligned}\boldsymbol{z} =&\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{y}/\sigma,\\ \boldsymbol{M}_\omega =&\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}_0^{-1/2},\\ \boldsymbol{M}_j =&\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{\Sigma}_0^{-1/2}.\end{aligned}$$

Note that $\boldsymbol{M}_j$ and $\boldsymbol{M}_\omega$ are symmetric and idempotent. Let $\boldsymbol{v}=\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{u}/\sigma$ and $\boldsymbol{\eta}=\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{X}\boldsymbol{\beta}^*/\sigma$. Then, it is seen that

$$\boldsymbol{M}_\omega\boldsymbol{\eta}=\boldsymbol{\eta},$$

for both underspecified and overspecified models, and that

$$\boldsymbol{M}_j\boldsymbol{\eta}\begin{cases}=\boldsymbol{\eta} & \text{for overspecified models,}\\ \neq\boldsymbol{\eta} & \text{for underspecified models,}\end{cases}$$

since $\boldsymbol{X}\boldsymbol{\beta}^* \in \mathcal{R}[\boldsymbol{X}_j]$ for the overspecified case. Thus W_0 can be rewritten as

$$W_0 = (\boldsymbol{\eta} + \boldsymbol{v})^t(\boldsymbol{I}_N - \boldsymbol{M}_\omega)(\boldsymbol{\eta} + \boldsymbol{v}) = \boldsymbol{v}^t(\boldsymbol{I}_N - \boldsymbol{M}_\omega)\boldsymbol{v}, \tag{3.3}$$

so that W_0 follows a chi-squared distribution with $N - p_\omega$ degrees of freedom, denoted by

$$W_0 \sim \chi^2_{N-p_\omega}.$$

Also, W_1 can be rewritten as

$$\begin{aligned} W_1 =& \boldsymbol{v}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta} \\ =& \boldsymbol{v}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2L + N\delta, \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} L =& \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v}, \\ \delta =& \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}/N. \end{aligned} \tag{3.5}$$

In the overspecified case, we have $W_1 \sim \chi^2_{p_\omega - p_j}$ since $\boldsymbol{M}_\omega\boldsymbol{\eta} = \boldsymbol{M}_j\boldsymbol{\eta} = \boldsymbol{\eta}$. In the underspecified case, W_1 follows a noncentral chi-squared distribution with $p_\omega - p_j$ degrees of freedom and with the noncentrality parameter $N\delta$, denoted by $W_1 \sim \chi^2_{p_\omega - p_j}(N\delta)$. Thus,

$$W_1 \sim \begin{cases} \chi^2_{p_\omega - p_j} & \text{for overspecified models,} \\ \chi^2_{p_\omega - p_j}(N\delta) & \text{for underspecified models.} \end{cases}$$

Since $\boldsymbol{u}^t\boldsymbol{\Sigma}_0^{-2}\boldsymbol{u} = \sigma^2\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}$ and

$$\boldsymbol{u}^t\boldsymbol{\Sigma}_0^{-2}(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*) = \sigma^2\left\{\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v} - \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}\right\},$$

we can rewrite (3.2) as

$$\Delta_{cAI} = N \cdot E\left[\frac{2N - \operatorname{tr}\boldsymbol{\Sigma}_0^{-1}}{W_0 + W_1} - \frac{\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}}{W_0 + W_1} + 2\frac{\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}}{W_0 + W_1} - 2\frac{\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}}{W_0 + W_1}\right]. \tag{3.6}$$

Although $W_0 + W_1$ has a central chi-squared distribution in the overspecified case, it has a noncentral chi-squared distribution in the underspecified case. Thus, we need to approximate the bias Δ_{cAI}. For the purpose, assume the following conditions:

(A1) $\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta} = O(N)$, which is the non-centrality parameter of W_1.

(A2) $\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta} = O(N)$.

The condition (A1) is equivalent to $\delta = O(1)$ given in (3.5). It is also noted that the condition (A2) is satisfied by (A1) if the maximum eigenvalue of $\boldsymbol{\Sigma}_0^{-1}$ is uniformly bounded. Under these assumptions, we can get the following theorem which will be proved in the Appendix.

Theorem 3.1 *In the overspecified case, the bias of cAIC is provided by the exact expression* $\Delta_{cAI} = B^*$, *where*

$$B^* = 2N \times \left\{ \frac{N - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}] + \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j]}{N - p_j - 2} + \frac{\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}] - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j]}{(N - p_j - 2)(N - p_j)} \right\}. \tag{3.7}$$

In the underspecified case, the bias of cAIC is approximated as

$$\Delta_{cAI} = B^* + B_1 + B_2 + B_3 + O(N^{-1}), \tag{3.8}$$

where B_1, B_2 *and* B_3 *are defined by*

$$B_1 = \frac{2N(\lambda - 1)}{N - p_j - 2}(N - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]), \tag{3.9}$$

$$\begin{aligned} B_2 =& 2p_j\lambda(\lambda - 1) - 4\lambda(\lambda - 1)^2 + 2\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j](\lambda - 1) \\ &+ 2(\lambda - 1)\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}] \times \frac{2\lambda^2 - (p_j + 1)\lambda + 1}{N}, \end{aligned} \tag{3.10}$$

$$B_3 = \frac{4\lambda^2}{N}\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}, \tag{3.11}$$

for $\lambda = 1/(1 + \delta)$.

It is noted that in the overspecified case the bias B^* given in (3.7) is identical to that in Vaida and Blanchard (2005). It is also noted that $B_1 = B_2 = B_3 = 0$ in the overspecified case, since $\lambda = 1$ and $\boldsymbol{M}_j\boldsymbol{\eta} = \boldsymbol{\eta}$.

3.2 Estimation of the bias

We now derive an asymptotically unbiased estimator of the bias Δ_{cAI}. It follows from Theorem 3.1 that it is sufficient to estimate $B^* + B_1 + B_2 + B_3$. Since B_1 and B_2 are linear functions of λ, λ^2 and λ^3, we begin by estimating these polynomials of λ.

Let us define $\hat{\lambda}$, $\widehat{\lambda^2}$ and $\widehat{\lambda^3}$ by

$$\hat{\lambda} = \frac{N - p_j}{N - p_\omega}\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}, \tag{3.12}$$

$$\widehat{\lambda^2} = \frac{(N - p_j)(N - p_j + 2)}{(N - p_\omega)(N - p_\omega + 2)}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^2, \tag{3.13}$$

$$\widehat{\lambda^3} = \frac{(N - p_j)(N - p_j + 2)(N - p_j + 4)}{(N - p_\omega)(N - p_\omega + 2)(N - p_\omega + 4)}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^3. \tag{3.14}$$

In the overspecifed case, it is noted that $N\hat{\sigma}_\omega^2 = \sigma^2 W_0 \sim \sigma^2\chi^2_{N-p_\omega}$, $W_1 \sim \chi^2_{p_\omega - p_j}$ and $N\hat{\sigma}_j^2 = \sigma^2(W_0 + W_1)$, so that

$$\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2} \sim Be\left(\frac{N - p_\omega}{2}, \frac{p_\omega - p_j}{2}\right),$$

where $Be(\cdot,\cdot)$ denotes the beta distribution. This implies that $E[\hat{\lambda}] = E[\widehat{\lambda^2}] = E[\widehat{\lambda^3}] = 1$ in the overspecified case. In the underspecified case, on the other hand, it follows that $E[(\hat{\sigma}_\omega^2/\hat{\sigma}_j^2)^k] \to \lambda^k$ as $N \to \infty$ for $k = 1, 2, 3$, where the brief proof is given in the Appendix.

Lemma 3.1 *In the overspecified case,* $E[\hat{\lambda}] = E[\widehat{\lambda^2}] = E[\widehat{\lambda^3}] = 1$*. In the underspecified case,* $\hat{\lambda}$*,* $\widehat{\lambda^2}$ *and* $\widehat{\lambda^3}$ *are asymptotically unbiased estimators of* λ*,* λ^2 *and* λ^3*, respectively.*

Using estimators (3.12), (3.13) and (3.14), we can estimate B_1 and B_2 in (3.9) and (3.10). However, because of $B_1 = O(N)$, a naive estimator that just substitutes $\hat{\lambda}$ for λ in B_1 has a bias with order $O(1)$. Then $E[\hat{\lambda}]$ can be expanded up to $O(N^{-1})$ as

$$\begin{aligned} E[\hat{\lambda}] =& \frac{N - p_j}{N - p_\omega} E\left[\frac{W_0}{W_0 + W_1}\right] \\ =& \lambda + \frac{-2\lambda^2(\lambda - 1) + p_j\lambda(\lambda - 1)}{N} + O(N^{-2}), \end{aligned} \tag{3.15}$$

where the proof is given in (A.9) in the Appendix.

Lemma 3.2 *Consider the following estimator for* B_1*:*

$$\widehat{B_1} = \frac{2N(N - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])}{N - p_j - 2}\left\{\hat{\lambda} - 1 + \frac{2(\widehat{\lambda^3} - \widehat{\lambda^2}) - p_j(\widehat{\lambda^2} - \hat{\lambda})}{N}\right\}. \tag{3.16}$$

Then, in the overspecified case, $E[\widehat{B_1}] = 0$*, and in the underspecified case,* $E[\widehat{B_1}] = B_1 + O(N^{-1})$*.*

We next obtain an estimator of $\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}$ which is a part of B_3. Define $\tilde{\sigma}_j^2$ by

$$\tilde{\sigma}_j^2 = (\boldsymbol{y} - \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j)^t\boldsymbol{\Sigma}_0^{-2}(\boldsymbol{y} - \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j).$$

From the fact that $\tilde{\sigma}_j^2 = \sigma^2(\boldsymbol{v} + \boldsymbol{\eta})^t(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)(\boldsymbol{v} + \boldsymbol{\eta})$, it follows that

$$E[\tilde{\sigma}_j^2] = \sigma^2\left\{\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)] + \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}\right\}.$$

Hence an estimator of $\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}$ is given by

$$\tilde{\sigma}_j^2/\hat{\sigma}_\omega^2 - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)].$$

Lemma 3.3 *Consider the following estimator for* B_3*:*

$$\widetilde{B_3} = \frac{4}{N}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^2 \times \left\{\frac{\tilde{\sigma}_j^2}{\hat{\sigma}_\omega^2} - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)]\right\}.$$

Then in the overspecified case, $E[\widetilde{B_3}] = O(N^{-1})$*. In the underspecified case,* $E[\widetilde{B_3}] = B_3 + O(N^{-1})$*.*

Lemma 3.3 implies that in both overspecified and underspecified cases, $\widetilde{B_3}$ is an asymptotically unbiased estimator of B_3 up to O(1), but $\widetilde{B_3}$ has a $O(N^{-1})$ bias that cannot be negligible for overspecified models. Since the cAIC by Vaida and Blanchard (2005) is an exact unbiased estimator of cAI, we want to adjust $\widetilde{B_3}$ so that the adjusted estimator can have a bias with order $O(N^{-2})$ in the overspecified case.

Lemma 3.4 *For B_3, consider the following estimator as a higher order unbiased estimator than $\widetilde{B_3}$:*

$$\widehat{B_3} = \widetilde{B_3} - \frac{4\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_n - \boldsymbol{M}_j)]p_\omega}{N^2} + \frac{8\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)]}{N^2}. \tag{3.17}$$

Then, in the overspecified case, $E[\widehat{B_3}] = O(N^{-2})$. In the underspecified case, $E[\widehat{B_3}] = B_3 + O(N^{-1})$.

Using Lemmas 3.1, 3.2 and 3.4, we can estimate the bias Δ_{cAI} by the estimator

$$\widehat{\Delta_{cAI}} = B^* + \widehat{B_1} + \widehat{B_2} + \widehat{B_3}. \tag{3.18}$$

The bias correction estimator can be used not only for overspecified models, but also for underspecified models. Thus, we get the modified conditional Akaike information criterion (McAIC) given by

$$McAIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j) + \widehat{\Delta_{cAI}}. \tag{3.19}$$

Theorem 3.2 *In the overspecified case, it follows that*

$$E[\widehat{\Delta_{cAI}}] = \Delta_{cAI} + O(N^{-2}) \quad \text{and} \quad E[McAIC] = cAI + O(N^{-2}).$$

In the underspecified case, it follows that

$$E[\widehat{\Delta_{cAI}}] = \Delta_{cAI} + O(N^{-1}) \quad \text{and} \quad E[McAIC] = cAI + O(N^{-1}).$$

Remark 3.1 In the derivation of McAIC, we assume that the covariance matrix of $\boldsymbol{b}$ is $\sigma^2\boldsymbol{G}_0$ for a known matrix $\boldsymbol{G}_0$. This setup seems restrictive, since $\sigma^2\boldsymbol{G}_0$ involves some unknown variance components in most linear mixed models. For example, we consider the nested error regression model which wil be treated in the next section for simulation. In this model, $\boldsymbol{G}_0$ is a function of $\psi = \tau^2/\sigma^2$ where τ^2 is a variance component of random effects. Since a consistent estimator $\widehat{\psi}$ for ψ is available, we can use the plug-in estimator $\boldsymbol{G}_0(\widehat{\psi})$ for $\boldsymbol{G}_0(\psi)$. Then, McAIC can be extended by replacing $\boldsymbol{G}_0(\psi)$ in (3.19) with $\boldsymbol{G}_0(\widehat{\psi})$. The influence by this replacement may be limited as long as one considers the problem of selecting only explanatory variables.

4 Simulation Study

In this section, we investigate the behaviors of the suggested criterion McAIC by simulation through two kinds of experiments.

In the first experiment, we consider a class of the nested models denoted by M_j, which is described by

$$M_j : \boldsymbol{y} = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{\beta}_j = (\beta_1, \ldots, \beta_j, 0, \ldots, 0)^t$, and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are distributed as the same as in the models in subsection 2.2 with $\boldsymbol{G}_0 = \boldsymbol{I}_m$ *i.e.* $q = 1$. We set $\boldsymbol{Z}_i = \boldsymbol{j}_{n_i}$ in (2.1) for $\boldsymbol{j}_{n_i} = (1, \ldots, 1)^t$, an n_i-vector of ones, and $n_1 = \cdots = n_m = n = N/m$. Let $\boldsymbol{X}$ be generated as $\mathbf{vec}\,(\boldsymbol{X}^t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N \otimes \boldsymbol{\Sigma_x})$ with $\boldsymbol{\Sigma_x} = (1 - \rho_x)\boldsymbol{I}_{p_\omega} + \rho_x\boldsymbol{J}_{p_\omega}$ for $\rho_x = 0.1$ and $\boldsymbol{J}_{p_\omega} = \boldsymbol{j}_{p_\omega}\boldsymbol{j}_{p_\omega}^t$. For the true model, $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p^*}, 0, \ldots, 0)^t$ and β_j is generated as $\beta_j = 2 \times ((-1)^j/(j + 0.7)) \times U(1, 2), 1 \leq j \leq p^*$ for a uniform random variable $U(1, 2)$ on the interval $(1, 2)$.

We here handle the case that $p_\omega = 7$, $p^* = 5$, $N = 50$ and $m = 10$. The performance of the criteria cAIC and McAIC is measured by the biases of estimating cAI and by the relative frequency of selecting the true model. The true values of cAI in each model are calculated from (A.2) based on 10,000 replications. The biases and the rates of selecting each model are computed as their averages based on 1,000 replications. The results are shown in Table 1 for $\sigma^2 = 1$ and in Table 2 for $\sigma^2 = 0.5$.

model	cAI	bias(cAIC)	bias(McAIC)	hit(cAIC)	hit(McAIC)
1	217.81	16.234	-0.63453	0	0
2	184.14	12.17	-0.55772	0	0
3	167.68	7.2864	-0.25035	0.005	0.025
4	163.66	4.7962	-0.15384	0.05	0.087
5	158.84	-0.16459	-0.13939	0.778	0.812
6	160.59	-0.22447	-0.14562	0.117	0.062
7	162.47	-0.23391	-0.15563	0.05	0.014

Table 1: $N = 50, m = 10, p^* = 5, p_\omega = 7, \sigma^2 = 1$

model	cAI	bias(cAIC)	bias(McAIC)	hit(cAIC)	hit(McAIC)
1	210.62	18.465	-0.53668	0	0
2	168.35	16.2	-0.47607	0	0
3	142.9	11.564	-0.20047	0	0.001
4	134.82	8.4066	-0.083391	0.004	0.014
5	124.18	-0.16459	-0.13939	0.824	0.907
6	125.94	-0.22447	-0.14562	0.122	0.063
7	127.81	-0.23391	-0.15563	0.05	0.015

Table 2: $N = 50, m = 10, p^* = 5, p_\omega = 7, \sigma^2 = 0.5$

Tables 1 and 2 show that although the conventional cAIC has large biases for underspecified models, namely model 1 to 4, our proposed McAIC has smaller biases for both underspecified and overspecified models. Especially, because cAIC overestimates the cAI for underspecified cases, cAIC tends to select larger models. The fact that McAIC can estimate with small biases for each model may imply that this criterion provides an appropriate weight vector for model averaging methods. We will check this hypothesis in the next experiment.

In the second experiment, we handle the case of unknown $\boldsymbol{G}_0$ and consider the model class which consists of all subsets of $\{\beta_1, \ldots, \beta_{p_\omega}\}$. The other set up is the same as in the first experiment except for $\boldsymbol{G}_0 = \boldsymbol{G}(\psi) = \psi \boldsymbol{I}_m$ where $\psi = \tau^2/\sigma^2$, namely $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I}_m)$. This model is known as the nested error regression model (NERM), and σ^2 and τ^2 are estimated by unbiased estimators proposed by Prasad and Rao (1990), which is given as follows: Let $S = \boldsymbol{y}^t(\boldsymbol{I}_N - \boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t)\boldsymbol{y}$ and $S_1 = \boldsymbol{y}^t(\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{E}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{E})\boldsymbol{y}$ where $\boldsymbol{E} = \mathrm{diag}\,(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_m)$ for $\boldsymbol{E}_i = \boldsymbol{I}_{n_i} - n_i^{-1}\boldsymbol{J}_{n_i}$. Then, unbiased estimators of σ^2 and τ^2 are given by

$$\hat{\sigma}_j^2 = S_1(N - m - p_j) \quad \text{and} \quad \hat{\tau}_j^2 = \left\{S - (N - p_j)\hat{\sigma}_j^2\right\}/N^*,$$

where $N^* = N - \mathrm{tr}\,[\boldsymbol{Z}^t\boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{Z}]$. Let $p_\omega = 5$ and $p^* = 3$. The parameter $\boldsymbol{\beta}$ is generated in the same way as in the first experiment. We here measure the performance of cAIC and McAIC via $\|\hat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}\|^2/N$ for $\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}}$, which is here called the prediction error because $\hat{\boldsymbol{b}}$ is a predictor of $\boldsymbol{b}$. The prediction errors are given as averages based on 1,000 replications.

In addition to cAIC and McAIC, we consider a model averaging procedure. The aim of model averaging is to predict a future value by a weighted mean of fitted values for the candidate models. The weighting functions are important in the model averaging method, and we use some optimal weights suggested in Burnham and Anderson (2002). In the context of McAIC, the weight is defined as follows: Let $McAIC_j$ denote the value of McAIC in model M_j and let $McAIC_{min}$ be the minimum McAIC value. Also, let $\Delta McAIC_j = McAIC_j - McAIC_{min}$. Then the weight is defined by

$$w_j = \frac{\exp\left(-\frac{1}{2}\Delta McAIC_j\right)}{\sum_k \exp\left(-\frac{1}{2}\Delta McAIC_k\right)}. \tag{4.2}$$

Based on the weights given in (4.2), we can obtain a model averaged fitted value

$$\hat{\boldsymbol{y}} = \sum_j w_j \hat{\boldsymbol{y}}_j,$$

where $\hat{\boldsymbol{y}}_j$ is the predictor based on model j, and the summation is taken over all the candidate models. We call this method ”Smoothed McAIC (S-McAIC)”. A similar method based on cAIC is called ”Smoothed cAIC (S-cAIC)”.

Table 3 reports the prediction errors for the best model selected by cAIC and McAIC and for the model averaged fitted values based on S-cAIC and S-McAIC. From the table, it can be seen that McAIC and the corresponding averaging procedure S-McAIC are better than cAIC and S-cAIC. Also, it is revealed that the prediction errors get smaller as the sample size is larger. This implies that the information criteria can estimate the cAI more accurately for the large sample size.

case	1	2	3	4	5	6
N	50	50	50	80	80	80
σ^2	1	0.5	0.5	1	0.5	0.5
τ^2	0.5	1	0.5	0.5	1	0.5
cAIC	0.23775	0.13346	0.13346	0.15529	0.084461	0.081887
McAIC	0.23473	0.13072	0.13072	0.15477	0.084100	0.081608
S-cAIC	0.23162	0.13068	0.13068	0.15150	0.082597	0.080020
S-McAIC	0.22942	0.12897	0.12897	0.15061	0.082037	0.079505

Table 3: Prediction errors of the fitted values based on cAIC, McAIC, S-cAIC and S-McAIC

We recall that all the possible candidate models are treated in the second experiment, while the nested models only are considered in the first experiment. As stated above, McAIC and the model averaging procedure S-McAIC have better performance than cAIC and S-cAIC for all the possible models. Although the details are omitted here, McAIC is not necessarily better than cAIC when candidate models are nested. This observation implies that cAIC is not bad as long as nested candidate models are considered, since cAIC is justified in the overspecified case.

5 Empirical Study

We apply the variable selection procedures cAIC and McAIC to the posted land price data along the Keikyu train line, which connects the suburbs in Kanagawa prefecture to the Tokyo metropolitan area. This data set was used by Kubokawa and Nagashima (2012) who studied parametric bootstrap methods in the linear mixed models.

We analyze the land price data in 2001 with covariates for 47 stations which we consider as small areas, namely $m = 47$. For the ith small area, there are data of n_i land spots, and the total sample size is $N = \sum_{i=1}^{m} n_i = 189$. The land price (Yen in hundreds of thousands) per m^2 of the k spot in the ith small area is denoted by y_{ik}, TRN_i is the time to take by train from the station i to the Tokyo station around 9:00 in the morning, DST_{ik} is the geographical distance from the spot k to the nearby station i, $FOOT_{ik}$ is the time to take on foot from the spot k to the nearby station i and FAR_{ik} denotes the floor-area ratio of the spot k. As explanatory variables, we consider nine variables $FAR_{ik}, TRN_i, TRN_i^2, DST_{ik}, DST_{ik}^2, FOOT_{ik}^2, TRN_i \times DST_{ik}$ and $TRN_i \times FOOT_{ik}$, which are denoted by $x_1, \ldots, x_9$ and x_0 denotes constant term.

We employ NERM, which we handle in the previous section, and estimate unknown parameters σ^2 and τ^2 with the Prasad-Rao estimators. The variable selection procedure is that regressors which minimizes the information criteria are added to the model based on the forward stepwise selection.

Table 4 reports values of cAIC and McAIC in each model. Both criteria select the

model	cAIC	McAIC
x_1	469.37	476.64
x_1, x_0	436.18	441.26
x_1, x_0, x_2	417.57	420.14
x_1, x_0, x_2, x_3	410.75	412.22
x_1, x_0, x_2, x_3, x_4	412.97	413.93
$x_1, x_0, x_2, x_3, x_4, x_5$	413.82	415.29

Table 4: Example of posted land price data

same model with $\{x_1, x_0, x_2, x_3\}$, namely,

$$y_{ik} = \beta_0 + FAR_i\beta_1 + TRN_i\beta_2 + (TRN_i)^2\beta_3 + v_i + \varepsilon_{ik},$$

where the parameters are estimated by $\hat{\sigma}^2 = 0.46152, \hat{\tau}^2 = 0.077363$ and $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3) = (5.0806, 6.3661 \times 10^{-3}, -1.0542 \times 10^{-1}, 6.4769 \times 10^{-4})$.

6 Concluding Remarks

In this paper, we have suggested the McAIC which has been derived by modifying the exact cAIC by Vaida and Blanchard (2005). It has been shown that McAIC is an asymptotically unbiased estimator of the conditional Akaike information cAI for both overspecified and underspecified models, while cAIC has a large bias for underspecified models. The asymptotic unbiasedness of McAIC has been confirmed numerically by simulation.

As an application of McAIC, we have suggested the model averaging procedure with the weights based on McAIC. When all the possible subsets of the full model are treated as candidate models, it has been shown numerically that McAIC and the corresponding model averaging procedure have better performance than cAIC.

A Appendix

A.1 Derivation of (3.2)

First compute the expectation with respect to $\tilde{\boldsymbol{y}} \sim f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\theta})$ in cAI. Then, cAI can be written as

$$\begin{aligned} cAI =& E\big[N\log(2\pi\hat{\sigma}_j^2) + N\sigma^2/\hat{\sigma}_j^2\big] \\ &+ E\Big[\big\{(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \boldsymbol{b})\big\}^t\big\{(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \boldsymbol{b})\big\}\Big/\hat{\sigma}_j^2\Big]. \quad \text{(A.1)} \end{aligned}$$

Note that $\boldsymbol{b}|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{G}_0\boldsymbol{Z}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{u}, \sigma^2(\boldsymbol{G}_0 - \boldsymbol{G}_0\boldsymbol{Z}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{Z}\boldsymbol{G}_0)\right)$ and that

$$\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \boldsymbol{b}) = (\boldsymbol{I}_N - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{u} + \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*) - \boldsymbol{Z}\boldsymbol{b}.$$

Taking the expectation with respect to $\boldsymbol{b}|\boldsymbol{y} \sim \pi(\boldsymbol{b}|\boldsymbol{y},\boldsymbol{\theta})$ in (A.1), we rewrite cAI as

$$\begin{aligned} cAI =& E\big[N\log(2\pi\hat{\sigma}_j^2) + (2N - \mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])\sigma^2/\hat{\sigma}_j^2\big] \\ &+ E\big[(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*)^t\boldsymbol{\Sigma}_0^{-2}(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*)/\hat{\sigma}_j^2\big]. \end{aligned} \tag{A.2}$$

Next, in a part of $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\theta}}_j)$ in (3.1), it is noted that

$$\begin{aligned} &\boldsymbol{y} - \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j \\ =&\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* - (\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*) - \boldsymbol{Z}\boldsymbol{G}_0\boldsymbol{Z}^t\boldsymbol{\Sigma}_0^{-1}\Big\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* - (\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*)\Big\} \\ =&\boldsymbol{\Sigma}_0^{-1}\boldsymbol{u} - \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}\boldsymbol{\beta}^*). \end{aligned} \tag{A.3}$$

Thus, from (A.2) and (A.3), we can see that $\Delta_{cAI} = cAI - E[-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\theta}}_j)]$ is expressed as (3.2). □

A.2 Proof of Theorem 3.1

For (3.6), we decompose Δ_{cAI} as

$$\Delta_{cAI} = b_1 + b_2 + b_3 + b_4,$$

where $b_1 = NE[(2N - \mathrm{tr}\,\boldsymbol{\Sigma}_0^{-1})/(W_0+W_1)]$, $b_2 = -NE[\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}/(W_0+W_1)]$, $b_3 = 2NE[\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}/(W_0+W_1)]$ and $b_4 = -2NE[\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}/(W_0+W_1)]$.

We begin by expanding $(W_0+W_1)^{-1}$ up to $O_p(N^{-2})$. Let $W = \boldsymbol{v}^t(\boldsymbol{I}_N - \boldsymbol{M}_\omega)\boldsymbol{v} + \boldsymbol{v}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v}$. Then, $W_0 + W_1 = W + 2L + N\delta$, $W \sim \chi^2_{N-p_j}$, $W = O_p(N)$ and $L = O_p(N^{1/2})$, so that we can write $(2L+N\delta)/W = \delta + D$ and $D = O_p(N^{-1/2})$. Thus, it follows that

$$\begin{aligned} (W_0+W_1)^{-1} =&(W+2L+N\delta)^{-1} = W^{-1}(1+\delta+D)^{-1} = \frac{\lambda}{W}(1+D\lambda)^{-1} \\ =&\frac{\lambda}{W}\left\{1 - D\lambda + (D\lambda)^2 + O_p(N^{-3/2})\right\}. \end{aligned} \tag{A.4}$$

Since $\delta\lambda = 1-\lambda$, it is seen that

$$D\lambda = (1 - \frac{N}{W})(\lambda - 1) + \frac{2L\lambda}{W}.$$

Let $A = W/(N-p_j) - 1$, which is of $O_p(N^{-1/2})$. then,

$$\begin{aligned} \frac{1}{W} =&\frac{1}{N-p_j}\left\{1 - A + A^2 + O_p(N^{-3/2})\right\} \\ =&\frac{1}{N}(1 - A + A^2 + \frac{p_j}{N}) + O_p(N^{-5/2}), \\ 1 - \frac{N}{W} =&A - A^2 - \frac{p_j}{N} + O_p(N^{-3/2}). \end{aligned} \tag{A.5}$$

Hence, $(W_0+W_1)^{-1}$ can be evaluated as

$$(W_0+W_1)^{-1}=\frac{\lambda}{N}\left\{1-(A+\frac{2L}{N})\lambda+A^2\lambda^2+\frac{p_j\lambda+4AL\lambda^2}{N}+\frac{4L^2\lambda^2}{N^2}\right\}+O_p(N^{-5/2}). \tag{A.6}$$

For any function $q(\cdot)$, we have $E[q(\boldsymbol{v}^t\boldsymbol{G}\boldsymbol{v})L]=0$ since $q(\boldsymbol{v}^t\boldsymbol{G}\boldsymbol{v})L$ is an odd function of $\boldsymbol{v}$. Also, $E[A]=0$, $E[A^2]=2/(N-p_j)$, $E[L^2]=N\delta$. Hence, it is observed that

$$E[(W_0+W_1)^{-1}]=\frac{\lambda}{N}\left\{1+\frac{-2\lambda^2+4\lambda+p_j\lambda}{N}\right\}+O(N^{-3}). \tag{A.7}$$

Using the expansions (A.6) and (A.7) , we can evaluate b_1, b_2, b_3 and b_4, respectively.

First, b_1 can be evaluated as

$$\begin{aligned}b_1 =&N(2N-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])\times\frac{\lambda}{N}\times\left\{1+\frac{-2\lambda^2+4\lambda+p_j\lambda}{N}\right\}+O(N^{-1})\\ =&b_1^*+(2N-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])(\lambda-1)\left\{\frac{N}{N-p_j-2}+\frac{-2\lambda^2+(p_j+2)\lambda}{N}\right\}+O(N^{-1}),\end{aligned}$$

where

$$b_1^*=\frac{N(2N-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])}{N-p_j-2},$$

which is the exact b_1 for overspecified models.

Next note that $\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]=O(N)$, $\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]=O_p(N^{1/2})$. Then, b_2 is evaluated as

$$\begin{aligned}b_2=&-NE\Bigg[\left\{\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]+(\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}])\right\}\\ &\times\frac{\lambda}{N}\left\{1-(A+\frac{2L}{N})\lambda+A^2\lambda^2+\frac{p_j\lambda+4AL\lambda^2}{N}+\frac{4L^2\lambda^2}{N^2}\right\}\Bigg]+O(N^{-1})\\ =&-\lambda\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]\left\{1+\frac{-2\lambda^2+4\lambda+p_j\lambda}{N}\right\}+\lambda^2E[\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}A]+O(N^{-1}).\end{aligned}$$

From the second order moment of quadratic forms of standard normal random vectors, it follows that

$$E[\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v}A]=\frac{2}{N-p_j}\left\{\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j]\right\}.$$

Using this equality, we can evaluate b_2 as

$$\begin{aligned}b_2=&-\lambda\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]\left\{1+\frac{-2\lambda^2+2\lambda+p_j\lambda}{N}\right\}+O(N^{-1})\\ =&b_2^*-\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}](\lambda-1)\left\{\frac{N}{N-p_j-2}+\frac{-2\lambda^2+p_j\lambda-2}{N}\right\}+O(N^{-1}),\end{aligned}$$

where

$$b_2^* = -N \times \left\{ \frac{\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}]}{N-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}] - 2\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j]}{(N-p_j)(N-p_j-2)} \right\},$$

which is the exact b_2 for overspecified models.

As for b_3, it can be decomposed as

$$\frac{\boldsymbol{v}^t\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}}{W_0+W_1} = \frac{\boldsymbol{v}^t\boldsymbol{M}_j\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}}{W_0+W_1} + \frac{\boldsymbol{v}^t(\boldsymbol{I}_N-\boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}}{W_0+W_1}.$$

Since $\boldsymbol{M}_j\boldsymbol{v}$ is independent of $(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{v}$ and W_0+W_1, and $E[\boldsymbol{M}_j\boldsymbol{v}] = 0$, it follows that

$$E\left[\frac{\boldsymbol{v}^t(\boldsymbol{I}_N-\boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}}{W_0+W_1}\right] = 0.$$

Further, because $\boldsymbol{v}^t\boldsymbol{M}_j\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}$ and W_0+W_1 are mutually independent, b_3 is evaluated as

$$\begin{aligned} b_3 =& 2N \times E[\boldsymbol{v}^t\boldsymbol{M}_j\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j\boldsymbol{v}] \times E[(W_0+W_1)^{-1}] \\ =& b_3^* + 2\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j](\lambda-1) + O(N^{-1}), \end{aligned}$$

where

$$b_3^* = \frac{2N\mathrm{tr}\,[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{M}_j]}{N-p_j-2},$$

which is the exact b_3 for overspecified models.

Finally, we evaluate b_4. Note that $\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v} = O_p(N^{1/2})$ from the assumption (A2). Then, b_4 can be expanded as

$$\begin{aligned} b_4 =& -2N \times E\left[\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v} \times \frac{\lambda}{N}\left\{1 - (A + \frac{2L}{N})\lambda\right\}\right] + O(N^{-1}) \\ =& \frac{4\lambda^2}{N}\boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta} + O(N^{-1}). \end{aligned}$$

Combining the evaluations of b_1, b_2, b_3 and b_4 yields the result in (3.8), where B^* is defined by $B^* = b_1^* + b_2^* + b_3^*$. □

A.3 Proof of Lemma 3.1

It follows from $W_0 = N + O_p(N^{1/2})$ and (A.6) that

$$E\left[\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^k\right] = E\left[\left(\frac{W_0}{W_0+W_1}\right)^k\right] \to \lambda^k \quad (N \to \infty), \tag{A.8}$$

which proves lemma 3.1. □

A.4 Proof of Lemma 3.2

In the overspecified case, it follows from Lemma 3.1 that $E[\hat{\lambda}] = E[\widehat{\lambda^2}] = E[\widehat{\lambda^3}] = 1$, so that $E[\widehat{B_1}] = 0$. In the underspecified case, we shall check (3.15). Using the expansion (A.6) of $(W_0 + W_1)^{-1}$, we can approximate $E[\hat{\lambda}]$ as

$$\begin{aligned} E[\hat{\lambda}] =& \frac{N-p_j}{N-p_\omega} \times E\left[\frac{W_0}{W_0+W_1}\right] \\ =& \frac{N-p_j}{N-p_\omega}\frac{\lambda}{N} \times E\Bigg[\left\{(N-p_\omega) + \boldsymbol{v}^t(\boldsymbol{I}_N - \boldsymbol{M}_\omega)\boldsymbol{v} - (N-p_\omega)\right\} \\ & \times \left\{1 - (A + \frac{2L}{N})\lambda + A^2\lambda^2 + \frac{p_j\lambda + 4AL\lambda^2}{N} + \frac{4L^2\lambda^2}{N^2}\right\}\Bigg] + O(N^{-2}). \end{aligned} \tag{A.9}$$

Evaluating (A.9) up to $O(N^{-1})$, we can get (3.15) and Lemma 3.2. □

A.5 Proof of Lemma 3.3

Let $c_1 = \operatorname{tr}[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)]$, $c_2 = \operatorname{tr}[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_\omega)]$ and

$$\begin{aligned} D_1 =& \boldsymbol{v}^t(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{v}, \\ D_2 =& 2\boldsymbol{\eta}^t(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{I}_N - \boldsymbol{M}_j)\boldsymbol{v}, \\ D_3 =& \boldsymbol{\eta}^t(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\eta}. \end{aligned} \tag{A.10}$$

Since $\tilde{\sigma}_j^2 = \sigma^2(D_1 + D_2 + D_3)$, we can rewrite $\widetilde{B_3}$ as

$$\begin{aligned} \widetilde{B_3} =& \frac{4W_0(D_1+D_2+D_3)}{(W_0+W_1)^2} - \frac{4c_1}{N}\frac{W_0^2}{(W_0+W_1)^2} \\ =& \widetilde{B_{31}} - \widetilde{B_{32}}. \quad \text{(say)} \end{aligned}$$

From the exapansion (A.6) of $(W_0 + W_1)^{-1}$, it follows that

$$\begin{aligned} E[\widetilde{B_{31}}] =& \frac{4\lambda^2}{N^2} \times \{E[W_0D_1] + E[W_0D_3]\} + O(N^{-1}) \\ =& \frac{4c_1\lambda^2}{N} + \frac{4\lambda^2 D_3}{N} + O(N^{-1}), \\ E[\widetilde{B_{32}}] =& \frac{4c_1}{N} \times \frac{\lambda^2}{N^2} \times E[W_0^2] + O(N^{-1}) \\ =& \frac{4c_1\lambda^2}{N} + O(N^{-1}), \end{aligned}$$

which proves Lemma 3.3. □

A.6 Proof of Lemma 3.4

It is noted that the adjustment term of $\widehat{B_3}$ is of order $O(N^{-1})$ from (3.17). Then it follows from Lemma 3.3 that $E[\widehat{B_3}] = B_3 + O(N^{-1})$. Thus it is sufficient to evaluate $E[\widetilde{B_3}]$ up to $O(N^{-1})$ for overspecified models.

In the overspecified case, it is noted that $N\hat{\sigma}_j^2 = \sigma^2 W$ and $\tilde{\sigma}_j^2 = \sigma^2 D_1$ for D_1 given in (A.10). Then,

$$\begin{aligned}\widetilde{B_3} =& 4 \times \frac{W_0 \times D_1}{W^2} - \frac{4c_1}{N}\left(\frac{W_0}{W}\right)^2 \\ =& \widetilde{B_{33}} - \widetilde{B_{34}}. \quad \text{(say)}\end{aligned}$$

From (A.5), W^{-2} is expanded as

$$\frac{1}{W^2} = \frac{1}{N^2}\left(1 - 2A + 3A^2 + \frac{2p_j}{N}\right) + O_p(N^{-7/2}).$$

Thus, $E[\widetilde{B_{33}}]$ is written as

$$\begin{aligned}E[\widetilde{B_{33}}] =& \frac{4}{N^2}E\Big[\{(N-p_\omega) + W_0 - (N-p_\omega)\}\{c_1 + (D_1 - c_1)\} \\ & \times \{1 - 2A + 3A^2 + \frac{2p_j}{N}\}\Big] + O(N^{-2}) \\ =& \frac{4}{N^2} \times \Big\{(N-p_\omega)c_1\big(1 + 3E[A^2] + \frac{2p_j}{N}\big) \\ & - 2c_1 E[W_0 A] - 2(N-p_\omega)E[D_1 A] + E[W_0 D_1] - c_1(N-p_\omega)\Big\} + O(N^{-2}) \\ =& \frac{4(N - p_\omega + 2p_j)c_1}{N^2} + \frac{8(c_2 - c_1)}{N^2} + O(N^{-2}), \qquad \text{(A.11)}\end{aligned}$$

since $W_0 - (N - p_\omega) = O_p(N^{1/2})$, $c_1 = O(N)$ and $D_1 - c_1 = O_p(N^{1/2})$. Noting that $W_0/W \sim Be((N-p_\omega)/2, (p_\omega - p_j)/2)$, we can evaluate $E[\widetilde{W_{34}}]$ as

$$\begin{aligned}E[\widetilde{B_{34}}] =& \frac{4c_1}{N}\frac{(N-p_\omega)(N-p_\omega+2)}{(N-p_j)(N-p_j+2)} \\ =& \frac{4(N - 2p_\omega + 2p_j)c_1}{N^2} + O(N^{-2}). \qquad \text{(A.12)}\end{aligned}$$

Combining (A.11) and (A.12) gives

$$E[\widetilde{B_3}] = \frac{4c_1 p_\omega}{N^2} + \frac{8(c_2 - c_1)}{N^2} + O(N^{-2}),$$

which shows Lemma 3.4. □

Acknowledgments.

Research of the second author was supported in part by Grant-in-Aid for Scientific Research (21540114 and 23243039) from Japan Society for the Promotion of Science.

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] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*, 2nd ed. New York: Springer.

[4] Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, **98**, 685-700.

[5] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707-716.

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

[7] Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed model. *J. Multivariate Anal.* **102**, 641-660.

[8] Kubokawa, T. and Nagashima, B. (2012). Parametric Bootstrap methods for bias correction in linear mixed models. *J. Multivariate Anal.* **106**, 1-16.

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

[10] Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

[11] Srivastava, M.S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *J. Multivariate Anal.*, **101**, 1970-1980.

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