Department of Social Systems and Management

Discussion Paper Series

No.1316

Variable Selection in a Bayesian Linear Regression Model via Generalized Bayesian Information Criterion

by

Satoshi KABE and Yuichiro KANAZAWA

March 2014

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Variable Selection in a Bayesian Linear Regression Model via Generalized Bayesian Information Criterion.

Satoshi KABE * Yuichiro KANAZAWA †

Department of Social Systems and Management,
University of Tsukuba, Ibaraki, Japan

In this paper, we consider the problem of variable selection in a Bayesian linear regression model with natural conjugate priors. Specifically, we first propose a variable selection criterion based on the generalized Bayesian information criterion (GBIC, Konishi et al., 2004). We then prove and show via simulation that the proposed criterion is consistent under the standard assumptions. We also compare the performance of our proposed criterion relative to other criteria (e.g., AIC, BIC, DIC) in small sample cases. The results of simulation studies show that the proposed GBIC-based criterion is not only consistent when the number of data increases, but also effective for small sample cases.

Keywords Variable selection; Bayesian linear regression model; Generalized Bayesian Information Criterion.
Mathematics Subject Classification Primary 62J05; Secondary 62F15.
Running title Variable selection for Bayesian linear model.

*Address correspondence to Satoshi Kabe Ph.D, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577, Japan; E-mail: k0420214@sk.tsukuba.ac.jp.
†Email: kanazawa@sk.tsukuba.ac.jp.

1 Introduction

The problem of variable selection in a multiple linear regression model is important in practice, and a number of model selection criteria have been proposed to evaluate the goodness-of-fit of candidate models. For example, Akaike's information criterion (AIC, Akaike, 1973) is widely used as a model selection criterion in terms of prediction: AIC is derived from the Kullback-Leibler divergence between unknown true density $f_Y(y)$ and the parametric model $g(y|\boldsymbol{\theta})$; AIC is designed to be an approximately unbiased estimator of expected log-likelihood. Sugiura (1978) suggested a finite bias-correction version of AIC (AIC_C) for a normal linear regression model: AIC_C is derived as an exact unbiased estimator of expected log-likelihood and asymptotically equivalent to AIC; Moreover, in small sample cases, AIC_C outperforms AIC (see Hurvich and Tsai, 1989). However, model selection based on AIC or AIC_C implicitly assumes that true density (or model) must be nested within the candidate model, i.e., $f_Y(y) \in \{g(y|\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ (see Hurvich and Tsai, 1991, p.500).

In Bayesian perspective, Schwarz (1978) proposed Bayesian information criterion (BIC). BIC selects a model M_r from the set of candidate models $\mathcal{M} \equiv \{M_1, M_2, \ldots, M_R\}$ based on the posterior probability $\Pr(M_r|\mathrm{Data})$: BIC is derived as an asymptotic approximation of the marginal likelihood and covers only models estimated by the maximum likelihood estimation. Unlike AIC and AIC_C derived from the Kullback-Leibler divergence, BIC does not require an assumption that candidate models contain the true model. Moreover, in cases true model exists in the set of candidate models, it is well-known that BIC is consistent, i.e., the posterior probability of choosing a true model converges to one when the number

of observed data goes to infinity (Nishii, 1984; Kubokawa and Srivastava, 2010). However it is known that BIC, though consistent for large samples, is not necessarily excellent in the sense of selecting variables in small sample sizes. Kubokawa and Srivastava (2010) suspected that one of the plausible reasons may be that BIC is far from the exact marginal distribution in small sample sizes. It is also known that BIC is asymptotically equivalent to a model selection based on Bayes factors (Kass and Raftery, 1995).

Bayes factor is defined as a ratio of marginal likelihoods for two different models evaluated at the observed data. Bayes factor enables us to introduce the prior information on the parameters. If the prior distribution is improper, however, it is well known that the Bayes factor does not work for model selection. To resolve this issue, many researchers (e.g., Aitkin, 1991; Gelfand and Dey, 1994; O'Hagan, 1995; Berger and Pericchi, 1996; Santis and Spezzaferri, 2001) proposed modifications of Bayes factor.

Bayes factors are possible criteria for the linear regression case. For example, variable selection via Bayes factors related to the Zellner (1986)'s g-prior is also consistent as seen in Fernández et al. (2001) and Liang et al. (2008). However, using a diffuse prior on the parameters in an effort to make it noninformative will lead to quite unexpected consequences. As was noted in Liang et al. (2008), "large spread of the prior induced by the noninformative choice of g has the unintended consequence of forcing the Bayes factor to favor the null model, the smallest model, regardless of the information in the data." and such phenomenon is called "Bartlett's paradox." Also, Bayes factors are known to be rather sensitive to the choice of the prior distributions on the parameters within each model. Even asymptotically, the influence of prior distributions does not vanish (see Kass and

Raftery, 1995; Fernández et al., 2001).

In hierarchical Bayesian perspective, deviance information criterion (DIC, Spiegelhalter et al., 2002) is widely used: DIC is easily computable and rather universally applicable Bayesian criterion for posterior predictive model comparison. Spiegelhalter et al. (2002) proposed a Bayesian measure of model complexity (i.e., effective number of parameters p_D) with respect to the hierarchical Bayesian model. This model complexity is obtained from the difference between the posterior mean of deviance and the deviance at the posterior mean of parameters. When the number of data is sufficiently large, DIC is given by adding p_D to the posterior mean of deviance. However, Ando (2007) shows that bias estimate of DIC tends to underestimate the true bias [1].

DIC can also be applied to the variable selection for non-hierarchical Bayesian linear regression models as well (van der Linde, 2005). However, as with AIC as pointed in Nishii (1984), even if the number of observed data increases, DIC will not consistently select the true model from a set of candidate models (see Spiegelhalter et al., 2002, p.613). Also, in small sample cases, DIC will need a bias corrected version comparable to AIC_C vis-a-vis AIC since Spiegelhalter et al. (2002) only gave an asymptotic justification of DIC.

Hence, it is desirable to search for a consistent Bayesian criterion in place of Bayes factor and DIC in the sense of selecting true variables for large samples and still performs well for small samples.

In this paper, we instead propose a generalized BIC (henceforth GBIC, Konishi

[1] As pointed by Robert and Titterington (2002) and Ando (2007), observed data is used twice in the construction of p_D: Indeed, observed data is used the first time to produce the posterior distribution from which the posterior mean of parameters is calculated; They are then used again to produce the posterior mean of deviance. This repeated use of observed data would appear to be a potential factor for overfitting (Robert and Titterington, 2002, p.621).

et al., 2004; Kawano and Konishi, 2009; Hirose et al., 2011; Matsui et al., 2013) based variable selection criterion with respect to the Bayesian linear regression model with natural conjugate priors. GBIC is derived as an approximation of marginal likelihood such as BIC, but, since GBIC includes terms discarded for the definition of BIC, such terms should improve the effectiveness of BIC in small sample cases (Neath and Cavanaugh, 1997).

We prove consistency of our proposed criterion under the standard assumptions and illustrate the proposed criterion is consistent in large sample cases. We then carry out performance comparisons of our proposed criterion relative to other prediction-base criteria such as AIC and DIC as well as the more traditional BIC in small sample cases to make our point clear.

The rest of this paper is organized as follows: Next section briefly describes the GBIC for Bayesian linear regression model with natural conjugate prior. In Section 3, we also prove the proposed criterion is consistent. Section 4 provides results of simulation study to illustrate the consistency of our proposed criterion and then shows the effectiveness for small sample cases via simulation. Finally, Section 5 discusses issues surrounding our proposed criterion.

2 Generalized Bayesian Information Criterion

With a finite set of R candidate models $\mathcal{M} \equiv \{M_1, M_2, \ldots, M_R\}$, posterior probability of choosing a model $M_k \in \mathcal{M}$ given the likelihood function $g_k(\boldsymbol{y}|\boldsymbol{\theta}_k) = \prod_{i=1}^{N} g_k(y_i|\boldsymbol{\theta}_k)$ and the prior distribution $\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)$ is obtained as

$$\Pr\left(M_k|\boldsymbol{y}, \boldsymbol{\psi}_k\right) = \frac{\Pr(M_k) \int g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)d\boldsymbol{\theta}_k}{\sum_{r=1}^{R} \Pr(M_r) \int g_r(\boldsymbol{y}|\boldsymbol{\theta}_r)\pi_r(\boldsymbol{\theta}_r|\boldsymbol{\psi}_r)d\boldsymbol{\theta}_r} \tag{2.1}$$

where $\boldsymbol{\theta}_k$ is a $d_k \times 1$ parameter vector and $\boldsymbol{\psi}_k$ is a $q_k \times 1$ hyper-parameter vector.

To show a heuristic derivation of GBIC, we assume that the prior probability of choosing a model M_k, $\Pr(M_k) = 1/R$ $(k = 1, 2, \ldots, R)$. Then GBIC selects a model among a finite set of candidate models $\mathcal{M}$ with maximizing the marginal likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\psi}_k, M_k) = \int g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)d\boldsymbol{\theta}_k. \tag{2.2}$$

To derive GBIC, we first rewrite the marginal likelihood in (2.2) as

$$\int g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)d\boldsymbol{\theta}_k = \int \exp\left[Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)\right] d\boldsymbol{\theta}_k \tag{2.3}$$

where $Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k) = \log\{g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)\}$. From the second-order Taylor expansion of $Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)$ in (2.3), we can approximate it around the posterior mode $\widehat{\boldsymbol{\theta}}_k = \text{argmax } Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)$ as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k) &\approx Q(\widehat{\boldsymbol{\theta}}_k; \boldsymbol{y}, \boldsymbol{\psi}_k) + (\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)' \left.\frac{\partial Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)}{\partial \boldsymbol{\theta}_k}\right|_{\boldsymbol{\theta}_k = \widehat{\boldsymbol{\theta}}_k} \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)' \left.\frac{\partial^2 Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'}\right|_{\boldsymbol{\theta}_k = \widehat{\boldsymbol{\theta}}_k} (\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k) \\ &= Q(\widehat{\boldsymbol{\theta}}_k; \boldsymbol{y}, \boldsymbol{\psi}_k) + \frac{1}{2}(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)' \left.\frac{\partial^2 Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'}\right|_{\boldsymbol{\theta}_k = \widehat{\boldsymbol{\theta}}_k} (\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k). \end{aligned} \tag{2.4}$$

From (2.3) and (2.4), $\exp\left[Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)\right]$ can be approximated as

$$\exp\left[Q(\boldsymbol{\theta}_k; \boldsymbol{y}, \boldsymbol{\psi}_k)\right] \approx g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k) \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)'(N\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k))(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)\right] \tag{2.5}$$

where

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k) = -\frac{1}{N}\left.\frac{\partial^2 \log\{g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)\}}{\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_k'}\right|_{\boldsymbol{\theta}_k=\widehat{\boldsymbol{\theta}}_k}. \tag{2.6}$$

From (2.5), the marginal likelihood in (2.2) is approximately

$$p(\boldsymbol{y}|\boldsymbol{\psi}_k, M_k) \approx g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\int \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)'(N\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k))(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)\right] d\boldsymbol{\theta}_k. \tag{2.7}$$

We can regard the integral in (2.7) as the multivariate normal distribution of $\boldsymbol{\theta}_k$ without normalizing constant and thus

$$\int (2\pi)^{-d_k/2}|N\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)'(N\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k))(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k)\right] d\boldsymbol{\theta}_k = 1$$

where $|\cdot|$ denotes the determinant of a matrix. Hence (2.6) can be rewritten as

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\psi}_k, M_k) &\approx g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\left[(2\pi)^{d_k/2}|N\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|^{-1/2}\right] \\ &= g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\left[(2\pi)^{d_k/2}N^{-d_k/2}|\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|^{-1/2}\right]. \end{aligned} \tag{2.8}$$

Taking logarithm on the both sides of (2.8) and then multiplying by -2, we have the generalized version of BIC [2] , presented by Konishi et al. (2004) such as

$$\begin{aligned} \mathrm{GBIC}(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_k, \boldsymbol{\psi}_k) = &-2\log\{g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\} + d_k\log(N) \\ &+ \log\{|\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|\} - d_k\log(2\pi). \end{aligned} \tag{2.9}$$

[2]Suppose that the number of observed data N is large enough and prior distribution $\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)$ is effectively uniform, we can treat $\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)$ as a constant (Burnham and Anderson, 2002) and the last two terms on the right-hand side of (2.9) can be ignored. Moreover, since $\pi_k(\boldsymbol{\theta}_k|\boldsymbol{\psi}_k)$ is effectively flat, posterior mode $\widehat{\boldsymbol{\theta}}_k$ can be replaced with the MLE $\widehat{\boldsymbol{\theta}}_k^{\mathrm{mle}}$ and we arrive at the BIC as

$$\mathrm{BIC}(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_k^{\mathrm{mle}}) = -2\log\{g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k^{\mathrm{mle}})\} + d_k\log(N).$$

In the fully Bayesian perspective, prior information on the parameter vector $\boldsymbol{\theta}_k$ plays an important role. As seen (2.9), the choice of the prior distribution $\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)$ has an influence on model comparison.

2.1 GBIC for Bayesian Linear Regression Model with Natural Conjugate Priors

We consider the linear regression model as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}_N, \sigma^2 \boldsymbol{I}_N\right) \tag{2.10}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of the response variable and $\boldsymbol{X}$ is a $N \times K$ non-stochastic matrix of explanatory variables. We assume that $\boldsymbol{X}$ has full column rank K. The parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ vector and error term $\boldsymbol{\varepsilon}$ follows a N-dimensional multivariate normal distribution $\mathcal{N}\left(\mathbf{0}_N, \sigma^2 \boldsymbol{I}_N\right)$ where $\mathbf{0}_N$ is a $N \times 1$ vector whose elements are zero and $\boldsymbol{I}_N$ is a $N \times N$ identity matrix. We assume that prior distribution of $\boldsymbol{\beta}$ is a K-dimensional multivariate normal distribution and that of σ^{-2} is a gamma distribution:

$$\boldsymbol{\beta}|\sigma^{-2} \sim \mathcal{N}\left(\boldsymbol{b}_0, \sigma^2 \boldsymbol{B}_0\right) \tag{2.11}$$

$$\sigma^{-2} \sim \mathcal{G}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right). \tag{2.12}$$

In particular, if the hyper-parameter $\boldsymbol{B}_0$ in (2.11) is specified as $\boldsymbol{B}_0 = (\kappa_0 \boldsymbol{X}'\boldsymbol{X})^{-1}$ where κ_0 (> 0) is unknown scalar hyper-parameter, then the prior distribution is well-known Zellner (1986)'s g-prior.

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

$$\boldsymbol{\beta}|\sigma^{-2}, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{b}_1, \sigma^2 \boldsymbol{B}_1\right) \tag{2.13}$$

$$\sigma^{-2}|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{G}\left(\frac{\nu_1}{2}, \frac{\lambda_1}{2}\right) \tag{2.14}$$

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

Applying the GBIC in (2.9) to the Bayesian linear regression model in (2.10) with natural conjugate priors in (2.11) and (2.12), we can obtain the $(K+1) \times (K+1)$ matrix $\boldsymbol{J}(\widehat{\boldsymbol{\theta}})$ in (2.6) as follows:

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}) = -\frac{1}{N}\begin{bmatrix} -\left(\frac{\nu_1 - 1}{\lambda_1}\right)\boldsymbol{B}_1^{-1} & \boldsymbol{0}_K \\ \boldsymbol{0}_K' & -\left(\frac{K}{2} + \frac{\nu_1}{2} - 1\right)\left(\frac{\nu_1 - 1}{\lambda_1}\right)^{-2} \end{bmatrix}. \tag{2.15}$$

The derivation of (2.15) is provided in Appendix A.

Taking determinant of (2.15) :

$$\begin{aligned} |\boldsymbol{J}(\widehat{\boldsymbol{\theta}})| &= \left|\frac{1}{N}\left(\frac{\nu_1 - 1}{\lambda_1}\right)\boldsymbol{B}_1^{-1}\right| \times \left(\frac{1}{N}\right)\left(\frac{K}{2} + \frac{\nu_1}{2} - 1\right)\left(\frac{\nu_1 - 1}{\lambda_1}\right)^{-2} \\ &= \left(\frac{\nu_1 - 1}{\lambda_1}\right)^{K-2}\left|\frac{1}{N}\boldsymbol{B}_1^{-1}\right| \times \left(\frac{1}{N}\right)\left(\frac{K}{2} + \frac{\nu_1}{2} - 1\right), \end{aligned} \tag{2.16}$$

we can compute the GBIC in (2.9) for the Bayesian linear regression model in (2.10) such as

$$\text{GBIC}(\boldsymbol{y}, \boldsymbol{X}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\psi}) = -2\log\{g(\boldsymbol{y}|\boldsymbol{X}, \widehat{\boldsymbol{\theta}})\pi(\widehat{\boldsymbol{\theta}}|\boldsymbol{\psi})\} + K\log(N)$$

$$+ (K-2)\log\left(\frac{\nu_1-1}{\lambda_1}\right) + \log\left|\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} + \frac{1}{N}\boldsymbol{B}_0^{-1}\right|$$
$$+ \log\left(\frac{K}{2} + \frac{\nu_1}{2} - 1\right) - (K+1)\log(2\pi) \tag{2.17}$$

where $\widehat{\boldsymbol{\theta}}$ is a vector of posterior modes of parameter $\boldsymbol{\beta}$ and σ^{-2} in (2.10), and $\boldsymbol{\psi}$ is a vector of hyper-parameters in (2.11) and (2.12). Substituting the posterior modes $\boldsymbol{b}_1$ and $(\nu_1-1)/\lambda_1$ into the parameters $\boldsymbol{\beta}$ and σ^{-2}, we have a log-likelihood $\log\{g(\boldsymbol{y}|\boldsymbol{X},\widehat{\boldsymbol{\theta}})\}$ and a logarithmic prior distribution $\log\{\pi(\widehat{\boldsymbol{\theta}}|\boldsymbol{\psi})\}$ as follows:

$$\log\{g(\boldsymbol{y}|\boldsymbol{X},\widehat{\boldsymbol{\theta}})\}$$
$$= -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\left(\frac{\nu_1-1}{\lambda_1}\right) - \frac{\nu_1-1}{2\lambda_1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_1)'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_1), \tag{2.18}$$

$$\log\{\pi(\widehat{\boldsymbol{\theta}}|\boldsymbol{\psi})\}$$
$$= -\frac{K}{2}\log(2\pi) + \frac{K}{2}\log\left(\frac{\nu_1-1}{\lambda_1}\right) - \frac{1}{2}\log|\boldsymbol{B}_0| - \frac{\nu_1-1}{2\lambda_1}(\boldsymbol{b}_1-\boldsymbol{b}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{b}_1-\boldsymbol{b}_0)$$
$$+ \frac{\nu_0}{2}\log\left(\frac{\lambda_0}{2}\right) - \log\Gamma\left(\frac{\nu_0}{2}\right) + \left(\frac{\nu_0}{2}-1\right)\log\left(\frac{\nu_1-1}{\lambda_1}\right) - \frac{\lambda_0}{2}\left(\frac{\nu_1-1}{\lambda_1}\right) \tag{2.19}$$

where $\Gamma(\cdot)$ is a gamma function.

3 Consistency of GBIC in (2.17)

Suppose that true model M_T exists within the finite set of candidate models (i.e., $M_T \in \mathcal{M}$), we denote the true model M_T as

$$M_T : \quad \boldsymbol{y} = \boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\varepsilon} \tag{3.1}$$

where $\boldsymbol{X}_T$ is a $N \times K_T$ matrix of true set of explanatory variables and $\boldsymbol{\beta}_T$ is a $K_T \times 1$ parameter vector. The error term $\boldsymbol{\varepsilon}$ is a $N \times 1$ normal random vector with mean zero and covariance matrix $\sigma^2 \boldsymbol{I}_N$ with the scalar σ^2 unknown.

We assume that the candidate model $M_r \in \mathcal{M}$ is specified in its stead as

$$M_r : \quad \boldsymbol{y} = \boldsymbol{X}_r \boldsymbol{\beta}_r + \boldsymbol{\varepsilon} \tag{3.2}$$

where $\boldsymbol{X}_r$ is a $N \times K_r$ matrix of explanatory variables and $\boldsymbol{\beta}_r$ is a $K_r \times 1$ parameter vector. Also, let us denote $\widehat{\boldsymbol{\theta}}_j^{\text{mle}}$ ($j = r$ or T) as a vector of the MLEs of unknown parameters $\boldsymbol{\beta}_j$ and σ^{-2} in (3.1) and (3.2), respectively.

Given the GBICs for the candidate model M_r in (3.2) and the true model M_T in (3.1), we show:

Theorem 3.1.

$$\Pr\left\{\text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_T, \widehat{\boldsymbol{\theta}}_T, \boldsymbol{\psi}_T) < \text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_r, \widehat{\boldsymbol{\theta}}_r, \boldsymbol{\psi}_r)\right\} \to 1 \tag{3.3}$$

as $N \to \infty$, where $M_r \neq M_T$ with quoted two lemmas in Fernández et al. (2001) in the appendix and th following assumptions

Assumption 3.1. *The elements of $\boldsymbol{X}_j$ ($j = r$ or T) in* (3.1) *and* (3.2) *are bounded and $\boldsymbol{X}_j' \boldsymbol{X}_j / N = O(1)$ as $N \to \infty$.*

Assumption 3.2. *The vector of hyper-parameters $\boldsymbol{\psi}_j$ ($j = r$ or T) is bounded, i.e., $\boldsymbol{\psi}_j = O(1)$ as $N \to \infty$.*

Assumption 3.3. *The logarithmic natural conjugate prior is bounded in probability, i.e., $\log\{\pi(\boldsymbol{\theta}_j | \boldsymbol{\psi}_j)\} = O_p(1)$ ($j = r$ or T) as $N \to \infty$.*

Proof. Using the fact that in large samples the posterior mode $\widehat{\boldsymbol{\theta}}_j$ ($j = r$ or T) of parameters $\boldsymbol{\beta}_j$ and σ^{-2} (i.e., $\boldsymbol{b}_{1j}$ and $(\nu_{1j}-1)/\lambda_{1j}$) is close to the MLE $\widehat{\boldsymbol{\theta}}_j^{\text{mle}}$, we here suppose that the number of data N is large enough so that the posterior mode $\widehat{\boldsymbol{\theta}}_j$ can be replaced with the MLE $\widehat{\boldsymbol{\theta}}_j^{\text{mle}}$. Then from (2.17), $-2\log\{g(\boldsymbol{y}|\boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_j)\pi(\widehat{\boldsymbol{\theta}}_j|\boldsymbol{\psi}_j)\}$ can be approximated by $-2\log\{g(\boldsymbol{y}|\boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_j^{\text{mle}})\pi(\widehat{\boldsymbol{\theta}}_j^{\text{mle}}|\boldsymbol{\psi}_j)\}$. Hence if N is large enough, $\text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\psi}_j)$ in (2.17) is equivalent to

$$\begin{aligned}\text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\psi}_j) = {} & N\log(2\pi) - N\log\left(\frac{N}{\text{RSS}_j}\right) + N \\ & - 2\log\{\pi(\widehat{\boldsymbol{\theta}}_j^{\text{mle}}|\boldsymbol{\psi}_j)\} + K_j\log(N) \\ & + (K_j - 2)\log\left(\frac{\nu_{1j}-1}{\lambda_{1j}}\right) + \log\left|\frac{1}{N}\boldsymbol{X}_j'\boldsymbol{X}_j + \frac{1}{N}\boldsymbol{B}_{0j}^{-1}\right| \\ & + \log\left(\frac{K_j}{2} + \frac{\nu_{1j}}{2} - 1\right) - (K_j + 1)\log(2\pi) \qquad (3.4)\end{aligned}$$

where $-2\log\{g(\boldsymbol{y}|\boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_j^{\text{mle}})\} = N\log(2\pi) - N\log(N/\text{RSS}_j) + N$.

Let us denote $\Delta_N = \text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_r, \widehat{\boldsymbol{\theta}}_r, \boldsymbol{\psi}_r) - \text{GBIC}(\boldsymbol{y}, \boldsymbol{X}_T, \widehat{\boldsymbol{\theta}}_T, \boldsymbol{\psi}_T)$ and from (3.4), we have

$$\Delta_N = -N\log\frac{\text{RSS}_T}{\text{RSS}_r} + (K_r - K_T)\log(N) + h_N \qquad (3.5)$$

where

$$\begin{aligned}h_N = {} & -2\log\{\pi(\widehat{\boldsymbol{\theta}}_r^{\text{mle}}|\boldsymbol{\psi}_r)\} + 2\log\{\pi(\widehat{\boldsymbol{\theta}}_T^{\text{mle}}|\boldsymbol{\psi}_T)\} \\ & + (K_r - 2)\log\left(\frac{\nu_{1r}-1}{\lambda_{1r}}\right) - (K_T - 2)\log\left(\frac{\nu_{1T}-1}{\lambda_{1T}}\right) \\ & + \log\frac{\left|\frac{1}{N}\boldsymbol{X}_r'\boldsymbol{X}_r + \frac{1}{N}\boldsymbol{B}_{0r}^{-1}\right|}{\left|\frac{1}{N}\boldsymbol{X}_T'\boldsymbol{X}_T + \frac{1}{N}\boldsymbol{B}_{0T}^{-1}\right|} + \log\frac{\left(\frac{K_r}{2} + \frac{\nu_{1r}}{2} - 1\right)}{\left(\frac{K_T}{2} + \frac{\nu_{1T}}{2} - 1\right)}\end{aligned}$$

$$- (K_r - K_T)\log(2\pi). \tag{3.6}$$

Using the fact that posterior mode $(\nu_{1j} - 1)/\lambda_{1j}$ ($j = r$ or T) is close to the MLE of σ^{-2} if N is large enough, we notice that h_N in (3.6) is bounded in probability under the assumptions 3.1, 3.2 and 3.3.

(i) $M_T \not\subset M_r$. In this case, it is sufficient to show that $\Delta_N \xrightarrow{p} \infty$ as $N \to \infty$. From lemma B.1 (ii), we notice that

$$\frac{\mathrm{RSS}_r/N}{\mathrm{RSS}_T/N} \xrightarrow{p} \frac{\sigma^2 + c_r}{\sigma^2} \ (> 1). \tag{3.7}$$

Therefore, we have from (3.5) and (3.6) the following limit

$$\begin{aligned} \Delta_N &= N\left(\log\frac{\mathrm{RSS}_r}{\mathrm{RSS}_T} + (K_r - K_T)\frac{\log(N)}{N} + \frac{h_N}{N}\right) \\ &\xrightarrow{p} \infty \end{aligned} \tag{3.8}$$

as $N \to \infty$, where $\lim_{N\to\infty} \frac{\log(N)}{N} = \lim_{N\to\infty} \frac{1/N}{1} = 0$.

(ii) $M_T \subset M_r$. Since we always have $K_r > K_T$, from lemma B.2 we notice that

$$\begin{aligned} \Pr\{0 < \Delta_N\} &= \Pr\left\{N\log\frac{\mathrm{RSS}_T}{\mathrm{RSS}_r} < (K_r - K_T)\log(N) + h_N\right\} \\ &\to \Pr\left\{\chi^2_{K_r - K_T} < \infty\right\} \\ &= 1. \end{aligned} \tag{3.9}$$

□

4 Simulation studies

In our simulation studies, we consider the candidate linear regression models by using three explanatory variables $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, where $\boldsymbol{x}_1$ is a $N \times 1$ vector whose elements are one and the other $N \times 1$ vectors $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are independently generated from the standard normal distributions. The explanatory variables in the candidate models are selected from the subsets of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ (i.e., $\{\boldsymbol{x}_1\}$, $\{\boldsymbol{x}_2\}$, $\{\boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$).

To examine the consistency of our proposed criterion based on GBIC in (2.17) via simulation, we generate the simulation data from true model $\boldsymbol{y} = 10\boldsymbol{x}_1 + 20\boldsymbol{x}_2 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_N, 5.0\boldsymbol{I}_N)$. We set the hyper-parameters in the natural conjugate priors in (2.11) and (2.12) such as $\boldsymbol{b}_0 = \mathbf{0}_K$, $\boldsymbol{B}_0 = 0.01\boldsymbol{I}_K$, $\nu_0 = 0.1$, and $\lambda_0 = 0.1$ and also for reference compare with DIC given by $\text{DIC} = -2\mathbf{E}_{\theta_k|y}[\log\{g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\}] + p_D$, where $\mathbf{E}_{\theta_k|y}[\cdot]$ denotes an expectation with respect to the posterior distribution and p_D is the effective number of parameters. In this simulation study, we generate 100 samples of two criteria, respectively, to examine the frequency of selecting the true model (see Figure 1). Figure 1 shows that GBIC consistently selects the true model and clearly outperforms DIC as the number of data N increases.

Next we carry out a simulation study to examine the performance of our proposed criterion in (2.17) for small sample cases ($N = 10, 20, 30$). In this simulation, we investigate effects of terms which are discarded as being asymptotically negligible, i.e., $\log\{\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\}$ and $\log\{|\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|\}$ in (2.9). We here compare the performance of our proposed criterion in (2.17) not only with DIC but also with

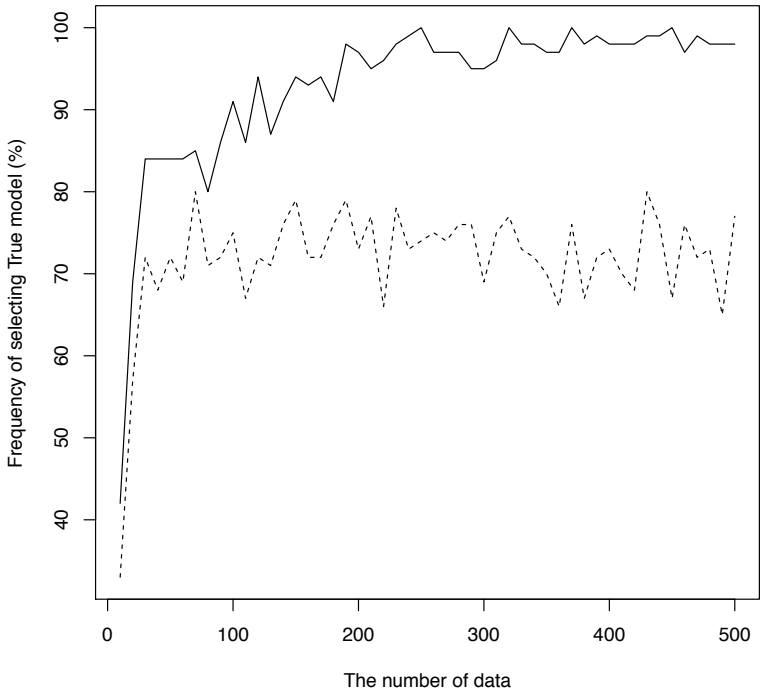

Figure 1: Frequency of selecting true model (%) with respect to GBIC and DIC. The solid line indicates GBIC and the dashed line indicates DIC.

the variants of GBIC in (2.9) such as Neath and Cavanaugh (1997) as follows:

$$\mathrm{GBIC}_1(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_k) = -2\log\{g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\} + d_k \log(N)$$
$$\mathrm{GBIC}_2(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_k) = -2\log\{g_k(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k)\} + d_k \log(N) + \log\{|\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|\}$$

where GBIC_1 throws out the effects of both $\log\{\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\}$ and $\log\{|\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_k)|\}$ such as BIC proposed by Schwarz (1978) and GBIC_2 only throws out the effect of $\log\{\pi_k(\widehat{\boldsymbol{\theta}}_k|\boldsymbol{\psi}_k)\}$. Notice that Neath and Cavanaugh (1997) evaluates the effects on BIC of truncation similarly under the maximum likelihood estimation. We also carry out the variable selection using AIC and BIC, which only deal with the mod-

els estimated by the maximum likelihood estimation, to compare the performance of AIC and BIC with that of GBIC.

In this simulation study, we set the true model $\boldsymbol{y} = 1.0\boldsymbol{x}_1 + 2.0\boldsymbol{x}_2 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_N, 0.5\boldsymbol{I}_N)$ to examine the performance of variable selection and also set the hyper-parameters in the natural conjugate priors in (2.11) and (2.12) such as $\boldsymbol{b}_0 = \mathbf{0}_K$, $\boldsymbol{B}_0 = \kappa_0 \boldsymbol{I}_K$ ($\kappa_0 = 1$ or 100), $\nu_0 = 0.1$, and $\lambda_0 = 0.1$. We here generate 500 samples of each criteria for the seven candidate models and report the frequency of selecting each candidate model in Table 1.

In the case of hyper-parameter $\kappa_0 = 1$, GBIC_2 and GBIC correctly select the true model (i.e., Model 4) as compared with the results of the other criteria. On the other hand, Table 1 also shows that AIC frequently selects the full model (i.e., Model 7) in all small sample cases ($N = 10, 20, 30$).

In the case of hyper-parameter $\kappa_0 = 100$, the performance of GBIC is quite well as compared with those of the other criteria. This result reflects the fact that BIC, GBIC_1 and GBIC_2 are far from the exact marginal likelihood relative to GBIC. Hence the terms discarded in the derivation of BIC should improve the performance of variable selection. On the other hand, AIC and DIC have a tendency to select the full model.

5 Conclusion and Discussion

In this paper, we consider variable selection in the Bayesian linear regression model with natural conjugate priors. In recent Bayesian modeling, prior information is aggressively applied to estimate the posterior distributions. For example, prior information from experts, theories, or other datasets plays an important role for

Table 1: Frequency of set of variables selected by each criteria in 500 samples for small sample cases ($N = 10, 20, 30$) when true set of variables is $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$.

Model	AIC	BIC	DIC	GBIC_1	GBIC_2	GBIC	AIC	BIC	DIC	GBIC_1	GBIC_2	GBIC
	$\kappa_0 = 1$, $N = 10$						$\kappa_0 = 100$, $N = 10$					
1. $\{\boldsymbol{x}_1\}$	0	0	0	0	0	0	0	0	0	0	0	0
2. $\{\boldsymbol{x}_2\}$	0	0	0	0	1	3	0	0	0	0	0	2
3. $\{\boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
4. $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$	377	392	446	475	477	478	386	398	383	398	442	485
5. $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
6. $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
7. $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	123	108	54	28	22	19	114	102	117	102	58	13
	$\kappa_0 = 1$, $N = 20$						$\kappa_0 = 100$, $N = 20$					
1. $\{\boldsymbol{x}_1\}$	0	0	0	0	0	0	0	0	0	0	0	0
2. $\{\boldsymbol{x}_2\}$	0	0	0	0	0	0	0	0	0	0	0	0
3. $\{\boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
4. $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$	393	442	429	473	484	488	406	440	397	440	471	494
5. $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
6. $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
7. $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	107	58	71	27	16	12	94	60	103	60	29	6
	$\kappa_0 = 1$, $N = 30$						$\kappa_0 = 100$, $N = 30$					
1. $\{\boldsymbol{x}_1\}$	0	0	0	0	0	0	0	0	0	0	0	0
2. $\{\boldsymbol{x}_2\}$	0	0	0	0	0	0	0	0	0	0	0	0
3. $\{\boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
4. $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$	400	450	420	465	482	485	416	459	413	459	485	500
5. $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
6. $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$	0	0	0	0	0	0	0	0	0	0	0	0
7. $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$	100	50	80	35	18	15	84	41	87	41	15	0

decision making (e.g., Rossi and Allenby, 2003). Bayesian analysis provides a unified and coherent way of thinking about decision problems and how to solve them using data and other information (Geweke, 2005).

In this paper, we first proved and illustrated via simulation that our GBIC-based criterion is consistent under the standard assumptions (see Figure 1). Figure 1 shows diverging asymptotic properties of GBIC and DIC: GBIC consistently selects the true model as the number of data increases, while DIC does not as expected. We then compare performance of the proposed GBIC in small sample cases (see Table 1) relative not only to more traditional AIC, BIC, and DIC, but also to the variants GBIC_1 and GBIC_2 to evaluate the effect of the terms discarded in the derivation of BIC on variable selection in linear regression setting.

Table 1 again shows that striking model selection performance differences between the proposed GBIC and DIC: When the prior is informative with $\kappa_0 = 1$, neither GBIC nor DIC can escape the influence of strong prior information when the numbers of sample are so small at $N = 10, 20$, and 30; however, it should be noted that the relative rates of misidentification of GBIC is far smaller than the rates of misidentification of DIC.

On the other hand, when the prior is noninformative with $\kappa_0 = 100$, GBIC's rates of misidentification are far smaller than the DIC regardless of the number of samples and the GBIC's misidentification rates rapidly decrease as the number of samples increases. Although GBIC has a very slight chance of choosing underspecified model $\{\boldsymbol{x}_2\}$ for $N = 10$ whether the prior is informative or noninformative, the overall small sample performance comparison clearly shows that GBIC is superior to DIC at least in this linear regression setting.

Table 1 also shows that our proposed criterion outperforms GBIC_1 and GBIC_2

in small sample cases as expected. Therefore in small sample cases where it is crucial to identify the true model under small sample, GBIC and not its truncated versions, GBIC_1 and GBIC_2, ought to be used. Overall our GBIC-based criterion is a useful Bayesian variable selection procedure in both large and small sample cases.

As an important direction for the future research, we would like to extend our proposed criterion to Bayesian econometric demand system models (e.g., Kabe and Kanazawa, 2013). Although such models are extensively used in empirical economic studies and policy decision making, their estimation is always constrained by the limited number - at most monthly, but quite often quarterly or semiannually - of data points and thus model performance comparisons must be carried out in small sample situations.

A Derivation of Eq. (2.15)

Using a fact that $p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi}) \propto g(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})$, we can rewrite the negative Hessian matrix in (2.6) as

$$\begin{aligned}\boldsymbol{J}(\widehat{\boldsymbol{\theta}}) &= -\frac{1}{N}\left.\frac{\partial^2 \log\{g(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\ &= -\frac{1}{N}\left.\frac{\partial^2 \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \end{aligned} \tag{A.1}$$

The logarithmic posterior distribution $\log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}$ in (A.1) for the Bayesian linear regression model in (2.10) is given from (2.13) and (2.14) as

$$\begin{aligned}\log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\} = &\frac{K}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) \\ &+ \left(\frac{\nu_1}{2}-1\right)\log\sigma^{-2} - \frac{\lambda_1}{2}\sigma^{-2} + \{\text{constant term}\}. \end{aligned} \tag{A.2}$$

where parameter vector $\boldsymbol{\theta}$ has the parameters $\boldsymbol{\beta}$ and σ^{-2}.

The first derivative of (A.2) with respect to $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\frac{\partial \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\left\{-\frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right\} \\ &= -\frac{\sigma^{-2}}{2}\frac{\partial}{\partial\boldsymbol{\beta}}\left\{\boldsymbol{\beta}'\boldsymbol{B}_1^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 - \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{\beta} + \boldsymbol{b}_1'\boldsymbol{B}_1^{-1}\boldsymbol{b}_1\right\} \\ &= -\frac{\sigma^{-2}}{2}\left\{2\boldsymbol{B}_1^{-1}\boldsymbol{\beta} - 2\boldsymbol{B}_1^{-1}\boldsymbol{b}_1 + 0\right\} \\ &= -\sigma^{-2}\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) \end{aligned} \tag{A.3}$$

where matrix $\boldsymbol{B}_1^{-1}$ is symmetric.

Next we take a first derivative of (A.2) with respect to σ^{-2} as follows:

$$\begin{aligned}
&\frac{\partial \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial \sigma^{-2}} \\
&= \frac{\partial}{\partial \sigma^{-2}} \left\{ \frac{K}{2}\log\sigma^{-2} - \frac{\sigma^{-2}}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) + \left(\frac{\nu_1}{2}-1\right)\log\sigma^{-2} - \frac{\lambda_1}{2}\sigma^{-2} \right\} \\
&= \frac{K}{2}\frac{1}{\sigma^{-2}} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) + \left(\frac{\nu_1}{2}-1\right)\frac{1}{\sigma^{-2}} - \frac{\lambda_1}{2} \\
&= \left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)(\sigma^{-2})^{-1} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) - \frac{\lambda_1}{2}. \qquad \text{(A.4)}
\end{aligned}$$

From (A.3), the second derivative of (A.2) with respect to $\boldsymbol{\beta}$ is obtained as

$$\begin{aligned}
\frac{\partial^2 \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} &= \frac{\partial}{\partial\boldsymbol{\beta}}\left\{-\sigma^{-2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}\right\} \\
&= -\sigma^{-2}\boldsymbol{B}_1^{-1} \qquad \text{(A.5)}
\end{aligned}$$

and with respect to σ^{-2} is also given by

$$\begin{aligned}
\frac{\partial^2 \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\boldsymbol{\beta}\partial\sigma^{-2}} &= \frac{\partial}{\partial\sigma^{-2}}\left\{-\sigma^{-2}\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right\} \\
&= -\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1). \qquad \text{(A.6)}
\end{aligned}$$

Similarly, from (A.4), we take a second derivative of (A.2) with respect to σ^{-2} as follows

$$\begin{aligned}
\frac{\partial^2 \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\sigma^{-2}\partial\sigma^{-2}} &= \frac{\partial}{\partial\sigma^{-2}}\left\{\left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)(\sigma^{-2})^{-1} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1) - \frac{\lambda_1}{2}\right\} \\
&= \frac{\partial}{\partial\sigma^{-2}}\left\{\left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)(\sigma^{-2})^{-1}\right\} \\
&= -\left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)\left(\sigma^{-2}\right)^{-2} \qquad \text{(A.7)}
\end{aligned}$$

and with respect to $\boldsymbol{\beta}'$ is given by

$$\begin{aligned}\frac{\partial^2 \log\{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\psi})\}}{\partial\sigma^{-2}\partial\boldsymbol{\beta}'} &= \frac{\partial}{\partial\boldsymbol{\beta}'}\left\{\left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)(\sigma^{-2})^{-1}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)-\frac{\lambda_1}{2}\right\}\\ &= \frac{\partial}{\partial\boldsymbol{\beta}'}\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1}(\boldsymbol{\beta}-\boldsymbol{b}_1)\right\}\\ &= -(\boldsymbol{\beta}-\boldsymbol{b}_1)'\boldsymbol{B}_1^{-1} \qquad \text{(A.8)}\end{aligned}$$

Substituting the posterior modes $\boldsymbol{b}_1$ and $(\nu_1-1)/\lambda_1$ into parameters $\boldsymbol{\beta}$ and σ^{-2} in (A.5), (A.6), (A.7) and (A.8), we have the negative Hessian matrix $\boldsymbol{J}(\widehat{\boldsymbol{\theta}})$ expressed by

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}) = -\frac{1}{N}\begin{bmatrix} -\left(\frac{\nu_1-1}{\lambda_1}\right)\boldsymbol{B}_1^{-1} & \boldsymbol{0}_K \\ \boldsymbol{0}_K' & -\left(\frac{K}{2}+\frac{\nu_1}{2}-1\right)\left(\frac{\nu_1-1}{\lambda_1}\right)^{-2}\end{bmatrix}, \qquad \text{(A.9)}$$

where $\boldsymbol{0}_K$ is a $K\times 1$ vector whose elements are zero.

B Fernández et al. (2001)'s lemmas

Lemma B.1.

(i) *If the true model M_T is nested within or is equal to the candidate model M_r* (i.e., $M_T \subseteq M_r$),

$$\frac{(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}})}{N} \xrightarrow{p} \sigma^2. \qquad \text{(B.1)}$$

(ii) *Assuming*

$$\frac{\boldsymbol{\beta}_T'\boldsymbol{X}_T'(\boldsymbol{I}_N-\boldsymbol{P}_r)\boldsymbol{X}_T\boldsymbol{\beta}_T}{N} \xrightarrow{p} c_r \in (0,\infty) \qquad \text{(B.2)}$$

that for any model M_r that does not nest M_T (i.e., $M_T \not\subset M_r$), *where* $\boldsymbol{P}_r = \boldsymbol{X}_r(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r'$, *we obtain*

$$\frac{(\boldsymbol{y} - \boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y} - \boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}})}{N} \xrightarrow{p} \sigma^2 + c_r. \tag{B.3}$$

Proof. Let us denote $\boldsymbol{P}_r = \boldsymbol{X}_r(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r'$, where $\boldsymbol{X}_r$ is a $N \times K_r$ design matrix of the candidate model $M_r \in \mathcal{M}$ and $\boldsymbol{P}_r$ and $(\boldsymbol{I}_N - \boldsymbol{P}_r)$ are known to be $N \times N$ symmetric and idempotent matrices. The residual sum of squares for candidate model M_r is given by

$$\begin{aligned}
&(\boldsymbol{y} - \boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y} - \boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\text{mle}}) \\
&= \boldsymbol{y}'\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\boldsymbol{y} \\
&= \left(\boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\varepsilon}\right)'\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\left(\boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\varepsilon}'\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\boldsymbol{\varepsilon} + 2\boldsymbol{\beta}_T'\boldsymbol{X}_T'\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\boldsymbol{\varepsilon} + \boldsymbol{\beta}_T'\boldsymbol{X}_T'\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\boldsymbol{X}_T\boldsymbol{\beta}_T.
\end{aligned} \tag{B.4}$$

(i) $M_T \subseteq M_r$. We assume that $\boldsymbol{X}_r$ is partitioned as $\boldsymbol{X}_r = [\boldsymbol{X}_T \quad \boldsymbol{Z}_r]$ where $\boldsymbol{Z}_r$ is a $N \times S_r$ matrix of additional explanatory variables. When the true variables $\boldsymbol{X}_T$ is regressed on $\boldsymbol{X}_r$, we have

$$\left(\boldsymbol{X}_r'\boldsymbol{X}_r\right)^{-1}\boldsymbol{X}_r'\boldsymbol{X}_T = \begin{bmatrix} \boldsymbol{I}_{K_T} \\ \boldsymbol{0}_{S_r \times K_T} \end{bmatrix}. \tag{B.5}$$

where $\boldsymbol{0}_{S_r \times K_T}$ is a $S_r \times K_T$ matrix whose elements are zero. Then we notice that

$$\begin{aligned}
\left(\boldsymbol{I}_N - \boldsymbol{P}_r\right)\boldsymbol{X}_T &= \left\{\boldsymbol{I}_N - \boldsymbol{X}_r\left(\boldsymbol{X}_r'\boldsymbol{X}_r\right)^{-1}\boldsymbol{X}_r'\right\}\boldsymbol{X}_T \\
&= \boldsymbol{X}_T - \boldsymbol{X}_r\left(\boldsymbol{X}_r'\boldsymbol{X}_r\right)^{-1}\boldsymbol{X}_r'\boldsymbol{X}_T
\end{aligned}$$

$$= \boldsymbol{X}_T - [\boldsymbol{X}_T \quad \boldsymbol{Z}_r] \begin{bmatrix} \boldsymbol{I}_{K_T} \\ \boldsymbol{0}_{S_r \times K_T} \end{bmatrix}$$

$$= \boldsymbol{0}_{N \times K_T}. \tag{B.6}$$

From (B.6), the residual sum of squares for candidate model M_r in (B.4) can be rewritten as

$$\begin{aligned}
&(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}}) \\
&= \boldsymbol{\varepsilon}' \left(\boldsymbol{I}_N - \boldsymbol{P}_r\right) \boldsymbol{\varepsilon} + 2\boldsymbol{\beta}_T' \boldsymbol{X}_T' \left(\boldsymbol{I}_N - \boldsymbol{P}_r\right) \boldsymbol{\varepsilon} + \boldsymbol{\beta}_T' \boldsymbol{X}_T' \left(\boldsymbol{I}_N - \boldsymbol{P}_r\right) \boldsymbol{X}_T \boldsymbol{\beta}_T \\
&= \boldsymbol{\varepsilon}' \left(\boldsymbol{I}_N - \boldsymbol{P}_r\right) \boldsymbol{\varepsilon}.
\end{aligned} \tag{B.7}$$

Since the expectation of (B.7) is given by

$$\begin{aligned}
\mathbf{E}\left[(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}})\right] &= \mathbf{E}\left[\boldsymbol{\varepsilon}' \left(\boldsymbol{I}_N - \boldsymbol{P}_r\right) \boldsymbol{\varepsilon}\right] \\
&= \sigma^2 \text{tr}\left\{\boldsymbol{I}_N - \boldsymbol{P}_r\right\} \\
&= (N - K_T - S_r)\sigma^2,
\end{aligned} \tag{B.8}$$

we have as $N \to \infty$, $N - K_r - S_r \approx N$ and

$$\frac{(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}})'(\boldsymbol{y} - \boldsymbol{X}_r \widehat{\boldsymbol{\beta}}_r^{\text{mle}})}{N} \xrightarrow{p} \sigma^2. \tag{B.9}$$

(ii) $M_T \not\subset M_r$. We suppose that

$$\frac{\boldsymbol{\beta}_T' \boldsymbol{X}_T' (\boldsymbol{I}_N - \boldsymbol{P}_r) \boldsymbol{X}_T \boldsymbol{\beta}_T}{N} \xrightarrow{p} c_r \in (0, \infty) \tag{B.10}$$

and expectation of (B.4) is given by

$$\begin{aligned}\mathbf{E}\left[(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\mathrm{mle}})'(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\mathrm{mle}})\right] &= \mathbf{E}\left[\boldsymbol{\varepsilon}'\left(\boldsymbol{I}_N-\boldsymbol{P}_r\right)\boldsymbol{\varepsilon}\right]+\boldsymbol{\beta}_T'\boldsymbol{X}_T'\left(\boldsymbol{I}_N-\boldsymbol{P}_r\right)\boldsymbol{X}_T\boldsymbol{\beta}_T \\ &= (N-K_r)\sigma^2+\boldsymbol{\beta}_T'\boldsymbol{X}_T'\left(\boldsymbol{I}_N-\boldsymbol{P}_r\right)\boldsymbol{X}_T\boldsymbol{\beta}_T. \end{aligned} \tag{B.11}$$

From (B.10) and (B.11), we have as $N\to\infty$, $N-K_r\approx N$ and

$$\frac{(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\mathrm{mle}})'(\boldsymbol{y}-\boldsymbol{X}_r\widehat{\boldsymbol{\beta}}_r^{\mathrm{mle}})}{N}\xrightarrow{p}\sigma^2+c_r. \tag{B.12}$$

□

Lemma B.2. *If the candidate model M_r nests the true model M_T* (i.e., $M_T\subset M_r$)*,*

$$N\log\frac{\mathrm{RSS}_T}{\mathrm{RSS}_r}\xrightarrow{d}\chi^2_{K_r-K_T}, \tag{B.13}$$

where RSS_j *is a residual sum of squares obtained by* $\mathrm{RSS}_j=(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j^{\mathrm{mle}})'(\boldsymbol{y}-\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j^{\mathrm{mle}})$ *and* $\chi^2_{K_r-K_T}$ *is a chi-square distribution with degree of freedom* K_r-K_T*.*

Proof. Please refer to likelihood ratio test literature (e.g., Amemiya, 1985). □

References

Aitkin, M. (1991). Posterior bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–142.

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

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 94(2):443–458.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information theoretic approach*. Springer(c2002), New York.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514.

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*, volume 537. Wiley. com.

Hirose, K., Kawano, S., Konishi, S., and Ichikawa, M. (2011). Bayesian information criterion and selection of the number of factors in factor analysis models. *Journal of Data Science*, 9(2):243–259.

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

Hurvich, C. M. and Tsai, C. L. (1991). Bias of the corrected aic criterion for underfitted regression and time series models. *Biometrika*, 78(3):499–509.

Kabe, S. and Kanazawa, Y. (2013). Estimating the markov-switching almost ideal demand systems: a bayesian approach. *Empirical Economics (forthcoming)*.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *The Journal of the American Statistical Association*, 90(430):773–795.

Kawano, S. and Konishi, S. (2009). Nonlinear logistic discrimination via regularized gaussian basis expansions. *Communications in Statistics - Simulation and Computation*, 38(7):1414–1425.

Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43.

Kubokawa, T. and Srivastava, M. S. (2010). An empirical bayes information criterion for selecting variables in linear mixed models. *J. Japan Statist. Soc*, 40(1):111–130.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481).

Matsui, H., Misumi, T., and Kawano, S. (2013). Model selection criteria for the varying-coefficient modelling via regularized basis expansions. *Journal of Statistical Computation and Simulation*, pages 1–10.

Neath, A. A. and Cavanaugh, J. E. (1997). Regression and time series model selection using variants of the schwarz information criterion. *Communications in Statistics-Theory and Methods*, 26(3):559–580.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–138.

Robert, C. P. and Titterington, D. M. (2002). Discussion of a paper by D.J. Spiegelhalter et al. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):621.

Rossi, P. E. and Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328.

Santis, F. D. and Spezzaferri, F. (2001). Consistent fractional bayes factor for nested normal linear models. *Journal of statistical planning and inference*, 97(2):305–321.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Mathematical Statistics*, 42(3):1003–1009.

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

van der Linde, A. (2005). Dic in variable selection. *Statistica Neerlandica*, 59(1):45–56.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.