Bayesian Analysis (2008) **3**, Number 3, pp. 585–614

Bayesian Variable Selection and Computation for Generalized Linear Models with Conjugate Priors

Ming-Hui Chen*, Lan Huang†, Joseph G. Ibrahim‡ and Sungduk Kim§

Abstract. In this paper, we consider theoretical and computational connections between six popular methods for variable subset selection in generalized linear models (GLM's). Under the conjugate priors developed by Chen and Ibrahim (2003) for the generalized linear model, we obtain closed form analytic relationships between the Bayes factor (posterior model probability), the Conditional Predictive Ordinate (CPO), the L measure, the Deviance Information Criterion (DIC), the Aikiake Information Criterion (AIC), and the Bayesian Information Criterion (BIC) in the case of the linear model. Moreover, we examine computational relationships in the model space for these Bayesian methods for an arbitrary GLM under conjugate priors as well as examine the performance of the conjugate priors of Chen and Ibrahim (2003) in Bayesian variable selection. Specifically, we show that once Markov chain Monte Carlo (MCMC) samples are obtained from the full model, the four Bayesian criteria can be simultaneously computed for all possible subset models in the model space. We illustrate our new methodology with a simulation study and a real dataset.

Keywords: Bayes factor, Conditional Predictive Ordinate, Conjugate prior, L measure, Poisson regression, Logistic regression

1 Introduction

Bayesian variable selection is still one of the most theoretically and computationally challenging problems encountered in practice due to issues regarding i) prior elicitation, ii) analytic evaluation of the model selection criterion, and iii) numerical computation of the criterion for all possible models in the model space. These issues have been discussed by many authors for various linear and generalized linear models including George and McCulloch (1993), Laud and Ibrahim (1995), George et al. (1996), Raftery (1996), Smith and Kohn (1996), George and McCulloch (1997), Raftery et al. (1997), Brown et al. (1998), Brown et al. (2002), Clyde (1999), Chen et al. (1999), Dellaportas and Forster (1999), Ibrahim et al. (1999), Chipman et al. (1998), Chipman et al. (2001), Chipman et al. (2003), George (2000), George and Foster (2000),

*Department of Statistics, University of Connecticut, Storrs, CT, http://www.stat.uconn.edu/~mhchen

†SRAB, National Cancer Institute, Rockville, MD, mailto:huangla@mail.nih.gov

‡Department of Biostatistics, University of North Carolina, Chapel Hill, NC, mailto:ibrahim@bios.unc.edu

§Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, Rockville, MD, mailto:kims2@mail.nih.gov

 DOI:10.1214/08-BA323

Ibrahim et al. (2000), Ntzoufras et al. (2003), and Chen et al. (2003). Clyde and George (2004) present an excellent review article on Bayesian model selection and uncertainty, and give an excellent exposition of the theoretical and computational issues involved in Bayesian variable selection and Bayesian model uncertainty in general. An entire monograph devoted to Bayesian model selection is given by Lahiri (2001).

One of the important unresolved issues in Bayesian model selection and Bayesian variable selection in particular is what the analytic or empirical connections are between the various methods. For example, it is not clear what the relationship is between BIC and DIC, or DIC and the L measure, and whether one is a monotonic function of the other, and whether one can compute BIC from DIC or vice versa. A related question is that if one has MCMC samples from the full model, how can those samples be used to obtain all four Bayesian criteria mentioned above. To answer these questions, we investigate the following in this paper: (i) for the normal linear model with conjugate priors, we obtain analytic relationships between the Bayes factor, CPO, the L measure, DIC, AIC, and BIC, and (ii) for the class of GLM's we show via the development of several theorems and identities how one can compute all of these Bayesian criteria simultaneously using only an MCMC sample from the full model.

The relationships obtained in (i) for the linear model shed light on the behavior and connections between these criteria for GLM's. The development of (ii) above is important and useful since it establishes the computational relationships in the model space for each of the four Bayesian criteria and shows that for variable subset selection in GLM's using the conjugate priors of Chen and Ibrahim (2003), we can compute the four Bayesian criteria for all possible 2^p subset models using only an MCMC sample from the full model with p covariates. Another important issue we examine in this paper is the performance of the conjugate priors proposed by Chen and Ibrahim (2003) in Bayesian variable subset selection. We demonstrate that these priors perform quite well in this context, and they are easy to specify and computationally feasible.

The rest of this paper is organized as follows. Section 2 gives formulas for each of the criteria under the conjugate priors of Chen and Ibrahim (2003) for GLM's and Section 3 develops the theoretical connections between the six criteria for the normal linear model. Section 4 establishes the computational connections in the model space for the four Bayesian criteria and several key identities and theorems that are needed. Section 5 presents a detailed simulation study examining various properties of the six criteria, and Section 6 presents a real data example. We conclude the article with brief remarks in Section 7. All proofs are given in the Appendix.

2 The Method

2.1 Model and Notation

Suppose that $\{(\mathbf{x}_i, y_i),\ i = 1, 2, \ldots, n\}$ are independent observations, where y_i is the response variable, and $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ik})'$ is a $(k+1) \times 1$ random vector of covariates. Let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m =$

$1, 2, \ldots, \mathcal{K}$, where $\mathcal{K}$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model. Also, let $\boldsymbol{\beta}^{(\mathcal{K})} = (\beta_0, \beta_1, \ldots, \beta_k)'$ denote the regression coefficients for the full model including an intercept, and let $\mathbf{x}_i^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ denote $k_m \times 1$ vectors of covariates and regression coefficients for model m with an intercept, and a specific choice of $k_m - 1$ covariates. We write $\mathbf{x}_i = (\mathbf{x}_i^{(m)'}, \mathbf{x}_i^{(-m)'})'$, and $\boldsymbol{\beta}^{(\mathcal{K})} = (\boldsymbol{\beta}^{(m)'}, \boldsymbol{\beta}^{(-m)'})'$, where $\mathbf{x}_i^{(-m)}$ is $\mathbf{x}_i$ with $\mathbf{x}_i^{(m)}$deleted and $\boldsymbol{\beta}^{(-m)}$ is $\boldsymbol{\beta}^{(\mathcal{K})}$ with $\beta^{(m)}$ deleted.

Under model m, the generalized linear model (GLM) is assumed for $[y_i|\mathbf{x}_i^{(m)}]$, which has the conditional density given by

$$f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)}, \tau) = \exp\left[a_i^{-1}(\tau)\{y_i\theta_i^{(m)} - b(\theta_i^{(m)})\} + c(y_i, \tau)\right], \ i = 1, 2, \ldots, n, \quad (1)$$

where $\theta_i^{(m)} = \theta(\eta_i^{(m)})$ is the canonical parameter, $\eta_i^{(m)} = \mathbf{x}_i^{(m)'}\boldsymbol{\beta}^{(m)}$, and τ is a dispersion parameter. The functions a, b and c determine a particular family in the class. The functions $a_i(\tau)$ are commonly of the form $a_i(\tau) = \tau^{-1}w_i^{-1}$, where the w_i's are known weights. For ease of exposition, we assume throughout that $\tau = 1$ and $w_i = 1$, as, for example, in logistic and Poisson regression. The methods proposed here can be easily extended to the case when τ is unknown. Under this assumption, (1) can be rewritten as

$$f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)}) = \exp\left\{y_i\theta_i^{(m)} - b(\theta_i^{(m)}) + c(y_i)\right\}, \ i = 1, 2, \ldots, n. \quad (2)$$

2.2 Prior and Posterior

In the context of Bayesian variable selection, a prior distribution for $\boldsymbol{\beta}^{(m)}$ needs to be specified for each model in the model space $\mathcal{M}$. To this end, we consider a conjugate prior for the GLM proposed by Chen and Ibrahim (2003). Under model m, the conjugate prior is of the form

$$\begin{aligned} &\pi(\boldsymbol{\beta}^{(m)}|\mathbf{y}_0, a_0, m) \\ &\propto \prod_{i=1}^{n} \exp\left[a_0\{y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)})\}\right] = \exp\left[a_0\{\mathbf{y}_0'\boldsymbol{\theta}^{(m)} - J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\}\right], \end{aligned} \quad (3)$$

where $a_0 > 0$ is a scalar prior parameter, $\mathbf{y}_0 = (y_{01}, \ldots, y_{0n})'$ is an $n \times 1$ vector of prior parameters, J is an $n \times 1$ vector of ones, and $\boldsymbol{b}(\boldsymbol{\theta}^{(m)}) = (b(\theta_1^{(m)}), \ldots, b(\theta_n^{(m)}))'$ is an $n \times 1$ vector of the $b(\theta_i^{(m)})$'s. As discussed in Chen and Ibrahim (2003), y_{0i} can be viewed as a prior prediction for the marginal mean of y_i at $\mathbf{x}_i$. Thus, in eliciting $\mathbf{y}_0$, the user must focus on a prediction (or guess) for $E(\mathbf{y})$, which narrows the possibilities for choosing $\mathbf{y}_0$. Moreover, the specification of all y_{0i} equal has an appealing interpretation. A prior specification with $y_{01} = \ldots = y_{0n}$ implies a prior in which the prior modes of the slopes in the regression model are the same, but the prior modes of intercepts in the regression model vary. For example, a prior with $y_{0i} = 0.5$ will have the same modes of slopes but a different mode of intercept than a prior with $y_{0i} = 0.1$. This is intuitively appealing since in this case the prior prediction on y_{0i} does not depend on the i^{th} subject's specific information. Mathematically, this result was established in Chen and Ibrahim (2003).

The details are as follows. Suppose we drop model index m. Let $\boldsymbol{\mu}_0$ be any prespecified $p \times 1$ vector, where $p = k + 1$. Suppose we take

$$\mathbf{y}_0 = \dot{\boldsymbol{b}}(\boldsymbol{\theta}) = \dot{\boldsymbol{b}}(\theta(X\boldsymbol{\mu}_0)),$$

where $\dot{\boldsymbol{b}}(\boldsymbol{\theta})$ is the gradient vector of $\boldsymbol{b}(\boldsymbol{\theta})$. Then, the conjugate prior yields a prior mode of $\boldsymbol{\beta}$ equal to $\boldsymbol{\mu}_0$. Now we can see that $\boldsymbol{\mu}_0 = (\beta_0, 0, \ldots, 0)'$ yields $y_{01} = y_{02} = \cdots = y_{0n} = \dot{b}(\theta(\beta_0))$. On the other hand, as under some mild conditions, the prior mode is unique, and, hence, the specification of $\mathbf{y}_0 = y_0\mathbf{1}$ leads to the prior mode $\boldsymbol{\mu}_0 = (\beta_0, 0, \ldots, 0)'$, where β_0 satisfies $\dot{b}(\theta(\beta_0)) = y_0$. For instance, under normal linear regression, we can show that the prior mode $\boldsymbol{\mu}_0$ of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\mu}_0 = (X'X)^{-1}X'\mathbf{y}_0.$$

If we specify $\mathbf{y}_0 = y_0\mathbf{1}$, we have

$$\boldsymbol{\mu}_0 = (y_0, 0, 0, \ldots, 0)',$$

which implies that all the slopes are 0 while the intercept is equal to y_0. This attractive feature allows us to do sensitivity analyses by varying the intercepts in the prior. The parameter a_0 in (3) can be generally viewed as a precision parameter that quantifies the strength of our prior belief in $\mathbf{y}_0$.

In the context of Bayesian variable selection, (3) specifies the priors for all models in $\mathcal{M}$ in an automatic and systematic fashion. Although various theoretical properties of (3) were examined in Chen and Ibrahim (2003) in a great detail, it is not clear how well this type of the prior performs in the context of Bayesian variable selection.

Now, under model m, the posterior distribution of $\boldsymbol{\beta}^{(m)}$ with the conjugate prior (3) is given by

$$\begin{aligned} \pi(\boldsymbol{\beta}^{(m)}|D, m) &\propto \exp\left\{\mathbf{y}'\boldsymbol{\theta}^{(m)} - J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\right\} \pi(\boldsymbol{\beta}^{(m)}|\mathbf{y}_0, a_0, m) \\ &\propto \exp\left\{(\mathbf{y} + a_0\mathbf{y}_0)'\boldsymbol{\theta}^{(m)} - (1 + a_0)J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\right\}, \end{aligned} \tag{4}$$

where $D = \{(y_i, \mathbf{x}_i),\ i = 1, 2, \ldots, n\}$ denotes the observed data. From (4), we can see that under the conjugate prior, the resulting posterior has a very attractive form. Furthermore, when $a_0 \to 0$, the posterior $\pi(\boldsymbol{\beta}^{(m)}|D, m)$ in (4) reduces to

$$\pi(\boldsymbol{\beta}^{(m)}|D, m) \propto \exp\left\{\mathbf{y}'\boldsymbol{\theta}^{(m)} - J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\right\},$$

which is the posterior distribution based on an improper uniform prior for $\boldsymbol{\beta}^{(m)}$.

2.3 Variable Selection Criteria

In this section, we consider four Bayesian model assessment criteria, namely, Conditional Predictive Ordinate (CPO) statistic (Geisser (1993); Gelfand et al. (1992);

and Gelfand and Dey (1994)), L measure (Ibrahim and Laud (1994); Laud and Ibrahim (1995); Gelfand and Ghosh (1998); Ibrahim et al. (2001a); and Chen et al. (2004)), Deviance Information Criterion (DIC) (Spiegelhalter et al. (2002)), and marginal likelihood (Bayes factor).

The CPO, L measure, and DIC are criterion based methods which can be attractive in the sense that they are well defined under improper priors as long as the posterior distribution is proper, and thus have an advantage over the marginal likelihood or Bayes factor approach in this sense. Because of this reason, these three criterion based methods can be directly compared to AIC (Akaike (1973)) and BIC (Schwarz (1978)). On the other hand, the marginal likelihood or the Bayes factor is well calibrated and relatively easy to interpret, but generally sensitive to vague proper priors. In the context of variable selection, it is not clear how these methods perform under the conjugate prior given in (3) for the GLM.

Under model m, for the i^{th} observation, we define the CPO statistic as follows:

$$\text{CPO}_i = f(y_i|\mathbf{x}_i, D^{(-i)}) = \int f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)})\pi(\boldsymbol{\beta}^{(m)}|D^{(-i)}, m)d\boldsymbol{\beta}^{(m)},$$

where $D^{(-i)}$ is D with the i^{th} observation deleted, and $\pi(\boldsymbol{\beta}|D^{(-i),m})$ is the posterior distribution based on the data $D^{(-i)}$. Due to the construction of the conjugate prior (3), it is more natural to define

$$\pi(\boldsymbol{\beta}^{(m)}|D^{(-i)}, m) \propto \prod_{j\neq i} \exp\left\{(y_j + a_0 y_{0j})\theta_j^{(m)} - (1+a_0)b(\theta_j^{(m)})\right\}.$$

After some messy algebra, we can show that CPO_i takes the following form:

$$\begin{aligned}\text{CPO}_i &= f(y_i|\mathbf{x}_i, D^{(-i)}) \\ &= \frac{\displaystyle\int \frac{1}{\exp[a_0\{y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)})\}]}\pi(\boldsymbol{\beta}^{(m)}|D, m)d\boldsymbol{\beta}^{(m)}}{\displaystyle\int \frac{1}{f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)})\exp[a_0\{y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)})\}]}\pi(\boldsymbol{\beta}^{(m)}|D, m)d\boldsymbol{\beta}^{(m)}}, \quad (5)\end{aligned}$$

where $f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)})$ is the density function given in (2). Also, we notice that the CPO defined in (5) is slightly different from the usual CPO (Geisser (1993) and Gelfand et al. (1992)), which is of the form

$$\left\{\int \frac{1}{f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)})}\pi(\boldsymbol{\beta}^{(m)}|D, m)d\boldsymbol{\beta}^{(m)}\right\}^{-1}.$$

However, these two forms will be identical as $a_0 \to 0$. As suggested in Ibrahim et al. (2001b), a natural summary statistic of the CPO_i's is the logarithm of the Pseudo-marginal likelihood (LPML) defined as

$$\text{LPML}_m = \sum_{i=1}^{n} \log(\text{CPO}_i).$$

We will use LPML_m as a criterion-based measure for variable selection.

The L measure criterion is another useful tool for model comparison and variable selection. The L measure is constructed from the posterior predictive distribution of the data. For the entire class of GLM's in (2), under model m, the L measure is defined as:

$$\begin{aligned} L_m(\nu) &= \sum_{i=1}^{n} \Big[E\{b''(\theta_i^{(m)})|D,m\} + \mathbf{Var}\{b'(\theta_i^{(m)})|D,m\} \Big] \\ &\quad + \nu \sum_{i=1}^{m} \Big[E\{b'(\theta_i^{(m)})|D,m\} - y_i \Big]^2, \end{aligned} \tag{6}$$

where $b'(.)$ and $b''(.)$ are the mean and variance functions of the GLM in (2), and all expectations and variances are taken with respect to the posterior distribution $\pi(\boldsymbol{\beta}^{(m)}|D,m)$ in (4). We note that for the GLM in (1), we need to modify $L_m(\nu)$ in (6) accordingly, and in this case, the L measure takes the form

$$\begin{aligned} L_m(\nu) &= \sum_{i=1}^{n} \Big[E\{a_i(\tau) b''(\theta_i^{(m)})|D,m\} + \mathbf{Var}\{b'(\theta_i^{(m)})|D,m\} \Big] \\ &\quad + \nu \sum_{i=1}^{m} \Big[E\{b'(\theta_i^{(m)})|D,m\} - y_i \Big]^2. \end{aligned} \tag{7}$$

The DIC criterion, proposed by Spiegelhalter et al. (2002), is given by

$$\text{DIC}_m = D(\bar{\boldsymbol{\beta}}^{(m)}) + 2p_D^{(m)}, \tag{8}$$

where

$$p_D^{(m)} = \overline{D(\boldsymbol{\beta}^{(m)})} - D(\bar{\boldsymbol{\beta}}^{(m)}),$$

$\bar{\boldsymbol{\beta}}^{(m)} = E[\boldsymbol{\beta}^{(m)}|D,m]$, and $\overline{D(\boldsymbol{\beta}^{(m)})} = E[D(\boldsymbol{\beta}^{(m)})|D,m]$. For the GLM in (2), under model m,

$$D(\boldsymbol{\beta}^{(m)}) = -2 \sum_{i=1}^{n} \Big\{ y_i \theta_i^{(m)} - b(\theta_i^{(m)}) \Big\}. \tag{9}$$

Similar to (6), under the GLM in (1), $D(\boldsymbol{\beta}^{(m)})$ needs to be modified accordingly.

In the spirit of marginal likelihoods, after ignoring the constants shared by all variable subset models in model space $\mathcal{M}$ for the GLM in (2), for the purpose of variable subset selection it suffices to compute the posterior normalizing constant

$$C_m(D) = \int \exp\Big\{ (\mathbf{y} + a_0 \mathbf{y}_0)' \boldsymbol{\theta}^{(m)} - (1 + a_0) J' \boldsymbol{b}(\boldsymbol{\theta}^{(m)}) \Big\} \, d\boldsymbol{\beta}^{(m)} \tag{10}$$

and the prior normalizing constant

$$C_{0m}(\mathbf{y}_0) = \int \exp\Big[a_0 \{ \mathbf{y}_0' \boldsymbol{\theta}^{(m)} - J' \boldsymbol{b}(\boldsymbol{\theta}^{(m)}) \} \Big] d\boldsymbol{\beta}^{(m)}. \tag{11}$$

Similar to the modification of (6) yielding (7), under the GLM in (1), $D(\boldsymbol{\beta}^{(m)})$ in (9), $C_m(D)$ in (10), and $C_{0m}(\mathbf{y}_0)$ in (11) need to be modified accordingly. In the context of variable selection, we select a variable subset model which yields the largest LPML_m under the CPO, the smallest $L_m(\nu)$ under the L measure, the smallest DIC_m under the DIC, and the largest $C_m(D)/C_{0m}(\mathbf{y}_0)$ or $\log[C_m(D)/C_{0m}(\mathbf{y}_0)]$ under the marginal likelihood.

3 Analytic Connections Between Variable Selection Criteria For the Normal Linear Regression Model

In this section, we consider the normal linear regression models given by

$$f(y_i|x_i^{(m)}, \boldsymbol{\beta}^{(m)}, \tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\Big\{ -\frac{\tau}{2}(y_i - \mathbf{x}_i^{(m)'}\boldsymbol{\beta}^{(m)})^2 \Big\}. \tag{12}$$

Let $X_m = (\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)}, \ldots, \mathbf{x}_n^{(m)})'$, which is the design matrix for the normal linear regression under model m. Assume X_m is of full rank k_m throughout. We focus only on the τ known case as analytical connections are more difficult to establish when τ is unknown. For the model in (12) with a known τ, the conjugate prior for $\boldsymbol{\beta}^{(m)}$ in (3) reduces to

$$[\boldsymbol{\beta}^{(m)}|\mathbf{y}_0, a_0, m] \sim N_{k_m}\Big((X_m'X_m)^{-1}X_m'\mathbf{y}_0, \frac{1}{\tau a_0}(X_m'X_m)^{-1}\Big), \tag{13}$$

and the posterior distribution for $\boldsymbol{\beta}^{(m)}$ is given by

$$[\boldsymbol{\beta}^{(m)}|D, m] \sim N_{k_m}\Big((X_m'X_m)^{-1}X_m'\frac{\mathbf{y} + a_0\mathbf{y}_0}{1 + a_0}, \frac{1}{\tau(1 + a_0)}(X_m'X_m)^{-1}\Big).$$

For (12), AIC and BIC under model m are given by

$$\text{AIC}_m = -2\log L(\hat{\boldsymbol{\beta}}^{(m)}|D) + 2k_m = -n\log\Big(\frac{\tau}{2\pi}\Big) + \tau\ \text{SSE}_m + 2k_m, \tag{14}$$

where $\hat{\boldsymbol{\beta}}^{(m)}$ is the maximum likelihood estimate of $\boldsymbol{\beta}^{(m)}$ and

$$\text{SSE}_m = \mathbf{y}'\{I - X_m(X_m'X_m)^{-1}X_m'\}\mathbf{y}$$

is the usual sum of squared errors, and

$$\text{BIC}_m = -2\log L(\hat{\boldsymbol{\beta}}^{(m)}|D) + \{\log(n)\}k_m = -n\log\Big(\frac{\tau}{2\pi}\Big) + \tau\ \text{SSE}_m + \{\log(n)\}k_m. \tag{15}$$

After some algebra, we can show that after putting back all normalizing constants, the logarithm of the marginal likelihood under model m is given by

$$\begin{aligned} &\log\{C_m(D)/C_{0m}(\mathbf{y}_0)\} \\ = \ &\frac{n}{2}\log\Big(\frac{\tau}{2\pi}\Big) - \frac{\tau}{2}\mathbf{y}'\mathbf{y} + \frac{\tau(1 + a_0)}{2}\Big(\frac{\mathbf{y} + a_0\mathbf{y}_0}{1 + a_0}\Big)'X_m(X_m'X_m)^{-1}X_m'\Big(\frac{\mathbf{y} + a_0\mathbf{y}_0}{1 + a_0}\Big) \\ &-\frac{\tau a_0}{2}\{\mathbf{y}_0'X_m(X_m'X_m)^{-1}X_m'\mathbf{y}_0\} + \Big(\frac{1}{2}\log\frac{a_0}{1 + a_0}\Big)k_m. \end{aligned} \tag{16}$$

When $\mathbf{y}_0 = \mathbf{0}$, the conjugate prior in (13) reduces to Zellner's g-prior (Zellner (1986)). For this special case, (16) becomes

$$\begin{aligned} & \log[C_m(D)/C_{0m}(\mathbf{0})] \\ = \; & \frac{n}{2}\log\Big(\frac{\tau}{2\pi}\Big) - \frac{\tau a_0}{2(1+a_0)}\mathbf{y}'\mathbf{y} - \frac{\tau}{2(1+a_0)}\mathrm{SSE}_m + \Big(\frac{1}{2}\log\frac{a_0}{1+a_0}\Big)k_m. \end{aligned} \tag{17}$$

Thus, we have

$$\begin{aligned} \mathcal{M}_m(a_0) \; &\equiv \; -2(1+a_0)\Big[\log\{C_m(D)/C_{0m}(\mathbf{0})\} + \frac{\tau a_0}{2(1+a_0)}\mathbf{y}'\mathbf{y}\Big] + a_0 n\log\Big(\frac{\tau}{2\pi}\Big) \\ &= \; -n\log\Big(\frac{\tau}{2\pi}\Big) + \tau\mathrm{SSE}_m + \Big\{(1+a_0)\log\frac{1+a_0}{a_0}\Big\}k_m. \end{aligned} \tag{18}$$

For purposes of variable selection, it suffices to compare $\mathcal{M}_m(a_0)$ and we then choose a model with the smallest $\mathcal{M}_m(a_0)$. From (18), we can see that

$$\mathcal{M}_m(a_0) = \begin{cases} \mathrm{AIC}_m & \text{if } (1+a_0)\log\frac{1+a_0}{a_0} = 2, \\ \mathrm{BIC}_m & \text{if } (1+a_0)\log\frac{1+a_0}{a_0} = \log n. \end{cases} \tag{19}$$

For (12), we use (7) to compute $L_m(\nu)$. In particular, we have $a_i(\tau) = 1/\tau$, $E[a_i(\tau)b''(\theta_i^{(m)})|D,m] = \frac{1}{\tau}$,

$$\begin{aligned} & \mathbf{Var}\big\{b'(\theta_i^{(m)})|D,m\big\} = \mathbf{Var}\big\{\mathbf{x}_i^{(m)\prime}\boldsymbol{\beta}^{(m)}|D,m\big\} \\ = \; & \mathbf{x}_i^{(m)\prime}\mathbf{Var}(\boldsymbol{\beta}^{(m)}|D,m)\mathbf{x}_i^{(m)} = \frac{1}{\tau(1+a_0)}\mathbf{x}_i^{(m)\prime}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}, \end{aligned}$$

and $E\{b'(\theta_i^{(m)})|D,m\} = E\{\mathbf{x}_i^{(m)\prime}\boldsymbol{\beta}^{(m)}|D,m\} = \mathbf{x}_i^{(m)\prime}(X_m'X_m)^{-1}X_m'\frac{\mathbf{y}+a_0\mathbf{y}_0}{1+a_0}$. Thus, we obtain

$$\begin{aligned} L_m(\nu) \; &= \; \frac{n}{\tau} + \frac{1}{\tau(1+a_0)}\sum_{i=1}^{n}\mathbf{x}_i^{(m)\prime}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)} \\ & \quad + \nu\sum_{i=1}^{n}\Big\{y_i - \mathbf{x}_i^{(m)\prime}(X_m'X_m)^{-1}X_m'\frac{\mathbf{y}+a_0\mathbf{y}_0}{1+a_0}\Big\}^2 \\ &= \; \frac{n}{\tau} + \frac{1}{\tau(1+a_0)}k_m + \nu\Big[\Big\{\mathbf{y} - X_m(X_m'X_m)^{-1}X_m'\frac{\mathbf{y}+a_0\mathbf{y}_0}{1+a_0}\Big\}' \\ & \quad \times\Big\{\mathbf{y} - X_m(X_m'X_m)^{-1}X_m'\frac{\mathbf{y}+a_0\mathbf{y}_0}{1+a_0}\Big\}\Big]. \end{aligned} \tag{20}$$

When $\mathbf{y}_0 = \mathbf{0}$, (20) reduces to

$$L_m(\nu) = \frac{n}{\tau} + \frac{1}{\tau(1+a_0)}k_m + \frac{\nu a_0^2}{(1+a_0)^2}\mathbf{y}'\mathbf{y} + \frac{\nu(1+2a_0)}{(1+a_0)^2}\mathrm{SSE}_m. \tag{21}$$

Write

$$\tilde{L}_m(\nu, a_0) = \frac{\tau(1+a_0)^2}{\nu(1+2a_0)}\Big\{L_m(\nu) - \frac{n}{\tau} - \frac{\nu a_0^2}{(1+a_0)^2}\mathbf{y}'\mathbf{y}\Big\} - n\log\Big(\frac{\tau}{2\pi}\Big). \tag{22}$$

Using (21) and (22), we obtain

$$\tilde{L}_m(\nu, a_0) = -n \log\left(\frac{\tau}{2\pi}\right) + \tau \, \mathrm{SSE}_m + \frac{1+a_0}{\nu(1+2a_0)} k_m,$$

and hence

$$\tilde{L}_m(\nu, a_0) = \begin{cases} \mathrm{AIC}_m & \text{if } \frac{1+a_0}{\nu(1+2a_0)} = 2, \\ \mathrm{BIC}_m & \text{if } \frac{1+a_0}{\nu(1+2a_0)} = \log n. \end{cases}$$

Note that in the context of variable selection, a model with the smallest $L_m(\nu)$ is the same model that has the smallest $\tilde{L}_m(\nu, a_0)$. Thus, in this sense, the L measure can be equivalent to AIC or BIC by appropriately tuning (ν, a_0). It is interesting to mention that in order to achieve $\tilde{L}_m(\nu, a_0) = \mathrm{AIC}_m$ or $\tilde{L}_m(\nu, a_0) = \mathrm{BIC}_m$, ν must be small, and hence when $\nu = 1$, the L measure always has a smaller dimensional penalty than both AIC and BIC. Unlike the marginal likelihood, a_0 plays a minimum role in controlling dimensional penalty in the L measure.

When $\mathbf{y}_0 = \mathbf{0}$, the posterior mean of $\boldsymbol{\beta}^{(m)}$ is given by $\bar{\boldsymbol{\beta}}^{(m)} = \frac{1}{1+a_0}(X_m' X_m)^{-1} X_m' \mathbf{y}$. Thus, we have $D(\boldsymbol{\beta}^{(m)}) = -n \log\left(\frac{\tau}{2\pi}\right) + \tau(\mathbf{y} - X_m \boldsymbol{\beta}^{(m)})'(\mathbf{y} - X_m \boldsymbol{\beta}^{(m)})$,

$$\begin{aligned} & \overline{D(\boldsymbol{\beta}^{(m)})} \\ = \; & E[D(\boldsymbol{\beta}^{(m)}) | D, m] = -n \log\left(\frac{\tau}{2\pi}\right) + \tau E\Big[\big\{\mathbf{y} - X_m \bar{\boldsymbol{\beta}}^{(m)} - X_m(\boldsymbol{\beta}^{(m)} - \bar{\boldsymbol{\beta}}^{(m)})\big\}' \\ & \times \big\{\mathbf{y} - X_m \bar{\boldsymbol{\beta}}^{(m)} - X_m(\boldsymbol{\beta}^{(m)} - \bar{\boldsymbol{\beta}}^{(m)})\big\} \Big| D, m\Big] \\ = \; & -n \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{1+a_0} k_m + \frac{\tau a_0^2}{(1+a_0)^2} \mathbf{y}'\mathbf{y} + \frac{\tau(1+2a_0)}{(1+a_0)^2} \mathrm{SSE}_m, \end{aligned} \tag{23}$$

and

$$\begin{aligned} D(\bar{\boldsymbol{\beta}}^{(m)}) &= -n \log\left(\frac{\tau}{2\pi}\right) + \tau(\mathbf{y} - X_m \bar{\boldsymbol{\beta}}^{(m)})'(\mathbf{y} - X_m \bar{\boldsymbol{\beta}}^{(m)}) \\ &= -n \log\left(\frac{\tau}{2\pi}\right) + \frac{\tau a_0^2}{(1+a_0)^2} \mathbf{y}'\mathbf{y} + \frac{\tau(1+2a_0)}{(1+a_0)^2} \mathrm{SSE}_m. \end{aligned} \tag{24}$$

Combining (23) and (24) gives

$$p_D^{(m)} = \overline{D(\boldsymbol{\beta}^{(m)})} - D(\bar{\boldsymbol{\beta}}^{(m)}) = \frac{1}{1+a_0} k_m. \tag{25}$$

Thus, the DIC_m for (12) is given by

$$\mathrm{DIC}_m = -n \log\left(\frac{\tau}{2\pi}\right) + \frac{\tau a_0^2}{(1+a_0)^2} \mathbf{y}'\mathbf{y} + \frac{\tau(1+2a_0)}{(1+a_0)^2} \mathrm{SSE}_m + \frac{2}{1+a_0} k_m. \tag{26}$$

Write

$$\mathrm{DIC}_m^*(a_0) = \frac{(1+a_0)^2}{1+2a_0} \Big\{ \mathrm{DIC}_m - \frac{\tau a_0^2}{(1+a_0)^2} \mathbf{y}'\mathbf{y} \Big\} + \frac{n a_0^2}{1+2a_0} \log\left(\frac{\tau}{2\pi}\right).$$

We have

$$\mathrm{DIC}^*_m(a_0) = -n\log\left(\frac{\tau}{2\pi}\right) + \tau\ \mathrm{SSE}_m + \frac{2(1+a_0)}{1+2a_0}k_m. \tag{27}$$

Therefore, when $a_0 = 0$, $\mathrm{DIC}^*_m(0) = \mathrm{DIC}_m = \mathrm{AIC}_m$, and when $a_0 > 0$, $\frac{2(1+a_0)}{1+2a_0} < 2$, which implies that $\mathrm{DIC}^*_m(a_0)$ has a smaller dimensional penalty than both AIC and BIC.

Similarly to DIC, we consider only $\mathbf{y}_0 = \mathbf{0}$. From (5), we have

$$\mathrm{LPML}_m = \sum_{i=1}^{n}\log(\mathrm{CPO}_i) = \sum_{i=1}^{n}\log\left(\mathrm{CPO}_{1i}\right) - \sum_{i=1}^{n}\log\left(\mathrm{CPO}_{2i}\right), \tag{28}$$

where $\mathrm{CPO}_{1i} = \int \exp\left\{\frac{a_0\tau}{2}\boldsymbol{\beta}^{(m)'}\mathbf{x}_i^{(m)}\mathbf{x}_i^{(m)'}\boldsymbol{\beta}^{(m)}\right\}\pi(\boldsymbol{\beta}^{(m)}|D,m)d\boldsymbol{\beta}^{(m)}$ and

$$\begin{aligned}\mathrm{CPO}_{2i} &= \left(\frac{\tau}{2\pi}\right)^{-1/2}\exp\left\{\frac{\tau}{2}y_i^2\right\}\int\exp\Big[\frac{\tau(1+a_0)}{2}\\ &\quad\times\left\{\boldsymbol{\beta}^{(m)'}\mathbf{x}_i^{(m)}\mathbf{x}_i^{(m)'}\boldsymbol{\beta}^{(m)} - \frac{2}{1+a_0}\boldsymbol{\beta}^{(m)'}\mathbf{x}_i^{(m)}y_i\right\}\Big]\pi(\boldsymbol{\beta}^{(m)}|D,m)d\boldsymbol{\beta}^{(m)}\end{aligned}$$

for $i = 1, 2, \ldots, n$. After some messy algebra, we obtain

$$\begin{aligned}\mathrm{CPO}_{1i} &= \left\{1 - \frac{a_0}{1+a_0}\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}\right\}^{-1/2}\\ &\quad\times\exp\left\{\frac{\tau a_0}{2(1+a_0)^2}\frac{\mathbf{y}'X_m(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}X_m'\mathbf{y}}{1-\frac{a_0}{1+a_0}\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}}\right\}\end{aligned}$$

and

$$\begin{aligned}\mathrm{CPO}_{2i} &= \left(\frac{\tau}{2\pi}\right)^{-1/2}\left\{1-\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}\right\}^{-1/2}\exp\left(\frac{\tau}{2}y_i^2\right)\\ &\times\exp\left[\frac{\tau}{2(1+a_0)}\left\{y_i\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}y_i - 2\mathbf{y}'X_m(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}y_i\right\}\right]\\ &\times\exp\left[\frac{\tau(X_m'\mathbf{y}-\mathbf{x}_i^{(m)}y_i)'(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}(X_m'\mathbf{y}-\mathbf{x}_i^{(m)}y_i)}{2(1+a_0)\{1-\mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}\}}\right].\end{aligned}$$

Let $\hat{\boldsymbol{\beta}}^{(m)} = (X_m'X_m)^{-1}X_m'\mathbf{y}$, $\hat{y}_i^{(m)} = \mathbf{x}_i^{(m)'}\hat{\boldsymbol{\beta}}^{(m)}$, and $h_{ii}^{(m)} = \mathbf{x}_i^{(m)'}(X_m'X_m)^{-1}\mathbf{x}_i^{(m)}$. Plugging CPO_{1i} and CPO_{2i} into (28) yields

$$\begin{aligned}\mathrm{LPML}_m &= \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\sum_{i=1}^{n}y_i^2 + \frac{1}{2}\sum_{i=1}^{n}\left\{\log(1-h_{ii}^{(m)}) - \log\left(1-\frac{a_0}{1+a_0}h_{ii}^{(m)}\right)\right\}\\ &\quad-\frac{\tau}{2(1+a_0)}\sum_{i=1}^{n}\left\{h_{ii}^{(m)}y_i^2 - 2y_i\hat{y}_i^{(m)}\right\} + \frac{\tau a_0}{2(1+a_0)^2}\sum_{i=1}^{n}\left\{\frac{\hat{y}_i^{(m)^2}}{1-\frac{a_0}{1+a_0}h_{ii}^{m}}\right\}\\ &\quad-\frac{\tau}{2(1+a_0)}\sum_{i=1}^{n}\frac{(\hat{y}_i^{(m)}-h_{ii}^{(m)}y_i)^2}{1-h_{ii}^{(m)}}.\end{aligned} \tag{29}$$

Using Taylor expansion and after some algebra, LPML_m in (29) can be rewritten as

$$\text{LPML}_m = \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau a_0^2}{2(1+a_0)^2}\mathbf{y}'\mathbf{y} - \frac{\tau(1+2a_0)}{2(1+a_0)^2}\text{SSE}_m - \frac{k_m}{2(1+a_0)} + R_m^*, \quad (30)$$

where

$$\begin{aligned} R_m^* &= -\frac{\tau}{2(1+a_0)}\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i^{(m)})^2 h_{ii}^{(m)}}{1-h_{ii}^{(m)}} + \frac{\tau a_0}{2(1+a_0)^2}\sum_{i=1}^{n}\frac{a_0}{1+a_0}\frac{h_{ii}^{(m)}\hat{y}_i^{(m)^2}}{1-\frac{a_0}{1+a_0}h_{ii}^{(m)}} \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=2}^{\infty}\left\{1-\left(\frac{a_0}{1+a_0}\right)^j\right\}\frac{(-1)^j h_{ii}^{(m)^j}}{j}. \end{aligned}$$

Write

$$\text{LPML}_m^* = -\frac{2(1+a_0)^2}{1+2a_0}\left\{\text{LPML}_m + \frac{\tau a_0^2}{2(1+a_0)^2}\mathbf{y}'\mathbf{y}\right\} + \frac{n a_0^2}{1+2a_0}\log\left(\frac{\tau}{2\pi}\right). \quad (31)$$

Using (30) and (31), we obtain

$$\text{LPML}_m^* = -n\log\left(\frac{\tau}{2\pi}\right) + \tau\ \text{SSE}_m + \frac{1+a_0}{(1+2a_0)}k_m + R_m,$$

where $R_m = -\frac{2(1+a_0)^2}{1+2a_0}R_m^*$. We choose a model with the smallest LPML_m^*. Note that the remainder term R_m is small when all $h_{ii}^{(m)}$'s are small. From (14), (15), and (27), we see that when R_m is small and does not vary much in the model space $\mathcal{M}$, LPML has a smaller dimensional penalty than DIC, AIC and BIC. In addition, when $a_0 = 0$, LPML_m in (30) is consistent with the one derived by Gelfand and Dey (1994) based on the asymptotic approximation.

Finally, we note that the quantities defined in (18), (22), (27) and (31) are linear transformations of those defined by (17), (21), (26) and (30), respectively. In these linear transformations, the relevant coefficients are independent of m. Thus, for the purposes of variable subset selection, these linearly transformed quantities act exactly like those original forms. With (18), (22), (27) and (31), we can much more clearly see the analytical connections to AIC and BIC. We also note that George and Foster (2000) provided some similar connections between model selection probabilities and various model selection criteria for this setup.

4 Computational Development: Theory and Implementation

For the purpose of variable selection, we need to compute LPML_m, $L_m(\nu)$, DIC_m, $C_m(D)$ and $C_{0m}(\mathbf{y}_0)$ for the Bayesian variable selection criteria described in the previous section for $m = 1, 2, \ldots, \mathcal{K}$. Due to the complexity and generality of the GLM in (2), the analytical evaluation of these measures does not appear possible. Thus, a Monte

Carlo (MC) based method is required for each of those measures under consideration. However, the MC methods currently available in the Bayesian computational literature require a Markov chain Monte Carlo (MCMC) sample from the posterior distribution $\pi(\boldsymbol{\beta}^{(m)}|D,m)$ in (4) under each variable subset model m. When the number of the models in $\mathcal{M}$ is large, sampling from the posterior distribution under each variable subset model can be expensive. Thus, the computation of these four measures for all submodels becomes a difficult and challenging task. Therefore, the development of an efficient Monte Carlo method for variable selection for the GLM is very essential.

After examining (5), (6), and (8), we observe that there is a common feature in computing LPML_m, $L_m(\nu)$, and DIC_m, i.e., all of these three measures require to compute

$$g_m = E\{g(\boldsymbol{\beta}^{(m)})|D,m\},$$

for various functions g, where the expectation is taken with respect to the joint posterior distribution in (4) under model m. Specifically, the functions required in these calculations include

(i) $g(\boldsymbol{\beta}^{(m)}) = \exp[-a_0\{y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)})\}]$ and $g(\boldsymbol{\beta}^{(m)}) = \Big(f(y_i|\mathbf{x}_i^{(m)},\boldsymbol{\beta}^{(m)})\exp[a_0\{y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)})\}]\Big)^{-1}$ for LPML_m;

(ii) $g(\boldsymbol{\beta}^{(m)}) = b'(\theta_i^{(m)})$, $g(\boldsymbol{\beta}^{(m)}) = \{b'(\theta_i^{(m)})\}^2$, and $g(\boldsymbol{\beta}^{(m)}) = b''(\theta_i^{(m)})$ for $L_m(\nu)$; and

(iii) $g(\boldsymbol{\beta}^{(m)}) = \boldsymbol{\beta}^{(m)}$ and $g(\boldsymbol{\beta}^{(m)}) = D(\boldsymbol{\beta}^{(m)})$ for DIC_m.

Write

$$L(\boldsymbol{\beta}^{(m)}|D,m) = \exp\Big\{(\mathbf{y}+a_0\mathbf{y}_0)'\boldsymbol{\theta}^{(m)} - (1+a_0)J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\Big\}$$

under model m and let $L(\boldsymbol{\beta}|D) = L(\boldsymbol{\beta}^{(\mathcal{K})}|D,\mathcal{K})$, $C(D) = C_{\mathcal{K}}(D)$, and $C_0(\mathbf{y}_0) = C_{0\mathcal{K}}(\mathbf{y}_0)$ under the full model. Here, we abuse the notation a little bit as $L(\boldsymbol{\beta}^{(m)}|D,m)$ is not a likelihood function in the usual sense. Then, for a given function g, mathematically, we have

$$g_m = E[g(\boldsymbol{\beta}^{(m)})|D,m] = \int g(\boldsymbol{\beta}^{(m)})\frac{L(\boldsymbol{\beta}^{(m)}|D,m)}{C_m(D)}d\boldsymbol{\beta}^{(m)},$$

where $C_m(D)$ is defined in (10). Now, we present a useful identity for g_m, which is formally stated in the following theorem.

Theorem 5. For any given function g, such that $E[|g(\boldsymbol{\beta}^{(m)})| \mid D,m] < \infty$, we have

$$g_m = \frac{C(D)}{C_m(D)}E\Big\{g(\boldsymbol{\beta}^{(m)})\frac{L(\boldsymbol{\beta}^{(m)}|D,m)w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big|D\Big\}, \tag{32}$$

where the expectation is taken with respect to the joint posterior distribution in (4) under the full model. Here, $w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})$ is a completely known conditional density,

whose support is contained in, or equal to, the support of the conditional density of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$ with respect to the joint posterior distribution in (4) under the full model.

Observing that when $g \equiv 1$, we have

$$1 = \frac{C(D)}{C_m(D)} E\Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \Big| D \Big\},$$

which leads to

$$\frac{C_m(D)}{C(D)} = E\Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \Big| D \Big\} \tag{33}$$

and

$$g_m = \frac{E\Big\{ \frac{g(\boldsymbol{\beta}^{(m)}) L(\boldsymbol{\beta}^{(m)}|D,m) w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \mid D \Big\}}{E\Big\{ \frac{L(\boldsymbol{\beta}^{(m)}|D,m) w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \mid D \Big\}}. \tag{34}$$

It is interesting to mention that the identity (33) is a by-product of this derivation and this identity can be used to compute the posterior normalizing constant under model m. The identities (33) and (34) play an important role in developing a novel Monte Carlo method for computing LPML_m, $L_m(\nu)$, DIC_m, and $C_m(D)$ simultaneously using a single MCMC sample from the joint posterior distribution under the full model. Towards this goal, we let $\{\boldsymbol{\beta}_s = ({\boldsymbol{\beta}^{(m)}}'_s, {\boldsymbol{\beta}^{(-m)}}'_s), s = 1, 2, \ldots, S\}$ denote a MCMC sample from the joint posterior distribution (4) under the full model, where S is the MCMC sample size. Then, an estimate of g_m is given by

$$\hat{g}_m = \frac{\sum_{s=1}^{S} \frac{g(\boldsymbol{\beta}_s^{(m)}) L(\boldsymbol{\beta}_s^{(m)}|D,m) w(\boldsymbol{\beta}_s^{(-m)}|\boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}}{\sum_{s=1}^{S} \frac{L(\boldsymbol{\beta}_s^{(m)}|D,m) w(\boldsymbol{\beta}_s^{(-m)}|\boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}}. \tag{35}$$

Under certain regularity conditions, such as ergodicity, we have

$$\lim_{S \to \infty} \hat{g}_m = g_m,$$

which indicates that $\hat{g}_m$ is consistent.

Letting

$$A_S = \frac{1}{S} \sum_{s=1}^{S} \frac{g(\boldsymbol{\beta}_s^{(m)}) L(\boldsymbol{\beta}_s^{(m)} \mid D, m) w(\boldsymbol{\beta}_s^{(-m)} \mid \boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)} \tag{36}$$

and

$$B_S = \frac{1}{S} \sum_{s=1}^{S} \frac{L(\boldsymbol{\beta}_s^{(m)} \mid D, m) w(\boldsymbol{\beta}_s^{(-m)} \mid \boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}, \tag{37}$$

we have

$$\lim_{S\to\infty} A_S = \frac{C_m(D)}{C(D)} g_m \equiv A, \tag{38}$$

and

$$\lim_{S\to\infty} B_S = \frac{C_m(D)}{C(D)} \equiv B. \tag{39}$$

From (38) and (39), we obtain

$$g_m = \frac{C(D)}{C_m(D)} A = \frac{A}{B}. \tag{40}$$

Using (36)-(40), we have

$$\hat{g}_m - g_m = \frac{A_S}{B_S} - g_m = \frac{A_S}{B_S} - \frac{A}{B} = \frac{A_S - \frac{A}{B}B_S}{B_S} = A\frac{\frac{A_S}{A} - \frac{B_S}{B}}{B_S} = g_m \frac{\frac{A_S}{A} - \frac{B_S}{B}}{\frac{B_S}{B}}. \tag{41}$$

In (41), $\lim_{S\to\infty} \frac{B_S}{B} = 1$ and

$$\lim_{S\to\infty} \left(\frac{A_S}{A} - \frac{B_S}{B}\right) = 0 \tag{42}$$

In addition, we have

$$\begin{aligned} \frac{A_S}{A} - \frac{B_S}{B} \quad = \quad & \frac{1}{S}\sum_{s=1}^{S} \Big[\frac{1}{A}\Big\{\frac{g(\boldsymbol{\beta}_s^{(m)})L(\boldsymbol{\beta}_s^{(m)} \mid D, m)w(\boldsymbol{\beta}_s^{(-m)} \mid \boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}\Big\} \\ & -\frac{1}{B}\Big\{\frac{L(\boldsymbol{\beta}_s^{(m)} \mid D, m)w(\boldsymbol{\beta}_s^{(-m)} \mid \boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}\Big\}\Big]. \end{aligned}$$

We are then led to the following theorem.

Theorem 6. Let $\{\boldsymbol{\beta}_s,\ s = 1, 2, \ldots, S\}$ be a random sample. Assume $A \neq 0$,

$$\begin{aligned} V_w(g_m) \quad = \quad & E\Big[\Big\{\frac{g(\boldsymbol{\beta}^{(m)})L(\boldsymbol{\beta}^{(m)} \mid D, m)w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{A\ L(\boldsymbol{\beta}|D)} \\ & -\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{B\ L(\boldsymbol{\beta}|D)}\Big\}^2 \Big| D\Big] < \infty \end{aligned} \tag{43}$$

and

$$E[\{g(\boldsymbol{\beta}^{(m)})\}^2 \mid D] < \infty, \tag{44}$$

where the expectation is taken with respect to the joint posterior distribution in (4) under the full model. Then we have

$$\lim_{S\to\infty} \Big[S \times E\Big\{\Big(\frac{\hat{g}_m - g_m}{g_m}\Big)^2\Big\}\Big] = V_w(g_m), \tag{45}$$

where $V_w(g_m)$ is defined by (43) and

$$\sqrt{S}(\hat{g}_m - g_m) \xrightarrow{\mathcal{D}} N(0, g_m^2 V_w(g_m)).$$

The proof of Theorem 6 directly follows from the proof of Theorem 3.1 of Chen and Shao (1997). Thus, the detail is omitted for brevity. From (45), we notice that $E[\frac{\hat{g}_m - g_m}{g_m}]^2$ is the relative mean-square error and Theorem 6 implies that when S is large,

$$E\Big(\frac{\hat{g}_m - g_m}{g_m}\Big)^2 \approx \frac{1}{S}\, V_w(g_m).$$

Remark 4.1: As discussed in Chen et al. (2000), the simulation standard error of $\hat{g}_m$ can be approximated by

$$se(\hat{g}_m) = \frac{|\hat{g}_m|}{\hat{A}} \sqrt{\frac{1}{S}\sum_{s=1}^{S}\Big[\{g(\boldsymbol{\beta}_s^{(m)}) - \hat{g}_m\}\frac{L(\boldsymbol{\beta}_s^{(m)} \mid D, m)w(\boldsymbol{\beta}_s^{(-m)} \mid \boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}\Big]^2},$$

where $\hat{A} = A_S$.

Remark 4.2: From (34), it is quite natural that one may think a more efficient way to obtain a MC estimate of g_m is by generating two MC samples from the posterior distribution so that one sample is used for computing $E\{\frac{g(\boldsymbol{\beta}^{(m)})L(\boldsymbol{\beta}^{(m)}|D,m)w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \mid D\}$ while the second sample is used for computing $E\{\frac{L(\boldsymbol{\beta}^{(m)}|D,m)w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \mid D\}$. In this remark, we show that the use of two MC samples in obtaining the MC estimate of g_m may not necessarily be more efficient than the use of just one MC sample. In addition, generating two MC samples requires more computing time. Specifically, suppose that $\{\boldsymbol{\beta}_{1,s},\ s = 1, 2, \ldots, S_1\}$ and $\{\boldsymbol{\beta}_{2,s},\ s = 1, 2, \ldots, S_2\}$ are two independent random samples from the joint posterior distribution (4) under the full model. Then g_m can be estimated by

$$\hat{g}_m^* = \frac{\frac{1}{S_1}\sum_{s=1}^{S_1}\frac{g(\boldsymbol{\beta}_{1,s}^{(m)})L(\boldsymbol{\beta}_{1,s}^{(m)}|D,m)w(\boldsymbol{\beta}_{1,s}^{(-m)}|\boldsymbol{\beta}_{1,s}^{(m)})}{L(\boldsymbol{\beta}_{1,s}|D)}}{\frac{1}{S_2}\sum_{s=1}^{S_2}\frac{L(\boldsymbol{\beta}_{2,s}^{(m)}|D,m)w(\boldsymbol{\beta}_{2,s}^{(-m)}|\boldsymbol{\beta}_{2,s}^{(m)})}{L(\boldsymbol{\beta}_{2,s}|D)}}. \tag{46}$$

By the δ-Method, we have

$$\begin{aligned}
E\Big(\frac{\hat{g}_m^* - g_m}{g_m}\Big)^2 &= \frac{\mathbf{Var}\Big\{\frac{1}{S_1}\sum_{s=1}^{S_1}\frac{g(\boldsymbol{\beta}_{1,s}^{(m)})L(\boldsymbol{\beta}_{1,s}^{(m)}|D,m)w(\boldsymbol{\beta}_{1,s}^{(-m)}|\boldsymbol{\beta}_{1,s}^{(m)})}{L(\boldsymbol{\beta}_{1,s}|D)}\Big\}}{A^2} \\
&\quad + \frac{\mathbf{Var}\Big\{\frac{1}{S_2}\sum_{s=1}^{S_2}\frac{L(\boldsymbol{\beta}_{2,s}^{(m)}|D,m)w(\boldsymbol{\beta}_{2,s}^{(-m)}|\boldsymbol{\beta}_{2,s}^{(m)})}{L(\boldsymbol{\beta}_{2,s}|D)}\Big\}}{B^2} + O\Big\{\frac{1}{(S_1+S_2)^2}\Big\} \\
&= \frac{1}{S_1\ A^2}\mathbf{Var}\Big\{\frac{g(\boldsymbol{\beta}^{(m)})L(\boldsymbol{\beta}^{(m)} \mid D, m)w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big\} \\
&\quad + \frac{1}{S_2\ B^2}\mathbf{Var}\Big\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big\} + O\Big\{\frac{1}{(S_1+S_2)^2}\Big\},
\end{aligned}$$

where the expectation and variance are taken with respect to the joint posterior distribution (4) under the full model.

Assuming that $S_1 = S_2 = S$, we have

$$\lim_{S\to\infty}\Big\{S\times E\Big(\frac{\hat{g}_m^*-g_m}{g_m}\Big)^2\Big\} = \frac{1}{A^2}\mathbf{Var}\Big\{\frac{g(\boldsymbol{\beta}^{(m)})L(\boldsymbol{\beta}^{(m)}\mid D,m)w(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big\} + \frac{1}{B^2}\mathbf{Var}\Big\{\frac{L(\boldsymbol{\beta}^{(m)}\mid D,m)w(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big\}. \quad (47)$$

Thus, if

$$E\Big\{\frac{g(\boldsymbol{\beta}^{(m)})L(\boldsymbol{\beta}^{(m)}\mid D,m)w(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)})}{A\ L(\boldsymbol{\beta}|D)}\times\frac{L(\boldsymbol{\beta}^{(m)}\mid D,m)w(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)})}{B\ L(\boldsymbol{\beta}|D)}\Big\}\ \geq\ 0, \quad (48)$$

we have

$$\lim_{S\to\infty}\Big\{S\times E\Big(\frac{\hat{g}_m^*-g_m}{g_m}\Big)^2\Big\}\geq\lim_{S\to\infty}\Big\{S\times E\Big(\frac{\hat{g}_m-g_m}{g_m}\Big)^2\Big\}.$$

It is easy to see that when $g(\boldsymbol{\beta}^{(m)})\geq 0$ or $g(\boldsymbol{\beta}^{(m)})\leq 0$, (48) automatically holds. Therefore, for many cases, it is unnecessary to use two MC samples instead of one MC sample in obtaining the MC estimate of g_m.

Note that the estimate $\hat{g}_m$ depends on $w(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)})$. It is reasonable to argue that the best choice of w should yield the smallest asymptotic variance of the estimate $\hat{g}_m$ among all possible w's. The following theorem precisely addresses this optimality issue.

Theorem 7. Let

$$w_{opt}=\pi(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)},D) \quad (49)$$

be the conditional posterior density of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$ under the full model, then we have

$$V_{w_{opt}}(g_m)\leq V_w(g_m) \quad (50)$$

for all w's, where $V_w(g_m)$ is defined by (43).

Remark 4.3: Note that (50) holds for any function g that satisfies the condition given in (44). Thus, for various functions g involved in LPML_m, $L_m(\nu)$ and DIC_m, the best choice of w is the same w_{opt} given in (49).

Remark 4.4: When we use $\hat{g}_m^*$ in (46), we can also show that $w_{opt}=\pi(\boldsymbol{\beta}^{(-m)}\mid\boldsymbol{\beta}^{(m)},D)$ yields the smallest asymptotic relative mean-square error of $\hat{g}_m^*$, for example, the one given by (47).

Remark 4.5: For computing CPO_i in (5) under model m, we do not need to compute $\frac{C(D)}{C_m(D)}$ in (32). In fact, it is easy to see that

$$\mathrm{CPO}_i^{(m)}=\frac{E\Big\{g_1(\boldsymbol{\beta}^{(m)})\frac{L(\boldsymbol{\beta}^{(m)}|D,m)w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big|D\Big\}}{E\Big\{g_2(\boldsymbol{\beta}^{(m)})\frac{L(\boldsymbol{\beta}^{(m)}|D,m)w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)}\Big|D\Big\}},$$

where $g_1(\boldsymbol{\beta}^{(m)}) = \exp[-a_0(y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)}))]$ and $g_2(\boldsymbol{\beta}^{(m)}) = \Big\{ f(y_i|\mathbf{x}_i^{(m)}, \boldsymbol{\beta}^{(m)}) \exp[a_0 (y_{0i}\theta_i^{(m)} - b(\theta_i^{(m)}))] \Big\}^{-1}$. Thus, given a MCMC sample $\{\boldsymbol{\beta}_s = (\boldsymbol{\beta}^{(m)'}_s, \boldsymbol{\beta}^{(-m)'}_s), s = 1, 2, \ldots, S\}$ from the joint posterior distribution (4), a MC estimate of CPO_i is given as follows:

$$\widehat{\text{CPO}}_i^{(m)} = \frac{\sum_{s=1}^{S} \frac{g_1(\boldsymbol{\beta}_s^{(m)}) L(\boldsymbol{\beta}_s^{(m)}|D,m) w(\boldsymbol{\beta}_s^{(-m)}|\boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}}{\sum_{s=1}^{S} \frac{g_2(\boldsymbol{\beta}_s^{(m)}) L(\boldsymbol{\beta}_s^{(m)}|D,m) w(\boldsymbol{\beta}_s^{(-m)}|\boldsymbol{\beta}_s^{(m)})}{L(\boldsymbol{\beta}_s|D)}}.$$

Following the proof of Theorem 7, we can easily show that the optimal choice of w for $\widehat{\text{CPO}}_i^{(m)}$ is still the same w_{opt} given in (49).

Remark 4.6: To compute $\text{LPML}_{\mathcal{K}}$, $L_{\mathcal{K}}(\nu)$ and $\text{DIC}_{\mathcal{K}}$ under the full model, we can simply take $\boldsymbol{\beta}^{(\mathcal{K})} = \boldsymbol{\beta}$ and $w(\boldsymbol{\beta}^{(-\mathcal{K})} \mid \boldsymbol{\beta}^{(\mathcal{K})}) = 1$. Then, for various functions g, given a MCMC sample $\{\boldsymbol{\beta}_s,\ s = 1, 2, \ldots, S\}$ (35) reduces to

$$\hat{g} = \frac{1}{S} \sum_{s=1}^{S} g(\boldsymbol{\beta}_s),$$

where $\{\boldsymbol{\beta}_s,\ s = 1, 2, \ldots, S\}$ is a MCMC sample from the posterior distribution (4) under the full model.

Remark 4.7: As shown in Theorem 7, the optimal choice of w is $w_{opt} = \pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)$. However, for the GLM in (2), w_{opt} is not available in closed form. Fortunately, for the GLM, a good $w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)})$, which is close to the optimal choice, can be constructed based on the asymptotic approximation to the joint posterior proposed by Chen (1985). Let $\hat{\boldsymbol{\beta}}$ denote the posterior mode of $\boldsymbol{\beta}$ under the full model, i.e.,

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}|D) = \arg\max_{\boldsymbol{\beta}} \{(\mathbf{y} + a_0\mathbf{y}_0)'\boldsymbol{\theta} - (1 + a_0)J'\mathbf{b}(\boldsymbol{\theta})\}.$$

Also let

$$\hat{\Sigma} = \Big\{ - \frac{\partial^2 \log L(\boldsymbol{\beta}|D)}{\partial\boldsymbol{\beta}\, \partial\boldsymbol{\beta}'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \Big\}^{-1}.$$

Then, the joint posterior $\pi(\boldsymbol{\beta} \mid D)$ under the full model can be approximated by

$$\hat{\pi}(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}, D) = (2\pi)^{-\frac{k+1}{2}} |\hat{\Sigma}|^{-\frac{1}{2}} \exp\Big\{ - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\hat{\Sigma}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \Big\}. \tag{51}$$

Using (51), we simply take $w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}) = \hat{\pi}(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, \hat{\boldsymbol{\beta}}, D)$, which is the conditional distribution of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$ with respect to the $(k+1)$-dimensional multivariate normal distribution in (51).

Remark 4.8: As a by-product, $C_m(D)/C(D)$ is ready to compute via the identity (33). It can also be shown that

$$\frac{C_{0m}(\mathbf{y}_0)}{C_0(\mathbf{y}_0)} = E\Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid \mathbf{y}_0, a_0, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|\mathbf{y}_0, a_0)} \Big| \mathbf{y}_0, a_0 \Big\}, \tag{52}$$

where $L(\boldsymbol{\beta}^{(m)} \mid \mathbf{y}_0, a_0, m) = \exp\left[a_0\{\mathbf{y}_0'\boldsymbol{\theta}^{(m)} - J'\boldsymbol{b}(\boldsymbol{\theta}^{(m)})\}\right]$ and the expectation is taken with respect to the prior distribution in (3) under the full model. After examining the construction of the conjugate prior and the form of the GLM in (2), we can also show that

$$B_m = \frac{C_m(D)/C(D)}{C_{0m}(\mathbf{y}_0)/C_0(\mathbf{y}_0)} = \frac{\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|D)}{\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|\mathbf{y}_0, a_0)}, \tag{53}$$

where $\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|D)$ and $\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|\mathbf{y}_0, a_0)$ are the marginal posterior density and the marginal prior density of $\boldsymbol{\beta}^{(-m)}$ evaluated at $\boldsymbol{\beta}^{(-m)} = \mathbf{0}$ under the full model. Furthermore, B_m in (53) is the Bayes factor for comparing model m to the full model. Thus, to compute B_m, we need to generate two MCMC samples, one from the posterior distribution and another one from the prior distribution of $\boldsymbol{\beta}$ under the full model, and then use (33) and (52).

Finally, we note that we derive w_{opt} under the independence assumption. We expect that this optimal choice will work well even when a dependent MCMC sample is used. Some related empirical studies have been reported and discussed in Meng and Wong (1996), Diciccio et al. (1997) and Meng and Schilling (2002). They suggested that the optimal or near-optimal procedures constructed under the independence assumption can work remarkably well in general, providing orders of magnitude improvement over other methods with similar computational effort. Alternatively, suppose we systematically take a 1-in-b subsample of size S from the Markov chain that is generated from the joint posterior distribution in (4). Then, following from Guha et al. (2004), we can show that (45) holds under some mild regularity conditions such as geometrical ergodicity and a sufficiently large b. Thus, if we take a MCMC sample in such a way, this MCMC sample can be treated as "a random sample."

5 A Simulation Study

In Section 3, we have established theoretical connections among AIC, BIC and the four Bayesian criteria in the normal linear regression setting. However, it does not appear possible that there are any analytic connections between AIC or BIC and the four Bayesian criteria for Poisson regression. For this reason, we present a simulation study for Poisson regression to empirically examine whether there exist any connections among these criteria and to examine the performance of conjugate priors in the context of variable selection. Suppose $y_i|\theta_i$ are independent Poisson observations with mean $e^{\mathbf{x}_i'\boldsymbol{\beta}}$, where $\mathbf{x}_i'$ is a $1 \times p$ vector, $i = 1, 2, \ldots, n$. The conjugate prior takes the form

$$\pi(\boldsymbol{\beta}|a_0, \mathbf{y}_0) \propto \exp\left\{\sum_{i=1}^{n} a_0\left(y_{0i}\mathbf{x}_i'\boldsymbol{\beta} - \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\right)\right\}, \tag{54}$$

where $\mathbf{y}_{0i}$ is the i^{th} component of $\mathbf{y}_0$. In the simulation, we assume that $x_{i0} = 1$, $x_{ij} \sim N(0, 1)$ independently for $j = 1, 2, 3$ and $i = 1, 2, \ldots, n$. In (54), we take $y_{0i} = 1$ for $i = 1, 2, \ldots, n$, which yields a prior mode of $\boldsymbol{\beta}$ to be $\mathbf{0}$, as shown in Chen and Ibrahim (2003). Further we use $\boldsymbol{\beta} = (-0.3, 0.3, 0, 0)'$, $\boldsymbol{\beta} = (-0.3, 0.3, 0.2, 0)'$, and

$\boldsymbol{\beta} = (-0.3, 0.3, 0.2, -0.15)'$ which correspond to the true models (x_1), (x_1, x_2), and (x_1, x_2, x_3) (full model), respectively. We also use the sample size of $n = 500$.

Under the simulation design, we independently generated $N = 500$ datasets. For each simulated dataset, we fit $2^3 = 8$ models. To compute the posterior model probabilities based on the conjugate priors, we implemented the Monte Carlo algorithm proposed in Section 4 with a Monte Carlo sample size of $S = 20,000$. For all of these 8 models, we computed BF, DIC, L measure, LPML, AIC, and BIC.

True Model	AIC	BIC
(x_1)	361	490
(x_1, x_2)	425	446
(x_1, x_2, x_3)	474	316

Table 1: Frequencies for Ranking the True Model as Best Using AIC and BIC Based on $n = 500$ and $N = 500$ Datasets

Tables 1 and 2 show results for the various methods. Our model performance evaluation criterion is a 0-1 loss function, the loss being 0 if the true model is selected and 1 otherwise. In Table 1, we see that BIC performs better than AIC in the number of times the true model is selected as best when the true model is a smaller model. For example, when (x_1) is the true model, AIC correctly identifies this model as best 361 times out of 500 and BIC correctly identifies this model as best 490 times. Table 2 compares the performance of the four other criteria under several values of a_0 from the conjugate prior as well as several values of ν for the L measure. We see from the table that, in general, for small values of a_0, which imply a noninformative prior, the Bayes factor results are quite consistent with DIC, the L measure, and LPML for small models being the true models, whereas when the full model is the true model, the Bayes factor tends to do worse for small a_0 compared to large a_0. In general, as a_0 increases, the performance of DIC, LPML, and the Bayes factor becomes worse, whereas for the L measure, it is fairly robust over several values of a_0. The L measure seems to perform best under moderate values of ν, such as $\nu = 0.5$.

6 A Real Data Example

Due to lack of analytic connections between AIC or BIC and the four Bayesian criteria for logistic regression, we consider the Chapman data from Los Angeles Heart Study of men ($n = 200$) presented in Dixon and Massey (1983) to empirically examine whether there exist any connections among these criteria.

In our analysis, we consider a coronary incident as a binary response variable (y), which takes the values 0 and 1, where a 1 denotes that an incident had occurred in the previous ten years and a 0 indicates otherwise. We consider five prognostic factors: age (Ag), systolic blood pressure in millimeters of mercury (S), diastolic blood pressure in millimeters of mercury (D), Cholesterol in milligrams per DL (Ch), and BMI =

True Model	a_0	LPML	DIC	BF	L Measure (ν) 0.1	0.3	0.5	0.7	0.9
(x_1)	0.001	395	361	492	398	396	359	318	276
	0.01	396	357	466	396	396	357	319	275
	0.1	377	332	386	408	396	352	304	268
	0.5	342	308	311	424	381	335	279	243
	1	320	299	288	424	372	321	264	222
(x_1, x_2)	0.001	425	425	436	164	347	390	380	356
	0.01	423	425	470	157	352	390	383	355
	0.1	417	417	443	195	370	399	372	353
	0.5	398	405	405	254	400	402	362	339
	1	382	394	391	269	410	390	359	329
(x_1, x_2, x_3)	0.001	475	474	291	88	371	456	475	480
	0.01	475	474	388	94	375	458	475	482
	0.1	479	475	460	125	402	466	480	488
	0.5	485	479	479	176	436	478	486	489
	1	486	481	481	214	453	483	487	490

Table 2: Frequencies for Ranking the True Model as Best Using BF, DIC, CPO and L measure for Various a_0 Based on $n = 500$ and $N = 500$ Datasets

(703.07Weight)/(Height^2).

Let x_1, x_2, x_3, x_4, and x_5 denote Ag, S, D, Ch, and BMIH. For the Chapman data, we fit a logistic regression model

$$\text{logit}\{P(y=1|\mathbf{x})\} = \log\left\{\frac{P(y=1|\mathbf{x})}{1-P(y=1|\mathbf{x})}\right\} = \mathbf{x}'\boldsymbol{\beta}. \tag{55}$$

The conjugate prior in (3) corresponding to the model (55) takes the form

$$\pi(\boldsymbol{\beta}|a_0, \mathbf{y}_0) \propto \exp\Big(\sum_{i=1}^{n} a_0\Big[y_{0i}\mathbf{x}_i'\boldsymbol{\beta} - \log\{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\}\Big]\Big), \tag{56}$$

where $y_{i0} = 0.5$, $i = 1, 2, \ldots, n$, to ensure the prior mode of $\boldsymbol{\beta}$ to be $\mathbf{0}$. We wish to compare the following 32 models: Intercept only, (x_1), $\ldots$, (x_5), (x_1, x_2), $\ldots$, $(x_1, x_2, x_3, x_4, x_5)$. We note that the notation $(x_1, x_2, x_3, x_4, x_5)$, for example, implies that $\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1\text{Ag}_i + \beta_2\text{S}_i + \beta_3\text{D}_i + \beta_4\text{Ch}_i + \beta_5\text{BMI}_i$ in (55). Thus, "Intercept only" is the model with zero predictors while $(x_1, x_2, x_3, x_4, x_5)$ is the *full model* with the largest model dimension. We also note that an intercept is included in every model. Further we denote that M_1 = (Int), M_2 = (Int, Ag), M_3 = (Int, S), M_4 = (Int, D), M_5 = (Int, Ch), M_6 =(Int, BMI), M_7 = (Int, Ag, S), M_8 =(Int, Ag, D), M_9 =(Int, Ag, Ch), M_{10} =(Int, Ag, BMI), M_{11} =(Int, S, D), M_{12} =(Int, S, Ch), M_{13} =(Int, S, BMI), M_{14} =(Int, D, Ch), M_{15} =(Int, D, BMI), M_{16} =(Int, Ch, BMI), M_{17} =(Int, Ag, S, D), M_{18} =(Int, Ag, S, Ch), M_{19} =(Int, Ag, S, BMI), M_{20} =(Int, Ag, D, Ch), M_{21} =(Int,

Ag, D, BMI), M_{22} =(Int, Ag, Ch, BMI), M_{23} =(Int, S, D, Ch), M_{24} =(Int, S, D, BMI), M_{25} =(Int, S, Ch, BMI), M_{26} = (Int, D, Ch, BMI), M_{27} =(Int, Ag, S, D, Ch), M_{28} =(Int, Ag, S, D, BMI), M_{29} = (Int, Ag, S, Ch, BMI), M_{30} =(Int, Ag, D, Ch, BMI), M_{31} = (Int, S, D, Ch, BMI), and M_{32} = (Int, Ag, S, D, Ch, BMI).

AIC		BIC	
M_k	Values	M_k	Values
M_{22}	142.75	M_2	153.34
M_{10}	143.73	M_{10}	153.63
M_{29}	144.69	M_9	155.83
M_{30}	144.75	M_{22}	155.94
M_{19}	145.57	M_{16}	155.99

Table 3: The Top Model Based on AIC and BIC for Chapman Data

	$a_0 = 0.001$		$a_0 = 0.01$		$a_0 = 0.1$		$a_0 = 0.5$		$a_0 = 1.0$	
Criterion	M_k	Values	M_k	Values	M_k	Values	M_k	Values	M_k	Values
PMP	M_2	0.57	M_2	0.25	M_{22}	0.14	M_{22}	0.07	M_{22}	0.06
DIC	M_{22}	142.83	M_{22}	142.67	M_{22}	144.74	M_{22}	165.65	M_{22}	186.77
LPML	M_2	-73.38	M_2	-73.30	M_{32}	-73.79	M_{32}	-83.10	M_{30}	-93.29
L($\nu = 0.1$)	M_{22}	21.47	M_{22}	21.98	M_{22}	26.96	M_{22}	38.92	M_{30}	45.21
L($\nu = 0.25$)	M_{22}	24.79	M_{22}	25.29	M_{22}	30.23	M_{22}	42.56	M_{30}	49.39
L($\nu = 0.5$)	M_{32}	30.20	M_{32}	30.73	M_{32}	35.66	M_{29}	48.59	M_{30}	56.36
L($\nu = 0.75$)	M_{32}	35.24	M_{32}	35.76	M_{32}	40.77	M_{32}	54.52	M_{30}	63.33
L($\nu = 0.9$)	M_{32}	38.26	M_{32}	38.78	M_{32}	43.83	M_{32}	58.06	M_{30}	67.51

Table 4: The Best Model Based on Posterior Model Probability (PMP), DIC, LPML, and L Measure for Chapman Data

To compute the posterior model probability (PMP), DIC, LPML, and L measure under various conjugate priors, we implemented the Monte Carlo algorithm proposed in Section 4 with a Monte Carlo sample size of $S = 20,000$. We see from Table 3 that M_{22} is selected as the best model by AIC and the fourth model by BIC, whereas M_{10} is selected as the second best model by both criteria. Table 4 shows the results of the L measure, posterior model probability (PMP), LPML, and DIC for several values of a_0, as well as several values of ν for the L measure. Table 3 reveals a similar story as the simulation study. Model M_{22} is selected as either the top model or second best model for most values of a_0 for DIC and PMP, as well as for the L measure under small values of ν. Under larger values of ν the L measure as well a LPML appear to favor model M_{32}. Finally, for small values of a_0, LPML and PMP appear to favor a smaller model, namely M_2. Thus, from these analyses, models $\{M_2, M_{22}, M_{32}\}$ appear to be the most promising based on all of these model selection criteria. Table 5 shows the top five models selected for each of the four variable selection criteria (PMP, DIC, L measure, LPML). Again we see a remarkable consistency between the four criteria, in which the ordering of the top models is similar for the four criteria for small, moderate, and large values of a_0, and for a wide range of ν values for the L measure.

	$a_0 = 0.001$		$a_0 = 0.01$		$a_0 = 0.1$		$a_0 = 0.5$		$a_0 = 1.0$	
Criterion	M_k	Values	M_k	Values	M_k	Values	M_k	Values	M_k	Values
PMP	M_2	0.57	M_2	0.25	M_{22}	0.14	M_{22}	0.07	M_{22}	0.06
	M_1	0.11	M_{10}	0.23	M_{10}	0.14	M_{10}	0.07	M_{10}	0.05
	M_5	0.07	M_9	0.08	M_9	0.06	M_{29}	0.05	M_{19}	0.04
	M_{10}	0.07	M_{22}	0.08	M_2	0.06	M_{19}	0.05	M_{29}	0.04
	M_6	0.06	M_{16}	0.07	M_{16}	0.06	M_{30}	0.05	M_{21}	0.04
DIC	M_{22}	142.83	M_{22}	142.67	M_{22}	144.74	M_{22}	165.65	M_{22}	186.77
	M_{10}	143.79	M_{10}	143.70	M_{10}	145.70	M_{10}	166.02	M_{10}	186.88
	M_{30}	144.85	M_{29}	144.74	M_{29}	146.43	M_{29}	166.42	M_{30}	187.10
	M_{29}	144.96	M_{30}	144.78	M_{30}	146.48	M_{19}	166.87	M_{21}	187.38
	M_{21}	145.63	M_{21}	145.59	M_{21}	147.29	M_{30}	166.90	M_{20}	187.75
LPML	M_2	-73.38	M_2	-73.30	M_{32}	-73.79	M_{32}	-83.10	M_{30}	-93.29
	M_5	-73.56	M_5	-73.50	M_2	-74.11	M_{29}	-83.32	M_{32}	-93.35
	M_4	-73.58	M_6	-73.55	M_7	-74.43	M_{10}	-83.34	M_{21}	-93.41
	M_6	-73.73	M_4	-73.64	M_8	-74.48	M_{19}	-83.55	M_{10}	-93.41
	M_{32}	-73.91	M_3	-73.65	M_9	-74.68	M_{21}	-83.56	M_{20}	-93.55
L	M_{22}	21.47	M_{22}	21.98	M_{22}	26.96	M_{22}	38.92	M_{30}	45.21
$\nu = 0.1$	M_{30}	22.04	M_{25}	22.63	M_{25}	27.33	M_{25}	38.99	M_{22}	45.23
	M_{26}	22.13	M_{30}	22.64	M_{30}	27.41	M_{29}	39.02	M_{26}	45.26
	M_{32}	22.15	M_{29}	22.65	M_{26}	27.45	M_{19}	39.16	M_{20}	45.28
	M_{10}	22.21	M_{10}	22.67	M_{29}	27.47	M_{10}	39.18	M_{21}	45.31
L	M_{22}	24.79	M_{22}	25.29	M_{22}	30.23	M_{22}	42.56	M_{30}	49.39
$\nu = 0.25$	M_{30}	25.17	M_{32}	25.69	M_{32}	30.56	M_{29}	42.61	M_{22}	49.45
	M_{32}	25.17	M_{30}	25.77	M_{30}	30.56	M_{25}	42.67	M_{20}	49.49
	M_{10}	25.43	M_{29}	25.78	M_{29}	30.62	M_{32}	42.73	M_{26}	49.51
	M_{20}	25.47	M_{10}	25.89	M_{25}	30.65	M_{19}	42.79	M_{21}	49.52
L	M_{32}	30.20	M_{32}	30.73	M_{32}	35.66	M_{29}	48.59	M_{30}	56.36
$\nu = 0.5$	M_{22}	30.31	M_{22}	30.80	M_{22}	35.70	M_{32}	48.62	M_{32}	56.47
	M_{30}	30.38	M_{30}	30.98	M_{30}	35.81	M_{22}	48.64	M_{22}	56.50
	M_{20}	30.77	M_{29}	31.00	M_{29}	35.88	M_{30}	48.78	M_{20}	56.50
	M_{10}	30.80	M_{10}	31.26	M_{10}	36.10	M_{25}	48.79	M_{21}	56.55
L	M_{32}	35.24	M_{32}	35.76	M_{32}	40.77	M_{32}	54.52	M_{30}	63.33
$\nu = 0.75$	M_{30}	35.59	M_{30}	36.20	M_{30}	41.06	M_{29}	54.56	M_{32}	63.41
	M_{22}	35.84	M_{29}	36.22	M_{29}	41.14	M_{22}	54.71	M_{20}	63.52
	M_{29}	36.04	M_{22}	36.31	M_{22}	41.16	M_{30}	54.78	M_{22}	63.54
	M_{20}	36.08	M_{10}	36.63	M_{10}	41.48	M_{19}	54.88	M_{21}	63.57
L	M_{32}	38.26	M_{32}	38.78	M_{32}	43.83	M_{32}	58.06	M_{30}	67.51
$\nu = 0.9$	M_{30}	38.72	M_{30}	39.33	M_{30}	44.21	M_{29}	58.15	M_{32}	67.58
	M_{22}	39.16	M_{29}	39.35	M_{29}	44.29	M_{22}	58.36	M_{20}	67.73
	M_{29}	39.17	M_{22}	39.61	M_{22}	44.44	M_{30}	58.37	M_{22}	67.77
	M_{20}	39.26	M_{27}	39.83	M_{10}	44.70	M_{19}	58.51	M_{21}	67.79

Table 5: The Top Five Models Based on PMP, DIC, LPML, and L Measure for Chapman Data

Table 6 shows the posterior means (Estimates), the posterior standard errors (SEs), and 95% HPD intervals for the β_j's under model M_{22} (Ag, Ch, BMI) and model M_{32} (Ag, S, D, Ch, BMI) when $a_0 = 0.01$. Table 6 also shows the corresponding maximum likelihood estimates (MLEs), the standard errors, and p-values. We see from Table 6 that the posterior estimates are very close to the MLEs, which is intuitively appealing, as a fairly noninformative ($a_0 = 0.01$) is used. We also see from this table that under these two "best" models, age and BMI are only two prognostic factors for the coronary incident, which are significant at the 5% significance level.

		Maximum Likelihood Estimates			Posterior Estimates		
							95% HPD
Model	Variable	Estimate	SE	p-value	Estimate	SE	Interval
M_{22}	Intercept	-2.252	0.275	$< .0001$	-2.265	0.272	(-2.805, -1.748)
	Ag	0.556	0.245	0.0230	0.554	0.242	(0.087, 1.032)
	Ch	0.405	0.233	0.0816	0.402	0.234	(-0.064, 0.854)
	BMI	0.470	0.204	0.0211	0.465	0.207	(0.069, 0.882)
M_{32}	Intercept	-2.248	0.274	$< .0001$	-2.292	0.273	(-2.828, -1.766)
	Ag	0.527	0.270	0.0507	0.531	0.270	(0.012, 1.067)
	S	0.106	0.336	0.7523	0.097	0.344	(-0.583, 0.757)
	D	-0.077	0.383	0.8417	-0.069	0.383	(-0.806, 0.687)
	Ch	0.404	0.235	0.0857	0.402	0.240	(-0.074, 0.866)
	BMI	0.474	0.226	0.0361	0.473	0.230	(0.028, 0.930)

Table 6: Estimates of the β under Model (Ag, Ch, BMI) and Model (Ag, S, D, Ch, BMI) for the Chapman Data when $a_0 = 0.01$

To examine performance of the proposed Monte Carlo method in Section 4, we first computed various model selection criteria under a sub-model using a MCMC sample from the full model. We then computed the same quantities using a MCMC sample directly from the posterior distribution under the same sub-model. For illustrative purposes, we considered a single variable sub-model M_2 = (Int, Ag) using the conjugate prior (56) with $a_0 = 0.01$. Using a MCMC sample size of $S = 20,000$, the Monte Carlo estimates (simulation standard errors) of DIC, LPML, L($\nu = 0.1$), L($\nu = 0.5$), and L($\nu = 0.9$) under model M_2 are 146.68 (0.08), -73.30 (0.04), 23.91 (0.05), 32.44 (0.06), and 40.96 (0.06), respectively, using the proposed Monte Carlo method via (35). With the same MC sample size, these quantities are 146.67 (0.02), -73.29 (0.01), 23.90 (0.02), 32.42 (0.02), and 40.95 (0.02), respectively, using the MC sample directly from the posterior distribution under model M_2. All simulation standard errors were computed using the overlapping batch statistics (OBS) method of Schmeiser et al. (1990). As expected, the simulation standard errors using the MC sample from the full model are slightly larger than those computed using the MC sample directly from model M_2. However, these two sets of the MC estimates are very close. This empirically demonstrates that the proposed MC method works quite well. Finally, we compared the computational times between the proposed Monte Carlo method and the exhaustive alternative. With 2,000 "burn-in" iterations and $S = 20,000$, the computational times of the proposed Monte Carlo method for 32 DIC's, LPML's, and L(ν)'s are 71.28, 100.11, and 76.36

seconds, respectively, on a Dell WS Xeon dual 2.4GHZ CPU Linux workstation. Using the same number of "burn-in" iterations, the same MC sample size, and the same computer, the computational times of the exhaustive alternative Monte Carlo method for 32 DIC's, LPML's, and L(ν)'s are 324.05, 357.97, and 322.13 seconds, respectively. Thus, it becomes apparent that the proposed Monte Carlo method leads to a substantial computational saving over the exhaustive alternative.

7 Concluding Remarks

We have examined and established theoretical and computational relationships between six commonly used methods for variable subset selection. These connections were facilitated from the class of conjugate priors of Chen and Ibrahim (2003). We saw that under this class of priors the four Bayesian criteria were quite similar in terms of model choice especially under small values of a_0, and the results were fairly robust under a wide choice of a_0 values. Further work remains to be done. In particular, it is of interest to obtain analytic connections between these criteria for specific GLM's, such as the logistic and Poisson regression models, as well as theoretically examine the small sample and large sample behavior of these methods. In Section 4, the theory and algorithm are developed for computing the four Bayesian criteria which are defined for the GLM in (2). With some straightforward modification, these theory and algorithm can be applied for computing the four Bayesian criteria that are defined for the general GLM in (1).

We note some philosophical issues about model selection that are worth noting. In this paper, we have evaluated the performance of all criteria based on how well they can pick up the true sampling model. However, there are other ways of defining the "Bayesian model." Many advocate that a Bayesian model is specified by the sampling density and the prior, not only by the sampling density. When one only evaluates the success of a criterion based on how well it picks up the sampling model, then a comparison between AIC (or BIC) and DIC is not meaningful when DIC is computed using an informative prior. Since AIC is equivalent to DIC based on a noninformative prior, a comparison of AIC (or BIC) to DIC is simply not meaningful when using informative priors. In general, one should avoid such comparisons, and only comparable criteria should be compared. For example, it is meaningful to compare AIC, BIC, DIC, LPML, the L-measure, and the Bayes factor based on noninformative priors. It is meaningful to compare DIC, the L-measure, LPML, and the Bayes factor based on informative priors. Finally, we note that most criteria for model assessment, especially the information criteria, are based on a well-defined utility function. If a utility function is chosen, a comparison to a criterion based on a different utility function is not justified. For example, the Bayes factor and BIC are prior predictive criteria aiming at the explanation of the data given the prior, whereas DIC (AIC as a special case) and LPML are posterior predictive criteria aiming at the explanation of replicate (unseen) data given the posterior. Thus, one must use caution in comparing these criteria in terms in picking up the true sampling model.

Appendix: Proofs of Theorems

Proof of Theorem 5:

Since $\int w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}) d\boldsymbol{\beta}^{(-m)} = 1$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(m)'}, \boldsymbol{\beta}^{(-m)'})'$, we have

$$\begin{aligned} g_m &= \int g(\boldsymbol{\beta}^{(m)}) \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C_m(D)} d\boldsymbol{\beta}^{(m)} \\ &= \int\int g(\boldsymbol{\beta}^{(m)}) \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C_m(D)} w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}) d\boldsymbol{\beta}^{(-m)} d\boldsymbol{\beta}^{(m)} \\ &= \frac{C(D)}{C_m(D)} \int g(\boldsymbol{\beta}^{(m)}) \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \frac{L(\boldsymbol{\beta}|D)}{C(D)} d\boldsymbol{\beta} \\ &= \frac{C(D)}{C_m(D)} E\Big\{ \frac{g(\boldsymbol{\beta}^{(m)}) L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \Big| D \Big\}, \end{aligned}$$

which completes the proof.

Proof of Theorem 7:

From (43), we have

$$V_w(g_m) = E\Big[\Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{L(\boldsymbol{\beta}|D)} \Big\}^2 \Big| D \Big]. \tag{A.1}$$

Plugging w_{opt} into (A.1), we have

$$\begin{aligned} & V_{w_{opt}}(g_m) \\ = & \int \Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C(D)} \Big\}^2 \frac{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)^2}{\pi(\boldsymbol{\beta}|D)} d\boldsymbol{\beta} \\ = & \int \Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C(D)} \Big\}^2 \frac{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D) \frac{\pi(\boldsymbol{\beta}^{(-m)}, \boldsymbol{\beta}^{(m)}|D)}{\pi(\boldsymbol{\beta}^{(m)}|D)}}{\pi(\boldsymbol{\beta}|D)} d\boldsymbol{\beta} \\ = & \int \Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C(D)} \Big\}^2 \frac{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)}{\pi(\boldsymbol{\beta}^{(m)}|D)} d\boldsymbol{\beta} \\ = & \int \Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{C(D)} \Big\}^2 \frac{d\boldsymbol{\beta}^{(m)}}{\pi(\boldsymbol{\beta}^{(m)}|D)} \int \pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D) d\boldsymbol{\beta}^{(-m)} \\ = & \int \Big\{ \frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B} \Big\}^2 \Big\{ \frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{\pi(\boldsymbol{\beta}^{(m)}|D) C(D)} \Big\}^2 \pi(\boldsymbol{\beta}^{(m)} \mid D) d\boldsymbol{\beta}^{(m)}, \end{aligned} \tag{A.2}$$

where $\pi(\boldsymbol{\beta}^{(m)} \mid D)$ denotes the marginal posterior distribution of $\boldsymbol{\beta}^{(m)}$ under the full

model. Thus, it suffices to show

$$\int \left\{\frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B}\right\}^2 \left\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{\pi(\boldsymbol{\beta}^{(m)}|D)}\right\}^2 \pi(\boldsymbol{\beta}^{(m)} \mid D) d\boldsymbol{\beta}^{(m)}$$

$$\leq \int \left\{\frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B}\right\}^2 \left\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\pi(\boldsymbol{\beta}|D)}\right\}^2 \pi(\boldsymbol{\beta} \mid D) d\boldsymbol{\beta}. \quad \text{(A.3)}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} 1 &= \left\{\int w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}) d\boldsymbol{\beta}^{(-m)}\right\}^2 \\ &= \left\{\int \frac{w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\sqrt{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)}} \sqrt{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)} d\boldsymbol{\beta}^{(-m)}\right\}^2 \\ &\leq \int \frac{w^2(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)} d\boldsymbol{\beta}^{(-m)} \int \pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D) d\boldsymbol{\beta}^{(-m)} \\ &= \int \frac{w^2(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)} d\boldsymbol{\beta}^{(-m)}. \end{aligned} \quad \text{(A.4)}$$

Using (A.4), the left-hand side of (A.3) becomes

$$\int \left\{\frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B}\right\}^2 \left\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m)}{\pi(\boldsymbol{\beta}^{(m)}|D)}\right\}^2 \pi(\boldsymbol{\beta}^{(m)} \mid D) d\boldsymbol{\beta}^{(m)}$$

$$\leq \int \left\{\frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B}\right\}^2 \left\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\pi(\boldsymbol{\beta}^{(m)}|D)}\right)^2 \frac{\pi(\boldsymbol{\beta}^{(m)} \mid D)}{\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, D)} d\boldsymbol{\beta}$$

$$= \int \left\{\frac{g(\boldsymbol{\beta}^{(m)})}{A} - \frac{1}{B}\right\}^2 \left\{\frac{L(\boldsymbol{\beta}^{(m)} \mid D, m) w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)})}{\pi(\boldsymbol{\beta}|D)}\right\}^2 \pi(\boldsymbol{\beta} \mid D) d\boldsymbol{\beta},$$

which exactly matches the right-hand side of (A.3).

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In Petrov, B. and Csaki, F. (eds.), *International Symposium on Information Theory*, 267–281. Budapest: Akademia Kiado. 589

Brown, P. J., Vanucci, M., and Fearn, T. (1998). "Multivariate Bayesian Variable Selection and Prediction." *Journal of the Royal Statistical Society, Series B*, 60: 627–641. 585

— (2002). "Bayes Model Averaging with Selection of Regresors." *Journal of the Royal Statistical Society, Series B*, 64: 519–536. 585

Chen, C. F. (1985). "On Asymptotic Normality of Limiting Density Functions with Bayesian Implications." *Journal of the Royal Statistical Society, Series B*, 47: 540–546. 601

Chen, M.-H., Dey, D. K., and Ibrahim, J. G. (2004). "Bayesian Criterion Based Model Assessment for Categorical Data." *Biometrika*, 91: 45–63. 589

Chen, M.-H. and Ibrahim, J. G. (2003). "Conjugate Priors for Generalized Linear Models." *Statistica Sinica*, 13: 461–476. 585, 586, 587, 588, 608

Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., and Weiss, R. E. (2003). "Prior Elicitation for Model Selection and Estimation in Generalized Linear Mixed Models." *Journal of Statistical Planning and Inference*, 111: 57–76. 586

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999). "Prior Elicitation, Variable Selection, and Bayesian Computation for Logistic Regression Models." *Journal of the Royal Statistical Society, Series B*, 61: 223–242. 585

Chen, M.-H. and Shao, Q.-M. (1997). "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants." *The Annals of Statistics*, 25: 1563–1594. 599

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag. 599

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). "Bayesian CART Model Search (with Discussion)." *Journal of the American Statistical Association*, 93: 935–960. 585

— (2001). "The practical Implementation of Bayesian Model Selection (with Discussion)." In Lahiri, P. (ed.), *Model Selection*, 63–134. Beachwood, Ohio: Institute of Mathematical Statistics. 585

— (2003). "Bayesian Treed Generalized Linear Models (with Discussion)." In Bernardo, J. M., Bayarri, M., Berger, J. O., Dawid, A. P., Heckerman, D., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 7, 85–103. Oxford: Oxford University Press. 585

Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies (with Discussion)." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 6, 157–185. Oxford: Oxford University Press. 585

Clyde, M. and George, E. I. (2004). "Model Uncertainty." *Statistical Science*, 19: 81–94. 586

Dellaportas, P. and Forster, J. J. (1999). "Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models." *Biometrika*, 86: 615–633. 585

Diciccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations." *Journal of the American Statistical Association*, 92: 903–915. 602

Dixon, W. J. and Massey, F. J. (1983). *Introduction to Statistical Analysis*. New York: McGraw-Hill, the fourth edition edition. 603

Geisser, S. (1993). *Predictive Inference: An Introduction*. London: Chapman & Hall. 588, 589

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian Model Choice: Asymptotics and Exact Calculations." *Journal of the Royal Statistical Society, Series B*, 56: 501–514. 589, 595

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model Determinating Using Predictive Distributions with Implementation via Sampling-based Methods (with Discussion)." In Bernado, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 4, 147–167. Oxford: Oxford University Press. 588, 589

Gelfand, A. E. and Ghosh, S. K. (1998). "Model Choice: A Minimum Posterior Predictive Loss Approach." *Biometrika*, 85: 1–13. 589

George, E. I. (2000). "The Variable Selection Problem." *Journal of the American Statistical Association*, 95: 1304–1308. 585

George, E. I. and Foster, D. P. (2000). "Calibration and Empirical Bayes Variable Selection." *Biometrika*, 87: 731–747. 585, 595

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 1304–1308. 585

— (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–374. 585

George, E. I., McCulloch, R. E., and Tsay, R. (1996). "Two Approaches to Bayesian Model Selection with Applications." In Berry, D., Chaloner, K., and Geweke, J. (eds.), *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, 339–348. New York: Wiley. 585

Guha, S., MacEachern, S. N., and Peruggia, M. (2004). "Benchmark Estimation for Markov Chain Monte Carlo Samples." *Journal of Computational and Graphical Statistics*, 13: 683–701. 602

Ibrahim, J. G., Chen, M.-H., and McEachern, S. N. (1999). "Bayesian Variable Selection for Proportional Hazards Models." *Canadian Journal of Statistics*, 27: 701–717. 585

Ibrahim, J. G., Chen, M.-H., and Ryan, L. M. (2000). "Bayesian Variable Selection for Time Series Count Data." *Statistica Sinica*, 10: 971–987. 586

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001a). "Criterion Based Methods for Bayesian Model Assessment." *Statistica Sinica*, 11: 419–443. 589

— (2001b). *Bayesian Survival Analysis*. New York: Springer-Verlag. 589

Ibrahim, J. G. and Laud, P. W. (1994). "A Predictive Approach to the Analysis of Designed Experiments." *Journal of the American Statistical Association*, 89: 309–319. 589

Lahiri, P. (2001). *Model Selection*. Beachwood, Ohio: Institute of Mathematical Statistics. 586

Laud, P. W. and Ibrahim, J. G. (1995). "Predictive Model Selection." *Journal of the Royal Statistical Society, Series B*, 57: 247–262. 585, 589

Meng, X.-L. and Schilling, S. (2002). "Warp Bridge Sampling." *Journal of Computational and Graphical Statistics*, 11: 552–586. 602

Meng, X.-L. and Wong, W. H. (1996). "Simulating Ratios of Normalizing Constants via A Simple Identity: A Theoretical Exploration." *Statistica Sinica*, 6: 831–860. 602

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). "Bayesian Variable and Link Determination for Generalised Linear Models." *Journal of Statistical Planning and Inference*, 111: 165–180. 586

Raftery, A. E. (1996). "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models." *Biometrika*, 83: 251–266. 585

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, 92: 179–191. 585

Schmeiser, B. W., Avramidis, A. N., and Hashem, S. (1990). "Overlapping Batch Statistics." In Balci, O., Sadowski, R. P., and Nance, R. E. (eds.), *Proceedings of the 1990 Winter Simulation Conference*, 395–398. San Diego, California: Society for Computer Simulation International. 607

Schwarz, G. (1978). "Estimating the Dimension of A Model." *The Annals of Statistics*, 6: 461–464. 589

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75: 317–343. 585

Spiegelhalter, D. J., Best, N. G., Carlin., B. P., and van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit (with Discussion)." *Journal of the Royal Statistical Society, Series B*, 62: 583–639. 589, 590

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions." In Goel, P. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques*, 233–243. Amsterdam: Elsevier Science Publishers B.V. 592

Acknowledgments

The authors wish to thank the Editor-in-Chief, the Editor, the Associate Editor, and the two referees for their helpful comments and suggestions, which have improved the paper. This research was partially supported by NIH grants #GM 70335 and #CA 74015.