CIRJE-F-971

A Variant of AIC Using Bayesian Marginal Likelihood

Yuki Kawakubo
Graduate School of Economics, The University of Tokyo

Tatsuya Kubokawa
The University of Tokyo

Muni S. Srivastava
University of Toronto

April 2015

A Variant of AIC Using Bayesian Marginal Likelihood

Yuki Kawakubo[*], Tatsuya Kubokawa[†] and Muni S. Srivastava[‡]

April 27, 2015

Abstract

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.

Key words and phrases: AIC, BIC, consistency, Kullback–Leibler divergence, linear regression model, residual information criterion, variable selection.

1 Introduction

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

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

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

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

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 paper, 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 $f(\boldsymbol{y}|\boldsymbol{\omega})$ for a vector of unknown parameters $\boldsymbol{\omega}$. Let $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ is a predictive density for $f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})$, where $\widetilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$. We here evaluate the predictive performance of $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ in terms of the following risk:

$$R(\boldsymbol{\omega};\hat{f}) = \int \left[\int \log \left\{ \frac{f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega})}{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})} \right\} f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega}) \mathrm{d}\widetilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{\omega}) \mathrm{d}\boldsymbol{y}. \tag{1.1}$$

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

$$\begin{aligned} I(\boldsymbol{\omega};\hat{f}) =& \iint -2\log\{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})\} f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega}) f(\boldsymbol{y}|\boldsymbol{\omega}) \mathrm{d}\widetilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y} \\ =& E_{\boldsymbol{\omega}} \left[-2\log\{\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})\} \right], \end{aligned} \tag{1.2}$$

which is a part of (1.1) (multiplied by 2), where $E_{\boldsymbol{\omega}}$ denotes the expectation with respect to the distribution of $f(\widetilde{\boldsymbol{y}},\boldsymbol{y}|\boldsymbol{\omega}) = f(\widetilde{\boldsymbol{y}}|\boldsymbol{\omega}) f(\boldsymbol{y}|\boldsymbol{\omega})$. Let $\Delta = I(\boldsymbol{\omega};\hat{f}) - E_{\boldsymbol{\omega}}[-2\log\{\hat{f}(\boldsymbol{y};\boldsymbol{y})\}]$. Then, the AIC variant based on the predictor $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ is defined by

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

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

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

(AIC) Put $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y}) = f(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\omega}})$ for the maximum likelihood estimator $\widehat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$. Then, $\mathrm{IC}(f(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\omega}}))$ 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 $-2\log\{f(\boldsymbol{y}|\widehat{\boldsymbol{\omega}})\} + 2\dim(\boldsymbol{\omega})$.

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

The criterion $\mathrm{IC}(\hat{f})$ 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 $\boldsymbol{\omega}$ is divided as $\boldsymbol{\omega} = (\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)^t$ for a p-dimensional parameter vector of interest $\boldsymbol{\beta}$ and a q-dimensional nuisance parameter vector $\boldsymbol{\theta}$. We assume that $\boldsymbol{\beta}$ has a prior density $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})$ with hyperparameter $\boldsymbol{\lambda}$. The model is described as

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

and $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are estimated by data. Inference based on such a model is called an empirical Bayes procedure. Put $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) = f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}}) = \int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}})\pi(\boldsymbol{\beta}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\mathrm{d}\boldsymbol{\beta}$ for some estimators $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\theta}}$. Then, the information in (1.2) is

$$I(\boldsymbol{\omega}; f_\pi) = \iint -2\log\{f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\} f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\mathrm{d}\widetilde{\boldsymbol{y}}\mathrm{d}\boldsymbol{y}, \tag{1.3}$$

and the resulting information criterion is

$$\mathrm{IC}(f_\pi) = -2\log\{f_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})\} + \widehat{\Delta}, \tag{1.4}$$

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

There are three motivations to consider the information $I(\boldsymbol{\omega}; f_\pi)$ in (1.3) and the information criterion $\mathrm{IC}(f_\pi)$ in (1.4).

Firstly, it is noted that the Bayesian predictor $f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\theta}})$ is evaluated by the risk $R(\boldsymbol{\omega}; f_\pi)$ in (1.1), which is based on a frequentist point of view. On the other hand, the Bayesian risk is

$$r(\boldsymbol{\psi}; \hat{f}) = \int R(\boldsymbol{\omega}; \hat{f})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}, \tag{1.5}$$

which measures the prediction error of $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y})$ under the assumption that the prior information is correct, where $\boldsymbol{\psi} = (\boldsymbol{\lambda}^t, \boldsymbol{\theta}^t)^t$. 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 $R(\boldsymbol{\omega}; f_\pi)$ can measure the prediction error of the Bayesian model from a standpoint of frequentists, however, the resulting criterion $\mathrm{IC}(f_\pi)$ is less influenced by the prior misspecification.

Secondly, we can construct the information criterion $\mathrm{IC}(f_\pi)$ when the prior distribution of $\boldsymbol{\beta}$ is improper, since the information $I(\boldsymbol{\omega}; f_\pi)$ in (1.3) can be defined formally for the corresponding improper marginal likelihood. Because the Bayesian risk $r(\boldsymbol{\psi}; f_\pi)$ 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 $\mathrm{IC}(f_\pi)$ 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 (1.4) 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 the paper is organized as follows. In Section 2, we obtain the information criterion (1.4) in linear regression model with general covariance structure and compare it with other related criteria. In Section 3, we prove the consistency of the criteria. In Section 4, we investigate the performance of the criteria through simulations. Section 5 concludes the paper with some discussions.

2 Proposed Criteria

2.1 Variable selection criteria for linear regression model

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of the response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of the explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the random errors. Throughout the paper, we assume that $\boldsymbol{X}$ has full column rank p. Here, the random error $\boldsymbol{\varepsilon}$ has the distribution $\mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{V})$, where σ^2 is an unknown scalar and $\boldsymbol{V}$ is a known positive definite matrix. We consider the problem of selecting the explanatory variables and assume that the true model is included in the family of the candidate models, that is the common assumption to obtain the criterion.

We shall construct the variable selection criteria for the regression model (2.1) which is of the form (1.4). 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 $\boldsymbol{W} = (\lambda\boldsymbol{X}^t\boldsymbol{X})^{-1}$ for $\lambda > 0$ when $\boldsymbol{V}$ is identity matrix, which is introduced by Zellner (1986), or more simply $\boldsymbol{W} = \lambda^{-1}\boldsymbol{I}_p$. Because the likelihood is $f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V})$, the marginal likelihood function is

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

where $\boldsymbol{A} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$. Note that $\boldsymbol{A} = (\boldsymbol{V} + \boldsymbol{B})^{-1}$ for $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^t$, namely $f_\pi(\boldsymbol{y}|\sigma^2) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{V} + \boldsymbol{B}))$. Then we take the predictive density as $\hat{f}(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) = f_\pi(\widetilde{\boldsymbol{y}}|\hat{\sigma}^2)$ and the information (1.3) can be written as

$$I_{\pi,1}(\boldsymbol{\omega}) = E_{\boldsymbol{\omega}}\left[n\log(2\pi\hat{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \widetilde{\boldsymbol{y}}^t\boldsymbol{A}\widetilde{\boldsymbol{y}}/\hat{\sigma}^2\right], \tag{2.2}$$

where $\hat{\sigma}^2 = \boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}/n$ and $E_{\boldsymbol{\omega}}$ denotes the expectation with respect to the distribution of $f(\widetilde{\boldsymbol{y}}, \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2)f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ for $\boldsymbol{\omega} = (\boldsymbol{\beta}^t, \sigma^2)^t$. Note that $\boldsymbol{\beta}$ is the parameter of interest and σ^2 is the nuisance parameter, which corresponds to $\boldsymbol{\theta}$ in the previous section. Then we propose the information criterion.

Proposition 2.1 *The information $I_{\pi,1}(\boldsymbol{\omega})$ in* (2.2) *is unbiasedly estimated by the information criterion*

$$\mathrm{IC}_{\pi,1} = -2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} + \frac{2n}{n-p-2}, \tag{2.3}$$

where

$$-2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\} = n\log\hat{\sigma}^2 + \log|\boldsymbol{V}| + \log|\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{I}_p| + \boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2 + (\text{const}),$$

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

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

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

when n is large.

Alternatively, the KL risk $r(\boldsymbol{\psi}; \hat{f})$ in (1.5) can be also used for evaluating the risk of the predictive density $f_\pi(\widetilde{\boldsymbol{y}}|\hat{\sigma}^2)$, since the prior distribution is proper. The resulting criterion is

$$\mathrm{IC}_{\pi,2} = n\log\hat{\sigma}^2 + \log|\boldsymbol{V}| + p\log n + p, \tag{2.4}$$

which is an asymptotically unbiased estimator of $I_{\pi,2}(\sigma^2) = E_\pi[I_{\pi,1}(\boldsymbol{\omega})]$ up to constant where E_π denotes the expectation with respect to the prior distribution $\pi(\boldsymbol{\beta}|\sigma^2)$, namely $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,2}) \to I_{\pi,2}(\sigma^2) + (\text{const})$ 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 $\boldsymbol{\beta} \sim uniform(\mathbb{R}^p)$. Though this is improper prior distribution, we can obtain the marginal likelihood function formally:

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

which is known as the residual likelihood (Patterson and Thompson, 1971), where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$. Then we take the predictive density as $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y}) = f_r(\widetilde{\boldsymbol{y}}|\tilde{\sigma}^2)$ and the information (1.3) can be written as

$$I_r(\boldsymbol{\omega}) = E_{\boldsymbol{\omega}}\left[(n-p)\log(2\pi\tilde{\sigma}^2) + \log|\boldsymbol{V}| + \log|\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}| + \widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}/\tilde{\sigma}^2\right], \qquad (2.5)$$

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

Proposition 2.2 *The information $I_r(\boldsymbol{\omega})$ in* (2.5) *is unbiasedly estimated by the infomation criterion*

$$\mathrm{IC}_r = -2\log\{f_r(\boldsymbol{y}|\tilde{\sigma}^2)\} + \frac{2(n-p)}{n-p-2}, \qquad (2.6)$$

where

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

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

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

$$\mathrm{IC}_r^* = (n-p)\log\tilde{\sigma}^2 + \log|\boldsymbol{V}| + p\log n + \frac{(n-p)^2}{n-p-2}, \qquad (2.7)$$

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 (2.7), which results in the RIC exactly. Note the criterion based on $f_r(\boldsymbol{y}|\sigma^2)$ and $r(\boldsymbol{\psi}; f_r)$ cannot be constructed because the KL risk of it diverges to infinity.

2.2 Extension to the case of unknown covariance

In the derivation of the criteria, we have assumed that the scaled covariance matrix $\boldsymbol{V}$ of the error terms vector are known. However, it is often the case that $\boldsymbol{V}$ is unknown and is some function of the unknown parameter $\boldsymbol{\phi}$, namely $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\phi})$. In that case, $\boldsymbol{V}$ in each criterion is replaced with its plug-in estimator $\boldsymbol{V}(\widehat{\boldsymbol{\phi}})$, 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 $\boldsymbol{V}$, $\boldsymbol{W}$ 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 (2.1), 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 $\boldsymbol{V} = \boldsymbol{I}_n$, (2.1) represents a multiple regression model with constant variance. In this model, the scaled covariance matrix $\boldsymbol{V}$ does not contain any unknown parameters.

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_2\boldsymbol{v}_2 + \cdots + \boldsymbol{Z}_r\boldsymbol{v}_r + \boldsymbol{\eta}, \tag{2.8}$$

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

$$y_{ij} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \eta_{ij}, \quad (i = 1, \ldots, m;\ j = 1, \ldots, n_i), \tag{2.9}$$

where v_i's and η_{ij}'s are mutually independently distributed as $v_i \sim \mathcal{N}(0, \tau^2)$ and $\eta_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $n = \sum_{i=1}^m n_i$. Note that the NERM in (2.9) is given by $\theta_1 = \sigma^2,\ \theta_2 = \tau^2,\ \boldsymbol{V}_1 = \boldsymbol{I}_n$ and $\boldsymbol{Z}_2 = \mathrm{diag}\,(\boldsymbol{j}_{n_1}, \ldots, \boldsymbol{j}_{n_m})$, where $\boldsymbol{j}_k$ is the k-dimensional vector of ones, for variance components model (2.8). This model is often used for the clustered data and v_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 $\boldsymbol{\varepsilon} = \boldsymbol{Z}_2\boldsymbol{v}_2 + \boldsymbol{\eta}$, $\boldsymbol{V} = \boldsymbol{V}(\phi) = \phi\boldsymbol{V}_2 + \boldsymbol{I}_n$ for (2.1), where $\phi = \tau^2/\sigma^2$ and $\boldsymbol{V}_2 = \boldsymbol{Z}_2\boldsymbol{Z}_2^t = \mathrm{diag}\,(\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{J}_{n_m})$ for $\boldsymbol{J}_k = \boldsymbol{j}_k\boldsymbol{j}_k^t$. In that case, our proposed variable selection procedure is valid.

[3] **regression model with autoregressive moving average errors**. Consider the regression model (2.1), assuming the random errors are generated by an ARMA(q, r) process defined by

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

where $\{u_i\}$ 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 $\varepsilon_1 \sim \mathcal{N}(0, \tau^2/(1-\phi^2))$, $\varepsilon_i = \phi\varepsilon_{i-1} + u_i$, $u_i \sim \mathcal{N}(0, \tau^2)$ for $i = 2, 3, \ldots, n$. When we define $\sigma^2 = \tau^2/(1-\phi^2)$, (i, j)-element of the scaled covariance matrix $\boldsymbol{V}$ in (2.1) is $\phi^{|i-j|}$.

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.

To discuss the consistency, we define the class of the candidate models and the true model more formally. Let $n \times p$ matrix $\boldsymbol{X}$ consist of all the explanatory variables and assume that rank $(\boldsymbol{X}) = p$. To define the candidate model by the index j, suppose that j denotes a subset of $\omega = \{1, \ldots, p\}$ containing p_j elements, namely $p_j = \#(j)$, and $\boldsymbol{X}_j$ consists of p_j columns of $\boldsymbol{X}$ indexed by the elements of j. Note that $\boldsymbol{X}_\omega = \boldsymbol{X}$ and $p_\omega = p$. We define the index set by $\mathcal{J} = \mathcal{P}(\omega)$, namely the power set of ω. Then the model j is

$$\boldsymbol{y} = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2 \boldsymbol{V}),$$

where $\boldsymbol{X}_j$ is $n \times p_j$ and $\boldsymbol{\beta}_j$ is $p_j \times 1$. The prior distribution of $\boldsymbol{\beta}_j$ is $\boldsymbol{\beta}_j \sim \mathcal{N}_{p_j}(\boldsymbol{0}, \sigma^2 \boldsymbol{W}_j)$. The true model is defined by j_0 and $\boldsymbol{X}_{j0}\boldsymbol{\beta}_{j_0}$ is abbreviated to $\boldsymbol{X}_0\boldsymbol{\beta}_0$, which is the true mean of $\boldsymbol{y}$. $\mathcal{J}$ is the collection of all the candidate models and divide $\mathcal{J}$ into two subsets $\mathcal{J}_+$ and $\mathcal{J}_-$, where $\mathcal{J}_+ = \{j \in \mathcal{J} : j_0 \subseteq j\}$ and $\mathcal{J}_- = \mathcal{J} \setminus \mathcal{J}_+$. Note that the true model j_0 is the smallest model in $\mathcal{J}_+$. Let $\hat{\jmath}$ denote the model selected by some criterion. Following Shi and Tsai (2002), we make the assumptions.

(A1) $E(\varepsilon_1^{4s}) < \infty$ for some positive integer s.

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

(A3) $\liminf_{n\to\infty} n^{-1} \inf_{j\in\mathcal{J}_-} \|\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{H}_j \boldsymbol{X}_0 \boldsymbol{\beta}_0\|^2 > 0$, where $\boldsymbol{H}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t$.

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

Theorem 1 *If assumptions* (A1)–(A3) *are satisfied,* $\mathcal{J}_+$ *is not empty, the* ε_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* $P(\hat{\jmath} = j_0) \to 1$ *as* $n \to \infty$.

The proof of Theorem 1 is given in Appendix B.

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, $\boldsymbol{V}$ and $\boldsymbol{W}_j$ are replaced with their plug-in estimators $\boldsymbol{V}(\widehat{\boldsymbol{\phi}})$ and $\boldsymbol{W}_j(\widehat{\boldsymbol{\lambda}}_j)$, respectively.

Theorem 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* $\boldsymbol{V}(\boldsymbol{\phi})$ *and* $\boldsymbol{W}_j(\boldsymbol{\lambda}_j)$ *are continuous functions of* $\boldsymbol{\phi}$ *and* $\boldsymbol{\lambda}_j$, *and* $\boldsymbol{V}(\boldsymbol{\phi})$ *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* $\boldsymbol{\varepsilon}$ *are replaced with* $\boldsymbol{V}^{-1/2}\boldsymbol{X}_j$ *and* $\boldsymbol{\varepsilon}^* = \boldsymbol{V}^{-1/2}\boldsymbol{\varepsilon}$ *respectively,* $\mathcal{J}_+$ *is not empty and the* ε_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 2, we can use the same techniques as those for the proof of Theorem 1.

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 (2.1) in Section 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 (2.1) with $\boldsymbol{\beta} = (1, 1, 1, 1, 0, 0, 0)^t$, 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 (SNR $= \{\mathrm{var}(\boldsymbol{x}_i^t\boldsymbol{\beta})/\mathrm{var}(\varepsilon_i)\}^{1/2}$) 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 $f_\pi(\boldsymbol{y}|\hat{\sigma}^2)$, where the estimate $\hat{\sigma}^2 = \boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}/n$ of σ^2 is plugged in. The unknown parameter ϕ involved in $\boldsymbol{V}$ 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 $S_0 = \boldsymbol{y}^t\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\}\boldsymbol{y}$ and $S_1 = \boldsymbol{y}^t\{\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{E}\}\boldsymbol{y}$ 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 $n^* = n - \mathrm{tr}\,[\boldsymbol{Z}_2^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Z}_2]$. 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 $\boldsymbol{V}(\hat{\phi})$.

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 $\|\boldsymbol{X}_{\hat{j}}\widehat{\boldsymbol{\beta}}_{\hat{j}} - \boldsymbol{X}_0\boldsymbol{\beta}_0\|^2/n$.

Tables 1–3 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 4–6 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 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 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 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
	$\text{IC}_{\pi,1}$	101	741	834	138	774	862	274	742	901
	$\text{IC}^*_{\pi,1}$	64	620	625	83	625	667	194	554	688
	$\text{IC}_{\pi,2}$	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
	$\text{IC}_{\pi,1}$	422	877	917	450	909	949	427	901	976
	$\text{IC}^*_{\pi,1}$	315	844	844	281	859	862	247	754	856
	$\text{IC}_{\pi,2}$	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
	$\text{IC}_{\pi,1}$	738	926	949	691	936	962	550	961	979
	$\text{IC}^*_{\pi,1}$	795	912	912	560	905	905	289	899	919
	$\text{IC}_{\pi,2}$	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 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
	$\mathrm{IC}_{\pi,1}$	1.70	0.111	0.0372
	$\mathrm{IC}^*_{\pi,1}$	1.92	0.122	0.0429
	$\mathrm{IC}_{\pi,2}$	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
	$\mathrm{IC}_{\pi,1}$	0.754	0.0523	0.0180
	$\mathrm{IC}^*_{\pi,1}$	1.05	0.0534	0.0192
	$\mathrm{IC}_{\pi,2}$	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
	$\mathrm{IC}_{\pi,1}$	0.265	0.0244	0.00841
	$\mathrm{IC}^*_{\pi,1}$	0.285	0.0245	0.00883
	$\mathrm{IC}_{\pi,2}$	0.285	0.0245	0.00883
	IC_r	0.270	0.0243	0.00841
	IC^*_r	0.267	0.0243	0.00840

Table 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: 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
	$\mathrm{IC}_{\pi,1}$	1.96	0.153	0.0492	1.78	0.182	0.0538	1.66	0.312	0.0831
	$\mathrm{IC}^*_{\pi,1}$	2.17	0.164	0.0559	1.95	0.211	0.0566	1.71	0.355	0.0962
	$\mathrm{IC}_{\pi,2}$	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
	$\mathrm{IC}_{\pi,1}$	0.839	0.0587	0.0204	0.832	0.0595	0.0207	1.07	0.136	0.0347
	$\mathrm{IC}^*_{\pi,1}$	1.15	0.0601	0.0216	1.10	0.0638	0.0218	1.14	0.202	0.0432
	$\mathrm{IC}_{\pi,2}$	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
	$\mathrm{IC}_{\pi,1}$	0.286	0.0262	0.00920	0.365	0.0279	0.00983	0.643	0.0530	0.0179
	$\mathrm{IC}^*_{\pi,1}$	0.331	0.0266	0.00958	0.566	0.0284	0.0102	0.914	0.0692	0.0181
	$\mathrm{IC}_{\pi,2}$	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

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 $R(\boldsymbol{\omega};\hat{f})$ in (1.1). We here explain them more clearly through comparison of the related Bayesian criteria. When the prior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\theta})$ is proper, we can treat the Bayesian prediction risk

$$r(\boldsymbol{\psi};\hat{f}) = \int R(\boldsymbol{\omega};\hat{f})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}$$

in (1.5). When $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are known, the predictive density $\hat{f}(\widetilde{\boldsymbol{y}};\boldsymbol{y})$ which minimizes $r(\boldsymbol{\psi};\hat{f})$ is the Bayesian predictive density (posterior predictive density) $\hat{f}_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\theta})$ given by

$$\int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta} = \frac{\int f(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}}{\int f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}}.$$

When $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are unknown, we can consider the Bayesian risk of the plug-in predictive density $\hat{f}_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{y},\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\theta}})$. 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 $r(\boldsymbol{\psi};\hat{f})$.

The Akaike's Bayesian information criterion (ABIC) (Akaike, 1980b) is another information criterion based on the Bayesian marginal likelihood, given by

$$\mathrm{ABIC} = -2\log\{f_\pi(\boldsymbol{y}|\widehat{\boldsymbol{\lambda}})\} + 2\dim(\boldsymbol{\lambda}),$$

where the nuisance parameter $\boldsymbol{\theta}$ is not considered. The ABIC measures the following KL risk:

$$\int\left[\int \log\left\{\frac{f_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{\lambda})}{f_\pi(\widetilde{\boldsymbol{y}}|\widehat{\boldsymbol{\lambda}})}\right\} f_\pi(\widetilde{\boldsymbol{y}}|\boldsymbol{\lambda})\mathrm{d}\widetilde{\boldsymbol{y}}\right] f_\pi(\boldsymbol{y}|\boldsymbol{\lambda})\mathrm{d}\boldsymbol{y},$$

which is not the same as either $R(\boldsymbol{\omega};\hat{f})$ or $r(\boldsymbol{\psi};\hat{f})$. 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 $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\theta})$, since the corresponding Bayesian prediction risks do not exist. On the other hand, we can construct the corresponding criteria based on $R(\boldsymbol{\omega};\hat{f})$, 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.

Acknowledgments.

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

A Derivations of the Criteria

In this section, we show the derivations of the criteria. To this end, we obtain the following lemma, which was shown in Section A.2 of Srivastava and Kubokawa (2010).

Lemma A.1 *Assume that $\boldsymbol{C}$ is an $n \times n$ symmetric matrix, $\boldsymbol{M}$ is an idempotent matrix of rank p and that $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then,*

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

A.1 Derivation of $\mathrm{IC}_{\pi,1}$ in (2.3)

It is sufficient to show that the bias correction $\Delta_{\pi,1} = I_{\pi,1}(\boldsymbol{\omega}) - E_{\boldsymbol{\omega}}[-2\log\{f_\pi(\boldsymbol{y}|\hat{\sigma}^2)\}]$ is $2n/(n-p-2)$, where $I_{\pi,1}(\boldsymbol{\omega})$ is given by (2.2). It follows that

$$\begin{aligned}\Delta_{\pi,1} =& E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{A}\widetilde{\boldsymbol{y}}/\hat{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2)\\ =& E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{A}\widetilde{\boldsymbol{y}}) \cdot E_{\boldsymbol{\omega}}(1/\hat{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2).\end{aligned}$$

Firstly,

$$\begin{aligned}E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{A}\widetilde{\boldsymbol{y}}) =& E_{\boldsymbol{\omega}}[(\widetilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{A}(\widetilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{\beta})]\\ =& \sigma^2\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}.\end{aligned} \tag{A.1}$$

Secondly, noting that $n\hat{\sigma}^2 = \boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y} = \sigma^2\boldsymbol{u}^t(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{u}$ for

$$\begin{aligned}\boldsymbol{u} =& \boldsymbol{V}^{-1/2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/\sigma,\\ \boldsymbol{M} =& \boldsymbol{I}_n - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1/2},\end{aligned} \tag{A.2}$$

and that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{0}$, we can obtain

$$\begin{aligned}E_{\boldsymbol{\omega}}(1/\hat{\sigma}^2) =& nE_{\boldsymbol{\omega}}\left(\frac{1}{\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}}\right) = nE_{\boldsymbol{\omega}}\left[\frac{1}{\sigma^2\boldsymbol{u}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{u}}\right]\\ =& \frac{n}{\sigma^2(n-p-2)}.\end{aligned} \tag{A.3}$$

Finally,

$$\begin{aligned}
E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2) =& nE_{\boldsymbol{\omega}}\left(\frac{\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}}{\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}}\right) = nE_{\boldsymbol{\omega}}\left[\frac{\sigma^2\boldsymbol{u}^t\boldsymbol{V}^{1/2}\boldsymbol{A}\boldsymbol{V}^{1/2}\boldsymbol{u}+\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2\boldsymbol{u}^t(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{u}}\right] \\
=& n\times\left\{\frac{\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V})}{n-p-2}-\frac{2\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V})}{(n-p)(n-p-2)}+\frac{\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2(n-p-2)}\right\}. \qquad \text{(A.4)}
\end{aligned}$$

The last equation in the above can be derived by Lemma A.1. Combining (A.1), (A.3) and (A.4), we get

$$\Delta_{\pi,1}=\frac{2n\cdot\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V})}{(n-p)(n-p-2)}.$$

We can see that

$$\begin{aligned}
\mathrm{tr}\,(\boldsymbol{A}\boldsymbol{V}\boldsymbol{P}\boldsymbol{V}) =& \mathrm{tr}\,\{(\boldsymbol{V}+\boldsymbol{B})^{-1}(\boldsymbol{V}+\boldsymbol{B}-\boldsymbol{B})\boldsymbol{P}\boldsymbol{V}\} \\
=& \mathrm{tr}\,(\boldsymbol{P}\boldsymbol{V})-\mathrm{tr}\,\{(\boldsymbol{V}+\boldsymbol{B})^{-1}\boldsymbol{B}\boldsymbol{P}\boldsymbol{V}\} \\
=& \mathrm{tr}\,(\boldsymbol{I}_n-\boldsymbol{M})=n-p, \qquad \text{(A.5)}
\end{aligned}$$

since $\boldsymbol{B}\boldsymbol{P}=\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^t\boldsymbol{P}=\boldsymbol{0}$, then we obtain $\Delta_{\pi,1}=2n/(n-p-2)$. □

A.2 Derivation of $\mathrm{IC}_{\pi,2}$ in (2.4)

From the fact that $E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1})=I_{\pi,1}(\boldsymbol{\omega})$ and that $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1})=E_\pi[I_{\pi,1}(\boldsymbol{\omega})]=I_{\pi,2}(\sigma^2)$, it suffices to show that $E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1})$ is approximated to

$$\begin{aligned}
E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,1}) \approx& E_\pi E_{\boldsymbol{\omega}}[n\log\hat{\sigma}^2+\log|\boldsymbol{V}|+p\log n+2+E_\pi E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2)] \\
\approx& E_\pi E_{\boldsymbol{\omega}}[n\log\hat{\sigma}^2+\log|\boldsymbol{V}|+p\log n+p]+(n+2)=E_\pi E_{\boldsymbol{\omega}}(\mathrm{IC}_{\pi,2})+(n+2),
\end{aligned}$$

when n is large. Note that $n+2$ is irrelevant to the model. It follows that

$$\begin{aligned}
& E_{\boldsymbol{\omega}}\left(\frac{\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}}{\hat{\sigma}^2}\right) \\
=& n\times E_{\boldsymbol{\omega}}\left[\frac{\boldsymbol{y}^t\{\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}{\boldsymbol{y}^t\{\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}\right] \\
=& n+n\times E_{\boldsymbol{\omega}}\left[\frac{\boldsymbol{y}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}}{\boldsymbol{y}^t\{\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\}\boldsymbol{y}}\right] \\
=& n+\frac{n}{\sigma^2(n-p-2)}\times E_{\boldsymbol{\omega}}\left[\boldsymbol{y}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}\right] \\
=& n+\frac{n}{\sigma^2(n-p-2)}\times\Big[\sigma^2\cdot\mathrm{tr}\,\{(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\} \\
& +\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}\Big],
\end{aligned}$$

and that

$$E_\pi[\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}\boldsymbol{\beta}]=\sigma^2\cdot\mathrm{tr}\,[\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}+\boldsymbol{W}^{-1})^{-1}].$$

If $n^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}$ converges to $p \times p$ positive definite matrix as $n \to \infty$, $\mathrm{tr}\,[(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}] \to 0$ and $\mathrm{tr}\,[\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{W}^{-1})^{-1}] \to p$. Then we can obtain $E_\pi E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}/\hat{\sigma}^2 - n) \to p$, which we want to show.

A.3 Derivation of IC_r in (2.6)

We shall show that the bias correction $\Delta_r = I_r(\boldsymbol{\omega}) - E_{\boldsymbol{\omega}}[-2\log\{f_r(\boldsymbol{y}|\tilde{\sigma}^2)\}]$ is $2(n-p)/(n-p-2)$, where $I_r(\boldsymbol{\omega})$ is given by (2.5). Then,

$$\begin{aligned}\Delta_r =& E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}/\tilde{\sigma}^2) - E_{\boldsymbol{\omega}}(\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y}/\tilde{\sigma}^2)\\ =& E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}) \cdot E_{\boldsymbol{\omega}}(1/\tilde{\sigma}^2) - (n-p).\end{aligned}$$

Since $E_{\boldsymbol{\omega}}(\widetilde{\boldsymbol{y}}^t\boldsymbol{P}\widetilde{\boldsymbol{y}}) = (n-p)\sigma^2$ and $E_{\boldsymbol{\omega}}(1/\tilde{\sigma}^2) = (n-p)/\{\sigma^2(n-p-2)\}$, we get $\Delta_r = 2(n-p)/(n-p-2)$. □

B Proof of Theorem 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

$$P(\hat{\jmath} = j) \le P\{\mathrm{IC}_{\pi,1}(j) < \mathrm{IC}_{\pi,1}(j_0)\}$$

for any $j \in \mathcal{J} \setminus \{j_0\}$, it suffices to show that $P\{\mathrm{IC}_{\pi,1} < \mathrm{IC}_{\pi,1}(j_0)\} \to 0$, or equivalently $P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_0) > 0\} \to 1$ as $n \to \infty$. When $\boldsymbol{V} = \boldsymbol{I}_n$, we obtain

$$\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_0) = I_1 + I_2 + I_3,$$

where

$$\begin{aligned}I_1 =& n\log(\hat{\sigma}_j^2/\hat{\sigma}_0^2) + \boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y}/\hat{\sigma}_0^2,\\ I_2 =& \log|\boldsymbol{X}_j^t\boldsymbol{X}_j + \boldsymbol{W}_j^{-1}| - \log|\boldsymbol{X}_0^t\boldsymbol{X}_0 + \boldsymbol{W}_0^{-1}|,\\ I_3 =& \log\{|\boldsymbol{W}_j|/|\boldsymbol{W}_0|\} + \frac{2n}{n-p_j-2} - \frac{2n}{n-p_0-2},\end{aligned}$$

for $\hat{\sigma}_j^2 = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{H}_j)\boldsymbol{y}/n$, $\hat{\sigma}_0^2 = \hat{\sigma}_{j_0}^2$, $\boldsymbol{A}_j = \boldsymbol{I}_n - \boldsymbol{X}_j(\boldsymbol{X}_j^t\boldsymbol{X}_j + \boldsymbol{W}_j^{-1})^{-1}\boldsymbol{X}_j^t$ and $\boldsymbol{H}_0 = \boldsymbol{H}_{j_0}$. We evaluate asymptotic behaviors of I_1, I_2 and I_3 for $j \in \mathcal{J}_-$ and $j \in \mathcal{J}_+ \setminus \{j_0\}$, separately.

[Case of $j \in \mathcal{J}_-$]. Firstly, we evaluate I_1. We decompose $I_1 = I_{11} + I_{12}$, where $I_{11} = n\log(\hat{\sigma}_j^2/\hat{\sigma}_0^2)$ and $I_{12} = \boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y}/\hat{\sigma}_0^2$. It follows that

$$\begin{aligned}\hat{\sigma}_j^2 - \hat{\sigma}_0^2 =& (\boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon})^t(\boldsymbol{I}_n - \boldsymbol{H}_j)(\boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon})/n - \boldsymbol{\varepsilon}^t(\boldsymbol{I}_n - \boldsymbol{H}_0)\boldsymbol{\varepsilon}/n\\ =& \|\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{H}_j\boldsymbol{X}_0\boldsymbol{\beta}_0\|^2/n + o_p(1),\end{aligned}$$

Then we can see that

$$n^{-1}I_{11} = \log\left(1 + \frac{\hat{\sigma}_j^2 - \hat{\sigma}_0^2}{\hat{\sigma}_0^2}\right) = \log\left\{1 + \frac{\|\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{H}_j\boldsymbol{X}_0\boldsymbol{\beta}_0\|^2}{n\sigma^2}\right\} + o_p(1), \tag{B.1}$$

and it follows from the assumption (A3) that

$$\liminf_{n\to\infty} \log\left\{1 + \frac{\|\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{H}_j\boldsymbol{X}_0\boldsymbol{\beta}_0\|^2}{n\sigma^2}\right\} > 0. \tag{B.2}$$

Because $\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y}/(n\hat{\sigma}_j^2) = 1 + o_p(1)$ and $\boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y}/(n\hat{\sigma}_0^2) = 1 + o_p(1)$, we obtain

$$n^{-1}I_{12} = o_p(1). \tag{B.3}$$

Secondly, we evaluate I_2. It follows that

$$\log|\boldsymbol{X}_j^t\boldsymbol{X}_j + \boldsymbol{W}_j^{-1}| = p_j\log n + \log|\boldsymbol{X}_j^t\boldsymbol{X}_j/n + \boldsymbol{W}_j^{-1}/n| = p_j\log n + O(1).$$

It can be also seen that $\log|\boldsymbol{X}_0^t\boldsymbol{X}_0 + \boldsymbol{W}_0^{-1}| = p_0\log n + O(1)$. Then,

$$n^{-1}I_2 = (p_j - p_0)n^{-1}\log n + o(1) = o(1). \tag{B.4}$$

Lastly, it is easy to see that

$$n^{-1}I_3 = o(1). \tag{B.5}$$

From (B.1)–(B.5), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_0) > 0\} \to 1, \tag{B.6}$$

for all $j \in \mathcal{J}-$.

[Case of $j \in \mathcal{J}_+ \setminus \{j_0\}$]. Firstly, we evaluate I_1. From the fact that

$$\hat{\sigma}_0^2 - \hat{\sigma}_j^2 = \boldsymbol{\varepsilon}^t(\boldsymbol{H}_j - \boldsymbol{H}_0)\boldsymbol{\varepsilon}/n = O_p(n^{-1}), \tag{B.7}$$

it follows that

$$\begin{aligned}(\log n)^{-1}I_{11} =&(\log n)^{-1}\cdot n\log\left\{\frac{\hat{\sigma}_0^2 - (\hat{\sigma}_0^2 - \hat{\sigma}_j^2)}{\hat{\sigma}_0^2}\right\}\\ =&(\log n)^{-1}\cdot n\cdot\log\{1 + O_p(n^{-1})\} = o_p(1).\end{aligned} \tag{B.8}$$

As for I_{12}, from (B.7) and $\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y} = O_p(1)$, we can obtain

$$\begin{aligned}I_{12} =&\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y}/\hat{\sigma}_j^2 - \boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y}/\hat{\sigma}_0^2\\ =&(\boldsymbol{y}^t\boldsymbol{A}_j\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{A}_0\boldsymbol{y})/\hat{\sigma}_0^2 + O_p(1) = O_p(1).\end{aligned}$$

Then,

$$(\log n)^{-1}I_{12} = o_p(1). \tag{B.9}$$

Secondly, we evaluate I_2. Since $p_j > p_0$ for all $j \in \mathcal{J}_+ \setminus \{j_0\}$,

$$\liminf_{n\to\infty} (\log n)^{-1} I_2 = p_j - p_0 > 0. \tag{B.10}$$

Finally, it is easy to see that

$$(\log n)^{-1} I_3 = o(1). \tag{B.11}$$

From (B.8)–(B.11), it follows that

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,2}(j_0) > 0\} \to 1, \tag{B.12}$$

for all $j \in \mathcal{J}_+ \setminus \{j_0\}$.

Combining (B.6) and (B.12), we obtain

$$P\{\mathrm{IC}_{\pi,1}(j) - \mathrm{IC}_{\pi,1}(j_0) > 0\} \to 1,$$

for all $j \in \mathcal{J} \setminus \{j_0\}$, which shows that $\mathrm{IC}_{\pi,1}$ is consistent. □

References

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 Trans. Autom. Contr.*, **AC-19**, 716–723.

Akaike, H. (1980a). On the use of predictive likelihood of a Gaussian model. *Ann. Inst. Statist. Math.*, **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.

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

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Soc.*, **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.

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.*, **101**, 157–167.

Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.*, **21**, 309–310.

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Commun. Statist. Theory Meth.*, **26**, 2223–2246.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221–264.

Shi, P. and Tsai, C.-L. (2002). Regression model selection—a residual likelihood approach. *J. Royal Statist. Soc.* B, **64**, 237–252.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

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

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

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theory Meth.*, **7**, 13–26.

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

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.