博士論文

Variable selection problem in mixed effects models with application to small area estimation

(混合効果モデルにおける変数選択問題と
小地域推定への応用)

Yuki KAWAKUBO

川久保友超

Abstract

This thesis studies the variable selection problem in mixed effects models, especially the method of information criteria is focused on. The conditional AIC proposed by Vaida and Blanchard (2005) is more appropriate for the focus on clusters than the conventional AIC is. This is useful in problems involving prediction of random effects such as small area estimation. Concerning the conditional AIC and the related fields, several problems are addressed and some new results are obtained. Firstly, the conditional AIC is modified for the underspecified model, which does not include the true model. Secondly, the variable selection problem in linear mixed model under covariate shift situation is considered. It is also shown that considering the covariate shift situation is meaningful in small area estimation problem. Thirdly, the conditional AIC in nonlinear mixed models based on natural exponential family is derived. Lastly, some variants of the AIC and the conditional AIC are proposed and their properties are discussed, which are also related to the Bayesian procedure as well as frequentists' methods.

Contents

Chapter 1

Introduction

This thesis studies the variable selection problem in mixed effects models, especially the method of information criteria is focused on. Vaida and Blanchard (2005) introduced the conditional Akaike information, which is related to the expected Kullback–Leibler divergence based on the conditional likelihood given random effects. The conditional Akaike information (cAI) and the corresponding criterion, called the conditional AIC (cAIC), are more appropriate for the focus on clusters than the conventional AIC (or marginal AIC, mAIC), which is based on the marginal likelihood integrating out the random effects. This is useful in problems involving prediction of random effects such as small area estimation. First of all, we here intruduce a class of mixed effects models and its application to small area estimation. Next, we review the variable selection problem in mixed effects models.

The class of mixed effects models has been studied for a long time from both theoretical and applied aspects. Especially, the linear mixed model and the best linear unbiased predictor (BLUP) introduced by Henderson (1950) provide flexible framework for modeling several types of data sets, whose applications are longitudinal data analysis in biostatistics, panel data analysis in econometrics, small area estimation in official statistics and others. The small area estimation problem is how to produce reliable estimates of some characteristic of interest for areas with small sample sizes. Model based estimator in small area estimation problem using mixed effects models can 'borrow information' from neighboring areas, which results in stable estimation of small area parameter. Datta and Ghosh (2012), Pfeffermann (2013) and Rao and Molina (2015) give good reviews about small area estimation.

There are several methods of variable selection in mixed effects models, which include the information criteria such as AIC or BIC, shrinkage methods such as LASSO, the Bayesian procedure, the Fence methods (Jiang et al., 2008), and others. Müller et al. (2013) is a good review of variable selection problem in linear mixed models. In this thesis, we focus on the information criteria, especially the cAIC. Since Vaida and Blanchard (2005), variable selection procedures using the cAIC have been developed and the properties of the cAIC have been discussed. Liang et al. (2008) proposed a different bias correction, who take into account estimation of the unknown parameters included in the covariance matrix of the vector of the random effects. It is noted that their bias correction is closely related to the generalized degrees of freedom of Ye (1998), while Vaida and Blanchard (2005) pointed out their bias correction is the same as the effective degrees of freedom of Hodges and Sargent (2001). Greven and Kneib (2010) derived analytical representation of the bias correction of Liang et al. (2008) and also applied to selecting the random effects. Srivastava and Kubokawa (2010) proposed other versions of the cAIC by changing the estimators of the regression coefficients and the variance parameter. Kubokawa (2011) derived the cAIC with a general covariance matrices of the vectors of the random effects

and the error terms. He also proposed a conditional version of Mallows' C_p (Mallows, 1973). The cAIC has been also applied to the variable selection problem in the generalized linear mixed models (GLMM). Donohue et al. (2011), Yu and Yau (2012) and Yu et al. (2013) derived the cAIC in the GLMM under the assumption that the sample size in each cluster goes to infinity. Saefken et al. (2014) proposed an exact unbiased estimator of the cAI, which is also justified for finite sample case, in a special case of the GLMM, Poisson mixed regression. As a related procedure to the cAIC, Zhang et al. (2014) proposed a model averaging method based on the prediction risk relative to the conditional Mallows' criterion. This procedure is an extension of the Mallows model averaging of Hansen (2007) to the linear mixed model and Zhang et al. (2014) also proved the model averaging estimator is asymptotically loss efficient under some regularity condition.

In this thesis, some problems are considered and new results are obtained. Firstly, the conditional AIC is modified for the underspecified model, which does not include the true model. Most of the Akaike-type information criteria put the assumption that the candidate model includes the true model, which we call overspecified assumption. Due to the assumption, the cAIC has large bias for estimating the cAI. This problem is considered in Chapter 3. Secondly, the variable selection problem in linear mixed model under covariate shift situation is considered. We also show that considering covariate shift situation is meaningful in small area estimation problem. We discuss this problem in Chapter 4. Thirdly, we derived the cAIC as a variable selection problem in a class of mixed effects models based on natural exponential family, which includes nonlinear mixed models such as Poisson-gamma model and binomial-beta model in Chapter 5. Lastly, some variants of the AIC and the cAIC are proposed and their properties are discussed, which are also related to the Bayesian procedure as well as frequentists' methods. We develop a variant of marginal AIC in Chapter 6 taking linear regression model as an example. In Chapter 7, variants of the cAIC are considered.

Chapter 2

Mixed effects models and conditional AIC

In this chapter, mixed effects models are introduced and the setup of the variable selection problem is explained. The mAIC and cAIC are also briefly summarized.

2.1 Mixed effects models

2.1.1 General

Let $\boldsymbol{y}$ be an observable random vector, and let $\boldsymbol{\theta}$ be an unobservable random vector. We treat continuous or discrete cases for $\boldsymbol{y}$ and $\boldsymbol{\theta}$. The conditional probability density (or mass) function of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ is $f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta})$ for a vector of unknown parameters $\boldsymbol{\eta}$, and the the probability density (or mass) function of $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, namely,

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\theta} &\sim f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}), \\ \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}|\boldsymbol{\eta}). \end{aligned} \tag{2.1}$$

When $\boldsymbol{\theta}$ denotes random effects, this expresses general mixed effects model. Because this model can be interpreted as a Bayesian model, we also use the terminology used in Bayes statistics. The marginal likelihood function of $\boldsymbol{y}$ and the conditional (or posterior) density function of $\boldsymbol{\theta}$ given $\boldsymbol{y}$ are

$$\begin{aligned} m(\boldsymbol{y}|\boldsymbol{\eta}) &= \int f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}, \\ p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\eta}) &= f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) / m(\boldsymbol{y}|\boldsymbol{\eta}). \end{aligned}$$

2.1.2 Linear mixed model

One of the most important examples of the mixed effects model is the linear mixed model, which is given as the following general form:

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

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of the response variables, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times p$ and $n \times q$ matrices of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, $\boldsymbol{b}$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. It is common to assume that $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are mutually independent and $\boldsymbol{b} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{Q})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{R})$.

The empirical best linear unbiased predictor (EBLUP) of the linear combination of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, which is of the form $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\beta}+\boldsymbol{d}^{\mathrm{T}}\boldsymbol{b}$ for a $p\times 1$ vector $\boldsymbol{c}$ and a $q\times 1$ vector $\boldsymbol{d}$, is $\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\boldsymbol{d}^{\mathrm{T}}\hat{\boldsymbol{b}}$, where

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y},$$
$$\hat{\boldsymbol{b}}=\boldsymbol{Q}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma}=\boldsymbol{Z}\boldsymbol{Q}\boldsymbol{Z}^{\mathrm{T}}+\boldsymbol{R}$.

2.2 Setup of variable selection problem

2.2.1 Candidate models, full model and true model

We focus on the variable selection of the fixed effects. To this end, we clarify candidate models, the true model, and the full model and use the following notations throughout the paper. We take a linear mixed model (2.2) for example.

First, we consider the collection of candidate models as follows. Let $n\times p_\omega$ matrix $\boldsymbol{X}(\omega)$ consist of all the explanatory variables and assume that $\mathrm{rank}\,(\boldsymbol{X}(\omega))=p_\omega$. In order to define candidate models by the index set j, suppose that j denotes a subset of $\omega=\{1,\ldots,p_\omega\}$ containing p_j elements, *i.e.* $p_j=\#(j)$, and $\boldsymbol{X}(j)$ consists of p_j columns of $\boldsymbol{X}(\omega)$ indexed by the elements of j. We define the index set by $\mathcal{J}=\mathcal{P}(\omega)$, namely the power set of ω, where we call ω the full model. Then we consider the candidate model j expressed as

$$\boldsymbol{y}=\boldsymbol{X}(j)\boldsymbol{\beta}_j+\boldsymbol{Z}\boldsymbol{b}_j+\boldsymbol{\varepsilon}_j,$$

where $\boldsymbol{\beta}_j$ is a $p_j\times 1$ vector of regression coefficients, $\boldsymbol{b}_j$ is a $q\times 1$ vector of random effects, $\boldsymbol{\varepsilon}_j$ is an $n\times 1$ vector of random errors.

Second, we assume that the true model exists in the collection of the candidate models $\mathcal{P}(\omega)$, which is denoted by j_*. It is noted that the dimension of the true model is p_{j_*}, which is abbreviated to p_*. Under the assumption, the true model can be written by using the full model design matrix, namely the true mean of $\boldsymbol{y}$ can be expressed as

$$E(\boldsymbol{y})=\boldsymbol{X}(\omega)\boldsymbol{\beta}_*,$$

where $\boldsymbol{\beta}_*$ is $p_\omega\times 1$ vector of regression coefficients, whose $p_\omega-p_*$ components are exactly 0 and the rest of components are not 0. Moreover, when the true model is included by the candidate model, the true mean of $\boldsymbol{y}$ can be also expressed as

$$E(\boldsymbol{y})=\boldsymbol{X}(j)\boldsymbol{\beta}_j^*,$$

where $\boldsymbol{\beta}_j^*$ is $p_j\times 1$ vector of regression coefficients, whose p_j-p_* components are exactly 0 and the rest of components are not 0. It is common to assume that the true model is included by the candidate model for the derivation of the Akaike-type information criteria.

2.2.2 Overspecified and underspecified

We introduce the terms overspecified and underspecified models. Candidate model j is overspecified if $\boldsymbol{X}(\omega)\boldsymbol{\beta}_*\in\mathcal{R}[\boldsymbol{X}(j)]$, which means that $\boldsymbol{X}(\omega)\boldsymbol{\beta}_*$ is in the column space of $\boldsymbol{X}(j)$ following Fujikoshi and Satoh (1997) or Kawakubo and Kubokawa (2014). The set of overspecified models are denoted by $\mathcal{J}_+=\{j\in\mathcal{J}|j_*\subseteq j\}$. On the other hand, candidate model j is called underspecified when $\boldsymbol{X}(\omega)\boldsymbol{\beta}_*\notin\mathcal{R}[\boldsymbol{X}(j)]$. The set of underspecified models are denoted by $\mathcal{J}_-=\mathcal{J}\setminus\mathcal{J}_+$.

2.3 Marginal and conditional AIC

2.3.1 Marginal AIC

The marginal AIC (mAIC) is related to the expected Kullback–Leibler (KL) divergence based on the marginal likelihood defined as

$$\int \left[\int \log \left\{ \frac{m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta})}{m(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\eta}})} \right\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] m(\boldsymbol{y}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{y}, \tag{2.3}$$

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ and $\widehat{\boldsymbol{\eta}}$ is the maximum likelihood estimator of $\boldsymbol{\eta}$. The marginal AIC (mAIC) is an (asymptotically) unbiased estimator of the following marginal Akaike information (mAI):

$$\mathrm{mAI} = \iint -2 \log\{m(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\eta}})\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta}) m(\boldsymbol{y}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y},$$

which is a part of (2.3) (multiplied by 2). Then, the mAIC is

$$\mathrm{mAIC} = -2 \log\{m(\boldsymbol{y}|\widehat{\boldsymbol{\eta}})\} + \Delta_{\mathrm{mAI}},$$

where Δ_{mAI} is bias correction (or penalty) term, which is given by

$$\Delta_{\mathrm{mAI}} = \mathrm{mAI} - E[-2 \log\{m(\boldsymbol{y}|\widehat{\boldsymbol{\eta}})\}].$$

Akaike (1973, 1974) proposed that the bias correction converges to $2p$ where p is the dimension of the candidate model.

2.3.2 Conditional AIC

The marginal AIC is not appropriate for the focus on the prediction of specific clusters or random effects. Taking this point into account, Vaida and Blanchard (2005) considered the expected KL divergence based on the conditional density, which is given by

$$\iint \left[\int \log \left\{ \frac{f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{\eta})}{f(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})} \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{\theta}, \tag{2.4}$$

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ given $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{\theta}}$ is some predictor of $\boldsymbol{\theta}$. The conditional AIC (cAIC) is an (asymptotically) unbiased estimator of the following conditional Akaike information (cAI):

$$\mathrm{cAI} = \iiint -2 \log\{f(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\} f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\eta}) \pi(\boldsymbol{\theta}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{\theta},$$

which is a part of (2.4) (multiplied by 2). Then, the cAIC is

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

where Δ_{cAI} is bias correction (or penalty) term, which is given as

$$\Delta_{\mathrm{cAI}} = \mathrm{cAI} - E[-2 \log\{f(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}})\}].$$

Chapter 3

Modified conditional AIC in linear mixed models

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 chapter, 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. This chapter is based on Kawakubo and Kubokawa (2014).

3.1 Motivation

The cAIC by Vaida and Blanchard (2005) is derived under the condition that the candidate model includes the true model. This assumption is called the overspecified assumption. 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 Table 3.1. Thus, the drawback of cAIC gives a motivation for addressing the query (II).

In this chapter, 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 the problem is explained in Section 3.2. In Section 3.3, 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 3.4, and it is shown that McAIC and the corresponding model averaging procedure are better than cAIC in terms of the prediction error. In Section 3.5, we apply the McAIC to estimate small area land prices. All the proofs are given in Section 3.6.

3.2 Setup of problem

We focus on the problem of selecting explanatory variables in linear mixed model, whose notations are given in Section 2.2.

First, we explain about the collection of candidate models. We consider the candidate model j, which is given as follows:

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j, \tag{3.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of the response variables, $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ are $n \times p_j$ and $n \times q$ matrices of covariates, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of regression coefficients, $\boldsymbol{b}_j$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_j$ is an $n \times 1$ vector of random errors. We here assume that $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ are mutually independent and that $\boldsymbol{b}_j \sim \mathcal{N}_r(\boldsymbol{0}, \sigma_j^2\boldsymbol{G})$, $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}_n)$ for a common unknown parameter σ_j^2 and a known matrix $\boldsymbol{G}$. It is important to point out that the random effects part remains the same over the models $\{j\}$. This means that we here consider the problem of selecting only the explanatory variables of the fixed effects. The conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}_j$ for model j is denoted by $f(\boldsymbol{y}|\boldsymbol{b}_j, \boldsymbol{\eta}_j)$, where $\boldsymbol{\eta}_j$ is the vector of unknown parameters, namely $\boldsymbol{\eta}_j = (\boldsymbol{\beta}_j^{\mathrm{T}}, \sigma_j^2)^{\mathrm{T}}$. The density function of $\boldsymbol{b}_j$ is denoted by $p(\boldsymbol{b}_j|\boldsymbol{\eta}_j)$.

Second, we assume that the data are generated from the true model which is given by

$$\boldsymbol{y} = \boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{Z}\boldsymbol{b}_* + \boldsymbol{\varepsilon}_*$$

for $\boldsymbol{b}_* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_*^2\boldsymbol{G}), \boldsymbol{\varepsilon}_* \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_*^2\boldsymbol{I}_n)$. Thus the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}(\omega)\boldsymbol{\beta}_*, \sigma_*^2\boldsymbol{\Sigma}), \tag{3.2}$$

where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{I}_n$. For the true model, the conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}_*$ and the density function of $\boldsymbol{b}_*$ are denoted by $f(\boldsymbol{y}|\boldsymbol{b}_*, \boldsymbol{\eta}_*)$ and $p(\boldsymbol{b}_*|\boldsymbol{\eta}_*)$, respectively, where $\boldsymbol{\eta}_* = (\boldsymbol{\beta}_*^{\mathrm{T}}, \sigma_*^2)^{\mathrm{T}}$.

Third, we assume that the collection of candidate models includes both underspecified and overspecified models, and that the full model ω includes the true model. This means that the set of overspecified models $\mathcal{J}_+$ is not empty set.

3.3 Modification of cAIC

3.3.1 Conditional Akaike information in linear mixed models

We begin with introducing the conditional Akaike information (cAI) in linear mixed models, which was proposed by Vaida and Blanchard (2005). The cAI is the estimand of the cAIC, and is related to the expected Kullback–Leibler (KL) divergence based on the conditional likelihood. The cAI for the setup explained in Section 3.2 is

$$\begin{aligned}
\mathrm{cAI} &= \iiint -2\log\{f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j)\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*, \boldsymbol{\eta}_*) f(\boldsymbol{y}|\boldsymbol{b}_*, \boldsymbol{\eta}_*) p(\boldsymbol{b}_*|\boldsymbol{\eta}_*) \mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b} \\
&= E^{(\boldsymbol{y},\boldsymbol{b}_*)} E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*} \left[n\log(2\pi\hat{\sigma}_j^2) + \|\tilde{\boldsymbol{y}} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j\|^2/\hat{\sigma}_j^2 \right],
\end{aligned}$$

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ given $\boldsymbol{b}$ and $E^{(\boldsymbol{y},\boldsymbol{b}_*)}$ 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 $\boldsymbol{b}_*$. The cAI measures the risk of plug-in predictive density $f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j)$, where $\widehat{\boldsymbol{\eta}}_j$ is the maximum likelihood estimator of $\boldsymbol{\eta}_j = (\boldsymbol{\beta}_j^{\mathrm{T}}, \sigma_j^2)^{\mathrm{T}}$ based on the candidate model j given as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j &= (\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \\
\hat{\sigma}_j^2 &= (\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j) \ / \ n,
\end{aligned} \tag{3.3}$$

and $\hat{\boldsymbol{b}}_j$ is the empirical Bayes estimator for quadratic loss of $\boldsymbol{b}_j$ given as

$$\hat{\boldsymbol{b}}_j = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j).$$

Then the cAIC is a bias corrected unbiased estimator of cAI, which is given as

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

where

$$\Delta_{\mathrm{cAI}} = \mathrm{cAI} - E^{(\boldsymbol{y},\boldsymbol{b}_*)}\big[-2\log\{f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j)\}\big], \tag{3.4}$$

which is called the bias correction (or penatly) term.

3.3.2 Evaluation of the bias of cAIC

Vaida and Blanchard (2005) evaluated the bias correction in (3.4) under the assumption that the candidate model j is overspecified. Then the cAIC is the exact unbiased estimator of cAI when the candidate model j is overspecified. However, when the candidate model j is underspecified, the cAIC has large bias for estimating the cAI. Thus we evaluate the bias correction in (3.4) both for overspecified and underspecified case.

First, it follows that

$$-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j) = n\log(2\pi\hat{\sigma}_j^2) + \|\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j\|^2/\hat{\sigma}_j^2. \tag{3.5}$$

Then, as shown in Section 3.6, the bias correction can be expressed as

$$\Delta_{\mathrm{cAI}} = E\left[\left\{(2n - \mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}])\sigma_*^2 - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{u} + 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*)\right\} \;/\; \hat{\sigma}_j^2\right], \tag{3.6}$$

where expectation is taken with respect to the joint distribution of $(\boldsymbol{y}, \boldsymbol{b}_*)$ and $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*$.

It is important to note that the distribution of $\hat{\sigma}_j^2$ for the underspecified case is different from that for 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

$$\begin{aligned} n\hat{\sigma}_j^2 &= \sigma_*^2\{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{z} + \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{z}\} \\ &= \sigma_*^2(K_0 + K_1), \end{aligned}$$

where $K_0 = \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{z}$, $K_1 = \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{z}$,

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}/\sigma_*, \\ \boldsymbol{M}_\omega &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}, \\ \boldsymbol{M}_j &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}. \end{aligned}$$

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

$$\boldsymbol{M}_j\boldsymbol{\xi}\begin{cases} = \boldsymbol{\xi} & \text{if } j \in \mathcal{J}_+, \\ \neq \boldsymbol{\xi} & \text{if } j \in \mathcal{J}_-, \end{cases}$$

since $\boldsymbol{X}_*\boldsymbol{\beta}_* \in \mathcal{R}[\boldsymbol{X}_j]$ if $j \in \mathcal{J}_+$. Thus K_0 can be rewritten as

$$K_0 = (\boldsymbol{\xi} + \boldsymbol{v})^{\mathrm{T}}(\boldsymbol{I}_N - \boldsymbol{M}_\omega)(\boldsymbol{\xi} + \boldsymbol{v}) = \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_N - \boldsymbol{M}_\omega)\boldsymbol{v}, \tag{3.7}$$

so that K_0 follows a chi-squared distribution with $n - p_\omega$ degrees of freedom, denoted by

$$K_0 \sim \chi^2_{n-p_\omega}.$$

Also, K_1 can be rewritten as

$$\begin{aligned} K_1 &= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi} \\ &= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2L + n\delta, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} L &= \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v}, \\ \delta &= \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi} \ / \ n. \end{aligned} \tag{3.9}$$

In the overspecified case, we have $K_1 \sim \chi^2_{p_\omega - p_j}$ since $\boldsymbol{M}_\omega\boldsymbol{\xi} = \boldsymbol{M}_j\boldsymbol{\xi} = \boldsymbol{\xi}$. In the underspecified case, K_1 follows a noncentral chi-squared distribution with $p_\omega - p_j$ degrees of freedom and with the noncentrality parameter $N\delta$, denoted by $K_1 \sim \chi^2_{p_\omega - p_j}(n\delta)$. Thus,

$$K_1 \sim \begin{cases} \chi^2_{p_\omega - p_j} & \text{if } j \in \mathcal{J}_+, \\ \chi^2_{p_\omega - p_j}(n\delta) & \text{if } j \in \mathcal{J}_-. \end{cases}$$

Since $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{u} = \sigma_*^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}$ and

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*) = \sigma_*^2\left\{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v} - \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}\boldsymbol{v}\right\},$$

we can rewrite (3.6) as

$$\Delta_{\mathrm{cAI}} = n{\cdot}E\left[(K_0 + K_1)^{-1}\left\{(2n - \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\right]) - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{v} + 2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v} - 2\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}\boldsymbol{v}\right\}\right]. \tag{3.10}$$

Although $K_0 + K_1$ 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}. To this end, we assume the following conditions:

(A1) $\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi} = O(n)$, which is the non-centrality parameter of K_1.

(A2) $\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi} = O(n)$.

The condition (A1) is equivalent to $\delta = O(1)$ given in (3.9). It is also noted that the condition (A2) is satisfied by (A1) if the maximum eigenvalue of $\boldsymbol{\Sigma}^{-1}$ is uniformly bounded. Under these assumptions, we can get the following theorem, which will be proved in Section 3.6.

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

$$B^* = 2n \times \left\{\frac{n - \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\right] + \mathrm{tr}\left[\boldsymbol{P}_j\right]}{n - p_j - 2} + \frac{\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\right] - \mathrm{tr}\left[\boldsymbol{P}_j\right]}{(n - p_j - 2)(n - p_j)}\right\}, \tag{3.11}$$

for $\boldsymbol{P}_j = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$*. In the underspecified case, the bias correction of* cAIC *is approximated as*

$$\Delta_{\mathrm{cAI}} = B^* + B_1 + B_2 + B_3 + O(n^{-1}), \tag{3.12}$$

where B_1, B_2 and B_3 are defined as

$$B_1 = \frac{2n(\lambda-1)}{n-p_j-2}(n - \mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}]), \tag{3.13}$$

$$B_2 = 2p_j\lambda(\lambda-1) - 4\lambda(\lambda-1)^2 + 2\mathrm{tr}\,[\boldsymbol{P}_j](\lambda-1) + 2(\lambda-1)\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}] \times \frac{2\lambda^2-(p_j+1)\lambda+1}{n}, \tag{3.14}$$

$$B_3 = \frac{4\lambda^2}{n}\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi}, \tag{3.15}$$

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

It is noted that in the overspecified case the bias B^* given in (3.11) 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 $\boldsymbol{M}_j\boldsymbol{\xi} = \boldsymbol{\xi}$.

3.3.3 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_1, B_2 and B_3 because B_* does not include any unknown parameters. 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 $\widehat{\lambda^3}$ as

$$\hat{\lambda} = \frac{n-p_j}{n-p_\omega}\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_j}, \tag{3.16}$$

$$\widehat{\lambda^2} = \frac{(n-p_j)(n-p_j+2)}{(n-p_\omega)(n-p_\omega+2)}\left(\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_j}\right)^2, \tag{3.17}$$

$$\widehat{\lambda^3} = \frac{(n-p_j)(n-p_j+2)(n-p_j+4)}{(n-p_\omega)(n-p_\omega+2)(n-p_\omega+4)}\left(\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_j}\right)^3. \tag{3.18}$$

In the overspecifed case, it is noted that $n\hat{\sigma}^2_\omega = \sigma^2_* K_0 \sim \sigma^2_*\chi^2_{n-p_\omega}$, $K_1 \sim \chi^2_{p_\omega-p_j}$ and $n\hat{\sigma}^2_j = \sigma^2_*(K_0+K_1)$, so that

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

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 $E[(\hat{\sigma}^2_\omega/\hat{\sigma}^2_j)^k] \to \lambda^k$ as $n \to \infty$ for $k = 1, 2, 3$, where the brief proof is given in Section 3.6.

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 (3.16), (3.17) and (3.18), we can estimate B_1 and B_2 in (3.13) and (3.14). 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

$$\begin{aligned} E[\hat{\lambda}] =& \frac{n-p_j}{n-p_\omega} E\left[\frac{K_0}{K_0+K_1}\right] \\ =& \lambda + \frac{-2\lambda^2(\lambda-1)+p_j\lambda(\lambda-1)}{n} + O(n^{-2}), \end{aligned} \tag{3.19}$$

where the proof is given in (3.33) in Section 3.6.

Lemma 3.2 *Consider the following estimator for B_1:*

$$\widehat{B_1} = \frac{2n(n-\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}])}{n-p_j-2}\left\{\hat{\lambda}-1+\frac{2(\widehat{\lambda^3}-\widehat{\lambda^2})-p_j(\widehat{\lambda^2}-\hat{\lambda})}{n}\right\}. \tag{3.20}$$

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 $\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\xi}$ which is a part of B_3. Define $\tilde{\sigma}_j^2$ by

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

From the fact that $\tilde{\sigma}_j^2=\sigma_*^2(\boldsymbol{v}+\boldsymbol{\xi})^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{I}_n-\boldsymbol{M}_j)(\boldsymbol{v}+\boldsymbol{\xi})$, it follows that

$$E[\tilde{\sigma}_j^2]=\sigma_*^2\left\{\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}-\boldsymbol{P}_j]+\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\xi}\right\}.$$

Hence an estimator of $\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega-\boldsymbol{M}_j)\boldsymbol{\xi}$ is given by

$$\tilde{\sigma}_j^2/\hat{\sigma}_\omega^2-\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}-\boldsymbol{P}_j].$$

Lemma 3.3 *Consider the following estimator for B_3:*

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

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 an $O(n^{-1})$ 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 $\widetilde{B_3}$:*

$$\widehat{B_3}=\widetilde{B_3}-\frac{4p_\omega\cdot\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}-\boldsymbol{P}_j]}{n^2}+\frac{8\mathrm{tr}\,[\boldsymbol{P}_\omega-\boldsymbol{P}_j]}{n^2}, \tag{3.21}$$

where $\boldsymbol{P}_\omega=\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. 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 correction Δ_{cAI} by the estimator

$$\widehat{\Delta_{\mathrm{cAI}}} = B^* + \widehat{B_1} + \widehat{B_2} + \widehat{B_3}. \tag{3.22}$$

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 as

$$\mathrm{McAIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) + \widehat{\Delta_{\mathrm{cAI}}}. \tag{3.23}$$

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

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

In the underspecified case, it follows that

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

Remark 3.1 In the derivation of McAIC, we assume that the covariance matrix of $\boldsymbol{b}_*$ is $\sigma_*^2\boldsymbol{G}$ for a known matrix $\boldsymbol{G}$. This setup seems restrictive, since $\sigma_*^2\boldsymbol{G}$ involves some unknown variance components in most linear mixed models. For example, we consider the nested error regression model which will be treated in the next section for simulation. In this model, $\boldsymbol{G}$ is a function of $\psi_* = \tau_*^2/\sigma_*^2$ where τ_*^2 is a variance component of random effects. We then propose to use a consistent estimator $\widehat{\psi} = \widehat{\psi}(\boldsymbol{y})$ which satisfies $\widehat{\psi}(\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\alpha}) = \widehat{\psi}(\boldsymbol{y})$ for any p-dimensional vector $\boldsymbol{\alpha}$, and replace $\boldsymbol{G}(\psi_*)$ in (3.23) by its plug-in estimator $\boldsymbol{G}(\widehat{\psi})$. For the location invariant property of $\widehat{\psi}$, the influence by replacement may be limited as long as we consider the problem of selecting only explanatory variables. Our recommendation is to estimate ψ_* by the full model. Estimating nuisance parameters by the full model is an methodology similar to Mallows' C_p.

3.4 Simulations

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

[**the first experiment**] In the first experiment, we compare the performance of the criteria cAIC and McAIC by measuring the biases of estimating cAI and the relative frequency of selecting the true model. It is important to note that Akaike-type criteria may be useful in getting the "best predictive model", but are not very suited for selecting the 'correct model' unlike the Bayesian information criterion (BIC) by Schwarz (1978). However, it has been reported that the performance of selecting the true model improves by modification or bias correction of Akaike-type criteria in Hurvich and Tsai (1989), Fujikoshi and Satoh (1997) and others. Thus it is meaningful to investigate the performance of selecting the true model, and we handle the experiment to measure the prediction error next.

We consider a class of the nested candidate models $j_\alpha = \{1, \ldots, \alpha\}$ for $\alpha = 1, \ldots, p_\omega$. The observed vector $\boldsymbol{y}$ is generated by the true model (3.2) with $\boldsymbol{Z} = \mathrm{diag}\,(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q), \boldsymbol{Z}_i = \boldsymbol{1}_r$ for $i = 1, \ldots, q$ and $\boldsymbol{G} = \boldsymbol{I}_q$, where $\boldsymbol{1}_r$ denotes an $r \times 1$ vector of ones and $r = n/q$. Note that this model is known as the nested error regression model (NERM), and that q and r denote the number of clusters (or areas) and the sample size in each cluster, respectively. Let $\boldsymbol{X}(\omega)$ be generated as $\mathbf{vec}\,(\boldsymbol{X}(\omega)^{\mathrm{T}}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_{\boldsymbol{x}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{x}} = 0.9\boldsymbol{I}_{p_\omega} + 0.1\boldsymbol{J}_{p_\omega}$ where $\boldsymbol{J}_{p_\omega} = \boldsymbol{1}_{p_\omega}\boldsymbol{1}_{p_\omega}^{\mathrm{T}}$. The true coefficient vector $\boldsymbol{\beta}_*$ is given by $\boldsymbol{\beta}_* = (\beta_1, \ldots, \beta_{p_*}, 0, \ldots, 0)^{\mathrm{T}}$ and β_l is generated as

$\beta_l = 2 \times ((-1)^l/(l+0.7)) \times U(1,2), 1 \leq l \leq p_*$ 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 $q = 10$. The true values of cAI in each model are calculated from (3.27) in Section 3.6 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 3.1. In each row, the several values, which are explained below, for each corresponding model j_α are reported. In the second column of the table, the true values of cAI in each model are reported. In the third and fourth columns, the averages of biases of estimating cAI by cAIC and McAIC are reported, respectively. In the fifth and sixth columns, the rates of selecting each model by cAIC and McAIC are reported, respectively.

		bias		**selection rates**	
model	cAI	cAIC	McAIC	cAIC	McAIC
pattern (a): $\sigma_*^2 = 1$					
j_1	217.81	16.234	-0.63453	0	0
j_2	184.14	12.170	-0.55772	0	0
j_3	167.68	7.2864	-0.25035	0.005	0.025
j_4	163.66	4.7962	-0.15384	0.050	0.087
j_5	158.84	-0.16459	-0.13939	0.778	0.812
j_6	160.59	-0.22447	-0.14562	0.117	0.062
j_7	162.47	-0.23391	-0.15563	0.050	0.014
pattern (b): $\sigma_*^2 = 0.5$					
j_1	210.62	18.465	-0.53668	0	0
j_2	168.35	16.200	-0.47607	0	0
j_3	142.90	11.564	-0.20047	0	0.001
j_4	134.82	8.4066	-0.083391	0.004	0.014
j_5	124.18	-0.16459	-0.13939	0.824	0.907
j_6	125.94	-0.22447	-0.14562	0.122	0.063
j_7	127.81	-0.23391	-0.15563	0.050	0.015

Table 3.1: Biases of estimating cAI by cAIC and McAIC, and the rates of selecting each model by cAIC and McAIC

Table 3.1 shows that although the conventional cAIC has large biases for underspecified models, namely model j_1 to j_4, our proposed McAIC has smaller biases for both underspecified and overspecified models. Especially, because cAIC overestimates the cAI for underspecified cases, cAIC does not tend to select smaller models. The McAIC procedures also have better performances of selecting the true model than the conventional cAIC procedures do. 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.

[**the second experiment**] In the second experiment, we investigate the prediction error of the best model chosen by cAIC and McAIC and that of model averaged predictor based on cAIC and McAIC. We here handle the case of unknown $\boldsymbol{G}$ and consider the model class which consists of all subsets of $\omega = \{1, \ldots, p_\omega\}$. The other setup is the same as that in the first experiment except for $\boldsymbol{G} = \boldsymbol{G}(\psi_*) = \psi_* \boldsymbol{I}_q$ where $\psi_* = \tau_*^2/\sigma_*^2$, namely $\boldsymbol{b}_* \sim \mathcal{N}(\boldsymbol{0}, \tau_*^2 \boldsymbol{I}_q)$. σ_*^2 and τ_*^2 are estimated by unbiased estimators proposed by Prasad and Rao (1990) under the full model,

which are explained as follows. Let $S = \boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{I}_n - \boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\}\boldsymbol{y}$ and $S_1 = \boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{E}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{E}\}\boldsymbol{y}$ where $\boldsymbol{E} = \mathrm{diag}\,(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_q), \boldsymbol{E}_i = \boldsymbol{I}_r - r^{-1}\boldsymbol{J}_r$ for $i = 1, \ldots, q$. Then, the Prasad–Rao estimators of σ_*^2 and τ_*^2 are given by

$$\hat{\sigma}^{2\mathrm{PR}} = S_1/(n - q - p_\omega) \quad \text{and} \quad \hat{\tau}^{2\mathrm{PR}} = \left\{S - (n - p_\omega)\hat{\sigma}^{2PR}\right\}/n^*, \tag{3.24}$$

where $n^* = n - \mathrm{tr}\,[\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{Z}]$. We estimate ψ_* by $\widehat{\psi} = \hat{\tau}^{2\mathrm{PR}}/\hat{\sigma}^{2\mathrm{PR}}$ and use the plug-in value of $\boldsymbol{G}(\widehat{\psi})$ for cAIC and McAIC. This $\widehat{\psi}$ satisfies the properties of consistency and location invariance mentioned in the Remark 3.1. It is important to point out that we use the ML estimator $\hat{\sigma}_j^2$ or $\hat{\sigma}_\omega^2$ in cAIC or McAIC given in (3.3), though we use $\hat{\sigma}^{2PR}$ to estimate ψ_* by substituting $\boldsymbol{\Sigma}(\widehat{\psi})$ for $\boldsymbol{\Sigma}$. We measure the performance of cAIC and McAIC via $\|\hat{\boldsymbol{y}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* - \boldsymbol{Z}\boldsymbol{b}_*\|^2/n$ for $\hat{\boldsymbol{y}}_j = \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j + \boldsymbol{Z}\hat{\boldsymbol{b}}_j$, which is here called the prediction error because $\hat{\boldsymbol{b}}_j$ is a predictor of $\boldsymbol{b}_*$. The prediction errors are given as averages based on 1,000 replications.

In addition to the procedures that select the best model by the cAIC and McAIC, we consider a model averaging method. The aim of model averaging is to predict a future value by a weighted mean of the predictors for the candidate models. The weighting functions are important in the model averaging method, and we use 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 the model j and let $\mathrm{McAIC}_{\min}$ be the minimum McAIC value. Also, let $\overline{\mathrm{McAIC}}_j = \mathrm{McAIC}_j - \mathrm{McAIC}_{\min}$. Then the weight is defined by

$$w_j = \frac{\exp(-\overline{\mathrm{McAIC}}_j/2)}{\sum_l \exp(-\overline{\mathrm{McAIC}}_l/2)}. \tag{3.25}$$

Based on the weights given in (3.25), we can obtain a model averaged predictor

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

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 3.2 reports the prediction errors for the best model selected by cAIC and McAIC and for the model averaged predictors based on S-cAIC and S-McAIC for the six cases in which n, σ_*^2 and τ_*^2 take different values. We set $p_\omega = 5$ and $p_* = 3$. The values in parentheses are the improvement over the prediction error by the cAIC procedure expressed in percentage. 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.

3.5 Real data example

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.

	cAIC	McAIC	S-cAIC	S-McAIC
pattern (a)	0.23198	0.22628	0.22581	0.22240
$N = 50$, $\sigma_*^2 = 1.0$, $\tau_*^2 = 0.5$		(2.46)	(2.66)	(4.13)
pattern (b)	0.13107	0.12818	0.12837	0.12629
$N = 50$, $\sigma_*^2 = 0.5$, $\tau_*^2 = 1.0$		(2.20)	(2.06)	(3.65)
pattern (c)	0.12516	0.12231	0.12234	0.12040
$N = 50$, $\sigma_*^2 = 0.5$, $\tau_*^2 = 0.5$		(2.28)	(2.25)	(3.80)
pattern (d)	0.15561	0.15344	0.15091	0.14930
$N = 80$, $\sigma_*^2 = 1.0$, $\tau_*^2 = 0.5$		(1.39)	(3.02)	(4.06)
pattern (e)	0.083838	0.082674	0.081682	0.080744
$N = 80$, $\sigma_*^2 = 0.5$, $\tau_*^2 = 1.0$		(1.39)	(2.57)	(3.69)
pattern (f)	0.081538	0.080418	0.079368	0.078477
$N = 80$, $\sigma_*^2 = 0.5$, $\tau_*^2 = 0.5$		(1.37)	(2.66)	(3.75)

Table 3.2: Prediction errors of the predictors based on cAIC, McAIC, S-cAIC and S-McAIC, and improvement over cAIC procedure

We analyze the land price data in 2001 with covariates for 47 stations which we consider as small areas and let $q = 47$. For the ith small area, there are data of n_i land spots, and the total sample size is $n = \sum_{i=1}^{q} n_i = 189$. 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, $\mathrm{TRN}_i \times \mathrm{DST}_{ik}$ and $\mathrm{TRN}_i \times \mathrm{FOOT}_{ik}$, which are denoted by $x_1, \ldots, x_9$ and x_0 denotes constant term.

When one wants to estimate the mean land prices of each small area, naive estimators such as sample mean may have large variabilities because the sample sizes in each small area are small. Then we use some model based small area estimation technique, which introduce random effects and borrow the strength from the information of other areas. For the details, see Rao and Molina (2015). One example is the nested error regression model (NERM), which we handled in the second experiment of the previous section. We here employ NERM to estimate the mean land prices in the places around each station. In order to specify the model, or to select the explanatory variables from 10 covariates $x_0, \ldots, x_9$, we use the variable selection criteria cAIC and McAIC. The procedure is that regressors which minimizes the information criteria are added to the model known as the forward stepwise selection. The unknown parameter $\psi_* = \tau_*^2/\sigma_*^2$ is estimated by $\hat{\tau}^{2\mathrm{PR}}/\hat{\sigma}^{2\mathrm{PR}}$ given in (3.24) by the full model.

Table 3.3 reports the values of cAIC and McAIC in each model. Both criteria select the same model $j = \{1, 0, 2, 3\}$, namely,

$$y_{ik} = \beta_0 + \mathrm{FAR}_i\beta_1 + \mathrm{TRN}_i\beta_2 + (\mathrm{TRN}_i)^2\beta_3 + b_i + \varepsilon_{ik},$$

where the parameters are estimated by $\hat{\sigma}^{2\mathrm{PR}} = 0.46382, \hat{\tau}^{2\mathrm{PR}} = 0.08199$, i.e., $\widehat{\psi} = 0.17722$ and $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3) = (5.0823, 6.3790 \times 10^{-3}, -1.0561 \times 10^{-1}, 6.4938 \times 10^{-4})$, $\hat{\sigma}_\omega^2 = 0.43889$, $\hat{\sigma}_j^2 = 0.44465$. This result demonstrates that the most important factor to decide land prices is the floor area ratio, which is very natural, and that the land prices are decreasing as a quadratic function of the time to take by train to the metropolitan area. It is interesting to see that the

model	cAIC	McAIC
$\{1\}$	531.89	512.08
$\{1,0\}$	482.63	467.98
$\{1,0,2\}$	419.18	417.31
$\{1,0,2,3\}$	410.07*	410.92*
$\{1,0,2,3,4\}$	412.21	413.01
$\{1,0,2,3,4,5\}$	413.09	413.79
$\{1,0,2,3,4,5,8\}$	415.02	415.54
$\{1,0,2,3,4,5,8,6\}$	417.14	417.60
$\{1,0,2,3,4,5,8,6,9\}$	419.02	419.23
ω	420.97	420.98

Table 3.3: The values of cAIC and McAIC in each model in the example of posted land price data

values of cAIC is bigger than that of McAIC for the models $\{1\}$, $\{1,0\}$ and $\{1,0,2\}$ and both criteria take similar values for the models bigger than $\{1,0,2,3\}$, which we take as the 'true' model. This result seems to consistent with the demonstration of the first numerical experiment in the previous section, in which cAIC tends to overestimate cAI for the underspecified models.

3.6 Proofs

3.6.1 Derivation of (3.6)

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

$$\begin{aligned}\mathrm{cAI} = & E\big[n\log(2\pi\hat{\sigma}_j^2) + n\sigma_*^2/\hat{\sigma}_j^2\big] \\ & + E\Big[\|(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_*)\|^2/\hat{\sigma}_j^2\Big]. \end{aligned} \tag{3.26}$$

Note that $\boldsymbol{b}_*|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}, \sigma_*^2(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G})\right)$ and that

$$\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_*) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*) - \boldsymbol{Z}(\boldsymbol{b}_* - \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}).$$

Taking the expectation with respect to $\boldsymbol{b}_*|\boldsymbol{y} \sim \pi(\boldsymbol{b}_*|\boldsymbol{y}, \boldsymbol{\eta}_*)$ in (3.26), we rewrite the cAI as

$$\begin{aligned}\mathrm{cAI} = & E\big[n\log(2\pi\hat{\sigma}_j^2) + (2n - \mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}])\sigma_*^2/\hat{\sigma}_j^2\big] \\ & + E\big[(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*)^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*)/\hat{\sigma}_j^2\big]. \end{aligned} \tag{3.27}$$

Next, for a part of $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j)$ in (3.5), it is noted that

$$\begin{aligned} & \boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j \\ = & \boldsymbol{y} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* - (\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*) - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{y} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* - (\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*)\right\} \\ = & \boldsymbol{\Sigma}^{-1}\boldsymbol{u} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*). \end{aligned} \tag{3.28}$$

Thus, from (3.27) and (3.28), we can see that $\Delta_{\mathrm{cAI}} = \mathrm{cAI} - E[-2\log f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\eta}}_j)]$ is expressed as (3.6). □

3.6.2 Proof of Theorem 3.1

For (3.10), we decompose Δ_{cAI} as

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

where $b_1 = nE[(2n - \mathrm{tr}\,\boldsymbol{\Sigma}^{-1})/(K_0+K_1)]$, $b_2 = -nE[\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}/(K_0+K_1)]$, $b_3 = 2nE[\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}/(K_0+K_1)]$ and $b_4 = -2nE[\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}\boldsymbol{v}/(K_0+K_1)]$.

We begin with expanding $(K_0+K_1)^{-1}$ up to $O_p(n^{-2})$. Let $= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{v} + \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v}$. Then, $K_0 + K_1 = K + 2L + n\delta$, $K \sim \chi^2_{n-p_j}$, $K = O_p(n)$ and $L = O_p(n^{1/2})$, so that we can write $(2L + n\delta)/K = \delta + D$ and $D = O_p(n^{-1/2})$. Thus, it follows that

$$\begin{aligned}(K_0 + K_1)^{-1} =& (K + 2L + n\delta)^{-1} = K^{-1}(1 + \delta + D)^{-1} = \frac{\lambda}{K}(1 + D\lambda)^{-1} \\ =& \frac{\lambda}{K}\left\{1 - D\lambda + (D\lambda)^2 + O_p(n^{-3/2})\right\}. \end{aligned} \tag{3.29}$$

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

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

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

$$\begin{aligned}\frac{1}{K} =& \frac{1}{n - p_j}\left\{1 - A + A^2 + O_p(n^{-3/2})\right\} \\ =& \frac{1}{n}(1 - A + A^2 + \frac{p_j}{n}) + O_p(n^{-5/2}), \\ 1 - \frac{n}{K} =& A - A^2 - \frac{p_j}{n} + O_p(n^{-3/2}). \end{aligned} \tag{3.30}$$

Hence, $(K_0 + K_1)^{-1}$ can be evaluated as

$$(K_0 + K_1)^{-1} = \frac{\lambda}{n}\left\{1 - (A + \frac{2L}{n})\lambda + A^2\lambda^2 + \frac{p_j\lambda + 4AL\lambda^2}{n} + \frac{4L^2\lambda^2}{n^2}\right\} + O_p(n^{-5/2}). \tag{3.31}$$

For any function $q(\cdot)$, we have $E[q(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{v})L] = 0$ since $q(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{v})L$ 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

$$E[(K_0 + K_1)^{-1}] = \frac{\lambda}{n}\left\{1 + \frac{-2\lambda^2 + (p_j + 4)\lambda}{n}\right\} + O(n^{-3}). \tag{3.32}$$

Using the expansions (3.31) and (3.32) , we can evaluate b_1, b_2, b_3 and b_4, respectively.

First, b_1 can be evaluated as

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

where

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

which is the exact b_1 for overspecified models.

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

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

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

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

Using this equality, we can evaluate b_2 as

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

where

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

which is the exact b_2 for overspecified models.

As for b_3, it can be decomposed as

$$\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}}{K_0 + K_1} = \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}}{K_0 + K_1} + \frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}}{K_0 + K_1}.$$

Since $\boldsymbol{M}_j\boldsymbol{v}$ is independent of $(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{v}$ and $K_0 + K_1$, and $E[\boldsymbol{M}_j\boldsymbol{v}] = 0$, it follows that

$$E\left[\frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}}{K_0 + K_1}\right] = 0.$$

Further, because $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}$ and $K_0 + K_1$ are mutually independent, b_3 is evaluated as

$$\begin{aligned} b_3 &= 2n \times E[\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j\boldsymbol{v}] \times E[(K_0 + K_1)^{-1}] \\ &= b_3^* + 2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_j](\lambda - 1) + O(n^{-1}), \end{aligned}$$

where

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

which is the exact b_3 for overspecified models.

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

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

Combining the evaluations of b_1, b_2, b_3 and b_4 yields the result in (3.12), where B^* is defined by $B^* = b_1^* + b_2^* + b_3^*$. □

3.6.3 Proof of Lemma 3.1

It follows from $K_0 = n + O_p(n^{1/2})$ and (3.31) that

$$E\left[\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^k\right] = E\left[\left(\frac{K_0}{K_0+K_1}\right)^k\right] \to \lambda^k \quad (n\to\infty),$$

which proves lemma 3.1. □

3.6.4 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 (3.19). Using the expansion (3.31) of $(K_0+K_1)^{-1}$, we can approximate $E[\hat{\lambda}]$ as

$$\begin{aligned} E[\hat{\lambda}] &= \frac{n-p_j}{n-p_\omega} \times E\left[\frac{K_0}{K_0+K_1}\right] \\ &= \frac{n-p_j}{n-p_\omega}\frac{\lambda}{n} \times E\Bigg[\{(n-p_\omega) + \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{v} - (n-p_\omega)\} \\ &\quad \times \left\{1 - (A+\frac{2L}{n})\lambda + A^2\lambda^2 + \frac{p_j\lambda + 4AL\lambda^2}{n} + \frac{4L^2\lambda^2}{n^2}\right\}\Bigg] + O(n^{-2}). \end{aligned} \tag{3.33}$$

Evaluating (3.33) up to $O(n^{-1})$, we can get (3.19) and Lemma 3.2. □

3.6.5 Proof of Lemma 3.3

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

$$\begin{aligned} D_1 &= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{v}, \\ D_2 &= 2\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{I}_n - \boldsymbol{M}_j)\boldsymbol{v}, \\ D_3 &= \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi}. \end{aligned} \tag{3.34}$$

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

$$\begin{aligned} \widetilde{B_3} &= \frac{4K_0(D_1+D_2+D_3)}{(K_0+K_1)^2} - \frac{4c_1}{n}\frac{K_0^2}{(K_0+K_1)^2} \\ &= \widetilde{B_{31}} - \widetilde{B_{32}}. \quad \text{(say)} \end{aligned}$$

From the exapansion (3.31) of $(K_0+K_1)^{-1}$, it follows that

$$\begin{aligned} E[\widetilde{B_{31}}] &= \frac{4\lambda^2}{n^2} \times \{E[K_0D_1] + E[K_0D_3]\} + O(n^{-1}) \\ &= \frac{4c_1\lambda^2}{n} + \frac{4\lambda^2 D_3}{n} + O(n^{-1}), \\ E[\widetilde{B_{32}}] &= \frac{4c_1}{n} \times \frac{\lambda^2}{n^2} \times E[K_0^2] + O(n^{-1}) \\ &= \frac{4c_1\lambda^2}{n} + O(n^{-1}), \end{aligned}$$

which proves Lemma 3.3. □

3.6.6 Proof of Lemma 3.4

It is noted that the adjustment term of $\widehat{B_3}$ is of order $O(n^{-1})$ from (3.21). 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 $n\hat{\sigma}_j^2 = \sigma_*^2 K$ and $\tilde{\sigma}_j^2 = \sigma_*^2 D_1$ for D_1 given in (3.34). Then,

$$\begin{aligned}\widetilde{B_3} &= 4 \times \frac{K_0 \times D_1}{K^2} - \frac{4c_1}{n}\left(\frac{K_0}{K}\right)^2 \\ &= \widetilde{B_{33}} - \widetilde{B_{34}}. \quad \text{(say)}\end{aligned}$$

From (3.30), K^{-2} is expanded as

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

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

$$\begin{aligned}E[\widetilde{B_{33}}] &= \frac{4}{n^2}E\Big[\{(n-p_\omega) + K_0 - (n-p_\omega)\}\{c_1 + (D_1 - c_1)\} \\ &\quad \times \{1 - 2A + 3A^2 + \frac{2p_j}{n}\}\Big] + O(n^{-2}) \\ &= \frac{4}{n^2} \times \Big\{(n-p_\omega)c_1\big(1 + 3E[A^2] + \frac{2p_j}{n}\big) \\ &\quad - 2c_1E[K_0A] - 2(n-p_\omega)E[D_1A] + E[K_0D_1] - c_1(n-p_\omega)\Big\} + O(n^{-2}) \\ &= \frac{4(n - p_\omega + 2p_j)c_1}{n^2} + \frac{8(c_2 - c_1)}{n^2} + O(n^{-2}), \end{aligned} \tag{3.35}$$

since $K_0 - (n - p_\omega) = O_p(n^{1/2})$, $c_1 = O(n)$ and $D_1 - c_1 = O_p(n^{1/2})$. Noting that $K_0/K \sim Be((n-p_\omega)/2, (p_\omega - p_j)/2)$, we can evaluate $E[\widetilde{B_{34}}]$ as

$$\begin{aligned}E[\widetilde{B_{34}}] &= \frac{4c_1}{n}\frac{(n-p_\omega)(n-p_\omega+2)}{(n-p_j)(n-p_j+2)} \\ &= \frac{4(n - 2p_\omega + 2p_j)c_1}{n^2} + O(n^{-2}). \end{aligned} \tag{3.36}$$

Combining (3.35) and (3.36) gives

$$E[\widetilde{B_3}] = \frac{4c_1p_\omega}{n^2} + \frac{8(c_2 - c_1)}{n^2} + O(n^{-2}),$$

which shows Lemma 3.4. □

Chapter 4

Conditional AIC under covariate shift with application to small area prediction

In this chapter, 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 AIC in terms of covariate shift. We especially focus on covariate shift in small area prediction and show usefulness of the proposed criterion.

4.1 Motivation

For prediction problems, 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. However, even when the information about the covariates in the predictive model can be used, most Akaike-type criteria do not use it. This is because most criteria put the assumption that the predictive model is the same as the observed model. We explain what this means in the context of the conditional AIC.

Vaida and Blanchard (2005) proposed the conditional AIC (cAIC) as an unbiased estimator of the cAI, which is given by

$$\mathrm{cAI} = \iiint -2\log\{f(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\}f(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{\eta})f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})\mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{\theta}, \tag{4.1}$$

in general, where the notations are the same as those in Chapter 2. We 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'. The cAI in (4.1) assumes that the conditional density of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ and that of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{\theta}$ are the same and both of them are denoted by $f(\cdot|\boldsymbol{\theta},\boldsymbol{\eta})$, which implies that the observed model and the predictive model are the same. However, under covariate shift, the conditional density of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{\theta}$ should be different from that of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ and is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{\eta})$. Then, we redefine the cAI under covariate shift as follows:

$$\mathrm{cAI} = \iiint -2\log\{g(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}})\}g(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{\eta})f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})\mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{\theta}.$$

In the next section, we define the cAI under covariate shift in linear mixed model.

4.2 Covariate shift conditional AIC

4.2.1 Observed model

The candidate observed model j is the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j, \tag{4.2}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of response variables, $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ are $n \times p_j$ and $n \times q$ matrices of covariates, respectively, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of regression coefficients, $\boldsymbol{b}_j$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_j$ is an $n \times 1$ vector of random errors. Let $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ be mutually independent and $\boldsymbol{b}_j \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_j^2\boldsymbol{G})$, $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2\boldsymbol{R})$, where $\boldsymbol{G}$ and $\boldsymbol{R}$ are $q \times q$ and $n \times n$ positive definite matrices and σ_j^2 is a scalar. We assume that $\boldsymbol{G}$ and $\boldsymbol{R}$ are known and σ_j^2 is unknown. The conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}_j$ and the density function of $\boldsymbol{b}_j$ for the model j are denoted by $f(\boldsymbol{y}|\boldsymbol{b}_j, \boldsymbol{\beta}_j, \sigma_j^2)$ and $p(\boldsymbol{b}_j|\sigma_j^2)$, respectively.

The true observed model j_* is

$$\boldsymbol{y} = \boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{Z}\boldsymbol{b}_* + \boldsymbol{\varepsilon}_*,$$

where $\boldsymbol{b}_* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_*^2\boldsymbol{G})$, $\boldsymbol{\varepsilon}_* \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_*^2\boldsymbol{R})$. Note that $\boldsymbol{X}(\omega)$ is $n \times p_\omega$ matrix of covariates for the full model ω and that $\boldsymbol{\beta}_*$ is $p_\omega \times 1$ vector of regression coefficients, whose $p_\omega - p_j$ components are exactly 0 and the rest of components are not 0, as explained in Section 2.2. Then the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}(\omega)\boldsymbol{\beta}_*, \sigma_*^2\boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{R}$. For the true model, the conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}_*$ and the density function of $\boldsymbol{b}_*$ are denoted by $f(\boldsymbol{y}|\boldsymbol{b}_*, \boldsymbol{\beta}_*, \sigma_*^2)$ and $p(\boldsymbol{b}_*|\sigma_*^2)$, respectively.

4.2.2 Predictive model

The candidate predictive model j is the linear mixed model which has the same regression coefficients $\boldsymbol{\beta}_j$ and random effects $\boldsymbol{b}_j$ as in the candidate observed model j, but different covariates, namely

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}(j)\boldsymbol{\beta}_j + \widetilde{\boldsymbol{Z}}\boldsymbol{b}_j + \tilde{\boldsymbol{\varepsilon}}_j, \tag{4.3}$$

where $\tilde{\boldsymbol{y}}$ is $m \times 1$ random vector of the target of prediction, $\widetilde{\boldsymbol{X}}(j)$ and $\widetilde{\boldsymbol{Z}}$ are $m \times p_j$ and $m \times q$ matrices of covariates whose columns correspond to those of $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$, respectively, and $\tilde{\boldsymbol{\varepsilon}}_j$ is $m \times 1$ vector of random errors, which is independent of $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ and is distributed as $\tilde{\boldsymbol{\varepsilon}}_j \sim \mathcal{N}_m(\boldsymbol{0}, \sigma_j^2\widetilde{\boldsymbol{R}})$, where $\widetilde{\boldsymbol{R}}$ is a known $m \times m$ positive definite matrix. We assume that we know the values of $\widetilde{\boldsymbol{X}}(j)$ and $\widetilde{\boldsymbol{Z}}$ in the predictive model and that they are not necessarily the same as those of $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ in the observed model. We call this situation covariate shift. The conditional density function of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_j$ for the model j is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_j, \boldsymbol{\beta}_j, \sigma_j^2)$.

The true predictive model j_* is

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* + \widetilde{\boldsymbol{Z}}\boldsymbol{b}_* + \tilde{\boldsymbol{\varepsilon}}_*,$$

where $\widetilde{\boldsymbol{X}}(\omega)$ is $m \times p_\omega$ matrix of covariates and $\tilde{\boldsymbol{\varepsilon}}_* \sim \mathcal{N}_m(\boldsymbol{0}, \sigma_*^2\widetilde{\boldsymbol{R}})$. Then the marginal distribution of $\tilde{\boldsymbol{y}}$ is

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}_m(\widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*, \sigma_*^2\widetilde{\boldsymbol{\Sigma}}),$$

where $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Z}}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^{\mathrm{T}} + \widetilde{\boldsymbol{R}}$. For the true model, the conditional density function of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_*$ is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*, \boldsymbol{\beta}_*, \sigma_*^2)$.

4.2.3 Conditional Akaike information

The conditional Akaike information (cAI) measures the prediction risk of the plug-in predictive density $g(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)$, where $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are maximum likelihood estimators of $\boldsymbol{\beta}_j$ and σ_j^2, which are given as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j &= (\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y},\\ \hat{\sigma}_j^2 &= (\boldsymbol{y}-\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)/n,\end{aligned}$$

and $\hat{\boldsymbol{b}}_j$ is empirical Bayes estimator of $\boldsymbol{b}_j$ for quadratic loss, which is given by

$$\hat{\boldsymbol{b}}_j = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j).$$

Then, the cAI under covariate shift is

$$\begin{aligned}\mathrm{cAI} &= E^{(\boldsymbol{y},\boldsymbol{b}_*)}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}\left[-2\log\{g(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)\}\right]\\ &= E^{(\boldsymbol{y},\boldsymbol{b}_*)}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}\left[m\log(2\pi\hat{\sigma}_j^2)+\log|\widetilde{\boldsymbol{R}}|+(\tilde{\boldsymbol{y}}-\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j-\widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}}-\widetilde{\boldsymbol{X}}(j))\widehat{\boldsymbol{\beta}}_j-\widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2\right],\end{aligned}$$

where $E^{(\boldsymbol{y},\boldsymbol{b}_*)}$ and $E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}$ denote expectation with respect to the joint distribution of $(\boldsymbol{y},\boldsymbol{b}_*)\sim f(\boldsymbol{y}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)p(\boldsymbol{b}_*|\sigma_*^2)$ and the conditional distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_*$, namely $\tilde{\boldsymbol{y}}|\boldsymbol{b}_*\sim g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)$, respectively. Taking expectation with respect to $\tilde{\boldsymbol{y}}|\boldsymbol{b}_*\sim g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)$ and $\boldsymbol{b}_*|\boldsymbol{y}\sim\mathcal{N}_q(\tilde{\boldsymbol{b}}_*,\sigma_*^2(\boldsymbol{G}-\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}))$ for $\tilde{\boldsymbol{b}}_*=\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}(\omega)\boldsymbol{\beta}_*)$, we can obtain

$$\mathrm{cAI} = E\left[m\log(2\pi\hat{\sigma}_j^2)+\log|\widetilde{\boldsymbol{R}}|+\mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda})\cdot\sigma_*^2/\hat{\sigma}_j^2+\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\hat{\sigma}_j^2\right], \tag{4.4}$$

where

$$\begin{aligned}\boldsymbol{\Lambda} &= \widetilde{\boldsymbol{\Sigma}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^{\mathrm{T}},\\ \boldsymbol{a} &= (\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j-\widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*)-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j-\boldsymbol{X}(\omega)\boldsymbol{\beta}_*).\end{aligned} \tag{4.5}$$

4.2.4 Criterion for overspecified model

In this subsection, we propose the covariate shift cAIC (CScAIC) as an unbiased estimator of cAI in (4.4) under the assumption that the candidate model is overspecified. When the candidate model j is overspecified, the true mean vector of $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ can be expressed as

$$E(\boldsymbol{y}) = \boldsymbol{X}(j)\boldsymbol{\beta}_j^*, \quad \text{and} \quad E(\tilde{\boldsymbol{y}}) = \widetilde{\boldsymbol{X}}(j)\boldsymbol{\beta}_j^*$$

where $\boldsymbol{\beta}_j^*$ is $p_j\times 1$ vector, whose p_j-p_* components are exactly 0 and the rest of components are not 0. In this subsection, we hereafter abbreviate $\boldsymbol{X}(j)$ to $\boldsymbol{X}$, $\widetilde{\boldsymbol{X}}(j)$ to $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{\beta}_j^*$ to $\boldsymbol{\beta}$ for notational convenience. Then, $\boldsymbol{a}$ in cAI given by (4.5) can be reduced to

$$\boldsymbol{a} = \boldsymbol{A}(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}),$$

where $\boldsymbol{A}=\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$. Furthermore, $n\hat{\sigma}_j^2/\sigma_*^2$ follows chi-squared distribution with $n-p_j$ degrees of freedom, denoted by $n\hat{\sigma}_j^2/\sigma_*^2\sim\chi^2_{n-p_j}$, which implies that

$$E[(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta})^{\mathrm{T}}] = \sigma_*^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1},$$

when the candidate model j is overspecified. Noting that $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are mutually independent, we can evaluate the cAI in (4.4) as

$$\text{cAI} = E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + \frac{n}{n-p_j-2}\Big\{\text{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}] + \text{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^{\mathrm{T}}]\Big\}. \quad (4.6)$$

We propose the CScAIC as an unbiased estimator of cAI as follows:

$$\text{CScAIC} = l(j) + \Delta_{\text{CS}}, \quad (4.7)$$

where $l(j)$ is likelihood (or goodness of fit) part and Δ_{CS} is bias correction, which are given by

$$l(j) = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2, \quad (4.8)$$

$$\Delta_{\text{CS}} = \frac{n}{n-p_j-2}\Big\{\text{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}] + \text{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^{\mathrm{T}}]\Big\} + \frac{n}{n-p_j}\left\{-\text{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] + \text{tr}\,[\boldsymbol{R}\boldsymbol{P}]\right\}, \quad (4.9)$$

where $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. Then we obtain the following theorem.

Theorem 4.1 *When the candidate model is overspecified, the covariate shift cAIC* (CScAIC) *in* (4.7) *is an unbiased estimator of the* cAI *in* (4.4).

Next corollary shows that our CScAIC includes the cAIC by Vaida and Blanchard (2005) as special case.

Corollary 4.1 *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 correction of the* CScAIC *in* (4.9) *is reduced to*

$$\Delta_{\text{CS}} = \frac{2n^2}{n-p_j-2} + \frac{2n(n-p_j-1)}{(n-p_j)(n-p_j-2)}\left\{-\text{tr}\,[\boldsymbol{\Sigma}^{-1}] + \text{tr}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}]\right\}, \quad (4.10)$$

which are identical to bias correction of the cAIC *by Theorem 2 in Vaida and Blanchard (2005).*

4.3 Modification of the criterion

4.3.1 Drawback of overspecified model assumption

Most of the Akaike-type criteria are derived under the assumption that 'the candidate model includes the true model', namely overspecified assumption. Although the assumption is too strong, the influence is restrictive in practical use. This is because the likelihood part of the criterion is a naive estimator of the risk function, namely the cAI in the context of the cAIC.

However, under the covariate shift situation, the likelihood part of the CScAIC is not a good estimator of the cAI. As a result, the CScAIC in (4.7) has large biases for estimating the cAI of the underspecified models as illustrated in simulations in Section 4.5.

Thus we evaluate and estimate the cAI directly both for the overspecified models and underspecified models in the following subsections.

4.3.2 Evaluation of cAI

We evaluate the cAI in (4.4) both for overspecified model and for underspecified model. We assume that the full model ω is overspecified, namely the collection of the overspecified models $\mathcal{J}_+$ is not empty set. We also assume that the size of response variable in predictive model m is of order $O(n)$.

When the candidate model j is overspecified, $n\hat{\sigma}_j^2/\sigma_*^2$ follows the chi-squared distribution. However, for the underspecified model, this is not true. We decompose $n\hat{\sigma}_j^2/\sigma_*^2$ as

$$\begin{aligned} n\hat{\sigma}_j^2/\sigma_*^2 &= \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{z} + \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{z} \\ &= K_0 + K_1 \quad \text{(say)}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}/\sigma_*, \\ \boldsymbol{M}_\omega &= \boldsymbol{W}_\omega(\boldsymbol{W}_\omega^{\mathrm{T}}\boldsymbol{W}_\omega)^{-1}\boldsymbol{W}_\omega^{\mathrm{T}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}, \\ \boldsymbol{M}_j &= \boldsymbol{W}_j(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{W}_j)^{-1}\boldsymbol{W}_j^{\mathrm{T}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}, \end{aligned}$$

for $\boldsymbol{W}_\omega = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\omega)/\sigma_*$ and $\boldsymbol{W}_j = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(j)/\sigma_*$. Note that $\boldsymbol{M}_j$ and $\boldsymbol{M}_\omega$ are projection matrices, namely symmetric and idempotent. Let $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}/\sigma_*$ and $\boldsymbol{\xi} = \boldsymbol{W}_\omega\boldsymbol{\beta}_*$. Then, it can be seen that $\boldsymbol{M}_\omega\boldsymbol{\xi} = \boldsymbol{\xi}$ and

$$\boldsymbol{M}_j\boldsymbol{\xi}\begin{cases} = \boldsymbol{\xi} & \text{if } j \in \mathcal{J}_+, \\ \neq \boldsymbol{\xi} & \text{if } j \in \mathcal{J}_-, \end{cases}$$

since $\boldsymbol{X}(\omega)\boldsymbol{\beta}_* \in \mathcal{R}[\boldsymbol{X}(j)]$ if $j \in \mathcal{J}_+$. Thus K_0 can be rewritten as

$$K_0 = (\boldsymbol{\xi} + \boldsymbol{v})^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)(\boldsymbol{\xi} + \boldsymbol{v}) = \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{v},$$

so that $K_0 \sim \chi^2_{n-p_\omega}$. Also, K_1 can be rewritten as

$$\begin{aligned} K_1 &= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi} \\ &= \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v} + 2L + n\delta, \end{aligned}$$

where

$$\begin{aligned} L &= \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{v}, \\ \delta &= \boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{\xi}/n. \end{aligned}$$

In the overspecified case, we have $K_1 \sim \chi^2_{p_\omega - p_j}$ since $\boldsymbol{M}_\omega\boldsymbol{\xi} = \boldsymbol{M}_j\boldsymbol{\xi} = \boldsymbol{\xi}$. In the underspecified case, K_1 follows a non-central chi-squared distribution with $p_\omega - p_j$ degrees of freedom and with the noncentrality parameter $n\delta$, denoted by $K_1 \sim \chi^2_{p_\omega - p_j}(n\delta)$. Thus,

$$K_1 \sim \begin{cases} \chi^2_{p_\omega - p_j} & \text{if } j \in \mathcal{J}_+, \\ \chi^2_{p_\omega - p_j}(n\delta) & \text{if } j \in \mathcal{J}_-. \end{cases}$$

Because $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are mutually independent, the cAI in (4.4) can be rewritten as

$$\mathrm{cAI} = E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + n \cdot E[(K_0 + K_1)^{-1}]\{\mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]\}. \quad (4.11)$$

Evaluating $E[(K_0 + K_1)^{-1}]$ and $E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]$, we can obtain the following theorem.

Theorem 4.2 *For the overspecified case, it follows that* cAI $= E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + R^*$, *where*

$$R^* = \frac{n\gamma}{n-p_j-2},$$

for $\gamma = \mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{A}^{\mathrm{T}}]$ *and* $\boldsymbol{A} = \widetilde{\boldsymbol{X}}(j) - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j)$. *For the underspecified case,* cAI *is approximated as*

$$\mathrm{cAI} = E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + R^* + R_1 + R_2 + R_3 + R_4 + O(n^{-1}), \tag{4.12}$$

where

$$\begin{aligned}
R_1 &= \gamma(\lambda-1),\\
R_2 &= \gamma\cdot n^{-1}\{-2\lambda^3 + (p_j+4)\lambda^2 - (p_j+2)\},\\
R_3 &= \lambda\cdot\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2,
\end{aligned}$$

and

$$R_4 = n^{-1}\{-2\lambda^3 + (p_j+4)\lambda^2\}\times\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2,$$

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

$$\begin{aligned}
\boldsymbol{B} &= \{\widetilde{\boldsymbol{P}}_j\boldsymbol{X}(\omega) - \widetilde{\boldsymbol{X}}(\omega) + \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\},\\
\boldsymbol{P}_j &= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1},\\
\boldsymbol{P}_\omega &= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1},
\end{aligned}$$

and

$$\widetilde{\boldsymbol{P}}_j = \widetilde{\boldsymbol{X}}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}.$$

When the candidate model j is overspecified, it follows that R_1, R_2, R_3 and R_4 are exactly 0.

4.3.3 Estimation of cAI

Because the approximation of cAI in (4.12) includes unknown parameters, we have to provide an estimator of cAI for practical use. Firstly, we obtain estimators of R_1 and R_2, which are polynomials of λ. We define $\hat{\lambda}$, $\widehat{\lambda^2}$ and $\widehat{\lambda^3}$ as

$$\begin{aligned}
\hat{\lambda} &= \frac{n-p_j}{n-p_\omega}\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2},\\
\widehat{\lambda^2} &= \frac{(n-p_j)(n-p_j+2)}{(n-p_\omega)(n-p_\omega+2)}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^2,
\end{aligned}$$

and

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

When $j\in\mathcal{J}_+$, it follows that $n\hat{\sigma}_\omega^2/\sigma_*^2 = K_0$, $n\hat{\sigma}_j^2/\sigma_*^2 = K_0+K_1$, and that K_0 and K_1 are mutually independent and distribute as $K_0\sim\chi^2_{n-p_\omega}$, $K_1\sim\chi^2_{p_\omega-p_j}$. Thus,

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

where $\mathrm{Be}(\cdot,\cdot)$ denotes the beta distribution. This implies that $E(\hat{\lambda}) = E(\widehat{\lambda^2}) = E(\widehat{\lambda^3}) = 1$ for the overspecified case. For the underspecified case, on the other hand, it follows that $E[(\hat{\sigma}_\omega^2/\hat{\sigma}_j^2)^k] = \lambda^k + O(n^{-1})$ as $n \to \infty$ for $k = 1, 2, 3$. Then we can obtain an estimator of R_2 in the approximation of cAI given by (4.12), which is given as follows:

$$\widehat{R_2} = \gamma \cdot \frac{-2\widehat{\lambda^3} + (p_j + 4)\widehat{\lambda^2} - (p_j + 2)}{n}. \tag{4.13}$$

Noting that $\gamma = O(n)$, we can get the following lemma.

Lemma 4.1 *When the candidate model j is underspecified, $\widehat{R_2}$ in (4.13) is an asymptotically unbiased estimator of R_2 whose bias is of order $O(n^{-1})$, namely*

$$E(\widehat{R_2}) = R_2 + O(n^{-1}).$$

When the candidate model j is overspecified, it follows that $E(\widehat{R_2}) = 0$.

Because R_1 is of order $O(n)$, we have to estimate λ with higher order accuracy in order to obtain an estimator of R_1 whose bias is of order $O(n^{-1})$ for the underspecified case. To this end, we expand $E(\hat{\lambda})$ up to $O(n^{-1})$ as

$$\begin{aligned} E(\hat{\lambda}) &= \frac{n - p_j}{n - p_\omega} \cdot E\left(\frac{K_0}{K_0 + K_1}\right) \\ &= \lambda + \frac{-2\lambda^3 + (p_j + 2)\lambda^2 - p_j\lambda}{n} + O(n^{-2}). \end{aligned}$$

Then we can obtain an estimator of R_1 given as

$$\widehat{R_1} = \gamma \cdot \left\{\hat{\lambda} - \frac{-2\widehat{\lambda^3} + (p_j + 2)\widehat{\lambda^2} - p_j\hat{\lambda}}{n} - 1\right\}. \tag{4.14}$$

Lemma 4.2 *When the candidate model j is underspecified, $\widehat{R_1}$ in (4.14) is an asymptotically unbiased estimator of R_1 whose bias is of order $O(n^{-1})$, namely*

$$E(\widehat{R_1}) = R_1 + O(n^{-1}).$$

When the candidate model j is overspecified, it follows that $E(\widehat{R_1}) = 0$.

We next consider estimation procedures of R_3 and R_4, which are complex functions of unknown parameters. We see R_3 and R_4 as functions of $\boldsymbol{\eta}_* = (\boldsymbol{\beta}_*^{\mathrm{T}}, \sigma_*^2)^{\mathrm{T}}$, namely $R_3 = R_3(\boldsymbol{\eta}_*)$, $R_4 = R_4(\boldsymbol{\eta}_*)$ and substitute their unbiased estimators $\widetilde{\boldsymbol{\eta}} = (\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \tilde{\sigma}^2)^{\mathrm{T}}$, which are given by

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}}_\omega = (\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \\ \tilde{\sigma}^2 &= (\boldsymbol{y} - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}})/(n - p_\omega). \end{aligned}$$

Then, plug-in estimators of R_3 and R_4 are

$$\begin{aligned} \widetilde{R_3} &= R_3(\widetilde{\boldsymbol{\eta}}) = \tilde{\lambda} \cdot \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\widetilde{\boldsymbol{\beta}}/\tilde{\sigma}^2, \\ \widetilde{R_4} &= R_4(\widetilde{\boldsymbol{\eta}}) = n^{-1}\{-2\tilde{\lambda}^3 + (p_j + 4)\tilde{\lambda}^2\} \times \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\widetilde{\boldsymbol{\beta}}/\tilde{\sigma}^2, \end{aligned} \tag{4.15}$$

where $\tilde{\lambda} = 1/(1+\tilde{\delta})$ for

$$\tilde{\delta} = \widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}(\omega)^{\mathrm{T}} (\boldsymbol{P}_\omega - \boldsymbol{P}_j) \boldsymbol{X}(\omega) \widetilde{\boldsymbol{\beta}} / (n\tilde{\sigma}^2).$$

Because R_3 is of order $O(n)$, plug-in estimator $\widetilde{R_3}$ has second order bias. Then we correct the bias by analytical method based on Taylor series expansions. We can see that expectation of the plug-in estimator $R_3(\widetilde{\boldsymbol{\eta}})$ is expanded as

$$E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) + O(n^{-2}), \tag{4.16}$$

where $B_1(\boldsymbol{\eta}_*)$ and $B_2(\boldsymbol{\eta}_*)$ are second- and third-order biases of $R_3(\widetilde{\boldsymbol{\eta}})$, respectively, namely $B_1(\boldsymbol{\eta}_*) = O(1)$ and $B_2(\boldsymbol{\eta}_*) = O(n^{-1})$. Because $\widetilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are independent, it follows that

$$B_1(\boldsymbol{\eta}_*) = \frac{1}{2} \cdot \mathrm{tr}\left[\frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}} E[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^{\mathrm{T}}] \right] + \frac{1}{2} \frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{(\partial \sigma_*^2)^2} E[(\tilde{\sigma}^2 - \sigma_*^2)^2], \tag{4.17}$$

where $E[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^{\mathrm{T}}] = \sigma_*^2 (\boldsymbol{X}(\omega)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}(\omega))^{-1}$ and $E[(\tilde{\sigma}^2 - \sigma_*^2)^2] = 2(\sigma_*^2)^2/(n - p_\omega)$. Second order partial derivatives of R_3 are given by the following lemma.

Lemma 4.3 *Second order partial derivative of $R_3(\boldsymbol{\eta}_*)$ with respect to $\boldsymbol{\beta}_*$ is*

$$\begin{aligned} \frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}} = \frac{\boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_*}{\sigma_*^2} \times \left\{ -\frac{2\boldsymbol{C}}{n\sigma_*^2(1+\delta)^2} + \frac{8\boldsymbol{C}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}}{n^2(\sigma_*^2)^2(1+\delta)^3} \right\} \\ - \frac{4\boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_* \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{C} + 4\boldsymbol{C}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B}}{n(\sigma_*^2)^2(1+\delta)^2} + 2\lambda \cdot \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} / \sigma_*^2, \end{aligned}$$

where $\boldsymbol{C} = \boldsymbol{X}(\omega)^{\mathrm{T}}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)$. Second order partial derivative of $R_3(\boldsymbol{\eta}_*)$ with respect to σ_*^2 is

$$\begin{aligned} \frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{(\partial \sigma_*^2)^2} = \frac{\boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_*}{\sigma_*^2} \times \left\{ -\frac{2\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}_*}{n(\sigma_*^2)^3(1+\delta)^2} + \frac{2(\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}_*)^2}{n^2(\sigma_*^2)^4(1+\delta)^3} \right\} \\ - \frac{2\boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_* \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{C} \boldsymbol{\beta}_*}{n(\sigma_*^2)^4(1+\delta)^2} + 2\lambda \cdot \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_* / (\sigma_*^2)^3. \end{aligned}$$

When the candidate model j is overspecified, second order bias $B_1(\boldsymbol{\eta}_*)$ can be simplified to

$$B_1(\boldsymbol{\eta}_*) = \mathrm{tr}\,[\boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} (\boldsymbol{X}(\omega)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}(\omega))^{-1}],$$

because $\boldsymbol{C}\boldsymbol{\beta}_* = \boldsymbol{B}\boldsymbol{\beta}_* = \boldsymbol{0}$ and $\lambda = 1$, which implies that $(\partial^2 R_3(\boldsymbol{\eta}_*))/(\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}) = 2\boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B}/\sigma_*^2$ and that $(\partial^2 R_3(\boldsymbol{\eta}_*))/(\partial \sigma_*^2)^2 = 0$. However, one cannot know which candidate models are overspecified. Then, we propose the following bias corrected estimator of R_3:

$$\widetilde{\widetilde{R_3}} = R_3(\widetilde{\boldsymbol{\eta}}) - B_1(\widetilde{\boldsymbol{\eta}}). \tag{4.18}$$

Lemma 4.4 *Both for the cases where the candidate model j is overspecified and where j is underspecified, $\widetilde{\widetilde{R_3}}$ and $\widetilde{R_4}$ in (4.18) and (4.15) are asymptotically unbiased estimators of R_3 and R_4, whose biases are of order $O(n^{-1})$, namely*

$$E(\widetilde{\widetilde{R_3}}) = R_3 + O(n^{-1}), \quad \text{and} \quad E(\widetilde{R_4}) = R_4 + O(n^{-1}).$$

Using $\widehat{R_1}$, $\widehat{R_2}$, $\widetilde{\widetilde{R_3}}$ and $\widetilde{R_4}$ given by (4.14), (4.13), (4.18) and (4.15), respectively, we can construct an estimator of cAI as follows:

$$\widehat{\mathrm{cAI}} = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + R^* + \widehat{R_1} + \widehat{R_2} + \widetilde{\widetilde{R_3}} + \widetilde{R_4}. \tag{4.19}$$

Theorem 4.3 *Both for the cases where the candidate model j is overspecified and where j is underspecified,* $\widehat{\mathrm{cAI}}$ *in* (4.19) *is a second-order asymptotically unbiased estimator of* cAI*, namely*

$$E(\widehat{\mathrm{cAI}}) = \mathrm{cAI} + O(n^{-1}).$$

Because the CScAIC in Section 4.2 is an exact unbiased estimator of the cAI when the candidate model is overspecified, the performance of $\widehat{\mathrm{cAI}}$ in (4.19) is not as good as that of the CScAIC for the overspecified model when n is small. Thus we should improve the estimator of R_3 and R_4 to remove the biases which are of order $O(n^{-1})$. To this end, we adopt parametric bootstrap method. Bootstrap sample $\boldsymbol{y}^\dagger$ is generated by

$$\boldsymbol{y}^\dagger = \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{b}^\dagger + \boldsymbol{\varepsilon}^\dagger,$$

where $\boldsymbol{b}^\dagger$ and $\boldsymbol{\varepsilon}^\dagger$ are generated by the following distribution:

$$\boldsymbol{b}^\dagger \sim \mathcal{N}(\boldsymbol{0}, \tilde{\sigma}^2\boldsymbol{G}), \quad \text{and} \quad \boldsymbol{\varepsilon}^\dagger \sim \mathcal{N}(\boldsymbol{0}, \tilde{\sigma}^2\boldsymbol{I}_n).$$

Then, we use the following estimator of R_4:

$$\widehat{R_4} = 2R_4(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_4(\widetilde{\boldsymbol{\eta}}^\dagger)] \tag{4.20}$$

where $E_{\widetilde{\boldsymbol{\eta}}}$ denotes expectation with respect to bootstrap distribution and $\widetilde{\boldsymbol{\eta}}^\dagger = ((\widetilde{\boldsymbol{\beta}}^\dagger)^{\mathrm{T}}, \tilde{\sigma}^{2\dagger})^{\mathrm{T}}$ is

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^\dagger &= (\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)\boldsymbol{\Sigma}^{-1}\boldsymbol{y}^\dagger, \\
\tilde{\sigma}^{2\dagger} &= (\boldsymbol{y}^\dagger - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}}^\dagger)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^\dagger - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}}^\dagger)/(n - p_\omega).
\end{aligned}$$

As for R_3, it follows from (4.16) that

$$E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] = R_3(\widetilde{\boldsymbol{\eta}}) + B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}}) + O_p(n^{-2}).$$

However, $B_1(\widetilde{\boldsymbol{\eta}})$ has bias which is of order $O(n^{-1})$, namely

$$E[B_1(\widetilde{\boldsymbol{\eta}})] = B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}),$$

where $B_{11}(\boldsymbol{\eta}_*) = O(n^{-1})$. Because this bias is not negligible when one wants to estimate R_3 with third order accuracy, we estimate this bias by bootstrap method as follows:

$$\widehat{B_{11}} = E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}}).$$

Then we obtain an estimator of R_3, which is given as

$$\begin{aligned}
\widehat{R_3} &= 2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + \widehat{B_{11}} \\
&= 2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}}).
\end{aligned} \tag{4.21}$$

Lemma 4.5 *Both for the cases where the candidate model j is overspecified and where j is underspecified, $\widehat{R_3}$ and $\widehat{R_4}$ in* (4.21) *and* (4.20) *are asymptotically unbiased estimators of R_3 and R_4, whose biases are of order $O(n^{-2})$, namely*

$$E(\widehat{R_3}) = R_3 + O(n^{-2}), \quad \text{and} \quad E(\widehat{R_4}) = R_4 + O(n^{-2}).$$

Using $\widehat{R_1}$, $\widehat{R_2}$, $\widehat{R_3}$ and $\widehat{R_4}$ given by (4.14), (4.13), (4.21) and (4.20), we can obtain an estimator of cAI as follows:

$$\widehat{\mathrm{cAI}}^{\dagger} = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + R^* + \widehat{R_1} + \widehat{R_2} + \widehat{R_3} + \widehat{R_4}, \tag{4.22}$$

which improves $\widehat{\mathrm{cAI}}$ in unbiasedness.

Theorem 4.4 *When the candidate model j is overspecified, $\widehat{\mathrm{cAI}}^{\dagger}$ in* (4.22) *is a third order asymptotically unbiased estimator of* cAI, *namely*

$$E(\widehat{\mathrm{cAI}}^{\dagger}) = \mathrm{cAI} + O(n^{-2}).$$

When the candidate model j is underspecified, $\widehat{\mathrm{cAI}}^{\dagger}$ is a second order asymptotically unbiased estimator of cAI, *namely*

$$E(\widehat{\mathrm{cAI}}^{\dagger}) = \mathrm{cAI} + O(n^{-1}).$$

4.4 Application to 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 (NERM) proposed by Battese et al. (1988).

Let Y_{ik} and $\boldsymbol{x}_{ik}(j)$ denote the value of a characteristic of interest and its p_j-dimensional auxiliary variable for the kth unit of the ith area for $i = 1, \ldots, q$ and $k = 1, \ldots, N_i$. Note that $\boldsymbol{x}_{ik}(j)$ is subvector of $\boldsymbol{x}_{ik}(\omega)$, which is the vector of the explanatory variables in the full model ω, and we hereafter abbreviate the model index j and write $\boldsymbol{x}_{ik}$ instead of $\boldsymbol{x}_{ik}(j)$, p instead of p_j and others. Then, the NERM is

$$Y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = 1, \ldots, N_i), \tag{4.23}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, b_i is a random effect for the ith area and b_i's and ε_{ik}'s are mutually independently distributed as $b_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$. We consider the situation that only n_i values of the Y_{ik}'s are observed through some sampling procedure. We define the number of the unobserved variables in the ith area by $N_i - n_i = r_i$ and let $n = n_1 + \cdots + n_q, r = r_1 + \cdots + r_q$. 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

$$y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = 1, \ldots, n_i), \tag{4.24}$$

which corresponds to (4.2) with $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n_i})^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})^{\mathrm{T}}$, $\boldsymbol{Z} = \mathrm{diag}\,(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q)$ for $\boldsymbol{Z}_i = \boldsymbol{1}_{n_i}$, $\boldsymbol{G} = \psi\boldsymbol{I}_q$ and $\boldsymbol{R} = \boldsymbol{I}_n$, where $\boldsymbol{1}_{n_i}$ denotes an $n_i \times 1$ vector of ones and $\psi = \tau^2/\sigma^2$. In the derivation of our proposed criteria,

we have assumed that the covariance matrix of $\boldsymbol{b}$ is $\sigma^2\boldsymbol{G}$ for a known matrix $\boldsymbol{G}$. However in the NERM, $\boldsymbol{G}$ 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.

We consider two types of predictive models. The first one can be used in the situation where all $\boldsymbol{x}_{ik}$'s are available. Then the predictive model, which we call the 'unit level predictive model', is defined by

$$Y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = n_i + 1, \ldots, N_i), \tag{4.25}$$

which corresponds to (4.3) with $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{y}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{y}}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\tilde{\boldsymbol{y}}_i = (Y_{i,n_i+1}, \ldots, Y_{i,N_i})^{\mathrm{T}}$, $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{X}}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\widetilde{\boldsymbol{X}}_i = (\boldsymbol{x}_{i,n_i+1}, \ldots, \boldsymbol{x}_{i,N_i}^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\boldsymbol{Z}} = \mathrm{diag}\,(\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_q)$ for $\widetilde{\boldsymbol{Z}}_i = \boldsymbol{1}_{r_i}$, $\widetilde{\boldsymbol{R}} = \boldsymbol{I}_r$. Note that $m = r$.

In the problem of small area prediction, we often encounter the situation where all $\boldsymbol{x}_{ik}$'s are not observed but the area mean $\bar{\boldsymbol{x}}_i = N_i^{-1}\sum_{k=1}^{N_i}\boldsymbol{x}_{ik}$ 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 second type of predictive model, which we call the 'area level predictive model', can be defined as

$$\overline{Y}_{i(u)} = \bar{\boldsymbol{x}}_{i(u)}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \bar{\varepsilon}_{i(u)} \quad (i = 1, \ldots, q), \tag{4.26}$$

where $\overline{Y}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i} Y_{ik}$, the mean of unobserved variables, $\bar{\boldsymbol{x}}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i}\boldsymbol{x}_{ik}$, calculated from $\bar{\boldsymbol{x}}_i$ and $(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$, and $\bar{\varepsilon}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i}\varepsilon_{ik}$ distributed as $\mathcal{N}(0, \sigma^2/r_i)$. The model (4.26) corresponds to (4.3) with $\tilde{\boldsymbol{y}} = (\overline{Y}_{1(u)}, \ldots, \overline{Y}_{q(u)})^{\mathrm{T}}$, $\widetilde{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_{1(u)}, \ldots, \bar{\boldsymbol{x}}_{q(u)})^{\mathrm{T}}$, $\widetilde{\boldsymbol{Z}} = \boldsymbol{I}_q$ and $\widetilde{\boldsymbol{R}} = \mathrm{diag}\,(\widetilde{R}_1, \ldots, \widetilde{R}_q)$ for $\widetilde{R}_i = 1/r_i$. Note that $m = q$.

After selecting explanatory variables with our proposed criteria, we predict $\overline{Y}_{i(u)}$ by the empirical best linear unbiased predictor $\widehat{\overline{Y}}_{i(u)} = \bar{\boldsymbol{x}}_{i(u)}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{b}_i$ and obtain a predictor of the finite population mean $\overline{Y}_i$, which is given as

$$\widehat{\overline{Y}}_i = \frac{1}{N_i}\left\{\sum_{k=1}^{n_i} y_{ik} + r_i\widehat{\overline{Y}}_{i(u)}\right\}. \tag{4.27}$$

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

4.5 Simulations

4.5.1 Simulations of measuring the biases of estimating the true cAI by the criteria

In this subsection, we compare the performance of the criteria by measuring the biases of estimating the cAI. We consider a class of the nested candidate models $j_\alpha = \{1, \ldots, \alpha\}$ for $\alpha = 1, \ldots, p_\omega$ where $p_\omega = 7$. The true observed model is the NERM in (4.24) with $\sigma^2 = \tau^2 = 1$ and $n_i = 3$ for $i = 1, \ldots, q$. We consider the unit level predictive model (4.25) for the first experiment and the area level predictive model (4.26) for the second experiment. The explanatory variables in the full model $\boldsymbol{x}_{ik}(\omega)$'s $(i = 1, \ldots, q;\ k = 1, \ldots, N_i)$ are independently generated by $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = 0.9\boldsymbol{I}_{p_\omega} + 0.1\boldsymbol{J}_{p_\omega}$ for $\boldsymbol{J}_{p_\omega} = \boldsymbol{1}_{p_\omega}\boldsymbol{1}_{p_\omega}^{\mathrm{T}}$. The true coefficient vector $\boldsymbol{\beta}_*$ is $\boldsymbol{\beta}_* = (\beta_1, \ldots, \beta_{p_*}, 0, 0)$

for $p_* = 5$ and β_l's $(1 \leq l \leq p_*)$ are generated by $\beta_l = 2 \times ((-1)^l/(l + 0.7)) \times U(1, 2)$ for a uniform random variable $U(1, 2)$ on the interval $(1, 2)$. The values of the explanatory variables $\boldsymbol{x}_{ik}$'s and the vector of regression coefficients $\boldsymbol{\beta}_*$ are fixed through simulations.

Table 4.1: Relative biases of estimating cAI by CScAIC, $\widehat{\text{cAI}}$ and $\widehat{\text{cAI}}^{\dagger}$ for unit level predictive model

	true value	relative bias		
model	cAI	CScAIC	$\widehat{\text{cAI}}$	$\widehat{\text{cAI}}^{\dagger}$
pattern (a): $q = 10$				
j_1	206.38	-33.439	-0.33371	-0.080972
j_2	152.81	-18.840	-0.2414	-0.16414
j_3	140.79	-18.556	-0.23026	-0.46789
j_4	132.61	-11.451	0.26445	-0.19514
j_5	116.51	-0.0019291	1.5979	0.49514
j_6	122.46	-0.050686	0.77756	0.15429
j_7	128.88	0.0086256	0.09468	0.09468
pattern (b): $q = 15$				
j_1	233.35	-0.22534	0.11255	0.1159
j_2	189.54	9.431	0.24145	0.21667
j_3	177.28	14.197	0.42246	0.37347
j_4	163.62	-0.76563	0.32597	0.02934
j_5	152.94	0.13627	0.8566	0.25115
j_6	156.73	0.068668	0.53817	0.15002
j_7	161.65	0.083897	0.015869	0.015869
pattern (c): $q = 20$				
j_1	299.12	4.3775	0.084838	0.084654
j_2	252.60	6.1677	0.24072	0.23581
j_3	250.27	-5.1825	-0.016634	0.010911
j_4	208.25	2.6115	0.36013	0.23929
j_5	197.52	0.25682	0.53321	0.26101
j_6	200.38	0.24977	0.40855	0.25647
j_7	203.57	0.22713	0.21272	0.21272

Table 4.1 and 4.2 report the true values of the cAI and the relative biases of estimating the cAI by the criteria CScAIC in (4.7), $\widehat{\text{cAI}}$ in (4.19) and $\widehat{\text{cAI}}^{\dagger}$ in (4.22), for the experiment using the unit level predictive model and for the experiment using the area level predictive model, respectively. We handle the cases where the number of the areas $q = 10,\ 15,\ 20$. The true values of cAI in each candidate model are calculated based on (4.4) with 10000 Monte Carlo iterations. The relative biases of estimating the cAI by the criteria is defined as

$$100 \times \frac{E[\text{IC}] - \text{cAI}}{\text{cAI}},$$

where IC = CScAIC, $\widehat{\text{cAI}}$, $\widehat{\text{cAI}}^{\dagger}$ and expectation is computed based on 1000 replications. From the tables, we can see the following facts. Firstly, the CScAIC has large biases for underspecified

Table 4.2: Relative biases of estimating cAI by CScAIC, $\widehat{\text{cAI}}$ and $\widehat{\text{cAI}}^{\dagger}$ for area level predictive model

	true value	relative bias		
model	cAI	CScAIC	$\widehat{\text{cAI}}$	$\widehat{\text{cAI}}^{\dagger}$
pattern (a): $q = 10$				
j_1	61.095	-10.429	-0.27157	-0.21359
j_2	46.635	5.4222	0.40292	-0.055653
j_3	50.744	-10.285	0.36735	-0.23857
j_4	47.323	-0.77484	1.3412	0.36198
j_5	45.735	-0.21108	2.2751	0.60073
j_6	49.452	-0.32466	0.82334	0.017969
j_7	52.805	-0.29278	-0.082743	-0.082743
pattern (b): $q = 15$				
j_1	95.393	-5.4716	0.12331	0.13930
j_2	70.056	17.521	0.39905	0.36599
j_3	66.412	21.039	0.59611	0.53872
j_4	61.310	5.174	0.58453	0.10853
j_5	60.532	0.23723	1.0853	0.25515
j_6	63.109	0.13276	0.50306	0.096935
j_7	65.379	0.16648	-0.0017143	-0.0017143
pattern (c): $q = 20$				
j_1	98.841	21.017	0.18161	0.17565
j_2	94.83	10.059	0.31607	0.30894
j_3	88.346	5.6464	0.13835	0.11302
j_4	78.383	7.8734	0.46942	0.27593
j_5	77.794	0.18930	0.54202	0.17009
j_6	79.277	0.18908	0.41365	0.18534
j_7	81.216	0.15395	0.11783	0.11783

models, namely j_1, j_2, j_3 and j_4, while the modified estimators of the cAI, $\widehat{\text{cAI}}$ and $\widehat{\text{cAI}}^{\dagger}$ have smaller biases for both overspecified and underspecified models. Secondly, $\widehat{\text{cAI}}^{\dagger}$ can estimate the cAI more unbiasedly than $\widehat{\text{cAI}}$ can for the case of small sample size because $\widehat{\text{cAI}}^{\dagger}$ is third order asymptotically unbiased estimator of the cAI. However, the relative biases of $\widehat{\text{cAI}}$, which is the second order asymptotically unbiased estimator of the cAI, gets smaller as the sample size is larger and the difference in performance between $\widehat{\text{cAI}}$ and $\widehat{\text{cAI}}^{\dagger}$ is not very important.

4.5.2 Simulations of predicting finite population mean

In this subsection, we investigate the numerical performance of the small are prediction problem explained in Section 4.4. We consider the model class which consists of all subsets of the full model $\omega = \{1, \ldots, p_\omega\}$ for $p_\omega = 5$. The true observed model is the NERM in (4.23) with $\sigma^2 = \tau^2 = 1$ and $n_i = 3$ for $i = 1, \ldots, q$, but we consider that τ^2 is unknown, namely $\boldsymbol{G} = \psi \boldsymbol{I}_q$ for $\psi = \tau^2/\sigma^2$ and ψ is the unknown parameter. Estimating procedure of ψ is the same as the one introduced in Section 3.4. We consider both the unit level predictive model (4.25) and the area level predictive model (4.26). The explanatory variables $\boldsymbol{x}_{ik}$'s for $k = 1, \ldots, n_i$, namely the observed model, are independently generated by $\mathcal{N}(4\mathbf{1}_{p_\omega}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = 0.9\boldsymbol{I}_{p_\omega} + 0.1\boldsymbol{J}_{p_\omega}$. On the other hand, the explanatory variables $\boldsymbol{x}_{ik}$'s for $k = n_i + 1, \ldots, N_i$, namely the predictive model, are independently generated by $\mathcal{N}(a\mathbf{1}_{p_\omega}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ for $a = 2, 4, 6$. Under this setting, we measure the simulated prediction error of the best model selected by the conventional cAIC of Vaida and Blanchard (2005) and $\widehat{\text{cAI}}$'s based on unit level predictive model and area level predictive model. The prediction error is measured by quadratic loss

$$\sum_{i=1}^{q} (\widehat{\overline{Y}}_i - \overline{Y})^2,$$

where $\widehat{\overline{Y}}$ is given by (4.27). The prediction errors are given as the averages based on 1000 replications.

Table 4.3: Prediction errors based on Vaida and Blanchard (2005)'s cAIC (VB), $\widehat{\text{cAI}}$ using unit level predictive model (unit) and $\widehat{\text{cAI}}$ using area level predictive model (area). The values in parentheses are the improvement over VB expressed in percentage.

	VB	unit	area
a=2	0.16966	0.16925	0.16887
		(0.24)	(0.47)
a=4	0.15419	0.15382	0.15322
		(0.24)	(0.63)
a=6	0.15346	0.15316	0.15257
		(0.20)	(0.58)

Table 4.3 reports the prediction errors of the best model selected by the cAIC of Vaida and Blanchard (2005) denoted by 'VB', $\widehat{\text{cAI}}$ using unit level predictive model denoted by 'unit', and $\widehat{\text{cAI}}$ using area level predictive model denoted by 'area'. The values in parentheses are the improvement over the prediction error based on the cAIC expressed in percentage. It can be seen that our proposed criteria are better than the cAIC in the sense that predicting the small

area mean. It is valuable to point out that the prediction error of the mean of finite population can be improved by using our proposed criteria, which motivates us to use them for variable selection in small area prediction of the finite population.

4.6 Proofs

Firstly, we introduce the following lemma, which was shown in Section A.2 of Srivastava and Kubokawa (2010).

Lemma 4.6 *Assume that $\boldsymbol{C}$ is an $n \times n$ symmetric matrix, $\boldsymbol{M}$ is an idempotent matrix of rank p and that $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then,*

$$E\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right] = \frac{\mathrm{tr}\,(\boldsymbol{C})}{n-p-2} - \frac{2\mathrm{tr}\,[\boldsymbol{C}(\boldsymbol{I}_n - \boldsymbol{M})]}{(n-p)(n-p-2)}.$$

4.6.1 Proof of Theorem 4.1

Because the cAI can be evaluated as (4.6) for the overspecified case, it suffices to show that

$$\begin{aligned} &E[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2] \\ &= \frac{n}{n-p_j}\left\{\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] - \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}]\right\}. \end{aligned} \tag{4.28}$$

It follows that

$$\begin{aligned} &\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j \\ &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j) \\ &= \boldsymbol{R}\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})\right\} \\ &= \boldsymbol{R}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}). \end{aligned}$$

Then, we can see that

$$\begin{aligned} &E[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2] \\ &= E[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\hat{\sigma}_j^2] + E[(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})/\hat{\sigma}_j^2] \\ &\quad - 2E[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})/\hat{\sigma}_j^2]. \end{aligned}$$

We define $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma_*$ and $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}$, which are the same notations as those in Section 4.3. Then, we can rewrite the equation above as

$$\begin{aligned} &E\big[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2\big] \\ &= nE\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}}{\boldsymbol{v}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right] + nE\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right] - 2nE\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right] \\ &= I_1 + I_2 - 2I_3 \quad \text{(say)}. \end{aligned}$$

It follows from Lemma 4.6 that

$$I_1 = n \times \left\{ \frac{\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}]}{n-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I}_n-\boldsymbol{M})]}{(n-p_j)(n-p_j-2)} \right\}$$
$$= n \times \left\{ \frac{\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}]}{n-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}] - 2\mathrm{tr}\,[\boldsymbol{R}\boldsymbol{P}]}{(n-p_j)(n-p_j-2)} \right\}. \tag{4.29}$$

As for I_2, because $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}$ is independent of $\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}$, we can obtain

$$I_2 = \frac{n}{n-p_j-2} \cdot \mathrm{tr}\,[\boldsymbol{R}\boldsymbol{P}].$$

To evaluate I_3, we rewrite

$$I_3 = nE\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}} \right] + nE\left[\frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}} \right]. \tag{4.30}$$

Because $\boldsymbol{M}\boldsymbol{v}$ 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 $I_3 = I_2$ and (4.28) form (4.29) and (4.30). □

4.6.2 Proof of Theorem of 4.2

For the overspecified case, the cAI is evaluated as (4.6), which is the same expression as Theorem 4.2. Thus we show that the cAI in (4.11) is evaluated as (4.12) for the underspecified case.

Because $E[(K_0+K_1)^{-1}]$ is evaluated as (3.32) in Chapter 3, it suffices to evaluate $E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]$. Let $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*$. Then, we can rewrite $\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*$ in $\boldsymbol{a}$ as

$$\begin{aligned}\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* &= \widetilde{\boldsymbol{P}}_j(\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{u}) - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* \\ &= (\widetilde{\boldsymbol{P}}_j\boldsymbol{X}(\omega) - \widetilde{\boldsymbol{X}}(\omega))\boldsymbol{\beta}_* + \sigma_*\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}.\end{aligned}$$

Next, we can rewrite $\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*$ in $\boldsymbol{a}$ as

$$\begin{aligned}\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* &= \sigma_*\boldsymbol{\Sigma}^{1/2}\{\boldsymbol{M}_j(\boldsymbol{W}_\omega\boldsymbol{\beta}_* + \boldsymbol{v}) - \boldsymbol{W}_\omega\boldsymbol{\beta}_*\} \\ &= \sigma_*\boldsymbol{\Sigma}^{1/2}\{-(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{W}_\omega\boldsymbol{\beta}_* + \boldsymbol{M}_j\boldsymbol{v}\} \\ &= -\boldsymbol{\Sigma}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \sigma_*\boldsymbol{\Sigma}^{1/2}\boldsymbol{M}_j\boldsymbol{v}.\end{aligned}$$

Then, we obtain

$$\boldsymbol{a} = \boldsymbol{B}\boldsymbol{\beta}_* + \sigma_*(\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_j)\boldsymbol{v}.$$

Moreover, it follows that

$$\begin{aligned}\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_j &= (\widetilde{\boldsymbol{X}}(j) - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{A}(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\end{aligned}$$

Thus, $E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]$ can be evaluated as

$$E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2] = \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{A}^{\mathrm{T}}] + \boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2. \tag{4.31}$$

It follows from (3.32) and (4.31) that

$$\begin{aligned}&n \cdot E[(K_0+K_1)^{-1}]\{\mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]\} \\ &= (\gamma + \boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2) \times \left\{ \lambda + \frac{-2\lambda^3 + (p_j+4)\lambda^2}{n} \right\} + O(n^{-1}),\end{aligned}$$

which shows that the cAI in (4.11) is approximated to (4.12). □

4.6.3 Proof of Lemma 4.1

When the candidate model j is overspecified, $E(\hat{\lambda}) = E(\widehat{\lambda^2}) = E(\widehat{\lambda^3}) = 1$ because $\hat{\sigma}^2_\omega/\hat{\sigma}^2_j \sim \mathrm{Be}((n-p_\omega)/2, (p_\omega - p_j)/2)$. Thus, it follows that $E(\widehat{R_2}) = 0$.

When the candidate model j is underspecified, it follows that

$$E\left[\left(\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_j}\right)^k\right] = E\left[\left(\frac{K_0}{K_0+K_1}\right)^k\right] = \lambda^k + O(n^{-1}), \tag{4.32}$$

for $k = 1, 2, 3$ because $(K_0 + K_1)^{-1}$ is expanded as (3.31) and $K_0 = \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M}_\omega)\boldsymbol{v}$. Thus, it follows that $\widehat{R_2} = R_2 + O(n^{-1})$. □

4.6.4 Proof of Lemma 4.2

When the candidate model j is overspecified, $E(\widehat{R_1}) = 0$ because $E(\hat{\lambda}) = E(\widehat{\lambda^2}) = E(\widehat{\lambda^3}) = 1$.

When the candidate model j is underspecified, it follows from (3.33) that

$$E(\hat{\lambda}) = \lambda + \frac{-\lambda^3 + (p_j+2)\lambda^2 - p_j\lambda}{n} + O(n^{-2}). \tag{4.33}$$

Thus it follows from (4.32) and (4.33) that $E(\widehat{R_1}) = R_1 + O(n^{-1})$. □

4.6.5 Proof of Lemma 4.3

Firstly, note that

$$R_3(\boldsymbol{\eta}_*) = \lambda \cdot \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B}\boldsymbol{\beta}_*/\sigma^2_*,$$

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

$$\delta = \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{X}(\omega)^{\mathrm{T}}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\boldsymbol{\beta}_*/(n\sigma^2_*).$$

Then, we can see that

$$\frac{\partial R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\eta}_*} = \frac{\partial \lambda}{\partial \delta} \cdot \frac{\partial \delta}{\partial \boldsymbol{\eta}_*}$$

and that $\partial\lambda/\partial\delta = -(1+\delta)^{-2}$. After some calculations, we can obtain Lemma 4.3. □

4.6.6 Proof of Lemma 4.4

Firstly, note that $E[R_3(\widetilde{\boldsymbol{\eta}})]$ is expanded as

$$E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + O(n^{-1}),$$

where $B_1(\boldsymbol{\eta}_*)$ is given as (4.17). Because $B_1(\boldsymbol{\eta}_*) = O(1)$, it follows that $B_1(\widetilde{\boldsymbol{\eta}}) = B_1(\boldsymbol{\eta}_*) + O(n^{-1})$, which shows that

$$E[\widetilde{\widetilde{R_3}}] = R_3 + O(n^{-1}).$$

In the same way, we can obtain $E[\widetilde{R_4}] = E[R_4(\widetilde{\boldsymbol{\eta}})] = R_4(\boldsymbol{\eta}_*) + O(n^{-1})$. □

4.6.7 Proof of Lemma 4.5

It follows from (4.21) that

$$E[\widehat{R_3}] = E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}})\right].$$

Because $E[R_3(\widetilde{\boldsymbol{\eta}})]$ is expanded as $E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) + O(n^{-2})$, we can see that

$$\begin{aligned} E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)]\right] &= 2\left\{R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*)\right\} - E\left[R_3(\widetilde{\boldsymbol{\eta}}) + B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}})\right] + O(n^{-2}) \\ &= R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) - E\left[B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}})\right] + O(n^{-2}). \end{aligned}$$

Moreover, because $E[B_1(\widetilde{\boldsymbol{\eta}})] = B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*) + O(n^{-2})$ and $E[B_2(\widetilde{\boldsymbol{\eta}})] = B_2(\boldsymbol{\eta}_*) + O(n^{-2})$, the equation above can be rewritten as

$$E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)]\right] = R_3(\boldsymbol{\eta}_*) - B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}). \tag{4.34}$$

Next, it is seen that

$$\begin{aligned} E\left[E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}})\right] &= E[B_1(\widetilde{\boldsymbol{\eta}}) + B_{11}(\widetilde{\boldsymbol{\eta}})] - \{B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*)\} + O(n^{-2}) \\ &= B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}). \end{aligned} \tag{4.35}$$

Thus, it follows from (4.34) and (4.35) that

$$E[\widehat{R_3}] = R_3(\boldsymbol{\eta}_*) + O(n^{-2}).$$

Similarly, we can show that $E[\widehat{R_4}] = R_4(\boldsymbol{\eta}_*) + O(n^{-2})$. □

4.6.8 Proof of Theorem 4.3 and Theorem 4.4

From Theorem 4.2 and Lemmas 4.1–4.5, we can easily show the theorems. □

Chapter 5

Conditional AIC in mixed effects models based on natural exponential family

In this chapter, we consider the variable selection problem for the class of mixed effects models based on natural exponential family, which includes useful nonlinear mixed models, Poisson-gamma model and binomial-beta model. We construct the conditional AIC in the models and show the usefulness for variable selection.

5.1 Motivation

For variable selection problem in mixed effects models, Vaida and Blanchard (2005) introduced the conditional Akaike information (cAI), which is related to the expected Kullback–Leibler divergence based on the conditional likelihood given random effects. The cAI and the resulting information criterion conditional AIC (cAIC) are appropriate when one is interested in predicting the random effects. Vaida and Blanchard (2005) proposed the cAIC for variable selection criterion in normal linear mixed model. Since then, the cAIC has been studied for various models, which include generalized linear mixed model (GLMM) as well as linear mixed model.

However, variable selection problem in nonlinear mixed model, for example Poisson-gamma model or binomial-beta model, has not been considered well. Although the conventional AIC (or marginal AIC, mAIC), which is based on the marginal likelihood integrating out the random effects, can be used as a variable selection criterion, the mAIC is not appropriate for predicting the random effects. Then, we consider the cAI for mixed effects models based on natural exponential family and construct the cAIC as an asymptotically unbiased estimator of the cAI.

The rest of this chapter is organized as follows. In Section 5.2, we explain about the mixed effects models based on natural exponential family and define the cAI for the class of the models. In Section 5.3, we evaluate and estimate the bias correction and construct the cAIC in three ways. The numerical performance of the proposed criteria is investigated by simulations in Section 5.4. Section 5.5 shows some results of analytical calculations and Section 5.6 gives proofs of lemmas and theorems.

5.2 Model and conditional AIC

5.2.1 Mixed effects models based on natural exponential family

Let $y_1, \ldots, y_m$ be mutually independent random variables where the conditional distribution of y_i given θ_i belongs to the following natural exponential family (NEF):

$$y_i|\theta_i \sim f(y_i|\theta_i) = \exp[n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)], \quad (i = 1, \ldots, m), \tag{5.1}$$

where θ_i is the natural parameter, n_i is a known scale parameter and $\psi(\cdot)$ and $c(\cdot, \cdot)$ are functions specific to each distribution. As an example of the random variable y_i with probability density (or mass) function (5.1), we consider the following situation.

For $i = 1, \ldots, m$, which denote clusters (or areas in the context of small area estimation), let the random variables $Z_{i1}, \ldots, Z_{i,n_i}$ be independent and identically distributed given θ_i with the common distribution belonging to the one parameter exponential family, namely,

$$P(Z_{ij} \in A) = \int_A \exp\{\theta_i z - \psi(\theta_i)\} \mathrm{d}F(z),$$

with F a Stieltjes measure on $\mathbb{R}$. When we define $y_i = (Z_{i1} + \cdots + Z_{i,n_i})/n_i$, the conditional distribution of y_i given θ_i has the density function of the form (5.1). This situation is also considered in Ghosh and Maiti (2008), who developed empirical Bayes (EB) confidence intervals for population means in each small area with NEF distributions. In their work, the result is based on an asymptotic theory in the sense that $n_i \to \infty$ and $m \to \infty$. We also consider the same setup to derive a criterion in Section 5.3.2.

From the property of NEF, the mean of y_i given θ_i is

$$\mu_i = E(y_i|\theta_i) = \psi'(\theta_i),$$

where $\psi'(\cdot)$ denotes the derivative of $\psi(\cdot)$. Define $Q(\mu_i) = \psi''(\theta_i)$, where $\psi''(\cdot)$ is the second derivative of $\psi(\cdot)$, then the variance of y_i given θ_i is

$$V(y_i|\theta_i) = \frac{\psi''(\theta_i)}{n_i} = \frac{Q(\mu_i)}{n_i}.$$

$Q(\cdot)$ is called the variance function and we hereafter assume that $Q(\cdot)$ is a quadratic function, namely $Q(x) = v_0 + v_1 x + v_2 x^2$ for known constants v_0, v_1 and v_2, which are not simultaneously zero. The family of such distributions is called natural exponential family with quadratic variance function (NEF-QVF), many properties of which are studied by Morris (1982, 1983).

As the random cluster (area) effect, we consider the conjugate prior for θ_i with the probability density function

$$p(\theta_i|\lambda, m_i) = \exp[\lambda(m_i\theta_i - \psi(\theta_i))]C(\lambda, m_i), \tag{5.2}$$

where λ is an unknown scalar hyperparameter and $C(\cdot, \cdot)$ is a function specific to each distribution. The mean and variance of $\mu_i = E(y_i|\theta_i)$ are

$$E(\mu_i|\lambda, m_i) = m_i \quad \text{and} \quad V(\mu_i|\lambda, m_i) = \frac{Q(m_i)}{\lambda - v_2},$$

respectively. When $p \times 1$ auxiliary variable $\boldsymbol{x}_i$ is available, one uses $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ as a predictor of m_i, where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients. We here consider the following link function $h(\cdot)$ between m_i and $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$:

$$m_i = h^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

If $h^{-1}(\cdot) = \psi'(\cdot)$, the link function $h(\cdot)$ is called canonical. We henceforth consider the canonical link. We define the vector of the unknown parameters in this model as $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \lambda)^{\mathrm{T}}$.

The posterior density function of θ_i given y_i, and the marginal density (or mass) function of y_i are

$$p(\theta_i|y_i, \lambda, m_i) = \exp[(n_i + \lambda)(\widehat{\mu}_i^{\mathrm{B}}\theta_i - \psi(\theta_i))]C(n_i + \lambda, \widehat{\mu}_i^{\mathrm{B}}), \tag{5.3}$$

$$m(y_i|\lambda, m_i) = \frac{C(\lambda, m_i)}{C(n_i + \lambda, \widehat{\mu}_i^{\mathrm{B}})} \exp[c(y_i, n_i)], \tag{5.4}$$

respectively, where $\widehat{\mu}_i^{\mathrm{B}}$ is the Bayes estimator of μ_i under quadratic loss, namely

$$\widehat{\mu}_i^{\mathrm{B}} = \widehat{\mu}_i(y_i, \boldsymbol{\eta}) = \frac{n_i y_i + \lambda m_i}{n_i + \lambda}, \quad m_i = \psi'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

We can get the empirical Bayes (EB) estimator of μ_i by substituting $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\lambda})^{\mathrm{T}}$ for $\boldsymbol{\eta}$ in $\widehat{\mu}_i(y_i, \boldsymbol{\eta})$ as follows:

$$\widehat{\mu}_i^{\mathrm{EB}} = \widehat{\mu}_i(y_i, \widehat{\boldsymbol{\eta}}) = \frac{n_i y_i + \hat{\lambda}\hat{m}_i}{n_i + \hat{\lambda}}, \quad \hat{m}_i = \psi'(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ are some estimators of $\boldsymbol{\beta}$ and λ based on the marginal distribution of $\boldsymbol{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$.

The model explained above is an example of general mixed effects model (2.1) for $\boldsymbol{y}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ and is called nonlinear mixed model unless $h(\cdot)$ is the identity function, which is the case of normal linear mixed model. Nonlinear mixed model is useful in the context of small area estimation and is used by Ghosh and Maiti (2004, 2008), Lohr and Rao (2009) and others. We provide two useful examples belonging to the nonlinear mixed model, one for the count and the other for binary data sets.

[1] Poisson-gamma mixture model. Let $z_1, \ldots, z_m$ be mutually independent random variables having

$$z_i|\mu_i \sim \mathrm{Po}(n_i\mu_i) \quad \text{and} \quad \mu_i \sim \mathrm{Ga}(\lambda m_i, \lambda^{-1}),$$

where $\mu_1, \ldots, \mu_m$ are mutually independent, $\mathrm{Po}(\mu)$ denotes Poisson distribution with mean μ, and $\mathrm{Ga}(a, b)$ denotes gamma distribution with density function

$$f(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-x/b), \quad x > 0,$$

where $\Gamma(\cdot)$ denotes gamma function. Let $y_i = z_i/n_i$. Then the probability mass function of y_i given μ_i is (5.1) with $\theta_i = \log(\mu_i)$, $\psi(\cdot) = \exp(\cdot)$ and the probability density function of θ_i is (5.2). In this model, canonical link function is $h(\cdot) = \psi'^{-1}(\cdot) = \log(\cdot)$ and the quadratic variance function is $Q(x) = x$, namely $v_1 = 1$, $v_0 = v_2 = 0$.

[2] binomial-beta mixture model. Let $z_1, \ldots, z_m$ be mutually independent random variables having

$$z_i|\mu_i \sim \mathrm{Bin}(n_i, \mu_i) \quad \text{and} \quad \mu_i \sim \mathrm{Beta}(\lambda m_i, \lambda(1 - m_i)),$$

where $\mu_1, \ldots, \mu_m$ are mutually independent, $\mathrm{Bin}(n, p)$ denotes binomial distribution, and $\mathrm{Beta}(a, b)$ denotes beta distribution. Let $y_i = z_i/n_i$. Then the probability mass function of y_i given μ_i is (5.1) with $\theta_i = \mathrm{logit}(\mu_i)$ and $\psi(\theta_i) = \log(1 + \exp(\theta_i))$ and the probability density function of θ_i is (5.2). In this model, canonical link function is $h(\cdot) = \psi'^{-1}(\cdot) = \mathrm{logit}(\cdot)$ and the quadratic variance function is $Q(x) = x - x^2$, namely $v_0 = 0$, $v_1 = 1$, $v_2 = -1$.

5.2.2 Variable selection problem in nonlinear mixed model

For the variable selection problem in the nonlinear mixed model, the marginal AIC (mAIC) can be easily used, because the marginal likelihood (5.4) is obtained analytically. However, when one wants to predict the random area effect θ_i, the mAIC based on the marginal likelihood is not appropriate (Vaida and Blanchard, 2005). Then we derive the cAIC for nonlinear mixed model.

In the same way as the previous chapters, we define candidate models by the index set j, which is a subset of $\omega = \{1, \ldots, p_\omega\}$. Then, we define $\boldsymbol{X}(j) = (\boldsymbol{x}_1(j), \ldots, \boldsymbol{x}_m(j))^{\mathrm{T}}$, where $\boldsymbol{x}_i(j)$ is a $p_j \times 1$ vector for $p_j = \#(j)$. Let j_* denote the true model and the dimension of the true model be p_{j_*}, which is abbreviated to p_*. To derive the criterion, we here assume that the candidate model j is overspecified. Under the assumption, the mean of the true model can be expressed as

$$m_i = \psi'(\boldsymbol{x}_i(j)^{\mathrm{T}}\boldsymbol{\beta}_j^*),$$

where $\boldsymbol{\beta}_j^*$ is $p_j \times 1$ vector of regression coefficients, whose $p_j - p_*$ components are exactly 0 and the rest of components are not 0. Thus we henceforth abbreviate $\boldsymbol{x}_i(j)$ to $\boldsymbol{x}_i$, $\boldsymbol{\beta}_j^*$ to $\boldsymbol{\beta}$ for notational convenience.

5.2.3 Conditional AIC in nonlinear mixed model

Vaida and Blanchard (2005) proposed the cAIC as an (asymptotically) unbiased estimator of a part of the prediction risk of the plug-in predictive density relative to the expected Kullback–Leibler divergence, which is called the conditional Akaike Information (cAI). Firstly, let us define the conditional Akaike Information (cAI) for the nonlinear mixed model. Let $\tilde{y}_i$ be an independent replication of y_i given θ_i. Then the logarithm of the plug-in predictive density is

$$\log\{f(\tilde{y}_i|\hat{\theta}_i^{\mathrm{EB}})\} = n_i(\hat{\theta}_i^{\mathrm{EB}}\tilde{y}_i - \psi(\hat{\theta}_i^{\mathrm{EB}})) + c(\tilde{y}_i, n_i),$$

where $\hat{\theta}_i^{\mathrm{EB}} = \psi'^{-1}(\widehat{\mu}_i^{\mathrm{EB}}) = h(\widehat{\mu}_i^{\mathrm{EB}})$. Because the second term of the equation above, namely $c(\tilde{y}_i, n_i)$, is irrelevant to the candidate model, we define the cAI in the nonlinear mixed model as

$$\begin{aligned}
\mathrm{cAI} = & -2E^{(\boldsymbol{y},\boldsymbol{\theta})}E^{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}\left[\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}\tilde{y}_i - \psi(\hat{\theta}_i^{\mathrm{EB}}))\right] \\
= & -2E^{(\boldsymbol{y},\boldsymbol{\theta})}\left[\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}\mu_i - \psi(\hat{\theta}_i^{\mathrm{EB}}))\right], \qquad (5.5)
\end{aligned}$$

where $E^{(\boldsymbol{y},\boldsymbol{\theta})}$ and $E^{\tilde{\boldsymbol{y}}|\boldsymbol{\theta}}$ denote expectation with respect to the joint distribution of $(\boldsymbol{y}, \boldsymbol{\theta})$ and the conditional distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{\theta}$ for $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)^{\mathrm{T}}$, respectively. When we estimate the cAI by $-2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}y_i - \psi(\hat{\theta}_i^{\mathrm{EB}}))$, the bias is

$$\begin{aligned}
E\left[-2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}y_i - \psi(\hat{\theta}_i^{\mathrm{EB}}))\right] - \mathrm{cAI} = & -2E\left[\sum_{i=1}^{m} n_i\hat{\theta}_i^{\mathrm{EB}}(y_i - \mu_i)\right] \\
= & -2B \quad \text{(say)}. \qquad (5.6)
\end{aligned}$$

Then we propose the cAIC for the nonlinear mixed model as bias corrected estimator of the cAI, which is given by

$$\mathrm{cAIC} = -2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}y_i - \psi(\hat{\theta}_i^{\mathrm{EB}})) + 2\widehat{B},$$

where $\widehat{B}$ is an asymptotically unbiased estimator of B, which we call the bias correction term or the penalty term. In the next section, we give an asymptotic approximation and an estimator of the penalty term B.

5.3 Approximation and estimation of penalty term

5.3.1 Decomposition of penalty term

Because it is difficult to evaluate the penalty term B exactly, we give second-order approximation of B for large m. Taylor series expansion of $\hat{\theta}_i^{\mathrm{EB}} = h\{\widehat{\mu}_i(y_i, \widehat{\boldsymbol{\eta}})\}$ around $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$ gives

$$h(\widehat{\mu}_i^{\mathrm{EB}}) = h(\widehat{\mu}_i^{\mathrm{B}}) + \frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \frac{1}{2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}} \frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + o_p(m^{-1}).$$

Then the penalty term B in (5.6) can be expanded as

$$\begin{aligned} B &= E\left[\sum_{i=1}^{m} n_i h(\widehat{\mu}_i^{\mathrm{B}})(y_i - \mu_i)\right] + E\left[\sum_{i=1}^{m} n_i (y_i - \mu_i) \frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right], \\ &\quad + \frac{1}{2} E\left[\sum_{i=1}^{m} n_i (y_i - \mu_i)(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}} \frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right] + o(1) \\ &= B_1(\boldsymbol{\eta}) + B_2(\boldsymbol{\eta}) + 2^{-1} B_3(\boldsymbol{\eta}) + o(1), \quad \text{(say)} \end{aligned} \tag{5.7}$$

where $B_1 = O(m)$, $B_2 = O(1)$ and $B_3 = O(1)$. We evaluate each term of (5.7) in the following subsections by three different methods. To this end, we consider the following conditions:

(C1) The number of clusters m goes to infinity, and for each $i = 1, \ldots, m$, n_i is sufficiently large, which is of order $n_i = O(m^{1/2+\delta_i})$ for some positive $\delta_i > 0$.

(C2) The number of clusters m goes to infinity, and for each $i = 1, \ldots, m$, $n_i = O(1)$.

In the next subsection, we evaluate and estimate B_1, B_2 and B_3 by analytical method under the condition (C1). In Section 5.3.3 and 5.3.4, we consider the case of constant n_i, namely the condition (C2). In Section 5.3.3, we propose a method using numerical integration and differentiation, and in Section 5.3.4 we give a numerical method based on parametric bootstrap.

5.3.2 Analytical method for the case of large n_i

Firstly, we evaluate B_1, which is given by

$$B_1(\boldsymbol{\eta}) = E\left[\sum_{i=1}^{m} n_i h(\widehat{\mu}_i^{\mathrm{B}})(y_i - \mu_i)\right]. \tag{5.8}$$

Though the Bayes estimator $\widehat{\mu}_i^{\mathrm{B}} = (n_i y_i + \lambda m_i)/(n_i + \lambda)$ is written as the linear function of y_i, the function $h(\cdot)$, which is the link between the mean parameter and the natural parameter, is nonlinear for most of the members of the natural exponential family except for the normal distribution. Thus it is difficult to evaluate B_1 exactly. We here use closeness of $\widehat{\mu}_i^{\mathrm{B}}$ and μ_i for large n_i, and expand $h(\widehat{\mu}_i^{\mathrm{B}}) = h(\mu_i + \widehat{\mu}_i^{\mathrm{B}} - \mu_i)$ in (5.8). Then $h(\widehat{\mu}_i^{\mathrm{B}})$ is approximated by polynomials of the linear function of random variables y_i and μ_i, and B_1 can be asymptotically approximated.

We give two examples of approximation of B_1, one is the Poisson-gamma mixture model, where $h(\cdot)$ is the logarithmic function, the other is the binomial-beta mixture model, where $h(\cdot)$ is the logit function. For the Poisson-gamma mixture model, $h(\widehat{\mu}_i^{\mathrm{B}}) = \log(\widehat{\mu}_i^{\mathrm{B}})$ is expanded as

$$\begin{aligned} h(\widehat{\mu}_i^{\mathrm{B}}) =& \log\left\{\mu_i\left(1+\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)\right\} \\ =& \theta_i + \frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i} - \frac{1}{2}\left(\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)^2 + \frac{1}{3}\left(\frac{\widehat{\mu}_i^{B}-\mu_i}{\mu_i}\right)^3 + \sum_{k=4}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)^k. \end{aligned}$$

Because $E\left[\{(\widehat{\mu}_i^{\mathrm{B}}-\mu_i)/\mu_i\}^r(y_i-\mu_i)\right] = O(n_i^{-3})$ for $r \geq 4$, if $n_i = O(m^{1/2+\delta_i})$ for some positive $\delta_i > 0$, it follows that

$$\begin{aligned} B_1(\boldsymbol{\eta}) =& \sum_{i=1}^{m} n_i E[\theta_i(y_i-\mu_i)] + \sum_{i=1}^{m} n_i E\left[\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}(y_i-\mu_i)\right] - \frac{1}{2}\sum_{i=1}^{m} n_i E\left[\left(\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)^2(y_i-\mu_i)\right] \\ &+ \frac{1}{3}\sum_{i=1}^{m} n_i E\left[\left(\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)^3(y_i-\mu_i)\right] + o(1). \end{aligned}$$

We can evaluate each term in the equation above exactly, noting that

$$\widehat{\mu}_i^{\mathrm{B}} - \mu_i = \frac{n_i}{n_i+\lambda}(y_i-\mu_i) - \frac{\lambda}{n_i+\lambda}(\mu_i - m_i).$$

After some calculations, we can obtain the following lemma.

Lemma 5.1 *Under the condition* (C1)*, for the Poisson-gamma mixture model,* B_1 *in* (5.8) *is approximated up to second-order as*

$$B_1(\boldsymbol{\eta}) = B_{11}(\boldsymbol{\eta}) + o(1),$$

where

$$B_{11}(\boldsymbol{\eta}) = m - \sum_{i=1}^{m}\frac{2\lambda^2 m_i - \lambda}{2n_i(\lambda m_i - 1)}. \tag{5.9}$$

For the binomial-beta mixture model, where $h(\cdot)$ is the logit function, $h(\widehat{\mu}_i^{\mathrm{B}}) = \log\{\widehat{\mu}_i^{\mathrm{B}}/(1-\widehat{\mu}_i^{\mathrm{B}})\}$ is expanded as

$$\begin{aligned} h(\widehat{\mu}_i^{\mathrm{B}}) =& \log\left\{\mu_i\left(1+\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right)\right\} - \log\left\{(1-\mu_i)\left(1-\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{1-\mu_i}\right)\right\} \\ =& \theta_i + \log\left(1+\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{\mu_i}\right) - \log\left(1-\frac{\widehat{\mu}_i^{\mathrm{B}}-\mu_i}{1-\mu_i}\right) \\ =& \theta_i + (C_i+D_i) - \frac{1}{2}(C_i^2-D_i^2) + \frac{1}{3}(C_i^3+D_i^3) + \sum_{k=4}^{\infty}\frac{1}{k}\left\{(-1)^{k+1}C_i^k + D_i^k\right\}, \end{aligned}$$

where $C_i = (\widehat{\mu}_i^{\mathrm{B}}-\mu_i)/\mu_i, D_i = (\widehat{\mu}_i^{\mathrm{B}}-\mu_i)/(1-\mu_i)$. If $n_i = O(m^{1/2+\delta_i}), \delta_i > 0$, it follows that

$$\begin{aligned} B_1 =& \sum_{i=1}^{m} n_i E[\theta_i(y_i-\mu_i)] + \sum_{i=1}^{m} n_i E[(C_i+D_i)(y_i-\mu_i)] - \frac{1}{2}\sum_{i=1}^{m} n_i E[(C_i^2-D_i^2)(y_i-\mu_i)] \\ &+ \frac{1}{3}\sum_{i=1}^{m} n_i E[(C_i^3+D_i^3)(y_i-\mu_i)] + o(1), \end{aligned}$$

where each term in the equation above can be evaluated exactly. After some calculations, we get the following lemma.

Lemma 5.2 *Under the condition* (C1)*, for the binomial-beta mixture model,* B_1 *in* (5.8) *is approximated up to second-order as*

$$B_1(\boldsymbol{\eta}) = B_{11}(\boldsymbol{\eta}) + o(1),$$

where

$$B_{11}(\boldsymbol{\eta}) = m - \sum_{i=1}^{m} \frac{\lambda+1}{n_i} + \sum_{i=1}^{m} \frac{4\lambda^2 m_i(1-m_i) - (\lambda+2)(\lambda-1)}{2n_i(\lambda m_i - 1)\{\lambda(1-m_i)-1\}}. \tag{5.10}$$

$B_{11}(\widehat{\boldsymbol{\eta}})$, where $B_{11}(\boldsymbol{\eta})$ is given by (5.9) or (5.10), is expanded as

$$B_{11}(\widehat{\boldsymbol{\eta}}) = B_{11}(\boldsymbol{\eta}) + \frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \frac{1}{2}\mathrm{tr}\left[\frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}\right] + o_p(1),$$

then we propose a bias corrected estimator of B_1 given by

$$\widehat{B}_1 = B_{11}(\widehat{\boldsymbol{\eta}}) - B_{12}(\widehat{\boldsymbol{\eta}}) - B_{13}(\widehat{\boldsymbol{\eta}}), \tag{5.11}$$

where

$$\begin{aligned} B_{12}(\boldsymbol{\eta}) &= \frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}} E(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \\ B_{13}(\boldsymbol{\eta}) &= \frac{1}{2}\mathrm{tr}\left[\frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}} E[(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}]\right]. \end{aligned}$$

When $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12) explained later, $E(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ and $E[(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}]$ are analytically approximated by (5.34) and (5.31). For the Poisson-gamma mixture model, $\partial B_{11}(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$ and $(\partial^2 B_{11}(\boldsymbol{\eta}))/(\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}})$ are

$$\frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \frac{\lambda^2 m_i}{2n_i(\lambda m_i - 1)^2}\boldsymbol{x}_i, \quad \frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \lambda} = \sum_{i=1}^{m} \frac{1}{n_i}\left\{\frac{1}{2(\lambda m_i - 1)^2} - 1\right\},$$

and

$$\begin{aligned} \frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} &= -\sum_{i=1}^{m} \frac{\lambda^2 m_i(\lambda m_i + 1)}{2n_i(\lambda m_i - 1)^3}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \quad \frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}\partial \lambda} = -\sum_{i=1}^{m} \frac{\lambda m_i}{n_i(\lambda m_i - 1)^3}\boldsymbol{x}_i, \\ \frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial \lambda^2} &= -\sum_{i=1}^{m} \frac{m_i}{n_i(\lambda m_i - 1)^3}. \end{aligned}$$

For the binomial-beta mixture model, it follows that

$$\begin{aligned} \frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{m} \frac{\lambda^2(\lambda-1)(\lambda-2)m_i(1-m_i)(1-2m_i)}{2n_i(\lambda m_i - 1)^2\{\lambda(1-m_i)-1\}^2}\boldsymbol{x}_i, \\ \frac{\partial B_{11}(\boldsymbol{\eta})}{\partial \lambda} &= -\sum_{i=1}^{m} \frac{1}{n_i} + \sum_{i=1}^{m} \frac{-\lambda(3\lambda-4)m_i(1-m_i) + (\lambda-1)^2}{2n_i(\lambda m_i - 1)^2\{\lambda(1-m_i)-1\}^2}, \end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \lambda^2(\lambda-1)(\lambda-2)\sum_{i=1}^{m}\Big\{\frac{\{1-6m_i(1-m_i)\}}{2n_i(\lambda m_i-1)^2\{\lambda(1-m_i)-1\}^2} \\
&\quad - \frac{\lambda^2 m_i(1-m_i)(1-2m_i)^2}{n_i(\lambda m_i-1)^3\{\lambda(1-m_i)-1\}^3}\Big\} m_i(1-m_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \\
\frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial\boldsymbol{\beta}\partial\lambda} &= -\sum_{i=1}^{m}\frac{\lambda(3\lambda-4)(1-2m_i)}{2n_i(\lambda m_i-1)^2\{\lambda(1-m_i)-1\}^2} \\
&\quad + \sum_{i=1}^{m}\frac{\lambda^2(1-2m_i)\{\lambda(3\lambda-4)m_i(1-m_i)-(\lambda-1)^2\}}{n_i(\lambda m_i-1)^3\{\lambda(1-m_i)-1\}^3} \\
\frac{\partial^2 B_{11}(\boldsymbol{\eta})}{\partial\lambda^2} &= \frac{\{2(\lambda-1)-m_i(1-m_i)(6\lambda-4)\}}{2n_i(\lambda m_i-1)^2\{\lambda(1-m_i)-1\}^2} \\
&\quad - \frac{\{2m_i(1-m_i)\lambda-1\}\{(\lambda-1)^2-\lambda(3\lambda-4)m_i(1-m_i)\}}{n_i(\lambda m_i-1)^3\{\lambda(1-m_i)-1\}^3}.
\end{aligned}$$

Next, we give an asymptotic approximation of B_2 and B_3 analytically when the hyperparameter $\boldsymbol{\eta}$ is estimated by the estimating equation suggested by Godambe and Thompson (1989). We define $\boldsymbol{g}_i = (g_{1i}, g_{2i})^{\mathrm{T}}$ for $g_{1i} = y_i - m_i$ and $g_{2i} = (y_i - m_i)^2 - \phi_i Q(m_i)$ and

$$\begin{aligned}
\boldsymbol{D}_i^{\mathrm{T}} &= E\left(-\frac{\partial \boldsymbol{g}_i^{\mathrm{T}}}{\partial\boldsymbol{\eta}}\right) \\
&= Q(m_i)\begin{bmatrix} \boldsymbol{x}_i & Q'(m_i)\phi_i\boldsymbol{x}_i \\ 0 & -(1+v_2/n_i)(\lambda-v_2)^{-2} \end{bmatrix}, \\
\boldsymbol{\Sigma}_i &= \mathbf{Cov}\left(\boldsymbol{g}_i\right) = \begin{bmatrix} \mu_{2i} & \mu_{3i} \\ \mu_{3i} & \mu_{4i}-\mu_{2i}^2 \end{bmatrix},
\end{aligned}$$

where $\mu_{ri} = E[(y_i - m_i)^r]$ and $\phi_i = (\lambda/n_i + 1)/(\lambda - v_2)$. Following Ghosh and Maiti (2004), the exact expressions of μ_{ri}'s for $r = 2, 3, 4$, are

$$\begin{aligned}
\mu_{2i} &= \phi_i Q(m_i), \quad \mu_{3i} = \frac{Q(m_i)Q'(m_i)(\lambda/n_i+1)(\lambda/n_i+2)}{(\lambda-v_2)(\lambda-2v_2)}, \\
\mu_{4i} &= (d_i+1)(2d_i+1)(3d_i+1)E[(\mu_i-m_i)^4] + 6n_i^{-1}Q'(m_i)(d_i+1)(2d_i+1)E[(\mu_i-m_i)^3] \\
&\quad + n_i^{-2}(d_i+1)\left[7\{Q'(m_i)\}^2 + 2n_i(4d_i+3)Q(m_i)\right]E[(\mu_i-m_i)^2] \\
&\quad + n_i^{-3}Q(m_i)\left[n_i(2d_i+3)Q(m_i) + \{Q'(m_i)\}^2\right],
\end{aligned}$$

where $d_i = v_2/n_i$ and $E[(\mu_i - m_i)^2] = Q(m_i)/(\lambda - v_2)$, $E[(\mu_i - m_i)^3] = 2Q(m_i)Q'(m_i)/\{(\lambda - v_2)(\lambda - 2v_2)\}$ and

$$E[(\mu_i-m_i)^4] = \frac{3Q(m_i)\left[(\lambda-2v_2)Q(m_i) + 2\{Q'(m_i)\}^2\right]}{(\lambda-v_2)(\lambda-2v_2)(\lambda-3v_2)}.$$

Then, Ghosh and Maiti (2004) derived the estimating equation:

$$\boldsymbol{s}(\boldsymbol{\eta}) = \mathbf{0} \quad \text{for} \quad \boldsymbol{s}(\boldsymbol{\eta}) = \sum_{i=1}^{m}\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{g}_i. \tag{5.12}$$

This method is also used in Ghosh and Maiti (2008) and Kubokawa et al. (2014). The estimating equation by Godambe and Thompson (1989) is an extension of quasi-likelihood methods

proposed by Wedderburn (1974) and $\boldsymbol{s}(\boldsymbol{\eta})$ is the 'extended quasi-score function'. In this context, 'quasi-Fisher information' is given by

$$\begin{aligned} E(\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}) = & \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} E(\boldsymbol{g}_i \boldsymbol{g}_i^{\mathrm{T}}) \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i \\ = & \boldsymbol{U} \quad \text{(say)}, \end{aligned}$$

and the asymptotic variance of $\widehat{\boldsymbol{\eta}}$, which is the solution of $\boldsymbol{s}(\boldsymbol{\eta}) = \boldsymbol{0}$, is $E(\widehat{\boldsymbol{\eta}}\widehat{\boldsymbol{\eta}}^{\mathrm{T}}) = \boldsymbol{U}^{-1} + o(m^{-1})$. Derivation of the asymptotic bias and variance of $\widehat{\boldsymbol{\eta}}$ are shown in Section 5.5.1 which gives a stochastic expansion of $\widehat{\boldsymbol{\eta}}$.

Following Ghosh and Maiti (2004), we define

$$\begin{aligned} & \boldsymbol{J}_r = \mathbf{Cov}\,(\boldsymbol{s}, \frac{\partial s_r}{\partial \boldsymbol{\eta}}), \quad \boldsymbol{K}_r = E\left(\frac{\partial^2 s_r}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathrm{T}}}\right), \quad r = 1, \ldots, p+1 \\ & \boldsymbol{a}^{\mathrm{T}} = [\mathrm{tr}\,(\boldsymbol{U}^{-1}\boldsymbol{J}_1), \ldots, \mathrm{tr}\,(\boldsymbol{U}^{-1}\boldsymbol{J}_{p+1})], \quad \boldsymbol{b}^{\mathrm{T}} = [\mathrm{tr}\,(\boldsymbol{U}^{-1}\boldsymbol{K}_1), \ldots, \mathrm{tr}\,(\boldsymbol{U}^{-1}\boldsymbol{K}_{p+1})], \end{aligned}$$

where $\boldsymbol{s} = (s_1, \ldots, s_{p+1})^{\mathrm{T}}$ and the expressions of $\boldsymbol{J}_r$ and $\boldsymbol{K}_r$ are given by (5.37) and (5.38), respectively. Then the asymptotic bias of $\widehat{\boldsymbol{\eta}}$ is $E(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \boldsymbol{U}^{-1}(\boldsymbol{a} + \boldsymbol{b}/2) + o(m^{-1})$. Furthermore, we decompose $\boldsymbol{U}^{-1}$ as

$$\boldsymbol{U}^{-1} = \begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix}, \quad \boldsymbol{U}^{-1} = \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{U}_{12}^{\mathrm{T}} & U_{22} \end{bmatrix},$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are $(p, p+1)$ and $(1, p+1)$ matrices, $\boldsymbol{U}_{11}, \boldsymbol{U}_{12}$ are (p, p) and $(p, 1)$ matrices, respectively, and U_{22} is a scaler. Now we can evaluate B_2 as the following lemma.

Lemma 5.3 *Under the condition* (C1)*, B_2 in* (5.7) *is approximated as*

$$B_2(\boldsymbol{\eta}) = B_{21}(\boldsymbol{\eta}) + B_{22}(\boldsymbol{\eta}) + o(1),$$

when $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12)*. B_{21} and B_{22} are*

$$\begin{aligned} B_{21}(\boldsymbol{\eta}) = & \sum_{i=1}^{m} \frac{n_i \lambda^2}{(n_i + \lambda)^2} Q(m_i) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{U}_1 \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \begin{bmatrix} \xi_{2i} \\ \xi_{3i} - \phi_i Q(m_i) \xi_{1i} \end{bmatrix} \\ & - \sum_{i=1}^{m} \frac{n_i^2 \lambda}{(n_i + \lambda)^3} \boldsymbol{U}_2 \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \begin{bmatrix} \xi_{3i} \\ \xi_{4i} - \phi_i Q(m_i) \xi_{2i} \end{bmatrix}, \\ B_{22}(\boldsymbol{\eta}) = & \sum_{i=1}^{m} \frac{n_i \lambda^2}{(n_i + \lambda)^2} Q(m_i) \xi_{1i} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{U}_1 (\boldsymbol{a} + 2^{-1}\boldsymbol{b}) - \sum_{i=1}^{m} \frac{n_i^2 \lambda}{(n_i + \lambda)^3} \xi_{2i} \boldsymbol{U}_2 (\boldsymbol{a} + 2^{-1}\boldsymbol{b}), \end{aligned}$$

where $\xi_{ri} = E[h'(\widehat{\mu}_i^B)(y_i - m_i)^r]$ for $r = 1, \ldots, 4$ and their approximations in the case of $n_i = O(m^{1/2+\delta_i}), \delta_i > 0$ are shown in Section 5.5.5.

By the lemma above, we can estimate B_2 by

$$\widehat{B_2} = B_{21}(\widehat{\boldsymbol{\eta}}) + B_{22}(\widehat{\boldsymbol{\eta}}), \tag{5.13}$$

which is second-order unbiased.

Next, B_3 can be evaluated as follows.

Lemma 5.4 *Under the condition* (C1)*, B_3 in* (5.7) *is approximated as*

$$B_3(\boldsymbol{\eta}) = B_{31}(\boldsymbol{\eta}) + 2B_{32}(\boldsymbol{\eta}) + o(1),$$

when $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12)*. B_{31} and B_{32} are*

$$B_{31}(\boldsymbol{\eta}) = \sum_{i=1}^{m} \frac{n_i\lambda^2}{(n_i+\lambda)^2} Q(m_i)Q'(m_i)\xi_{1i}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_{11}\boldsymbol{x}_i,$$

$$B_{32}(\boldsymbol{\eta}) = \sum_{i=1}^{m} \frac{n_i^2\lambda}{(n_i+\lambda)^3} Q(m_i)\xi_{1i}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_{12},$$

where $\xi_{1i} = E[h'(\widehat{\mu}_i^B)(y_i - m_i)]$ and its approximation in the case of $n_i = O(m^{1/2+\delta_i}), \delta_i > 0$ are shown in Section 5.5.5.

By the lemma above, we can estimate B_3 by

$$\widehat{B_3} = B_{31}(\widehat{\boldsymbol{\eta}}) + 2B_{32}(\widehat{\boldsymbol{\eta}}), \tag{5.14}$$

which is second-order unbiased.

Using $\widehat{B_1}$, $\widehat{B_2}$ and $\widehat{B_3}$ given by (5.11), (5.13) and (5.14), we propose the cAIC in nonlinear mixed model as follows:

$$\mathrm{cAIC} = -2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}} y_i - \psi(\hat{\theta}_i^{\mathrm{EB}})) + 2(\widehat{B_1} + \widehat{B_2} + 2^{-1}\widehat{B_3}). \tag{5.15}$$

Theorem 5.1 *Under the condition* (C1)*, the* cAIC *in* (5.15) *is a second-order unbiased estimator of* cAI *in* (5.5)*, namely*

$$E(\mathrm{cAIC}) = \mathrm{cAI} + o(1),$$

when $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12)*.*

5.3.3 Method for constant n_i by using numerical integration and differentiation

In the last subsection, we consider the condition of large n_i to derive the criterion. However, this condition is not appropriate for many cases in real data. The most typical example is small area estimation, where the number of observations in each area is not very large. Then, in this and the next subsection, we propose alternative methods of estimating the penalty term, which does not need the assumption of large n_i.

The first method is based on stochastic expansion of $\widehat{\boldsymbol{\eta}}$, which is the same as the method of the previous subsection. However we cannot approximate $h(\widehat{\mu}_i^{\mathrm{B}})$ by polynomials of $\widehat{\mu}_i^{\mathrm{B}}$ for the case of $n_i = O(1)$. Then, we alternatively use numerical integration to evaluate the expectation of complex functions of random variables. We firstly estimate B_1. By the law of iterated expectations and the fact that $y_i - \widehat{\mu}_i^{\mathrm{B}} = \lambda(y_i - m_i)/(n_i + \lambda)$, B_1 can be rewritten as

$$B_1(\boldsymbol{\eta}) = \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E[h(\widehat{\mu}_i^{\mathrm{B}})(y_i - m_i)].$$

Because it is hard to obtain closed form expression of $B_1(\boldsymbol{\eta})$, we propose to calculate $B_1(\widehat{\boldsymbol{\eta}})$ by Monte Carlo simulation, or which can be seen as parametric bootstrap method. Bootstrap sample $\boldsymbol{y}^* = (y_1^*, \ldots, y_m^*)^{\mathrm{T}}$ is generated by

$$y_i^* \sim f(y_i^*|\theta_i^*), \quad \text{and} \quad \theta_i^* \sim p(\theta_i^*|\hat{\lambda}, \hat{m}_i). \tag{5.16}$$

Then $B_1(\widehat{\boldsymbol{\eta}})$ can be written as

$$B_1(\widehat{\boldsymbol{\eta}}) = \sum_{i=1}^{m} \frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}} E_*[h(\widehat{\mu}_i^{\mathrm{B}*})(y_i^* - \hat{m}_i) \mid \boldsymbol{y}], \tag{5.17}$$

where $\widehat{\mu}_i^{\mathrm{B}*} = (n_i y_i^* + \hat{\lambda}\hat{m}_i)/(n_i+\hat{\lambda})$ and E_* denotes expectation with respect to the distribution of $\boldsymbol{y}^*$ given $\boldsymbol{y}$. Monte Carlo approximation of (5.17) is

$$B_1(\widehat{\boldsymbol{\eta}}) \approx \widetilde{B_1}^* = B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m} \frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}} h(\widehat{\mu}_i^{\mathrm{B}*}(b))(y_i^*(b) - \hat{m}_i), \tag{5.18}$$

for large B, where $y_i^*(b)$ is the bth bootstrap sample based on (5.16) and $\widehat{\mu}_i^{\mathrm{B}*}(b)$ is the version of $\widehat{\mu}_i^{\mathrm{B}*}$ based on the bootstrap sample.

Because $B_1(\widehat{\boldsymbol{\eta}})$ has second-order bias, we consider bias correction as follows:

$$B_1(\widehat{\boldsymbol{\eta}}) - B_{14}(\widehat{\boldsymbol{\eta}}) - B_{15}(\widehat{\boldsymbol{\eta}}),$$

where

$$B_{14}(\boldsymbol{\eta}) = \frac{\partial B_1(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\boldsymbol{U}(\boldsymbol{\eta})^{-1}(\boldsymbol{a}(\boldsymbol{\eta}) + \boldsymbol{b}(\boldsymbol{\eta})/2), \quad \text{and} \quad B_{15}(\boldsymbol{\eta}) = \frac{1}{2}\mathrm{tr}\left[\frac{\partial^2 B_1(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\boldsymbol{U}(\boldsymbol{\eta})^{-1}\right],$$

noting that $E(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \boldsymbol{U}^{-1}(\boldsymbol{a}+\boldsymbol{b}/2) + o(m^{-1})$ and $E[(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}] = \boldsymbol{U}^{-1} + o(m^{-1})$ when $\widehat{\boldsymbol{\eta}}$ is estimated by the estimating equation (5.12). The values of the first- and second-derivative of $B_1(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$ is calculated by numerical differentiation method. We propose the following procedure:

$$\left.\frac{\partial B_1(\boldsymbol{\eta})}{\partial \eta_k}\right|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} \approx \frac{\widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} + \varepsilon\boldsymbol{e}_k) - \widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} - \varepsilon\boldsymbol{e}_k)}{2\varepsilon}, \quad (k = 1, \ldots, p+1),$$

and

$$\left.\frac{\partial^2 B_1(\boldsymbol{\eta})}{\partial\eta_k\partial\eta_l}\right|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}}$$
$$\approx \begin{cases} \widetilde{B_1}^{*(kk)} = \{\widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} + \varepsilon\boldsymbol{e}_k) + \widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} - \varepsilon\boldsymbol{e}_k) - 2\widetilde{B_1}^*(\widehat{\boldsymbol{\eta}})\}/\varepsilon^2 \quad (\text{if } k = l), \\ \widetilde{B_1}^{*(kl)} = \big[\{\widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} + \varepsilon(\boldsymbol{e}_k + \boldsymbol{e}_l)) + \widetilde{B_1}^*(\widehat{\boldsymbol{\eta}} - \varepsilon(\boldsymbol{e}_k + \boldsymbol{e}_l)) - 2\widetilde{B_1}^*(\widehat{\boldsymbol{\eta}})\} \\ \qquad\qquad -\varepsilon^2\{\widetilde{B_1}^{*(kk)} + \widetilde{B_1}^{*(ll)}\}\big]/(2\varepsilon^2) \quad (\text{if } k \neq l), \end{cases}$$

for small positive ε, where $\boldsymbol{e}_k$ is the $(p+1)\times 1$ vector, with the kth component equal to 1 and the other components equal to 0. Using these numerical integration and differentiation, we can approximate $B_{14}(\widehat{\boldsymbol{\eta}})$ and $B_{15}(\widehat{\boldsymbol{\eta}})$, which we call $\widehat{B_{14}}^*$ and $\widehat{B_{15}}^*$. Then we can obtain an estimator of B_1 as follows:

$$\widehat{B_1}^* = \widetilde{B_1}^* - \widehat{B_{14}}^* - \widehat{B_{15}}^*. \tag{5.19}$$

Next we estimate B_2 and B_3. By Lemma 5.3, we can approximate B_2 as

$$B_2(\boldsymbol{\eta}) = B_{21}(\boldsymbol{\eta}) + B_{22}(\boldsymbol{\eta}) + o(1),$$

which is also valid under the condition (C2). In the previous subsection, $\xi_{ri} = E[h'(\widehat{\mu}_i^{\mathrm{B}})(y_i - m_i)^r]$ $(r = 1, \ldots, 4)$ in B_{21} and B_{22} are approximated by Taylor series expansion of $h(\cdot)$ for large n_i.

However, we now consider the case of $n_i = O(1)$. Then we calculate $\xi_{ri}(\widehat{\boldsymbol{\eta}})$ by Monte Carlo approximation based on bootstrap samples (5.16) as follows:

$$\xi_{ri}(\widehat{\boldsymbol{\eta}}) \approx B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m} h'(\widehat{\mu}_i^{\mathrm{B}*}(b))(y_i^*(b) - \hat{m}_i)^r.$$

Calculating $\xi_{ri}(\widehat{\boldsymbol{\eta}})$'s, we can obtain $B_{21}(\widehat{\boldsymbol{\eta}})$ and $B_{22}(\widehat{\boldsymbol{\eta}})$ based on Monte Carlo approximation, which we call $\widehat{B_{21}}^*$ and $\widehat{B_{22}}^*$, and the following estimator of B_2:

$$\widehat{B_2}^* = \widehat{B_{21}}^* + \widehat{B_{22}}^*. \tag{5.20}$$

As for B_3, Lemma 5.4 gave an asymptotic approximation. However, this approximation is based on the condition (C1), we have to modify the evaluation. Then we give the following lemma.

Lemma 5.5 *Under the condition* (C2)*, B_3 in* (5.7) *is approximated as*

$$B_3(\boldsymbol{\eta}) = B_{33}(\boldsymbol{\eta}) + 2B_{34}(\boldsymbol{\eta}) + B_{35}(\boldsymbol{\eta}) + o(1),$$

when $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12)*. $B_{33}(\boldsymbol{\eta})$, $B_{34}(\boldsymbol{\eta})$ and $B_{35}(\boldsymbol{\eta})$ are*

$$\begin{aligned}
B_{33}(\boldsymbol{\eta}) &= B_{31}(\boldsymbol{\eta}) + \sum_{i=1}^{m}\frac{n_i\lambda^3}{(n_i+\lambda)^3}\{Q(m_i)\}^2\nu_{1i}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_{11}\boldsymbol{x}_i,\\
B_{34}(\boldsymbol{\eta}) &= B_{32}(\boldsymbol{\eta}) - \sum_{i=1}^{m}\frac{n_i^2\lambda^2}{(n_i+\lambda)^4}Q(m_i)\nu_{2i}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_{12},\\
B_{35}(\boldsymbol{\eta}) &= \sum_{i=1}^{m}\left\{\frac{2n_i^2\lambda}{(n_i+\lambda)^4}\xi_{2i} + \frac{n_i^3\lambda}{(n_i+\lambda)^5}\nu_{3i}\right\}\times U_{22},
\end{aligned}$$

where $\nu_{ri} = E[h''(\widehat{\mu}_i^{\mathrm{B}})(y_i - m_i)^r]$ for $r = 1, \ldots, 3$.

We calculate $\nu_{ri}(\widehat{\boldsymbol{\eta}})$ in the lemma above by Monte Carlo approximation based on bootstrap samples (5.16) as follows:

$$\nu_{ri}(\widehat{\boldsymbol{\eta}}) \approx B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m} h''(\widehat{\mu}_i^{\mathrm{B}*}(b))(y_i^*(b) - \hat{m}_i)^r$$

Calculating $\nu_{ri}(\widehat{\boldsymbol{\eta}})$'s, we can obtain $B_{33}(\widehat{\boldsymbol{\eta}})$, $B_{34}(\widehat{\boldsymbol{\eta}})$ and $B_{35}(\widehat{\boldsymbol{\eta}})$, which we call $\widehat{B_{33}}^*$, $\widehat{B_{34}}^*$ and $\widehat{B_{35}}^*$, and the following estimator of B_3:

$$\widehat{B_3}^* = \widehat{B_{33}}^* + 2\widehat{B_{34}}^* + \widehat{B_{35}}. \tag{5.21}$$

Using $\widehat{B_1}^*$, $\widehat{B_2}^*$ and $\widehat{B_3}^*$ given by (5.19), (5.20) and (5.21), we propose the following cAIC* using numerical integration and differentiation:

$$\mathrm{cAIC}^* = -2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}y_i - \psi(\hat{\theta}_i^{\mathrm{EB}})) + 2(\widehat{B_1}^* + \widehat{B_2}^* + 2^{-1}\widehat{B_3}^*). \tag{5.22}$$

Theorem 5.2 *Under the condition* (C2)*, the* cAIC* *in* (5.22) *is second-order asymptotically unbiased estimator of* cAI *in* (5.5)*, namely*

$$E(\mathrm{cAIC}^*) = \mathrm{cAI} + o(1),$$

when $\boldsymbol{\eta}$ is estimated by the estimating equation (5.12)*.*

5.3.4 Numerical method for constant n_i based on parametric bootstrap

The second method to estimate the penalty term without the assumption of large n_i is based on parametric bootstrap. This method does not need stochastic expansion of $\widehat{\boldsymbol{\eta}}$.

We firstly estimate B_1. The plug-in estimator $B_1(\widehat{\boldsymbol{\eta}})$ can be calculated by Monte Carlo approximation $\widetilde{B_1}^*$ given by (5.18). However, this naive estimator has second-order bias. When it is hard to obtain analytical form of asymptotic bias and variance of $\widehat{\boldsymbol{\eta}}$, we cannot correct the bias of $B_1(\widehat{\boldsymbol{\eta}})$ analytically. Then we propose numerical bias correction based on parametric bootstrap method as follows:

$$2B_1(\widehat{\boldsymbol{\eta}}) - E_*[B_1(\widehat{\boldsymbol{\eta}}^*) \mid \boldsymbol{y}].$$

The second term of the equation above can be written as

$$E_*[B_1(\widehat{\boldsymbol{\eta}}^*) \mid \boldsymbol{y}] = \sum_{i=1}^{m} E_*\left[\frac{n_i\hat{\lambda}^*}{n_i+\hat{\lambda}^*} \cdot E_{**}\left[h(\widehat{\mu}_i^{\mathrm{B}**})(y_i^{**} - \hat{m}_i^*) \mid (\boldsymbol{y}^*, \boldsymbol{y})\right] \mid \boldsymbol{y}\right], \tag{5.23}$$

where the distribution of y_i^{**} is

$$y_i^{**} \sim f(y_i^{**}|\theta_i^{**}), \quad \text{and} \quad \theta_i^{**} \sim p(\theta_i^{**}|\hat{\lambda}^*, \hat{m}_i^*), \tag{5.24}$$

$\widehat{\mu}_i^{\mathrm{B}**} = (n_i y_i^{**} + \hat{\lambda}^* \hat{m}_i^*)/(n_i + \hat{\lambda}^*)$, and E_{**} denotes expectation with respect to the distribution of $\boldsymbol{y}^{**} = (y_1^{**}, \ldots, y_m^{**})^{\mathrm{T}}$ given $\boldsymbol{y}^*$ and $\boldsymbol{y}$. Monte Carlo approximation of (5.23) is

$$E_*[B_1(\widehat{\boldsymbol{\eta}}^*) \mid \boldsymbol{y}] \approx \widetilde{B_1}^{**} = (BC)^{-1} \sum_{b=1}^{B}\sum_{c=1}^{C}\sum_{i=1}^{m} \frac{n_i\hat{\lambda}^*(b)}{n_i+\hat{\lambda}^*(b)} h\{\widehat{\mu}_i^{\mathrm{B}**}(c(b))\}\{y_i^{**}(c(b)) - \hat{m}_i^*(b)\},$$

for large B and C, where $y_i^{**}(c(b))$ is the cth double bootstrap sample from the bth bootstrap sample based on (5.24), and $\widehat{\mu}_i^{\mathrm{B}**}(c(b))$ is the version of $\widehat{\mu}_i^{\mathrm{B}**}$ based on the bootstrap sample. Then, we propose the following bias corrected estimator of B_1:

$$\widehat{B_1}^{**} = 2\widetilde{B_1}^* - \widetilde{B_1}^{**}. \tag{5.25}$$

We next estimate B_2 and B_3. It is noted that $B_2(\boldsymbol{\eta})$ and $B_3(\boldsymbol{\eta})$ can be rewritten as

$$\begin{aligned}
B_2(\boldsymbol{\eta}) &= \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E\left[(y_i - m_i)\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right], \\
B_3(\boldsymbol{\eta}) &= \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E\left[(y_i - m_i)(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right],
\end{aligned}$$

by the law of iterated expectations and the fact that $y_i - \widehat{\mu}_i^{\mathrm{B}} = \lambda(y_i - m_i)/(n_i + \lambda)$. Then $B_2(\widehat{\boldsymbol{\eta}})$ and $B_3(\widehat{\boldsymbol{\eta}})$ can be written as

$$\begin{aligned}
B_2(\widehat{\boldsymbol{\eta}}) &= \sum_{i=1}^{m} \frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}} E_*\left[(y_i^* - \hat{m}_i)\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial \widehat{\boldsymbol{\eta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}}) \mid \boldsymbol{y}\right], \\
B_3(\widehat{\boldsymbol{\eta}}) &= \sum_{i=1}^{m} \frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}} E_*\left[(y_i^* - \hat{m}_i)(\widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}})^{\mathrm{T}}\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial \widehat{\boldsymbol{\eta}}\partial \widehat{\boldsymbol{\eta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}}) \mid \boldsymbol{y}\right],
\end{aligned} \tag{5.26}$$

where $\widehat{\boldsymbol{\eta}}^*$ is an estimator of $\boldsymbol{\eta}$ based on bootstrap sample $\boldsymbol{y}^*$. Note that the exact expressions of $\partial h(\widehat{\mu}_i^{\mathrm{B}*})/\partial \widehat{\boldsymbol{\eta}}^{\mathrm{T}}$ and $\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})/\partial \widehat{\boldsymbol{\eta}}\partial \widehat{\boldsymbol{\eta}}^{\mathrm{T}}$ are

$$\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial \widehat{\boldsymbol{\eta}}} = \begin{bmatrix} h'(\widehat{\mu}_i^{\mathrm{B}*})\hat{\lambda}(n_i+\hat{\lambda})^{-1}Q(\hat{m}_i)\boldsymbol{x}_i \\ -h'(\widehat{\mu}_i^{\mathrm{B}*})n_i(n_i+\hat{\lambda})^{-2}(y_i^* - \hat{m}_i) \end{bmatrix}$$

and

$$\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial\widehat{\boldsymbol{\eta}}\partial\widehat{\boldsymbol{\eta}}^{\mathrm{T}}} = \begin{bmatrix} \partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})/(\partial\widehat{\boldsymbol{\beta}}\partial\widehat{\boldsymbol{\beta}}^{\mathrm{T}}) & \partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})/(\partial\widehat{\boldsymbol{\beta}}\partial\hat{\lambda}) \\ \partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})/(\partial\hat{\lambda}\partial\widehat{\boldsymbol{\beta}}^{\mathrm{T}}) & \partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})/(\partial\hat{\lambda})^2, \end{bmatrix}$$

where

$$\begin{aligned}
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial\widehat{\boldsymbol{\beta}}\partial\widehat{\boldsymbol{\beta}}^{\mathrm{T}}} &= \frac{\hat{\lambda}}{n_i+\hat{\lambda}} Q(\hat{m}_i)\left\{ h''(\widehat{\mu}_i^{\mathrm{B}*})\frac{\hat{\lambda}}{n_i+\hat{\lambda}} Q(\hat{m}_i) + h'(\widehat{\mu}_i^{\mathrm{B}*})Q'(\hat{m}_i)\right\}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \\
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial\widehat{\boldsymbol{\beta}}\partial\hat{\lambda}} &= \frac{n_i}{(n_i+\hat{\lambda})^2} Q(\hat{m}_i)\left\{ -h''(\widehat{\mu}_i^{\mathrm{B}*})\frac{\hat{\lambda}}{n_i+\hat{\lambda}}(y_i^*-\hat{m}_i) + h'(\widehat{\mu}_i^{\mathrm{B}*})\right\}\boldsymbol{x}_i, \\
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*})}{\partial\hat{\lambda}^2} &= \frac{n_i}{(n_i+\hat{\lambda})^3}(y_i^*-\hat{m}_i)\left\{ h''(\widehat{\mu}_i^{\mathrm{B}*})\frac{n_i}{n_i+\hat{\lambda}}(y_i^*-\hat{m}_i) + 2h'(\widehat{\mu}_i^{\mathrm{B}*})\right\}.
\end{aligned}$$

Monte Carlo approximation of (5.26) is

$$\begin{aligned}
B_2(\widehat{\boldsymbol{\eta}}) \approx \widehat{B_2}^{**} &= B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m}\frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}}(y_i^*(b)-\hat{m}_i)\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}*}(b))}{\partial\widehat{\boldsymbol{\eta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}^*(b)-\widehat{\boldsymbol{\eta}}), \\
B_3(\widehat{\boldsymbol{\eta}}) \approx \widehat{B_3}^{**} &= B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m}\frac{n_i\hat{\lambda}}{n_i+\hat{\lambda}}(y_i^*(b)-\hat{m}_i)(\widehat{\boldsymbol{\eta}}^*(b)-\widehat{\boldsymbol{\eta}})^{\mathrm{T}}\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}*}(b))}{\partial\widehat{\boldsymbol{\eta}}\partial\widehat{\boldsymbol{\eta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}^*(b)-\widehat{\boldsymbol{\eta}}),
\end{aligned} \tag{5.27}$$

for large B, where $y_i^*(b)$ is the bth bootstrap sample based on (5.16) and $\widehat{\mu}_i^{\mathrm{B}*}(b)$ and $\widehat{\boldsymbol{\eta}}^*(b)$ are the versions of $\widehat{\mu}_i^{\mathrm{B}*}$ and $\widehat{\boldsymbol{\eta}}^*$ based on the bootstrap sample. Because $B_2(\boldsymbol{\eta})$ and $B_3(\boldsymbol{\eta})$ are of order $O(1)$, $\widehat{B_2}^{**}$ and $\widehat{B_3}^{**}$ are asymptotically unbiased estimators of B_2 and B_3 whose biases are of order $o(1)$.

Using $\widehat{B_1}^{**}$, $\widehat{B_2}^{**}$ and $\widehat{B_3}^{**}$ given by (5.25) and (5.27), we propose the following cAIC** based on parametric bootstrap:

$$\mathrm{cAIC}^{**} = -2\sum_{i=1}^{m} n_i(\hat{\theta}_i^{\mathrm{EB}}y_i - \psi(\hat{\theta}_i^{\mathrm{EB}})) + 2(\widehat{B_1}^{**} + \widehat{B_2}^{**} + 2^{-1}\widehat{B_3}^{**}). \tag{5.28}$$

Theorem 5.3 *Under the condition* (C2)*, the* cAIC** *in* (5.28) *is a second-order asymptotically unbiased estimator of* cAI *in* (5.5)*, namely*

$$E(\mathrm{cAIC}^{**}) = \mathrm{cAI} + o(1).$$

Though the cAIC** does not depend on the method of estimating $\boldsymbol{\eta}$, this requires parametric bootstrap, which is computationally harder than cAIC and cAIC*. Especially, estimation of B_1 needs double bootstrap method. However, because each step in estimation of B_1 does not include maximization algorithm, namely B_1 does not require estimating the hyperparameter, the computational load of estimating B_1 is the same level as estimating B_2 and B_3, which only requires single bootstrap.

5.4 Simulations

In this section, we compare the numerical performance between the conventional mAIC and the proposed criteria. We handle Poisson-gamma model introduced in Section 5.2. We consider a class of the nested candidate models $j_\alpha = \{1,\ldots,\alpha\}$ for $\alpha = 1,\ldots,p_\omega$. Let $\boldsymbol{x}_i(\omega)$'s be

independently generated as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = 0.9\boldsymbol{I}_{p_\omega} + 0.1\boldsymbol{J}_{p_\omega}$ for $\boldsymbol{J}_{p_\omega} = \mathbf{1}_{p_\omega}\mathbf{1}_{p_\omega}^{\mathrm{T}}$. The true vector of coefficients $\boldsymbol{\beta}_*$ is given by $\boldsymbol{\beta}_* = (1, 1, 1, 0, 0)^{\mathrm{T}}$, namely $p_\omega = 5$ and $p_* = 3$. Scale parameters are given by $\lambda = 4$ and $n_i = 11$ for $i = 1, \ldots, m$, where the sample size m varies with simulations.

We compare the performance between the mAIC

$$\mathrm{mAIC} = -2\sum_{i=1}^{m} \log\{m(y_i|\hat{\lambda}, \hat{m}_i)\} + 2(p_j + 1),$$

where $m(y_i|\lambda, m_i)$ is given in (5.4), and the proposed criteria, based on 500 replications. We handle the cases of $m = 100$ and $m = 70$. Note that the case of $m = 100$ satisfies the condition (C1) but the case of $m = 70$ does not.

Table 5.1: The number of selecting each model based on 500 replications and prediction error of the best model selected by the criteria

	the number of selection					prediction error
	j_1	j_2	j_3	j_4	j_5	
			$m = 100$			
mAIC	91	102	211	58	38	7.42962
cAIC	59	98	223	72	48	7.42248
			$m = 70$			
mAIC	112	114	190	59	25	5.30810
cAIC	97	163	145	62	33	5.30644

Table 5.1 reports the number of selecting each model and prediction error of the best model selected by the mAIC and the cAIC in (5.15). The prediction error is measured by quadratic loss

$$\sum_{i=1}^{m} (\widehat{\mu}_i^{\mathrm{EB}} - \mu_i)^2,$$

and the values are given as the averages based on 500 replications. For the case of $m = 100$, where the condition (C1) is satisfied, the performance of the cAIC is better than that of the mAIC in terms of both selecting the true model and prediction error. For the case of $m = 70$, where the condition (C1) is not satisfied, the performance of the cAIC is comparable to that of the mAIC in terms of the prediction error, but is not in terms of selecting the true model. Although we conducted simulations which investigate cAIC* in (5.22) and cAIC** in (5.28), the performance of the criteria was numerically instable. This is the future work to be resolved.

5.5 Some results of analytical calculations

5.5.1 Stochastic Expansion of $\widehat{\boldsymbol{\eta}}$

We give a stochastic expansion of $\widehat{\boldsymbol{\eta}}$ and derive the asymptotic bias and variance of $\widehat{\boldsymbol{\eta}}$. Let $\boldsymbol{t} = (t_1, \ldots, t_{p+1})^{\mathrm{T}}$ where

$$t_r = (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}} \frac{\partial^2 s_r(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^{\mathrm{T}}} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \quad r = 1, \ldots, p + 1.$$

It follows from the Taylor series expansion of $\boldsymbol{s}(\widehat{\boldsymbol{\eta}})$ around $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$ that

$$\mathbf{0} = \boldsymbol{s}(\boldsymbol{\eta}) + \frac{\partial \boldsymbol{s}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \frac{1}{2}\boldsymbol{t} + o_p(1),$$

thus we can get the expression

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} = \left(-\frac{\partial \boldsymbol{s}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right)^{-1} \left\{\boldsymbol{s}(\boldsymbol{\eta}) + \frac{1}{2}\boldsymbol{t} + o_p(1)\right\}. \tag{5.29}$$

It follows that

$$\begin{aligned} -\frac{\partial \boldsymbol{s}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}} =& \sum_{i=1}^{m} \left(-\frac{\partial \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right) (\boldsymbol{I}_{p+1} \otimes \boldsymbol{g}_i) + \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \left(-\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right) \\ & (\equiv \boldsymbol{V}_1 + \boldsymbol{V}_2), \end{aligned}$$

and $E(\boldsymbol{V}_1) = \mathbf{0}, \boldsymbol{V}_1 = O_p(m^{1/2}), E(\boldsymbol{V}_2) = \boldsymbol{U} = O(m)$. Let $\boldsymbol{W} = \boldsymbol{V}_1 + \boldsymbol{V}_2 - \boldsymbol{U}$, then noting that $\boldsymbol{V}_2 - \boldsymbol{U} = O_p(m^{1/2}), \boldsymbol{W} = O_p(m^{1/2})$, we can expand $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}$ in (5.29) as

$$\begin{aligned} \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} &= (\boldsymbol{U} + \boldsymbol{W})^{-1} \left\{\boldsymbol{s} + \frac{1}{2}\boldsymbol{t} + o_p(1)\right\} \\ &= \left\{\boldsymbol{U}^{-1} - \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{U}^{-1} + o_p(m^{-3/2})\right\} \left\{\boldsymbol{s} + \frac{1}{2}\boldsymbol{t} + o_p(1)\right\} \\ &= \boldsymbol{U}^{-1}\boldsymbol{s} - \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s} + \frac{1}{2}\boldsymbol{U}^{-1}\boldsymbol{t} + o_p(m^{-1}) \\ &\equiv \widehat{\boldsymbol{\eta}}^{(1)} + \widehat{\boldsymbol{\eta}}^{(2)} + \widehat{\boldsymbol{\eta}}^{(3)} + o_p(m^{-1}), \end{aligned} \tag{5.30}$$

where $\widehat{\boldsymbol{\eta}}^{(1)} = O_p(m^{-1/2}), \widehat{\boldsymbol{\eta}}^{(2)} = O_p(m^{-1}), \widehat{\boldsymbol{\eta}}^{(3)} = O_p(m^{-1})$. Now we evaluate each term in (5.30).

First we evaluate the first and the second moment of $\widehat{\boldsymbol{\eta}}^{(1)} = \boldsymbol{U}^{-1} \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{g}_i$. It is easy to see that $E(\widehat{\boldsymbol{\eta}}^{(1)}) = \mathbf{0}$. The second moment is

$$\begin{aligned} E[\widehat{\boldsymbol{\eta}}^{(1)}(\widehat{\boldsymbol{\eta}}^{(1)})^{\mathrm{T}}] &= \sum_{i=1}^{m} \boldsymbol{U}^{-1} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} E[\boldsymbol{g}_i \boldsymbol{g}_i^{\mathrm{T}}] \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i \boldsymbol{U}^{-1} \\ &= \sum_{i=1}^{m} \boldsymbol{U}^{-1} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i \boldsymbol{U}^{-1} \\ &= \boldsymbol{U}^{-1}, \end{aligned}$$

thus it follows that

$$E[(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}] = \boldsymbol{U}^{-1} + o(m^{-1}), \tag{5.31}$$

where $\boldsymbol{U}^{-1}$ is the asymptotic variance of $\widehat{\boldsymbol{\eta}}$.

Next we see $\widehat{\boldsymbol{\eta}}^{(2)}$. Let $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{p+1})^{\mathrm{T}}$, then $\widehat{\boldsymbol{\eta}}^{(2)}$ can be rewritten as

$$\begin{aligned} \widehat{\boldsymbol{\eta}}^{(2)} =& \boldsymbol{U}^{-1}(-\boldsymbol{W})\boldsymbol{U}^{-1}\boldsymbol{s} \\ =& \boldsymbol{U}^{-1} \left[\mathrm{tr}\,[\boldsymbol{U}^{-1}\boldsymbol{s}(-\boldsymbol{w}_1^{\mathrm{T}})], \ldots, \mathrm{tr}\,[\boldsymbol{U}^{-1}\boldsymbol{s}(-\boldsymbol{w}_{p+1}^{\mathrm{T}})]\right]^{\mathrm{T}}. \end{aligned}$$

It is noted that $E[\boldsymbol{s}(-\boldsymbol{w}_r^{\mathrm{T}})] = \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{w}_r) = \mathbf{Cov}\,(\boldsymbol{s}, \partial s_r/\partial\boldsymbol{\eta}) = \boldsymbol{J}_r$, then

$$E(\widehat{\boldsymbol{\eta}}^{(2)}) = \boldsymbol{U}^{-1}\boldsymbol{a}. \tag{5.32}$$

Finally we evaluate $\widehat{\boldsymbol{\eta}}^{(3)} = 2^{-1}\boldsymbol{U}^{-1}\boldsymbol{t}$. It can be seen that

$$\begin{aligned} t_r &= (\widehat{\boldsymbol{\eta}}^{(1)})^{\mathrm{T}} \frac{\partial^2 s_r}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}} \widehat{\boldsymbol{\eta}}^{(1)} + o_p(1) \\ &= \mathrm{tr}\left[\frac{\partial^2 s_r}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}} \boldsymbol{U}^{-1}\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{U}^{-1}\right] + o_p(1), \end{aligned}$$

Because $E(\boldsymbol{s}) = \mathbf{0}$ and y_i's are mutually independent, it follows that

$$\begin{aligned} E(t_r) &= \mathrm{tr}\left[E\left(\frac{\partial^2 s_r}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\right)\boldsymbol{U}^{-1}E(\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}})\boldsymbol{U}^{-1}\right] + o(1) \\ &= \mathrm{tr}\,[\boldsymbol{K}_r\boldsymbol{U}^{-1}] + o(1). \end{aligned}$$

Thus we can get

$$E(\widehat{\boldsymbol{\eta}}^{(3)}) = 2^{-1}\boldsymbol{U}^{-1}\boldsymbol{b} + o(m^{-1}). \tag{5.33}$$

From (5.32) and (5.33), the bias of $\widehat{\boldsymbol{\eta}}$ is

$$E(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \boldsymbol{U}^{-1}(\boldsymbol{a} + 2^{-1}\boldsymbol{b}) + o(m^{-1}). \tag{5.34}$$

5.5.2 Expressions of $\boldsymbol{J}_r$ and $\boldsymbol{K}_r$

First we get the expression of $\boldsymbol{J}_r$. Let $\boldsymbol{V}_1 = [\boldsymbol{v}_{11}, \ldots, \boldsymbol{v}_{1,p+1}]^{\mathrm{T}}$, $\boldsymbol{V}_2 = [\boldsymbol{v}_{21}, \ldots, \boldsymbol{v}_{2,p+1}]^{\mathrm{T}}$. It follows from $\boldsymbol{W} = \boldsymbol{V}_1 + \boldsymbol{V}_2 - \boldsymbol{U}$ that

$$\boldsymbol{J}_r = \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{w}_r) = \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{1r}) + \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{2r}).$$

Let $\boldsymbol{D}_i^{\mathrm{T}} = [\boldsymbol{d}_{i1}, \ldots, \boldsymbol{d}_{i,p+1}]^{\mathrm{T}}$. Then $\mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{1r})$ can be written as

$$\begin{aligned} \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{1r}) &= E\left[\sum_{i,j=1}^{m} \boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{g}_i \left(\frac{\partial \boldsymbol{d}_{jr}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}}{\partial\boldsymbol{\eta}^{\mathrm{T}}}\right)(\boldsymbol{I}_{p+1} \otimes \boldsymbol{g}_j)\right] \\ &= \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}E(\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}) \left(\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}}{\partial\boldsymbol{\eta}}\right)^{\mathrm{T}} \\ &= \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \left\{\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial\boldsymbol{\eta}}\boldsymbol{\Sigma}_i^{-1} + (\boldsymbol{I}_{p+1} \otimes \boldsymbol{d}_{ir}^{\mathrm{T}})\frac{\partial\boldsymbol{\Sigma}_i^{-1}}{\partial\boldsymbol{\eta}}\right\}^{\mathrm{T}}. \end{aligned} \tag{5.35}$$

$\mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{2r})$ is

$$\begin{aligned} \mathbf{Cov}\,(\boldsymbol{s}, -\boldsymbol{v}_{2r}) &= E\left[\sum_{i,j=1}^{m} \boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{g}_i\boldsymbol{d}_{jr}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\frac{\partial\boldsymbol{g}_j}{\partial\boldsymbol{\eta}^{\mathrm{T}}}\right] \\ &= E\left[\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{g}_i\boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{g}_i}{\partial\boldsymbol{\eta}^{\mathrm{T}}}\right]. \end{aligned}$$

Here $\partial \boldsymbol{g}_i/\partial \boldsymbol{\eta}^{\mathrm{T}}$ is expressed as

$$\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{\eta}^{\mathrm{T}}} = -2Q(m_i)(y_i - m_i)\begin{bmatrix} \mathbf{0} & 0 \\ \boldsymbol{x}_i^{\mathrm{T}} & 0 \end{bmatrix} - Q(m_i)\begin{bmatrix} \boldsymbol{x}_i^{\mathrm{T}} & 0 \\ \phi_i Q'(m_i)\boldsymbol{x}_i^{\mathrm{T}} & -(1+v_2/n_i)(\lambda - v_2)^{-2} \end{bmatrix},$$

thus it follows that

$$\mathbf{Cov}\left(\boldsymbol{s}, -\boldsymbol{v}_{2r}\right) = -2\sum_{i=1}^{m} Q(m_i)\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\mu_{2i}\\ \mu_{3i}\end{bmatrix}\boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix} \mathbf{0} & 0 \\ \boldsymbol{x}_i^{\mathrm{T}} & 0 \end{bmatrix}. \tag{5.36}$$

From (5.35) and (5.36),

$$\boldsymbol{J}_r = \sum_{i=1}^{m}\boldsymbol{D}_i^{\mathrm{T}}\left\{\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial \boldsymbol{\eta}}\boldsymbol{\Sigma}_i^{-1} + (\boldsymbol{I}_{p+1}\otimes \boldsymbol{d}_{ir}^{\mathrm{T}})\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \boldsymbol{\eta}}\right\}^{\mathrm{T}} - 2\sum_{i=1}^{m} Q(m_i)\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\mu_{2i}\\ \mu_{3i}\end{bmatrix}\boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix} \mathbf{0} & 0 \\ \boldsymbol{x}_i^{\mathrm{T}} & 0 \end{bmatrix}, \tag{5.37}$$

where the expressions of $\partial \boldsymbol{d}_{ir}^{\mathrm{T}}/\partial \boldsymbol{\eta}$ and $\partial \boldsymbol{\Sigma}_i^{-1}/\partial \boldsymbol{\eta}$ are given in Section 5.5.3 and 5.5.4.

Next we get the expression of $\boldsymbol{K}_r = E[(\partial^2 s_r)/(\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}})]$. It can be seen that

$$\begin{aligned}\boldsymbol{K}_r &= \sum_{i=1}^{m}\left\{2\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial \boldsymbol{\eta}}\boldsymbol{\Sigma}_i^{-1}E\left(\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right) + 2(\boldsymbol{I}_{p+1}\otimes \boldsymbol{d}_{ir}^{\mathrm{T}})\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \boldsymbol{\eta}}E\left(\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right) + \{\boldsymbol{I}_{p+1}\otimes(\boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1})\}E\left(\frac{\partial^2 \boldsymbol{g}_i}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}}\right)\right\}\\ &= \sum_{i=1}^{m}\left\{-2\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial \boldsymbol{\eta}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{D}_i - 2(\boldsymbol{I}_{p+1}\otimes \boldsymbol{d}_{ir}^{\mathrm{T}})\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \boldsymbol{\eta}}\boldsymbol{D}_i + \{\boldsymbol{I}_{p+1}\otimes(\boldsymbol{d}_{ir}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1})\}E\left(\frac{\partial^2 \boldsymbol{g}_i}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}}\right)\right\},\end{aligned} \tag{5.38}$$

where the expressions of $\partial \boldsymbol{d}_{ir}^{\mathrm{T}}/\partial \boldsymbol{\eta}$ and $\partial \boldsymbol{\Sigma}_i^{-1}/\partial \boldsymbol{\eta}$ are given in Section 5.5.3 and 5.5.4, and $E[(\partial^2 \boldsymbol{g}_i)/(\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}})] = [\boldsymbol{G}_1^{\mathrm{T}}, \ldots, \boldsymbol{G}_{p+1}^{\mathrm{T}}]^{\mathrm{T}}$ is written as

$$\begin{aligned}&[\boldsymbol{G}_1^{\mathrm{T}}, \ldots, \boldsymbol{G}_p^{\mathrm{T}}]^{\mathrm{T}}\\ &= -Q(m_i)\boldsymbol{x}_i \otimes \begin{bmatrix} Q'(m_i)\boldsymbol{x}_i^{\mathrm{T}} & 0 \\ \{-2Q(m_i) + (Q'(m_i))^2 + 2v_2\phi_i Q(m_i)\}\boldsymbol{x}_i^{\mathrm{T}} & -Q'(m_i)(1+v_2/n_i)(\lambda - v_2)^{-2} \end{bmatrix},\\ &\boldsymbol{G}_{p+1}\\ &= -Q(m_i)\begin{bmatrix} \mathbf{0} & 0 \\ -(1+v_2/n_i)(\lambda - v_2)^{-2}Q'(m_i)\boldsymbol{x}_i^{\mathrm{T}} & 2(1+v_2/n_i)(\lambda - v_2)^{-3} \end{bmatrix}.\end{aligned}$$

5.5.3 Expression of $\partial \boldsymbol{d}_{ir}^{\mathrm{T}}/\partial \boldsymbol{\eta}$

First, we give the expression of $\partial \boldsymbol{d}_{ir}^{\mathrm{T}}/\partial \boldsymbol{\eta}$ for $r = 1, \ldots, p$. It follows that

$$\boldsymbol{d}_{ir}^{\mathrm{T}} = Q(m_i)\left[x_{ir}, Q'(m_i)\phi_i x_{ir}\right],$$

then we get

$$\begin{aligned}\frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial \boldsymbol{\beta}} &= Q(m_i)x_{ir}\left[Q'(m_i), \left\{(Q'(m_i))^2 + 2v_2Q(m_i)\right\}\phi_i\right]\otimes \boldsymbol{x}_i,\\ \frac{\partial \boldsymbol{d}_{ir}^{\mathrm{T}}}{\partial \lambda} &= Q(m_i)Q'(m_i)x_{ir}\left[0, -(1+v_2/n_i)(\lambda - v_2)^{-2}\right].\end{aligned}$$

Next, we give the expression of $\partial \boldsymbol{d}_{i,p+1}^{\mathrm{T}}/\partial \boldsymbol{\eta}$. It follows that

$$\boldsymbol{d}_{i,p+1}^{\mathrm{T}} = Q(m_i)\left[0, -(1+v_2/n_i)(\lambda - v_2)^{-2}\right],$$

then we can get

$$\begin{aligned}\frac{\partial \boldsymbol{d}_{i,p+1}^{\mathrm{T}}}{\partial \boldsymbol{\beta}} &= -Q'(m_i)Q(m_i)(1+v_2/n_i)(\lambda-v_2)^{-2}[\mathbf{0};\boldsymbol{x}_i],\\ \frac{\partial \boldsymbol{d}_{i,p+1}^{\mathrm{T}}}{\partial \lambda} &= \left[0, 2Q(m_i)(1+v_2/n_i)(\lambda-v_2)^{-3}\right].\end{aligned}$$

5.5.4 Expression of $\partial\boldsymbol{\Sigma}_i^{-1}/\partial\boldsymbol{\eta}$

It follows that

$$\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \boldsymbol{\eta}} = -(\boldsymbol{I}_{p+1}\otimes\boldsymbol{\Sigma}_i^{-1})\frac{\partial \boldsymbol{\Sigma}_i}{\partial \boldsymbol{\eta}}\boldsymbol{\Sigma}_i^{-1},$$

thus it suffices to evaluate $\partial\boldsymbol{\Sigma}_i/\partial\boldsymbol{\eta}$.

Let $\partial\boldsymbol{\Sigma}_i/\partial\boldsymbol{\eta} = [\boldsymbol{H}_1^{\mathrm{T}},\ldots,\boldsymbol{H}_{p+1}^{\mathrm{T}}]^{\mathrm{T}}$. For $r=1,\ldots,p$,

$$\boldsymbol{H}_r = \begin{bmatrix} \partial\mu_{2i}/\partial\beta_r & \partial\mu_{3i}/\partial\beta_r \\ \partial\mu_{3i}/\partial\beta_r & \partial\mu_{4i}/\partial\beta_r - \partial\mu_{2i}^2/\partial\beta_r \end{bmatrix},$$

where

$$\begin{aligned}\frac{\partial\mu_{2i}}{\partial\beta_r} &= \phi_i Q(m_i)Q'(m_i)x_{ir},\\ \frac{\partial\mu_{2i}^2}{\partial\beta_r} &= 2\phi_i^2\left\{Q(m_i)\right\}^2 Q'(m_i)x_{ir},\\ \frac{\partial\mu_{3i}}{\partial\beta_r} &= \frac{(\lambda/n_i+1)(\lambda/n_i+2)}{(\lambda-v_2)(\lambda-2v_2)}\left[\left\{Q'(m_i)\right\}^2 + 2v_2Q(m_i)\right]Q(m_i)x_{ir}.\end{aligned}$$

Let $d_i = v_2/n_i$, then

$$\begin{aligned}\frac{\partial\mu_{4i}}{\partial\beta_r} = &\left\{\frac{6(d_i+1)(2d_i+1)(3d_i+1)}{(\lambda-v_2)(\lambda-2v_2)(\lambda-3v_2)} + \frac{12(d_i+1)(2d_i+1)}{n_i(\lambda-v_2)(\lambda-2v_2)} + \frac{7(d_i+1)}{n_i^2(\lambda-v_2)} + \frac{1}{n_i^3}\right\}J_1,\\ &+\left\{\frac{3(d_i+1)(2d_i+1)(3d_i+1)}{(\lambda-v_2)(\lambda-3v_2)} + \frac{2(d_i+1)(4d_i+3)}{n_i(\lambda-v_2)} + \frac{2d_i+3}{n_i^2}\right\}J_2,\end{aligned}$$

where

$$\begin{aligned}J_1 &= \left[\left\{Q'(m_i)\right\}^2 + 4v_2Q(m_i)\right]Q(m_i)Q'(m_i)x_{ir},\\ J_2 &= 2\left\{Q(m_i)\right\}^2 Q'(m_i)x_{ir}.\end{aligned}$$

It can be seen that

$$\boldsymbol{H}_{p+1} = \begin{bmatrix} \partial\mu_{2i}/\partial\lambda & \partial\mu_{3i}/\partial\lambda \\ \partial\mu_{3i}/\partial\lambda & \partial\mu_{4i}/\partial\lambda - \partial\mu_{2i}^2/\partial\lambda \end{bmatrix},$$

where

$$\begin{aligned}
\frac{\partial \mu_{2i}}{\partial \lambda} =& -Q(m_i)\frac{v_2/n_i+1}{(\lambda-v_2)^2},\\
\frac{\partial \mu_{2i}^2}{\partial \lambda} =& -2\left\{Q(m_i)\right\}^2\frac{(\lambda/n_i+1)(v_2/n_i+1)}{(\lambda-v_2)^3},\\
\frac{\partial \mu_{3i}}{\partial \lambda} =& Q(m_i)Q'(m_i)\frac{(2\lambda+3n_i)(\lambda-v_2)(\lambda-2v_2)-(\lambda+n_i)(\lambda+2n_i)(2\lambda-3v_2)}{n_i^2(\lambda-v_2)^2(\lambda-2v_2)^2},\\
\frac{\partial \mu_{4i}}{\partial \lambda} =& 6(d_i+1)(2d_i+1)(3d_i+1)J_3+\frac{12}{n_i}(d_i+1)(2d_i+1)Q(m_i)\left\{Q'(m_i)\right\}^2J_4\\
&-\left\{\frac{2(d_i+1)(4d_i+3)}{n_i}\left\{Q(m_i)\right\}^2+\frac{7(d_i+1)}{n_i^2}Q(m_i)\{Q'(m_i)\}^2\right\}\frac{1}{(\lambda-v_2)^2},
\end{aligned}$$

for

$$\begin{aligned}
J_3 =& -\left\{Q(m_i)\right\}^2\frac{\lambda-2v_2}{(\lambda-v_2)^2(\lambda-3v_2)^2}-Q(m_i)\left\{Q'(m_i)\right\}^2\frac{3\lambda^2-12v_2\lambda+11v_2^2}{(\lambda-v_2)^2(\lambda-2v_2)^2(\lambda-3v_2)^2},\\
J_4 =& -\frac{2\lambda-3v_2}{(\lambda-v_2)^2(\lambda-2v_2)^2}.
\end{aligned}$$

5.5.5 Approximations of ξ_{ri}

It suffices to evaluate ξ_{ri} up to $O(1)$, namely the remainder is of order $O(n_i^{-1})$. It is noted that $\xi_{ri}=E[h'(\widehat{\mu}_i^B)(y_i-m_i)^r]=E[h'(\mu_i)(y_i-m_i)^r]+O(n_i^{-1})$. After some calculations, we can get ξ_{ri}. For the Poisson-gamma mixture model,

$$\begin{aligned}
\xi_{1i} =& -\frac{1}{\lambda m_i-1}+O(n_i^{-1}), \quad \xi_{2i}=\frac{m_i}{\lambda m_i-1}+O(n_i^{-1}),\\
\xi_{3i} =& -\frac{m_i}{\lambda(\lambda m_i-1)}+O(n_i^{-1}), \quad \xi_{4i}=\frac{3\lambda m_i^2-2m_i}{\lambda^2(\lambda m_i-1)}+O(n_i^{-1}),
\end{aligned}$$

for the binomial-beta mixture model,

$$\begin{aligned}
\xi_{1i} =& -\frac{(\lambda-1)(1-2m_i)}{(\lambda m_i-1)\{\lambda(1-m_i)-1\}}+O(n_i^{-1}),\\
\xi_{2i} =& \frac{(\lambda-2)m_i(1-m_i)}{(\lambda m_i-1)\{\lambda(1-m_i)-1\}}+O(n_i^{-1}),\\
\xi_{3i} =& -\frac{m_i(1-m_i)(1-2m_i)}{(\lambda m_i-1)\{\lambda(1-m_i)-1\}}+O(n_i^{-1}),\\
\xi_{4i} =& \frac{3\lambda m_i^2(1-m_i)^2-2m_i(1-m_i)(m_i^2-m_i+1)}{(\lambda m_i-1)\{\lambda(1-m_i)-1\}}+O(n_i^{-1}).
\end{aligned}$$

5.6 Proofs

5.6.1 Proof of Lemma 5.3

First, B_2 can be rewritten as

$$\begin{aligned} B_2 &= \sum_{i=1}^{m} n_i E\left[E\left[(y_i-\mu_i)\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}) \mid \boldsymbol{y}\right]\right] \\ &= \sum_{i=1}^{m} n_i E\left[(y_i-\widehat{\mu}_i^{\mathrm{B}})\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta})\right] \\ &= \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E\left[\boldsymbol{f}_i^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta})\right], \end{aligned} \tag{5.39}$$

where

$$\boldsymbol{f}_i = (y_i-m_i)\frac{\partial h(\widehat{\mu}_i^{\mathrm{B}})}{\partial \boldsymbol{\eta}} = \left[\frac{\lambda Q(m_i)h'(\widehat{\mu}_i^{\mathrm{B}})(y_i-m_i)}{n_i+\lambda}\boldsymbol{x}_i^{\mathrm{T}}, -\frac{n_i h'(\widehat{\mu}_i^{\mathrm{B}})(y_i-m_i)^2}{(n_i+\lambda)^2}\right]^{\mathrm{T}}.$$

Using the stochastic expansion of $\widehat{\boldsymbol{\eta}}$ in Section 5.5.1, we decompose B_2 as

$$\begin{aligned} B_2 &= \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E(\boldsymbol{f}_i^{\mathrm{T}}\widehat{\boldsymbol{\eta}}^{(1)}) + \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda} E[\boldsymbol{f}_i^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}^{(2)}+\widehat{\boldsymbol{\eta}}^{(3)})] + o(1) \\ &\equiv I_1 + I_2 + o(1). \end{aligned}$$

$E(\boldsymbol{f}_i^{\mathrm{T}}\widehat{\boldsymbol{\eta}}^{(1)})$ is evaluated as

$$\begin{aligned} E(\boldsymbol{f}_i^{\mathrm{T}}\widehat{\boldsymbol{\eta}}^{(1)}) &= E\left[\left[\frac{\lambda Q(m_i)h'(\widehat{\mu}_i^{\mathrm{B}})(y_i-m_i)}{n_i+\lambda}\boldsymbol{x}_i^{\mathrm{T}}, -\frac{n_i h'(\widehat{\mu}_i^{\mathrm{B}})(y_i-m_i)^2}{(n_i+\lambda)^2}\right]\boldsymbol{U}^{-1}\sum_{j=1}^{m}\boldsymbol{D}_j^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{g}_j\right] \\ &= \frac{\lambda}{n_i+\lambda}Q(m_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_1\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\xi_{2i}\\ \xi_{3i}-\phi_i Q(m_i)\xi_{1i}\end{bmatrix} \\ &\quad - \frac{n_i}{(n_i+\lambda)^2}\boldsymbol{U}_2\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\begin{bmatrix}\xi_{3i}\\ \xi_{4i}-\phi_i Q(m_i)\xi_{2i}\end{bmatrix}, \end{aligned}$$

which yields $I_1 = B_{21}$. From (5.32) and (5.33), $E[\boldsymbol{f}_i^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}^{(2)}+\widehat{\boldsymbol{\eta}}^{(3)})]$ is evaluated as

$$\begin{aligned} E[\boldsymbol{f}_i^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}^{(2)}+\widehat{\boldsymbol{\eta}}^{(3)})] &= E(\boldsymbol{f}_i^{\mathrm{T}})E(\widehat{\boldsymbol{\eta}}^{(2)}+\widehat{\boldsymbol{\eta}}^{(3)}) + o(m^{-1}) \\ &= \left[\frac{\lambda}{n_i+\lambda}Q(m_i)\xi_{1i}\boldsymbol{x}_i^{\mathrm{T}}, -\frac{n_i}{(n_i+\lambda)^2}\xi_{2i}\right]\boldsymbol{U}^{-1}(\boldsymbol{a}+2^{-1}\boldsymbol{b}) + o(m^{-1}) \\ &= \frac{\lambda}{n_i+\lambda}Q(m_i)\xi_{1i}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{U}_1(\boldsymbol{a}+2^{-1}\boldsymbol{b}) - \frac{n_i}{(n_i+\lambda)^2}\xi_{2i}\boldsymbol{U}_2(\boldsymbol{a}+2^{-1}\boldsymbol{b}) + o(m^{-1}), \end{aligned}$$

which yields $I_2 = B_{22} + o(1)$. □

5.6.2 Proof of Lemmas 5.4 and 5.5

In the same manner as (5.39), B_3 can be rewritten as

$$B_3 = \sum_{i=1}^{m} \frac{n_i\lambda}{n_i+\lambda}\mathrm{tr}\left[E\left[\boldsymbol{F}_i(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta})(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta})^{\mathrm{T}}\right]\right],$$

where

$$\boldsymbol{F}_i = (y_i - m_i)\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}} = (y_i - m_i)\begin{bmatrix} \partial^2 h(\widehat{\mu}_i^{\mathrm{B}})/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}) & \partial^2 h(\widehat{\mu}_i^{\mathrm{B}})/(\partial\boldsymbol{\beta}\partial\lambda) \\ \partial^2 h(\widehat{\mu}_i^{\mathrm{B}})/(\partial\lambda\partial\boldsymbol{\beta}^{\mathrm{T}}) & \partial^2 h(\widehat{\mu}_i^{\mathrm{B}})/(\partial\lambda)^2 \end{bmatrix},$$

for

$$\begin{aligned}
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \frac{\lambda}{n_i+\lambda}Q(m_i)Q'(m_i)h'(\widehat{\mu}_i^{B})\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} + \left(\frac{\lambda}{n_i+\lambda}\right)^2\{Q(m_i)\}^2 h''(\widehat{\mu}_i^{\mathrm{B}})\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \\
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial\boldsymbol{\beta}\partial\lambda} &= \frac{n_i}{(n_i+\lambda)^2}Q(m_i)h'(\widehat{\mu}_i^{\mathrm{B}})\boldsymbol{x}_i - \frac{n_i\lambda}{(n_i+\lambda)^3}Q(m_i)h''(\widehat{\mu}_i^{\mathrm{B}})\boldsymbol{x}_i, \\
\frac{\partial^2 h(\widehat{\mu}_i^{\mathrm{B}})}{\partial\lambda\partial\lambda} &= \frac{2n_i}{(n_i+\lambda)^3}h'(\widehat{\mu}_i^{\mathrm{B}})(y_i - m_i) + \frac{n_i^2}{(n_i+\lambda)^4}h''(\widehat{\mu}_i^{\mathrm{B}})(y_i - m_i)^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
E\left[\boldsymbol{F}_i(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}\right] &= E(\boldsymbol{F}_i)E[\widehat{\boldsymbol{\eta}}^{(1)}(\widehat{\boldsymbol{\eta}}^{(1)})^{\mathrm{T}}] + o(m^{-1}) \\
&= E(\boldsymbol{F}_i)\boldsymbol{U}^{-1} + o(m^{-1}),
\end{aligned}$$

which yields $B_3 = B_{31} + 2B_{32} + o(1)$ under the condition (C1), and $B_3 = B_{33} + 2B_{34} + B_{35} + o(1)$ under the condition (C2), respectively. □

Chapter 6

A variant of AIC using Bayesian marginal likelihood

In this chapter, we propose an information criterion which measures the prediction risk of the predictive density based on the Bayesian marginal likelihood from a frequentist point of view. We derive the criteria for selecting variables in linear regression models by putting the prior on the regression coefficients, and discuss the relationship between the proposed criteria and other related ones. There are three advantages of our method. Firstly, this is a compromise between the frequentist and Bayesian standpoint because it evaluates the frequentist's risk of the Bayesian model. Thus it is less influenced by prior misspecification. Secondly, non-informative improper prior can be also used for constructing the criterion. When the uniform prior is assumed on the regression coefficients, the resulting criterion is identical to the residual information criterion (RIC) of Shi and Tsai (2002). Lastly, the criteria have the consistency property for selecting the true model.

6.1 Motivation

The problem of selecting appropriate models has been extensively studied in the literature since Akaike (1973, 1974), who derived so called the Akaike information criterion (AIC). Since the AIC and their variants are based on the risk of the predictive densities with respect to the Kullback–Leibler (KL) divergence, they can select a good model in the light of prediction. It is known, however, that the AIC-type criteria do not have the consistency property, namely, the probability that the criteria select the true model does not converges to 1. Another approach to model selection is Bayesian procedures such as Bayes factors and the Bayesian information criterion (BIC) suggested by Schwarz (1978), both of which are constructed based on the Bayesian marginal likelihood. Bayesian procedures for model selection have the consistency property in some specific models, while they do not select models in terms of prediction. In addition, it is known that Bayes factors do not work for improper prior distributions and that the BIC does not use any specific prior information. In this chapter, we provide a unified framework to derive an information criterion for model selection so that it can produce various information criteria including AIC, BIC and the residual information criterion (RIC) suggested by of Shi and Tsai (2002). Especially, we propose an intermediate criterion between AIC and BIC using the empirical Bayes method.

To explain the unified framework in the general setup, let $\boldsymbol{y}$ be an n-variate observable random vector whose density is $m(\boldsymbol{y}|\boldsymbol{\eta})$ for a vector of unknown parameters $\boldsymbol{\eta}$. Let $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})$ be

a predictive density for $m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta})$, where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$. We here evaluate the predictive performance of $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})$ in terms of the following risk:

$$R(\boldsymbol{\eta};\hat{m}) = \int \left[\int \log \left\{ \frac{m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta})}{\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})} \right\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] m(\boldsymbol{y}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{y}. \tag{6.1}$$

Since this is interpreted as a risk with respect to the KL divergence, we call it the KL risk. The spirit of AIC suggests that we can provide an information criterion for model selection as an (asymptotically) unbiased estimator of the information

$$\begin{aligned} I(\boldsymbol{\eta};\hat{m}) &= \iint -2\log\{\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta}) m(\boldsymbol{y}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y} \\ &= E_{\boldsymbol{\eta}}\left[-2\log\{\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})\} \right], \end{aligned} \tag{6.2}$$

which is a part of (6.1) (multiplied by 2), where $E_{\boldsymbol{\eta}}$ denotes the expectation with respect to the distribution of $m(\tilde{\boldsymbol{y}},\boldsymbol{y}|\boldsymbol{\eta}) = m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta}) m(\boldsymbol{y}|\boldsymbol{\eta})$. Let $\Delta = I(\boldsymbol{\eta};\hat{m}) - E_{\boldsymbol{\eta}}[-2\log\{\hat{m}(\boldsymbol{y};\boldsymbol{y})\}]$. Then, the AIC variant based on the predictor $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})$ is defined by

$$\mathrm{IC}(\hat{m}) = -2\log\{\hat{m}(\boldsymbol{y};\boldsymbol{y})\} + \widehat{\Delta},$$

where $\widehat{\Delta}$ is an (asymptotically) unbiased estimator of Δ.

It is interesting to point out that $\mathrm{IC}(\hat{m})$ produces AIC and BIC for specific predictors.

(AIC) Put $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y}) = m(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\eta}})$ for the maximum likelihood estimator $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$. Then, $\mathrm{IC}(m(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\eta}}))$ is the exact AIC or the corrected AIC suggested by Sugiura (1978) and Hurvich and Tsai (1989), which is approximated by AIC of Akaike (1973, 1974) as $-2\log\{m(\boldsymbol{y}|\widehat{\boldsymbol{\eta}})\} + 2\dim(\boldsymbol{\eta})$.

(BIC) Put $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y}) = m_{\pi_0}(\tilde{\boldsymbol{y}}) = \int m(\tilde{\boldsymbol{y}}|\boldsymbol{\eta})\pi_0(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}$ for a proper prior distribution $\pi_0(\boldsymbol{\eta})$. Since it can be easily seen that $I(\boldsymbol{\eta};m_{\pi_0}) = E_{\boldsymbol{\eta}}[-2\log\{m_{\pi_0}(\boldsymbol{y})\}]$, we have $\Delta = 0$ in this case, so that $\mathrm{IC}(m_{\pi_0}) = -2\log\{m_{\pi_0}(\boldsymbol{y})\}$, which is the Bayesian marginal likelihood. It is noted that $-2\log\{m_{\pi_0}(\boldsymbol{y})\}$ is approximated by $\mathrm{BIC} = -2\log\{m(\boldsymbol{y}|\widehat{\boldsymbol{\eta}})\} + \log(n)\cdot\dim(\boldsymbol{\eta})$.

The criterion $\mathrm{IC}(\hat{m})$ can produce not only the conventional information criteria AIC and BIC, but also various criteria between AIC and BIC. For example, it is supposed that $\boldsymbol{\eta}$ is divided as $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$ for a p-dimensional parameter vector of interest $\boldsymbol{\beta}$ and an r-dimensional nuisance parameter vector $\boldsymbol{\omega}$. We assume that $\boldsymbol{\beta}$ has a prior density $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\omega})$ with hyperparameter $\boldsymbol{\lambda}$. The model is described as

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta} &\sim m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\omega}), \\ \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\omega}), \end{aligned}$$

and $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$ are estimated by data. Inference based on such a model is called an empirical Bayes procedure. Put $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y}) = m_{\pi}(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\omega}}) = \int f(\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\widehat{\boldsymbol{\omega}})\pi(\boldsymbol{\beta}|\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\omega}})\mathrm{d}\boldsymbol{\beta}$ for some estimators $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\omega}}$. Then, the information in (6.2) is

$$I(\boldsymbol{\eta};m_{\pi}) = \iint -2\log\{m_{\pi}(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\omega}})\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\boldsymbol{\omega}) m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\omega}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y}, \tag{6.3}$$

and the resulting information criterion is

$$\mathrm{IC}(m_{\pi}) = -2\log\{m_{\pi}(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\omega}})\} + \widehat{\Delta}, \tag{6.4}$$

where $\widehat{\Delta}$ is an (asymptotically) unbiased estimator of $\Delta = I(\boldsymbol{\eta};m_{\pi}) - E_{\boldsymbol{\eta}}[-2\log\{m_{\pi}(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\omega}})\}]$.

There are three motivations to consider the information $I(\boldsymbol{\eta}; m_\pi)$ in (6.3) and the information criterion $\mathrm{IC}(m_\pi)$ in (6.4).

Firstly, it is noted that the Bayesian predictor $m_\pi(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\omega}})$ is evaluated by the risk $R(\boldsymbol{\eta}; m_\pi)$ in (6.1), which is based on a frequentist point of view. On the other hand, the Bayesian risk is

$$r(\boldsymbol{\psi}; \hat{m}) = \int R(\boldsymbol{\eta}; \hat{m}) \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\beta}, \tag{6.5}$$

which measures the prediction error of $\hat{m}(\tilde{\boldsymbol{y}}; \boldsymbol{y})$ under the assumption that the prior information is correct, where $\boldsymbol{\psi} = (\boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$. The resulting Bayesian criteria such as PIC (Kitagawa, 1997) and DIC (Spiegelhalter et al., 2002) are sensitive to the prior misspecification, since they depend on the prior information. Because $R(\boldsymbol{\eta}; m_\pi)$ can measure the prediction error of the Bayesian model from a standpoint of frequentists, however, the resulting criterion $\mathrm{IC}(m_\pi)$ is less influenced by the prior misspecification.

Secondly, we can construct the information criterion $\mathrm{IC}(m_\pi)$ when the prior distribution of $\boldsymbol{\beta}$ is improper, since the information $I(\boldsymbol{\eta}; m_\pi)$ in (6.3) can be defined formally for the corresponding improper marginal likelihood. Because the Bayesian risk $r(\boldsymbol{\psi}; m_\pi)$ does not exist for the improper prior, however, we cannot obtain the corresponding Bayesian criteria and cannot use the Bayesian risk. Objective Bayesians want to avoid informative prior and many non-informative priors are improper. The suggested criterion $\mathrm{IC}(m_\pi)$ can respond to such a request. For example, objective Bayesians assume the uniform improper prior on regression coefficients $\boldsymbol{\beta}$ in linear regression models. It is interesting to note that the resulting variable selection criterion (6.4) is identical to the residual information criterion (RIC) of Shi and Tsai (2002), which is shown in the next section.

Lastly, this criterion has the consistency property. We derive the criterion for the variable selection problem in general linear regression model and prove that the criterion selects the true model with probability tending to one. The BIC or marginal likelihood are known to have the consistency (Nishii, 1984), while most AIC-type criteria are not consistent. But AIC-type criteria have the property to choose a good model in the sense of minimizing the prediction error (Shibata, 1981; Shao, 1997). Our proposed criterion is consistent for selection of the parameters of interest $\boldsymbol{\beta}$ and selects a good model in the light of prediction based on the empirical Bayes model.

The rest of this chapter is organized as follows. In Section 6.2, we obtain the information criterion (6.4) in linear regression model with general covariance structure and compare it with other related criteria. In Section 6.3, we prove the consistency of the criteria. In Section 6.4, we investigate the performance of the criteria through simulations. Section 6.5 concludes the chapter with some discussions.

6.2 Proposed criteria

6.2.1 Variable selection criteria for linear regression model

We consider the linear regression model as the candidate model, which is given as

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j + \boldsymbol{u}_j, \tag{6.6}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of the response variables, $\boldsymbol{X}(j)$ is an $n \times p_j$ matrix of the explanatory variables, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of the regression coefficients, $\boldsymbol{u}_j$ is an $n \times 1$ vector

of the random errors, and j is the index set which denotes the candidate model. Throughout the chapter, we assume that $\boldsymbol{X}(j)$ has full column rank p_j. Here, the random error $\boldsymbol{u}_j$ has the distribution $\mathcal{N}_n(\boldsymbol{0}, \sigma_j^2\boldsymbol{\Sigma})$, where σ_j^2 is an unknown scalar and $\boldsymbol{\Sigma}$ is a known positive definite matrix. The linear regression model (6.6) includes the linear mixed model (3.1) in Chapter 3 as a special case.

We consider the problem of selecting the explanatory variables and assume that the true model is included by the candidate model, namely all the candidate model is overspecified. This is the common assumption to obtain the criterion. Under this assumption, the true mean of $\boldsymbol{y}$ can be written as

$$E(\boldsymbol{y}) = \boldsymbol{X}(j)\boldsymbol{\beta}_j^*,$$

where $\boldsymbol{\beta}_j^*$ is $p_j \times 1$ vector whose $p_j - p_*$ components are exactly 0 and the rest of components are not 0. We hereafter abbreviate the model index j for notational convenience. We also abbreviate $\boldsymbol{\beta}_j^*$ as $\boldsymbol{\beta}$ and write the true variance parameter as σ^2.

We shall construct the variable selection criteria for the regression model (6.6) which is of the form (6.4). We consider the following two situations.

[i] Normal prior for $\boldsymbol{\beta}$. We first assume the prior distribution of $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{W}),$$

where $\boldsymbol{W}$ is a $p \times p$ matrix suitably chosen with full rank. Examples of $\boldsymbol{W}$ are $\boldsymbol{W} = (\lambda\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ for $\lambda > 0$ when $\boldsymbol{\Sigma}$ is identity matrix, which is introduced by Zellner (1986), or more simply $\boldsymbol{W} = \lambda^{-1}\boldsymbol{I}_p$. Because the likelihood is $m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma})$, the marginal likelihood function is

$$\begin{aligned} m_\pi(\boldsymbol{y}|\sigma^2) &= \int m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\mathrm{d}\boldsymbol{\beta} \\ &= (2\pi\sigma^2)^{-n/2} \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot |\boldsymbol{W}|^{-1/2} \cdot |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1}|^{-1/2} \cdot \exp\left\{-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/(2\sigma^2)\right\}, \end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. Note that $\boldsymbol{A} = (\boldsymbol{\Sigma} + \boldsymbol{B})^{-1}$ for $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}$, namely $m_\pi(\boldsymbol{y}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma} + \boldsymbol{B}))$. Then we take the predictive density as $\hat{m}(\tilde{\boldsymbol{y}}; \boldsymbol{y}) = m_\pi(\tilde{\boldsymbol{y}}|\hat{\sigma}^2)$ and the information (6.3) can be written as

$$I_{\pi,1}(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}}\left[n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{y}}/\hat{\sigma}^2\right], \tag{6.7}$$

where $\hat{\sigma}^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/n$ for $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$ and $E_{\boldsymbol{\eta}}$ denotes the expectation with respect to the distribution of $m(\tilde{\boldsymbol{y}}, \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2)m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ for $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$. Note that $\boldsymbol{\beta}$ is the parameter of interest and σ^2 is the nuisance parameter, which corresponds to $\boldsymbol{\omega}$ in the previous section. Then we propose the information criterion.

Proposition 6.1 *The information $I_{\pi,1}(\boldsymbol{\eta})$ in (6.7) is unbiasedly estimated by the information criterion*

$$\mathrm{IC}_{\pi,1} = -2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} + \frac{2n}{n-p-2}, \tag{6.8}$$

where

$$-2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} = n\log\hat{\sigma}^2 + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2 + (\mathrm{const}),$$

namely, $E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,1}) = I_{\pi,1}(\boldsymbol{\eta})$.

If $n^{-1}\boldsymbol{W}^{1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{W}^{1/2}$ converges to a $p\times p$ positive definite matrix as $n\to\infty$, $\log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p|$ can be approximated to $p\log n$. In that case, $\mathrm{IC}_{\pi,1}$ is approximately expressed as

$$\mathrm{IC}^*_{\pi,1}=n\log\hat{\sigma}^2+\log|\boldsymbol{\Sigma}|+p\log n+2+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2,$$

when n is large.

Alternatively, the KL risk $r(\boldsymbol{\psi};\hat{m})$ in (6.5) can be also used for evaluating the risk of the predictive density $m_\pi(\tilde{\boldsymbol{y}}|\hat{\sigma}^2)$, since the prior distribution is proper. The resulting criterion is

$$\mathrm{IC}_{\pi,2}=n\log\hat{\sigma}^2+\log|\boldsymbol{\Sigma}|+p\log n+p, \tag{6.9}$$

which is an asymptotically unbiased estimator of $I_{\pi,2}(\sigma^2)=E_\pi[I_{\pi,1}(\boldsymbol{\eta})]$ up to constant where E_π denotes the expectation with respect to the prior distribution $\pi(\boldsymbol{\beta}|\sigma^2)$, namely $E_\pi E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,2})\to I_{\pi,2}(\sigma^2)+(\mathrm{const})$ as $n\to\infty$. It is interesting to point out that $\mathrm{IC}_{\pi,2}$ is analogous to the criterion proposed by Bozdogan (1987) known as the consistent AIC, who suggested to replace the penalty term $2p$ in the AIC with $p+p\log n$.

[ii] Uniform prior for $\boldsymbol{\beta}$. We next assume the uniform prior for $\boldsymbol{\beta}$, namely $\boldsymbol{\beta}\sim\mathrm{uniform}(\mathbb{R}^p)$. Though this is improper prior distribution, we can obtain the marginal likelihood function formally:

$$\begin{aligned}m_r(\boldsymbol{y}|\sigma^2)=&\int m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\mathrm{d}\boldsymbol{\beta}\\ =&(2\pi\sigma^2)^{-(n-p)/2}\cdot|\boldsymbol{\Sigma}|^{-1/2}\cdot|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2}\cdot\exp\left\{-\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\boldsymbol{y}/(2\sigma^2)\right\},\end{aligned}$$

which is known as the residual likelihood (Patterson and Thompson, 1971). Then we take the predictive density as $\hat{m}(\tilde{\boldsymbol{y}};\boldsymbol{y})=m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}^2)$ and the information (6.3) can be written as

$$I_r(\boldsymbol{\eta})=E_{\boldsymbol{\eta}}\left[(n-p)\log(2\pi\tilde{\sigma}^2)+\log|\boldsymbol{\Sigma}|+\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|+\tilde{\boldsymbol{y}}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\tilde{\boldsymbol{y}}/\tilde{\sigma}^2\right], \tag{6.10}$$

where $\tilde{\sigma}^2=\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\boldsymbol{y}/(n-p)$, which is the residual maximum likelihood (REML) estimator of σ^2 based on the residual likelihood $m_r(\boldsymbol{y}|\sigma^2)$. Then we propose the information criterion.

Proposition 6.2 *The information $I_r(\boldsymbol{\eta})$ in (6.10) is unbiasedly estimated by the infomation criterion*

$$\mathrm{IC}_r=-2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}^2)\}+\frac{2(n-p)}{n-p-2}, \tag{6.11}$$

where

$$-2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}^2)\}=(n-p)\log\tilde{\sigma}^2+\log|\boldsymbol{\Sigma}|+\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|+\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\boldsymbol{y}/\tilde{\sigma}^2+(\mathrm{const}),$$

namely, $E_{\boldsymbol{\eta}}(\mathrm{IC}_r)=I_r(\boldsymbol{\eta})$.

Note that $\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\boldsymbol{y}/\tilde{\sigma}^2=n-p$. If $n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ converges to $p\times p$ positive definite matrix as $n\to\infty$, $\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|$ can be approximated to $p\log n$. In that case, we can approximately write

$$\mathrm{IC}^*_r=(n-p)\log\tilde{\sigma}^2+\log|\boldsymbol{\Sigma}|+p\log n+\frac{(n-p)^2}{n-p-2}, \tag{6.12}$$

when n is large. It is important to note that IC^*_r is identical to the RIC proposed by Shi and Tsai (2002) up to constant. Since $(n-p)^2/(n-p-2)=(n+2)+\{4/(n-p-2)-p\}$ and $n+2$ is irrelevant to the model, we can subtract $n+2$ from IC^*_r in (6.12), which results in the RIC exactly. Note the criterion based on $m_r(\boldsymbol{y}|\sigma^2)$ and $r(\boldsymbol{\psi};m_r)$ cannot be constructed because the KL risk of it diverges to infinity.

6.2.2 Extension to the case of unknown covariance

In the derivation of the criteria, we have assumed that the scaled covariance matrix $\boldsymbol{\Sigma}$ of the error terms vector are known. However, it is often the case that $\boldsymbol{\Sigma}$ is unknown and is some function of the unknown parameter $\boldsymbol{\phi}$, namely $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\phi})$. In that case, $\boldsymbol{\Sigma}$ in each criterion is replaced with its plug-in estimator $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\phi}})$, where $\widehat{\boldsymbol{\phi}}$ is some consistent estimator of $\boldsymbol{\phi}$. This strategy is also used in many other studies, for example in Shi and Tsai (2002), who proposed the RIC. We suggest that the $\boldsymbol{\phi}$ is estimated based on the full model. The method to estimate the nuisance parameters by the full model is similar to the C_p criterion by Mallows (1973). The scaled covariance matrix $\boldsymbol{W}$ of the prior distribution of $\boldsymbol{\beta}$ is also assumed to be known. In practice, its structure should be specified and we have to estimate the parameters $\boldsymbol{\lambda}$ involved in $\boldsymbol{W}$ from the data. In the same manner as $\boldsymbol{\Sigma}$, $\boldsymbol{W}$ in each criterion is replaced with $\boldsymbol{W}(\widehat{\boldsymbol{\lambda}})$. We propose that $\boldsymbol{\lambda}$ is estimated based on each candidate model under consideration because the structure of $\boldsymbol{W}$ depends on the model.

We here give three examples for the regression model (6.6), a regression model with constant variance, a variance components model, and a regression model with ARMA errors, where the second and the third ones include the unknown parameter in the covariance matrix.

[1] regression model with constant variance. In the case where $\boldsymbol{\Sigma} = \boldsymbol{I}_n$, (6.6) represents a multiple regression model with constant variance. In this model, the scaled covariance matrix $\boldsymbol{\Sigma}$ does not contain any unknown parameters.

[2] variance components model. Consider a variance components model (Henderson, 1950) described by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_2\boldsymbol{b}_2 + \cdots + \boldsymbol{Z}_r\boldsymbol{b}_r + \boldsymbol{\varepsilon}, \tag{6.13}$$

where $\boldsymbol{Z}_i$ is an $n \times m_i$ matrix with $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}$, $\boldsymbol{b}_i$ is an $m_i \times 1$ random vector having the distribution $\mathcal{N}_{m_i}(\boldsymbol{0}, \theta_i\boldsymbol{I}_{m_i})$ for $i \geq 2$, $\boldsymbol{\varepsilon}$ is an $n \times 1$ random vector with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma}_0 + \theta_1\boldsymbol{\Sigma}_1)$ for known $n \times n$ matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, and $\boldsymbol{e}$, $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_r$ are mutually independently distributed. The nested error regression model (NERM) is a special case of variance components model given by

$$y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik}, \quad (i = 1, \ldots, m;\ k = 1, \ldots, n_i), \tag{6.14}$$

where b_i's and ε_{ik}'s are mutually independently distributed as $b_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$ and $n = \sum_{i=1}^m n_i$. Note that the NERM in (6.14) is given by $\theta_1 = \sigma^2$, $\theta_2 = \tau^2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_n$ and $\boldsymbol{Z}_2 = \mathrm{diag}\,(\boldsymbol{1}_{n_1}, \ldots, \boldsymbol{1}_{n_m})$, where $\boldsymbol{1}_l$ is the l-dimensional vector of ones, for variance components model (6.13). This model is often used for the clustered data and b_i can be seen as the random effect of the cluster (Battese et al., 1988). For such a model, when one is interested in the specific cluster or predicting the random effects, the conditional AIC proposed by Vaida and Blanchard (2005), which is based on the conditional likelihood given the random effects, is appropriate. However, when the aim of the analysis is focused on the population, the NERM can be seen as linear regression model and the random effects are involved in the error term, namely we can treat $\boldsymbol{u} = \boldsymbol{Z}_2\boldsymbol{b}_2 + \boldsymbol{\varepsilon}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\phi) = \phi\boldsymbol{\Sigma}_2 + \boldsymbol{I}_n$ for (6.6), where $\phi = \tau^2/\sigma^2$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{Z}_2\boldsymbol{Z}_2^{\mathrm{T}} = \mathrm{diag}\,(\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{J}_{n_m})$ for $\boldsymbol{J}_l = \boldsymbol{1}_l\boldsymbol{1}_l^{\mathrm{T}}$. In that case, our proposed variable selection procedure is valid.

[3] regression model with autoregressive moving average errors. Consider the regression model (6.6), assuming the random errors are generated by an ARMA(q, r) process defined by

$$u_i - \phi_1 u_{i-1} - \cdots - \phi_q u_{i-q} = \varepsilon_i - \varphi_1\varepsilon_{i-1} - \cdots - \varphi_r\varepsilon_{i-r},$$

where $\{\varepsilon_i\}$ is a sequence of independent normal random variables having mean 0 and variance τ^2. A special case of this model is the regression model with AR(1) errors satisfying $u_1 \sim \mathcal{N}(0, \tau^2/(1-\phi^2))$, $u_i = \phi u_{i-1} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$ for $i = 2, 3, \ldots, n$. When we define $\sigma^2 = \tau^2/(1-\phi^2)$, (i,j)-element of the scaled covariance matrix $\boldsymbol{\Sigma}$ in (6.6) is $\phi^{|i-j|}$.

6.3 Consistency of the criteria

In this section, we prove that the proposed criteria have the consistency property. Our asymptotic framework is that n goes to infinity and the true dimension of the regression coefficients p is fixed. Following Shi and Tsai (2002), we first show the criteria are consistent for the regression model with constant variance and the prespecified $\boldsymbol{W}$, and then extend the result to the regression model with general covariance matrix and the case where $\boldsymbol{W}$ is estimated.

Let $\hat{j}$ denote the model selected by some criterion. Following Shi and Tsai (2002), we make the assumptions.

(A1) $E(u_1^4) < \infty$.

(A2) $0 < \liminf_{n\to\infty} \min_{j\in\mathcal{J}} |\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{X}(j)/n|$ and $\limsup_{n\to\infty} \max_{j\in\mathcal{J}} |\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{X}(j)/n| < \infty$.

(A3) $\liminf_{n\to\infty} n^{-1} \inf_{j\in\mathcal{J}_-} \|\boldsymbol{X}(\omega)\boldsymbol{\beta}_* - \boldsymbol{H}_j\boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2 > 0$, where $\boldsymbol{H}_j = \boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}$.

We can now obtain asymptotic properties of the criteria for the regression model with constant variance.

Theorem 6.1 *If assumptions* (A1)–(A3) *are satisfied, $\mathcal{J}_+$ is not empty, the u_i's are independent and identically distributed (iid) and $\boldsymbol{W}_j$ in the prior distribution of $\boldsymbol{\beta}_j$ is prespecified, then the criteria* $\mathrm{IC}_{\pi,1}$, $\mathrm{IC}^*_{\pi,1}$, $\mathrm{IC}_{\pi,2}$, IC_r *and* IC^*_r *are consistent, namely $P(\hat{j} = j_*) \to 1$ as $n \to \infty$.*

The proof of Theorem 6.1 is given in Section 6.7.

We next consider the regression model with a general covariance structure and the case where $\boldsymbol{W}_j$ is estimated by the data. In this case, $\boldsymbol{\Sigma}$ and $\boldsymbol{W}_j$ are replaced with their plug-in estimators $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\phi}})$ and $\boldsymbol{W}_j(\widehat{\boldsymbol{\lambda}}_j)$, respectively.

Theorem 6.2 *Assume that $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0$ and $\widehat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_{j,0}$ tend to 0 in probability as $n \to \infty$ for all $j \in \mathcal{J}$. In addition, assume that the elements of $\boldsymbol{\Sigma}(\boldsymbol{\phi})$ and $\boldsymbol{W}_j(\boldsymbol{\lambda}_j)$ are continuous functions of $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}_j$, and $\boldsymbol{\Sigma}(\boldsymbol{\phi})$ and $\boldsymbol{W}_j(\boldsymbol{\lambda}_j)$ is positive definite in the neighborhood of $\boldsymbol{\phi}_0$ and $\boldsymbol{\lambda}_{j,0}$ for all $j \in \mathcal{J}$. If assumptions* (A1)–(A3) *are satisfied when $\boldsymbol{X}(j)$ and $\boldsymbol{u}$ are replaced with $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(j)$ and $\boldsymbol{u}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}$ respectively, $\mathcal{J}_+$ is not empty and the u_i^*'s are iid, then the criteria* $\mathrm{IC}_{\pi,1}$, $\mathrm{IC}^*_{\pi,1}$, $\mathrm{IC}_{\pi,2}$, IC_r *and* IC^*_r *are consistent.*

For the proof of Theorem 6.2, we can use the same techniques as those for the proof of Theorem 6.1.

6.4 Simulations

In this section, we compare the numerical performance of the proposed criteria with some other conventional ones, which are AIC, BIC, the corrected AIC (AICC) by Sugiura (1978) and Hurvich and Tsai (1989). We shall consider the three regression models—regression model with constant variance, NERM, and regression model with AR(1) errors—which are taken as examples of (6.6) in Section 6.2.2. For the NERM, we consider the balanced sample case, namely

$n_1 = \cdots = n_m (= n_0)$. In each simulation, 1000 realizations are generated from (6.6) with $\boldsymbol{\beta} = (1, 1, 1, 1, 0, 0, 0)^{\mathrm{T}}$, namely the full model is seven-dimensional and the true model is four-dimensional. All explanatory variables are randomly generated from the standard normal distribution. The signal-to-noise ratio ($\mathrm{SNR} = \{\mathrm{var}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\mathrm{var}(u_i)\}^{1/2}$) is controlled at 1, 3, and 5. In the NERM, three cases of variance ratio $\phi = \tau^2/\sigma^2$ are considered with $\phi = 0.5,\ 1$ and 2. In the regression model with AR(1) errors, three correlation structures are considered with AR parameter $\phi = 0.1,\ 0.5$ and 0.8.

When deriving the criteria $\mathrm{IC}_{\pi,1}$, $\mathrm{IC}^*_{\pi,1}$ and $\mathrm{IC}_{\pi,2}$, we set the prior distribution of $\boldsymbol{\beta}$ as $\mathcal{N}_p(\boldsymbol{0}, \sigma^2\lambda^{-1}\boldsymbol{I}_p)$, namely $\boldsymbol{W} = \lambda^{-1}\boldsymbol{I}_p$. The hyperparameter λ is estimated by maximizing the marginal likelihood $m_\pi(\boldsymbol{y}|\hat{\sigma}^2)$, where the estimate $\hat{\sigma}^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/n$ of σ^2 is plugged in. The unknown parameter ϕ involved in $\boldsymbol{\Sigma}$ is estimated by some consistent estimator based on the full model. In the NERM, $\phi = \tau^2/\sigma^2$ is estimated by $\hat{\tau}^{2\mathrm{PR}}/\hat{\sigma}^{2\mathrm{PR}}$, where $\hat{\tau}^{2\mathrm{PR}}$ and $\hat{\sigma}^{2\mathrm{PR}}$ are unbiased estimators proposed by Prasad and Rao (1990). Let $S_0 = \boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\}\boldsymbol{y}$ and $S_1 = \boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{E}\}\boldsymbol{y}$ where $\boldsymbol{E} = \mathrm{diag}\,(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_m)$, $\boldsymbol{E}_i = \boldsymbol{I}_{n_0} - n_0^{-1}\boldsymbol{J}_{n_0}$ for $i = 1, \ldots, m$. Then, the Prasad–Rao estimators of σ^2 and τ^2 are

$$\hat{\sigma}^{2\mathrm{PR}} = S_1/(n - m - p), \quad \hat{\tau}^{2\mathrm{PR}} = \left\{S_0 - (n - p)\hat{\sigma}^{2\mathrm{PR}}\right\}/n^*$$

where $n^* = n - \mathrm{tr}\,[\boldsymbol{Z}_2^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}_2]$. In the regression model with AR(1) errors, the AR parameter ϕ is estimated by the maximum likelihood estimator based. Note that ϕ is estimated based on the full model and that σ^2 and λ is estimated based on each candidate model using the plug-in version of $\boldsymbol{\Sigma}(\hat{\phi})$.

The candidate models include all the subsets of the full model and select the model by the criteria. The performance of the criteria is measured by the number of selecting the true model and the prediction error of the selected model based on quadratic loss, namely $\|\boldsymbol{X}(\hat{\jmath})\widehat{\boldsymbol{\beta}}_{\hat{\jmath}} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2/n$.

Tables 6.1–6.3 give the number of selecting the true model by the criteria and the average prediction error of the selected model by each criterion is shown in Tables 6.4–6.6 for each of the regression models. From these tables, we can see the following three facts. Firstly, the number of selecting the true model approaches 1000 for all the proposed criteria, that is the numerical evidence of the consistency of the criteria. Though the BIC is also consistent, the small sample performance is not as good as our criteria. Secondly, the proposed criteria are not only consistent but also have smaller prediction error even when the sample size is small. Especially, $\mathrm{IC}_{\pi,1}$ is the best for the most of the experiments except when both the sample size and SNR are small. AIC and AICC have good performance in that situation in terms of prediction error. Thirdly, $\mathrm{IC}_{\pi,1}$ and IC_r have better performance than their approximation $\mathrm{IC}^*_{\pi,1}$ and IC^*_r, respectively, but the difference gets smaller as n becomes larger.

Table 6.1: The number of selecting the true model by the criteria in 1000 realizations of the regression model with constant variance.

		SNR = 1	SNR = 3	SNR = 5
$n = 20$	AIC	130	428	428
	BIC	118	587	588
	AICC	89	749	755
	$\mathrm{IC}_{\pi,1}$	115	843	905
	$\mathrm{IC}^*_{\pi,1}$	73	732	738
	$\mathrm{IC}_{\pi,2}$	73	731	737
	IC_r	143	797	882
	IC^*_r	147	828	898
$n = 40$	AIC	419	536	536
	BIC	424	800	800
	AICC	470	687	687
	$\mathrm{IC}_{\pi,1}$	472	900	938
	$\mathrm{IC}^*_{\pi,1}$	353	876	876
	$\mathrm{IC}_{\pi,2}$	352	876	876
	IC_r	462	895	934
	IC^*_r	478	899	941
$n = 80$	AIC	546	553	553
	BIC	827	872	872
	AICC	604	613	613
	$\mathrm{IC}_{\pi,1}$	750	934	968
	$\mathrm{IC}^*_{\pi,1}$	839	928	928
	$\mathrm{IC}_{\pi,2}$	838	928	928
	IC_r	722	937	968
	IC^*_r	739	936	969

Table 6.2: The number of selecting the true model by the criteria in 1000 realizations of the nested error regression model.

		$\phi = 0.5$			$\phi = 1$			$\phi = 2$		
	SNR	1	3	5	1	3	5	1	3	5
$n_0 = 5$	AIC	75	382	385	104	387	393	137	393	403
$m = 4$	BIC	60	546	564	90	539	574	134	556	586
	AICC	57	694	736	86	693	742	140	691	748
	$\mathrm{IC}_{\pi,1}$	71	821	902	119	829	915	181	835	941
	$\mathrm{IC}^*_{\pi,1}$	35	646	702	66	656	715	115	662	721
	$\mathrm{IC}_{\pi,2}$	34	642	701	70	653	715	116	657	720
	IC_r	244	789	888	310	818	900	432	838	924
	IC^*_r	78	723	912	106	723	922	149	698	941
$n_0 = 5$	AIC	220	458	458	235	465	465	259	473	473
$m = 8$	BIC	169	731	731	208	739	741	240	746	750
	AICC	219	607	607	251	612	612	284	625	627
	$\mathrm{IC}_{\pi,1}$	319	891	936	369	913	943	436	928	953
	$\mathrm{IC}^*_{\pi,1}$	107	838	839	151	836	841	199	843	853
	$\mathrm{IC}_{\pi,2}$	112	838	839	158	836	841	202	841	853
	IC_r	436	890	934	512	903	943	593	925	953
	IC^*_r	209	901	944	230	901	949	247	905	962
$n_0 = 5$	AIC	418	522	522	417	528	528	418	545	545
$m = 16$	BIC	394	853	853	407	859	859	416	866	866
	AICC	452	594	594	447	603	603	443	616	616
	$\mathrm{IC}_{\pi,1}$	622	926	955	618	941	959	624	951	968
	$\mathrm{IC}^*_{\pi,1}$	299	911	910	332	913	913	354	915	915
	$\mathrm{IC}_{\pi,2}$	300	910	910	331	913	913	358	915	915
	IC_r	691	925	954	695	942	960	709	953	968
	IC^*_r	443	932	959	438	946	964	421	956	972

Table 6.3: The number of selecting the true model by the criteria in 1000 realizations of the regression model with AR(1) errors.

		$\phi = 0.1$			$\phi = 0.5$			$\phi = 0.8$		
	SNR	1	3	5	1	3	5	1	3	5
$n = 20$	AIC	110	347	346	125	373	372	193	377	414
	BIC	91	482	482	118	523	542	209	502	587
	AICC	84	642	646	117	652	688	228	584	715
	$\mathrm{IC}_{\pi,1}$	101	741	834	138	774	862	274	742	901
	$\mathrm{IC}^*_{\pi,1}$	64	620	625	83	625	667	194	554	688
	$\mathrm{IC}_{\pi,2}$	63	618	624	88	621	667	197	553	685
	IC_r	123	698	801	224	769	846	562	808	901
	IC^*_r	122	702	790	144	715	826	241	646	825
$n = 40$	AIC	365	483	483	356	544	544	283	533	551
	BIC	369	756	756	334	791	793	286	737	797
	AICC	416	642	642	373	671	672	308	668	700
	$\mathrm{IC}_{\pi,1}$	422	877	917	450	909	949	427	901	976
	$\mathrm{IC}^*_{\pi,1}$	315	844	844	281	859	862	247	754	856
	$\mathrm{IC}_{\pi,2}$	314	844	844	282	858	862	255	752	854
	IC_r	430	866	917	507	903	945	685	918	974
	IC^*_r	412	865	918	377	881	932	316	785	932
$n = 80$	AIC	516	521	521	483	552	552	333	553	553
	BIC	789	851	851	598	865	865	334	859	868
	AICC	586	593	593	525	614	614	367	637	638
	$\mathrm{IC}_{\pi,1}$	738	926	949	691	936	962	550	961	979
	$\mathrm{IC}^*_{\pi,1}$	795	912	912	560	905	905	289	899	919
	$\mathrm{IC}_{\pi,2}$	792	912	912	559	905	905	296	889	919
	IC_r	713	921	951	724	938	961	692	966	980
	IC^*_r	714	920	951	600	911	952	382	908	958

Table 6.4: The prediction error of the best model selected by the criteria for the regression model with constant variance.

	SNR	1	3	5
$n = 20$	AIC	1.59	0.141	0.0504
	BIC	1.74	0.131	0.0468
	AICC	1.77	0.120	0.0421
	$\text{IC}_{\pi,1}$	1.70	0.111	0.0372
	$\text{IC}^*_{\pi,1}$	1.92	0.122	0.0429
	$\text{IC}_{\pi,2}$	1.92	0.122	0.0430
	IC_r	1.56	0.116	0.0383
	IC^*_r	1.57	0.114	0.0374
$n = 40$	AIC	0.708	0.0660	0.0238
	BIC	0.862	0.0568	0.0205
	AICC	0.732	0.0609	0.0219
	$\text{IC}_{\pi,1}$	0.754	0.0523	0.0180
	$\text{IC}^*_{\pi,1}$	1.05	0.0534	0.0192
	$\text{IC}_{\pi,2}$	1.05	0.0534	0.0192
	IC_r	0.716	0.0524	0.0181
	IC^*_r	0.718	0.0522	0.0179
$n = 80$	AIC	0.292	0.0321	0.0115
	BIC	0.265	0.0260	0.00936
	AICC	0.283	0.0310	0.0112
	$\text{IC}_{\pi,1}$	0.265	0.0244	0.00841
	$\text{IC}^*_{\pi,1}$	0.285	0.0245	0.00883
	$\text{IC}_{\pi,2}$	0.285	0.0245	0.00883
	IC_r	0.270	0.0243	0.00841
	IC^*_r	0.267	0.0243	0.00840

Table 6.5: The prediction error of the best model selected by the criteria for the nested error regression model.

		$\phi = 0.5$			$\phi = 1$			$\phi = 2$		
	SNR	1	3	5	1	3	5	1	3	5
$n_0 = 5$	AIC	1.80	0.150	0.0524	1.61	0.145	0.0494	1.44	0.140	0.0463
$m = 4$	BIC	1.96	0.159	0.0496	1.76	0.159	0.0473	1.50	0.158	0.0447
	AICC	2.00	0.172	0.0458	1.78	0.174	0.0443	1.53	0.179	0.0428
	$\mathrm{IC}_{\pi,1}$	1.93	0.151	0.0417	1.72	0.156	0.0410	1.47	0.159	0.0400
	$\mathrm{IC}^*_{\pi,1}$	2.14	0.190	0.0470	1.88	0.188	0.0452	1.60	0.186	0.0434
	$\mathrm{IC}_{\pi,2}$	2.14	0.192	0.0470	1.87	0.190	0.0452	1.59	0.186	0.0434
	IC_r	1.48	0.135	0.0422	1.37	0.137	0.0413	1.21	0.139	0.0404
	IC^*_r	1.91	0.251	0.0413	1.72	0.271	0.0434	1.52	0.316	0.0471
$n_0 = 5$	AIC	0.983	0.0696	0.0251	0.907	0.0659	0.0237	0.824	0.0619	0.0223
$m = 8$	BIC	1.25	0.0622	0.0224	1.11	0.0615	0.0216	1.01	0.0613	0.0209
	AICC	1.10	0.0655	0.0236	0.989	0.0627	0.0226	0.899	0.0610	0.0215
	$\mathrm{IC}_{\pi,1}$	0.989	0.0567	0.0197	0.888	0.0554	0.0196	0.804	0.0565	0.0194
	$\mathrm{IC}^*_{\pi,1}$	1.41	0.0594	0.0211	1.21	0.0613	0.0207	1.07	0.0635	0.0202
	$\mathrm{IC}_{\pi,2}$	1.40	0.0594	0.0211	1.20	0.0613	0.0207	1.07	0.0652	0.0202
	IC_r	0.743	0.0568	0.0197	0.666	0.0557	0.0196	0.602	0.0565	0.0194
	IC^*_r	1.16	0.0609	0.0196	1.07	0.0691	0.0195	1.01	0.0841	0.0193
$n_0 = 5$	AIC	0.451	0.0341	0.0123	0.440	0.0325	0.0117	0.434	0.0308	0.0111
$m = 16$	BIC	0.740	0.0294	0.0106	0.711	0.0289	0.0104	0.684	0.0284	0.0102
	AICC	0.489	0.0333	0.0120	0.483	0.0319	0.0115	0.481	0.0304	0.0109
	$\mathrm{IC}_{\pi,1}$	0.433	0.0280	0.00980	0.435	0.0277	0.00983	0.438	0.0275	0.00980
	$\mathrm{IC}^*_{\pi,1}$	0.864	0.0283	0.0102	0.812	0.0281	0.0101	0.770	0.0279	0.0100
	$\mathrm{IC}_{\pi,2}$	0.862	0.0283	0.0102	0.811	0.0281	0.0101	0.766	0.0279	0.0100
	IC_r	0.348	0.0280	0.00981	0.346	0.0277	0.00982	0.344	0.0274	0.00980
	IC^*_r	0.658	0.0278	0.00977	0.664	0.0276	0.00979	0.683	0.0281	0.00978

Table 6.6: The prediction error of the best model selected by the criteria for the regression model with AR(1) errors.

		$\phi = 0.1$			$\phi = 0.5$			$\phi = 0.8$		
	SNR	1	3	5	1	3	5	1	3	5
$n = 20$	AIC	1.85	0.175	0.0628	1.71	0.173	0.0613	1.75	0.256	0.0759
	BIC	2.01	0.170	0.0595	1.82	0.192	0.0583	1.73	0.305	0.0802
	AICC	2.04	0.163	0.0549	1.85	0.203	0.0577	1.69	0.343	0.0909
	$\text{IC}_{\pi,1}$	1.96	0.153	0.0492	1.78	0.182	0.0538	1.66	0.312	0.0831
	$\text{IC}^*_{\pi,1}$	2.17	0.164	0.0559	1.95	0.211	0.0566	1.71	0.355	0.0962
	$\text{IC}_{\pi,2}$	2.17	0.164	0.0559	1.95	0.216	0.0566	1.72	0.354	0.0972
	IC_r	1.79	0.154	0.0505	1.59	0.157	0.0516	1.71	0.224	0.0720
	IC^*_r	1.84	0.159	0.0507	1.72	0.212	0.0557	1.77	0.361	0.114
$n = 40$	AIC	0.783	0.0733	0.0264	0.812	0.0685	0.0246	1.09	0.124	0.0365
	BIC	0.973	0.0639	0.0230	0.992	0.0640	0.0225	1.13	0.162	0.0408
	AICC	0.824	0.0681	0.0245	0.881	0.0662	0.0236	1.12	0.135	0.0360
	$\text{IC}_{\pi,1}$	0.839	0.0587	0.0204	0.832	0.0595	0.0207	1.07	0.136	0.0347
	$\text{IC}^*_{\pi,1}$	1.15	0.0601	0.0216	1.10	0.0638	0.0218	1.14	0.202	0.0432
	$\text{IC}_{\pi,2}$	1.15	0.0601	0.0216	1.10	0.0639	0.0218	1.15	0.202	0.0441
	IC_r	0.782	0.0592	0.0203	0.711	0.0598	0.0208	0.919	0.118	0.0347
	IC^*_r	0.813	0.0592	0.0203	0.857	0.0631	0.0209	1.12	0.195	0.0446
$n = 80$	AIC	0.311	0.0342	0.0123	0.365	0.0329	0.0118	0.666	0.0520	0.0187
	BIC	0.301	0.0282	0.0102	0.500	0.0290	0.0105	0.861	0.0613	0.0182
	AICC	0.303	0.0331	0.0119	0.377	0.0322	0.0116	0.693	0.0526	0.0186
	$\text{IC}_{\pi,1}$	0.286	0.0262	0.00920	0.365	0.0279	0.00983	0.643	0.0530	0.0179
	$\text{IC}^*_{\pi,1}$	0.331	0.0266	0.00958	0.566	0.0284	0.0102	0.914	0.0692	0.0181
	$\text{IC}_{\pi,2}$	0.333	0.0266	0.00958	0.566	0.0284	0.0102	0.910	0.0692	0.0181
	IC_r	0.290	0.0264	0.00918	0.330	0.0278	0.00984	0.514	0.0499	0.0179
	IC^*_r	0.292	0.0264	0.00919	0.414	0.0283	0.00991	0.778	0.0662	0.0180

6.5 Discussion

We have derived the variable selection criteria for linear regression model relative to the frequentist KL risk of the predictive density based on the Bayesian marginal likelihood. We have proved the consistency of the criteria and have showed that they perform well also in the sense of the prediction through simulations.

We gave some advantages of the approach based on frequentist's risk $R(\boldsymbol{\eta}; \hat{m})$ in (6.1). We here explain them more clearly through comparison of the related Bayesian criteria. When the prior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega})$ is proper, we can treat the Bayesian prediction risk

$$r(\boldsymbol{\psi}; \hat{m}) = \int R(\boldsymbol{\eta}; \hat{m}) \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\beta}$$

in (6.5). When $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ are known, the predictive density $\hat{m}(\tilde{\boldsymbol{y}}; \boldsymbol{y})$ which minimizes $r(\boldsymbol{\psi}; \hat{m})$ is the Bayesian predictive density (posterior predictive density) $\hat{m}_\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\omega})$ given as

$$\int m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\omega}) \pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\beta} = \frac{\int m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\omega}) m(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\omega}) \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\beta}}{\int m(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\omega}) \pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\beta}}.$$

When $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ are unknown, we can consider the Bayesian risk of the plug-in predictive density $\hat{m}_\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\omega}})$. Then the resulting criterion is known as the predictive likelihood (Akaike, 1980a) or the PIC (Kitagawa, 1997). The deviance information criterion (DIC) of Spiegelhalter et al. (2002) and the Bayesian predictive information criterion (BPIC) of Ando (2007) are related criteria based on the Bayesian prediction risk $r(\boldsymbol{\psi}; \hat{m})$.

The Akaike's Bayesian information criterion (ABIC) (Akaike, 1980b) is another information criterion based on the Bayesian marginal likelihood, given by

$$\mathrm{ABIC} = -2 \log\{m_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}})\} + 2 \dim(\boldsymbol{\lambda}),$$

where the nuisance parameter $\boldsymbol{\omega}$ is not considered. The ABIC measures the following KL risk:

$$\int \left[\int \log \left\{ \frac{m_\pi(\tilde{\boldsymbol{y}}|\boldsymbol{\lambda})}{m_\pi(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}})} \right\} m_\pi(\tilde{\boldsymbol{y}}|\boldsymbol{\lambda}) \mathrm{d}\tilde{\boldsymbol{y}} \right] m_\pi(\boldsymbol{y}|\boldsymbol{\lambda}) \mathrm{d}\boldsymbol{y},$$

which is not the same as either $R(\boldsymbol{\eta}; \hat{m})$ or $r(\boldsymbol{\psi}; \hat{m})$. The ABIC is the criterion for choosing the hyperparameter $\boldsymbol{\lambda}$ in the same sense as the AIC. However, it is noted that the ABIC works as a model selection criterion for $\boldsymbol{\beta}$ because it is based on the Bayesian marginal likelihood.

A drawback of such Bayesian criteria is that we cannot construct them for improper prior distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega})$, since the corresponding Bayesian prediction risks do not exist. On the other hand, we can construct the corresponding criteria based on $R(\boldsymbol{\eta}; \hat{m})$, because the approach suggested in this paper measures the prediction risk in the framework of frequentists. In fact, putting the uniform improper prior on regression coefficients $\boldsymbol{\beta}$ in the linear regression model, we get the RIC of Shi and Tsai (2002). Note that the criteria based on improper marginal likelihood works as variable selection only when the marginal likelihood itself does. For the case where the improper priors cannot be used for model selection, intrinsic prior was proposed in the literature (Berger and Pericchi, 1996; Casella and Moreno, 2006, and others), which is an objective and automatic procedure. As future work, it is worthwhile to consider such an automatic procedure in the framework of our proposed criteria.

6.6 Derivations of the criteria

6.6.1 Derivation of $\mathrm{IC}_{\pi,1}$ in (6.8)

It is sufficient to show that the bias correction $\Delta_{\pi,1} = I_{\pi,1}(\boldsymbol{\eta}) - E_{\boldsymbol{\eta}}[-2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}^2)\}]$ is $2n/(n-p-2)$, where $I_{\pi,1}(\boldsymbol{\eta})$ is given by (6.7). It follows that

$$\begin{aligned}\Delta_{\pi,1} =& E_{\boldsymbol{\eta}}(\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{y}}/\hat{\sigma}^2) - E_{\boldsymbol{\eta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2)\\ =& E_{\boldsymbol{\eta}}(\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{y}}) \cdot E_{\boldsymbol{\eta}}(1/\hat{\sigma}^2) - E_{\boldsymbol{\eta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2).\end{aligned}$$

Firstly,

$$\begin{aligned}E_{\boldsymbol{\eta}}(\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{y}}) =& E_{\boldsymbol{\eta}}[(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{A}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta})]\\ =& \sigma^2\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}.\end{aligned} \tag{6.15}$$

Secondly, noting that $n\hat{\sigma}^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y} = \sigma^2\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}$ for

$$\begin{aligned}\boldsymbol{v} =& \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma,\\ \boldsymbol{M} =& \boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2},\end{aligned} \tag{6.16}$$

and that $\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v} \sim \chi^2_{n-p}$, we can obtain

$$\begin{aligned}E_{\boldsymbol{\eta}}(1/\hat{\sigma}^2) =& nE_{\boldsymbol{\eta}}\left[\frac{1}{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}}\right] = nE_{\boldsymbol{\eta}}\left[\frac{1}{\sigma^2\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right]\\ =& \frac{n}{\sigma^2(n-p-2)}.\end{aligned} \tag{6.17}$$

Finally,

$$\begin{aligned}E_{\boldsymbol{\eta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2) =& nE_{\boldsymbol{\eta}}\left[\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}}\right] = nE_{\boldsymbol{\eta}}\left[\frac{\sigma^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{1/2}\boldsymbol{v} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right]\\ =& n \times \left\{\frac{\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{\Sigma})}{n-p-2} - \frac{2\mathrm{tr}\,[\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}]}{(n-p)(n-p-2)} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2(n-p-2)}\right\}.\end{aligned} \tag{6.18}$$

The last equation in the above can be derived by Lemma 4.6. Combining (6.15), (6.17) and (6.18), we get

$$\Delta_{\pi,1} = \frac{2n \cdot \mathrm{tr}\,[\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}]}{(n-p)(n-p-2)}.$$

We can see that

$$\begin{aligned}\mathrm{tr}\,[\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}] &= \mathrm{tr}\,\{(\boldsymbol{\Sigma} + \boldsymbol{B})^{-1}(\boldsymbol{\Sigma} + \boldsymbol{B} - \boldsymbol{B})(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}\}\\ &= \mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}] - \mathrm{tr}\,\{(\boldsymbol{\Sigma} + \boldsymbol{B})^{-1}\boldsymbol{B}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}\}\\ &= \mathrm{tr}\,(\boldsymbol{I}_n - \boldsymbol{M}) = n - p,\end{aligned} \tag{6.19}$$

since $\boldsymbol{B}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P}) = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P}) = \boldsymbol{0}$, then we obtain $\Delta_{\pi,1} = 2n/(n-p-2)$. □

6.6.2 Derivation of $\mathrm{IC}_{\pi,2}$ in (6.9)

From the fact that $E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,1}) = I_{\pi,1}(\boldsymbol{\eta})$ and that $E_\pi E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,1}) = E_\pi[I_{\pi,1}(\boldsymbol{\eta})] = I_{\pi,2}(\sigma^2)$, it suffices to show that $E_\pi E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,1})$ is approximated to

$$\begin{aligned} E_\pi E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,1}) &\approx E_\pi E_{\boldsymbol{\eta}}[n\log\hat{\sigma}^2 + \log|\boldsymbol{\Sigma}| + p\log n + 2 + E_\pi E_{\boldsymbol{\eta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2)] \\ &\approx E_\pi E_{\boldsymbol{\eta}}[n\log\hat{\sigma}^2 + \log|\boldsymbol{\Sigma}| + p\log n + p] + (n+2) = E_\pi E_{\boldsymbol{\eta}}(\mathrm{IC}_{\pi,2}) + (n+2), \end{aligned}$$

when n is large. Note that $n+2$ is irrelevant to the model. It follows that

$$\begin{aligned} &E_{\boldsymbol{\eta}}\left(\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}}{\hat{\sigma}^2}\right) \\ =&n \times E_{\boldsymbol{\eta}}\left[\frac{\boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\}\boldsymbol{y}}\right] \\ =&n + n \times E_{\boldsymbol{\eta}}\left[\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\}\boldsymbol{y}}\right] \\ =&n + \frac{n}{\sigma^2(n-p-2)} \times E_{\boldsymbol{\eta}}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right] \\ =&n + \frac{n}{\sigma^2(n-p-2)} \times \Big[\sigma^2 \cdot \mathrm{tr}\left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\right\} \\ &+ \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}\Big], \end{aligned}$$

and that

$$E_\pi[\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}] = \sigma^2 \cdot \mathrm{tr}\,[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}].$$

If $n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ converges to $p\times p$ positive definite matrix as $n \to \infty$, $\mathrm{tr}\,[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}] \to 0$ and $\mathrm{tr}\,[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}] \to p$. Then we can obtain $E_\pi E_{\boldsymbol{\eta}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2 - n) \to p$, which we want to show.

6.6.3 Derivation of IC_r in (6.11)

We shall show that the bias correction $\Delta_r = I_r(\boldsymbol{\eta}) - E_{\boldsymbol{\eta}}[-2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}^2)\}]$ is $2(n-p)/(n-p-2)$, where $I_r(\boldsymbol{\eta})$ is given by (6.10). Then,

$$\begin{aligned} \Delta_r &= E_{\boldsymbol{\eta}}[\tilde{\boldsymbol{y}}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\tilde{\boldsymbol{y}}/\tilde{\sigma}^2] - E_{\boldsymbol{\eta}}[\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/\tilde{\sigma}^2] \\ &= E_{\boldsymbol{\eta}}[\tilde{\boldsymbol{y}}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\tilde{\boldsymbol{y}}] \cdot E_{\boldsymbol{\eta}}(1/\tilde{\sigma}^2) - (n-p). \end{aligned}$$

Since $E_{\boldsymbol{\eta}}[\tilde{\boldsymbol{y}}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\tilde{\boldsymbol{y}}] = (n-p)\sigma^2$ and $E_{\boldsymbol{\eta}}(1/\tilde{\sigma}^2) = (n-p)/\{\sigma^2(n-p-2)\}$, we get $\Delta_r = 2(n-p)/(n-p-2)$. □

6.7 Proof of Theorem 6.1

We only prove the consistency of $\mathrm{IC}_{\pi,1}$. The proof of the consistency of the other criteria can be done in the same manner. Because we see that

$$P(\hat{j} = j) \le P\{\mathrm{IC}_{\pi,1}(j) < \mathrm{IC}_{\pi,1}(j_*)\}$$

for any $j \in \mathcal{J} \setminus \{j_*\}$, it suffices to show that $P\{\mathrm{IC}_{\pi,1}(j) < \mathrm{IC}_{\pi,1}(j_*)\} \to 0$, or equivalently $P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) > 0\} \to 1$ as $n \to \infty$. When $\boldsymbol{\Sigma} = \boldsymbol{I}_n$, we obtain

$$\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) = I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 =& n \log(\hat{\sigma}_j^2/\hat{\sigma}_*^2) + \boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_j \boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_* \boldsymbol{y}/\hat{\sigma}_*^2, \\ I_2 =& \log |\boldsymbol{X}(j)^{\mathrm{T}} \boldsymbol{X}(j) + \boldsymbol{W}_j^{-1}| - \log |\boldsymbol{X}(j_*)^{\mathrm{T}} \boldsymbol{X}(j_*) + \boldsymbol{W}_*^{-1}|, \\ I_3 =& \log\{|\boldsymbol{W}_j|/|\boldsymbol{W}_*|\} + \frac{2n}{n - p_j - 2} - \frac{2n}{n - p_* - 2}, \end{aligned}$$

for $\hat{\sigma}_j^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}_j)\boldsymbol{y}/n$, $\hat{\sigma}_*^2 = \hat{\sigma}_{j_*}^2$, $\boldsymbol{A}_j = \boldsymbol{I}_n - \boldsymbol{X}(j)(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{X}(j) + \boldsymbol{W}_j^{-1})^{-1}\boldsymbol{X}(j)^{\mathrm{T}}$, $\boldsymbol{A}_* = \boldsymbol{A}_{j_*}$ and $\boldsymbol{W}_* = \boldsymbol{W}_{j_*}$. We evaluate asymptotic behaviors of I_1, I_2 and I_3 for $j \in \mathcal{J}_-$ and $j \in \mathcal{J}_+ \setminus \{j_*\}$, separately.

[Case of $j \in \mathcal{J}_-$]. Firstly, we evaluate I_1. We decompose $I_1 = I_{11} + I_{12}$, where $I_{11} = n \log(\hat{\sigma}_j^2/\hat{\sigma}_*^2)$ and $I_{12} = \boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_j \boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_* \boldsymbol{y}/\hat{\sigma}_*^2$. It follows that

$$\begin{aligned} \hat{\sigma}_j^2 - \hat{\sigma}_*^2 =& (\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{u})^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}_j)(\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{u})/n - \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}_*)\boldsymbol{u}/n \\ =& \|\boldsymbol{X}(\omega)\boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2/n + o_p(1), \end{aligned}$$

where $\boldsymbol{H}_* = \boldsymbol{H}_{j_*}$. Then we can see that

$$n^{-1} I_{11} = \log\left(1 + \frac{\hat{\sigma}_j^2 - \hat{\sigma}_*^2}{\hat{\sigma}_*^2}\right) = \log\left\{1 + \frac{\|\boldsymbol{X}(\omega)\boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2}{n\sigma^2}\right\} + o_p(1), \tag{6.20}$$

and it follows from the assumption (A3) that

$$\liminf_{n\to\infty} \log\left\{1 + \frac{\|\boldsymbol{X}(\omega)\boldsymbol{\beta}_* - \boldsymbol{H}_j \boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2}{n\sigma^2}\right\} > 0. \tag{6.21}$$

Because $\boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_j \boldsymbol{y}/(n\hat{\sigma}_j^2) = 1 + o_p(1)$ and $\boldsymbol{y}^{\mathrm{T}} \boldsymbol{A}_* \boldsymbol{y}/(n\hat{\sigma}_*^2) = 1 + o_p(1)$, we obtain

$$n^{-1} I_{12} = o_p(1). \tag{6.22}$$

Secondly, we evaluate I_2. It follows that

$$\log |\boldsymbol{X}(j)^{\mathrm{T}} \boldsymbol{X}(j) + \boldsymbol{W}_j^{-1}| = p_j \log n + \log |\boldsymbol{X}(j)^{\mathrm{T}} \boldsymbol{X}(j)/n + \boldsymbol{W}_j^{-1}/n| = p_j \log n + O(1).$$

It can be also seen that $\log |\boldsymbol{X}(j_*)^{\mathrm{T}} \boldsymbol{X}(j_*) + \boldsymbol{W}_*^{-1}| = p_* \log n + O(1)$. Then,

$$n^{-1} I_2 = (p_j - p_*) n^{-1} \log n + o(1) = o(1). \tag{6.23}$$

Lastly, it is easy to see that

$$n^{-1} I_3 = o(1). \tag{6.24}$$

From (6.20)–(6.24), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) > 0\} \to 1, \tag{6.25}$$

for all $j \in \mathcal{J}-$.

[Case of $j \in \mathcal{J}_+ \setminus \{j_*\}$]. Firstly, we evaluate I_1. From the fact that

$$\hat{\sigma}_*^2 - \hat{\sigma}_j^2 = \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{H}_j - \boldsymbol{H}_*)\boldsymbol{u}/n = O_p(n^{-1}), \tag{6.26}$$

it follows that

$$\begin{aligned}(\log n)^{-1} I_{11} =& (\log n)^{-1} \cdot n \log \left\{ \frac{\hat{\sigma}_*^2 - (\hat{\sigma}_*^2 - \hat{\sigma}_j^2)}{\hat{\sigma}_*^2} \right\} \\ =& (\log n)^{-1} \cdot n \cdot \log\{1 + O_p(n^{-1})\} = o_p(1).\end{aligned} \tag{6.27}$$

As for I_{12}, from (6.26) and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_0\boldsymbol{y} = O_p(1)$, we can obtain

$$\begin{aligned}I_{12} &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_*\boldsymbol{y}/\hat{\sigma}_*^2 \\ &= (\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}_*\boldsymbol{y})/\hat{\sigma}_*^2 + O_p(1) = O_p(1).\end{aligned}$$

Then,

$$(\log n)^{-1} I_{12} = o_p(1). \tag{6.28}$$

Secondly, we evaluate I_2. Since $p_j > p_*$ for all $j \in \mathcal{J}_+ \setminus \{j_*\}$,

$$\liminf_{n\to\infty} (\log n)^{-1} I_2 = p_j - p_* > 0. \tag{6.29}$$

Finally, it is easy to see that

$$(\log n)^{-1} I_3 = o(1). \tag{6.30}$$

From (6.27)–(6.30), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,2}(j_*) > 0\} \to 1, \tag{6.31}$$

for all $j \in \mathcal{J}_+ \setminus \{j_*\}$.

Combining (6.25) and (6.31), we obtain

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_*) > 0\} \to 1,$$

for all $j \in \mathcal{J} \setminus \{j_*\}$, which shows that $\mathrm{IC}_{\pi,1}$ is consistent. □

Chapter 7

Variants of conditional AIC in linear mixed models

In this chapter, we consider information criteria which measure prediction risks of several predictive densities in terms of expected Kullback–Leibler divergence based on conditional likelihood given random effects for variable selection problem in linear mixed model. When the predictive density considered is plug-in predictive density, the resulting criterion is the conditional AIC proposed by Vaida and Blanchard (2005). We consider two types of predictive densities, both of which are superior to plug-in predictive density in some sense. The first one is the Bayesian predictive density, which is the best predictive density in the sense of minimizing the expected Kullback–Leibler divergence. The second one is the predictive density based on the Bayesian marginal likelihood. The resulting criterion is related to the ones introduced in the last chapter.

7.1 Motivation

For variable selection problem in linear mixed model, Vaida and Blanchard (2005) introduced the conditional Akaike information (cAI), which is relevant to expected Kullback–Leibler (KL) divergence based on the conditional likelihood given random effects. This risk function is appropriate when one is interested in predicting the random effects. The cAI or its estimator conditional AIC (cAIC) measures the risk of the plug-in predictive density. However, there is no reason to restrict the predictive density to plug-in predictive density. Then, in this chapter, we consider two types of predictive densities.

The first one is the Bayesian (posterior) predictive density, which is the best predictive density among any predictive density in the sense of minimizing the expected KL divergence. Information criterion based on the Bayesian predictive density is known as the predictive likelihood (Akaike, 1980a) or the predictive information criterion (PIC) (Kitagawa, 1997). We construct PIC's in linear mixed model for two situations. In one situation, we assume that the vector of regression coefficients is unknown parameter. In the other situation, we consider the prior distribution of regression coefficients.

The second one is a predictive density based on Bayesian marginal likelihood assuming prior distribution on the regression coefficients of fixed effects. The resulting criterion is relevant to the AIC variant using Bayesian marginal likelihood, which was introduced in the last chapter. Like the AIC variant, there are three advantages of this method. Firstly, this criterion is less influenced by prior misspecification because risk function does not take expectation with respect to the prior distribution of regression coefficients. Secondly, non-informative improper prior can

be also used for constructing the criterion. When the uniform prior is assumed on the regression coefficients, the Bayesian marginal likelihood is identical to residual likelihood. Then we call the resulting criterion the conditional residual information criterion (cRIC). Lastly, the criteria have the consistency property for selecting the true model.

The rest of this chapter is organized as follows. In Section 7.2, we explain about the setup of variable selection and about the concept of cAIC variants in detail. In Section 7.3, we propose two PIC's in linear mixed model. In Section 7.4, we propose the cAIC variants based on Bayesian marginal likelihood. The numerical performance of the proposed criteria is investigated by simulations in Section 7.5.

7.2 Variants of conditional Akaike information

7.2.1 Setup of variable selection

The candidate model j is the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j, \tag{7.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of response variables, $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ are $n \times p_j$ and $n \times q$ matrices of covariates, respectively, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of coefficients, $\boldsymbol{b}_j$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_j$ is an $n \times 1$ vector of random errors. Let $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ be mutually independent and $\boldsymbol{b}_j \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_j^2\boldsymbol{G})$, $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}_n)$, where $\boldsymbol{G}$ is a $q \times q$ positive definite matrix and σ_j^2 is a scalar. We assume that $\boldsymbol{G}$ is known and σ_j^2 is unknown.

To derive the criterion, we assume that the true model j_* is included by each candidate model j, namely each candidate model j is overspecified. In that situation, the true model j_* can be expressed by

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j^* + \boldsymbol{Z}\boldsymbol{b}_* + \boldsymbol{\varepsilon}_*, \tag{7.2}$$

where $\boldsymbol{b}_* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_*^2\boldsymbol{G})$, $\boldsymbol{\varepsilon}_* \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_*^2\boldsymbol{I}_n)$ and $\boldsymbol{\beta}_j^*$ is a $p_j \times 1$ vector of regression coefficients, whose $p_j - p_*$ components are exactly 0 and the rest of components are not 0, for $p_* = p_{j_*}$. Henceforth, we abbreviate the model index j for notational convenience when there is no confusion. We also abbreviate $\boldsymbol{\beta}_j^*$ as $\boldsymbol{\beta}$, $\boldsymbol{b}_*$ as $\boldsymbol{b}$ and σ_*^2 as σ^2. Let $f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)$ and $p(\boldsymbol{b}|\sigma^2)$ denote the conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}$ and the density function of $\boldsymbol{b}$, respectively.

7.2.2 Conditional Kullback–Leibler risk

A conventional method of selecting the explanatory variable in linear mixed model is to use the marginal AIC (mAIC). The mAIC is based on the marginal likelihood integrating out the random effects $\boldsymbol{b}$, which is given by

$$\mathrm{mAIC} = -2\log\{m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\} + 2p_j,$$

where $m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \int f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{b}|\sigma^2)\mathrm{d}\boldsymbol{b}$ and $\widehat{\boldsymbol{\beta}}_j$ is the maximum likelihood estimator of $\boldsymbol{\beta}_j$. However, the mAIC is not appropriate for the focus on the prediction of specific clusters or random effects. Then, Vaida and Blanchard (2005) considered the expected Kullback–Leibler (KL) divergence based on the conditional density, which is given by

$$R_c(\boldsymbol{\eta}; \hat{f}_j) = \iint \left[\int \log \left\{ \frac{f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)}{\hat{f}_j(\tilde{\boldsymbol{y}}; \boldsymbol{y})} \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)\mathrm{d}\tilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{b}|\sigma^2)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b}, \tag{7.3}$$

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ given $\boldsymbol{b}$, $\hat{f}_j(\tilde{\boldsymbol{y}};\boldsymbol{y})$ is some predictive density of $f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2)$ based on the candidate model j, and $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}},\sigma^2)^{\mathrm{T}}$. We call $R_c(\boldsymbol{\eta};\hat{f}_j)$ in (7.3) the conditional KL risk. We can provide an information criterion as an (asymptotically) unbiased estimator of the information

$$I_c(\boldsymbol{\eta};\hat{f}_j) = \iiint -2\log\left\{\hat{f}_j(\tilde{\boldsymbol{y}};\boldsymbol{y})\right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) p(\boldsymbol{b}|\sigma^2)\mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b},$$

which is a part of (7.3) (multiplied by 2). Vaida and Blanchard (2005) proposed the conditional AIC (cAIC) as an unbiased estimator of $I_c(\boldsymbol{\eta};\hat{f}_j)$ for $\hat{f}_j = f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)$, which is called plug-in predictive density, where $\hat{\boldsymbol{b}}_j$, $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are some predictor of $\boldsymbol{b}_j$ and estimators of $\boldsymbol{\beta}_j$ and σ_j^2, respectively. The cAIC is of the form

$$\mathrm{cAIC} = -2\log\{f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)\} + \Delta_{\mathrm{cAI}},$$

where $\Delta_{\mathrm{cAI}} = I_c(\boldsymbol{\eta};\hat{f}_j) - E[-2\log\{f(\boldsymbol{y}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)\}]$.

Although the conditional KL risk $R_c(\boldsymbol{\eta};\hat{f}_j)$ is appropriate for variable selection in linear mixed model, there is no reason to restrict the predictive density $\hat{f}_j$ to plug-in predictive density. Thus we consider two types of predictive densities: the first one is the Bayesian (posterior) predictive density and the second one is a predictive density based on Bayesian marginal likelihood, both of which are superior to the plug-in predictive density in some sense. In the following sections, we derive information criteria based on these predictive densities and discuss their properties.

7.3 Predictive information criterion

7.3.1 Bayesian predictive density in linear mixed model

Aitchison (1975) showed that the best predictive density $\hat{f}_j$ which minimizes the risk $R_c(\boldsymbol{\eta};\hat{f}_j)$ is

$$\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta}) = \frac{\int f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) p(\boldsymbol{b}|\sigma^2)\mathrm{d}\boldsymbol{b}}{\int f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) p(\boldsymbol{b}|\sigma^2)\mathrm{d}\boldsymbol{b}}, \tag{7.4}$$

when $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}},\sigma^2)^{\mathrm{T}}$ is known. The predictive density $\hat{f}^{\mathrm{BP}}$ is known as the Bayesian (or posterior) predictive density. Note that the linear mixed model can be seen as a Bayesian model where the random effects $\boldsymbol{b}$ has prior distribution $p(\boldsymbol{b}|\sigma^2)$. The cAIC measures the risk $R_c(\boldsymbol{\eta};\hat{f}_j)$ for $\hat{f}_j = f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}^2)$, which is plug-in predictive density, and it follows that

$$R_c(\boldsymbol{\eta};f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j,\boldsymbol{\beta},\sigma^2)) \geq R_c(\boldsymbol{\eta};\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta})),$$

when $\boldsymbol{\eta}$ is known. Thus it is natural to use the Bayesian predictive density to measure the prediction risk of the model. Information criterion based on the Bayesian predictive density is known as the predictive likelihood (Akaike, 1980a) or the predictive information criterion (PIC) (Kitagawa, 1997).

The next proposition shows the Bayesian predictive distribution in linear mixed model (7.2).

Proposition 7.1 *In the linear mixed model* (7.2)*, the Bayesian predictive distribution whose density function is given by* (7.4) *is*

$$\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}},\sigma^2\boldsymbol{V}), \tag{7.5}$$

where $\hat{\boldsymbol{b}}^{\mathrm{B}} = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ *and* $\boldsymbol{V} = 2\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1}$.

7.3.2 Derivation of PIC in linear mixed model

In this subsection, we derive the PIC in linear mixed model based on the Bayesian predictive distribution (7.5). Because $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ is unknown, we estimate it by maximum likelihood based on each candidate model j as follows:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j &= (\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y},\\ \hat{\sigma}_j^2 &= (\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)/n.\end{aligned}$$

Then we consider the following information:

$$\begin{aligned}I_c(\boldsymbol{\eta}; \hat{f}_j^{\mathrm{BP}}) &= \iiint -2\log\left\{\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{b}|\sigma^2) \mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b},\\ &= \int \left[\int -2\log\left\{\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\} \hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}}\right] m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\mathrm{d}\boldsymbol{y}\\ &= \mathrm{PI} \quad \text{(say)}, \end{aligned} \tag{7.6}$$

We want to construct an information criterion as an unbiased estimator of PI, which is of the form

$$\mathrm{PIC} = -2\log\left\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\} + \Delta_{\mathrm{PI}},$$

where

$$\begin{aligned}-2\log\left\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\} &= n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2,\\ \Delta_{\mathrm{PI}} &= \mathrm{PI} - E[-2\log\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\}]\end{aligned}$$

for $\hat{\boldsymbol{b}}_j = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j)$. Then we have to evaluate the bias correction Δ_{PI}.

Firstly, taking expectation of

$$-2\log\left\{\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\} = n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}| + (\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\boldsymbol{V}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2,$$

with respect to the distribution of $\tilde{\boldsymbol{y}}|\boldsymbol{y} \sim \hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}}, \sigma^2\boldsymbol{V})$, we can rewrite the PI in (7.6) as

$$\begin{aligned}\mathrm{PI} &= E\Big[n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}| + n\sigma^2/\hat{\sigma}_j^2\\ &\qquad + \left\{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \hat{\boldsymbol{b}}^{\mathrm{B}})\right\}^{\mathrm{T}}\boldsymbol{V}^{-1}\left\{\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \hat{\boldsymbol{b}}^{\mathrm{B}})\right\}\Big]\\ &= E\Big[n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}| + n\sigma^2/\hat{\sigma}_j^2\\ &\qquad + (\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})/\hat{\sigma}_j^2\Big],\end{aligned}$$

noting that $\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}) + \boldsymbol{Z}(\hat{\boldsymbol{b}}_j - \hat{\boldsymbol{b}}^{\mathrm{B}}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})$. Next, $-2\log\left\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\}$ is rewritten as

$$\begin{aligned}&-2\log\left\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\right\}\\ =&n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}| + \left\{\boldsymbol{u} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})\right\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{u} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})\right\}/\hat{\sigma}_j^2,\end{aligned}$$

where $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. Then the bias correction Δ_{PI} is

$$\begin{aligned}\Delta_{\mathrm{PI}} &= E\left[n\sigma^2/\hat{\sigma}_j^2 - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}/\hat{\sigma}_j^2 + 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})/\hat{\sigma}_j^2\right]\\ &= I_1 - I_2 + 2I_3 \quad \text{(say)}.\end{aligned}$$

Thus, it suffices to evaluate I_1, I_2 and I_3.

Noting that $n\hat{\sigma}_j^2/\sigma^2 \sim \chi^2_{n-p_j}$, we can evaluate I_1 as

$$I_1 = \frac{n^2}{n-p_j-2}.$$

Next, I_2 is rewritten as

$$I_2 = n \cdot E\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right],$$

where $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}/\sigma$ and $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}$. Then, it follows from Lemma 4.6 that

$$\begin{aligned} I_2 &= n \times \left\{\frac{\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1})}{n-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I}_n-\boldsymbol{M})]}{(n-p_j)(n-p_j-2)}\right\} \\ &= \frac{n\cdot\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1})}{n-p_j} + \frac{2n\cdot\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{P})}{(n-p_j)(n-p_j-2)}, \end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. Lastly, I_3 can be rewritten as

$$\begin{aligned} I_3 &= n \cdot E\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] \\ &= n \cdot E\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] + E\left[\frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right]. \end{aligned}$$

Because $\boldsymbol{M}\boldsymbol{v}$ and $(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}$ are independent and $E(\boldsymbol{M}\boldsymbol{v}) = \boldsymbol{0}$, the second term of the equation above is 0. Then,

$$I_3 = \frac{n\cdot\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{P})}{n-p_j-2}.$$

Thus we can obtain

$$\Delta_{\mathrm{PI}} = \frac{n^2}{n-p_j-2} - \frac{n\cdot\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1})}{n-p_j} + \frac{2n(n-p_j-1)\mathrm{tr}\,(\boldsymbol{V}^{-1}\boldsymbol{P})}{(n-p_j)(n-p_j-2)},$$

and propose the following PIC:

$$\mathrm{PIC} = -2\log\left\{\hat{f}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y},\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)\right\} + \Delta_{\mathrm{PI}}. \tag{7.7}$$

Theorem 7.1 *The* PIC *in* (7.7) *is an unbiased estimator of* PI *in* (7.6)*, namely* $E(\mathrm{PIC}) = \mathrm{PI}$.

7.3.3 Another PIC putting prior on regression coefficients

Although the PIC in (7.7) works as a variable selection criterion, the predictive density function $\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)$ seems not to be very appropriate for evaluation of the risk of the candidate model j. This is because the Bayesian predictive density $\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta})$, which minimizes the conditional KL risk, is derived under the condition that $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}},\sigma^2)^{\mathrm{T}}$ is known, although $\boldsymbol{\beta}$ varies with the model j. Then, we consider in this subsection another approach to construct the PIC.

To this end, we put the prior on the regression coefficients $\boldsymbol{\beta}$ as well as on the random effects $\boldsymbol{b}$. We assume that the prior distribution of $\boldsymbol{\beta}$ is $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}|\sigma^2) = \mathcal{N}_p(\mathbf{0}, \sigma^2\boldsymbol{W})$, where $\boldsymbol{W}$ is a known positive definite matrix, and that $\boldsymbol{\beta}$ is independent of $\boldsymbol{b}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}})^{\mathrm{T}}$ and $\psi(\boldsymbol{\theta}|\sigma^2)$ denote the prior density function of $\boldsymbol{\theta}$. Then, we can see the linear mixed model as the Bayesian model

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\theta} &\sim g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma^2\boldsymbol{I}_n), \\ \boldsymbol{\theta} &\sim \psi(\boldsymbol{\theta}|\sigma^2) = p(\boldsymbol{b}|\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2) = \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{G}) \cdot \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{W}), \end{aligned} \tag{7.8}$$

In this model, the Bayesian predictive density is

$$\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2) = \frac{\int g(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2) g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{\theta}}{\int g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{\theta}}, \tag{7.9}$$

where $\boldsymbol{z}$ is independent replication of $\boldsymbol{y}$ given $\boldsymbol{\theta}$. Note that $\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2)$ is the predictive density which minimizes the following conditional KL risk:

$$\overline{R}_c(\sigma^2; \hat{g}) = \iint \left[\int \log \left\{ \frac{g(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2)}{\hat{g}(\boldsymbol{z}; \boldsymbol{y})} \right\} g(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2)\mathrm{d}\boldsymbol{z} \right] g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{\theta},$$

when σ^2 is known, namely for any predictive density $\hat{g}(\boldsymbol{z}; \boldsymbol{y})$ the following inequality holds:

$$\overline{R}_c(\sigma^2; \hat{g}(\boldsymbol{z}; \boldsymbol{y})) \geq \overline{R}_c(\sigma^2; \hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2))$$

It is also important to note that $\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2)$ implicitly depends on the candidate model j. Thus an information criterion based on $\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2)$ is appropriate for variable selection.

The next proposition shows the Bayesian predictive distribution in the Bayesian model (7.8).

Proposition 7.2 *In the Bayesian model* (7.8)*, the Bayesian predictive distribution whose density function is given by* (7.9) *is*

$$\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2) = \mathcal{N}\big(\{\boldsymbol{I}_n - (\boldsymbol{B} + \boldsymbol{\Sigma})^{-1}\}\boldsymbol{y}, \sigma^2\boldsymbol{V}_2\big), \tag{7.10}$$

where $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}$ *and* $\boldsymbol{V}_2 = 2\boldsymbol{I}_n - (\boldsymbol{B} + \boldsymbol{\Sigma})^{-1}$.

We consider the following information:

$$\begin{aligned} \mathrm{PI}_2 &= \iiint -2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \hat{\sigma}_j^2)\right\} g(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2) g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{\theta} \\ &= \int \left[\int -2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \hat{\sigma}_j^2)\right\} \hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y}, \sigma^2)\mathrm{d}\boldsymbol{z} \right] g_\psi(\boldsymbol{y}|\sigma^2)\mathrm{d}\boldsymbol{y}, \end{aligned} \tag{7.11}$$

where $g_\psi(\boldsymbol{y}|\sigma^2) = \int g(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{\theta}$. We want to construct an information criterion as an unbiased estimator of PI_2, which is of the form

$$\mathrm{PIC}_2 = -2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \hat{\sigma}_j^2)\right\} + \Delta_{\mathrm{PI}_2},$$

where

$$\begin{aligned} -2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \hat{\sigma}_j^2)\right\} &= n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{V}_2| + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{B} + \boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B} + \boldsymbol{\Sigma})^{-1}\boldsymbol{y}/\hat{\sigma}_j^2, \\ \Delta_{\mathrm{PI}_2} &= \mathrm{PI}_2 - E_{g_\psi}[-2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y}, \hat{\sigma}_j^2)\right\}], \end{aligned}$$

and E_{g_ψ} denotes the expectation with respect to the distribution of $g_\psi(\boldsymbol{y}|\sigma^2)$. Then we have to evaluate the bias correction Δ_{PI_2}.

Firstly, taking expectation of

$$\begin{aligned}&-2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y},\hat{\sigma}_j^2)\right\}\\&=n\log(2\pi\hat{\sigma}_j^2)+\log|\boldsymbol{V}_2|+\left[\boldsymbol{z}-\{\boldsymbol{I}_n-(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\}\boldsymbol{y}\right]^{\mathrm{T}}\boldsymbol{V}_2^{-1}\left[\boldsymbol{z}-\{\boldsymbol{I}_n-(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\}\boldsymbol{y}\right]/\hat{\sigma}_j^2,\end{aligned}$$

with respect to the distribution of $\hat{g}^{\mathrm{BP}}(\boldsymbol{z}|\boldsymbol{y},\sigma^2)=\mathcal{N}(\{\boldsymbol{I}_n-(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\}\boldsymbol{y},\sigma^2\boldsymbol{V}_2)$, we can rewrite PI_2 in (7.11) as

$$\mathrm{PI}_2=E_{g_\psi}\left[n\log(2\pi\hat{\sigma}_j^2)+\log|\boldsymbol{V}_2|+n\sigma^2/\hat{\sigma}_j^2\right].$$

Then, the bias correction Δ_{PI_2} can be expressed as

$$\begin{aligned}\Delta_{\mathrm{PI}_2}&=E_{g_\psi}\left[n\sigma^2/\hat{\sigma}_j^2-\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y}/\hat{\sigma}_j^2\right]\\&=I_4-I_5\quad\text{(say)}.\end{aligned}$$

Thus, it suffices to evaluate I_4 and I_5.

Noting the fact that

$$\begin{aligned}g_\psi(\boldsymbol{y}|\sigma^2)=&\int g(\boldsymbol{y}|\boldsymbol{\theta},\sigma^2)\psi(\boldsymbol{\theta}|\sigma^2)\mathrm{d}\boldsymbol{\theta}=\iint f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b},\sigma^2)p(\boldsymbol{b}|\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\mathrm{d}\boldsymbol{b}\mathrm{d}\boldsymbol{\beta}\\=&\int m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\mathrm{d}\boldsymbol{\beta},\end{aligned}$$

I_4 can be rewritten as

$$I_4=E_{g_\psi}[n\sigma^2/\hat{\sigma}_j^2]=n^2E_{\boldsymbol{\beta}}E_{\boldsymbol{y}|\boldsymbol{\beta}}[n\sigma^2/\hat{\sigma}_j^2],$$

where $E_{\boldsymbol{y}|\boldsymbol{\beta}}$ and $E_{\boldsymbol{\beta}}$ denote the expectation with respect to the distribution of $\boldsymbol{y}|\boldsymbol{\beta}\sim m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)$ and $\boldsymbol{\beta}\sim\pi(\boldsymbol{\beta}|\sigma^2)$, respectively. Then, I_4 is evaluated as

$$I_4=\frac{n^2}{n-p_j-2}.$$

Next, I_5 can be decomposed as

$$\begin{aligned}I_5&=nE_{g_\psi}\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right]+nE_{g_\psi}\left[\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2\cdot\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right]\\&=I_{51}+I_{52}\quad\text{(say)}.\end{aligned}$$

Using Lemma 4.6, we can evaluate I_{51} as

$$\begin{aligned}I_{51}&=\frac{n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\right]}{n-p_j-2}-\frac{2n\cdot\mathrm{tr}\left[\boldsymbol{\Sigma}^{1/2}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}_n-\boldsymbol{M})\right]}{(n-p_j)(n-p_j-2)}\\&=\frac{n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\right]}{n-p_j-2}-\frac{2n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}+\boldsymbol{I}_n)^{-1}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}+\boldsymbol{I}_n)^{-1}\right]}{(n-p_j)(n-p_j-2)}.\end{aligned}$$

I_{52} can be evaluated as

$$\begin{aligned} I_{52} &= nE_{\boldsymbol{\beta}}E_{\boldsymbol{y}|\boldsymbol{\beta}}\left[\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2\cdot\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right] \\ &= nE_{\boldsymbol{\beta}}\left[\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2(n-p_j-2)}\right] \\ &= \frac{n\cdot\mathrm{tr}\left[\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{X}\boldsymbol{W}\right]}{n-p_j-2} \\ &= \frac{n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\right]}{n-p_j-2}. \end{aligned}$$

Then, it follows that

$$I_5 = \frac{n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\right]}{n-p_j-2} - \frac{2n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}+\boldsymbol{I}_n)^{-1}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}+\boldsymbol{I}_n)^{-1}\right]}{(n-p_j)(n-p_j-2)}$$

Thus we can obtain

$$\Delta_{\mathrm{PI}_2} = \frac{n^2}{n-p_j-2} - \frac{n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{B}+\boldsymbol{\Sigma})^{-1}\right]}{n-p_j-2} + \frac{2n\cdot\mathrm{tr}\left[\boldsymbol{V}_2^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}+\boldsymbol{I}_n)^{-1}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}+\boldsymbol{I}_n)^{-1}\right]}{(n-p_j)(n-p_j-2)},$$

and propose the following information criterion:

$$\mathrm{PIC}_2 = -2\log\left\{\hat{g}^{\mathrm{BP}}(\boldsymbol{y}|\boldsymbol{y},\hat{\sigma}_j^2)\right\} + \Delta_{\mathrm{PI}_2}. \tag{7.12}$$

Theorem 7.2 *The* PIC_2 *in* (7.12) *is an unbiased estimator of* PI_2 *in* (7.11)*, namely* $E_{g_\psi}(\mathrm{PIC}_2) = \mathrm{PI}_2$*.*

7.4 Conditional AIC variant based on Bayesian marginal likelihood

7.4.1 Conditional KL risk of predictive density based on Bayesian marginal likelihood

In Chapter 6, we considered the KL risk of predictive density based on Bayesian marginal likelihood. We derived criteria for variable selection in linear regression model by putting the prior on regression coefficients. The key point of the criteria is that the risk of the predictive density is measured by KL divergence not considering the prior distribution of the regression coefficients. In other words, we considered the risk of the Bayesian model (marginal likelihood) by frequentist point of view.

In this section, we consider the conditional KL risk (7.3) for the predictive density $\hat{f}_j$ which is based on the Bayesian marginal likelihood putting the prior on the unknown parameter of interest. Especially, we assume the prior distribution of the vector of the regression coefficients and derive variable selection criteria in linear mixed model.

7.4.2 Case of normal prior

In this subsection, we consider normal prior on regression coefficients. We assume the prior distribution of $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{W}),$$

where $\boldsymbol{W}$ is a $p \times p$ matrix suitably chosen with full rank. Then the marginal likelihood of $\boldsymbol{y}$ is

$$\begin{aligned} m_\pi(\boldsymbol{y}|\sigma^2) &= \iint f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{b}|\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\mathrm{d}\boldsymbol{b}\mathrm{d}\boldsymbol{\beta} = \int m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2)\mathrm{d}\boldsymbol{\beta} \\ &= (2\pi\sigma^2)^{-n/2}\cdot|\boldsymbol{\Sigma}|^{-1/2}\cdot|\boldsymbol{W}|^{-1/2}\cdot|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1}|^{-1/2}\cdot\exp\{-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/(2\sigma^2)\}, \end{aligned}$$

where $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. Note that $\boldsymbol{A} = (\boldsymbol{\Sigma}+\boldsymbol{B})^{-1}$ for $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}$, namely $m_\pi(\boldsymbol{y}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{\Sigma}+\boldsymbol{B}))$. Then we consider the following information:

$$\begin{aligned} I_c(\boldsymbol{\eta}; m_\pi) &= \iiint -2\log\left\{m_\pi(\tilde{\boldsymbol{y}}|\hat{\sigma}_j^2)\right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{b}|\sigma^2)\mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b} \\ &= \int\left[\int -2\log\{m_\pi(\tilde{\boldsymbol{y}}|\hat{\sigma}_j^2)\}\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta})\mathrm{d}\tilde{\boldsymbol{y}}\right] m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\mathrm{d}\boldsymbol{y}. \end{aligned} \tag{7.13}$$

We want to construct an information criterion as an unbiased estimator $I_c(\boldsymbol{\eta}; m_\pi)$, which is of the form

$$\mathrm{IC}_{c,\pi} = -2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\} + \Delta_{c,\pi},$$

where

$$\begin{aligned} -2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\} &= n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p| + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}_j^2, \\ \Delta_{c,\pi} &= I_c(\boldsymbol{\eta}; m_\pi) - E[-2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\}]. \end{aligned}$$

Note that the expectation in the equation above is the one with respect to the distribution of $\boldsymbol{y} \sim m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)$. Then we have to evaluate the bias correction $\Delta_{c,\pi}$.

Firstly, taking expectation of

$$-2\log\{m_\pi(\tilde{\boldsymbol{y}}|\hat{\sigma}_j^2)\} = n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p| + \tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{A}\tilde{\boldsymbol{y}}/\hat{\sigma}_j^2$$

with respect to the distribution of $\tilde{\boldsymbol{y}}|\boldsymbol{y} \sim \hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}}, \sigma^2\boldsymbol{V})$, we can rewrite $I_c(\boldsymbol{\eta}; m_\pi)$ in (7.13) as

$$\begin{aligned} I_c(\boldsymbol{\eta}; m_\pi) &= E\big[n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p| + \mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V})\cdot\sigma^2/\hat{\sigma}_j^2 \\ &\qquad + (\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}})/\hat{\sigma}_j^2\big] \\ &= E\big[n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p| + \mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V})\cdot\sigma^2/\hat{\sigma}_j^2 \\ &\qquad + \{\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})\boldsymbol{u}\}^{\mathrm{T}}\boldsymbol{A}\{\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})\boldsymbol{u}\}/\hat{\sigma}_j^2\big], \end{aligned}$$

noting that $\boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})\boldsymbol{u}$. Next, $-2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\}$ is rewritten as

$$\begin{aligned} &-2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\} \\ &= n\log(2\pi\hat{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\boldsymbol{I}_p| + (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})/\hat{\sigma}_j^2 \end{aligned}$$

Then the bias correction $\Delta_{c,\pi}$ is

$$\begin{aligned} \Delta_{c,\pi} &= E\left[\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V})\cdot\sigma^2/\hat{\sigma}_j^2 + \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1} - 2\boldsymbol{A}\boldsymbol{\Sigma}^{-1})\boldsymbol{u}/\hat{\sigma}_j^2\right] \\ &= J_1 + J_2 \quad \text{(say)}. \end{aligned}$$

Thus, it suffices to evaluate J_1 and J_2.

Noting that $n\hat{\sigma}_j^2/\sigma^2 \sim \chi^2_{n-p_j}$, we can evaluate J_1 as

$$J_1 = \frac{n \cdot \mathrm{tr}\,(\boldsymbol{AV})}{n-p_j-2} = -\frac{n \times \{\mathrm{tr}\,(\boldsymbol{A\Sigma}^{-1}) - 2\mathrm{tr}\,(\boldsymbol{A})\}}{n-p_j-2}$$

Next, J_2 is rewritten as

$$J_2 = n \cdot E\left[\frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A\Sigma}^{-1/2} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{A\Sigma}^{-1/2} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{A\Sigma}^{1/2})\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{v}}\right],$$

which can be evaluated by Lemma 4.6 as

$$\begin{aligned}
J_2 &= n \times \left\{\frac{\mathrm{tr}\,(\boldsymbol{A\Sigma}^{-1}) - 2\mathrm{tr}\,(\boldsymbol{A})}{n-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A\Sigma}^{-1/2}(\boldsymbol{I}_n - \boldsymbol{M}) - 2\boldsymbol{\Sigma}^{1/2}\boldsymbol{A\Sigma}^{-1/2}(\boldsymbol{I}_n - \boldsymbol{M})]}{(n-p_j)(n-p_j-2)}\right\} \\
&= n \times \left\{\frac{\mathrm{tr}\,(\boldsymbol{A\Sigma}^{-1}) - 2\mathrm{tr}\,(\boldsymbol{A})}{n-p_j-2} - \frac{2\mathrm{tr}\,[\boldsymbol{A}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})] - 4\mathrm{tr}\,[\boldsymbol{A}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{\Sigma}]}{(n-p_j)(n-p_j-2)}\right\} \\
&= n \times \left\{\frac{\mathrm{tr}\,(\boldsymbol{A\Sigma}^{-1}) - 2\mathrm{tr}\,(\boldsymbol{A})}{n-p_j} + \frac{2\mathrm{tr}\,(\boldsymbol{AP}) - 4\mathrm{tr}\,(\boldsymbol{AP\Sigma})}{(n-p_j)(n-p_j-2)}\right\}
\end{aligned}$$

Thus we can obtain

$$\Delta_{c,\pi} = n \times \left\{\frac{-2\mathrm{tr}\,[\boldsymbol{A}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})] + 4\mathrm{tr}\,[\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{P\Sigma})]}{(n-p_j)(n-p_j-2)}\right\},$$

and propose the following information criterion:

$$\mathrm{IC}_{c,\pi} = -2\log\{m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)\} + \Delta_{c,\pi}. \tag{7.14}$$

Theorem 7.3 *The information criterion* $\mathrm{IC}_{c,\pi}$ *in* (7.14) *is an unbiased estimator of* $I_c(\boldsymbol{\eta}; m_\pi)$ *in* (7.13)*, namely* $E(\mathrm{IC}_{c,\pi}) = I_c(\boldsymbol{\eta}; m_\pi)$*.*

7.4.3 Conditional RIC

In this subsection, we assume the uniform prior for $\boldsymbol{\beta}$, namely $\boldsymbol{\beta} \sim \mathrm{uniform}(\mathbb{R}^p)$. Although this is improper prior distribution, we can obtain the marginal likelihood function formally as follows:

$$\begin{aligned}
m_r(\boldsymbol{y}|\sigma^2) &= \int m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\mathrm{d}\boldsymbol{\beta} \\
&= (2\pi\sigma^2)^{-(n-p)/2} \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2} \cdot \exp\{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/(2\sigma^2)\},
\end{aligned}$$

which is the same as the residual likelihood (Patterson and Thompson, 1971). In the last chapter, we measured the risk of predictive density $m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)$ in terms of the following KL divergence:

$$R(\boldsymbol{\eta}; m_r) = \int\left[\int \log\left\{\frac{m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2)}{m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)}\right\} m(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2)\mathrm{d}\tilde{\boldsymbol{y}}\right] m(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\mathrm{d}\boldsymbol{y},$$

and showed that the resulting criterion is identical to residual information criterion (RIC) proposed by Shi and Tsai (2002). Azari et al. (2006) applied the RIC to variable selection in linear mixed model for longitudinal data analysis. However, the KL risk $R(\boldsymbol{\eta}; m_r)$ is not appropriate

when one is interested in predicting random effects, which was pointed out by Vaida and Blanchard (2005). Then we propose to measure the prediction risk of $m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)$ by the conditional KL risk $R_c(\boldsymbol{\eta};\hat{f}_j)$ in (7.3) for $\hat{f}_j = m_r(\boldsymbol{y}|\tilde{\sigma}_j)$ and call the resulting criterion the conditional RIC (cRIC).

Then we consider the following information:

$$\begin{aligned} I_c(\boldsymbol{\eta}; m_r) &= \iiint -2\log\{m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) p(\boldsymbol{b}|\sigma^2) \mathrm{d}\tilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{b} \\ &= \int \left[\int -2\log\{m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)\} \hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\mathrm{d}\boldsymbol{y} \\ &= \mathrm{cRI} \quad \text{(say)}, \end{aligned} \tag{7.15}$$

where $\tilde{\sigma}_j^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/(n-p_j)$ is the residual maximum likelihood (REML) estimator of σ_j^2. We want to construct an information criterion as an unbiased estimator of $I_c(\boldsymbol{\eta}; m_r)$, which is of the form

$$\mathrm{cRIC} = -2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}_j^2)\} + \Delta_{\mathrm{cRI}},$$

where

$$\begin{aligned} -2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}_j^2)\} &= (n-p_j)\log(2\pi\tilde{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{y}/\tilde{\sigma}_j^2, \\ \Delta_{\mathrm{cRI}} &= \mathrm{cRI} - E[-2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}_j^2)\}]. \end{aligned}$$

Note that the expectation in the equation above is the one with respect to the distribution of $\boldsymbol{y} \sim m(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)$. Then we have to evaluate the bias correction Δ_{cRI}.

Firstly, taking expectation of

$$\begin{aligned} &-2\log\{m_r(\tilde{\boldsymbol{y}}|\tilde{\sigma}_j^2)\} \\ &= (n-p_j)\log(2\pi\tilde{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \tilde{\boldsymbol{y}}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\tilde{\boldsymbol{y}}/\tilde{\sigma}_j^2, \end{aligned}$$

with respect to the distribution of $\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\eta}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}}, \sigma^2\boldsymbol{V})$, we can rewrite the cRI in (7.15) as

$$\begin{aligned} \mathrm{cRI} &= E\Big[(n-p_j)\log(2\pi\tilde{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \mathrm{tr}\,[\boldsymbol{V}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})]\cdot\sigma^2/\tilde{\sigma}_j^2 \\ &\qquad + \{\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})\boldsymbol{u}\}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\{\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})\boldsymbol{u}\}/\tilde{\sigma}_j^2\Big] \\ &= E\Big[(n-p_j)\log(2\pi\tilde{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + \mathrm{tr}\,[\boldsymbol{V}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})]\cdot\sigma^2/\tilde{\sigma}_j^2 \\ &\qquad + \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})\boldsymbol{u}/\tilde{\sigma}_j^2\Big], \end{aligned}$$

noting that $(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$. From the fact that

$$\begin{aligned} &-2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}_j^2)\} \\ &= (n-p_j)\log(2\pi\tilde{\sigma}_j^2) + \log|\boldsymbol{\Sigma}| + \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| + (n-p_j), \end{aligned}$$

the bias correction can be reduced to

$$\begin{aligned} \Delta_{\mathrm{cRI}} &= E\Big[\mathrm{tr}\,[\boldsymbol{V}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})]\cdot\sigma^2/\tilde{\sigma}_j^2 + \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}^{-1} - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1})\boldsymbol{u}/\tilde{\sigma}_j^2\Big] - (n-p_j) \\ &= J_3 + J_4 - (n-p_j), \quad \text{(say)}. \end{aligned}$$

Thus, it suffices to evaluate J_3 and J_4.

It is easy to see that

$$J_3 = \frac{(n-p_j)\cdot \mathrm{tr}\,[\boldsymbol{V}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})]}{n-p_j-2} = \frac{(n-p_j)\cdot \mathrm{tr}\,[(2\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})]}{n-p_j-2},$$

because $(n-p_j)\tilde{\sigma}_j^2/\sigma^2 \sim \chi^2_{n-p_j}$. Next, J_4 is rewritten as

$$J_4 = (n-p_j)\cdot E\left[\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{v}}\right],$$

which can be evaluated by Lemma 4.6 as

$$\begin{aligned}
J_4 &= (n-p_j) \\
&\quad\times\left\{\frac{\mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})]}{n-p_j-2} - \frac{2\mathrm{tr}\,[(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})(\boldsymbol{I}_n-\boldsymbol{M})(\boldsymbol{I}_n-\boldsymbol{\Sigma}^{-1})(\boldsymbol{I}_n-\boldsymbol{M})]}{(n-p_j)(n-p_j-2)}\right\} \\
&= (n-p_j) \\
&\quad\times\left\{\frac{\mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{\Sigma}-2\boldsymbol{I}_n+\boldsymbol{\Sigma}^{-1})]}{n-p_j-2} - \frac{2\mathrm{tr}\,[(\boldsymbol{\Sigma}-\boldsymbol{I}_n)(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{\Sigma}-\boldsymbol{I}_n)(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})]}{(n-p_j)(n-p_j-2)}\right\}.
\end{aligned}$$

Thus we can obtain

$$\Delta_{\mathrm{cRI}} = \frac{2(n-p_j)}{n-p_j-2} - \frac{2}{n-p_j-2}\mathrm{tr}\,[(\boldsymbol{\Sigma}-\boldsymbol{I}_n)(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})(\boldsymbol{\Sigma}-\boldsymbol{I}_n)(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})]$$

and propose the following cRIC:

$$\mathrm{cRIC} = -2\log\{m_r(\boldsymbol{y}|\tilde{\sigma}_j^2)\} + \Delta_{\mathrm{cRI}}. \tag{7.16}$$

Theorem 7.4 *The* cRIC *in* (7.16) *is an unbiased estimator of* cRI *in* (7.15)*, namely* $E(\mathrm{cRIC}) = \mathrm{cRI}$.

7.5 Simulations

In this section, we compare the numerical performance of the proposed criteria, PIC, PIC_2, $\mathrm{IC}_{c,\pi}$ and cRIC with the conventional cAIC of Vaida and Blanchard (2005). We handle the nested error regression model (NERM) and consider the same setting as that of Section 6.4. When we derive the criteria PIC_2 and $\mathrm{IC}_{c,\pi}$, we set the prior distribution of $\boldsymbol{\beta}$ as $\mathcal{N}_p(\boldsymbol{0}, \sigma^2\lambda^{-1}\boldsymbol{I}_p)$, namely $\boldsymbol{W} = \lambda^{-1}\boldsymbol{I}_p$. The hyperparameter λ is estimated by maximizing the marginal likelihood $m_\pi(\boldsymbol{y}|\hat{\sigma}_j^2)$, where the estimate $\hat{\sigma}_j^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1}-\boldsymbol{P})\boldsymbol{y}/n$ of σ^2 is plugged in. The unknown parameter $\phi = \tau^2/\sigma^2$ included by $\boldsymbol{\Sigma}$ is estimated by consistent estimator based on the full model in the same way as Section 6.4. The class of the candidate models includes all the subsets of the full model and select the model by the criteria. The performance of the criteria is measured by the number of selecting the true model and the prediction error of the selected model based on quadratic loss, namely $\|\boldsymbol{X}(\hat{\jmath})\widehat{\boldsymbol{\beta}}_{\hat{\jmath}} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*\|^2/n$.

Table 7.1 and 7.2 show the number of selecting the true model by the criteria and the average prediction error of the selected model by each criterion, respectively. From the table we can see the following facts. Firstly, the number of selecting the true model approaches 1000 for $\mathrm{IC}_{c,\pi}$ and cRIC, which is the numerical evidence of the consistency of the criteria. Especially, $\mathrm{IC}_{c,\pi}$ performs well for almost all situations in terms of prediction error as well as selecting the true model. However, for the case of small sample size and noisy data, namely signal-to-noise ratio (SNR) is small, PIC_2 performs the best among the criteria.

Table 7.1: The number of selecting the true model by the criteria in 1000 realizations

		$\phi = 0.5$			$\phi = 1$			$\phi = 2$		
	SNR	1	3	5	1	3	5	1	3	5
$n_0 = 5$	cAIC	121	718	732	140	694	729	175	677	740
$m = 4$	PIC	124	706	715	157	683	703	194	673	713
	PIC_2	155	555	558	171	550	557	201	539	559
	$\text{IC}_{c,\pi}$	146	850	905	176	849	913	237	849	937
	cRIC	20	531	730	41	590	758	96	679	804
$n_0 = 5$	cAIC	355	631	631	357	645	645	337	652	653
$m = 8$	PIC	364	602	602	359	593	593	359	601	601
	PIC_2	330	479	479	322	487	489	330	495	496
	$\text{IC}_{c,\pi}$	427	894	938	448	918	952	489	934	960
	cRIC	153	702	850	234	754	863	358	802	895
$n_0 = 5$	cAIC	576	625	625	542	630	630	492	629	629
$m = 16$	PIC	561	596	596	535	595	595	507	596	596
	PIC_2	475	509	511	452	505	507	438	514	515
	$\text{IC}_{c,\pi}$	736	938	958	720	947	963	699	955	971
	cRIC	436	832	900	498	856	921	578	893	935

Table 7.2: The prediction error of the best model selected by the criteria

		$\phi = 0.5$			$\phi = 1$			$\phi = 2$		
	SNR	1	3	5	1	3	5	1	3	5
$n_0 = 5$	cAIC	1.44	0.120	0.0430	1.11	0.0962	0.0343	0.708	0.0652	0.0233
$m = 4$	PIC	1.43	0.121	0.0433	1.10	0.0968	0.0346	0.698	0.0658	0.0235
	PIC_2	1.35	0.127	0.0455	1.05	0.101	0.0362	0.675	0.0689	0.0247
	$\text{IC}_{c,\pi}$	1.41	0.115	0.0400	1.09	0.0915	0.0320	0.705	0.0615	0.0214
	cRIC	1.28	0.127	0.0431	1.01	0.100	0.0341	0.679	0.0664	0.0229
$n_0 = 5$	cAIC	0.846	0.0854	0.0307	0.638	0.0672	0.0242	0.418	0.0455	0.0164
$m = 8$	PIC	0.845	0.0859	0.0309	0.638	0.0679	0.0244	0.422	0.0460	0.0166
	PIC_2	0.833	0.0880	0.0317	0.644	0.0695	0.0250	0.429	0.0472	0.0170
	$\text{IC}_{c,\pi}$	0.842	0.0793	0.0280	0.628	0.0621	0.0220	0.405	0.0418	0.0149
	cRIC	0.840	0.0840	0.0289	0.655	0.0656	0.0228	0.437	0.0439	0.0153
$n_0 = 5$	cAIC	0.564	0.0622	0.0224	0.459	0.0508	0.0183	0.322	0.0356	0.0128
$m = 16$	PIC	0.566	0.0625	0.0225	0.461	0.0510	0.0184	0.323	0.0358	0.0129
	PIC_2	0.572	0.0633	0.0228	0.466	0.0516	0.0186	0.327	0.0362	0.0130
	$\text{IC}_{c,\pi}$	0.547	0.0582	0.0208	0.443	0.0476	0.0170	0.308	0.0334	0.0120
	cRIC	0.577	0.0599	0.0212	0.466	0.0488	0.0173	0.323	0.0340	0.0121

7.6 Proofs

7.6.1 Proof of Proposition 7.1

The marginal joint distribution of $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ is

$$\begin{pmatrix}\tilde{\boldsymbol{y}}\\ \boldsymbol{y}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{X}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12}\\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{pmatrix}\right),$$

where $\boldsymbol{\Lambda}_{11} = \boldsymbol{\Lambda}_{22} = \sigma^2\boldsymbol{\Sigma}$,

$$\begin{aligned}&\boldsymbol{\Lambda}_{12} = \boldsymbol{\Lambda}_{21}\\ &= E[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}]\\ &= E[(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon})(\boldsymbol{Z}\boldsymbol{b} + \tilde{\boldsymbol{\varepsilon}})^{\mathrm{T}}]\\ &= \sigma^2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}},\end{aligned}$$

and $\tilde{\boldsymbol{\varepsilon}}$ is independent replication of $\boldsymbol{\varepsilon}$, which is also independent of $\boldsymbol{b}$. Then, from the property of multivariate normal distribution, it follows that

$$\hat{f}^{\mathrm{BP}}(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}),$$

where $\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Z}\hat{\boldsymbol{b}}^{\mathrm{B}}$ and

$$\begin{aligned}\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21} &= \sigma^2\boldsymbol{\Sigma} - \sigma^2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\\ &= \sigma^2\boldsymbol{\Sigma} - \sigma^2(\boldsymbol{\Sigma} - \boldsymbol{I}_n)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{I}_n)\\ &= \sigma^2(2\boldsymbol{I}_n - \boldsymbol{\Sigma}^{-1}) = \sigma^2\boldsymbol{V},\end{aligned}$$

which shows Proposition 7.1. □

7.6.2 Proof of Proposition 7.2

The marginal joint distribution of $\boldsymbol{z}$ and $\boldsymbol{y}$ is

$$\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{y}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12}\\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22}\end{pmatrix}\right),$$

where $\boldsymbol{\Gamma}_{11} = \boldsymbol{\Gamma}_{22} = \sigma^2(\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{I}_n) = \sigma^2(\boldsymbol{C} + \boldsymbol{I}_n)$,

$$\begin{aligned}&\boldsymbol{\Gamma}_{12} = \boldsymbol{\Gamma}_{21}\\ &= E[\boldsymbol{y}\boldsymbol{z}^{\mathrm{T}}]\\ &= \sigma^2\boldsymbol{C},\end{aligned}$$

and $\boldsymbol{C} = \boldsymbol{B} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}$. Then, form the property of multivariate normal distribution, it follows that

$$\hat{g}^{\mathrm{BP}} = \mathcal{N}(\boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{y}, \boldsymbol{\Gamma}_{11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{21}),$$

where

$$\begin{aligned}\boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1} &= \boldsymbol{C}(\boldsymbol{C} + \boldsymbol{I}_n)^{-1}\\ &= (\boldsymbol{C} + \boldsymbol{I}_n - \boldsymbol{I}_n)(\boldsymbol{C} + \boldsymbol{I}_n)^{-1}\\ &= \boldsymbol{I}_n - (\boldsymbol{C} + \boldsymbol{I}_n)^{-1}\\ &= \boldsymbol{I}_n - (\boldsymbol{B} + \boldsymbol{\Sigma})^{-1},\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Gamma}_{11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{21} &= \sigma^2(\boldsymbol{C} + \boldsymbol{I}_n) - \sigma^2\boldsymbol{C}(\boldsymbol{C} + \boldsymbol{I}_n)^{-1}\boldsymbol{C} \\ &= \sigma^2(\boldsymbol{C} + \boldsymbol{I}_n) - \sigma^2(\boldsymbol{C} + \boldsymbol{I}_n - \boldsymbol{I}_n)(\boldsymbol{C} + \boldsymbol{I}_n)^{-1}(\boldsymbol{C} + \boldsymbol{I}_n - \boldsymbol{I}_n) \\ &= \sigma^2\{2\boldsymbol{I}_n - (\boldsymbol{C} + \boldsymbol{I}_n)^{-1}\} \\ &= \sigma^2\{2\boldsymbol{I}_n - (\boldsymbol{B} + \boldsymbol{\Sigma})^{-1}\} = \sigma^2\boldsymbol{V}_2,\end{aligned}$$

which shows Proposition 7.2.

Acknowledgment

I thank my supervisor Prof. Tatsuya Kubokawa for all of his help and lots of encouragement to me during the master's and doctoral courses. I also thank Prof. Malay Ghosh, Prof. Muni Srivastava and Prof. J.N.K. Rao for their valuable comments and suggestions. Prof. Naoto Kunitomo, Prof. Yoshihiro Yajima, Prof. Yasuhiro Omori, Prof. Katsumi Shimotsu, Prof. Akimichi Takemura, Prof. Fumiyasu Komaki, Dr. Yuzo Maruyama and Dr. Kengo Kato gave me many helpful comments about my research. Mr. Shinichiro Shirota, Mr. Masaaki Imaizumi and Mr. Shonosuke Sugasawa made my student life very meaningful and enjoyable in various forms. I am also grateful for the discussion with Dr. Genya Kobayashi, Dr. Kota Ogawasara and all of my academic colleagues. I was financially supported by JSPS. Lastly, I thank my father, mother, sister, grand parents and their families for their support.

Bibliography

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

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.

Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Akaike, H. (1980a). On the use of predictive likelihood of a Gaussian model. *Annals of the Institute of Statistical Mathematics*, **32**, 311–324.

Akaike, H. (1980b). Likelihood and the Bayes procedure. In *Bayesian Statistics*, (N.J. Bernard, M.H. Degroot, D.V. Lindaley and A.F.M. Simith, eds.), Valencia, Spain, University Press, 141–166.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**, 443–458.

Azari, R., Li, L. and Tsai, C.-L. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis*, **50**, 3053–3066.

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. *Journal of the American Statistical Associaton*, **83**, 28–36.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Associaton*, **91**, 109–122.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

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

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Associaton*, **101**, 157–167.

Datta, G. and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science*, **27**, 95–114.

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.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

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

Ghosh, M. and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, **91**, 95–112.

Ghosh, M. and Maiti, T. (2008). Empirical Bayes confidence intervals for means of natural exponential family-quadratic variance function distributions with application to small area estimation. *Scandinavian Journal of Statistics*, **35**, 484–495.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference*, **22**, 137–152.

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

Henderson, C.R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, **21**, 309–310.

Hodges, J.S. and Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and the richly-parameterised models. *Biometrika*, **88**, 367–379.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, **36**, 1669–1692.

Hansen, B.E. (2007). Least squares model averaging. *Econometrica*, **75**, 1175–1189.

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kawakubo, Y. and Kubokawa, T. (2014). Modified conditional AIC in linear mixed models. *Journal of Multivariate Analysis*, **129**, 44–56.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics — Theory and Methods*, **26**, 2223–2246.

Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed model. *Journal of Multivariate Analysis*, **102**, 641–660.

Kubokawa, T., Hasukawa, M. and Takahashi, K. (2014). On measuring uncertainty of benchmarked predictors with application to disease risk estimate. *Scandinavian Journal of Statistics*, **41**, 394–413.

Kubokawa, T. and Nagashima, B. (2012). Parametric Bootstrap methods for bias correction in linear mixed models. *Journal of Multivariate Analysis*, **106**, 1–16.

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

Lohr, S.L. and Rao, J.N.K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457–468.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, **10**, 65–80.

Morris, C. (1983). Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, **11**, 515–529.

Müller, S., Scealy, J.L. and Welsh, A.H. (2013). Model selection in linear mixed models. *Statistical Science*, **28**, 135–167.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley.

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. *Electronic Journal of Statistics*, **8**, 201–225.

Satoh, K. (1997). AIC-type model selection criterion for multivariate linear regression with a future experiment. *Journal of the Japan Statistical Society*, **27**, 135–140.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221–264.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society series B*, **64**, 237–252.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**, 227–244.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society series B*, **64**, 583–639.

Srivastava, M.S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**, 1970–1980.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics — Theory and Methods*, **7**, 13–26.

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

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**, 439–447.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Yu, D. and Yau, K.K.W. (2012). Conditional Akaike information criterion of generalized linear mixed models. *Computational Statistics and Data Analysis*, **56**, 629–644.

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. *Journal of Multivariate Analysis*, **116**, 245–262.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (P.K. Goel and A. Zellner, eds.), pp. 233–243, Amsterdam: North-Holland/Elsevier.

Zhang, X., Zou, G. and Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, **101**, 205-218.