CIRJE-F-944

Conditional AIC under Covariate Shift with Application to Small Area Prediction

Yuki Kawakubo
The University of Tokyo
Shonosuke Sugasawa
The University of Tokyo
Tatsuya Kubokawa
The University of Tokyo

October 2014

Conditional AIC under Covariate Shift with Application to Small Area Prediction

Yuki Kawakubo*, Shonosuke Sugasawa†and Tatsuya Kubokawa‡
University of Tokyo

October 21, 2014

Abstract

In this paper, we consider the problem of selecting explanatory variables of fixed effects in linear mixed models under covariate shift, which is the situation that the values of covariates in the predictive model are different from those in the observed model. We construct a variable selection criterion based on the conditional Akaike information introduced by Vaida and Blanchard (2005) and the proposed criterion is generalization of the conditional Akaike information criterion (conditional AIC) in terms of covariate shift. We especially focus on covariate shift in small area prediction and show usefulness of the proposed criterion through simulation studies.

Key words and phrases: Akaike information criterion, conditional AIC, covariate shift, linear mixed model, small area estimation, variable selection.

1 Introduction

Linear mixed models have been studied for a long time theoretically, and also have many applications, for example longitudinal data analysis in biostatistics, panel data analysis in econometrics, small area estimation in official statistics, and others. The problem of selecting explanatory variables in linear mixed models is important and many literatures have investigated this problem. Müller et al. (2013) is a good survey about the model selection in linear mixed models.

When the purpose of the variable selection is to find a set of significant variables for a good prediction, Akaike-type information criteria (Akaike; 1973, 74) are well-known methods. However, the Akaike information criterion (AIC) based on the marginal likelihood,

*Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, Research Fellow of Japan Society for the Promotion of Science, E-Mail: y.k.5.58.2010@gmail.com

†Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: shonosuke622@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

which integrates out the likelihood with respect to random effects, is not appropriate when the prediction is focused on random effects. Then, Vaida and Blanchard (2005) proposed to consider the Akaike-type information based on the conditional density given the random effects and proposed the conditional AIC. To give a brief explanation about the concept of the conditional AIC, we introduce some notations as follows. Let $\boldsymbol{y}$ be an observable random vector of the response variables, $\boldsymbol{\theta}$ be a vector of the unknown parameters, $\boldsymbol{b}$ be a random vector of the random effects. The conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}$ is denoted by $f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\theta})$, and the density function of $\boldsymbol{b}$ is denoted by $\pi(\boldsymbol{b}|\boldsymbol{\theta})$. Then, Vaida and Blanchard (2005) proposed to measure the prediction risk of the plug-in predictive density $f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}},\widehat{\boldsymbol{\theta}})$ relative to the Kullback–Leibler divergence given as follows:

$$\iint \left[\int \log \left\{ \frac{f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\theta})}{f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}},\widehat{\boldsymbol{\theta}})} \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\theta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\theta})\pi(\boldsymbol{b}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b}, \tag{1}$$

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ given $\boldsymbol{b}$, and $\hat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{\theta}}$ is some predictor or estimator of $\boldsymbol{b}$ and $\boldsymbol{\theta}$, respectively. The conditional AIC is an (asymptotically) unbiased estimator of a part of the risk in (1), which is called the conditional Akaike information (cAI) given as follows:

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

The conditional AIC as the variable selection criterion in linear mixed models has been studied by Liang et al. (2008), Greven and Kneib (2010), Srivastava and Kubokawa (2010), Kubokawa (2011), Kubokawa and Nagashima (2012), Kawakubo and Kubokawa (2014) and others. Furthermore, the conditional AIC has been constructed as a variable selection criterion in generalized linear mixed models by Donohue et al. (2011), Yu and Yau (2012), Yu et al. (2013), Saefken et al. (2014) and others.

Considering the prediction problem, it is often the case that the values of covariates in the predictive model are different from those in the observed model, which we call covariate shift. We here call the model in which $\boldsymbol{y}$ is the vector of the response variables the 'observed model', and call the model in which $\tilde{\boldsymbol{y}}$ is the vector of the response variables the 'predictive model'. It is noted that the terminology 'covariate shift' was first used by Shimodaira (2000), who defined it as the situation that the distribution of the covariates in the predictive model is different from that in the observed model. In this paper, though we treat the covariates as non-random, we use the same terminology 'covariate shift' as Shimodaira (2000). Even when the information about the covariates in the predictive model can be used, most variable selection criteria do not use it. This is because most criteria put the assumption that the predictive model is the same as the observed model. As for the conditional AIC explained above, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{b}$ and that of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$ are the same, and both of them are denoted by $f(\cdot|\boldsymbol{b},\boldsymbol{\theta})$. On the other hand, under the covariate shift, the conditional density of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$ is different from that of $\boldsymbol{y}$ given $\boldsymbol{b}$ and is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\theta})$. When the aim of the variable selection is to choose the best predictive model, it is not appropriate to use the covariates only in the observed model. Then we redefine the cAI under covariate shift as follows:

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

and construct an information criterion as an unbiased estimator of the cAI. The proposed criterion includes the original conditional AIC by Vaida and Blanchard (2005).

The term 'prediction' in this context includes not only future forecast but also interpolation. We especially focus on covariate shift in the context of small area prediction which is based on finite-super population model. We consider the situation that we are interested in finite population mean of some characteristic and that some values in the population are observed through some sampling procedure. When the sample size is small, the problem is called small area estimation. For the detail about small area estimation, see Rao (2003), Datta and Ghosh (2012), Pfeffermann (2013) and others. The model based approach in small area estimation often assumes that the finite population which has the super-population with random effects and borrow the strength from other areas to estimate (predict) the small area (finite population) mean. The well-known unit level model is the nested error regression model, which is a kind of linear mixed models and discussed in Battese et al. (1988). The nested error regression model can be used when the values of the auxiliary variables for the units whose values of characteristic of interest (response variable in the model) are observed through survey sampling. This is the observed model in the framework of our variable selection procedure. On the other hand, we consider an area level model as the predictive model, which can be used under the situation that each mean of the auxiliary variables are known for each small area. This situation is often the case in official statistics and the model introduced by Fay and Herriot (1979) is often used in this case.

The rest of this paper is organized as follows. In Section 2, the setup of the problem is explained and the variable selection criterion for the problem is proposed. In Section 3, we give an example of covariate shift, which is focused on small are prediction, and investigate the numerical performance of the problem. In Section 4, concluding remarks are given. All the proofs are given in the Appendix.

2 Problem of selecting variables

2.1 Model focus

The observed model we treat is the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{y}$ is an n-variate observation vector of response variables, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times p$ and $n \times r$ matrices of covariates, respectively, $\boldsymbol{\beta}$ is a p-variate vector of regression coefficients, $\boldsymbol{b}$ is an r-variate vector of random effects, and $\boldsymbol{\varepsilon}$ is an n-variate vector of random errors. Let $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ be mutually independent and $\boldsymbol{b} \sim \mathcal{N}_r(\boldsymbol{0}, \sigma^2\boldsymbol{G})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{R})$, where $\boldsymbol{G}$ and $\boldsymbol{R}$ are $r \times r$ and $n \times n$ positive definite matrices and σ^2 is a scalar. We assume that $\boldsymbol{G}$ and $\boldsymbol{R}$ are known and handle the two cases that σ^2 is known and unknown. The marginal distribution of $\boldsymbol{y}$ is $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^t + \boldsymbol{R}$. The conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}$ is denoted by $f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)$, and the density of $\boldsymbol{b}$ is $\pi(\boldsymbol{b}|\sigma^2)$.

The predictive model is the linear mixed model which has the same regression coeffi-

cients $\boldsymbol{\beta}$ and random effects $\boldsymbol{b}$ as in the observed model, but different covariates, namely

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{Z}}\boldsymbol{b} + \tilde{\boldsymbol{\varepsilon}}, \tag{3}$$

where $\tilde{\boldsymbol{y}}$ is an m-variate random vector of the target of prediction, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Z}}$ are $m \times p$ and $m \times r$ matrices of covariates, and $\tilde{\boldsymbol{\varepsilon}}$ is an m-variate vector of random errors, independent of $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$, and distributed as $\tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}_m(\boldsymbol{0}, \sigma^2\widetilde{\boldsymbol{R}})$, where $\widetilde{\boldsymbol{R}}$ is a known $m \times m$ positive definite matrix. We assume that we know the values of $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Z}}$ in the predictive model and that they are not necessarily the same as those of $\boldsymbol{X}$ and $\boldsymbol{Z}$ in the observed model. We call this situation covariate shift. The marginal distribution of $\tilde{\boldsymbol{y}}$ is $\tilde{\boldsymbol{y}} \sim \mathcal{N}_m(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}, \widetilde{\boldsymbol{\Sigma}})$, where $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Z}}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^t + \widetilde{\boldsymbol{R}}$. The conditional density function of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$ is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)$.

The regression coefficient $\boldsymbol{\beta}$ and the random effect $\boldsymbol{b}$ are estimated by the maximum likelihood estimator and the empirical Bayes estimator, respectively, given as follows:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} =& (\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{y},\\ \hat{\boldsymbol{b}} =& \boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).\end{aligned}$$

When the variance parameter σ^2 is unknown, we consider to estimate it by the maximum likelihood estimator given by

$$\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n. \tag{4}$$

2.2 Proposed Criterion

We now derive the conditional Akaike information criterion under the covariate shift in the two cases of known and unknown σ^2.

[σ^2 **is known**] Firstly we consider the simple case that σ^2 is known. Because -2 times logarithm of the plug-in predictive density is

$$-2\log\{g(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}, \widehat{\boldsymbol{\beta}})\} = m\log(2\pi\sigma^2) + (\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})/\sigma^2,$$

the cAI is expressed as

$$\mathrm{cAI} = m\log(2\pi\sigma^2) + E^{\boldsymbol{y},\boldsymbol{b}}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}}\left[(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})\right]/\sigma^2,$$

where $E^{\boldsymbol{y},\boldsymbol{b}}$ and $E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}}$ denote the expectation with respect to the joint distribution of $(\boldsymbol{y}, \boldsymbol{b})$ and the conditional distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$. Then the conditional AIC under covariate shift (CScAIC) is defined by a bias corrected unbiased estimator of the cAI as follows:

$$\mathrm{CScAIC} = m\log(2\pi\sigma^2) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})/\sigma^2 + B_c, \tag{5}$$

where B_c is bias correction given by

$$\begin{aligned}B_c =& \mathrm{cAI} - E\left[m\log(2\pi\sigma^2) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})/\sigma^2\right]\\ =& E\Big[(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}}) - (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})\Big]/\sigma^2,\end{aligned} \tag{6}$$

which can be exactly evaluated in the following theorem.

Theorem 1 *When the variance parameter σ^2 is known, the bias correction B_c of CScAIC in (6) is*

$$\begin{aligned} B_c =& \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}] + \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t] \\ & - \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] + \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}], \end{aligned} \tag{7}$$

where $\boldsymbol{\Lambda} = \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^t$.

[σ^2 **is unknown**] Next we handle the case that σ^2 is unknown and estimated by the maximum likelihood estimator (4). In this case, the cAI is expressed as

$$\mathrm{cAI} = E^{\boldsymbol{y},\boldsymbol{b}}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}}\left[m\log(2\pi\hat{\sigma}^2) + (\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})/\hat{\sigma}^2\right].$$

Then the covariate shift conditional AIC is defined by a bias corrected unbiased estimator of the cAI as follows:

$$\mathrm{CScAIC} = m\log(2\pi\hat{\sigma}^2) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})/\hat{\sigma}^2 + B_c^*, \tag{8}$$

where B_c^* is bias correction given by

$$\begin{aligned} B_c^* =& cAI - E\left[m\log(2\pi\hat{\sigma}^2) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})/\hat{\sigma}^2\right] \\ =& E\Big[(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})/\hat{\sigma}^2 - (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})/\hat{\sigma}^2\Big], \end{aligned} \tag{9}$$

which can be exactly evaluated in the following theorem.

Theorem 2 *When the variance parameter σ^2 is unknown, the bias correction B_c^* of CScAIC in (9) is*

$$\begin{aligned} B_c^* =& \frac{n}{n-p-2}\left\{\mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}] + \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t]\right\} \\ & + \frac{n}{n-p}\left\{-\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] + \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}]\right\}, \end{aligned} \tag{10}$$

Next corollary shows that our covariate shift conditional AIC includes the conditional AIC by Vaida and Blanchard (2005) as special case.

Corollary 1 *Suppose that covariate shift does not occur, namely $\widetilde{\boldsymbol{X}} = \boldsymbol{X}, \widetilde{\boldsymbol{Z}} = \boldsymbol{Z}$ and $n = m$. In addition, let the covariance matrix of $\boldsymbol{\varepsilon}$ and $\tilde{\boldsymbol{\varepsilon}}$ be both $\sigma^2\boldsymbol{I}_n$, namely $\boldsymbol{R} = \widetilde{\boldsymbol{R}} = \boldsymbol{I}_n$. Then the bias corrections of the covariate shift conditional AIC in (7) and (10) are reduced to*

$$B_c = 2n - 2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}] + 2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}], \tag{11}$$

$$B_c^* = \frac{2n^2}{n-p-2} + \frac{2n(n-p-1)}{(n-p)(n-p-2)}\left\{-\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}] + \mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}]\right\}, \tag{12}$$

which are identical to the bias corrections of the conditional AIC by Theorem 1 and 2 in Vaida and Blanchard (2005).

3 Examples

3.1 Covariate shift in small area prediction

A typical example of the covariate shift situation appears in small area prediction problem. The model for small area prediction supposes that the observed small area data have the finite population which has the super-population model with random effects, one of which is the well-known nested error regression model (Battese *et al.*, 1988). Let Y_{ij} and $\boldsymbol{x}_{ij}$ denote the value of a characteristic of interest and its p-dimensional auxiliary variable for the j-th unit of the i-th finite population where $i = 1, \ldots, k, j = 1, \ldots, N_i$. Then, the nested error regression model is

$$Y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, N_i,$$

where $\boldsymbol{\beta}$ is a p-variate vector of regression coefficients, b_i is a random effect for the i-th finite population and b_i's and ε_{ij}'s are mutually independently distributed as $b_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. We consider the situation that only n_i value of the Y_{ij}'s are observed through some sampling procedure. We define the number of the unobserved variables in the i-th population by $N_i - n_i = m_i$ and let $n = n_1 + \cdots + n_k, m = m_1 + \cdots + m_k$. Suppose, without loss of generality, the first n_i elements of $\{Y_{i1}, \ldots, Y_{i,N_i}\}$ are observed, which are denoted by $y_1, \ldots, y_{i,n_i}$, and $Y_{i,n_i+1}, \ldots, Y_{i,N_i}$ are unobserved. Then the observed model is defined as

$$y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \tag{13}$$

which corresponds to (2) with $\boldsymbol{y} = (\boldsymbol{y}_1^t, \ldots, \boldsymbol{y}_k^t)^t$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n_i})^t$, $\boldsymbol{X} = (\boldsymbol{X}_1^t, \ldots, \boldsymbol{X}_k^t)^t$ for $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})^t$, $\boldsymbol{Z} = \mathrm{diag}\,(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k)$ for $\boldsymbol{Z}_i = \boldsymbol{j}_{n_i}$, $\boldsymbol{G} = \psi \boldsymbol{I}_k$ and $\boldsymbol{R} = \boldsymbol{I}_n$, where $\boldsymbol{j}_{n_i}$ denotes an n_i-vector of ones and $\psi = \tau^2/\sigma^2$. Note that $r = k$. In the derivation of our proposed criterion, we have assumed that the covariance matrix of $\boldsymbol{b}$ is $\sigma^2 \boldsymbol{G}$ for a known matrix $\boldsymbol{G}$. However in the nested error regression model, $\boldsymbol{G}$ includes the parameter ψ, which is usually unknown and has to be estimated. In this case, we propose that $\boldsymbol{G}$ in the bias correction should be replaced with its plug-in estimator $\boldsymbol{G}(\widehat{\psi})$. The influence caused by the replacement may be limited because ψ is the nuisance parameter when one is interested in selecting only explanatory variables. Kawakubo and Kubokawa (2014) discussed the problem in their remark 3.1.

In the problem of small area prediction, we often encounter the situation where all $\boldsymbol{x}_{ij}$'s are not observed but the area mean $\bar{\boldsymbol{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$ is known and we are interested in predicting $\overline{Y}_i$, which is the mean of finite population $\{Y_{i1}, \ldots, Y_{i,N_i}\}$, by using the value of $\bar{\boldsymbol{x}}_i$. Then the predictive model can be defined as

$$\overline{Y}_{i(u)} = \bar{\boldsymbol{x}}_{i(u)}^t \boldsymbol{\beta} + b_i + \bar{\varepsilon}_{i(u)}, \quad i = 1, \ldots, k, \tag{14}$$

where $\overline{Y}_{i(u)} = m_i^{-1} \sum_{j=n_i+1}^{N_i} y_{ij}$, the mean of unobserved variables, $\bar{\boldsymbol{x}}_{i(u)} = m_i^{-1} \sum_{j=n_i+1}^{N_i} \boldsymbol{x}_{ij}$, calculated from $\bar{\boldsymbol{x}}_i$ and $(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$, and $\bar{\varepsilon}_{i(u)} = m_i^{-1} \sum_{j=n_i+1}^{N_i} \varepsilon_{ij}$ distributed as $\mathcal{N}(0, \sigma^2/m_i)$. The model (14) corresponds to (3) with $\tilde{\boldsymbol{y}} = (\overline{Y}_{1(u)}, \ldots, \overline{Y}_{k(u)})^t$, $\widetilde{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_{1(u)}, \ldots, \bar{\boldsymbol{x}}_{k(u)})^t$, $\widetilde{\boldsymbol{Z}} = \boldsymbol{I}_k$ and $\widetilde{\boldsymbol{R}} = \mathrm{diag}\,(\widetilde{R}_1, \ldots, \widetilde{R}_k)$ for $\widetilde{R}_i = 1/m_i$. After selecting explanatory variables

with our proposed criterion, we predict $\overline{Y}_{i(u)}$ by the empirical best linear unbiased predictor $\widehat{\overline{Y}}_{i(u)} = \bar{\boldsymbol{x}}^t_{i(u)}\widehat{\boldsymbol{\beta}} + \hat{b}_i$ and obtain a predictor of the mean of finite polulation, $\overline{Y}_i$, as

$$\widehat{\overline{Y}}_i = \frac{1}{N_i}\sum_{i=1}^{n_i} y_{ij} + \frac{m_i}{N_i}\widehat{\overline{Y}}_{i(u)}. \tag{15}$$

Thus, covariate shift appears in standard models for small area prediction and the proposed criterion is important and useful in such a situation.

3.2 Simulation Study

In this subsection, we investigate numerical performances of the small area prediction problem explained in the previous subsection. We consider the nested error regression model and we use the same notation in the previous subsection. The observed vector $\boldsymbol{y}$ is generated by the observed model (13) with $p = 5$, $n_i = 3$, $k = 30$ (so that $n = 90$), for $i = 1, \ldots, k$ and $\tau^2 = \sigma^2$. Let $\boldsymbol{X}$ be generated as $\mathbf{vec}\,(\boldsymbol{X}^t) \sim \mathcal{N}(4\boldsymbol{j}_{pn}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_x)$ for $\boldsymbol{\Sigma}_x = 0.9\boldsymbol{I}_p + 0.1\boldsymbol{J}_p$ where $\boldsymbol{J}_p = \boldsymbol{j}_p\boldsymbol{j}_p^t$, and fixed through the simulation. The true coefficient vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, 0, 0)^t$ where $\beta_l, l = 1, 2, 3$ is generated as $\beta_l = U(1, 2)$ for a uniform random variable $U(1, 2)$ on the interval $(1, 2)$, and true variance is $\sigma^2 = 1$.

We consider the predictive model (14) with $m_i = 10$. We generated samples $\boldsymbol{x}_{ij}$, $j = n_i + 1, \ldots, N_i$ independently from $\mathcal{N}(a\boldsymbol{j}_p, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ for $a = 2, 4, 6$ and calculate $\boldsymbol{x}_{i(u)}$, fixed through simulation. Moreover, using generated $\boldsymbol{x}_{ij}$'s, we generate $y_{ij}, j = n_i, \ldots, N_i$ and calculate $\overline{Y}$ from generated samples, which is simulated mean of finite population. Under this settings, we investigate selection rates of the true model and calculate simulated prediction error of the best model chosen by our CScAIC defined as (9) and (10), and the original conditional AIC (cAIC), whose bias correction is (12). The prediction error is measured by

$$\|\widehat{\overline{Y}} - \overline{Y}\|^2, \qquad \overline{Y} = (\overline{Y}_1, \ldots, \overline{Y}_k)^t,$$

where $\widehat{\overline{Y}} = (\widehat{\overline{Y}}_1, \ldots, \widehat{\overline{Y}}_k)^t$ and $\widehat{\overline{Y}}_i$ for $i = 1, \ldots, k$ is calculated from (15). The prediction errors are given as averages based on 1000 replications.

Table 1 reports the selecting rates and prediction errors for the best model selected by two criteria. The values in parentheses are the improvement over the prediction error by the cAIC procedure expressed in percentage. As a candidate model, we consider the all $2^5 - 1$ combination of explanatory variables, but the only four models were selected in our simulation, so that we report the result regarding four models. M_1, M_2, M_3 and M_4 denotes the models with explanatory variables $\{1, 2, 3\}$, $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$ and $\{1, 2, 3, 4, 5\}$, respectively, and M_1 is the true model. From the table, it can be seen that the CScAIC is better than the cAIC in both cases. It is valuable to point out that the prediction error of the mean of finite population can be improved by using our proposed criterion, which motivate us to use it for variable selection in small area prediction of the finite population.

Table 1: Selecting rates and prediction errors based on cAIC and CScAIC, and improvement over cAIC procedure. M_1 is the true model.

		selecting rates				
		M_1	M_2	M_3	M_4	prediction error
a=2	cAIC	0.727	0.129	0.125	0.019	0.2216
	CScAIC	0.961	0.028	0.010	0.001	0.2190 (1.17)
a=4	cAIC	0.748	0.116	0.116	0.020	0.2195
	CScAIC	0.982	0.012	0.006	0.000	0.2163 (1.42)
a=6	cAIC	0.724	0.111	0.143	0.022	0.2199
	CScAIC	0.963	0.016	0.020	0.001	0.2163 (1.66)

4 Concluding Remarks

In this paper, we have proposed a variable selection criterion under covariate shift based on the conditional Akaike information proposed by Vaida and Blanchard (2005) and the proposed criterion includes the original conditional AIC as a special case where the covariate shift does not occur. We have pointed out that covariate shift is essential issue in small area prediction of the mean of finite population and proposed to use our criterion for variable selection. We have confirmed through a simulation study that the proposed criterion performs better than the original conditional AIC in small area prediction.

Acknowledgments.

The research of the first author was supported in part by Grant-in-Aid for Scientific Research (26-10395) from Japan Society for the Promotion of Science (JSPS). The research of the third author was supported in part by Grant-in-Aid for Scientific Research (23243039 and 26330036) from JSPS.

Appendix

A.1 Proof of Theorem 1. We decompose B_c in (6) as follows:

$$\begin{aligned} B_c =& E\left[(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t \widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})\right]/\sigma^2 - E\left[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^t \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})\right]/\sigma^2 \\ =& B_{c1} - B_{c2}. \quad \text{(say)} \end{aligned}$$

For the evaluation of B_{c1}, we first take the expectation with respect to the conditional distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$. Since $\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}} = \tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \widetilde{\boldsymbol{Z}}\boldsymbol{b} - \widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}} - \boldsymbol{b})$,

$$B_{c1} = m + E\left[\left\{\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}} - \boldsymbol{b})\right\}^t \widetilde{\boldsymbol{R}}^{-1}\left\{\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}} - \boldsymbol{b})\right\}\right]/\sigma^2.$$

It can be easily seen that

$$\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}} - \boldsymbol{b}) = (\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \widetilde{\boldsymbol{Z}}(\boldsymbol{b} - E(\boldsymbol{b}|\boldsymbol{y})),$$

where $E(\boldsymbol{b}|\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. Noting that $Var(\boldsymbol{b}|\boldsymbol{y}) = \sigma^2(\boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G})$ and $E[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t] = \sigma^2(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$, we have

$$\begin{aligned}
&E\left[\left\{\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}}-\boldsymbol{b})\right\}^t\widetilde{\boldsymbol{R}}^{-1}\left\{\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\widetilde{\boldsymbol{Z}}(\hat{\boldsymbol{b}}-\boldsymbol{b})\right\}\right]/\sigma^2\\
=&\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t\right]\\
&+\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}\widetilde{\boldsymbol{Z}}(\boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G})\widetilde{\boldsymbol{Z}}^t\right].
\end{aligned}$$

The second term of the right hand side of the above equation is rewritten as

$$\begin{aligned}
&\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{\Sigma}}-\widetilde{\boldsymbol{R}})-\widetilde{\boldsymbol{R}}^{-1}\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^t\right]\\
=&-m+\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{\Sigma}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^t)\right]\\
=&-m+\mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}].
\end{aligned}$$

Thus we can obtain

$$B_{c1} = \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}]+\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t\right]. \quad (16)$$

Next we evaluate B_{c2} as follows:

$$\begin{aligned}
B_{c2} =&E\left[\left\{\boldsymbol{R}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})-\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}^t\boldsymbol{R}^{-1}\left\{\boldsymbol{R}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})-\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}\right]/\sigma^2\\
=&\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}]-\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}]. \quad (17)
\end{aligned}$$

From (16) and (17), we can obtain (7). □

A.1 Proof of Theorem 2. In the same way as the proof of Theorem 1, we decompose B_c^* in (9) as follows:

$$\begin{aligned}
B_c^* =&E\left[(\tilde{\boldsymbol{y}}-\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})^t\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}}-\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}})/\hat{\sigma}^2\right]\\
&-E\left[(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\hat{\boldsymbol{b}})^t\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\hat{\boldsymbol{b}})/\hat{\sigma}^2\right]\\
=&B_{c1}^*-B_{c2}^*. \quad \text{(say)}
\end{aligned}$$

Firstly, we evaluate B_{c1}^*. From the proof of Theorem 1, it can be easily seen that

$$B_{c1}^* =\sigma^2\mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}]E[1/\hat{\sigma}^2]+E\left[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\hat{\sigma}^2\right].$$

Since $n\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$ and $\widehat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$, we can obtain

$$B_{c1}^* =\frac{n}{n-p-2}\left\{\mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}]+\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^t\right]\right\}. \quad (18)$$

Next we calculate B_{c2}^*. It is easily seen that

$$\begin{aligned}
B_{c2}^* =&E\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/\hat{\sigma}^2\right]+E\left[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\hat{\sigma}^2\right]\\
&-2E\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\hat{\sigma}^2\right].
\end{aligned}$$

We define $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/\sigma$ and $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1/2}$, then $\boldsymbol{v} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\boldsymbol{M}$ is idempotent. Using this notation, we rewrite B_{c2} as

$$\begin{aligned} B^*_{c2} =& nE\left[\frac{\boldsymbol{v}^t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}}{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] + nE\left[\frac{\boldsymbol{v}^t\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] - 2nE\left[\frac{\boldsymbol{v}^t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] \\ =& B^*_{c21} + B^*_{c22} - 2B^*_{c23}. \quad \text{(say)} \end{aligned}$$

By Lemma A.1. in Srivastava and Kubokawa (2010),

$$B^*_{c21} = n \times \left\{\frac{\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}]}{n-p-2} - \frac{2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I}_n-\boldsymbol{M})]}{(n-p)(n-p-2)}\right\}$$

Since $\boldsymbol{v}^t\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}$ is independent of $\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}$, we can get

$$B^*_{c22} = \frac{n}{n-p-2}\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}].$$

To evaluate B^*_{c23}, we rewrite

$$B^*_{c23} = nE\left[\frac{\boldsymbol{v}^t\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] + nE\left[\frac{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right].$$

Since $\boldsymbol{M}\boldsymbol{v}$ is independent of $(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}$ and $E[\boldsymbol{M}\boldsymbol{v}] = \boldsymbol{0}$, the second term of the right hand side of the above equation is 0. Then we get $B^*_{c23} = B^*_{c22}$. Combining B^*_{c21}, B^*_{c22} and B^*_{c23}, we can obtain

$$B^*_{c2} = \frac{n}{n-p}\left\{\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] - \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}]\right\}. \tag{19}$$

From (18) and (19), (10) follows. □

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] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

[4] Datta, G. and Ghosh, M. (2012). Small are shrinkage estimation. *Statist. Sci.*, **27**, 95-114.

[5] 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.

[6] Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James.Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

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

[8] Kawakubo, Y. and Kubokawa, T. (2014). Modified conditional AIC in linear mixed models. *J. Multivariate Anal.*, **129**, 44-56.

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

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

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

[12] Müller, S., Scealy, J.L. and Welsh, A.H. (2013). Model selection in linear mixed models. *Statist. Sci.*, **28**, 135-167.

[13] Pfeffermann, D. (2013). New important developments in small area estimation. *Statist. Sci.*, **28**, 40-68.

[14] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley, New Jersey.

[15] Saefken, B., Kneib, T., van Waveren, C.S. and Greven, S. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electron. J. Statist.*, **8**, 201-225.

[16] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, **90**, 227-244.

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

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

[19] Yu, D. and Yau, K.K.W. (2012). Conditional Akaike information criterion for generalized linear mixed models. *Comput. Statist. Data Anal.*, **56**, 629-644.

[20] Yu, D., Zhang, X. and Yau, K.K.W. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *J. Multivariate Anal.*, **116**, 245-262.