数理解析研究所講究録
第1860巻 2013年 58-72

Variable Selection for Bayesian Linear Regression Model in a Finite Sample Size

Satoshi KABE [1]
Graduate School of Systems and Information Engineering,
University of Tsukuba.

Yuichiro KANAZAWA [2]
Graduate School of Systems and Information Engineering,
University of Tsukuba.

1 Introduction

Data analysis often involves a comparison of several candidate models. Because true model is seldom known a priori, there is a need for a simple, effective, and objective methods for the selection of the best approximating model. Akaike (1973) proposed an information criterion later known to be Akaike's information criterion (AIC) based on the concept of minimizing the Kullback-Leibler Information, a measure of discrepancy between the true density (or model) and approximating model: The discrepancy between two models or two probability densities is expressed by the expected log-likelihood with respect to the true density. AIC is designed to be an approximately unbiased estimator of the expected log-likelihood under the assumption that "true model" is one of the candidate models being considered.

Hurvich and Tsai (1989) illustrated that AIC can be dramatically biased when the sample size is small by comparing AIC with the finite bias-corrected version of AIC (AIC_C) proposed by Sugiura (1978). This criterion adjusts AIC to be an exact unbiased estimator of the expected log-likelihood. They then extend AIC_C so that it can be employed for autoregressive (AR) model and autoregressive moving average (ARMA) model. Later Bedrick and Tsai (1994) further extend Sugiura (1978) to multivariate regression cases where response variable is expressed by a matrix.

In the 'fully' Bayesian data analysis, the deviance information criterion (DIC) is widely used for the model selection. Spiegelhalter et al. (2002) proposed a Bayesian measure of model complexity (i.e., effective number of parameters p_D) obtained from the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters. When the number of data is sufficiently large, DIC is given by adding p_D to the posterior mean of the deviance. Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in cases where the number of observations is large relative to the number of parameters.

In an discussion to Spiegelhalter et al. (2002), Burnham (2002) questioned "[a] lesson that we learned was that, if sample size n is small or the number of estimated parameters p is large relative to n, a modified AIC should be used, such as $\mathrm{AIC}_C = \mathrm{AIC} + 2p(p+1)/(n-p-1)$. I wonder whether DIC needs such a modification or if it really automatically adjusts for a small sample size or large p, relative to n." We write this article to answer this question, at least partially for a very important case of linear regression. More concretely,

[1]E-mail: k0420214@sk.tsukuba.ac.jp.
[2]E-mail: kanazawa@sk.tsukuba.ac.jp.

we propose a finite-sample bias corrected information criterion for the Bayesian linear regression models with normally distributed error, and then we implement simulation studies when the sample size is relatively small.

The rest of this article is organized as follows: Next section briefly describes the Bayesian linear regression model. In Section 3, we propose our information criterion for the Bayesian linear regression model. Section 4 shows the results of simulation studies to show the validity of our proposed information criterion when the sample size is small. Finally, Section 5 draws some conclusions concerning our proposed criterion.

2 Bayesian Linear Regression Model

We consider the linear regression model as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N\right) \tag{2.1}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector and $\boldsymbol{X}$ is a $N \times K$ non-stochastic matrix. The parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ vector and error term $\boldsymbol{\varepsilon}$ follows a N-dimensional multivariate normal distribution $\mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N\right)$.

We assume that prior distribution of $\boldsymbol{\beta}$ is a K-dimensional multivariate normal distribution and that of σ^{-2} is a gamma distribution:

$$\boldsymbol{\beta}|\sigma^{-2} \sim \mathcal{N}\left(\boldsymbol{b}_0, \sigma^2 \boldsymbol{B}_0\right) \tag{2.2}$$

$$\sigma^{-2} \sim \mathcal{G}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right). \tag{2.3}$$

Then posterior distributions of parameters $\boldsymbol{\beta}$ and σ^{-2} are expressed as

$$\boldsymbol{\beta}|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{b}_1, \sigma^2 \boldsymbol{B}_1\right) \tag{2.4}$$

$$\sigma^{-2}|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{G}\left(\frac{\nu_1}{2}, \frac{\lambda_1}{2}\right) \tag{2.5}$$

where $\boldsymbol{b}_1 = \boldsymbol{B}_1\left(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right)$, $\boldsymbol{B}_1 = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}$, $\nu_1 = \nu_0 + N + K$, $\lambda_1 = \lambda_0 + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_N) + (\boldsymbol{b}_0 - \hat{\boldsymbol{\beta}}_N)'[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}_0]^{-1}(\boldsymbol{b}_0 - \hat{\boldsymbol{\beta}}_N)$ and $\hat{\boldsymbol{\beta}}_N = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

3 Finite-Sample Bias Correction and Variable Selection Criterion

Let us denote unknown true density as $f_{\boldsymbol{Y}}(\cdot)$ and approximating model as $g(\cdot|\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$. Then Kullback-Leibler Information between $f_{\boldsymbol{Y}}(\cdot)$ and $g(\cdot|\boldsymbol{\theta})$ can be expressed as follows

$$I(f_{\boldsymbol{Y}}, g(\cdot|\boldsymbol{\theta})) = \int f_{\boldsymbol{Y}}(\boldsymbol{z}) \log\left\{\frac{f_{\boldsymbol{Y}}(\boldsymbol{z})}{g(\boldsymbol{z}|\boldsymbol{\theta})}\right\} d\boldsymbol{z} \tag{3.1}$$

and (3.1) can be rewritten as

$$I(f_{\boldsymbol{Y}}, g(\cdot|\boldsymbol{\theta})) = \mathbf{E}_z\left[\log\{f_{\boldsymbol{Y}}(\boldsymbol{z})\}\right] - \mathbf{E}_z\left[\log\{g(\boldsymbol{z}|\boldsymbol{\theta})\}\right]. \tag{3.2}$$

Even though the true density $f_{\mathbf{Y}}(\cdot)$ is unknown, the first term on the right-hand side of Kullback-Leibler Information in (3.2) can be regarded as a constant since the variable $\boldsymbol{z}$ is integrated out. Then we select the best approximating model with maximizing the expected log-likelihood.

In Bayesian perspective, parameters follow the posterior distributions estimated by observed data $\boldsymbol{y}$. Hence we consider the posterior mean of (3.2):

$$\mathbf{E}_{\theta|y}\left[I(f_{\mathbf{Y}}, g(\cdot|\boldsymbol{\theta}))\right] = \mathbf{E}_z\left[\log\{f_{\mathbf{Y}}(\boldsymbol{z})\}\right] - \mathbf{E}_{\theta|y}\left[\mathbf{E}_z\left[\log\{g(\boldsymbol{z}|\boldsymbol{\theta})\}\right]\right] \tag{3.3}$$

and as in Spiegelhalter et al. (2002) and Ando (2007), our proposed information criterion is constructed based on the posterior mean of expected log-likelihood.

From the Bayesian linear regression model (2.1), we use $\boldsymbol{y}$ as observed data of sample size N obtained from the unknown true density $f_{\mathbf{Y}}(\boldsymbol{y})$ to estimate the posterior distributions of parameters $\boldsymbol{\beta}$ and σ^{-2}, while we also use $\boldsymbol{z}$ as replicate data of sample size N generated from the unknown true density $f_{\mathbf{Y}}(\boldsymbol{z})$ to evaluate the goodness of fit of approximating model $g(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^{-2})$. Then posterior mean of expected log-likelihood $\mathcal{T}$ is defined as

$$\mathcal{T} \equiv \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\log\left\{g(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^{-2})\right\}\right]\right] \tag{3.4}$$

where expectation $\mathbf{E}_{\beta,\sigma^{-2}|y,X}[\,\cdot\,]$ can be calculated by $\mathbf{E}_{\sigma^{-2}|y,X}[\ \mathbf{E}_{\beta|\sigma^{-2}y,X}[\,\cdot\,]\ ]$ from the posterior distributions (2.4) and (2.5).

To estimate the posterior mean of expected log-likelihood $\mathcal{T}$ in (3.4), we use the posterior mean of observed log-likelihood $\widehat{\mathcal{T}}_N$:

$$\widehat{\mathcal{T}}_N \equiv \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\log\left\{g(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^{-2})\right\}\right]. \tag{3.5}$$

and the bias-corrected estimator is obtained as $\widehat{\mathcal{T}}_N - \widehat{b}_N$, where $\widehat{b}_N$ is an estimate of bias $b_\Theta \equiv \mathbf{E}_y[\widehat{\mathcal{T}}_N - \mathcal{T}] \neq 0$. Then we propose information criterion (IC) of the form

$$\mathrm{IC} \equiv -2\widehat{\mathcal{T}}_N + 2\widehat{b}_N. \tag{3.6}$$

Ignoring the constant term, we can express the log-likelihood function for the replicate data $\boldsymbol{z}$ such as

$$\log\left\{g(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^{-2})\right\} = \frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.7}$$

where parameters $\boldsymbol{\beta}$ and σ^{-2} follow the posterior distributions estimated by observed data $\boldsymbol{y}$ and $\boldsymbol{X}$. Then the posterior mean of expected log-likelihood $\mathcal{T}$ in (3.4) is rewritten as

$$\begin{aligned}
\mathcal{T} &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{N}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \\
&\quad - \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_z\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&\quad + \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]
\end{aligned}$$

$$= \widehat{\mathcal{T}}_N - C_1 + C_2 \tag{3.8}$$

and the bias b_Θ with respect to the true density $f_{\mathbf{Y}}(\boldsymbol{y})$ is obtained by $b_\Theta \equiv \mathbf{E}_y[\widehat{\mathcal{T}}_N - \mathcal{T}] = \mathbf{E}_y[C_1 - C_2]$.

First we evaluate C_1 in (3.8). However, true density $f_{\mathbf{Y}}(\boldsymbol{z})$ is seldom known in practice, so that expectation with respect to the true density is not analytically obtained. In the previous studies, Kitagawa (1997) replaced the unknown true density by the prior predictive density to construct the predictive information criterion (PIC) for the Bayesian linear Gaussian model, while Laud and Ibrahim (1995), Gelfand and Ghosh (1998), and Ibrahim et al. (2001) considered using the posterior predictive density to generate the replicate data $\boldsymbol{z}$ for model assessment. In this article, we use the posterior predictive density to evaluate the expectation with respect to the true density $f_{\mathbf{Y}}(\boldsymbol{z})$ in C_1 because the prior predictive density is far more sensitive to the selection of prior distribution.

To evaluate C_1 in (3.8), we replace the true density $f_{\mathbf{Y}}(\cdot)$ with a (conditional) posterior predictive density

$$\boldsymbol{z}|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{b}_1, \sigma^2\boldsymbol{\Sigma}_0\right), \tag{3.9}$$

where σ^{-2} follows the posterior distribution (2.5) and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'$. From (3.9), we estimate C_1 in (3.8) as

$$\begin{aligned}
\widehat{C}_1 &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&\quad + \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right].
\end{aligned} \tag{3.10}$$

The first term on the right-hand side of (3.10) can be rewritten as follows

$$\begin{aligned}
&\mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}\mathrm{tr}\left\{\mathbf{E}_{z|\sigma^{-2},y,X}\left[(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'\right]\right\}\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}\mathrm{tr}\left\{\sigma^2\boldsymbol{\Sigma}_0\right\}\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}_N + \boldsymbol{X}\boldsymbol{B}_1\boldsymbol{X}'\right\} \\
&= \frac{N}{2} + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}.
\end{aligned} \tag{3.11}$$

The second term on the right-hand side of (3.10) is similarly obtained from the posterior distribution (2.4) as follows

$$\begin{aligned}
&\mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right] \\
&= \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\sigma^2\boldsymbol{B}_1\right\}\right]
\end{aligned}$$

$$= \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}. \tag{3.12}$$

From (3.11) and (3.12), $\widehat{C}_1$ in (3.10) is evaluated as follows

$$\begin{aligned}
\widehat{C}_1 &= \mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\mathbf{E}_{z|\sigma^{-2},y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{b}_1)\right]\right] \\
&\quad + \mathbf{E}_{\sigma^{-2}|y,X}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{X}'\boldsymbol{X})\mathbf{E}_{\beta|\sigma^{-2},y,X}\left[(\boldsymbol{\beta}-\boldsymbol{b}_1)(\boldsymbol{\beta}-\boldsymbol{b}_1)'\right]\right\}\right] \\
&= \frac{N}{2} + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} + \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} \\
&= \frac{N}{2} + \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}.
\end{aligned} \tag{3.13}$$

Then we notice that $\widehat{C}_1$ does not depend on any data $\boldsymbol{y}$, i.e., $\mathbf{E}_y(\widehat{C}_1) = \widehat{C}_1$.

Suppose that interchange of order of integrations is valid, we can rewrite $\mathbf{E}_y(C_2)$ such as

$$\begin{aligned}
\mathbf{E}_y(C_2) &= \mathbf{E}_y\left[\mathbf{E}_{\beta,\sigma^{-2}|y,X}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right] \\
&= \mathbf{E}_{\beta,\sigma^{-2}}\left[\mathbf{E}_{y|X,\beta,\sigma^{-2}}\left[\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right]\right]
\end{aligned} \tag{3.14}$$

where $\mathbf{E}_{\beta,\sigma^{-2}}[\,\cdot\,]$ is an expectation with respect to the joint prior distribution and $\mathbf{E}_{y|X,\beta,\sigma^{-2}}[\,\cdot\,]$ is an expectation with respect to N-dimensional multivariate normal distribution with mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2\boldsymbol{I}_N$. Since $\mathbf{E}_{y|X,\beta,\sigma^{-2}}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'] = \sigma^2\boldsymbol{I}_N$, $\mathbf{E}_y(C_2)$ in (3.14) can be evaluated as

$$\begin{aligned}
\mathbf{E}_y(C_2) &= \mathbf{E}_{\beta,\sigma^{-2}}\left[\mathbf{E}_{y|X,\beta,\sigma^{-2}}\left[\mathrm{tr}\left\{\frac{\sigma^{-2}}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\right\}\right]\right] \\
&= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}_N\right\} \\
&= \frac{N}{2}.
\end{aligned} \tag{3.15}$$

Therefore $\widehat{b}_N$ is given by

$$\begin{aligned}
\widehat{b}_N &= \mathbf{E}_y\left[\widehat{C}_1 - C_2\right] \\
&= \widehat{C}_1 - \mathbf{E}_y(C_2) \\
&= \frac{N}{2} + \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} - \frac{N}{2} \\
&= \mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\}.
\end{aligned} \tag{3.16}$$

Then $\widehat{b}_N$ in (3.16) can be regarded as a ratio of variance-covariance matrices $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\sigma^2\boldsymbol{B}_1$.

Multiplying -2 to the bias-corrected estimator $\widehat{\mathcal{T}}_N - \widehat{b}_N$, our proposed information criterion for variable selection in the Bayesian linear regression model (IC_{BL}) is obtained by

$$\mathrm{IC}_{BL} = -2\widehat{\mathcal{T}}_N + 2\mathrm{tr}\left\{(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_1\right\} \tag{3.17}$$

where $\boldsymbol{B}_1 = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}$.

For simplicity, let us denote the parameter $\boldsymbol{B}_0$ in (2.2) as $\boldsymbol{B}_0 = \kappa_0 \boldsymbol{I}_K$ ($\kappa_0 > 0$) and the bias term $\widehat{b}_N$ can be rewritten as

$$\widehat{b}_N = K - \mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_0 + \boldsymbol{I}_K)^{-1}\} \tag{3.18}$$

from the matrix inversion lemma [3]. Then if the sample size $N \to \infty$, the last term in (3.18) is expected to be zero because $\mathrm{tr}\{(\kappa_0 \boldsymbol{X}'\boldsymbol{X}/N + \boldsymbol{I}_N/N)^{-1}/N\} \to 0$ (i.e., $\widehat{b}_N \to K$) when each element of $\boldsymbol{X}'\boldsymbol{X}/N$ does not diverge. Furthermore, if κ_0 is sufficiently large (i.e., non-informative prior), we also have $\mathrm{tr}\{(\kappa_0 \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_K)^{-1}\} \to 0$ in (3.18).

4 Simulation Experiments

4.1 Deviance Information Criterion (DIC)

We conduct two simulation studies to compare small sample performances of our proposed information criterion (IC_{BL}) in (3.17) with the deviance information criterion (DIC) :

$$\mathrm{DIC} = -2\widehat{\mathcal{T}}_N + p_D, \tag{4.1}$$

where Spiegelhalter et al. (2002) termed p_D as the effective number of parameters defined to be $p_D \equiv 2\log\{g(\boldsymbol{y}|\boldsymbol{X}, \bar{\boldsymbol{\beta}}, \bar{\sigma}^{-2})\} - 2\widehat{\mathcal{T}}_N$ evaluated at the posterior means of parameters $\bar{\boldsymbol{\beta}}$ ($= \boldsymbol{b}_1$) and $\bar{\sigma}^{-2}$ ($= \nu_1/\lambda_1$) in (2.4) and (2.5).

4.2 Simulation studies

In this article, we present two small sample simulation studies: The first one is designed to examine cases where the set of candidate models includes many over-specified models; the second one examines performances with respect to under-specified candidate models.

Case 1

As in Hurvich and Tsai (1989), we consider the nested candidate models by using seven explanatory variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7$. In our simulation study, $\boldsymbol{x}_1$ is a $N \times 1$ vector whose elements are ones (i.e., intercept term) and the other $N \times 1$ vectors $\boldsymbol{x}_i$ ($2 \leq i \leq 7$) are generated from the uniform distribution $\mathcal{U}(-2, 2)$. These variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7$ are included into the candidate models in a sequentially nested fashion. The candidate models are linear regression models given by $\boldsymbol{y} = \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \cdots + \beta_K\boldsymbol{x}_K + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$. The candidate model with $K = 1$ has only intercept term and with $K = 7$ is the full model. In this simulation study, we determine the number of variables K by using our proposed information criterion (IC_{BL}) in (3.17) and DIC in (4.1) in small sample cases $N = 25, 50, 100$ with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors. To

[3] For any matrices $\boldsymbol{A}$ ($m \times m$), $\boldsymbol{B}$ ($m \times n$), $\boldsymbol{C}$ ($n \times m$), and $\boldsymbol{D}$ ($n \times n$), we have

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ and $\boldsymbol{D}$ are nonsingular matrices.

examine the small sample performance in the Bayesian linear regression cases, we generate a sample of $\boldsymbol{y}$ from the true model ($K = 3$):

$$\boldsymbol{y} = 1.0\boldsymbol{x}_1 + 2.0\boldsymbol{x}_2 + 3.0\boldsymbol{x}_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, 1.0\boldsymbol{I}_N). \tag{4.2}$$

In Table 1, we examine the performance of IC_{BL} and DIC for the small sample cases ($N = 25, 50, 100$). The parameters of prior distributions in (2.2) and (2.3) are set to be $\boldsymbol{b}_0 = \boldsymbol{0}$, $\boldsymbol{B}_0 = \kappa_0 \boldsymbol{I}_K$ ($\kappa_0 = 0.1$ or 100), and $\nu_0 = \lambda_0 = 0.1$. The simulation considered each combination of $N = 25, 50, 100$ and $\kappa_0 = 0.1, 100$. We draw 50,000 MCMC samples from the posterior distributions in (2.4) and (2.5) to compute the posterior mean of observed log-likelihood $\widehat{T}_N$ in (3.5). For each combination of (N, κ_0), we generate 100 observations of IC_{BL} and DIC, and record the number of selected models (i.e., the candidate model with minimum value for the two criteria).

Table 1 shows that our proposed information criterion (IC_{BL}) identifies the true model ($K = 3$) for the small sample cases ($N = 25, 50, 100$) with informative prior ($\kappa_0 = 0.1$) far better than DIC, on the other hand DIC tends to overfit the true model. For the non-informative prior ($\kappa_0 = 100$), both criteria tend to overfit the true model for the sample size $N = 25$ and 50, but nevertheless our proposed information criterion (IC_{BL}) far outperformes DIC at the sample size $N = 100$.

In Tables 2 and 3, we show the results of average criteria in 100 observations for the small sample cases ($N = 25, 50, 100$) with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors. Both criteria selected the true model ($K = 3$) in all cases, but the difference between $2\widehat{b}_N$ and p_D becomes more apparent along with an increase in the number of explanatory variables. Hence the effective number of parameters p_D in DIC tends to underestimate the model complexity compared with the bias term $2\widehat{b}_N$ in IC_{BL}.

Case 2

Next we consider the cases where explanatory variables in the candidate models are selected from the subsets of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, (i.e., $\{\boldsymbol{x}_1\}$, $\{\boldsymbol{x}_2\}$, $\{\boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$). Then we specify a true model such as

$$\boldsymbol{y} = 1.0\boldsymbol{x}_1 + 2.0\boldsymbol{x}_2 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, 1.0\boldsymbol{I}_N). \tag{4.3}$$

We implement a simulation study to examine the performance of IC_{BL} in (3.17) and DIC in (4.1) for under-specified candidate models in small sample cases.

The parameters of prior distributions in (2.2) and (2.3) are set to be $\boldsymbol{b}_0 = \boldsymbol{0}$, $\boldsymbol{B}_0 = \kappa_0 \boldsymbol{I}_K$ ($\kappa_0 = 0.1$ or 100), and $\nu_0 = \lambda_0 = 0.1$. We draw 50,000 MCMC samples from the posterior distributions in (2.4) and (2.5) and compute the posterior mean of observed log-likelihood $\widehat{T}_N$ in (3.5). For each combination of (N, κ_0), we generate 100 observations of IC_{BL} and DIC and record the number of selected models such as Case 1.

Table 4 shows that DIC tends to select a full model (i.e., $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$) as compared with IC_{BL} in small sample cases. On the other hand, under-specified candidate models (i.e., $\{\boldsymbol{x}_1\}$, $\{\boldsymbol{x}_2\}$, $\{\boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$) are not selected at all. Hence, DIC shows a tendency to overfit the true model in this simulation study.

In Tables 5 and 6, average values of model complexity in IC_{BL} and DIC with respect to the candidate models $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, and $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ are close to each other. However,

true model (i.e., $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$) has the largest value of posterior mean of observed log-likelihood $\widehat{\mathcal{T}}_N$ among three candidate models. Therefore, average values of IC_{BL} and DIC successfully select the true model (4.3) in all cases.

5 Conclusion and Discussion

In Bayesian data analysis, DIC (Spiegelhalter et al., 2002) is widely used for the model selection, since this criterion is relatively easy to calculate and applicable to a wide range of statistical models. Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in cases where the number of observations is large relative to the number of parameters. However, Burnham (2002) questioned if the DIC needs a modification for small sample size. In this article, we proposed a variable selection criterion IC_{BL} for the Bayesian linear regression models in small sample cases, as Sugiura (1978) proposed AIC_C in frequentist framework. We then examined the performance of our proposed information criterion (IC_{BL}) relative to the DIC for small sample cases.

In our simulation studies, we found that DIC often showed tendency to overfit the true model (see Tables 1 and 4), whereas our proposed information criterion (IC_{BL}) performs well for small sample cases ($N = 25, 50, 100$). We also found that the measure of model complexity $2\widehat{b}_N$ was mostly larger than p_D (see Tables 2 and 3), leading us to conclude bias correction of DIC underestimates the model complexity in small sample cases.

An important directions for the further research would be to extend our information criterion to the several types of Bayesian regression models (e.g., the hierarchical Bayesian regression models, Bayesian regression models with serially correlated error, and Bayesian Markov-switching models) because empirical analysis is generally performed under the limitation of data availability.

Table 1: The number of selected models by IC_{BL} and DIC for small sample cases $N = 25, 50, 100$ with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors in Case 1 (100 observations).

	Informative prior ($\kappa_0 = 0.1$)						Non-informative prior ($\kappa_0 = 100$)					
	$N = 25$		$N = 50$		$N = 100$		$N = 25$		$N = 50$		$N = 100$	
Model (K)	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	97	79	95	74	95	76	81	65	89	75	90	75
4	2	9	5	15	5	14	8	10	5	13	7	12
5	1	6	0	7	0	3	7	9	4	8	3	3
6	0	5	0	2	0	5	0	3	0	1	0	6
7	0	1	0	2	0	2	4	13	2	3	0	4

Table 2: Average values of IC_{BL} and DIC in 100 observations for small sample cases $N = 25, 50, 100$ with informative prior ($\kappa_0 = 0.1$) in Case 1.

	Informative prior ($\kappa_0 = 0.1$)														
	$N = 25$					$N = 50$					$N = 100$				
Model (K)	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
1	99.304	99.558	-48.938	1.429	1.683	196.367	196.519	-97.350	1.667	1.819	394.064	394.147	-196.123	1.818	1.902
	(6.596)	(6.595)	(3.298)	(0.000)	(0.008)	(9.404)	(9.404)	(4.702)	(0.000)	(0.009)	(12.972)	(12.972)	(6.486)	(0.000)	(0.009)
2	93.196	92.664	-45.134	2.928	2.396	181.946	181.215	-89.275	3.396	2.665	361.471	360.616	-178.897	3.677	2.822
	(5.943)	(5.945)	(2.974)	(0.077)	(0.041)	(7.360)	(7.362)	(3.682)	(0.030)	(0.021)	(9.739)	(9.737)	(4.868)	(0.012)	(0.013)
3	55.425	54.115	-25.502	4.420	3.111	86.241	84.633	-40.565	5.111	3.503	136.230	134.440	-65.349	5.531	3.742
	(2.984)	(2.984)	(1.486)	(0.112)	(0.065)	(5.324)	(5.326)	(2.663)	(0.043)	(0.046)	(10.446)	(10.448)	(5.223)	(0.019)	(0.038)
4	56.745	54.676	-25.436	5.874	3.805	87.682	85.198	-40.430	6.823	4.338	137.879	135.149	-65.248	7.382	4.653
	(3.030)	(3.029)	(1.516)	(0.146)	(0.075)	(5.385)	(5.383)	(2.693)	(0.058)	(0.050)	(10.578)	(10.589)	(5.289)	(0.027)	(0.042)
5	58.004	55.200	-25.357	7.290	4.486	89.188	85.832	-40.333	8.521	5.165	139.577	135.916	-65.171	9.234	5.573
	(3.088)	(3.075)	(1.536)	(0.208)	(0.105)	(5.474)	(5.477)	(2.743)	(0.063)	(0.051)	(10.477)	(10.480)	(5.239)	(0.031)	(0.046)
6	59.228	55.703	-25.272	8.684	5.159	90.776	86.568	-40.283	10.211	6.002	141.136	136.540	-65.026	11.083	6.487
	(3.141)	(3.131)	(1.560)	(0.234)	(0.119)	(5.456)	(5.465)	(2.735)	(0.084)	(0.052)	(10.495)	(10.492)	(5.248)	(0.034)	(0.041)
7	60.470	56.221	-25.201	10.068	5.820	92.206	87.137	-40.160	11.886	6.817	142.716	137.185	-64.895	12.925	7.395
	(3.094)	(3.069)	(1.530)	(0.269)	(0.139)	(5.446)	(5.454)	(2.729)	(0.095)	(0.057)	(10.512)	(10.519)	(5.256)	(0.038)	(0.046)

Standard deviations in parentheses

Table 3: Average values of IC_{BL} and DIC in 100 observations for small sample cases $N = 25, 50, 100$ with non-informative prior ($\kappa_0 = 100$) in Case 1.

	Non-informative prior ($\kappa_0 = 100$)														
	$N = 25$					$N = 50$					$N = 100$				
Model (K)	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
1	98.416	98.386	-48.208	1.999	1.970	197.876	197.861	-97.938	2.000	1.984	393.852	393.846	-195.926	2.000	1.993
	(6.588)	(6.589)	(3.294)	(0.000)	(0.010)	(8.148)	(8.147)	(4.074)	(0.000)	(0.010)	(12.035)	(12.034)	(6.017)	(0.000)	(0.010)
2	93.019	91.952	-44.510	3.999	2.932	182.333	181.298	-89.167	3.999	2.965	360.553	359.537	-178.277	4.000	2.984
	(5.732)	(5.732)	(2.866)	(1.65E-04)	(0.011)	(7.243)	(7.243)	(3.622)	(4.67E-05)	(0.011)	(10.999)	(10.998)	(5.500)	(1.49E-05)	(0.011)
3	31.506	29.409	-12.754	5.998	3.901	56.363	54.310	-25.182	5.999	3.946	108.119	106.092	-51.060	5.999	3.973
	(7.097)	(7.097)	(3.549)	(2.90E-04)	(0.012)	(12.093)	(12.090)	(6.047)	(7.37E-05)	(0.013)	(13.496)	(13.498)	(6.748)	(2.42E-05)	(0.012)
4	33.519	30.392	-12.761	7.997	4.869	58.419	55.352	-25.210	7.999	4.932	110.077	107.040	-51.039	7.999	4.962
	(7.281)	(7.282)	(3.641)	(3.70E-04)	(0.014)	(12.074)	(12.074)	(6.037)	(1.05E-04)	(0.016)	(13.381)	(13.381)	(6.691)	(2.78E-05)	(0.014)
5	35.083	30.928	-12.544	9.996	5.841	60.275	56.190	-25.138	9.998	5.914	111.959	107.913	-50.980	9.999	5.952
	(7.372)	(7.373)	(3.686)	(4.98E-04)	(0.015)	(12.271)	(12.270)	(6.136)	(1.35E-04)	(0.015)	(13.369)	(13.370)	(6.685)	(3.29E-05)	(0.016)
6	37.234	32.048	-12.620	11.995	6.809	62.496	57.395	-25.249	11.998	6.897	113.978	108.924	-50.990	11.999	6.945
	(7.633)	(7.634)	(3.817)	(6.20E-04)	(0.016)	(12.272)	(12.273)	(6.136)	(1.66E-04)	(0.020)	(13.766)	(13.766)	(6.883)	(3.61E-05)	(0.016)
7	38.753	32.543	-12.379	13.994	7.784	64.278	58.159	-25.140	13.997	7.878	115.613	109.551	-50.807	13.999	7.937
	(8.174)	(8.177)	(4.087)	(7.31E-04)	(0.018)	(12.289)	(12.289)	(6.144)	(1.85E-04)	(0.018)	(14.019)	(14.020)	(7.010)	(3.84E-05)	(0.019)

Standard deviations in parentheses

Table 4: The number of selected models by IC_{BL} and DIC for small sample cases $N = 25, 50, 100$ with informative ($\kappa_0 = 0.1$) and non-informative ($\kappa_0 = 100$) priors in Case 2 (100 observations).

	Informative prior ($\kappa_0 = 0.1$)						Non-informative prior ($\kappa_0 = 100$)					
	$N = 25$		$N = 50$		$N = 100$		$N = 25$		$N = 50$		$N = 100$	
Model	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC	IC_{BL}	DIC
$\{\boldsymbol{x}_1\}$	0	0	0	0	0	0	0	0	0	0	0	0
$\{\boldsymbol{x}_2\}$	0	0	0	0	0	0	0	0	0	0	0	0
$\{\boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
$\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$	91	80	93	84	93	85	89	79	91	80	96	90
$\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
$\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	9	20	7	16	7	15	11	21	9	20	4	10

Table 5: Average values of IC_{BL} and DIC in 100 observations for small sample cases $N = 25, 50, 100$ with informative prior ($\kappa_0 = 0.1$) in Case 2.

	Informative prior ($\kappa_0 = 0.1$)														
	$N = 25$					$N = 50$					$N = 100$				
Model	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
$\{x_1\}$	73.331	73.588	-35.951	1.429	1.685	145.125	145.278	-71.729	1.667	1.819	287.602	287.685	-142.892	1.818	1.901
	(5.347)	(5.347)	(2.674)	(0.000)	(0.009)	(7.837)	(7.837)	(3.919)	(0.000)	(0.009)	(10.764)	(10.765)	(5.382)	(0.000)	(0.009)
$\{x_2\}$	49.125	49.324	-23.794	1.537	1.736	89.938	90.054	-44.100	1.738	1.854	173.414	173.476	-85.777	1.860	1.922
	(4.651)	(4.649)	(2.322)	(0.068)	(0.041)	(8.048)	(8.049)	(4.023)	(0.029)	(0.022)	(12.967)	(12.968)	(6.483)	(0.012)	(0.016)
$\{x_3\}$	76.914	77.122	-37.694	1.525	1.733	152.120	152.239	-75.193	1.733	1.852	302.167	302.229	-150.154	1.859	1.921
	(5.308)	(5.311)	(2.656)	(0.060)	(0.031)	(8.618)	(8.617)	(4.309)	(0.033)	(0.019)	(13.281)	(13.282)	(6.641)	(0.013)	(0.010)
$\{x_1, x_2\}$	39.725	39.186	-18.390	2.944	2.406	64.713	63.985	-30.658	3.396	2.668	112.735	111.877	-54.530	3.675	2.817
	(4.009)	(4.009)	(2.003)	(0.078)	(0.043)	(7.585)	(7.583)	(3.792)	(0.033)	(0.031)	(11.646)	(11.646)	(5.823)	(0.013)	(0.023)
$\{x_1, x_3\}$	74.736	74.206	-35.903	2.929	2.399	146.744	146.015	-71.676	3.392	2.663	289.677	288.823	-143.001	3.675	2.821
	(5.423)	(5.425)	(2.713)	(0.067)	(0.034)	(7.816)	(7.817)	(3.909)	(0.036)	(0.018)	(10.954)	(10.952)	(5.476)	(0.013)	(0.015)
$\{x_2, x_3\}$	50.140	49.553	-23.550	3.041	2.454	91.506	90.742	-44.022	3.463	2.699	175.133	174.255	-85.708	3.717	2.839
	(4.971)	(4.974)	(2.489)	(0.088)	(0.047)	(7.982)	(7.988)	(3.992)	(0.047)	(0.032)	(13.000)	(12.999)	(6.500)	(0.018)	(0.019)
$\{x_1, x_2, x_3\}$	40.917	39.608	-18.247	4.423	3.115	66.202	64.591	-30.545	5.113	3.502	114.416	112.622	-54.443	5.529	3.736
	(4.183)	(4.188)	(2.094)	(0.099)	(0.056)	(7.639)	(7.641)	(3.825)	(0.051)	(0.039)	(11.596)	(11.595)	(5.797)	(0.020)	(0.026)

Standard deviations in parentheses

Table 6: Average values of IC_{BL} and DIC in 100 observations for small sample cases $N = 25, 50, 100$ with non-informative prior ($\kappa_0 = 100$) in Case 2.

	Non-informative prior ($\kappa_0 = 100$)														
	$N = 25$					$N = 50$					$N = 100$				
Model	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D	IC_{BL}	DIC	$\widehat{T}_N$	$2\widehat{b}_N$	p_D
$\{\boldsymbol{x}_1\}$	73.185	73.156	-35.593	1.999	1.970	145.721	145.705	-71.861	2.000	1.984	288.977	288.969	-143.489	2.000	1.992
	(6.051)	(6.052)	(3.025)	(0.000)	(0.009)	(8.455)	(8.454)	(4.227)	(0.000)	(0.010)	(11.986)	(11.986)	(5.993)	(0.000)	(0.010)
$\{\boldsymbol{x}_2\}$	45.601	45.571	-21.801	1.999	1.969	86.823	86.809	-42.412	2.000	1.985	172.104	172.095	-85.052	2.000	1.991
	(6.308)	(6.308)	(3.154)	(1.0.E-04)	(0.009)	(9.168)	(9.168)	(4.584)	(4.2.E-05)	(0.009)	(11.558)	(11.559)	(5.779)	(1.5.E-05)	(0.009)
$\{\boldsymbol{x}_3\}$	77.356	77.326	-37.678	1.999	1.969	153.189	153.174	-75.595	2.000	1.985	303.755	303.749	-150.878	2.000	1.993
	(7.055)	(7.054)	(3.527)	(1.1.E-04)	(0.010)	(8.989)	(8.990)	(4.495)	(4.0.E-05)	(0.008)	(11.788)	(11.788)	(5.894)	(1.4.E-05)	(0.009)
$\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$	29.687	28.622	-12.844	3.999	2.933	55.125	54.092	-25.563	3.999	2.966	104.365	103.350	-50.183	4.000	2.984
	(7.380)	(7.381)	(3.690)	(1.5.E-04)	(0.010)	(9.710)	(9.711)	(4.855)	(5.3.E-05)	(0.011)	(13.881)	(13.883)	(6.941)	(1.6.E-05)	(0.009)
$\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$	75.212	74.147	-35.607	3.999	2.934	147.694	146.661	-71.848	3.999	2.966	291.063	290.048	-143.532	4.000	2.984
	(6.228)	(6.230)	(3.114)	(1.3.E-04)	(0.011)	(8.535)	(8.537)	(4.267)	(4.3.E-05)	(0.009)	(12.005)	(12.004)	(6.003)	(1.5.E-05)	(0.011)
$\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$	47.295	46.227	-21.648	3.999	2.931	88.602	87.569	-42.302	3.999	2.966	174.231	173.213	-85.115	4.000	2.982
	(6.349)	(6.349)	(3.174)	(1.8.E-04)	(0.011)	(9.404)	(9.406)	(4.702)	(5.6.E-05)	(0.010)	(11.594)	(11.596)	(5.797)	(2.3.E-05)	(0.012)
$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	31.448	29.349	-12.725	5.998	3.898	56.957	54.907	-25.479	5.999	3.949	106.577	104.549	-50.289	5.999	3.971
	(7.589)	(7.590)	(3.794)	(2.2.E-04)	(0.013)	(10.021)	(10.023)	(5.010)	(6.9.E-05)	(0.014)	(13.769)	(13.769)	(6.884)	(2.3.E-05)	(0.013)

Standard deviations in parentheses

References

Akaike, H. (1973) 'Information theory as an extension of the maximum likelihood principle.' Second International Symposium on Information Theory Akademiai Kiado, Budapest pp. 267–281

Ando, T. (2007) 'Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models.' *Biometrika* 94(2), 443–458

Bedrick, E. J., and C. L. Tsai (1994) 'Model selection for multivariate regression in small samples.' *Biometrics* pp. 226–231

Burnham, K. P. (2002) 'Discussion of a paper by D.J. Spiegelhalter et al.' *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(4), 629

Gelfand, A. E., and S. K. Ghosh (1998) 'Model choice: A minimum posterior predictive loss approach.' *Biometrika* 85(1), 1–11

Hurvich, C.M., and C.L. Tsai (1989) 'Regression and time series model selection in small samples.' *Biometrika* 76(2), 297–307

Ibrahim, J. G., M. H. Chen, and D. Sinha (2001) 'Criterion-based methods for bayesian model assessment.' *Statistica Sinica* 11(2), 419–444

Kitagawa, G. (1997) 'Information criteria for the predictive evaluation of bayesian models.' *Communications in Statistics-Theory and Methods* 26(9), 2223–2246

Laud, P. W., and J. G. Ibrahim (1995) 'Predictive model selection.' *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 247–262

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002) 'Bayesian measures of model complexity and fit.' *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64(4), 583–639

Sugiura, N. (1978) 'Further analysts of the data by akaike's information criterion and the finite corrections.' *Communications in Statistics-Theory and Methods* 7(1), 13–26