Approximate Predictive Densities and Fully Bayes Variable Selection in Generalized Linear Models

Xinlei WANG and Min CHEN*

May 2004

Abstract

Exact calculations of model posterior probabilities or related quantities are often infeasible due to the analytical intractability of predictive densities. Here new approximations to obtain predictive densities are proposed and contrasted with those based on the Laplace method. The attractive features of the proposed methods include ease of implementation, computational efficiency, and accuracy over a wide range of hyperparameters. In the context of variable selection in GLMs, they are employed to facilitate the implementation of a Fully Bayes approach under three classes of informative priors on regression coefficients, namely, normal, conjugate and power priors. Metropolis-Hastings MCMC algorithms are used for stochastically searching high posterior models. An illustrative application demonstrates the effectiveness of our selection procedure.

Keywords: Laplace Approximation; Hierarchical Models; Normal Prior; Power Prior; Conjugate Prior; Asymptotic Normality; Markov Chain Monte Carlo; Stochastic Search; Logistic Regression.

**Xinlei Wang is Assistant Professor, Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, Texas 75275-0332, swang@mail.smu.edu. Min Chen is a Ph.D. student, Department of Management Science and Information Systems, University of Texas, Austin, Texas 78712, minchen@mail.utexas.edu.*

1 Introduction

Bayesian applications in a number of statistical problems need to be able to evaluate marginal probability distributions of data, often called predictive distributions, or their ratios known as Bayes factors for a set of competing models, which are often analytically intractable. Calculations of such quantities have been addressed by several authors, including sampling or Monte Carlo methods (e.g., Gelfand & Smith 1990, Verdinelli & Wasserman 1995, Han & Carlin 2001) and analytic approximations based on the Laplace method (e.g., Tierney & Kadane 1986, Tierney et al. 1989, Gelfand & Dey 1994, Raftery 1996). The first part of this paper presents new methods to approximate these predictive distributions, where we also discuss and compare their theoretical properties with those of the Laplace method.

The other major part of the paper is devoted to the problem of Bayesian variable selection in Generalized Linear Models (GLMs). There has been considerable recent work in this field, for example, Carlin et al. (1992), George et al. (1994), Raftery (1996), Bedrick et al. (1997), Kuo & Mallick (1998), Clyde & Parmigiani (1998), Clyde (1999), Chen et al. (1999), Ibrahim et al. (2000), Meyer & Laud (2002), Ntzoufras et al. (2003), Wang & George (2004), etc. Here, we consider a Fully Bayes approach under a hierarchical mixture setup for model uncertainty, where the proposed approximation methods are used to facilitate the computation of posterior probabilities as well as stochastic search of high posterior models. In particular, we review and discuss informative prior classes for regression coefficients and obtain a unified framework to take into account the uncertainty of unknown hyperparameters.

The remainder of the paper is organized as follows. In Section 2, we introduce new analytic methods for approximating predictive distributions. Section 3 describes every necessary "brick" of a hierarchical Bayesian formulation for GLMs. We introduce settings and notations, address prior and hyperprior specification, derive analytical approximations for the marginal densities of the data, and propose MCMC sampling schemes for posterior computation and stochastic search. Section 4 presents two examples, one providing a simulation evaluation and comparison of various approximations, and the other illustrating an application of our FB approach of variable selection. Section 5 concludes with a discussion.

2 Methods for Approximating Predictive Distributions

2.1 The Methods with Normal Priors

We begin with an integral for a general predictive distribution based on normal priors,

$$I = \int L_n(\boldsymbol{\beta})\pi(\boldsymbol{\beta})\mathbf{d}\boldsymbol{\beta} \tag{2.1}$$

where $\boldsymbol{\beta}$ is a parameter vector with a domain Ω being $\mathbf{R}^m$, $L_n(\boldsymbol{\beta})$ is a likelihood function based on n observations, and $\pi(\boldsymbol{\beta})$ is a prior *normal* density on $\boldsymbol{\beta}$ with a mean vector $\boldsymbol{\beta}_0$ and a covariance matrix $\lambda\boldsymbol{\Sigma}_0$, $\lambda > 0$. The hyperparameter λ quantifies the strength of subjective prior belief in $\boldsymbol{\beta}_0$. When $\lambda = 0$, π reduces to a point mass at $\boldsymbol{\beta}_0$. When $\lambda \to +\infty$, π reduces to a flat prior for $\boldsymbol{\beta}$ but still integrates to 1.

Theorem 2.1. *Suppose* $\{l_n = \log L_n : n = 1, 2, \ldots\}$ *is a Laplace-regular sequence of log-likelihood functions, having strict local maxima* $\{\hat{\boldsymbol{\beta}}_n : n = 1, 2, \ldots\}$ *and positive definite matrices* $\{\boldsymbol{\Sigma}_n = [-l_n''(\hat{\boldsymbol{\beta}}_n)]^{-1} : n = 1, 2, \ldots\}$. *Let*

$$\tilde{I} \;=\; L_n(\hat{\boldsymbol{\beta}}_n) \cdot \left|\lambda\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_n^{-1} + \mathbf{I}\right|^{-\frac{1}{2}} \exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T(\lambda\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_n)^{-1}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)}{2}\right\} \tag{2.2}$$

and I *is defined in (2.1) under the normal prior. Then* $\tilde{I} = I(1 + O(n^{-1}))$.

Laplace regularity has been discussed in depth in Kass et al. (1990), where they pointed out it is straightforward to verify Laplace regularity for a wide range of situations, such as exponential and curved exponential families, certain mixture models, etc. By assuming this for $\{l_n\}$, we know from Lemma 2 in Kass et al. (1990), $\exists\ c > 0$ and $\delta > 0$, such that for all sufficient large n,

$$\int_{\Omega - B_\delta(\hat{\boldsymbol{\beta}}_n)} \exp\{l_n(\boldsymbol{\beta}) - l_n(\hat{\boldsymbol{\beta}}_n)\}\pi(\boldsymbol{\beta})\mathbf{d}\boldsymbol{\beta} < \exp\{-nc\} \tag{2.3}$$

where $B_\delta(\hat{\boldsymbol{\beta}}_n)$ denotes the open ball of radius δ centered at $\hat{\boldsymbol{\beta}}_n$. This reduces our consideration of I to the integral of $L_n(\boldsymbol{\beta})\pi(\boldsymbol{\beta})$ over a neighborhood $B_\delta(\hat{\boldsymbol{\beta}}_n)$ because (2.3) assures that I depends only on the behaviour of L_n near its maximum when n is large. Then $\tilde{I}$ can be obtained by first approximating $l_n(\boldsymbol{\beta})$ with a second-order Taylor series expanded at $\hat{\boldsymbol{\beta}}_n$, say $\hat{l}_n(\boldsymbol{\beta})$, over the region $B_\delta(\hat{\boldsymbol{\beta}}_n)$, then inserting it in (2.1) and integrating out $\boldsymbol{\beta}$ from the

approximate integrand $\exp\{\hat{l}(\boldsymbol{\beta})\}\pi(\boldsymbol{\beta})$. As a result, $\tilde{I}$ is derived by taking advantage of the normality assumed for $\pi(\boldsymbol{\beta})$. A full proof of the theorem is given in Appendix A.

Under the conditions of Theorem 2.1, the well-known Laplace's method is also legitimate for the integral I. The resulting approximation to (2.1) is not unique and depends on how one defines the Laplace-regular sequence $\{l_n\}$. Defining $l_n = \log L_n$ yields a standard-form Laplace approximation, $\tilde{I}_L = I(1 + O(n^{-1}))$, namely,

$$\tilde{I}_L = L_n(\hat{\boldsymbol{\beta}}_n) \cdot \lambda^{-\frac{m}{2}} \left|\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_n^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)}{2\lambda}\right\}. \tag{2.4}$$

Contrasting $\tilde{I}$ with $\tilde{I}_L$ yields a couple of interesting results. Firstly, if $L_n(\boldsymbol{\beta})$ is proportional to a normal pdf for $\boldsymbol{\beta}$, then the approximation $\tilde{I}$ in (2.2) is exact, namely, $\tilde{I} = I$. This can be seen by noting that in the normal case, the second-order approximation to the log-likelihood is exactly itself. However, it is easy to verify $\tilde{I}_L \neq I$ in this case. Secondly, for large λ, $\tilde{I} \approx \tilde{I}_L$. This follows directly from $\lim_{\lambda \to +\infty} \tilde{I}/\tilde{I}_L = 1$. But for small λ, $\tilde{I}$ may differ substantially from $\tilde{I}_L$. For $\tilde{I}_L$, we have

$$\lim_{\lambda \to 0} \tilde{I}_L = \begin{cases} 0 & \text{if } \boldsymbol{\beta}_0 \neq \hat{\boldsymbol{\beta}}_n \\ +\infty & \text{if } \boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_n \end{cases}; \tag{2.5}$$

for $\tilde{I}$, we have

$$\lim_{\lambda \to 0} \tilde{I} = L_n(\hat{\boldsymbol{\beta}}_n) \cdot \exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_n^{-1} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)}{2}\right\}. \tag{2.6}$$

Based on these limits, it appears that when λ is small, $\tilde{I}$ is better than $\tilde{I}_L$ for approximating I. For example, the value of I when $\lambda = 0$ is $L_n(\boldsymbol{\beta}_0)$ since $\boldsymbol{\beta}$ is fixed at $\boldsymbol{\beta}_0$ in this case. Comparing this with (2.6), we see that

$$\lim_{n \to +\infty} \lim_{\lambda \to 0} \tilde{I} = I(\lambda = 0) \tag{2.7}$$

whenever $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ as $n \to +\infty$, which occurs with probability 1 under mild regularity conditions for many common-used models including GLMs if $\hat{\boldsymbol{\beta}}_n$ is the MLE . This limiting equality does not hold for $\tilde{I}_L$.

We can also consider Laplace's method in a fully exponential form (Tierney & Kadane, 1986; Tierney et al., 1989) by defining $l_n = \log L_n + \log \pi$. Suppose $\{l_n = \log L_n + \log \pi : n = 1, 2, \ldots\}$ have strict local maxima $\{\tilde{\boldsymbol{\beta}}_n : n = 1, 2, \ldots\}$ and positive definite matrices

$\{\Xi_n = [-l_n''(\tilde{\boldsymbol{\beta}}_n)]^{-1} : n = 1, 2, \ldots\}$, then we have the approximation

$$\tilde{I}_{LF} = L_n(\tilde{\boldsymbol{\beta}}_n) \cdot \left|\lambda \boldsymbol{\Sigma}_0 \Xi_n^{-1} + \mathbf{I}\right|^{-\frac{1}{2}} \exp\left\{-\frac{(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)}{2\lambda}\right\} \tag{2.8}$$

and $\tilde{I}_{LF} = I(1 + O(n^{-1}))$. Like $\tilde{I}$, if $L_n(\boldsymbol{\beta})$ is proportional to a normal pdf for $\boldsymbol{\beta}$, $\tilde{I}_{LF}$ is exact. But unfortunately, the limiting equality (2.7) does not hold for $\tilde{I}_{LF}$ because for each n,

$$\lim_{\lambda \to 0} \tilde{I}_{LF} = \begin{cases} 0 & \text{if } \boldsymbol{\beta}_0 \neq \hat{\boldsymbol{\beta}}_n \\ L_n(\boldsymbol{\beta}_0) & \text{if } \boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_n \end{cases}. \tag{2.9}$$

It appears that $\tilde{I}$ is better than either $\tilde{I}_L$ or $\tilde{I}_{LF}$ as λ approaches 0, which has been confirmed in our limited experiments, as will be discussed in Section 4.1. We should note the limiting behaviors discussed above may be useful under situations where we are interested in integrals (2.1) for a wide range of λ, such as a sensitivity analysis on λ or a Fully Bayes (FB) approach that entails choosing a prior on λ and integrating it out of I over the support $(0, +\infty)$.

2.2 The Methods with Non-normal Priors

We proceed to discuss methods for approximating the integral in (2.1) with a non-normal prior. Here, the domain Ω is an open subset of $\mathbf{R}^m$. A straightforward extension of Theorem 2.1 leads to the following result.

Corollary 2.1. *Suppose $\{l_n = \log L_n : n = 1, 2, \ldots\}$ satisfies the same condition in Theorem 2.1; $\pi(\boldsymbol{\beta})$ is a four-time continuous differentiable nonnormal prior on $\boldsymbol{\beta}$ with the mode $\boldsymbol{\beta}_0$; and $\boldsymbol{\Sigma}_0 = [-\lambda(\log \pi)'']^{-1}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is positive definite, $\lambda > 0$. Let*

$$\tilde{I}^{NN} = \tilde{I} + \tilde{I}_L^{NN} - \tilde{I}_L \tag{2.10}$$

where $\tilde{I}$ is given in (2.2), $\tilde{I}_L$ is given in (2.4) and $\tilde{I}_L^{NN} = (2\pi)^{\frac{m}{2}} |\boldsymbol{\Sigma}_n|^{\frac{1}{2}} L_n(\hat{\boldsymbol{\beta}}_n) \cdot \pi(\hat{\boldsymbol{\beta}}_n)$; then $\tilde{I}^{NN} = I(1 + O(n^{-1}))$.

Proof. Let $\pi^N(\boldsymbol{\beta})$ denote the pdf of $N(\boldsymbol{\beta}_0, \lambda \boldsymbol{\Sigma}_0)$. Note that we can write

$$I = \int L_n(\boldsymbol{\beta}) \pi^N(\boldsymbol{\beta}) d\boldsymbol{\beta} + \int L_n(\boldsymbol{\beta})[\pi(\boldsymbol{\beta}) - \pi^N(\boldsymbol{\beta})] \mathbf{d}\boldsymbol{\beta}. \tag{2.11}$$

Applying Theorem 2.1 to the first integral at the right-hand side of (2.11) and applying the standard-form Laplace approximation to the second integral yields (2.10) immediately. □

One can treat $\tilde{I}^{NN}$ as an improved Laplace approximation to I under the condition $\pi(\boldsymbol{\beta})$ can be approximated by $\pi^N(\boldsymbol{\beta})$, which is often satisfied when the prior is constructed from the likelihood or posterior of $\boldsymbol{\beta}$ based on historical or imaginary data with valid asymptotic normality. This is because in (2.10), $\tilde{I}_L^{NN}$ is indeed the standard-form Laplace approximation to I; the remaining term $\tilde{I} - \tilde{I}_L$ provides a correction factor, using the difference between the Laplace's method and Theorem 2.1 based on the normality of $\pi^N(\boldsymbol{\beta})$. This correction is useful when λ is small or when the prior sample size n_0 (i.e., the size of historical data) is close to or even larger than n. Under this case, the standard-form Laplace approximation may not work well as L_n does not dominate π; $\tilde{I}^{NN}$ can achieve better performance because $\pi - \pi^N$ is dominated by L_n when π is well approximated by π^N. On the other hand, if this approximate normality does not hold for π, no significant improvement can be achieved through the correction so $\tilde{I}^{NN}$ would be similar to the standard-form Laplace approximation.

Like $\tilde{I}$ for normal priors, an advantage of $\tilde{I}^{NN}$ for nonnormal priors over a Laplace approximation to I is its nice limiting property. For example, when λ equals 0, $\pi(\boldsymbol{\beta})$ reduces to a point mass at its mode $\boldsymbol{\beta}_0$ so that $I(\lambda = 0) = L_n(\boldsymbol{\beta}_0)$; in this case, the Laplace approximation to the second integral in (2.11) is zero because $\pi(\boldsymbol{\beta}) = \pi^N(\boldsymbol{\beta})$; this leads to $\tilde{I}^{NN} = \tilde{I}$ and the equality $\lim_{n\to+\infty} \lim_{\lambda\to 0} \tilde{I}^{NN} = I(\lambda = 0)$ follows directly from (2.7). We also note that if $\hat{\boldsymbol{\beta}}_n$ is within a small neighborhood of $\boldsymbol{\beta}_0$, then $\tilde{I}_L^{NN} \approx \tilde{I}_L$. This occurs sometimes in practice, for example, when the information contained in the current data agrees well with that in the historical data where the prior of $\boldsymbol{\beta}$ is from. In this case, I can be approximated by $\tilde{I}$ only.

3 Application to Bayes Variable Selection in GLMs

3.1 Basic Formulation

Suppose $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^T$ are independent observations and follow an exponential family distribution

$$p\left(\mathbf{Y}|\boldsymbol{\theta}, \phi\right) = \exp\left\{\frac{\boldsymbol{\theta}\mathbf{W}\mathbf{Y} - \mathbf{b}(\boldsymbol{\theta})\mathbf{W}\mathbf{J}}{\phi} + \mathbf{c}(\mathbf{Y}, \phi)\mathbf{J}\right\} \tag{3.1}$$

indexed by the dispersion parameter ϕ and the unknown canonical parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots \theta_n)$ that may depend on p observed covariates $\mathbf{X}_1$,..., $\mathbf{X}_p$. The functions $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), b(\theta_2), \cdots, b(\theta_n))$ and $\mathbf{c}(\mathbf{Y}, \phi) = (c(y_1, \phi,), c(y_2, \phi), \cdots, c(y_n, \phi))$, assumed to be known, jointly determine the

type of the distribution. The $n \times n$ matrix $\mathbf{W}$ is diagonal with its ith diagonal element being w_i, a known weight for the ith observation. $\mathbf{J}$ is the $n \times 1$ vector of all 1's.

To fix notation, let $\gamma = 1, 2, \ldots, 2^p$ index all subsets of the covariates and let q_γ be the size of the γth subset. The problem here is to select the "best" model of the form $g(E(\mathbf{Y})) = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma$, where g is a known link function that by definition is monotonic and differentiable, $\mathbf{X}_\gamma$ is a $n \times (q_\gamma + 1)$ covariate matrix with 1's in the first column and the γth subset of $\mathbf{X}_j$'s in the remaining columns, and $\boldsymbol{\beta}_\gamma$ is a $(q_\gamma + 1) \times 1$ vector of regression coefficients. Based on (3.1), the γth model for $\mathbf{Y}$ in (3.5) may be expressed as

$$p\left(\mathbf{Y}|\gamma, \boldsymbol{\beta}_\gamma, \phi\right) = \exp\left\{\frac{\boldsymbol{\theta}(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)\mathbf{W}\mathbf{Y} - \mathbf{b}(\boldsymbol{\theta}(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma))\mathbf{W}\mathbf{J}}{\phi} + \mathbf{c}(\mathbf{Y}, \phi)\mathbf{J}\right\} \tag{3.2}$$

Here, we denote $\boldsymbol{\theta}(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)$ explicitly for $\boldsymbol{\theta}$ because $\boldsymbol{\theta} = b'^{-1} \circ g^{-1}(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma)$ holds under model γ, where $\circ$ denotes function composition.

A full Bayesian solution to variable selection uncertainty for the GLM setup proceeds as follows. Consider prior formulations of the form

$$\pi(\gamma, \boldsymbol{\beta}_\gamma, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2|\phi) = \pi(\gamma|\boldsymbol{\psi}_1)\pi(\boldsymbol{\beta}_\gamma|\gamma, \boldsymbol{\psi}_2, \phi)\pi(\boldsymbol{\psi}_1)\pi(\boldsymbol{\psi}_2) \tag{3.3}$$

where $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are unknown hyperparameters indexing the priors on γ and $\boldsymbol{\beta}_\gamma$, respectively; and $\pi(\boldsymbol{\psi}_1)$ and $\pi(\boldsymbol{\psi}_2)$ are the the hyperprior distributions on $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, respectively. Such prior distributions lead to posterior distributions over γ of the form:

$$\begin{aligned} \pi(\gamma|\mathbf{Y}, \phi) &\propto p\left(\mathbf{Y}, \gamma|\phi\right) \\ &= \pi(\gamma) \int p\left(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2, \phi\right)\pi(\boldsymbol{\psi}_2)\, \mathbf{d}\boldsymbol{\psi}_2 \end{aligned} \tag{3.4}$$

where $\pi(\gamma) = \int \pi(\gamma|\boldsymbol{\psi}_1)\pi(\boldsymbol{\psi}_1)\, \mathbf{d}\boldsymbol{\psi}_1$ is the unconditional prior density of γ and

$$p\left(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2, \phi\right) = \int p\left(\mathbf{Y}|\gamma, \boldsymbol{\beta}_\gamma, \phi\right)\pi(\boldsymbol{\beta}_\gamma|\gamma, \boldsymbol{\psi}_2, \phi)\, \mathbf{d}\boldsymbol{\beta}_\gamma \tag{3.5}$$

is the predictive distribution for model γ given $\boldsymbol{\psi}_2$ and ϕ. Note in (3.3) - (3.5), we treat ϕ as a constant instead of a parameter. This indeed occurs in Poisson, Binomial and Negative Binomial GLMs, where ϕ is equal to 1. For other members within the exponential family such as Normal, Gamma and Inverse Gaussian GLMs, ϕ is unknown so we might further consider a prior $\pi(\phi)$ on ϕ to formally account for its uncertainty and obtain $\pi(\gamma|\mathbf{Y})$ from

$\pi(\gamma|\mathbf{Y}) \propto \int p(\mathbf{Y}, \gamma|\phi)\, \pi(\phi)\, d\phi$; or we might proceed as before but with ϕ replaced by an estimate (Raftery 1996 and Wang & George 2004).

In what follows, we demonstrate that the asymptotic methods proposed in the previous section can facilitate the implementation of the above fully Bayes framework for various prior choices of $\boldsymbol{\beta}_\gamma$. We believe these methods are preferable here to Laplace methods (Raftery 1996 and Gelfand & Dey 1994) on grounds of the behaviors at small λ and ease of implementation.

3.2 Prior Distributions on Regression Coefficients

Choices of prior forms for model-specific parameters $\boldsymbol{\beta}_\gamma$ have been fruitfully explored by existing literature. Here we restrict attention to informative priors only. Our hope is to achieve some unification of the informative priors on $\boldsymbol{\beta}_\gamma$ where we base our discussions on.

There are in general three classes of informative priors on $\boldsymbol{\beta}_\gamma$ that we are aware of, normal, conjugate and power priors. Before we proceed to discuss these priors, for notation simplicity, we denote the whole data of the current analysis as $\mathcal{D} = (n, \mathbf{Y}, \mathbf{X}, \mathbf{W})$ and the subset data for model γ as $\mathcal{D}_\gamma = (n, \mathbf{Y}, \mathbf{X}_\gamma, \mathbf{W})$, denote the likelihood (3.2) as $L(\boldsymbol{\beta}_\gamma, \phi|\mathbf{Y}, \mathbf{X}_\gamma, \mathbf{W})$ or simply $L(\boldsymbol{\beta}_\gamma, \phi|\mathcal{D}_\gamma)$, and denote the Hessian matrix of $L(\boldsymbol{\beta}_\gamma, 1|\mathcal{D}_\gamma)$ as $\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}_\gamma)$.

1. The normal prior denoted π^N has been widely used for GLMs (Dellaportas & Smith 1993 and Meyer & Laud 2002). Here we consider the general form

$$\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda^N \sim \mathbf{N}(\mathbf{m}_\gamma, \lambda^N \phi \mathbf{U}_\gamma) \quad \text{for } \lambda^N > 0 \tag{3.6}$$

   where λ^N is a hyperparameter reflecting the importance given to the prior mean $\mathbf{m}_\gamma$ and $\mathbf{U}_\gamma$ is a multiple of the prior covariance matrix of $\boldsymbol{\beta}_\gamma$.

2. For the conjugate prior denoted π^C, we follow Meyer & Laud (2002) and look it as the likelihood for parameters $(\boldsymbol{\beta}_\gamma, \lambda^C\phi)$ with $\mathcal{D}_\gamma^C = (n, \boldsymbol{\mu}_0, \mathbf{X}_\gamma, \mathbf{W})$ as the data:

$$\pi^C(\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda^C) \propto L(\boldsymbol{\beta}_\gamma, \lambda^C\phi|\mathcal{D}_\gamma^C) \quad \text{for } \lambda^C > 0 \tag{3.7}$$

   where $\boldsymbol{\mu}_0$ is the prior guess for $\mathbf{Y}$, and λ^C is a hyperparameter that reflects the quality of the information conveyed by $\boldsymbol{\mu}_0$. The theoretical properties of this prior are discussed in Chen & Ibrahim (2003).

3. The power prior denoted π^P, proposed in Chen et al. (2000a) with its optimality properties described in Ibrahim et al. (2003), is based on historical data sets containing the same response and covariates as the current study. Without loss of generality, we restrict our attention to a single historical data set $\mathcal{D}^P = (n_0, \mathbf{Y}_0, \mathbf{X}_0, \mathbf{W}_0)$. Then π^P can be expressed as

$$\pi^P(\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda^P) \quad \propto \quad L^{1/\lambda^P}(\boldsymbol{\beta}_\gamma, \phi|\mathcal{D}^P_\gamma) \propto L(\boldsymbol{\beta}_\gamma, \lambda^P\phi|\mathcal{D}^P_\gamma) \quad \text{for } \lambda^P > 0 \tag{3.8}$$

where λ^P is a hyperparameter weighing the likelihood of the historical data relative to that of the current study. This meaningful prior provides a natural route to quantify historical data and incorporate them into the current study.

Although seemingly different in forms, the three classes of priors on $\boldsymbol{\beta}_\gamma$ are closely related via their large sample properties. The conjugate prior π^C is asymptotically normal as the sample size $n \to \infty$:

$$\pi^C(\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda^C) \to \mathbf{N}(\hat{\boldsymbol{\beta}}^C_{0\gamma}, \lambda^C\phi\hat{\mathbf{V}}^C_{0\gamma}) \tag{3.9}$$

where $\hat{\boldsymbol{\beta}}^C_{0\gamma}$ is the MLE of $\boldsymbol{\beta}_\gamma|\gamma$ using $\boldsymbol{\mu}_0$ rather than $\mathbf{Y}$ as the response vector; and $\hat{\mathbf{V}}^C_{0\gamma}$ is minus the inverse of $\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}^C_\gamma)$, the Hessian matrix of the likelihood $L(\boldsymbol{\beta}_\gamma, 1|\boldsymbol{\mu}_0, \mathbf{X}_\gamma, \mathbf{W})$, evaluated at $\hat{\boldsymbol{\beta}}^C_{0\gamma}$. This result can be obtained from Theorem 2.1 in Chen (1985) under some mild normality conditions. Similarly, the power prior π^P is asymptotically normal as the historical sample size $n_0 \to \infty$:

$$\pi^P(\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda^P) \to \mathbf{N}(\hat{\boldsymbol{\beta}}^P_{0\gamma}, \lambda^P\phi\hat{\mathbf{V}}^P_{0\gamma}) \tag{3.10}$$

where $\hat{\boldsymbol{\beta}}^P_{0\gamma}$ is the MLE of $\boldsymbol{\beta}_\gamma$ based on the historical data $\mathcal{D}^P_\gamma$, and $\hat{\mathbf{V}}^P_{0\gamma}$ is minus the inverse of $\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}^P_\gamma)$, evaluated at $\hat{\boldsymbol{\beta}}^P_{0\gamma}$.

Comparing the three normal distributions given by (3.6) , (3.9) and (3.10) yields insightful findings. Firstly, the hyperparameters λ^N, λ^C and λ^P essentially play the same role in the three classes of priors, weighing the impact of the prior information relative to the current data. This provides a formal justification that a unified hyperprior can be chosen for any λ no matter which prior class it is from. From now on, we ignore the superscripts of λ's and simply use λ to denote any of λ^N, λ^C and λ^P. Secondly, the conjugate and power priors are asymptotically special cases of the normal prior. (3.9) and (3.10) indeed sheds light on

choosing $\mathbf{m}_\gamma$ and $\mathbf{U}_\gamma$ in the normal prior (3.6) when real and meaningful prior information is available. A reasonable choice for $\mathbf{m}_\gamma$ is $\hat{\boldsymbol{\beta}}^P_{0\gamma}$ and for $\mathbf{U}_\gamma$ is $\hat{\mathbf{V}}^P_{0\gamma}$ when historical data exist; when the prior guess $\boldsymbol{\mu}_0$ for $\mathbf{Y}$ can be obtained from a prior prediction based on theoretical models, expert opinions etc., a reasonable choice for $\mathbf{m}_\gamma$ is $\hat{\boldsymbol{\beta}}^C_{0\gamma}$ and for $\mathbf{U}_\gamma$ is $\hat{\mathbf{V}}^C_{0\gamma}$, or their surrogates requiring less computing efforts (Laud & Ibrahim, 1996 and Meyer & Laud, 2002).

Under the situation where strong prior information does not exist, a natural default choice for $\mathbf{m}_\gamma$ is $(\bar{\beta}_0, 0, \ldots, 0)^T$ (Chipman et al., 2003), where $\bar{\beta}_0$ is the MLE of β_0 under the null model, namely $g(\bar{Y})$ for any link function g or specifically $b'^{-1}(\bar{Y})$ for a canonical link . A simple choice for $\mathbf{U}_\gamma$ in this case is the identity or diagonal matrix that assumes the apriori independence thus completely ignores the correlation structure among $\boldsymbol{\beta}_{\gamma i}$'s. A more realistic choice of $\mathbf{U}_\gamma$ is minus the inverse of $\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}_\gamma)$, evaluated at $\hat{\boldsymbol{\beta}}_\gamma$ (i.e., the MLE of $\boldsymbol{\beta}_\gamma$ based on the current data $\mathcal{D}_\gamma$). This choice, using the correlation structure estimated from the data, leads to great analytical tractability of model posteriors under the fully Bayes framework described in Section 3.1, as shown in Wang & George (2004).

The prior specification for $\boldsymbol{\beta}_\gamma$ is completed by choosing a hyperprior distribution for λ. For posterior probabilities, Bayes factors or related quantities to be well defined in the context of variable selection, a proper joint prior for $\boldsymbol{\beta}_\gamma$ and λ, i.e., $\pi(\boldsymbol{\beta}_\gamma, \lambda|\gamma, \phi)$, is desirable. Under our prior structure (3.3), it is easy to verify such a joint prior is proper when both $\pi(\boldsymbol{\beta}_\gamma|\gamma, \phi, \lambda)$ and $\pi(\lambda)$ are proper. The propriety of the conjugate prior (3.7) and the power prior (3.8) on $\boldsymbol{\beta}_\gamma$ has been established for GLMs under some very general conditions in Meyer & Laud (2002) and Chen et al. (2000a). In addition, a natural proper prior for λ is an inverse gamma $IG(a, b)$, which leads to a proper joint prior under the conjugate or power priors of the form

$$\pi(\boldsymbol{\beta}_\gamma, \lambda|\gamma, \phi) \propto \frac{L(\boldsymbol{\beta}_\gamma, \lambda\phi|\mathcal{D}_{0\gamma})}{\int L(\boldsymbol{\beta}_\gamma, \lambda\phi|\mathcal{D}_{0\gamma})\mathbf{d}\boldsymbol{\beta}_\gamma}\lambda^{-(a+1)}\exp(-\frac{b}{\lambda}) \tag{3.11}$$

where $\mathcal{D}_0$ is $\mathcal{D}^C$ for the conjugate prior and is $\mathcal{D}^P$ for the power prior. Note here the normalizing part $\int L(\boldsymbol{\beta}_\gamma, \lambda\phi|\mathcal{D}_{0\gamma})\mathbf{d}\boldsymbol{\beta}_\gamma$ of π^C or π^P is a function of λ and ϕ and varies from model to model, so cannot be ignored from the the joint prior (3.11) when model uncertainty is under consideration. This makes (3.11) quite different from those considered in Chen & Ibrahim (2003) and Chen et al. (2000a) that ignored the normalizing part directly.

In practice, the range of all plausible values of λ can be used to guide the choice of the hyperprior parameters values for (a, b). Unless the prior information on $\boldsymbol{\beta}_\gamma$ is extremely

important, as will be rarely the case, λ is often set to be ≥ 1, assigning less or equal importance to the prior compared to the data. Also, Meyer & Laud (2002) suggested a guide value for λ is n/n_π, where n_π is a sample size judged to be equivalent to the information in the prior. Based on this, it is reasonable to expect that λ will be smaller than n. Using these values as guides, a and b would be chosen so that the prior on λ assigns large probability to the interval $(1, n)$ and meanwhile allows for a reasonable spread within the interval. In any case, it may be appropriate to explore the consequences of several different hyperprior choices.

3.3 Elicitation of Prior Model Probabilities

The prior on γ often takes the form of prior evidence for the inclusion of a variable rather than an individual model (Raftery & Richardson, 1993, Clyde, 1999, etc.), namely

$$\pi(\gamma|\omega_1, \omega_2, \ldots, \omega_p) = \prod_{i=1}^{p} \omega_i^{\gamma_i}(1-\omega_i)^{1-\gamma_i} \tag{3.12}$$

where ω_i is the prior probability that X_i is present in a model and γ_i is the indicator of whether X_i is present in the γth model. Under conjugate hyperpriors $\omega_i \sim beta(\alpha_i, \beta_i)$, the unconditional prior distribution of γ is then

$$\pi(\gamma) \propto \prod_{i=1}^{p} \left\{\Gamma(\gamma_i + \alpha_i)\Gamma(1 - \gamma_i + \beta_i)/\Gamma(1 + \alpha_i + \beta_i)\right\}. \tag{3.13}$$

In a practical situation, there may be subjective prior information about the importance of a covariate available from previous studies or expert systems. This can easily be taken into account by adjusting the choice of (α_i, β_i) for each ω_i in the prior (3.13), even if such information is verbal and vague. Another sensible prior on γ, proposed in Chen et al. (1999) and Ibrahim et al. (2000), is the posterior probability of model γ based on historical data or its generalization. This allows for objective inference and efficient use of historical information. However, it requires extensive computing and also is limited by the existence and quality of historical data. In this paper, we consider the prior (3.13) only.

3.4 Approximate Predictive Distributions

We proceed to derive approximate representations, using the asymptotic methods proposed in Section 2, for the marginal likelihoods or the predictive distributions in (3.5) with $\boldsymbol{\psi}_2$ replaced

by λ, based on the three classes of meaningful priors on $\boldsymbol{\beta}_\gamma$. As will be shown next, the analytic asymptotics we present have advantages of conceptual simplicity and ease of implementation for users with standard computing resources.

A direct application of Theorem 2.1 yields an approximation for $p\left(\mathbf{Y}|\gamma,\lambda,\phi\right)$ based on the normal prior (3.6), namely

$$\tilde{p}(\mathbf{Y}|\gamma,\lambda,\phi) \;\; = L(\hat{\boldsymbol{\beta}}_\gamma,\phi|\mathcal{D}_\gamma)\left|\lambda\mathbf{U}_\gamma\hat{\mathbf{V}}_\gamma^{-1}+\mathbf{I}\right|^{-\frac{1}{2}}\exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)^T(\lambda\mathbf{U}_\gamma+\hat{\mathbf{V}}_\gamma)^{-1}(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)}{2\phi}\right\} \quad (3.14)$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the MLE of $\boldsymbol{\beta}_\gamma$ using the current data $\mathcal{D}_\gamma=(n,\mathbf{Y},\mathbf{X}_\gamma,\mathbf{W})$, and $\hat{\mathbf{V}}_\gamma=-\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}}_\gamma|\mathcal{D}_\gamma)$. When $\mathbf{Y}$ is normally distributed so that the canonical link GLM is the familiar normal linear model, this approximation is exact, i.e. $\tilde{p}\left(\mathbf{Y}|\gamma,\lambda,\phi\right)=p\left(\mathbf{Y}|\gamma,\lambda,\phi\right)$. Another attractive feature of (3.14) is that, if $\mathbf{U}_\gamma$ is set to $\hat{\mathbf{V}}_\gamma$, it can yield great analytical tractability when further integrating λ out . This can be seen from

$$\tilde{p}(\mathbf{Y}|\gamma,\lambda,\phi) \;\; = L(\hat{\boldsymbol{\beta}}_\gamma,\phi|\mathcal{D}_\gamma)(\lambda+1)^{-\frac{q_\gamma+1}{2}}\exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)^T\hat{\mathbf{V}}_\gamma^{-1}(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)}{2\phi(\lambda+1)}\right\}$$

that is conjugate to a prior of λ in form of $1/(\lambda+1)\sim Truncated\ Gamma(a,b)$. As shown in Wang & George (2004), the resulting model posterior has computational simplicity and adaptive performance in selection.

Corollary 2.1 can be employed to approximate $p(\mathbf{Y}|\gamma,\lambda,\phi)$ in (3.5) with the conjugate or power prior of $\boldsymbol{\beta}_\gamma$. As we only know $\pi(\boldsymbol{\beta}_\gamma|\gamma,\phi,\lambda)$ up to a normalizing constant, we first calculate the constant from the Laplace method, namely

$$\int L(\boldsymbol{\beta}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})\mathbf{d}\boldsymbol{\beta}_\gamma \approx (2\pi)^{\frac{q_\gamma+1}{2}}\left|\lambda\phi\hat{\mathbf{V}}_{0\gamma}\right|^{\frac{1}{2}}\cdot L(\hat{\boldsymbol{\beta}}_{0\gamma},\lambda\phi|\mathcal{D}_{0\gamma}) \quad (3.15)$$

where $\mathcal{D}_0$ is defined in (3.11); for the conjugate prior (3.7), $\hat{\mathbf{V}}_{0\gamma}$ and $\hat{\boldsymbol{\beta}}_{0\gamma}$ are $\hat{\mathbf{V}}_{0\gamma}^C$ and $\hat{\boldsymbol{\beta}}_{0\gamma}^C$ in (3.9), respectively; for the power prior (3.8), $\hat{\mathbf{V}}_{0\gamma}$ and $\hat{\boldsymbol{\beta}}_{0\gamma}$ are $\hat{\mathbf{V}}_{0\gamma}^P$ and $\hat{\boldsymbol{\beta}}_{0\gamma}^P$ in (3.10), respectively. Then Corollary 2.1 yields

$$\begin{aligned}\tilde{p}(\mathbf{Y}|\gamma,\lambda,\phi) = \;\; & L(\hat{\boldsymbol{\beta}}_\gamma,\phi|\mathcal{D}_\gamma)\cdot\left\{\left|\lambda\hat{\mathbf{V}}_{0\gamma}\hat{\mathbf{V}}_\gamma^{-1}+\mathbf{I}\right|^{-\frac{1}{2}}\exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\hat{\boldsymbol{\beta}}_{0\gamma})^T(\lambda\hat{\mathbf{V}}_{0\gamma}+\hat{\mathbf{V}}_\gamma)^{-1}(\hat{\boldsymbol{\beta}}_\gamma-\hat{\boldsymbol{\beta}}_{0\gamma})}{2\phi}\right\}+\right.\\ & \left.\left|\lambda\hat{\mathbf{V}}_{0\gamma}\hat{\mathbf{V}}_\gamma^{-1}\right|^{-\frac{1}{2}}\left[\frac{L(\hat{\boldsymbol{\beta}}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})}{L(\hat{\boldsymbol{\beta}}_{0\gamma},\lambda\phi|\mathcal{D}_{0\gamma})}-\exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\hat{\boldsymbol{\beta}}_{0\gamma})^T\hat{\mathbf{V}}_{0\gamma}^{-1}(\hat{\boldsymbol{\beta}}_\gamma-\hat{\boldsymbol{\beta}}_{0\gamma})}{2\lambda\phi}\right\}\right]\right\} \quad (3.16)\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_\gamma$ and $\hat{\mathbf{V}}_\gamma$ are defined in (3.14). Due to the asymptotic normality in (3.9) and (3.10), (3.16) is in general better than the standard-form Laplace approximation, as discussed in

Section 2.2. When $\hat{\boldsymbol{\beta}}_\gamma$ is close to $\hat{\boldsymbol{\beta}}_{0\gamma}$, the second term in the brace of (3.16) is about zero so $p(\mathbf{Y}|\gamma,\lambda,\phi)$ can be approximated only by the first term, which simplifies the calculation in some degree.

One can easily calculate (3.14) and (3.16) using statistical packages for GLMs, in which MLEs and estimated covariance matrices are often standard outputs. For example, $\hat{\mathbf{V}}_\gamma$ can be calculated from the estimated covariance matrix divided by $\hat{\phi}_\gamma$ that is fitted to estimate $\phi|\gamma$ by the software, based on the current data $\mathcal{D}=(n,\mathbf{Y},\mathbf{X},\mathbf{W})$; $\hat{\mathbf{V}}_{0\gamma}$ can be calculated in the same way but based on the prior guess data $\mathcal{D}^C=(n,\boldsymbol{\mu}_0,\mathbf{X},\mathbf{W})$ for the conjugate prior of $\boldsymbol{\beta}_\gamma$, and based on the historical data $\mathcal{D}^P=(n_0,\mathbf{Y}_0,\mathbf{X}_0,\mathbf{W}_0)$ for the power prior.

Raftery (1996) proposed methods for approximating Bayes factors based on Laplace's method and Newton's method. Like those derived here, his methods use only the output of standard computer program for GLMs. In this context, the idea can be extended directly for approximating the marginal likelihood (3.5) with the conjugate or power prior of $\boldsymbol{\beta}_\gamma$. To see this, first apply Laplace's method in a fully exponential form along with the approximation (3.15) to the normalizing constant to obtain

$$p(\mathbf{Y}|\gamma,\lambda,\phi)\approx\left|\lambda\hat{\mathbf{V}}_{0\gamma}\tilde{\boldsymbol{\Psi}}_\gamma^{-1}\right|^{-\frac{1}{2}}\frac{L(\tilde{\boldsymbol{\beta}}_\gamma,\phi|\mathcal{D}_\gamma)L(\tilde{\boldsymbol{\beta}}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})}{L(\hat{\boldsymbol{\beta}}_{0\gamma},\lambda\phi|\mathcal{D}_{0\gamma})} \tag{3.17}$$

where $\tilde{\boldsymbol{\beta}}_\gamma$ is the posterior mode of $\boldsymbol{\beta}_\gamma$ under model γ given ϕ and λ, $\mathcal{D}_0$ is defined as in (3.15), $\boldsymbol{\Psi}_\gamma=-[\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}_\gamma)+\mathbf{H}(\boldsymbol{\beta}_\gamma|\mathcal{D}_{0\gamma})/\lambda]^{-1}$ and $\tilde{\boldsymbol{\Psi}}_\gamma$ is $\boldsymbol{\Psi}_\gamma$ evaluated at $\boldsymbol{\beta}_\gamma=\tilde{\boldsymbol{\beta}}_\gamma$. Next, apply one-step Newton's method to approximate $\tilde{\boldsymbol{\beta}}_\gamma$ from $\hat{\boldsymbol{\beta}}_\gamma$, namely

$$\tilde{\boldsymbol{\beta}}_\gamma\approx\hat{\boldsymbol{\beta}}_\gamma+\hat{\boldsymbol{\Psi}}_\gamma\mathbf{X}_{0\gamma}^T\mathbf{A}_{0\gamma}^{-1}[\mathbf{Y}_0-\mathbf{b}'(\boldsymbol{\theta}(\mathbf{X}_{0\gamma}\hat{\boldsymbol{\beta}}_\gamma))] \tag{3.18}$$

where $\hat{\boldsymbol{\Psi}}_\gamma$ is $\boldsymbol{\Psi}_\gamma$ evaluated at $\boldsymbol{\beta}_\gamma=\hat{\boldsymbol{\beta}}_\gamma$; $\mathbf{A}_{0\gamma}$ is diagonal with its ith diagonal element being $a_{\gamma i}=\lambda b''(\hat{\theta}_{0\gamma i})g'(b'(\hat{\theta}_{0\gamma i}))/w_i$, and $\hat{\theta}_{0\gamma i}=b'^{-1}\circ g^{-1}(\mathbf{X}_{0\gamma i}\hat{\boldsymbol{\beta}}_\gamma)$. In (3.18), $\mathbf{Y}_0=\boldsymbol{\mu}_0$ and $\mathbf{X}_0=\mathbf{X}$ for the conjugate prior of $\boldsymbol{\beta}_\gamma$. Now noting $\tilde{\boldsymbol{\Psi}}_\gamma\approx\hat{\boldsymbol{\Psi}}_\gamma=[\hat{\mathbf{V}}_\gamma^{-1}+(\lambda\hat{\mathbf{V}}_{0\gamma})^{-1}]^{-1}$ and substituting this and (3.18) in (3.17) yields an approximation for $p(\mathbf{Y}|\gamma,\lambda,\phi)$

$$\tilde{p}_{MR}(\mathbf{Y}|\gamma,\lambda,\phi)\approx\left|\lambda\hat{\mathbf{V}}_{0\gamma}\hat{\mathbf{V}}_\gamma^{-1}+\mathbf{I}\right|^{-\frac{1}{2}}\frac{L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma,\phi|\mathcal{D}_\gamma)L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})}{L(\hat{\boldsymbol{\beta}}_{0\gamma},\lambda\phi|\mathcal{D}_{0\gamma})} \tag{3.19}$$

where $\hat{\tilde{\boldsymbol{\beta}}}_\gamma$ is the right hand side of (3.18). Based on Raftery (1996), rather than using $L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma,\phi|\mathcal{D}_\gamma)$ to approximate $L(\tilde{\boldsymbol{\beta}}_\gamma,\phi|\mathcal{D}_\gamma)$ and $L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})$ to approximate $L(\tilde{\boldsymbol{\beta}}_\gamma,\lambda\phi|\mathcal{D}_{0\gamma})$

in (3.17), we should further approximate them by their second-order and first-order Taylor series expanded around $\hat{\boldsymbol{\beta}}_\gamma$, respectively. This is not employed here because $L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma, \phi|\mathcal{D}_\gamma)$ and $L(\hat{\tilde{\boldsymbol{\beta}}}_\gamma, \lambda\phi|\mathcal{D}_{0\gamma})$ are quite easy to compute given (3.18) and such steps do not reduce computing efforts but cause less accuracy than (3.19). Due to this difference, we refer (3.19) as a modified Raftery method.

The approximation (3.19) appears less accurate than the full-exponential Laplace approximation (3.17) because of the Newton's step (3.18). However, it avoids the need of calculating the posterior mode $\tilde{\boldsymbol{\beta}}_\gamma$, to which an user often does not have a direct access using standard statistical software. And it provides an alternative way to approximate $p(\mathbf{Y}|\gamma, \lambda, \phi)$ in case that the prior of $\boldsymbol{\beta}_\gamma$ cannot be approximated by a normal density well.

3.5 Posterior Computation and Stochastic Search

In situations where both the hyperparameter λ and the dispersion parameter ϕ are known or can be reasonably specified, the analytic asymptotics we have derived offer the advantage of analytical simplification which allows for exhaustive posterior evaluation in moderately sized problems. On the other hand, when a single value of λ is difficult to specify a priori or empirically, a prior on λ is necessary. Except for certain restricted examples, one cannot integrate λ out from the marginal density of the data $p(\mathbf{Y}|\gamma, \lambda, \phi)$ or its approximate representations $\tilde{p}(\mathbf{Y}|\gamma, \lambda, \phi)$, so the model posterior $\pi(\gamma|\mathbf{Y})$ (up to a normalizing constant) is analytical intractable. MCMC methods have become a standard workhorse for calculating such posterior probabilities. Even if $\pi(\gamma|\mathbf{Y})$ is in close form, MCMC can be employed to stochastically search for high posterior models, to avoid the burden of calculating the posterior probabilities of all 2^p models. For recent discussion and comparison of available MCMC methods for Bayesian variable selection, see George & McCulloch (1997), Chen et al. (2000b), Dellaportas et al. (2002), Han & Carlin (2001) and the references therein.

Without loss of generality, suppose that λ and ϕ are both unknown here. An MCMC algorithm for computing $\pi(\gamma|\mathbf{Y})$ typically operates over the sampling space created by model indicators and parameters jointly (Carlin & Chib 1995, Green 1995, etc.), i.e., $(\gamma, \boldsymbol{\beta}_\gamma, \lambda, \phi)$ in this context. As we now proceed to show, the analytic approximations $\tilde{p}(\mathbf{Y}|\gamma, \lambda, \phi)$ for $p(\mathbf{Y}|\gamma, \lambda, \phi)$ derived in Section 3.4 greatly reduce the dimension of the sampling space by sim-

ulating a Markov chain with limiting distribution $\tilde{p}\,(\gamma, \lambda, \phi|\mathbf{Y}) \propto \tilde{p}\,(\mathbf{Y}|\gamma, \lambda, \phi)\pi(\gamma)\pi(\lambda)\pi(\phi)$.

We begin with a simple Metropolis-Hastings (MH) algorithm that updates γ, λ and ϕ simultaneously. Starting with an initial state S^0, this algorithm generates each transition from $S^t = (\lambda^t, \phi^t, \gamma^t)$ to $S^{t+1} = (\lambda^{t+1}, \phi^{t+1}, \gamma^{t+1})$ as follows.

1. Generate candidate values $S^* = (\lambda^*, \phi^*, \gamma^*)$ with probability distribution $q(S^t, S^*)$.

2. Set $S^{t+1} = S^*$ with probability

$$\alpha(S^t, S^*) = \min\left\{\frac{\tilde{p}\,(\mathbf{Y}|S^*)\pi(\gamma^*)\pi(\lambda^*)\pi(\phi^*)q(S^*, S^t)}{\tilde{p}\,(\mathbf{Y}|S^t)\pi(\gamma^t)\pi(\lambda^t)\pi(\phi^t)q(S^t, S^*)},\ 1\right\}; \tag{3.20}$$

   otherwise, set $S^{t+1} = S^t$.

There are many variations or simpler versions of this MH algorithm, depending on how the transition kernel $q(S^t, S^*)$ is specified. In particular, a practical scheme is, (1) generate model γ^* from γ^t by randomly selecting one or more covariates and switching their status in γ^t (present or absent); (2) simulate a random walk from $\log\lambda^t$ to $\log\lambda^*$, where the step of the move, $\log(\lambda^*/\lambda^t)$, is proposed based on a symmetric distribution; (3) propose $\log\phi^*$ from $\log\phi^t$ in a similar way. Under this case, $q_\gamma(\gamma^t, \gamma^*)$ is symmetric in (γ^*,γ^t), so the acceptance probability (3.20) becomes

$$\alpha(S^t, S^*) = \min\left\{\frac{\tilde{p}\,(\mathbf{Y}|S^*)\pi(\gamma^*)\pi(\lambda^*)\pi(\phi^*)\lambda^*\phi^*}{\tilde{p}\,(\mathbf{Y}|S^t)\pi(\gamma^t)\pi(\lambda^t)\pi(\phi^t)\lambda^t\phi^t},\ 1\right\}. \tag{3.21}$$

To avoid traps at local maxima, achieve better mixing behaviours and increase sampling efficiency, one can easily adopt methods of parallel tempering (Geyer 1991b, Geyer & Thompson 1995) or evolutionary Monte Carlo (Liang & Wong 2000, Liang & Wong 2001) here. The basic idea of such methods is, instead of using a single long chain, one can simulate a population of Markov chains in parallel where each chain is attached to a different temperature; the population is then updated by both within-chain (mutation) and between-chain operations (crossover or exchange). For example, an MCMC algorithm with parallel tempering that entails mutation and exchange operations can be described as follows.

1. Initialize a population of size M, $\mathbf{S}^0 = \{S_1^0, ..., S_M^0\}$ at random, and decide a temperature ladder $\boldsymbol{\tau} = \{\tau_1, ..., \tau_M\}$ with $\tau_1 < \cdots < \tau_M$ and one of them is set to 1.

2. For each member of the population $\mathbf{S}^t$ at the tth iteration (say member m), run a MH algorithm to generate a sample S_m^{t+1}. Note the transition probability in (3.20) now is

$$\alpha^{PT}(S^t, S^*) = \min\left\{\left[\frac{\tilde{p}\,(\mathbf{Y}|S^*)\pi(\gamma^*)\pi(\lambda^*)\pi(\phi^*)}{\tilde{p}\,(\mathbf{Y}|S^t)\pi(\gamma^t)\pi(\lambda^t)\pi(\phi^t)}\right]^{1/\tau_m}\frac{q(S^*, S^t)}{q(S^t, S^*)},\ 1\right\}. \tag{3.22}$$

3. Exchange S_l^{t+1} with S_k^{t+1} for M pairs (l, k) with probability

$$\alpha_E^{PT}(S_l^{t+1}, S_k^{t+1}) = \min\left\{\left[\frac{\tilde{p}\,(\mathbf{Y}|S_k^{t+1})\pi(\gamma_k^{t+1})\pi(\lambda_k^{t+1})\pi(\phi_k^{t+1})}{\tilde{p}\,(\mathbf{Y}|S_l^{t+1})\pi(\gamma_l^{t+1})\pi(\lambda_l^{t+1})\pi(\phi_l^{t+1})}\right]^{1/\tau_l - 1/\tau_k},\ 1\right\} \tag{3.23}$$

   where l is sampled uniformly on $\{1, \cdots, N\}$; for $1 < l < M$, $k = l \pm 1$ with probability 0.5, for $l = 1$, then $k = 2$ and for $l = M$, then $k = M - 1$.

4. Repeat step 3 and 4 for the $(t + 1)$th iteration until the chains converge.

For the purpose of variable selection, models with high posterior probabilities are selected according to their frequencies in the simulated Markov chain with temperature 1. In situations where a single model is needed, the most-frequent model is selected.

4 Examples

4.1 Evaluation in a Simple Case

We first study the performance potential of the method proposed in Theorem 2.1 on a very simple Poisson linear model with a canonical link function. Two datasets, with $n = 10$ and 100 respectively, were simulated by generating (x_{1i}, x_{2i}) from $N(\mathbf{0}, \mathbf{I})$ and each y_i from independent Poisson with mean μ_i given by $\log \mu_i = 1 + x_{1i} - 0.5x_{2i}$ for $i = 1, \ldots, n$. Note $\boldsymbol{\beta} = (1, 1, -0.5)^T$, and $\phi = 1$, $b(\theta_i) = \mu_i = \exp(\theta_i)$ and $c(y_i, \phi) = -\log(y_i!)$ in (3.1). For comparison, we used various methods to calculate the marginal likelihoods of the simulated data, $p\,(\mathbf{Y}|\lambda) = \int p\,(\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\lambda)\,\mathbf{d}\boldsymbol{\beta}$, based on normal priors on $\boldsymbol{\beta}$ in form of $\mathbf{N}(\mathbf{m}, \lambda\mathbf{I})$. For each dataset, we considered two prior means, $\mathbf{m} = (0, 0, 0)^T$ and $\mathbf{m} = (1, 1, -0.5)^T$, and seven different λ values varying from very small to large. The methods include SL (standard-form Laplace), FEL (full-exponential Laplace), R (Raftery's method) and IS (Importance Sampling) in addition to the proposed one. The formulas for these methods are available in Appendix B. As noted

in Green (1992), it is hard to find an importance sampling distribution that works well for a wide range of λ. To overcome the difficulty, we adopt a mixture distribution to generate $\boldsymbol{\beta}$ when applying IS (Geyer 1991a). We also keep generating samples until the estimate of IS for each λ stabilizes within a small range, so the results from IS can be treated as surrogates of exact values. For implementation details about IS, see Appendix B.

The results are shown in Table 1. All methods work equally well when $\lambda \geq 1$. But their performance differs greatly for $\lambda < 1$. In this case, FEL and the proposed method are much better than SL and R, especially when n is small or λ gets closer to 0. From results for $\lambda = 1\text{E}-20$ in Table 1, we can see that as $\lambda \to 0$, no analytic method gives results converging to exact values except for the proposed one under the case of $\mathbf{m} = \boldsymbol{\beta}$. When $\mathbf{m} \neq \boldsymbol{\beta}$, as will be typically the case in the practice, both FEL and the proposed method give reasonable estimates for λ as small as 0.01.

It may be helpful to discuss the methods from a computational perspective. SL, R and the proposed method only involve the MLE of $\boldsymbol{\beta}$, estimated covariance matrix and likelihood, so they are easy to program using standard statistical software. FEL requires more programming effort because an user needs to calculate the posterior mode of $\boldsymbol{\beta}$ via an optimization algorithm and the corresponding posterior covariance matrix involving detailed formulas. IS, as a sample-based method, is easy to code but requires fine tune for fast convergence. Turning to the CPU time, based on calculations for the dataset of size 100 and a fixed λ on a workstation with 1.8GHz Xeon processor and 1GB of RAM, it is 33 ms for SL, R and the proposed method, 68 ms for FEL, and for IS to get an estimate with an error within ± 0.05, the time varies from 270 ms to 51 seconds for different λ. Overall, it appears that the proposed method is promising because of greater accuracy, less human effort and faster computing speed.

4.2 Intensive Care Unit Data

We consider a dataset from Hosmer & Lemeshow (1989) with 200 subjects, who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The response STA is the vital status of a patient at the time of hospital discharge (0=Lived, 1=Died). There are 19 predictor variables in the dataset: (1) Age; (2) Sex; (3) Race (White/Black/Other); (4) service at ICU admission (SER, Medical/Surgical); (5) cancer part

of Present Problem (CAN, No/Yes); (6) History of Chronic Renal Failure (CRN, No/Yes); (7) Infection Probable at ICU Admission (INF, No/Yes); (8) CPR prior to ICU Admission (CPR, No/Yes); (9) Systolic Blood Pressure at ICU Admission (SYS); (10) Heart Rate at ICU Admission (HRA); (11) Previous Admission to an ICU within 6 Months (PRE, No/Yes); (12) Type of Admission (TYP, Elective/Emergency); (13) Long Bone, Multiple, Neck, Single Area, or Hip Fracture (FRA, No/Yes); (14) PO2 from Initial Blood Gases ($PO2$, $> 60/ \leq 60$); (15) PH from Initial Blood Gases (PH,$\geq 7.25/ < 7.25$); (16) PCO2 from Initial Blood Gases (PCO, $\leq 45/ > 45$); (17) Bicarbonate from Initial Blood Gases (BIC, $\geq 18/ < 18$); (18) Creatinine from Initial Blood Gases (CRE, $\leq 60/ > 60$); (19) Level of Consciousness at ICU Admission (LOC, No Coma/Deep Stupor/Coma). Our aim is to select models with highest posterior probabilities out of 2^{19} or 524288 possible logistic regression models to predict the probability of survival and study the risk factors associated with ICU mortality. Note for Logistic regression, $\phi = 1$, $b(\theta_i) = \log(1 + e^{\theta_i})$ and $c(y_i, \phi) = 0$ in (3.1).

To begin with, we randomly split the dataset into two parts, one with 120 subjects for conducting variable selection and the remaining 80 subjects for cross validation. Here, we considered conjugate priors (3.7) on regression coefficients and inverse gamma hyperpriors $IG(a, b)$ on the hyperparameter λ, whose joint prior density is in form of (3.11). For illustrative purposes, the prior guess $\boldsymbol{\mu}_0$ was obtained from a prior prediction using the logistic regression model reported in Lemeshow et al. (1988), namely

$$\begin{aligned} \log \frac{\mu_{0i}}{1-\mu_{0i}} &= -1.37 + 2.44LOC_i + 1.81TYP_i + 1.49CAN_i + 0.974CPR_i + 0.965INF_i \\ &\quad +0.0368AGE_i - 0.0606SYS_i + 0.000175SYS_i^2. \end{aligned} \tag{4.1}$$

This model was fitted from the data collected on 755 patients admitted to the ICU at Baystate Medical Center in Springfield, Massachusetts between February 1 and August 15, 1983. In an actual ICU data analysis, $\boldsymbol{\mu}_0$ can be easily supplied by subjective models based on variables and associated weights determined by panels of medical "experts", such as APS and SAPS systems (see Lemeshow et al. 1988 and the references therein). For the choices of a and b, we explored four sets of values with prior means of λ at 2, 5, 10, and 50, respectively: (i) $(a, b) = (3, 4)$; (ii)$(a, b) = (2.5, 7.5)$; (iii)$(a, b) = (2.5, 15)$; and (iv) $(a, b) = (2.25, 62.5)$. For prior model probabilities, to reflect no real prior information, we took $\alpha_i = \beta_i = 1$ in (3.13) (i.e., $uniform(0, 1)$ distribution on each ω_i), which is equivalent to assigning equal probability

to each possible model. To efficiently search high posterior models, parallel tempering was applied in this example. For each pair of (a, b), four chains were simulated in parallel with a temperature ladder $\boldsymbol{\tau} = \{1, 2, 3, 4\}$ and each running 50,000 iterations; the overall acceptance rate of local updating were about 0.4, and the overall exchange rate were about 0.6.

Table 2 shows summary statistics for the prior and posterior distributions of λ for the four sets of (a, b). Overall, the posterior of λ is sensitive to the choice of (a, b). For the first set $(a, b) = (3, 4)$, the upper bound of the posterior HPD is bigger than that of the prior HPD, indicating the first prior may incorrectly concentrate on small λ values so do not assign enough probability mass to larger values (e.g., $\lambda > 4.90$); for the other three sets, the posterior HPDs are all tighter than the corresponding prior HPDs. To further decide a reasonable choice of (a, b), we calculated the average misclassification rate on the test dataset for the top 50 models from the MC chains under each hyperprior. From Table 2, $(a, b) = (2.25, 62.5)$ gives the lowest misclassification rate although $(a, b) = (2.5, 7.5)$ and $(a, b) = (2.5, 15)$ also give similar rates. Finally, we chose $(a, b) = (2.25, 62.5)$ also because the corresponding hyperprior has a heavier right tail than the others. This choice indicates the prior guess (4.1) is not to have much impact compared to the data.

Table 3 reports the top 50 models from the MC chain of temperature 1 under $(a, b) = (2.25, 62.5)$. The posterior probability of model γ was estimated from the frequency of γ in the chain divided by 40,000 (the first 10,000 iterations were discarded for the burn-in process). For comparison, we also give AIC and BIC values and ranks for each of the models: $\text{AIC}_\gamma = \log L(\hat{\boldsymbol{\beta}}_\gamma|\mathcal{D}_\gamma) - q_\gamma$ and $\text{BIC}_\gamma = \log L(\hat{\boldsymbol{\beta}}_\gamma|\mathcal{D}_\gamma) - q_\gamma \log n/2$. We can see from Table 3 that the "best" model selected by our FB procedure contains 6 covariates: $RACE$, SER, CAN, TYP, FRA, and LOC, denoted γ_{FB}; the "best" model given by AIC contains 10 covariates: AGE, $RACE$, SER, CAN, PRE, TYP, FRA, PH, PCO and LOC, denoted γ_A; and the "best" model given by BIC contains only three covariates: CAN, TYP, and LOC, denoted γ_B. It is well known that AIC tends to favor large models while BIC tends to favor small models, so it is interesting to see that $\gamma_B \subset \gamma_{FB} \subset \gamma_A$ here. In addition, γ_{FB} is the very model that AIC and BIC agree most (i.e., the model with the smallest rank sum). Another interesting feature for this dataset is, none of the top 50 models selected by AIC agrees with the top 50 models selected by BIC; for example, γ_A ranks 8098 in BIC, γ_B ranks 2588 in AIC,

but both of them are in the top 50 list of our procedure, which implies that our procedure agrees partially with both AIC and BIC, as would be expected for a good selection procedure.

We shall note that the "best" model γ_{FB} represents only 0.47% of the total posterior probability, indicating a fair amount of model uncertainty in the ICU data. For better predictive performance, Bayesian model averaging based on the top models would be recommended.

5 Discussion

In this paper, we propose new methods to approximate predictive distributions, and compare them with several existing methods. The proposed methods, when applicable, are accurate over a wide range of hyperparameter values, easy to implement and computationally efficient; in contrast, none of the other methods possess all these advantages. In the context of variable selection in GLMs, the proposed methods are employed to facilitate the implementation of a Fully Bayes approach under informative priors on regression coefficients. Ways of specifying hyperprior distributions are suggested. MCMC algorithms that operate in a sampling space with a fixed low dimension (≤ 3) are presented for posterior exploration. An illustrative example is provided to demonstrate the feasibility and usefulness of our selection procedure.

We mention that our approach to variable selection in GLMs is different from the related previous Bayesian work. Raftery (1996) presented asymptotic analytics to approximate Bayes factors and accounted for model uncertainty in GLMs; but he did not take into account the uncertainty due to the unknown hyperparameters. He also assumed independent normal priors on regression coefficients, which is a special case of our approach. Chen et al. (1999) and Ibrahim et al. (2000) concentrated on variable selection for logistic and Poisson regression models respectively, so their approaches are case specific. Also, their computation of model posteriors is purely sample-based without using any analytic approximations. Wang & George (2004) proposed several closed-form FB selection criteria for GLMs, using specific normal priors on regression coefficients. An integrated Laplace method was used to achieve analytical tractability, which is a special case of our proposed methods; the classes of conjugate and power priors were not discussed there. In conclusion, what distinguishes our work from these papers is the *generality* of our Fully Bayes approach and the novelty of our analytical approximations.

A Proof of Theorem 2.1

This proof is similar in part to that of Theorem 1 in Kass et al. (1990). Without loss of generality, we consider the case $m = 1$ for simplicity. The higher-dimensional case involves straightforward modifications. Let $h_n(\beta) \equiv -l_n(\beta)/n$ so that the integrand $L_n(\beta)\pi(\beta)$ of (2.1) can be written as $\exp[-nh_n(\beta)]\pi(\beta)$, and let $u \equiv n^{1/2}(\beta - \hat{\beta}_n)$ so that for a fixed u, $(\beta - \hat{\beta}_n)^k$ is of $O(n^{-k/2})$. Now expanding $nh_n(\beta)$ about $\hat{\beta}_n$ and e^{-x} about zero to the terms of order smaller than $O(1)$ yields

$$\exp[-nh_n(\beta)]\pi(\beta) = \exp\left[-nh_n(\hat{\beta}_n) - \frac{1}{2}h_n''(\hat{\beta}_n)u^2\right] \cdot \left\{1 - \frac{1}{6}n^{-1/2}h_n^{(3)}(\hat{\beta}_n)u^3 + R_n(u)\right\}\pi(\beta)$$

where $R_n(u)$ is of order $O(n^{-1})$ uniformly on $B_\delta(\hat{\beta}_n)$ defined in (2.3). Then

$$\int_{B_\delta(\hat{\beta}_n)} \exp[-nh_n(\beta)]\pi(\beta)\mathbf{d}\beta = \exp\left[-nh_n(\hat{\beta}_n)\right] \cdot (E_1 + E_2)$$

where

$$E_1 = \int_{B_\delta(\hat{\beta}_n)} \exp\left[-\frac{1}{2}h_n''(\hat{\beta}_n)u^2\right]\pi(\beta)\mathbf{d}\beta$$

and

$$E_2 = \int_{B_\delta(\hat{\beta}_n)} \exp\left[-\frac{1}{2}h_n''(\hat{\beta}_n)u^2\right] \cdot \left\{-\frac{1}{6}n^{-1/2}h_n^{(3)}(\hat{\beta}_n)u^3 + R_n(u)\right\}\pi(\beta)\mathbf{d}\beta.$$

Let's look at E_1 first. Note by changing the variable from β to u, we have

$$E_1 = \int_{B_{\delta(n)}(0)} \exp\left[-\frac{1}{2}h_n''(\hat{\beta}_n)u^2\right] \cdot \pi(n^{-1/2}u + \hat{\beta}_n) \cdot n^{-1/2}\mathbf{d}u \tag{A.1}$$

where $\delta(n) = n^{1/2}\delta$. Since the integration region $B_{\delta(n)}(0)$ is expanding at the rate $O(n^{1/2})$ as $n \to +\infty$, replacing this region by the whole real line incurs an error of exponentially decreasing order for the integral in (A.1), and yields (after some algebra)

$$E_1 \approx \exp\left[nh_n(\hat{\beta}_n)\right] \cdot \tilde{I} \tag{A.2}$$

in which $\tilde{I}$ is given in (2.2) . For E_2, expanding $\pi(\beta)$ about $\hat{\beta}_n$ and changing the variable from β to u, we have

$$\begin{aligned} E_2 &= \int_{B_{\delta(n)}(0)} \exp\left[-\frac{1}{2}h_n''(\hat{\beta}_n)u^2\right] \cdot \left\{-\frac{1}{6}n^{-1/2}h_n^{(3)}(\hat{\beta}_n)u^3 + R_n(u)\right\} \\ &\quad \cdot \left\{\pi(\hat{\beta}_n) + n^{-1/2}\pi'(\hat{\beta}_n)u + S_n(u)\right\} n^{-1/2}\mathbf{d}u \end{aligned} \tag{A.3}$$

where $S_n(u)$ is of order $O(n^{-1})$ uniformly on $B_\delta(\hat{\beta}_n)$. Using the same reasoning as for E_1, we replace $B_{\delta(n)}(0)$ by the real line in (A.3) and note that the third central moment of a normal distribution vanishes, so

$$E_2 = O(n^{-1}) \tag{A.4}$$

holds as long as the ith derivative of $h_n(\ i \leq 4)$ is uniformly bounded, which is satisfied automatically from the Laplace regularity of l_n. Combining (A.2) , (A.4) and (2.3) yields $\tilde{I} = I(1 + O(n^{-1}))$, which completes the proof.

B Formulas For Calculations in Section 4.1

For a GLM described by (3.2), consider calculating the marginal density of the data $p\left(\mathbf{Y}|\gamma, \lambda, \phi\right)$ based on a normal prior $\pi(\boldsymbol{\beta}_\gamma)$ in form of $\mathbf{N}(\mathbf{m}_\gamma, \lambda\phi\mathbf{U}_\gamma)$, $\lambda > 0$. Below list the formulas we derive for approximating $p\left(\mathbf{Y}|\gamma, \lambda, \phi\right)$ using different methods:

1. Standard-form Laplace

$$\tilde{p}_L\left(\mathbf{Y}|\gamma, \lambda, \phi\right) \ = L(\hat{\boldsymbol{\beta}}_\gamma, \phi|\mathcal{D}_\gamma)\left|\lambda\mathbf{U}_\gamma\hat{\mathbf{V}}_\gamma^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)^T(\lambda\mathbf{U}_\gamma)^{-1}(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)}{2\phi}\right\}$$

2. Full-exponential Laplace

$$\tilde{p}_{LF}\left(\mathbf{Y}|\gamma, \lambda, \phi\right) \ = L(\tilde{\boldsymbol{\beta}}_\gamma, \phi|\mathcal{D}_\gamma)\cdot\left|\lambda\mathbf{U}_\gamma\tilde{\mathbf{V}}_\gamma^{-1} + \mathbf{I}\right|^{-\frac{1}{2}} \exp\left\{-\frac{(\tilde{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)^T(\lambda\mathbf{U}_\gamma)^{-1}(\tilde{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)}{2\phi}\right\}$$

   where $\tilde{\boldsymbol{\beta}}_\gamma$ is the posterior mode of $\boldsymbol{\beta}_\gamma$; $\tilde{\mathbf{V}}_\gamma = (\mathbf{X}_\gamma^T\mathbf{D}_\gamma\mathbf{X}_\gamma)^{-1}$ where $\mathbf{D}_\gamma$ is diagonal with its ith diagonal element being $d_{\gamma i}$, and $\tilde{\theta}_{\gamma i} = b'^{-1} \circ g^{-1}(\mathbf{X}_{\gamma i}\tilde{\boldsymbol{\beta}}_\gamma)$,

$$d_{\gamma i} = \frac{1}{b''(\tilde{\theta}_{\gamma i})[g'(b'(\tilde{\theta}_{\gamma i}))]^2} + [y_i - b'(\tilde{\theta}_{\gamma i})]\frac{[b''(\tilde{\theta}_{\gamma i})]^2 g''(b'(\tilde{\theta}_{\gamma i})) + b^{(3)}(\tilde{\theta}_{\gamma i})g'(b'(\tilde{\theta}_{\gamma i}))}{[b''(\tilde{\theta}_{\gamma i})]^3[g'(b'(\tilde{\theta}_{\gamma i}))]^3}.$$

3. Raftery's method

$$\tilde{p}_R(\mathbf{Y}|\gamma, \lambda, \phi) \approx \left|\lambda\mathbf{U}_\gamma\hat{\mathbf{V}}_\gamma^{-1} + \mathbf{I}\right|^{-\frac{1}{2}} L(\hat{\boldsymbol{\beta}}_\gamma, \phi|\mathcal{D}_\gamma)$$
$$\cdot \exp\left\{-\frac{(\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)^T(\lambda\mathbf{U}_\gamma+\hat{\mathbf{V}}_\gamma)^{-1}[\hat{\mathbf{V}}_\gamma(\lambda\mathbf{U}_\gamma+\hat{\mathbf{V}}_\gamma)^{-1}+\mathbf{I}-\hat{\mathbf{V}}_\gamma(\lambda\mathbf{U}_\gamma)^{-1}](\hat{\boldsymbol{\beta}}_\gamma-\mathbf{m}_\gamma)}{2\phi}\right\}.$$

   This formula is derived from equation (11) in Raftery (1996).

4. Importance Sampling:

$$\tilde{p}_{IS}(\mathbf{Y}|\gamma,\lambda,\phi) = \frac{1}{M}\sum_{t=1}^{M}\frac{p\left(\mathbf{Y}|\gamma,\boldsymbol{\beta}_\gamma^{(t)},\phi\right)\pi(\boldsymbol{\beta}_\gamma^{(t)})}{h(\boldsymbol{\beta}_\gamma^{(t)})}$$

where $\boldsymbol{\beta}_\gamma^{(t)}$, $t = 1, ..., M$, are independent samples, each with probability r generated from $\pi(\boldsymbol{\beta}_\gamma)$, i.e., $\mathbf{N}(\mathbf{m}_\gamma, \lambda\phi\mathbf{U}_\gamma)$ and with probability $1-r$ generated from $\mathbf{N}(\hat{\boldsymbol{\beta}}_\gamma, \phi\hat{\mathbf{V}}_\gamma)$; and $h(\boldsymbol{\beta}_\gamma)$ is the density of the mixture, $h(\boldsymbol{\beta}_\gamma) = r\pi(\boldsymbol{\beta}_\gamma) + (1-r)f(\boldsymbol{\beta}_\gamma)$, where $f(\boldsymbol{\beta}_\gamma)$ is the pdf of $\mathbf{N}(\hat{\boldsymbol{\beta}}_\gamma, \phi\hat{\mathbf{V}}_\gamma)$. In our experiment, for λ of $1\mathrm{E}-20$, we set r equal to 0.8; for λ of 0.001, $r = 0.1$ and for any larger λ, $r = 0.5$; M is chosen to be large enough so that the estimate $\tilde{p}_{IS}(\mathbf{Y}|\gamma,\lambda,\phi)$ stabilizes within a small region.

References

Bedrick, E. J., Christensen, R., & Johnson, W. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, 51, 211–218.

Carlin, B. P. & Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society*, 57(3), 473–484.

Carlin, B. P., Kass, R. E., Lerch, F. J., & Huguenard, B. R. (1992). Predicting working memory failure: A subjective bayesian approach to model selection. *Journal of the American Statistical Association*, 87, 319–327.

Chen, M.-H. & Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica*, 13, 461–476.

Chen, M.-H., Ibrahim, J. G., & Shao, Q.-M. (2000a). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84, 121–137.

Chen, M.-H., Ibrahim, J. G., & Yiannoutsos, C. (1999). Prior elicitation, variable selection and bayesian computation for logistic regression models. *Journal of the Royal Statistical Society*, 61, 223–242.

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000b). *Monte Carlo Methods in Bayesian Computation*. Springer.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2003). *Bayesian Treed Generalized Linear Models*. Oxford: Clarendon Press.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 157–185). Oxford: University Press.

Clyde, M. A. & Parmigiani, G. (1998). Protein construct storage: Bayesian variable selection and prediction with mixtures. *Journal of Biopharmaceutical Statistics*, 8, 431–443.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12, 27–36.

Dellaportas, P. & Smith, A. F. M. (1993). Bayesian inference for generalised linear and proportional hazards models via gibbs sampling. *Applied Statistics*, 42, 443–459.

Gelfand, A. E. & Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society*, 56(3), 501–514.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.

George, E. I. & McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7, 339–373.

George, E. I., McCulloch, R. E., & Tsay, R. (1994). Two approaches to bayesian model selection with applications. In D. A. Berry, K. M. Chaloner, & J. F. Geweke (Eds.), *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*. New York.

Geyer, C. J. (1991a). *Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo*. Technical report, University of Minnesota.

Geyer, C. J. (1991b). Markov chain monte carlo maximum likelihood. In E. M. Keramigas (Ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (pp. 156–163). Fairfax, VA.

Geyer, C. J. & Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431), 909–920.

Green, P. (1992). Discussion of the paper by geyer and thompson. *Journal of the Royal Statistical Society*, 54, 683–684.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4), 711–732.

Han, C. & Carlin, B. P. (2001). Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455), 1122–1132.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression.* John Wiley |& Sons.

Ibrahim, J. G., Chen, M.-H., & Ryan, L. M. (2000). Bayesian variable selection for time series count data. *Statistica Sinica*, 10(3), 971–987.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association*, 98, 204–213.

Kass, R. E., Tierney, L., & Kadane, J. B. (1990). The validity of posterior expansions based on laplace's method. In S. Geisser, J. S. Hodges, S. J. Press, & A. Zellner (Eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics* (pp. 473–483). Amsterdam: Elsevier Science.

Kuo, L. & Mallick, B. (1998). Variable selection for regression models. *Sankhya*, 60, 65–81.

Laud, P. W. & Ibrahim, J. G. (1996). Predicitve specification of prior model probabilities in variable selection. *Biometrika*, 83, 267–274.

Lemeshow, S., Teres, D., Avrunin, J. S., & Pastides, H. (1988). Predicting the outcome of intensive care unit patients. *Journal of American Statistical Association*, 83(402), 348–356.

Liang, F. & Wong, W. H. (2000). Evolutionary monte carlo: Applications to cp model sampling and change point problem. *Statistica Sinica*, 10, 317–342.

Liang, F. & Wong, W. H. (2001). Real parameter evolutionary monte carlo with application to bayesian mixture models. *Journal of the American Statistical Association*, 96(454), 653–666.

Meyer, M. C. & Laud, P. W. (2002). Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, 97(459), 859–871.

Ntzoufras, I., Dellaportas, P., & Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111, 165–180.

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83, 251–266.

Raftery, A. E. & Richardson, S. (1993). Model selection for generalized linear models via glib, with application to epidemiology. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian Biostatistics*. New York: Marcel Dekker.

Tierney, L. & Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika*, 76, 425–437.

Verdinelli, I. & Wasserman, L. (1995). Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430), 614–618.

Wang, X. & George, E. I. (2004). A hierarchical bayes approach to variable selection for generalized linear models. *Submitted*.

Table 1: Approximate Log Marginal Likelihoods For the Poisson Regression Model

n	$\mathbf{m}$	λ	SL	FEL	R	Proposed	IS
10	(0,0,0)	1.0E-20	-8.5E+19	-2.0E+12	8.5E+19	-24.3	-20.5
		0.001	-851.8	-20.4	799.4	-24.1	-20.4
		0.01	-94.1	-20.1	43.4	-20.9	-20.1
		0.1	-21.4	-18.1	-16.7	-18.6	-18.1
		1	-17.2	-17.3	-17.3	-17.3	-17.3
		10	-19.9	-19.9	-19.9	-19.9	-19.9
		100	-23.3	-23.3	-23.3	-23.3	-23.3
	(1,0.5,-1)	1.0E-20	-6.0E+18	-4.4E+10	6.0E+18	-14.0	-14.1
		0.001	-66.3	-14.1	44.1	-14.0	-14.1
		0.01	-15.5	-14.1	-9.7	-14.0	-14.1
		0.1	-13.5	-14.5	-14.3	-14.4	-14.5
		1	-16.4	-16.6	-16.6	-16.6	-16.6
		10	-19.8	-19.9	-19.9	-19.9	-19.9
		100	-23.3	-23.3	-23.3	-23.3	-23.3
100	(0,0,0)	1.0E-20	-1.1E+20	-4.8E+11	1.1E+20	-1853.9	-939.4
		0.001	-1283.7	-710.9	-657.6	-841.1	-862.2
		0.01	-292.9	-285.3	-285.2	-285.9	-285.5
		0.1	-197.0	-196.9	-196.9	-196.9	-196.9
		1	-190.5	-190.5	-190.5	-190.5	-190.5
		10	-192.9	-192.9	-192.9	-192.9	-192.9
		100	-196.3	-196.3	-196.3	-196.3	-196.3
	(1,0.5,-1)	1.0E-20	-5.5E+17	-5.3E+09	5.5E+17	-181.2	-181.2
		0.001	-184.5	-181.7	-177.6	-181.7	-181.7
		0.01	-183.0	-183.2	-183.1	-183.2	-183.2
		0.1	-186.0	-186.0	-186.0	-186.0	-186.0
		1	-189.4	-189.4	-189.4	-189.4	-189.4
		10	-192.8	-192.8	-192.8	-192.8	-192.8
		100	-196.3	-196.3	-196.3	-196.3	-196.3

Table 2: Summary Statistics For the Prior and Posterior Distributions of λ

(a, b)	Prior				Posterior				Misclass.
	Mean	Mode	SD	95% HPD	Mean	Mode	SD	95% HPD	Rate (%)
(3, 4)	2	1	2	(0.29, 4.90)	2.49	1.83	1.21	(0.52, 6.13)	15.64
(2.5, 7.5)	5	2.14	7.07	(0.58, 13.11)	4.69	3.00	2.98	(0.94, 10.01)	12.97
(2.5, 15)	10	4.29	14.14	(1.17, 26.21)	6.91	3.54	4.78	(1.74, 13.65)	12.72
(2.25, 62.5)	50	19.23	100	(5.07, 135.99)	17.45	10.20	11.89	(6.01, 33.49)	12.55

Table 3: Top 50 Models From MCMC for the ICU Data

Model	No. Variables	Est. Posterior	AIC (rank)	BIC (rank)
3,4,5,12,13,19	**6**	**4.70**	**-51.20 (11)**	**-63.75 (61)**
2,3,4,5,12,13,19	7	4.67	-51.72 (44)	-65.66 (593)
3,4,5,12,19	5	3.95	-52.28 (173)	-63.43 (41)
3,4,5,11,12,13,19	7	3.87	-51.12 (8)	-65.06 (312)
2,3,4,5,11,12,13,19	8	3.15	-51.74 (45)	-67.07 (2326)
3,4,5,6,12,13,19	7	2.70	-51.83 (58)	-65.76 (657)
1,3,4,5,8,11,12,13,16,19	10	2.55	-51.64 (33)	-69.76 (16867)
3,4,5,6,11,12,13,15,16,19	10	2.50	-51.79 (51)	-69.91 (18400)
1,3,4,5,11,12,13,19	8	2.45	-51.28 (12)	-66.61 (1501)
1,3,4,5,12,13,16,19	8	2.35	-51.14 (9)	-66.47 (1318)
2,3,4,5,11,12,13,18,19	9	2.20	-52.41 (228)	-69.13 (11251)
3,5,12,19	4	2.15	-53.44 (1435)	-63.19 (30)
3,4,5,12,13,15,19	7	2.07	-51.96 (87)	-65.89 (752)
3,4,5,6,10,12,13,19	8	2.00	-52.78 (470)	-68.11 (5364)
3,4,5,8,11,12,13,16,19	9	2.00	-52.00 (98)	-68.72 (8428)
3,4,5,12,13,16,19	7	1.97	-51.60 (31)	-65.54 (530)
2,3,4,5,6,12,19	7	1.87	-53.16 (916)	-67.10 (2382)
3,4,5,8,11,12,19	7	1.80	-52.87 (562)	-66.81 (1786)
3,4,5,11,12,13,14,19	8	1.80	-51.91 (78)	-67.25 (2696)
5,8,12,15,16,19	6	1.77	-53.42 (1399)	-64.57 (174)
3,4,5,11,12,19	6	1.77	-52.57 (301)	-65.11 (325)
2,3,4,5,6,8,11,12,13,19	10	1.75	-52.99 (701)	-71.10 (36334)
3,4,5,10,11,12,13,19	8	1.72	-52.10 (115)	-67.43 (3114)
5,12,19	**3**	**1.67**	**-53.81 (2588)**	**-60.78 (1)**
3,5,10,12,19	5	1.67	-54.17 (4250)	-65.32 (413)
1,2,3,4,5,12,13,15,16,19	10	1.65	-51.39 (19)	-69.51 (14521)
1,3,4,5,9,11,12,13,16,18,19	11	1.65	-52.79 (478)	-72.30 (64620)
3,4,5,12,15,19	6	1.57	-52.86 (556)	-65.41 (450)
1,2,3,4,5,12,13,16,19	9	1.57	-51.69 (37)	-68.41 (6714)
2,3,4,5,11,12,13,15,16,19	10	1.57	-51.38 (17)	-69.50 (14404)
3,4,5,7,8,10,12,13,19	9	1.55	-53.56 (1730)	-70.28 (23084)
1,3,4,5,12,13,15,16,19	9	1.50	-51.08 (7)	-67.81 (4238)
3,4,5,11,12,13,16,19	8	1.47	-51.30 (13)	-66.63 (1526)
3,4,5,8,12,19	6	1.45	-52.78 (468)	-65.32 (415)
3,4,5,8,12,13,19	7	1.42	-51.96 (89)	-65.90 (761)
3,4,5,11,12,15,19	7	1.32	-53.03 (742)	-66.96 (2065)
3,4,5,9,11,12,19	7	1.30	-52.97 (679)	-66.90 (1948)
1,3,4,5,11,12,13,15,16,19	**10**	**1.30**	**-50.55 (1)**	**-68.67 (8098)**
2,3,4,5,12,13,18,19	8	1.27	-52.51 (265)	-67.84 (4350)
2,3,4,5,12,19	6	1.25	-52.86 (549)	-65.40 (446)
5,12,15,19	4	1.22	-53.84 (2723)	-62.20 (7)
3,4,5,7,12,13,19	7	1.22	-51.70 (40)	-65.63 (579)
1,3,4,5,11,12,13,17,19	9	1.22	-52.27 (168)	-68.99 (10156)
3,4,5,9,11,12,13,15,16,19	10	1.22	-51.57 (28)	-69.69 (16167)
4,5,12,19	4	1.20	-53.02 (727)	-61.38 (2)
3,4,5,7,11,12,13,19	8	1.20	-51.66 (35)	-66.99 (2126)
3,4,5,8,11,12,13,16,17,19	10	1.20	-52.96 (669)	-71.08 (35795)
1,4,5,12,19	5	1.17	-53.11 (837)	-62.87 (14)
3,5,8,12,19	5	1.17	-53.47 (1519)	-64.62 (188)
3,4,5,6,9,12,19	7	1.17	-52.86 (554)	-66.80 (1776)

Note: The estimated posterior probabilities $\hat{\pi}(\gamma|\mathbf{Y})$ were multipled by 1,000.