Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 4, 691–709
DOI: 10.1214/15-BJPS299

Consistency of hyper-g-prior-based Bayesian variable selection for generalized linear models

Ho-Hsiang Wu[a], **Marco A. R. Ferreira**[b] **and Matthew E. Gompper**[a]

[a]*University of Missouri*
[b]*Virginia Polytechnic Institute and State University*

Abstract. We study the consistency of a Bayesian variable selection procedure for generalized linear models. Specifically, we consider the consistency of a Bayes factor based on g-priors proposed by Sabanés Bové and Held [*Bayesian Analysis* **6** (2011) 387–410]. The integrals necessary for the computation of this Bayes factor are performed with Laplace approximation and Gaussian quadrature. We show that, under certain regularity conditions, the resulting Bayes factor is consistent. Furthermore, a simulation study confirms our theoretical results. Finally, we illustrate this model selection procedure with an application to a real ecological dataset.

1 Introduction

Generalized Linear Model (GLM, see McCullagh and Nelder, 1989) is a ubiquitous tool in all areas of science to study the relationship between explanatory variables and response variables of various types, for example proportions, binary, ordinal, multinomial, and count variables. An important part of the application of GLMs is variable selection, an approach used to decide what explanatory variables should be included in the model. For that aim, Sabanés Bové and Held (2011) have developed promising methodology that uses Bayes factors based on hyper-g priors. However, to the best of our knowledge, the large sample properties of these hyper-g-priors-based Bayes factors have not been studied.

Our contribution in this article is two-fold. First, we develop a more efficient algorithm than the one provided in Sabanés Bové and Held (2011). Sabanés Bové and Held (2011) adopted in their procedure the Laplace approximation (Tierney and Kadane, 1986; Raudenbush, Yang and Yosef, 2000) and the Gauss–Hermite quadrature (Pinheiro and Bates, 1995), and both two approximations require evaluation of the first two moments. While Sabanés Bové and Held (2011) use R routine to perform line search and numerical differentiation, we provide the explicit form of a Newton–Raphson-type algorithm that speeds up the computational time and locates the moments simultaneously. Second, we assume regularity conditions and present results on the variable selection consistency of hyper-g-prior-based Bayes factors.

Key words and phrases. Bayes factor, hyper-g/n prior, GLMs, model selection consistency, Zellner–Siow prior.

Received April 2014; accepted August 2015.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote a data set of n subjects which can be modeled by a GLM with density function given by

$$f(\mathbf{y}|\beta_0, \boldsymbol{\beta}) \propto \exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right),$$

where $\theta_i = \theta_i(\eta_i)$ is the canonical parameter, $b(\theta_i)$ is a function such that the mean of the response $E(y_i) = \partial b(\theta_i)/\partial\theta_i$, the variance $\text{var}(y_i) = \phi_i \partial^2 b(\theta_i)/\partial\theta_i^2$, and all higher order derivatives exist. Let $\eta_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} = h(\mu_i)$ denote the linear predictor, where $h(\cdot)$ denotes the link function that relates the mean response, $\mu_i = E(y_i)$ to a linear expression of the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, and p is the number of available covariates. Let $\mathcal{B} = \{(\beta_0, \boldsymbol{\beta}) : \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} \in \Theta, i = 1, 2, \ldots, n\}$, where Θ denotes the canonical parameter space. Moreover, we follow Sabanés Bové and Held (2011) and assume that the dispersion parameter $\phi_i = \phi/w_i$ is known and may incorporate a weight w_i. To simplify the exposition without loss of generality, we assume the canonical link function, that is, $\theta_i = \eta_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}$, unless otherwise specified.

In variable selection problems (see George, 2000), we need to select a subset of explanatory variables that are predictive of response variables from the original set of p explanatory variables. Let γ denote the model indicator so that model M_γ has a linear predictor $\eta_{\gamma i} = \beta_0 + \mathbf{x}_{\gamma i}^T\boldsymbol{\beta}_\gamma$, where $\mathbf{x}_{\gamma i} = (x_{i\gamma_1}, \ldots, x_{i\gamma_{p_\gamma}})^T$, with p_γ the number of covariates of model M_γ. Further, let $\mathbf{X}_\gamma = (\mathbf{x}_{\gamma 1}, \ldots, \mathbf{x}_{\gamma n})^T$ be the $n \times p_\gamma$ design matrix under model M_γ.

Bayesian variable selection often proceeds by comparing the model posterior probabilities. On a side note, there are other Bayesian model selection procedures that do not use model posterior probabilities such as, for example, the deviance information criterion (DIC, Spiegelhalter et al., 2002) and the full Bayesian significance test (Pereira et al., 2008). However, we focus on procedures based on the comparison of posterior probabilities of 2^p possible models, and we assume enumerating every model within the model space is computationally feasible. When all the models have the same prior probabilities, the posterior probabilities are proportional to the predictive densities. Then, the Bayes factor, that is the ratio of the predictive density of each two models, can be used for model comparison.

Since the Bayes factor is prior sensitive (e.g., see Kass and Raftery, 1995), it is crucial to assign a prior to regression coefficients that leads to desirable variable selection properties. For the estimation of Gaussian linear models, Zellner (1986) proposed the g-prior for the regression coefficients, that is

$$\boldsymbol{\beta}_\gamma|\phi, g, M_\gamma \sim N\big(0, g\phi(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\big).$$

This prior has been widely adopted in the Bayesian variable selection literature for a long time, since it results in closed form marginal likelihood density. Different choices of g have also been intensively studied ever since. Zellner and Siow (1980) proposed the now well-known Zellner–Siow prior that is obtained by, in addition

to using the above prior for $\boldsymbol{\beta}_\gamma$, assigning an inverse-gamma hyperprior with shape parameter $a = 0.5$ and scale parameter $b = 0.5n$ to g. After integrating out g, the resulting marginal prior for $\boldsymbol{\beta}_\gamma$ is a multivariate Cauchy distribution. George and Foster (2000) proposed empirical Bayes methodology for estimation of g. Liang et al. (2008) compared several common choices of prior for g for Gaussian linear models, and investigated two frequentist properties, the variable selection consistency and prediction consistency. Their studies showed that the procedures based on the Zellner–Siow prior and on the hyper-g/n prior given by $\pi(g) = \frac{1}{n}(1 + \frac{g}{n})^{-2}$ result in both variable selection and prediction consistency while still providing computation efficiency.

Sabanés Bové and Held (2011) developed an extension of the g prior to GLMs, called the "hyper-g prior". As shown in Sabanés Bové and Held (2011, Section 2), the hyper-g prior originates from Chen and Ibrahim (2003, Formula 2.6) prior, and is strongly associated with the unit information prior approach of Kass and Wasserman (1995). With the aid of fast and accurate numerical approximations, the hyper-g priors can be shown to provide promising solutions to the variable selection problem in GLMs. It is henceforth desirable to derive a justification to these empirical findings.

The remainder of this paper is organized as follows. Section 2 presents the Sabanés Bové and Held's hyper-g prior for GLMs and then shows that, under certain regularity conditions, the model selection procedure for GLMs proposed by Sabanés Bové and Held (2011) is consistent. Section 3 covers not only the original algorithm of Sabanés Bové and Held (2011), but also a competing algorithm we propose to increase computational efficiency. Section 4 confirms the theoretical results of Section 2 with a Monte Carlo study using binomial and Poisson models. Section 5 illustrates the variable selection procedure with an application to the distribution of parasites across a wildlife host population. Section 6 concludes with a brief discussion. For clarity of exposition, all proofs are presented in the Appendix.

2 Main results

Sabanés Bové and Held (2011) assign a flat prior on the intercept β_0 that is common to all models and assume a conditionally Gaussian prior for $\boldsymbol{\beta}_\gamma$:

$$\boldsymbol{\beta}_\gamma | g, M_\gamma \sim N_{p_\gamma}(\mathbf{0}, g\phi c(\mathbf{X}_\gamma^T \mathbf{W} \mathbf{X}_\gamma)^{-1}), \tag{1}$$

where ϕ is the known dispersion parameter, $\mathbf{W} = \text{diag}\{(w_1, \ldots, w_n)^T\}$, and c denotes a scale factor determined by

$$c = -\mathbf{E}\left[\frac{\partial^2 \log f(y|\eta)}{\partial \eta^2}\right]\Bigg|_{\eta=0}.$$

For instance, $c = 1$ under Poisson model with log link, $c = 4$ under binomial model with logit link, and $c = \pi/2$ under binary model with the probit link.

Note, the precision matrix of the prior, $(\mathbf{X}_\gamma^T \mathbf{W} \mathbf{X}_\gamma)/\phi c$, is the expected Fisher information matrix evaluated at $\boldsymbol{\beta} = \mathbf{0}$. Given an imaginary sample y_0 such that $y_0 = h(\mathbf{0})$, Sabanés Bové and Held (2011) showed that the Chen and Ibrahim (2003) prior has mode at $\boldsymbol{\beta} = \mathbf{0}$ and converges to (1) as the sample size increases, which justifies the choice of the precision matrix.

For the prior of hyperparameter g, Sabanés Bové and Held (2011) consider two hyperpriors:

$$\pi^{\mathrm{ZS}}(g) = IG\left(\frac{1}{2}, \frac{n}{2}\right)$$

and

$$\pi^{\mathrm{Hyp}}(g) = \frac{1}{n}\left(1 + \frac{g}{n}\right)^{-2}.$$

We also note that the same hyperpriors for g have been considered by Liang et al. (2008) for variable selection in Gaussian linear models.

To investigate the asymptotic property of Bayesian variable selection procedures, we consider the definition of consistency used in Fernandez, Ley and Steel (2001). Let $P(M_\gamma|\mathbf{y})$ be the posterior probability of model M_γ. Then the variable selection procedure is said to be consistent if

$$\lim_{n\to\infty} P(M_\gamma|\mathbf{y}) = 1, \quad \text{when } M_\gamma \text{ is the true model.}$$

Liang et al. (2008) have shown this consistency property can be equivalently defined in terms of the Bayes factor for comparing any model $M_{\gamma'}$ with the true model M_γ, that is,

$$\lim_{n\to\infty} BF[\gamma', \gamma] = \lim_{n\to\infty} \frac{BF[\gamma', N]}{BF[\gamma, N]} = 0,$$

when M_γ is the true model $\forall M_{\gamma'} \neq M_\gamma$, where $BF[\gamma, N]$, the Bayes factor for comparing model M_γ with the null model M_N, is defined in Zellner and Siow (1980) by

$$BF[\gamma, N] = f(\mathbf{y}|M_\gamma)/f(\mathbf{y}|M_N),$$

with $f(\mathbf{y}|M_\gamma)$ being the marginal likelihood density, defined as

$$f(\mathbf{y}|M_\gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(\mathbf{y}|\beta_0, \boldsymbol{\beta}_\gamma, \gamma) \int_{\mathbb{R}_+} \pi(\boldsymbol{\beta}_\gamma|g, \gamma)\pi(g)\, dg\, d\beta_0\, d\boldsymbol{\beta}_\gamma. \tag{2}$$

In general, under GLMs, there is no closed form representation for (2). To obtain a partial analytical result, valid numerical approximation is required. Application of a $(p_\gamma + 1)$-dimensional Laplace approximation (Tierney and Kadane, 1986) on (2) with respect to $(\beta_0, \boldsymbol{\beta}_\gamma)$ yields the integrated likelihood of g, with an accuracy of order $O(1/n)$. Consequently, the marginal likelihood density becomes

$$f(\mathbf{y}|M_\gamma) = \int (2\pi)^{(p_\gamma+1)/2} |\mathbf{R}^*_{0\gamma}|^{-1/2} f(\mathbf{y}|\boldsymbol{\beta}^*_{0\gamma}, \gamma)\pi(\boldsymbol{\beta}^*_\gamma|g, \gamma)\pi(g)\, dg + \mathbf{e}_n,$$

where $\mathbf{e}_n = O(n^{-1})$ is the resulting Laplace approximation error correction term. In addition, $\boldsymbol{\beta}_\gamma^*$ is the conditional posterior mode of $\boldsymbol{\beta}_\gamma$. Finally, $\boldsymbol{\beta}_{0\gamma}^*$ and $\mathbf{R}_{0\gamma}^*$ are respectively, the conditional posterior mode vector and the conditional posterior precision matrix of $(\beta_0, \boldsymbol{\beta}_\gamma)$.

We introduce additional notation. For simplicity of exposition, we incorporate the intercept β_0 into the parameter vector $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}_\gamma$, $\boldsymbol{\beta}_N$ and $\boldsymbol{\beta}_F$ denote the parameter vectors in the true data generating model M_γ, the null model M_N, and the full model M_F, respectively. We denote by $\hat{\boldsymbol{\beta}}_\gamma$ the Maximum Likelihood Estimator (MLE) of β_γ calculated from model M_γ, and similarly, denote by $\boldsymbol{\beta}_\gamma^*$ the posterior mode of β_γ. Further, we denote by $\boldsymbol{\beta}_\gamma^*(g)$ the conditional posterior mode of β_γ given g. In addition, we denote the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_t$, augmented by zeros to a $(p+1) \times 1$ vector. Further, let us denote the Fisher information matrix for the parameter $\boldsymbol{\beta}$ by $\mathbf{F}(\boldsymbol{\beta})$. Let $\mathbf{A}^{T/2}$ be a right square root of a positive definite matrix $\mathbf{A}$ such that $\mathbf{A}^{1/2}(\mathbf{A}^{T/2}) = \mathbf{A}$. Then we define the δ-neighborhood of $\boldsymbol{\beta}_t$ as the set $N(\delta)$ such that $N(\delta) = \{\boldsymbol{\beta} : \|\mathbf{F}(\boldsymbol{\beta}_t)^{T/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_t)\| \leq \delta\}$.

Now we consider setting the regularity conditions under which we can expect the hyper-g-prior-based Bayes factors to be consistent. Research on regularity conditions and implications in GLMs has progressed mostly from a frequentist point of view. For example, Fahrmeir and Kaufmann (1985) established regularity conditions for consistency and asymptotic normality for the estimation problem in GLMs. Some of these regularity conditions are useful to study the consistency of the variable selection procedure. Qian and Field (2002) provided regularity conditions and corresponding asymptotic results in the binomial model with canonical link. Qian and Wu (2006) exploited the binomial model with non-canonical link. Chen and Chen (2012) studied the variable selection consistency under a large p small n setting. We have adopted 3 regularity conditions that have been used in the above mentioned references. These conditions and some of their interpretations are stated as follows.

Condition C1. *The eigenvalues of* $\mathbf{X}'\mathbf{W}\mathbf{X}$ *satisfy* $0 < a_1 n \leq \zeta_1(\mathbf{X}'\mathbf{W}\mathbf{X}) \leq \cdots \leq \zeta_{p+1}(\mathbf{X}'\mathbf{W}\mathbf{X}) \leq a_2 n$ *for some constants* a_1 *and* a_2, *where* $\zeta_1(\mathbf{X}'\mathbf{W}\mathbf{X}) \leq \zeta_2(\mathbf{X}'\mathbf{W}\mathbf{X}) \leq \cdots \leq \zeta_{p+1}(\mathbf{X}'\mathbf{W}\mathbf{X})$ *are the ordered eigenvalues of* $\mathbf{X}'\mathbf{W}\mathbf{X}$.

Condition C2. *The eigenvalues of* $\mathbf{F}(\boldsymbol{\beta}_t)$ *satisfy* $0 < a_3 n \leq \zeta_1(\mathbf{F}(\boldsymbol{\beta}_t)) \leq \zeta_{p+1} \times (\mathbf{F}(\boldsymbol{\beta}_t)) \leq a_4 n$ *for some constants* a_3 *and* a_4.

Condition C3. *For any* $\delta > 0, \varepsilon > 0$, *there exists* n_1 *such that* $|\boldsymbol{v}'\mathbf{F}(\boldsymbol{\beta})\boldsymbol{v} - \boldsymbol{v}'\mathbf{F}(\boldsymbol{\beta}_t)\boldsymbol{v}| \leq \varepsilon \boldsymbol{v}'\mathbf{F}(\boldsymbol{\beta}_t)\boldsymbol{v}$, *for all* $\boldsymbol{v} \in \mathbb{R}^{p+1}$, $\boldsymbol{\beta} \in N(\delta)$, *and* $n > n_1$.

Remark 2.1. Condition C1 essentially requires the predictors $\mathbf{X}$ to be non trivial and finite, so that the information could be accumulated at the same rate as the sample size increases for all $\boldsymbol{\beta} \in \mathcal{B}$. For example, simply consider an n-observation

experiment. For the first k observations, we randomly assign each of their covariates a random vector, and for the remaining $n-k$ observations, we assign each of their covariates a vector of 1. Then Condition C1 fails since the information does not accumulate in a proper manner.

Remark 2.2. Condition C2 is commonly used in the literature, for example, by Qian and Wu (2006) and recently by Chen and Chen (2012), and can be thought of as a slightly stronger statement of condition (**D**) in Fahrmeir and Kaufmann (1985) for estimation consistency. Condition C2 fails whenever $0 < \partial\boldsymbol{\mu}/\partial\boldsymbol{\eta} < \infty$ is not satisfied, for example, when there are too many zero observations in a Poisson or a binomial model, or when there is a perfect data separation in a binary logistic model (Hosmer and Lemeshow, 2004). Condition C3 is equivalent to condition (**N**) in Fahrmeir and Kaufmann (1985), which states the following

$$\max_{\boldsymbol{\beta}\in N(\delta)} \left\|\mathbf{F}(\boldsymbol{\beta}_t)^{-1/2}\mathbf{F}(\boldsymbol{\beta})\mathbf{F}(\boldsymbol{\beta}_t)^{-T/2} - \mathbf{I}\right\| \to 0, \quad \forall\delta > 0.$$

Remark 2.3. Condition C3 requires the relative difference in information, within a small neighborhood of $\boldsymbol{\beta}_t$, to be arbitrarily small in all directions. In other words, Condition C3 extends Condition C2 to the neighborhood of $\boldsymbol{\beta}_t$. Condition C3 can be verified through Condition C2 and the condition $\max_{1\le i\le n}(\|\mathbf{x}_i\|^2/n) \to 0$ (Mielniczuk and Teisseyre, 2012). Thus, assuming Condition C2, C3 becomes mild.

Note that all conditions are for a general class of GLMs since no specification of link function is involved. For $\mathbf{x}'s$ drawn from a probability density, conditions follow naturally from the Law of Large Numbers (Fahrmeir and Kaufmann, 1985; Chen and Chen, 2012).

Lemma 1. *If Conditions C2 and C3 are satisfied, then the following properties hold,*

1. $\hat{\boldsymbol{\beta}}_{\gamma'} \to p\boldsymbol{\beta}_t, \forall\gamma' \supseteq \gamma$,
2. $\mathbf{F}(\boldsymbol{\beta}_t)^{T/2}(\hat{\boldsymbol{\beta}}_{\gamma'} - \boldsymbol{\beta}_t) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}), \forall\gamma' \supseteq \gamma$.

The above lemma is proved in Fahrmeir and Kaufmann (1985) (see Theorem 1 and Theorem 3 therein).

Remark 2.4. The posterior mode $\boldsymbol{\beta}^*_{\gamma'}$ is derived, as suggested by Sabanés Bové and Held (2011), using Bayesian iterative reweighted least squares (IRLS) (West, 1985; Gamerman, 1997), and hence it is easy to see that the posterior mode g^* plays a role of shrinkage factor that shrinks the posterior mode $\boldsymbol{\beta}^*$ toward the MLE $\hat{\boldsymbol{\beta}}$ almost surely, as long as g^* has at least order $O(n^{\varepsilon})$ for some $\varepsilon > 0$ and n

is large enough. To make sure g^* has order $O(n^{\varepsilon})$, we first note that the likelihood function of $\boldsymbol{\beta}^*(g)$ is strictly increasing function of g such that

$$\arg\max_g f\big(\mathbf{y}|\boldsymbol{\beta}^*(g),\gamma\big)\pi\big(\boldsymbol{\beta}^*(g)|g,\gamma\big)\pi(g) > \arg\max_g \pi\big(\boldsymbol{\beta}^*(g)|g,\gamma\big)\pi(g).$$

As a consequence, we can focus on solving $\arg\max_g \pi(\boldsymbol{\beta}^*(g)|g,\gamma)\pi(g)$ only. When $\pi^{\mathrm{ZS}}(g)$ is assigned, it is straightforward to show g^* has at least order $O(n)$ by conjugacy; when $\pi^{\mathrm{Hyp}}(g)$ is assigned, we show the assumption holds by solving the following polynomial equation,

$$\frac{\partial}{\partial g}\log\left\{g^{-a_5}\exp\left(-\frac{a_6}{g}\right)\left(1+\frac{g}{n}\right)^{-2}\right\}=0, \quad \text{for some constants } a_5, a_6 > 0.$$

Consequently, $\boldsymbol{\beta}^*_{\gamma'} \to p\boldsymbol{\beta}_t, \forall \gamma' \supseteq \gamma$.

Finally, the following theorem provides a general result of the hyper-g-prior-based Bayes factor when the regularity conditions are satisfied.

Theorem 1. *Under Conditions* C1, C2 *and* C3, *the Bayes factors under the Zellner–Siow prior and the hyper-g/n prior are consistent for variable selection.*

The proof of the theorem is given in the Appendix. In the proof of Theorem 1, we show that, to obtain consistency, the integral with respect to g needs to be a decreasing function of the sample size n. This condition is satisfied under both the Zellner–Siow prior and the hyper-g/n prior.

When the true model is not the null model, the integrals with respect to g under the Zellner–Siow prior and hyper-g/n prior penalize the model fit in a similar manner, that is, both hyper-g priors lead to more penalty on more complex models. However, when the true model is the null model, the integral under hyper-g/n prior tends to penalize the model fit equally regardless of the model complexity, while the Zellner–Siow prior still penalizes more heavily more complex models. This difference is illustrated by the simulation study presented in Section 5 and by the application presented in Section 6.

3 Computation

For the computation of the marginal likelihood of each model, $\boldsymbol{\beta}$ and g need to be integrated out numerically. Sabanés Bové and Held (2011) perform such integration with a two-step approximation procedure (Rue, Martino and Chopin, 2009): First, they compute the integrated likelihood of g by a Laplace approximation to integrate out the model parameters $\boldsymbol{\beta}$. Second, they integrate out g by Gauss–Hermite quadrature. In contrast, we compare Gauss–Hermite quadrature and Laplace approximation to integrate out g.

As mentioned earlier in Section 2, the integrated likelihood of g, $\widetilde{f}(\mathbf{y}|g,\gamma)$, results from an application of $(p_\gamma+1)$-dimensional Laplace approximation. To integrate out g, Sabanés Bové and Held (2011) first set $z=\log(g)$ and then apply the Gauss–Hermite quadrature approximation (Naylor and Smith, 1982) with respect to z to obtain the approximate marginal likelihood for model M_γ

$$\tilde{f}_1(\mathbf{y}|M_\gamma)=\sum_{j=1}^{N} m_j f_z(z_j,\mathbf{y}|\gamma), \tag{3}$$

where $f_z(z,\mathbf{y}|\gamma)=\widetilde{f}_g(\mathbf{y}|z,\gamma)f_g(z)|\mathbf{J}|_{g\to z}$ is the unnormalized approximate posterior density, z^* and σ^* are respectively, the mode and standard deviation derived from $f_z(z,\mathbf{y}|\gamma)$, $m_j=w_j\exp(t_j^2)\sqrt{2}\sigma^*$ are the weights, and $z_j=z^*+\sqrt{2}\sigma^* t_j$ are the nodes, $j=1,\ldots,N$.

An alternative approach favored by many authors (e.g., see Liang et al., 2008) for approximating the marginal likelihood of model M_γ is the univariate Laplace approximation

$$\tilde{f}_2(\mathbf{y}|M_\gamma)=\sqrt{2\pi}\hat{\sigma}_z^* f\left(\mathbf{y}|z^*,\gamma\right)f_z\left(z^*\right), \tag{4}$$

where z^* is the mode of $\log(\widetilde{f}(\mathbf{y}|z,\gamma)f_z(z))$, and $(\hat{\sigma}_z^*)^2$ is the negative inverse second derivative of $\log(\widetilde{f}(\mathbf{y}|z,\gamma)f_z(z))$. As the sample size n grows large, this approximation obtains an accuracy of order $O(1/n)$ (Tierney and Kadane, 1986).

Note, the Laplace approximation can be seen as a special case of Gauss–Hermite quadrature when the number of nodes equals one. Thus, the Laplace approximation always provides faster computation, while the Gauss–Hermite quadrature always provides higher accuracy if the target density follows normal distribution. To see the difference, we have implemented both approximation methods given in equations (3) and (4) for integrating out g in the simulation study that we present in Section 4.

Finally, we address the computation of the posterior mode and the precision of z, which are required by both the Gauss–Hermite quadrature and the Laplace approximation. While Sabanés Bové and Held (2011) computed these two quantities by applying the R function *optimize* and C function *dfridr* (Press et al., 2007, p. 231) on the unnormalized log-posterior of z, we opt for a Newton–Raphson algorithm.

For simple exposition, we drop the subscription of model index if there is no risk of confusion. Following Bayesian IRLS (West, 1985; Gamerman, 1997), we have

$$\boldsymbol{\beta}^*=(\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}\hat{\boldsymbol{\beta}},$$

where $\mathbf{R}=\text{blockdiag}(0,(\exp(z)\phi c)^{-1}\mathbf{X}'\mathbf{W}\mathbf{X})$ and $\mathbf{F}$ denotes the Fisher information matrix. As a consequence, the profile likelihood of z becomes

$$\log f_z(z,\mathbf{y}|\gamma)\propto -\frac{1}{2}\log(\mathbf{F}+\mathbf{R})+\ell(\boldsymbol{\beta}^*)+\left(1-\frac{p}{2}\right)z-\frac{\boldsymbol{\beta}^{*'}\mathbf{R}\boldsymbol{\beta}^*}{2}+\log\left(\pi(z)\right),$$

where $\ell(\cdot)$ denotes the log likelihood density.

Based on this profile likelihood, the explicit expressions of score function and Hessian matrix for the implementation of the Newton–Raphson algorithm can be derived as the following,

$$\begin{aligned}
&\frac{\partial \log f_z(z, \mathbf{y}|\gamma)}{\partial z} \\
&\quad = \frac{1}{2}\big[\mathrm{vec}\big((\mathbf{F}+\mathbf{R})^{-1}\big)\big]^T \mathrm{vec}(\mathbf{R}) \\
&\qquad + (\mathbf{y}-\boldsymbol{\mu})^T \mathbf{X}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad - \frac{1}{2}\widehat{\boldsymbol{\beta}}^T \mathbf{F}\{2(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad + \frac{1}{2}\widehat{\boldsymbol{\beta}}^T \mathbf{F}\{(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad + \mathbf{Q}_1,
\end{aligned}$$

where

$$\mathbf{Q}_1 = \begin{cases} -\frac{1}{2}\big(p+1-n\exp(-z)\big) & \text{if } \boldsymbol{\beta} \sim \text{Zellner–Siow prior,} \\ -\frac{1}{2}\big(p-2+4\big(1+n\exp(-z)\big)^{-1}\big) & \text{if } \boldsymbol{\beta} \sim \text{Hyper } g/n \text{ prior,} \end{cases}$$

$$\begin{aligned}
&\frac{\partial^2 \log f_z(z, \mathbf{y}|\gamma)}{\partial z^2} \\
&\quad = \frac{1}{2}\big\{\big[\mathrm{vec}\big((\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\big)\big]^T \mathrm{vec}(\mathbf{R}) - \big[\mathrm{vec}\big((\mathbf{F}+\mathbf{R})^{-1}\big)\big]^T \mathrm{vec}(\mathbf{R})\big\} \\
&\qquad - \widehat{\boldsymbol{\beta}}^T \mathbf{F}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad + (\mathbf{y}-\boldsymbol{\mu})^T \mathbf{X}\{2(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1} \\
&\qquad - (\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad - \frac{1}{2}\widehat{\boldsymbol{\beta}}^T \mathbf{F}\{6(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad + \frac{1}{2}\widehat{\boldsymbol{\beta}}^T \mathbf{F}\{6(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad - \frac{1}{2}\widehat{\boldsymbol{\beta}}^T \mathbf{F}\{(\mathbf{F}+\mathbf{R})^{-1}\mathbf{R}(\mathbf{F}+\mathbf{R})^{-1}\}\mathbf{F}\widehat{\boldsymbol{\beta}} \\
&\qquad + \mathbf{Q}_2,
\end{aligned}$$

where

$$\mathbf{Q}_2 = \begin{cases} -\frac{1}{2}n\exp(-z) & \text{if } \boldsymbol{\beta} \sim \text{Zellner–Siow prior,} \\ -2n\exp(z)\big(n+\exp(z)\big)^{-2} & \text{if } \boldsymbol{\beta} \sim \text{hyper } g/n \text{ prior.} \end{cases}$$

For comparison, we have simulated 1000 datasets from a Poisson regression model, each with 400 observations and 8 predictors. We compare the proposed Newton–Raphson algorithm with a combination of the two R routines *optimize* and *Hessian* with griding interval set at (0, 10). The computational time of our implementation is on average 20% less than that of using *optimize* and *Hessian*. Note that while *optimize* and *Hessian* are coded in C, our implementation is coded in R. Thus, further computational time reductions may be obtained with a C implementation of our procedure. Therefore in Sections 4 and 5 we have implemented only our proposed Newton–Raphson algorithm for finding the mode and precision to be used in the Laplace approximation and in the Gauss–Hermite quadrature. Lastly, in practical implementation, to avoid the divergence due to the initial value of z, we suggest scaling the step size. We have found that a default scale factor of 0.5 works well in practice.

4 Simulation study

To illustrate the performance of the Bayesian variable selection procedures considered here, we present a Monte Carlo simulation study on two commonly used GLM classes: the binomial and the Poisson models.

We first consider a logistic model for binomial response such that $y_i|\mathbf{x}_i, \boldsymbol{\beta} \sim \text{Binomial}(m_i, p_i)$ with $\text{logit}(p_i) = 2 + \mathbf{x}_i^T\boldsymbol{\beta}$, where we generate m_i from the discrete uniform density on $\{1, 2, \ldots, 40\}$. Under the logit link, $\mathbf{W} = \text{diag}\{m_1, \ldots, m_n\}$ and $c = 4$. The second model we consider is a Poisson model with the log link. We have $y_i|\mathbf{x}_i, \boldsymbol{\beta} \sim \text{Poisson}(\lambda_i)$ with $\log(\lambda_i) = 2 + \mathbf{x}_i^T\boldsymbol{\beta}$. Under log link, $\mathbf{W}$ is the identity matrix and $c = 1$.

Procedures: For each GLM class, we consider two hyper-g-prior-based procedures: the Zellner–Siow's prior based procedure, denoted by ZS, and the hyper-g/n prior based procedure, denoted by Hyp. For each procedure, we have implemented Laplace approximation and Gauss–Hermite quadrature for integrating out the hyperparameter g, and thus we index these two numerical methods by subscription *LA* and *GH*, respectively. For the comparison, we consider the standard default criterion BIC (Schwarz, 1978).

General setting: We fix $p = 8$ across GLM classes. For each GLM class, we consider two values of p_γ, $p_\gamma = 0$ and $p_\gamma = 3$. For each p_γ, we consider three sample sizes $n = 30$, 100 and 400. For each sample size, we perform 1000 Monte Carlo experiments.

Assessment of performance: We use the highest model posterior probability as selection criterion. Since we assume all models share the same model prior probability, selecting the highest model posterior probability is equivalent to selecting the highest marginal likelihood. For BIC, the criterion is also to select highest marginal likelihood among competing models, where the BIC-based approximate marginal likelihood for M_γ is defined by $f(\mathbf{y}|M_\gamma) = \exp(-\frac{1}{2}\,\text{BIC}(\gamma))$ (see Kass

and Raftery, 1995). For each selection procedure, we measure its performance by the proportion of Monte Carlo experiments for which the true model is selected as the best model.

Generating X: We generate $p = 8$ covariates from the multivariate $N(0_n, I_n)$ and then apply Gram–Schmidt orthogonalization on the generated covariates such that $\mathbf{X}^T\mathbf{X}$ is a diagonal matrix with all the diagonal elements being of order n.

Values of $\boldsymbol{\beta}_\gamma$: We compare two possible values of $\boldsymbol{\beta}_\gamma$, $\boldsymbol{\beta}_\gamma = (1, 1, 1, 0, 0, 0, 0, 0)$ and $\boldsymbol{\beta}_\gamma = (0.3, 0.3, 0.3, 0, 0, 0, 0, 0)$.

Simulation results: Table 1 summarizes the results under the binomial logistic model, and Table 2 summarizes the results under the Poisson model.

We begin with the binomial logistic model. In the case of $p_\gamma = 3$ and $\boldsymbol{\beta}_\gamma = (1, 1, 1, 0, 0, 0, 0, 0)$, hyper-$g$ prior procedures outperform the BIC across sample sizes. The *LA* and *GH* approximations perform in a similar manner. In the case of $p_\gamma = 3$ and $\boldsymbol{\beta}_\gamma = (0.3, 0.3, 0.3, 0, 0, 0, 0, 0)$, BIC performs better at $n = 30$, but hyper-g prior procedures perform better as n increases over 100. Again, the *LA* and *GH* approximations perform in a similar manner. In the case of $p_\gamma = 0$, the ZS performs similarly to BIC for medium and large sample sizes, slightly outperforms

Table 1 *Binomial logistic model*: *success rate* (*SR*) *of selecting the true model as the best model*

	$p_\gamma = 3$ $\boldsymbol{\beta}_\gamma = (1, 1, 1, 0, 0, 0, 0, 0)$			$p_\gamma = 3$ $\boldsymbol{\beta}_\gamma = (0.3, 0.3, 0.3, 0, 0, 0, 0, 0)$			$p_\gamma = 0$		
	Sample size			Sample size			Sample size		
Procedure	30	100	400	30	100	400	30	100	400
ZS_{LA}	0.820	0.937	0.964	0.482	0.868	0.910	0.614	0.766	0.875
ZS_{GH}	0.839	0.929	0.965	0.481	0.860	0.888	0.610	0.739	0.870
Hyp_{LA}	0.804	0.931	0.957	0.482	0.864	0.910	0.504	0.645	0.793
Hyp_{GH}	0.821	0.920	0.960	0.480	0.853	0.867	0.490	0.622	0.788
BIC	0.695	0.857	0.925	0.502	0.844	0.877	0.593	0.768	0.887

Table 2 *Poisson model*: *success rate* (*SR*) *of selecting the true model as the best model*

	$p_\gamma = 3$ $\boldsymbol{\beta}_\gamma = (1, 1, 1, 0, 0, 0, 0, 0)$			$p_\gamma = 3$ $\boldsymbol{\beta}_\gamma = (0.3, 0.3, 0.3, 0, 0, 0, 0, 0)$			$p_\gamma = 0$		
	Sample size			Sample size			Sample size		
Procedure	30	100	400	30	100	400	30	100	400
ZS_{LA}	0.949	0.964	0.992	0.773	0.901	0.962	0.893	0.908	0.922
ZS_{GH}	0.947	0.964	0.991	0.775	0.902	0.969	0.880	0.906	0.920
Hyp_{LA}	0.946	0.962	0.990	0.728	0.869	0.941	0.809	0.884	0.911
Hyp_{GH}	0.946	0.962	0.990	0.726	0.866	0.941	0.788	0.873	0.907
BIC	0.748	0.854	0.927	0.684	0.833	0.890	0.587	0.762	0.845

BIC for small sample sizes, and consistently outperforms the Hyp regardless of sample sizes. In terms of numerical approximations, *LA* performs considerably better than *GH*.

We proceed to discuss the example of Poisson model. For all cases, hyper-g prior procedures outperform BIC. In particular, ZS outperforms Hyp. In terms of numerical approximations, the *LA* and *GH* approximations perform in similar manner.

Overall, we recommend the use of the ZS_{LA} since it yields satisfactory frequentist property across different simulation settings at cost of less computational time than ZS_{GH}.

5 Application: Raccoon-ticks interaction

In this section, we consider a real ecological dataset to investigate whether parasite infrapopulation size can be associated with certain host, parasite and environmental explanatory variables. An infrapopulation is defined as the population of parasites on a single host. The data consist of counts of the number of ticks (*Dermacentor variabilis*) parasitizing $n = 228$ raccoons (*Procyon lotor*) caught within the central Missouri, USA, with each host observation providing information on 6 variables: total number of ticks, number of replete ticks (that is, ticks that have consumed a blood meal, which is a necessary step for reproduction), sex ratio of the tick infrapopulation, host sex, host age, and time of the year when the raccoon was caught. For further information on the ecological system, see Monello and Gompper (2007), Monello and Gompper (2010) and Ruiz-López et al. (2012). Here, parasite success will be indicated by whether the parasite has fed and is replete. Let m_i be the total number of ticks on raccoon i. Further, let y_i denote the number of replete ticks on raccoon i, $i = 1, \ldots, n$. We fit a logistic model with $p = 8$ explanatory variables that are described in Table 3. To assess the robustness with respect to link function, we have also fitted a probit model. Because the results under the probit and logistic models are similar, for the sake of brevity, we only

Table 3 *Description of the variables in the Missouri raccoon parasite dataset*

Variables	Description
m_i	Total number of ticks on host i
y_i	Number of replete ticks on host i
x_1	Proportion of male ticks
x_2	Sex of host (1 = female, 0 = male)
x_3	Season index (from -1 to 1)
x_4, x_5, x_6, x_7, x_8	Raccoon age indicators, total of 6 groups (x_4: youngest group; x_8: 2nd oldest group; oldest group is used as baseline.)

Table 4 *Highest posterior probability models. The symbol $*$ indicates that the corresponding variable is included in the highest posterior probability model*

	$n = 30$			$n = 100$			$n = 228$		
Variables	ZS_{LA}	Hyp_{LA}	BIC	ZS_{LA}	Hyp_{LA}	BIC	ZS_{LA}	Hyp_{LA}	BIC
x_1					*				
x_2	*	*	*	*	*		*	*	*
x_3									
x_4				*	*	*			
x_5					*	*			
x_6				*	*	*	*	*	*
x_7	*	*	*	*	*	*	*	*	*
x_8					*	*			

present results using the logistic model. In the logistic model, we are interested in the replete probability p_i.

We consider BIC, hyper-g/n-prior-based and Zellner–Siow-prior-based variable selection procedures computed with Laplace approximation. Further, to investigate the effect of changes of sample size in different variable selection procedures, we consider two random subsamples of sizes 30 and 100 of the original dataset, as well as the original sample of size 228. The subsamples have been restricted to the class of subsamples for which the matrix $\mathbf{X}'\mathbf{X}$ has full rank. In addition, to study the effect of accumulation of information, the subsample of size 30 is a subset of the subsample of size 100.

Table 4 displays the variables included in the highest posterior probability model by BIC, ZS_{LA}, and Hyp_{LA}, respectively, under different sample sizes. We can see that when the complete dataset is used, BIC, ZS_{LA} and Hyp_{LA} all clearly select variables x_2, x_6, and x_7. Thus we consider the model that includes x_2, x_6 and x_7 as the benchmark model. These results are similar to previous findings that indicate host sex and age as strong predictors of the number of ticks and replete ticks on a host (Monello and Gompper, 2007; Monello and Gompper, 2010), and also shows that among the different age groups, it is parasitism of the oldest individuals that are especially informative to the patterns of parasitism across the host population. When the sample size is $n = 30$, the three procedures coincide by having x_2 and x_7 in their highest posterior probability model. When the sample size increases to $n = 100$, Hyp_{LA} selects x_1, x_2, x_4, x_5, x_6, x_7, and x_8, that is, Hyp_{LA} includes many more variables than the benchmark model. This undesirable behavior of the Hyp_{LA} procedure seems to be a result of a weaker penalty for more complex models imposed by the hyper-g/n prior. Meanwhile, for $n = 100$, the BIC selects x_4, x_5, x_6, x_7, and x_8, but leaves out the important variable x_2 (sex of the host). Finally, for $n = 100$, ZS_{LA} selects x_2, x_4, x_6, and x_7, which is really close to the benchmark model. Therefore, the Zellner–Siow-prior-based Bayes factor provides results that are much more stable across different sample sizes.

6 Discussion

In this paper, we have shown that under certain mild conditions, the hyper-g-prior-based Bayes factors proposed by Sabanés Bové and Held (2011) for model selection in GLMs are consistent. We confirm this consistency result with a simulation study with both binomial logistic and Poisson models. Under the binomial logistic model, both hyper-g-prior-based variable selection procedures perform comparably to the BIC for large sample size. Under the Poisson model, both hyper-g-prior-based variable selection procedures outperform the BIC across sample sizes. Overall, Zellner–Siow-prior-based procedure provide satisfactory performance in most cases. In terms of the numerical approximations, both the *LA* and *GH* approximations perform in a similar manner.

Note, although we have two different numerical methods implemented in the simulation study, we did not study the accuracy of their approximation to the marginal likelihood. Either the use of Laplace approximation or that of Gauss–Hermite quadrature for calculating the marginal likelihood have the practical implication of limiting the computation of the integral to the region of the parameter space closer to the posterior mode. As a consequence, the well-known sensitivity of the marginal likelihood to the tail behavior of the prior density is greatly reduced by either numerical methods. For example, given a regression model, if we assign a uniform prior on the unknown parameter, the exact marginal likelihood is undefined. However, an application of Laplace approximation yields the BIC. As a consequence, we should focus on the variable selection performance instead of the approximation accuracy.

There are many possible directions for future research. One possible future research area is the study of consistency of hyper-g-prior-based Bayes factors for GLMs under the large p small n setting, where the number of regressors p is no longer fixed. Another possible research direction is the development of a Bayes factor for the case when the dispersion parameter is unknown. In that case, a prior density would be assigned for the dispersion parameter. An interesting question would be under what conditions would the resulting Bayes factor be consistent. A particularly promising research direction is the study of variable selection procedures in the presence of overdispersion.

A possible way to deal with overdispersion is through the use of random effects in the context of generalized linear mixed models (GLMM). Accordingly, we have started the study on the application of g priors under the GLMM framework, including the study of consistency, and expect to present the results elsewhere in the future.

Appendix: Proof of Theorem 1

Let C with different numbering of subscription represent distinct constants. Lowercase c still denotes the scale factor defined in the Section 2.

Recall that the Bayes factor for comparing any other model $M_{\gamma'}$ with the true model M_γ is

$$BF[\gamma', \gamma] = \frac{BF[\gamma', N]}{BF[\gamma, N]}.$$

Using Laplace approximation, we can approximate $BF[\gamma', N]$ and have

$$\begin{aligned} BF[\gamma', N] &= C_1 \int_0^\infty \Big\{ \exp[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_N)] |\mathbf{X}_{\gamma'} \mathbf{W} \mathbf{X}_{\gamma'}|^{1/2} \Big(\frac{|\mathbf{F}_N + \mathbf{R}_N|}{|\mathbf{F}_{\gamma'} + \mathbf{R}_{\gamma'}|} \Big)^{1/2} \\ &\qquad \times g^{-p_{\gamma'}/2} \exp\Big(-\frac{\boldsymbol{\beta}^{*'}_{\gamma'} \mathbf{X}_{\gamma'} \mathbf{W} \mathbf{X}_{\gamma'} \boldsymbol{\beta}^*_{\gamma'}}{2g\phi c} \Big) \pi(g) \Big\} dg \\ &= C_2 \exp[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_N)] \\ &\qquad \times \int_0^\infty \Big(1 + \frac{1}{g} \Big)^{-(p_{\gamma'}+1)/2} g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g} \Big) \pi(g)\, dg, \end{aligned}$$

with $B^*_{\gamma'}$ denoting the quadratic form $(\boldsymbol{\beta}^{*'}_{\gamma'} \mathbf{X}_{\gamma'} \mathbf{W} \mathbf{X}_{\gamma'} \boldsymbol{\beta}^*_{\gamma'})/\phi c$. Note, the second equality follows from Condition C1 and Remark 2.2. Next, for the sake of simplicity, we define

$$F_a(p, B^*, \pi) = C_3 \int_a^\infty \Big(1 + \frac{1}{g} \Big)^{-(p+1)/2} g^{-p/2} \exp\Big(-\frac{B^*}{2g} \Big) \pi(g)\, dg.$$

Consequently, the Bayes factor $BF[\gamma', \gamma]$ becomes

$$BF[\gamma', \gamma] = C \exp[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_\gamma)] \Big\{ \frac{F_0(p_{\gamma'}, B^*_{\gamma'}, \pi)}{F_0(p_\gamma, B^*_\gamma, \pi)} \Big\}.$$

We organize this proof in two cases: $M_\gamma \neq M_{\text{Null}}$ and $M_\gamma = M_{\text{Null}}$. Moreover, for the case $M_\gamma \neq M_{\text{Null}}$, we consider two subcases: $M_\gamma \subset M_{\gamma'}$ and $M_\gamma \nsubseteq M_{\gamma'}$.

Case 1. $M_\gamma \neq M_{\text{Null}}$.

We first show for both the Zellner–Siow prior and the hyper-g/n prior that $F_0(p_{\gamma'}, B^*_{\gamma'}, \pi)$ has order $O(n^{-p_{\gamma'}/2})$. The proof is by the squeeze theorem. First, note that regularity Condition C1 implies that $B^*_{\gamma'}$ is of order $O(n)$. Therefore, we can write $B^*_{\gamma'} = nd_n$ where $d_n = O(1)$. We now define

$$\begin{aligned} I_a(p, n, h_n) &= C_4 \int_a^\infty g^{-p/2-1} \exp\Big(-\frac{h_n n}{2g} \Big) dg \\ &= C_5 n^{-p/2} \Big(1 - \int_{(h_n n)/a}^\infty t^{p/2-1} e^{-t}\, dt \Big), \quad a, p \in \mathbb{N}. \end{aligned}$$

Note that when $h_n = O(1)$ then $I_a(p, n, h_n) = O(n^{-p/2})$.

Then, if the Zellner–Siow prior is used, we have

$$\begin{aligned} F_0\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{ZS}}\big) &\geq F_1\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{ZS}}\big) \\ &\geq C_6 \int_1^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\pi^{\mathrm{ZS}}(g)\,dg \\ &= C_7 n^{1/2} \int_1^\infty g^{-(p_{\gamma'}+1)/2-1} \exp\Big[-\frac{n(d_n+1)}{2g}\Big]dg \\ &= C_7 n^{1/2} I_1(p_{\gamma'}+1, n, d_n+1) = O\big(n^{-p_{\gamma'}/2}\big) \end{aligned}$$

and

$$\begin{aligned} F_0\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{ZS}}\big) &\leq C_8 \int_0^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\pi^{\mathrm{ZS}}(g)\,dg \\ &= C_9 n^{1/2} I_0(p_{\gamma'}+1, n, d_n+1) = O\big(n^{-p_{\gamma'}/2}\big). \end{aligned}$$

Therefore, $F_0(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{ZS}}) = O(n^{-p_{\gamma'}/2})$.

Similarly, if the hyper-g/n prior is used, we have

$$\begin{aligned} F_0\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{Hyp}}\big) &\geq F_n\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{Hyp}}\big) \\ &\geq C_{10} \int_n^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\frac{1}{n}\Big(1+\frac{g}{n}\Big)^{-2} dg \\ &\geq C_{10} \int_n^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\frac{1}{n}\Big(\frac{2g}{n}\Big)^{-2} dg \\ &\geq C_{11} n \int_n^\infty g^{-(p_{\gamma'}+2)/2-1} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big) dg \\ &= C_{11} n I_n(p_{\gamma'}+2, n, d_n) = O\big(n^{-p_{\gamma'}/2}\big) \end{aligned}$$

and

$$\begin{aligned} F_0\big(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{Hyp}}\big) &\leq C_{12} \int_0^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\frac{1}{n}\Big(1+\frac{g}{n}\Big)^{-2} dg \\ &\leq C_{12} \int_0^\infty g^{-p_{\gamma'}/2} \exp\Big(-\frac{B^*_{\gamma'}}{2g}\Big)\frac{1}{n}\Big(\frac{g}{n}\Big)^{-2} dg \\ &= C_{12} n I_0(p_{\gamma'}+2, n, d_n) = O\big(n^{-p_{\gamma'}/2}\big). \end{aligned}$$

Therefore, $F_0(p_{\gamma'}, B^*_{\gamma'}, \pi^{\mathrm{Hyp}}) = O(n^{-p_{\gamma'}/2})$. *Then for both Zellner–Siow and hyper-g/n priors, the Bayes factor* $BF[\gamma', \gamma]$ *becomes*

$$BF\big[\gamma', \gamma\big] = C_{13} \exp\big[\ell\big(\boldsymbol{\beta}^*_{\gamma'}\big) - \ell\big(\boldsymbol{\beta}^*_{\gamma}\big)\big] O\big(n^{-(p_{\gamma'}/2 - p_\gamma/2)}\big).$$

Consider the following two subcases.

(a) $M_\gamma \subset M_{\gamma'}$.

When Conditions C2 and C3 are satisfied and the true model is nested in the competing model, $[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_{\gamma})]$ *converges to an asymptotic Chi-square distribution with number of degrees of freedom equal to* $(p_{\gamma'} - p_\gamma)$ *(see* Ferguson, 1996*). Then for both Zellner–Siow and hyper-*g/n *priors, we have*

$$BF[\gamma', \gamma] = C_{14} \exp\bigl(\chi^2_{p_{\gamma'} - p_\gamma}\bigr) O\bigl(n^{-(p_{\gamma'}/2 - p_\gamma/2)}\bigr).$$

Since $\frac{p_{\gamma'}}{2} - \frac{p_\gamma}{2} > 0$, $BF[\gamma', \gamma]$ *converges to zero.*

(b) $M_\gamma \nsubseteq M_{\gamma'}$.

When Conditions C2 and C3 are satisfied but the true model is not nested in the competing model, $\exp[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_{\gamma})]$ *converges to zero exponentially fast with respect to* n*, no matter what value* $(p_{\gamma'} - p_\gamma)$ *may take. As a consequence, the Bayes factor* $BF[\gamma', \gamma]$ *converges to zero exponentially fast.*

Case 2. $M_\gamma = M_{\text{Null}}$.

When the true model is the null model, we have $M_\gamma \subset M_{\gamma'}$ *as well as* $BF[\gamma', \gamma] = BF[\gamma', N]$*. Moreover,* $[\ell(\boldsymbol{\beta}^*_{\gamma'}) - \ell(\boldsymbol{\beta}^*_{N})]$ *converges to an asymptotic Chi-square distribution with* $p_{\gamma'}$ *degrees of freedom. Further, the Bayes factor* $BF[\gamma', N]$ *is determined by* $F_0(p_{\gamma'}, B^*_{\gamma'}, \pi)$*, i.e.,*

$$BF[\gamma', \gamma] = BF[\gamma', N] = C_{15} \exp\bigl(\chi^2_{p_{\gamma'}}\bigr) F_0\bigl(p_{\gamma'}, B^*_{\gamma'}, \pi\bigr).$$

If the Zellner–Siow prior is used, it is straightforward to show the order of the integral $F_0(p_{\gamma'}, B^*_{\gamma'}, \pi^{\text{ZS}})$ *is* $O(n^{-p_{\gamma'}/2})$*. In addition, in this case* $B^*_{\gamma'}$ *has order* $O(1)$ *instead of order* $O(n)$*. As a consequence, if the hyper-*g/n *prior is used, we have*

$$\begin{aligned} F_0\bigl(p_{\gamma'}, B^*_{\gamma'}, \pi^{\text{Hyp}}\bigr) &\le C_{16} \int_0^\infty (1 + n\alpha)^{-p_{\gamma'}/2} (1 + \alpha)^{-2}\, d\alpha \\ &\le C_{17} \int_0^{1/n} (1 + \alpha)^{-2}\, d\alpha + C_{18} \int_{1/n}^1 (n\alpha)^{-p_{\gamma'}/2}\, d\alpha \\ &\quad + C_{19} \int_1^\infty n^{-p_{\gamma'}/2} \alpha^{-p_{\gamma'}/2 - 2}\, d\alpha \\ &= O\bigl(n^{-1}\bigr), \end{aligned}$$

where $\alpha = 1/g$*. As a result, for both Zellner–Siow prior and hyper-*g/n *prior, the Bayes factor* $BF[\gamma', N]$ *converges to zero.*

Acknowledgments

The authors would like to thank the Editor, the Associate Editor and the two referees for their meticulous review and comments which greatly improved the paper.

References

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* **22**, 555–574. MR2954352

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica* **13**, 461–476. MR1977737

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**, 342–368. MR0773172

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall/CRC. MR1699953

Fernandez, C., Ley, E. and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**, 381–427. MR1820410

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7**, 57–68.

George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747. MR1813972

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308. MR1825282

Hosmer, D. W. Jr. and Lemeshow, S. (2004). *Applied Logistic Regression*. New York: Wiley.

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

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934. MR1354008

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423. MR2420243

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Model*. London: Chapman & Hall/CRC. MR3223057

Mielniczuk, J. and Teisseyre, P. (2012). Model selection in logistic regression using p-values and greedy search. In *Security and Intelligent Information Systems, Vol. 7053* (P. Bouvry, M. Kłopotek, F. Leprévost, M. Marciniak, A. Mykowiecka and H. Rybiński, eds.) 128–141. Berlin Heidelberg: Springer-Verlag GmbH.

Monello, R. and Gompper, M. (2007). Biotic and abiotic predictors of tick (Dermacentor variabilis) abundance and engorgement on free-ranging raccoons (Procyon lotor). *Parasitology* **134**, 2053–2062.

Monello, R. J. and Gompper, M. E. (2010). Differential effects of experimental increases in sociality on ectoparasites of free-ranging raccoons. *Journal of Animal Ecology* **79**, 602–609.

Naylor, J. C. and Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics* **31**, 214–225. MR0694917

Pereira, C. A. D. B., Stern, J. M. and Wechsler, S. (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis* **3**, 79–100. MR2383252

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge: Cambridge Univ. Press. MR2371990

Qian, G. and Field, C. (2002). Law of iterated logarithm and consistent model selection criterion in logistic regression. *Statistics & Probability Letters* **56**, 101–112. MR1881536

Qian, G. and Wu, Y. (2006). Strong limit theorems on model selection in generalized linear regression with binomial responses. *Statistica Sinica* **16**, 1335–1365. MR2327494

Raudenbush, S. W., Yang, M.-L. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–157. MR1826278

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319–392. MR2649602

Ruiz-López, M. J., Monello, R. J., Gompper, M. E. and Eggert, L. S. (2012). The effect and relative importance of neutral genetic diversity for predicting parasitism varies across parasite taxa. *PLoS ONE* **7**, e45404.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis* **6**, 387–410. MR2843537

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. MR0468014

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 583–639. MR1979380

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86. MR0830567

West, M. (1985). Generalized linear models: Scale parameters, outlier accomodation, and prior distributions. In *Bayesian Statistics, Vol. 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 531–558. Amsterdam: North-Holland. MR0862501

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti, Vol. 6*, 233–243. Amsterdam: North-Holland. MR0881437

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603.

Ho-Hsiang Wu
Department of Statistics
University of Missouri—Columbia
146 Middlebush Hall
Columbia, MO 65211-6100
E-mail: hwd23@mail.missouri.edu

Marco A. R. Ferreira
Department of Statistics
Virginia Tech
210-B Hutcheson Hall (MC 0439)
250 Drillfield Drive
Blacksburg, Virginia 24061
USA
E-mail: marf@vt.edu

Matthew E. Gompper
Department of Fisheries and Wildlife Sciences
University of Missouri—Columbia
302 Anheuser-Busch Natural Resources Building
Columbia, MO 65211-6100
E-mail: GompperM@missouri.edu