# A generalization of BIC for the general exponential family

Arijit Chakrabarti[a,*], Jayanta K. Ghosh[b,c]

[a]The Ohio State University, Department of Statistics, 1958 Neil Avenue, Columbus, OH 43210, USA
[b]Purdue University, Department of Statistics, 150 North University Street, West Lafayette, IN 47907-2067, USA
[c]Indian Statistical Institute, 203 B. T. Road, Calcutta 700108, India

Received 27 October 2003; accepted 18 January 2005
Available online 14 March 2005

**Abstract**

In a normal example of Stone (1979, J. Roy. Statist. Soc. Ser. B 41, 276–278), Berger et al. (2003, J. Statist. Plann. Inference 112, 241–258) showed BIC may be a poor approximation to the logarithm of Bayes Factor. They proposed a Generalized Bayes Information Criterion (GBIC) and a Laplace approximation to the log Bayes factor in that problem. We consider a fairly general case where one has $p$ groups of observations coming from an arbitrary general exponential family with each group having a different parameter and $r$ observations. We derive a GBIC and a Laplace approximation to the integrated likelihood, under the assumption that $p \to \infty$ and $r \to \infty$ (and some additional restrictions, which vary from example to example). The general derivation clarifies the structure of GBIC. A general theorem is presented to prove the accuracy of approximation, and the worst possible approximation error is derived for several examples. In several numerical examples, the Laplace approximation and GBIC are seen to be quite good. They perform much better than BIC.

*MSC:* 62F15

*Keywords:* Bayes factor; BIC; AIC; Posterior normality; Laplace approximation

## 1. Introduction

Approximations to Bayes factors by Schwarz (1978), BIC, further improvements in approximation, and interesting applications have been studied by Kass and Vaidyanathan (1992), Kass and Wasserman (1995), Kass and Raftery (1995), Raftery (1996), Pauler (1998), and Volinsky and Raftery (2000). As in Schwarz, the highest dimension $p$ is fixed, while the sample size tends to infinity. The BIC also emerges as an approximation in model selection via computational complexity or equivalently a Kullback–Leibler loss function in Rissanen (1978, 1983). A very interesting paper by Spiegelhalter et al. (2002) is based on the last criterion. However, the approximation and asymptotic qualities of their new method are yet to be studied.

Our interest here is in developing an approximation that throws light on the behaviour of the Bayes Factor when both the dimension and the sample size tend to infinity.

In the first study of this kind, Berger Ghosh and Mukhopadhyay (2003), henceforth abbreviated as BGM, have shown that the BIC may not be a good approximation to the Bayes Factor (see Tables 1 and 2 of BGM) in a problem of Stone (1979). A much better approximation is provided by their new information criterion Generalized Bayes Information Criterion (GBIC), which reduces to the BIC when $p$ (the dimension) is fixed. Stone (1979) had shown earlier that BIC is an inconsistent model selection criterion in the same problem.

BGM consider an equivalent formulation of Stone's problem with $y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, r$, $n = pr$ and $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$ and the models being compared are $M_1 : \mu_i = 0$ for $i = 1, 2, \ldots, p$ vs. $M_2 : \boldsymbol{\mu} \in \mathbb{R}^p$. A fully Bayesian method of model selection is then proposed based on the Bayes Factor,

$$\mathrm{BF}_{21} = \frac{\int \mathrm{e}^{L_2(\boldsymbol{\mu})}\pi(\boldsymbol{\mu})\,\mathrm{d}\boldsymbol{\mu}}{\mathrm{e}^{L_1(\mathbf{0})}}. \tag{1}$$

$L_j(\boldsymbol{\mu})$ is the log-likelihood function under model $M_j$ and $\pi(\boldsymbol{\mu})$ the well-known Zellner–Siow (1980) multivariate Cauchy prior. We focus on numerator since the denominator is easy to calculate. A major result in BGM is a rigorous Laplace approximation to log $\mathrm{BF}_{21}$ which leads to the GBIC.

Stone proved the inconsistency of $(p/2)\log n$ as the BIC penalty term. It can be easily verified that Stone's counterexample holds for BIC even under the more appropriate penalty $(p/2)\log r$.

This paper considers a whole new class of examples where a Laplace approximation, and hence GBIC (see Eqs. (4) and (9) in Section 2.1), are valid for certain sets of values of $p$ and $r$. Our proof is new and sheds light on the structure displayed in (4) and (9). Instead of a normal family, we consider a general exponential family, and instead of the Cauchy prior, we consider a general mixture of conjugate priors. In this generality we apply the Laplace approximation in two steps (see Section 2.1) and are able to prove a general theorem (Theorem 1) and verify the conditions and the order of magnitude of the remainder in various special cases, namely, binomial, normal, exponential and Poisson. The worst possible orders for the remainders for the four distributions mentioned above are, respectively, $\max\{1/p, p(\log r)^3/\sqrt{r}\}$, $\max\{1/p, p(\log r)^{1+\gamma}/\sqrt{r}\}$ for some $0 < \gamma < 1$, $\max\{1/p, p/r^{1/3+\gamma}\}$ for some $0 < \gamma < \frac{1}{6}$, and $\max\{1/p, p(\log r)^{4+\gamma}/\sqrt{r}\}$

for some $0 < \gamma < 1$. Our result holds when $p \to \infty$ and $r \to \infty$ and the derivatives of the loglikelihood satisfy certain conditions near the boundaries of the natural parameter space. The specific assumptions change somewhat from example to example as illustrated in Section 3. The advantage of this approach is that it not only generalizes the approximations used in BGM, but explains why this kind of structure holds and where the different terms come from.

Sections 2–4 address the development of a two-stage Laplace approximation and a GBIC. They present details about the assumptions made, the application of the general result in specific examples and illustrations of the use of the results in the context of some model selection problems, and several inferential implications of our approximation method. Section 5 provides a numerical study of different approximations. Further heuristic generalizations are included in Section 6, which considers the case where the observations need not be distributed as an exponential family, and where the number of parameters also vary across different groups. At this level of generality, it is unlikely that the approximation can be rigorously justified. But we have included a limited numerical study that suggests that our approximations are quite good.

## 2. Generalization of BIC, assumptions, and proof of the main result

### 2.1. Notations and development of GBIC

Consider observations $y_{ij}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, r$, $n = pr$ with $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{ir})'$ having a joint density, given by

$$f(\mathbf{y}_i) = \exp\left\{\sum_{j=1}^{r} \boldsymbol{\psi}(y_{ij})'\boldsymbol{\theta}_i + r A(\boldsymbol{\theta}_i)\right\} \prod_{j=1}^{r} h(y_{ij}), \tag{2}$$

where $\boldsymbol{\theta}_i \in \Theta \subset \mathbb{R}^k$, $k \geqslant 1$, and $\Theta$ is the natural parameter space. For notational simplicity, we shall henceforth write $h(\mathbf{y}_i) = \prod_{j=1}^{r} h(y_{ij})$. In the above representation $\boldsymbol{\psi}(\cdot)$ is a vector-valued function, while $h(\cdot)$ and $A(\cdot)$ are scalar-valued functions. We are considering hypotheses $M_1 : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_p = \boldsymbol{\theta}_0 \in \mathbb{R}^k$, where $\boldsymbol{\theta}_0$ is known, vs. $M_2 : (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) \in \mathbb{R}^{pk}$.

Our approximation of the numerator of $\mathrm{BF}_{21}$ for the above testing problem, which we denoted by $m_2$, is presented below.

Let $\boldsymbol{t}_i = \sum_{j=1}^{r} \boldsymbol{\psi}(y_{ij})$, a $(k \times 1)$ vector. Consider a mixture of conjugate priors $\pi(\boldsymbol{\theta}_i | \boldsymbol{\alpha}, \beta)$ and assume the $\boldsymbol{\theta}_i$'s are i.i.d. given $(\boldsymbol{\alpha}, \beta)$ having a common conjugate prior $\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \beta) = C(\boldsymbol{\alpha}, \beta) \exp\{\boldsymbol{\alpha}'\boldsymbol{\theta} + \beta A(\boldsymbol{\theta})\}$, where $C(\boldsymbol{\alpha}, \beta)$ is a normalizing constant and $\pi_1(\boldsymbol{\alpha}, \beta)$ is the mixing prior on the hyperparameters. It is important to note that in the above prior $\boldsymbol{\alpha}$ is a vector and $\beta$ is a scalar. As pointed out by Ghosh and Samanta (2002), in high-dimensional problems, this is a common and natural choice of an objective prior. Note that $m_2$ is given by

$$m_2 = \left\{\prod_{i=1}^{p} h(\mathbf{y}_i)\right\} \int \cdots \int C^p(\boldsymbol{\alpha}, \beta) \left\{\prod_{i=1}^{p} \exp[(\boldsymbol{t}_i + \boldsymbol{\alpha})'\boldsymbol{\theta}_i + (r + \beta)A(\boldsymbol{\theta}_i)]\right\} \times \pi_1(\boldsymbol{\alpha}, \beta)\, \mathrm{d}\boldsymbol{\theta}_1 \cdots \mathrm{d}\boldsymbol{\theta}_p\, \mathrm{d}\boldsymbol{\alpha}\, \mathrm{d}\beta. \tag{3}$$

Let $\hat{\boldsymbol{\theta}}_i$ be the m.l.e. of the original parameters and $|-rA''(\hat{\boldsymbol{\theta}}_i)|$ be the determinant of the Hessian of the quantity $-rA(\boldsymbol{\theta})$, a $(k \times k)$ matrix, evaluated at $\hat{\boldsymbol{\theta}}_i$. Also let $|M(\hat{\boldsymbol{\alpha}}, \hat{\beta})|$ be the determinant of the Hessian of $-p \log C(\boldsymbol{\alpha}, \beta)$, a $(k+1 \times k+1)$ matrix, evaluated at $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$, where $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ will be defined through Eqs. (7) and (8) later in this section. Also denote the $(k \times 1)$ vector $\boldsymbol{t}_i/r$ by $\bar{\boldsymbol{t}}_i$. Then, by Laplace approximation, we have

$$
\begin{aligned}
\log \hat{m}_2 = {} & rp \left\{ \frac{1}{p} \sum_{i=1}^{p} (\bar{\boldsymbol{t}}_i' \hat{\boldsymbol{\theta}}_i + A(\hat{\boldsymbol{\theta}}_i)) \right\} + \left\{ \sum_{i=1}^{p} \log(h(\boldsymbol{y}_i)) \right\} \\
& - \frac{p}{2} \left\{ \frac{1}{p} \sum_{i=1}^{p} \log(|-rA''(\hat{\boldsymbol{\theta}}_i)|) \right\} \\
& + p \left\{ \frac{\hat{\boldsymbol{\alpha}}'}{p} \sum_{i=1}^{p} \hat{\boldsymbol{\theta}}_i + \frac{\hat{\beta}}{p} \sum_{i=1}^{p} A(\hat{\boldsymbol{\theta}}_i) + \log C(\hat{\boldsymbol{\alpha}}, \hat{\beta}) + \frac{k}{2} \log 2\pi \right\} \\
& - \frac{1}{2} \log(|M(\hat{\boldsymbol{\alpha}}, \hat{\beta})|) \\
& + \left\{ \frac{k+1}{2} \log 2\pi + \log \pi_1(\hat{\boldsymbol{\alpha}}, \hat{\beta}) \right\}. \qquad (4)
\end{aligned}
$$

We can write

$$
m_2 = \left\{ \prod_{i=1}^{p} h(\boldsymbol{y}_i) \right\} \int \int \left\{ \prod_{i=1}^{p} I_i(\boldsymbol{\alpha}, \beta) \right\} C^p(\boldsymbol{\alpha}, \beta) \pi_1(\boldsymbol{\alpha}, \beta) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\beta, \qquad (5)
$$

where $I_i(\boldsymbol{\alpha}, \beta) = \int \exp\{\boldsymbol{t}_i' \boldsymbol{\theta}_i + rA(\boldsymbol{\theta}_i)\} \exp\{\boldsymbol{\alpha}' \boldsymbol{\theta}_i + \beta A(\boldsymbol{\theta}_i)\} \, \mathrm{d}\boldsymbol{\theta}_i$. As $r \to \infty$ the integral $I_i(\boldsymbol{\alpha}, \beta)$, for each $i$, is approximated by

$$
\hat{I}_i(\boldsymbol{\alpha}, \beta) = \frac{\exp\{\boldsymbol{t}_i' \hat{\boldsymbol{\theta}}_i + rA(\hat{\boldsymbol{\theta}}_i)\} \exp\{\boldsymbol{\alpha}' \hat{\boldsymbol{\theta}}_i + \beta A(\hat{\boldsymbol{\theta}}_i)\}}{\sqrt{|-rA''(\hat{\boldsymbol{\theta}}_i)|}} (\sqrt{2\pi})^k \qquad (6)
$$

with

$$
A'(\hat{\boldsymbol{\theta}}_i) = -\frac{\boldsymbol{t}_i}{r}.
$$

We make a second-stage Laplace approximation to estimate $\{\prod_{i=1}^{p} h(\boldsymbol{y}_i)\} \int \int \{\prod_{i=1}^{p} \hat{I}_i(\boldsymbol{\alpha}, \beta)\} C^p(\boldsymbol{\alpha}, \beta) \pi_1(\boldsymbol{\alpha}, \beta) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\beta$, getting

$$
\left\{ \prod_{i=1}^{p} h(\boldsymbol{y}_i) \right\} \frac{\{\prod_{i=1}^{p} \hat{I}_i(\hat{\boldsymbol{\alpha}}, \hat{\beta})\} C^p(\hat{\boldsymbol{\alpha}}, \hat{\beta}) \pi_1(\hat{\boldsymbol{\alpha}}, \hat{\beta})}{\sqrt{|M(\hat{\boldsymbol{\alpha}}, \hat{\beta})|}} (\sqrt{2\pi})^{k+1}.
$$

The values of $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$ are obtained by solving the equations

$$\frac{\partial \log C(\boldsymbol{\alpha}, \beta)}{\partial \boldsymbol{\alpha}} = -\frac{1}{p} \sum_{i=1}^{p} \hat{\boldsymbol{\theta}}_i, \tag{7}$$

$$\frac{\partial \log C(\boldsymbol{\alpha}, \beta)}{\partial \beta} = -\frac{1}{p} \sum_{i=1}^{p} A(\hat{\boldsymbol{\theta}}_i). \tag{8}$$

So, putting the two Laplace approximations together, $\log m_2$ is approximated by (4). Note that the first two groups of terms in (4) arise from the maximized log-likelihood of the data and the first group is an $\mathrm{O}(pr)$ term. The third group arises from the first-stage Laplace approximations of the $p$ integrals $I_i(\boldsymbol{\alpha}, \beta)$ described above. It is $\mathrm{O}(p \log r)$ and corresponds to the penalty term in BIC. In general, the second group will be different from the penalty in BIC because of its dependence on $A''(\hat{\boldsymbol{\theta}}_i)$. The first three groups come entirely from the data and the model through the Laplace approximation. The fourth group collects the $\mathrm{O}(p)$ terms and they come from the prior, the data and the first and second-stage Laplace approximations. The first three terms in this group give the maximized value of the logarithm of the conjugate prior w.r.t. the hyperparameters, evaluated at the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_i$ of the parameters $\boldsymbol{\theta}_i$. The term $(pk/2) \log 2\pi$ arises from the $p$ first-stage $k$-dimensional Laplace approximations where $k$ is the number of components in $\boldsymbol{\theta}$. The fifth group comes from the second-stage Laplace approximation and is the logarithm of the determinant of $p$ times the information matrix from the first-stage prior and is of order $\log p$. These, namely the third, fourth and fifth groups of terms provide an adjustment to the BIC penalty. Some of its inferential implications are presented in Section 4. The sixth group has two terms, both $\mathrm{O}(1)$. The first term comes from the second-stage Laplace approximation and involves the dimension $k$ of $\boldsymbol{\alpha}$. The second term is a contribution from the second stage prior.

Finally, we propose GBIC in the general case as

$$\begin{aligned}
\mathrm{GBIC} &= \log \hat{m}_2 - C_1 \\
&= rp \left\{ \frac{1}{p} \sum_{i=1}^{p} (\vec{t}_i' \hat{\boldsymbol{\theta}}_i + A(\hat{\boldsymbol{\theta}}_i)) \right\} + \left\{ \sum_{i=1}^{p} \log(h(\mathbf{y}_i)) \right\} \\
&\quad - \frac{p}{2} \left\{ \frac{1}{p} \sum_{i=1}^{p} \log(|-rA''(\hat{\boldsymbol{\theta}}_i)|) \right\} \\
&\quad + p \left\{ \frac{\hat{\boldsymbol{\alpha}}'}{p} \sum_{i=1}^{p} \hat{\boldsymbol{\theta}}_i + \frac{\hat{\beta}}{p} \sum_{i=1}^{p} A(\hat{\boldsymbol{\theta}}_i) + \log C(\hat{\boldsymbol{\alpha}}, \hat{\beta}) + \frac{k}{2} \log 2\pi \right\} \\
&\quad - \frac{k+1}{2} \log p,
\end{aligned} \tag{9}$$

where $C_1 = [(k+1)/2] \log 2\pi + \log \pi_1(\hat{\boldsymbol{\alpha}}, \hat{\beta}) - \frac{1}{2} \log(M_1(\hat{\boldsymbol{\alpha}}, \hat{\beta}))$, where $M_1(\hat{\boldsymbol{\alpha}}, \hat{\beta}) = (1/p) M(\hat{\boldsymbol{\alpha}}, \hat{\beta})$.

For the problem in BGM, our results lead to a GBIC that is exactly the same as theirs but under different assumptions on $p$ and $r$. It can also be shown that with $r \to \infty$, our approximation of $\log m_2$ differs from that in BGM by $o(1)$.

Note that if one chooses to take $\boldsymbol{\alpha} \equiv 0$ for the conjugate prior (as in BGM), then the last term in GBIC above will become $-\log p/2$ and the first term in $C_1$ will become $\log 2\pi/2$, since the second-stage approximation will be done with respect to a one-dimensional argument, namely $\beta$.

### 2.2. The main result on Laplace approximation and the remainder

Before presenting the general theorem, we first list the assumptions needed for proving it. All our assumptions are verifiable from the given data (unlike the probabilistic assumptions in BGM). We give a rigorous proof for $k=1$. In principle, a similar line of argument should work for $k>1$. Let us first define $A_{1i}=\{\theta : |\theta-\hat{\theta}_i| \leqslant c \log r/\sqrt{r}\}$, for some suitably chosen constant $c>0$.

**Assumptions.** (A) For each $i=1,2,\ldots,p$, $\hat{\theta}_i$ falls in the interior of the natural parameter space $\Theta$, and $\hat{\theta}_i$'s are such that there is a constant $c>0$ such that $A_{1i}=\{\theta : |\theta-\hat{\theta}_i| \leqslant c \log r/\sqrt{r}\} \subset \Theta$ for all large $r$ for each $i=1,2,\ldots,p$.

(B) $(\log r/\sqrt{r})\sup_{\theta\in A_{1i}}|A'(\theta)| = o(1)$ uniformly in $i$, as $r \to \infty$.

(C.1) There exist positive constants $c_1 > c_2 > 0$, such that $c_1 A''(\hat{\theta}_i) < A''(\theta) < c_2 A''(\hat{\theta}_i)$ for all $\theta \in A_{1i}$ uniformly in $i$, as $r \to \infty$.

(C.2) $|-A''(\hat{\theta}_i)|$ is at most of the order of $r^{c_3(\log r)^{1-\delta}-1/2}$ where $c_3=c^2/2 \times c_2$, $c$ and $c_2$ being defined above, and is at least of the order of $1/(\log r)^{\delta}$ for some $0<\delta<1$, uniformly in $i$ as $r \to \infty$.

(D.1) There exist positive constants $c_5 > c_4 > 0$, such that $c_4|A'''(\hat{\theta}_i)| \leqslant \sup_{\theta\in A_{1i}}|A'''(\theta)| \leqslant c_5|A'''(\hat{\theta}_i)|$ as $r \to \infty$, uniformly in $i$.

(D.2) $[(\log r)^3/\sqrt{r}]|A'''(\hat{\theta}_i)| = o(1)$ uniformly in $i$ as $r \to \infty$.

(E) $\sup_{\theta\in A_{1i}} \pi(\theta) \leqslant c_6 \pi(\hat{\theta}_i)$ uniformly in $i, \alpha, \beta$ as $r \to \infty$ for some constant $c_6>0$ and $\pi(\hat{\theta}_i)$ is at least of the order of $1/r^{(\log r)^{1-\delta'}}$, for some $0<\delta<\delta'<1$, uniformly in $i, \alpha, \beta$ as $r \to \infty$, where $c_3$ and $\delta$ are as in (C.2).

(F) Support of the mixing prior is finite and $(\hat{\alpha}, \hat{\beta})$ falls in the interior of the support of the mixing prior. The derivatives of $\log c(\alpha, \beta)$ and $\pi_1(\alpha, \beta)$, up to the sixth and fourth orders, respectively, are bounded in a neighbourhood of $(\hat{\alpha}, \hat{\beta})$, and $\pi_1(\alpha, \beta)$ and the determinant of the Hessian of $-\log c(\alpha, \beta)$ are bounded away from zero at $(\hat{\alpha}, \hat{\beta})$.

(G) $\sum_{i=1}^{p} h_{ir} = o(1)$ uniformly in $(\alpha, \beta)$ as $p, r \to \infty$, where $h_{ir} = E_1 + E_2 + E_3 + E_4$ is obtained by replacing $\hat{\theta}$ by $\hat{\theta}_i$ and $A_1$ by $A_{1i}$ in the definitions of $E_1$, $E_2$, $E_3$ and $E_4$ in Eqs. (22)–(25) below.

For each $i=1,2,\ldots,p$, conditions (A) through (E) determine a certain range $S \subset \Theta$ within which $\hat{\theta}_i$ must lie. Our approximation goes through for those data sets for which $\{\hat{\theta}_i \in S,\ i=1,2,\ldots,p\}$ and also conditions (F) and (G) are satisfied. Condition (F) is in the spirit of Kass et al. (1990).

We first prove the theorem under these assumptions and then discuss what can be done if some of the assumptions are violated.

**Theorem 1.** *Suppose we have data* $\mathbf{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{ir}\}'$, $i = 1, 2, \ldots, p$, *with* $\mathbf{y}_i$ *having joint density as shown in* (1), *for* $i = 1, 2, \ldots, p$. *Further assume that conditions* (A) *through* (E) *above are satisfied for each* $\hat{\theta}_i$, $i = 1, 2, \ldots, p$ *and also conditions* (F) *and* (G) *hold. Then* $\log \hat{m}_2 - \log m_2 = \mathrm{o}(1)$.

We write the proof so as to indicate how the assumptions arise. A key implicit step in the proof is to allow $\hat{\theta}_i$ to get close to the boundaries of the natural parameter space $\Theta$ and then control the growth or decay of derivatives of the likelihood. Then some standard arguments for handling the remainder for the Laplace approximation to $m_2$ under a fixed true $\theta_0$ (see, for example, pp. 35–39 of Ghosh and Ramamoorti, 2002) break down. New appropriate assumptions and arguments have to be invoked to make the remainder negligible. The general assumptions lead to specific orders for the remainder in several examples, namely, normal, binomial, Poisson and exponential. In a sense these orders are the worst possible ones. One can also get more precise orders that depend on the values of the $\hat{\theta}_i$'s. A numerical illustration of this is given in Section 5.

As we are giving a proof for $k = 1$, we are dealing with scalars only. Lemma 1 is a key step in the proof of the theorem. We begin with its statement and proof.

**Lemma 1.** *Under conditions* (A) *through* (E),

$$I_i(\alpha, \beta) = \hat{I}_i(\alpha, \beta)(1 + \mathrm{O}(h_{ir})),$$

*where* $h_{ir} \to 0$ *uniformly in i as* $r \to \infty$, *and* $\hat{I}_i(\alpha, \beta)$ *has been defined in Eq.* (6) *in Section* 2.1.

**Proof.** We suppress the $i$ for notational convenience. For the same reason, we shall use notations $I$ and $\hat{I}$ instead of $I(\alpha, \beta)$, $\hat{I}(\alpha, \beta)$ throughout the proof of Lemma 1. Also, by an abuse of notation, we use $\pi(\theta)$ to denote $\exp\{\alpha\theta + \beta A(\theta)\}$ for this proof.

Here, we are trying to estimate the quantity $|I - \hat{I}| = |\int_\Theta \mathrm{e}^{L(\theta)}\pi(\theta)\,\mathrm{d}\theta - \mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta}) \int_{\mathbb{R}} \mathrm{e}^{(r/2)(\theta-\hat{\theta})^2 A''(\hat{\theta})}\,\mathrm{d}\theta|$, where $L(\theta) = \{t\theta + rA(\theta)\}$ and $\pi(\theta) = \exp\{\alpha\theta + \beta A(\theta)\}$. Define

$$f(\theta) = \begin{cases} \mathrm{e}^{L(\theta) - L(\hat{\theta})} & \text{if } \theta \in \Theta, \\ 0 & \text{if } \theta \in \Theta^c. \end{cases} \tag{10}$$

Then $|I - \hat{I}| = \mathrm{e}^{L(\hat{\theta})} |\int_{\mathbb{R}} \{f(\theta)\pi(\theta) - \pi(\hat{\theta})\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2 A''(\hat{\theta})}\}\,\mathrm{d}\theta|$.

Let $A_1 = \{\theta : |\theta - \hat{\theta}| \leqslant c \log r/\sqrt{r}\}$ and $A_2 = \{\theta : |\theta - \hat{\theta}| > \frac{c \log r}{\sqrt{r}}\}$ for some constant $c > 0$ chosen so that $A_1 \subset \Theta$, for large enough $r$. Then

$$|I - \hat{I}| \leqslant \mathrm{e}^{L(\hat{\theta})} \left\{ \int_{A_1} |f(\theta)\pi(\theta) - \pi(\hat{\theta})\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2 A''(\hat{\theta})}|\,\mathrm{d}\theta \right.$$
$$\left. + \int_{A_2} f(\theta)\pi(\theta)\,\mathrm{d}\theta + \pi(\hat{\theta}) \int_{A_2} \mathrm{e}^{(r/2)(\theta-\hat{\theta})^2 A''(\hat{\theta})}\,\mathrm{d}\theta \right\}. \tag{11}$$

We shall first look at $\mathrm{e}^{L(\hat{\theta})}\int_{A_2} f(\theta)\pi(\theta)\,\mathrm{d}\theta \leqslant \mathrm{e}^{L(\hat{\theta})}\{\sup_{\theta\in A_2} f(\theta)\}\int_{\theta\in A_2}\pi(\theta)\,\mathrm{d}\theta$. By the concavity of $L(\theta)$, we have

$$\begin{aligned}\sup_{\theta\in A_2} f(\theta) \leqslant \sup_{\theta\in A_2}\frac{\mathrm{e}^{L(\theta)}}{\mathrm{e}^{L(\hat{\theta})}} &\leqslant \max\left\{\frac{\mathrm{e}^{L(\hat{\theta}+c\log r/\sqrt{r})}}{\mathrm{e}^{L(\hat{\theta})}}, \frac{\mathrm{e}^{L(\hat{\theta}-c\log r/\sqrt{r})}}{\mathrm{e}^{L(\hat{\theta})}}\right\}\\ &= \max\{\mathrm{e}^{(c^2/2)(\log r)^2 A''(\theta_1')}, \mathrm{e}^{(c^2/2)(\log r)^2 A''(\theta_2')}\},\end{aligned}\tag{12}$$

where $\theta_1'$ is a point between $\hat{\theta}$ and $\hat{\theta}+c\log r/\sqrt{r}$ and $\theta_2'$ is a point between $\hat{\theta}-c\log r/\sqrt{r}$ and $\hat{\theta}$. We can rewrite the rightmost expression in the string of expressions in (12) as $\max\{r^{(c^2/2)\log r A''(\theta_1')}, r^{(c^2/2)\log r A''(\theta_2')}\}$.

Noting that we need to show that $\mathrm{e}^{L(\hat{\theta})}\int_{A_2} f(\theta)\pi(\theta)\,\mathrm{d}\theta = \mathrm{o}(\hat{I})$, we need $r^{c_3\log r A''(\hat{\theta})}\sqrt{|-rA''(\hat{\theta})|}/\pi(\hat{\theta}) \to 0$ as $r\to\infty$, where $c_3 = c^2/2\times c_2$, because we have assumed that $\hat{\theta}$ falls in such a range that condition (C.1) is satisfied. Now by assumption (C.2) on the order of $A''(\hat{\theta})$, we see that $r^{c_3\log r A''(\hat{\theta})}\sqrt{|-rA''(\hat{\theta})|}$ is at most of the order of $1/r^{(c_3/2)(\log r)^{1-\delta}-1/4}$.

So, using the fact that $\int_{\theta\in A_2}\pi(\theta)\,\mathrm{d}\theta\leqslant 1$, we have that

$$\frac{\mathrm{e}^{L(\hat{\theta})}\int_{\theta\in A_2} f(\theta)\pi(\theta)\,\mathrm{d}\theta}{\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})\sqrt{2\pi}/(\sqrt{|-rA''(\hat{\theta})|})}$$

is at most of the order of

$$\frac{1}{\left\{r^{\frac{c_3(\log r)^{1-\delta}}{2}-\frac{1}{4}}\right\}\pi(\hat{\theta})} = \mathrm{O}\left(\frac{1}{r^{\frac{c_3(\log r)^{1-\delta}}{4}-\frac{1}{4}}}\right) \quad \text{as } r\to\infty,$$

by assumption (E).

Let us now look at

$$\begin{aligned}&\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})\int_{\theta\in A_2}\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2 A''(\hat{\theta})}\,\mathrm{d}\theta\\ &\quad= \frac{2\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})}{\sqrt{|-rA''(\hat{\theta})|}}\left\{1-\Phi\left(\frac{c\log r}{\sqrt{r}}\sqrt{|-rA''(\hat{\theta})|}\right)\right\}\\ &\quad\leqslant \frac{2\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})}{\sqrt{|-rA''(\hat{\theta})|})}\,\frac{1}{c\log r\sqrt{|-A''(\hat{\theta})|}}\,\frac{1}{\sqrt{2\pi}\,r^{(c^2/2)\log r|A''(\hat{\theta})|}}.\end{aligned}\tag{13}$$

The last expression in (13) is at most of the order of

$$\frac{\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})}{\sqrt{|-rA''(\hat{\theta})|}}\,\frac{1}{(\log r)^{1-\frac{\delta}{2}}}\,\frac{1}{r^{(c^2/2)(\log r)^{1-\delta}}} = \mathrm{o}(\hat{I}),\tag{14}$$

by assumption (C.2).

Now we are going to estimate

$$\mathrm{e}^{L(\hat{\theta})}\left\{\int_{A_1}|f(\theta)\pi(\theta)-\pi(\hat{\theta})\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}|\,\mathrm{d}\theta\right\}.$$

The quantity above is less than

$$\mathrm{e}^{L(\hat{\theta})}\left\{\int_{A_1}|f(\theta)-\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}|\pi(\theta)\,\mathrm{d}\theta+\int_{A_1}|\pi(\hat{\theta})-\pi(\theta)|\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}\,\mathrm{d}\theta\right\}.$$

First note

$$\begin{aligned}
&\mathrm{e}^{L(\hat{\theta})}\int_{A_1}|f(\theta)-\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}|\pi(\theta)\,\mathrm{d}\theta\\
&\quad=\mathrm{e}^{L(\hat{\theta})}\int_{A_1}\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}|\mathrm{e}^{(r/6)(\theta-\hat{\theta})^3A'''(\theta_1)}-1|\pi(\theta)\,\mathrm{d}\theta\\
&\quad\leqslant\mathrm{e}^{L(\hat{\theta})}\left\{\sup_{\theta\in A_1}\pi(\theta)\right\}\sup_{\theta\in A_1}|\mathrm{e}^{R(\theta)}-1|\int_{A_1}\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}\,\mathrm{d}\theta,
\end{aligned}\tag{15}$$

where $\theta_1$ is a point between $\theta$ and $\hat{\theta}$, and

$$R(\theta)=\frac{r}{6}(\theta-\hat{\theta})^3A'''(\theta_1).\tag{16}$$

Note that

$$\int_{A_1}\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}\,\mathrm{d}\theta\leqslant\frac{\sqrt{2\pi}}{\sqrt{|-rA''(\hat{\theta})|}}.\tag{17}$$

Also $\sup_{\theta\in A_1}|\mathrm{e}^{R(\theta)}-1|\leqslant\{\sup_{\theta\in A_1}|R(\theta)|\}\times\{\sup_{\theta\in A_1}\mathrm{e}^{|R(\theta)|}\}$. Now by assumptions (D.1) and (D.2), we have

$$\sup_{\theta\in A_1}|R(\theta)|\to 0\quad\text{and}\quad\sup_{\theta\in A_1}\mathrm{e}^{|R(\theta)|}\to 1\quad\text{as } r\to\infty.$$

Again, by assumption (E), we see that the last expression in the string of equality/inequalities (15) is $\hat{I}\,\mathrm{o}(1)$, where the o(1) term is determined by the order of $[(\log r)^3/\sqrt{r}]|A'''(\hat{\theta})|$, by assumption (D.1).

Now look at

$$\begin{aligned}
&\mathrm{e}^{L(\hat{\theta})}\int_{A_1}|\pi(\hat{\theta})-\pi(\theta)|\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}\,\mathrm{d}\theta\\
&\quad=\mathrm{e}^{L(\hat{\theta})}\int_{A_1}|\theta-\hat{\theta}||\pi'(\theta_1^*)|\mathrm{e}^{(r/2)(\theta-\hat{\theta})^2A''(\hat{\theta})}\,\mathrm{d}\theta,
\end{aligned}\tag{18}$$

where $\theta_1^*$ is a point between $\theta$ and $\hat{\theta}$. The above quantity is less than

$$\mathrm{e}^{L(\hat{\theta})}\frac{c\log r}{\sqrt{r}}\times\frac{\sqrt{2\pi}}{\sqrt{|-rA''(\hat{\theta})|}}\times\sup_{\theta\in A_1}|\pi'(\theta)|.\tag{19}$$

Note that $\pi'(\theta) = \{\alpha + \beta A'(\theta)\}\pi(\theta)$, so we have

$$\sup_{\theta\in A_1} |\pi'(\theta)| \leqslant \left\{ |\alpha| + |\beta| \sup_{\theta\in A_1} |A'(\theta)| \right\} \times \sup_{\theta\in A_1} \pi(\theta). \tag{20}$$

The expression on the r.h.s. above is at most of the order of $\pi(\hat{\theta}) \times (1 + \sup_{\theta\in A_1} |A'(\theta)|)$ by assuming condition (E) and assuming $\alpha$ and $\beta$ fall inside a bounded range. By assumption (B), the quantity in (19) is $\mathrm{o}(\hat{I})$.

Finally combining all the different approximation rates, we get

$$I = \hat{I}(1 + \mathrm{O}(h_r)), \tag{21}$$

where $h_r$ is given by $h_r = E_1 + E_2 + E_3 + E_4$, where

$$E_1 = \frac{r^{c3 \log r A''(\hat{\theta})} \times \sqrt{|-rA''(\hat{\theta})|}}{\pi(\hat{\theta})}, \tag{22}$$

$$E_2 = \frac{r^{(c^2/2) \log r A''(\hat{\theta})}}{\log r \sqrt{|-A''(\hat{\theta})|}}, \tag{23}$$

$$E_3 = \frac{(\log r)^3}{\sqrt{r}} \times \sup_{\theta\in A_1} |A'''(\theta)| \times \mathrm{e}^{(\log r)^3/\sqrt{r} \times \sup_{\theta\in A_1} |A'''(\theta)|}, \tag{24}$$

$$E_4 = \frac{\log r}{\sqrt{r}} \times \left( 1 + \sup_{\theta\in A_1} |A'(\theta)| \right). \tag{25}$$

By assumptions (A) through (E), $h_r \to 0$ uniformly in $i$ as $r \to \infty$. So Lemma 1 is proved. □

We now complete the proof of the theorem.

**Proof of Theorem 1.** Using Lemma 1, for $i = 1, 2, \ldots, p$, we have

$$\prod_{i=1}^{p} \hat{I}_i(\alpha, \beta) = \left( \prod_{i=1}^{p} I_i(\alpha, \beta) \right) \times \left\{ \prod_{i=1}^{p} (1 + \mathrm{O}(h_{ir})) \right\}. \tag{26}$$

Then under assumption (G), and using also the fact that the $\mathrm{O}(h_{ir})$ terms above are in fact all $\mathrm{o}(1)$ uniformly in $i$, $\alpha$ and $\beta$ as $r \to \infty$, it is easy to verify that

$$\prod_{i=1}^{p} \hat{I}_i(\alpha, \beta) = \left\{ \prod_{i=1}^{p} I_i(\alpha, \beta) \right\} \times \left\{ 1 + \mathrm{O}\left( \sum_{i=1}^{p} h_{ir} \right) \right\} = \left\{ \prod_{i=1}^{p} I_i(\alpha, \beta) \right\} \times (1 + \mathrm{o}(1)), \tag{27}$$

as $p, r \to \infty$. Note that the o(1) term in (27) converges to 0 uniformly in $(\alpha, \beta)$ as $p, r \to \infty$. We will now look at the approximation of

$$\int\int \left\{\prod_{i=1}^{p} \hat{I}_i(\alpha, \beta)\right\} C^p(\alpha, \beta)\pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

$$= \frac{(\sqrt{2\pi})^{pk} \mathrm{e}^{\{t_i \sum_{i=1}^{p} \hat{\theta}_i + r\sum_{i=1}^{p} A(\hat{\theta}_i)\}}}{\prod_{i=1}^{p} \sqrt{|-rA''(\hat{\theta}_i)|}}$$

$$\times \int\int \mathrm{e}^{\{p \log C(\alpha, \beta) + \alpha \sum_{i=1}^{p} \hat{\theta}_i + \beta \sum_{i=1}^{p} A(\hat{\theta}_i)\}} \pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta. \tag{28}$$

We only need to approximate

$$I_{\text{second}} = \int\int \mathrm{e}^{\{p \log C(\alpha, \beta) + \alpha \sum_{i=1}^{p} \hat{\theta}_i + \beta \sum_{i=1}^{p} A(\hat{\theta}_i)\}} \pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta. \tag{29}$$

By assumption (F), the unique maximum $(\hat{\alpha}, \hat{\beta})$ of the quantity $p \log C(\alpha, \beta) + \alpha \sum_{i=1}^{p} \hat{\theta}_i + \beta \sum_{i=1}^{p} A(\hat{\theta}_i)$ falls in the interior of the support of $\pi_1(\alpha, \beta)$. Then using standard arguments for proving Laplace approximations for integrals of exponents of concave functions (noting that $p \log C(\alpha, \beta) + \alpha \sum_{i=1}^{p} \hat{\theta}_i + \beta \sum_{i=1}^{p} A(\hat{\theta}_i)$ is a concave function), we get

$$\hat{I}_{\text{second}} = \mathrm{e}^{\{p \log C(\hat{\alpha}, \hat{\beta}) + \hat{\alpha} \sum_{i=1}^{p} \hat{\theta}_i + \hat{\beta} \sum_{i=1}^{p} A(\hat{\theta}_i)\}} \times \frac{\pi_1(\hat{\alpha}, \hat{\beta})}{\sqrt{|M(\hat{\alpha}, \hat{\beta})|}} \times 2\pi$$

$$= I_{\text{second}} \left(1 + \mathrm{O}\left(\frac{1}{p}\right)\right), \tag{30}$$

as by assumption (F), the functions $\log C(\alpha, \beta)$ and $\pi_1(\alpha, \beta)$ are nice. Finally, combining all the error rates, the overall rate of approximation is given by

$$\hat{m}_2 = m_2 \left\{1 + \mathrm{O}\left(\sum_{i=1}^{p} h_{ir}\right)\right\} \left\{1 + \mathrm{O}\left(\frac{1}{p}\right)\right\}$$

$$= m_2(1 + \mathrm{o}(1)), \tag{31}$$

as $p \to \infty$ and $r \to \infty$, finishing the proof of Theorem 1. □

### 2.3. Discussion

As mentioned before, all assumptions can be checked from the data. Assumption (B) will hold unless the mean of any block, i.e. $\bar{t_i}$ for any $i$, is too large. For example, it is satisfied if $\bar{t_i}$ for each $i$ is, say, at most of the order of $r^{1/2-\delta}/\log r$, for some $0 < \delta < \frac{1}{2}$. Assumptions (C.2) and (D.2) will hold for practically all $\hat{\theta}_i$'s, unless some of them are extremely close to the boundary of the interior of $\Theta$ or extremely large in magnitude. Assumptions (C.1), (D.1) and (E) are basically assumptions on the continuity of the derivatives of the loglikelihood and the prior. Condition (E) requires that the derivatives of $\pi$ are not very large at $\hat{\theta}$. For usual priors, it sometimes means that $\hat{\theta}$ is not extremely close to the boundary of the interior of $\Theta$ and $\pi$ is not too sharply peaked at $\hat{\theta}$.

We crucially need assumptions (A) and (F) for our approximation to be valid. For the m.l.e. $\hat{\theta}_i$ of $\theta_i$ to exist and fall in the interior $\Theta^o$ of $\Theta$, it is sufficient that $t_i$ falls in $-r A'(\Theta^o)$. If the boundaries of $A'(\Theta^o)$ have zero mass under each $\theta$, this will happen with probability 1. All the above facts follow easily from Theorem 3.6, Theorem 5.5 and the discussion on p. 149 following Theorem 5.5 in Brown (1986).

We now study the existence of $(\hat{\alpha}, \hat{\beta})$. First observe that

$$\exp\left\{p \log C(\alpha,\beta) + \alpha \sum_{i=1}^{p} \theta_i + \beta \sum_{i=1}^{p} A(\theta_i)\right\}$$

is the joint density of $(\theta_1, \theta_2, \ldots, \theta_p)$ given $(\alpha, \beta)$, where $\theta_i \sim \pi(.|\alpha, \beta)$. If we were to solve the m.l.e. of $(\alpha, \beta)$ from the above density, by exactly the same argument as above for $\hat{\theta}_i$ and noting that the distribution of sufficient statistics $(\sum_{i=1}^{p} \theta_i, \sum_{i=1}^{p} A(\theta_i))'$ will be absolutely continuous for $p > 1$, there will be solution to the equations

$$\frac{\partial \log C(\alpha,\beta)}{\partial \alpha} = -\frac{1}{p} \sum_{i=1}^{p} \theta_i,$$

$$\frac{\partial \log C(\alpha,\beta)}{\partial \beta} = -\frac{1}{p} \sum_{i=1}^{p} A(\theta_i),$$

with probability 1 for any random sample of size p from $\pi(.|\alpha, \beta)$. But we have here $\hat{\theta}_i$'s and $A(\hat{\theta}_i)$'s instead of $\theta_i$'s and $A(\theta_i)$'s. However, by assumption $(A)$, all the $\hat{\theta}_i$'s are in the interior of $\Theta$ and so we can as well treat them as if they are a sample from $\pi(.|\alpha, \beta)$, and so the $(\hat{\alpha}, \hat{\beta})$ will exist with probability 1 in this case also.

The proof holds under the assumption that $\{\hat{\theta}_i \in S, i = 1, 2, \ldots, p\}$ and that $(\hat{\alpha}, \hat{\beta})$ falls in the interior of the support of $\pi_1(\alpha, \beta)$. If, for some data sets, one has $\hat{\theta}_i \in S^c$ for a few indices $i$, one can use numerical integration for those indices and do our method of approximation for the remaining indices. In specific examples like binomial or Poisson, one can use Stirling's approximation for those co-ordinates when one has $\hat{\theta}_i$'s falling on the boundary of the natural parameter space, e.g. if the number of successes in a block is either 0 or $r$, in the Bernoulli case. One potential problem with $(\hat{\alpha}, \hat{\beta})$ is that we are restricting $\pi_1(\alpha, \beta)$ to be within a bounded rectangle, not the entire natural parameter space. So, in some cases, $(\hat{\alpha}, \hat{\beta})$ will not be inside this rectangle. In those cases, one needs to use direct numerical calculations to evaluate the second-stage integral. We do not report calculations based on these alternative methods in the present paper.

One natural question might be, why do we use $A_1 = \{\theta : |\theta - \hat{\theta}| \leqslant c \log r/\sqrt{r}\}$ ? Other possible choices are $\sqrt{\log r}/\sqrt{r}$, $(\log r)^{1/2+\delta}/\sqrt{r}$ for any $\delta > 0$, and $r^d/\sqrt{r}$ where $d > \frac{1}{2}$.

Note that $\sqrt{\log r}/\sqrt{r}$ is chosen by Ghosh and Ramamoorti (2002, pp. 35–39, although there is a typo). This is the smallest set of this kind for which the complement has negligible probability under the approximating normal. A common choice is $r^d/\sqrt{r}$. This option is excluded since in this case the error terms will become extremely large.

The problem with the choice of $\sqrt{\log r}/\sqrt{r}$ is as follows. If one is using $\sqrt{\log r}/\sqrt{r}$, one has $\mathrm{e}^{L(\hat{\theta})}\int_{\theta\in A_2} f(\theta)\pi(\theta)\,\mathrm{d}\theta \leqslant \mathrm{e}^{L(\hat{\theta})}/r^{d|A''(\hat{\theta})|}$, for some constant $d>0$, using assumption (C.1). For our approximation to work, we need

$$\frac{\dfrac{\mathrm{e}^{L(\hat{\theta})}}{r^{d|A''(\hat{\theta})|}}}{\dfrac{\sqrt{2\pi}\mathrm{e}^{L(\hat{\theta})}\pi(\hat{\theta})}{\sqrt{|-rA''(\hat{\theta})|}}} \to 0 \quad \text{as } r\to\infty.$$

A careful inspection reveals that, in general, unless we allow $|A''(\hat{\theta})|$ extremely close to zero or have $|A''(\hat{\theta})|>d'$, for some positive constant $d'$, we need $\pi(\hat{\theta})$ to be large (at least having some polynomial rate of $r$) for the above to happen. Using $(\log r)^{1/2+\delta}/\sqrt{r}$, for any $\delta>0$, removes these difficulties, as we can make $\hat{\theta}$ go reasonably close to the boundary (without having to assume that it is bounded below by some constant) and also we do not need to assume anything unreasonable on the magnitude of $\pi(\hat{\theta})$. We choose $\delta=\frac{1}{2}$ to make our argument neater.

Finally, it is worthwhile to mention that in specific examples, it is possible to estimate the expected proportion of $i$'s (at least an upper bound of it) for which $\hat{\theta}_i\notin S$, under the assumption that $M_2$ is correct and $\theta_i \overset{\text{i.i.d.}}{\sim} \pi(.|\alpha,\beta)$, given $(\alpha,\beta)$ and $(\alpha,\beta)\sim\pi_1(.)$. An example is presented in the Appendix.

## 3. Discussion of the examples

### 3.1. Verification of assumptions and rates of convergence for specific distributions

(1) *Bernoulli distribution*: In this case, $A(\theta)=-\log(1+\mathrm{e}^{\theta})$ and $-\infty<\theta<\infty$. It is easy to see that $|A''(\theta)|\leqslant 1$ always. So, the first requirement in (C.2) is always satisfied. Noting that as $\hat{\theta}\to\infty$, $|A''(\hat{\theta})|$ behaves like $\mathrm{e}^{-\hat{\theta}}$ and as $\hat{\theta}\to-\infty$, $|A''(\hat{\theta})|$ behaves like $\mathrm{e}^{\hat{\theta}}$, we need $|\hat{\theta}|$ to be at most of the order of $\log(\log r)^{\delta}$ for some $0<\delta<1$, to meet the second requirement of assumption (C.2). Now noting that $|A'(\theta)|\leqslant 1$ and $A'''(\theta)=A''(\theta)\{2A'(\theta)+1\}$, we have $[(\log r)^3/\sqrt{r}]\sup_{\theta\in A_1}|A'''(\theta)| \to 0$ as $r\to\infty$. So we do not really need assumption (D.1) in this case. Condition (B) is also trivially satisfied. For condition (C.1), we need to look at

$$\frac{A''(\theta)}{A''(\hat{\theta})}=\mathrm{e}^{\delta}\left(\frac{(1+\mathrm{e}^{\hat{\theta}})}{(1+\mathrm{e}^{\hat{\theta}+\delta})}\right)^2>\frac{1}{2} \tag{32}$$

for all large values of $r$ and all $\theta \in A_1$, where $\delta = \theta - \hat{\theta}$. A similar inequality will hold for the quantity $A''(\hat{\theta})/A''(\theta)$, hence (C.1) is satisfied.

Now let us look at condition (E). Consider

$$\frac{\pi(\theta)}{\pi(\hat{\theta})} = (\mathrm{e}^{\theta-\hat{\theta}})^{\alpha}\left(\frac{1+\mathrm{e}^{\hat{\theta}}}{1+\mathrm{e}^{\theta}}\right)^{\beta}. \tag{33}$$

Note that by making $r$ large enough,

$$\sup_{\theta\in A_1}\frac{\pi(\theta)}{\pi(\hat{\theta})} \leqslant \sup_{\theta\in A_1}(\mathrm{e}^{\theta-\hat{\theta}})^{\alpha}\sup_{\theta\in A_1}\left(\frac{1+\mathrm{e}^{\hat{\theta}}}{1+\mathrm{e}^{\theta}}\right)^{\beta} \leqslant 2 \tag{34}$$

for any bounded range of $\alpha$ and $\beta$. We can also verify that the other requirement of (E) is also satisfied by all $|\hat{\theta}|$'s that are at most of the order of $\log(\log r)^{\delta}$.

We see that all the conditions for proving Theorem 1 are satisfied for all $|\hat{\theta}_i|$'s that are at most of the order of $\log(\log r)^{\delta}$ for some $0<\delta<1$, as $r \to \infty$ uniformly in $i$, provided $(\hat{\alpha}, \hat{\beta})$ satisfies (F).

We shall now look at the worst error rates achieved for the first-stage Laplace approximations. Quantity $E_1$ is at most of the order of

$$\frac{1}{\{r^{\frac{c_3(\log r)^{1-\delta}}{2}-\frac{1}{4}}\}\pi(\hat{\theta}_i)} = \mathrm{O}\left(\frac{1}{r^{\frac{c_3(\log r)^{1-\delta}}{4}-\frac{1}{4}}}\right) \tag{35}$$

by condition (E). Quantity $E_2$ is at most of the order of $\frac{1}{r^{(c^2/2)(\log r)^{1-\delta}}(\log r)^{1-\delta/2}}$. The quantity $E_3 \leqslant [(\log r)^3/\sqrt{r}]\mathrm{e}^{(\log r)^3/\sqrt{r}} \sim (\log r)^3/\sqrt{r}$ for large $r$. Quantity $E_4$ is at most of the order of $\log r/\sqrt{r}$ for a bounded range of values of $(\alpha, \beta)$. So, $E_1+E_2+E_3+E_4 = \mathrm{O}((\log r)^3/\sqrt{r})$ as $r \to \infty$ for each $i = 1, \ldots, p$. Hence

$$\prod_{i=1}^{p}\hat{I}_i(\alpha,\beta) = \left\{\prod_{i=1}^{p} I_i(\alpha,\beta)\right\}\left(1+\mathrm{O}\left(\frac{p(\log r)^3}{\sqrt{r}}\right)\right) \tag{36}$$

as $r \to \infty$, $p \to \infty$, assuming $p(\log r)^3/\sqrt{r} \to 0$ as $r, p \to \infty$. Now the error rate for the second-stage Laplace approximation is $\mathrm{O}(1/p)$. So, the overall error rate for the two-stage approximation will be obtained from the fact that

$$\begin{aligned}\hat{m}_2 &= m_2\left(1+\mathrm{O}\left(\frac{p(\log r)^3}{\sqrt{r}}\right)\right)\left(1+\mathrm{O}\left(\frac{1}{p}\right)\right)\\ &= m_2(1+\mathrm{O}(J_{r,p})),\end{aligned} \tag{37}$$

where $J_{r,p} = \max\{\frac{1}{p}, \frac{p(\log r)^3}{\sqrt{r}}\}$.

(2) *Normal distribution*: As in BGM, we consider the case where the variance is known to be equal to 1 and there is a prior on the mean parameter. In this case, $A(\theta) = -\theta^2/2$,

$A''(\theta) = -1$, and $A'''(\theta) = 0$ while $-\infty < \theta < \infty$. Thus, conditions (C.1), (C.2), (D.1) and (D.2) are trivially satisfied. For condition (B), we need

$$\frac{\log r}{\sqrt{r}}\left(\hat{\theta} + c\frac{\log r}{\sqrt{r}}\right) \to 0 \quad \text{as } r \to \infty \text{ if } \hat{\theta} > 0$$

and

$$\frac{\log r}{\sqrt{r}}\left(\hat{\theta} - c\frac{\log r}{\sqrt{r}}\right) \to 0 \quad \text{as } r \to \infty \text{ if } \hat{\theta} < 0.$$

So, if we have $|\hat{\theta}|$ at most of the order of $\frac{r^{1/2-\gamma'}}{\log r}$ for some $0 < \gamma' < \frac{1}{2}$, that will be sufficient for the above to hold. But for condition (E) to hold, we need that $|\hat{\theta}|$ be at most of the order of $(\log r)^{(2-\delta')/2}$ for some $0 < \delta' < 1$, noting that (C.2) is now satisfied with $\delta = 0$. Also (C.1) is satisfied with $c_2 = 1$. So we have, in this case,

$$E_1 = \frac{\sqrt{r}}{r^{c^2 \log r/2}\pi(\hat{\theta})} = \mathrm{O}\left(\frac{1}{r^{c^2(\log r)/4-\frac{1}{4}}}\right), \quad E_2 = \frac{1}{r^{(c^2/2)\log r}\log r}, \quad E_3 = 0$$

and $E_4 = \mathrm{O}((\log r)^{1+\gamma}/\sqrt{r})$ where $\gamma = (2-\delta')/2$.

Therefore, the error in approximation for the first-stage Laplace approximation is $\mathrm{O}(p(\log r)^{1+\gamma}/\sqrt{r})$, provided $p(\log r)^{1+\gamma}/\sqrt{r} \to 0$, and the overall rate in the Normal case is $\mathrm{O}(J_{r,p})$, where

$$J_{r,p} = \max\left\{\frac{1}{p},\, p\frac{(\log r)^{1+\gamma}}{\sqrt{r}}\right\}.$$

In the above proof, we have assumed $\alpha$ and $\beta$ are inside a bounded range. In the special case of the normal distribution this assumption can be relaxed to include the Zellner-Siow prior (1980). Nonetheless, our general method is less efficient for the normal case than in BGM, as we are using approximations in both the stages, while the first-stage integral is evaluated directly for the normal case in BGM.

(3) *Exponential distribution*: In this case $A(\theta) = \log(-\theta)$ and $\theta < 0$. Choosing the proper range of $\hat{\theta}_i$'s so that the assumptions for Theorem 1 are satisfied, the worst possible error rate is $\mathrm{O}(J_{r,p})$, where $J_{r,p} = \max\{1/p,\, p/r^{1/3+\gamma}\}$, for some $0 < \gamma < \frac{1}{6}$, assuming $p/r^{1/3+\gamma} \to 0$ as $r,\ \ p \to \infty$.

(4) *Poisson distribution*: In this case $A(\theta) = -\mathrm{e}^{\theta}$ and $-\infty < \theta < \infty$. The worst possible rate in this is $\mathrm{O}(J_{r,p})$, where $J_{r,p} = \max\{1/p,\, p(\log r)^{4+\gamma}/\sqrt{r}\}$, for some $1 - \delta < \gamma < 1$ provided $p(\log r)^{4+\gamma}/\sqrt{r} \to 0$ as $r,\ \ p \to \infty$. Here $0 < \delta < 1$ is chosen to satisfy condition (C.2).

### 3.2. Illustration of application of results in some model selection problems

(a) *Multiple regression problem*: Consider the problem of selecting a model from among several multiple regression models. Typically in such studies, the object of interest is the dependence of an observed variable on several explanatory variables through multiple regression models. Often the study would be undertaken in closely related populations in similar geographical areas, like the states of the US or the countries of Europe or smaller such local regions. For example, one might want to study the dependence of the price of a commodity on a fixed number of possible economic parameters like average income in the populations, demand, production etc. Let us consider, for population $i$, the regression model, expressed in a Linear Model as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}^i + \epsilon_i, \tag{38}$$

where $\boldsymbol{\beta}^i = (\beta_1^i, \beta_2^i, \ldots, \beta_K^i)'$, $\boldsymbol{X}_i$ is an $N \times K$ matrix such that the columns of $X_i$ are orthogonal to each other and are of length 1, $(N \geqslant K)$ and $\epsilon_i \sim \mathrm{N}(\boldsymbol{0}, I_{N \times N})$. For the $i$th population, we have $r$ observation vectors; denoted $\boldsymbol{Y}_{i,1}, \boldsymbol{Y}_{i,2}, \ldots, \boldsymbol{Y}_{i,r}$, where given $i$, $\boldsymbol{Y}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(X_i \boldsymbol{\beta}^i, I_{N \times N})$ for $j = 1, 2, \ldots, r$. We consider $p$ such populations and for each population we have $r$ observation vectors. We can express the whole thing in a big linear model/multiple regression setup as

$$\boldsymbol{Y} = \mathbf{Z}\boldsymbol{\beta} + \epsilon, \tag{39}$$

where $\boldsymbol{Y} = \{(\boldsymbol{Y}_{1,1})', (\boldsymbol{Y}_{1,2})', \ldots, (\boldsymbol{Y}_{1,r})', \ldots, (\boldsymbol{Y}_{p,1})', (\boldsymbol{Y}_{p,2})', \ldots, (\boldsymbol{Y}_{p,r})'\}'$. $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = \{(\boldsymbol{\beta}^i)', (\boldsymbol{\beta}^2)', \ldots, (\boldsymbol{\beta}^p)'\}'$ and Z is a matrix of block-diagonal structure, given by

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_r \otimes \boldsymbol{X}_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_r \otimes \boldsymbol{X}_2 & \mathbf{0} & \ldots & \mathbf{0} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}_r \otimes \boldsymbol{X}_p \end{pmatrix}.$$

So we can also write

$$\boldsymbol{Y} = \sum_{i=1}^{p} \mathbf{Z}_i \boldsymbol{\beta}^i + \epsilon, \tag{40}$$

where each $\mathbf{Z}_i$ is an $(prN \times K)$ matrix. Notice that in this formulation, each element of $E(\boldsymbol{Y})$ is expressed as a linear combination of $pK$ parameters, although the design matrix $\mathbf{Z}$ is such that only $K$ of these will have possibly nonzero multipliers. We will consider model selection in this setup.

As the populations are similar, we can assume that $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^p$ are exchangeable. Let $\hat{\boldsymbol{\beta}}_j^i$ be the sample regression coefficients based on the $j$th data vector in the $i$th group. (Note each $\hat{\boldsymbol{\beta}}_j^i$ is a $K$-dimensional vector.) Then, for a given $i$, $\hat{\boldsymbol{\beta}}_j^i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\boldsymbol{\beta}^i, (X_i'X_i)^{-1} = I_{K \times K})$; for $j = 1, 2, \ldots, r$. We will work with $\hat{\boldsymbol{\beta}} = \{(\hat{\boldsymbol{\beta}}^1)', (\hat{\boldsymbol{\beta}}^2)', \ldots, (\hat{\boldsymbol{\beta}}^p)'\}'$ where $\hat{\boldsymbol{\beta}}^i = \{(\hat{\boldsymbol{\beta}}_1{}^i)', \ldots, (\hat{\boldsymbol{\beta}}_r{}^i)'\}'$.

The competing models we consider here correspond to which variables are included in the regression (in Eq. (40)). So, we consider a nested sequence of models $M_1 \subset M_2 \subset \cdots \subset M_K$; where under $M_l$, $\{\beta_h^i = 0$ for $h = l+1, \ldots, K$ and for each $i = 1, 2, \ldots, p\}$. This is a reasonable hypothesis to consider since the populations are expected to be similar. Note that under $M_l$, each vector $\hat{\boldsymbol{\beta}}_j^i$ will have a mean vector such that the last $(K - l)$ elements are zero.

Let $\pi(M_l) = 1/K$, $1 \leqslant K \leqslant l$, i.e. we assume that each model is a priori equally likely. Then a Bayesian selects a model $l^*$ with highest posterior probability $\pi(M_l^*|\hat{\boldsymbol{\beta}}) = \max_{1 \leqslant l \leqslant K} \pi(M_l|\hat{\boldsymbol{\beta}})$; or equivalently the model $l^*$ such that the integrated likelihood $m(l^*) = \max_{1 \leqslant l \leqslant K} m(l)$. Since $m(l)$ cannot be calculated exactly in general, this can be approximated by the two-stage Laplace approximation method and GBIC described in Section 2. Then the model $l^{\max}$ for which one has $\text{GBIC}(l^{\max}) = \max_l(\text{GBIC}(l))$ or $\log \hat{m}(l^{\max}) = \max_l\{\log \hat{m}(l)\}$ should be selected; where $\hat{m}(l)$ and $\text{GBIC}(l)$ have obvious meanings.

Towards that, under each model $M_l$, $1 \leqslant l \leqslant K$, we assume a prior distribution on $\boldsymbol{\beta}$ which gives point mass 1 at 0 for each coordinate of $\boldsymbol{\beta}$ which becomes zero under the model and for the remaining $lp$ coordinates, denoted by $\boldsymbol{\beta}_l$, assumes a conjugate mixture prior of the form

$$\pi_l(\boldsymbol{\beta}_l) = \int_{\boldsymbol{\alpha}_l, \gamma} \int \prod_{i=1}^{p} \{\exp[\boldsymbol{\alpha}_l' \boldsymbol{\beta}_l^i + \gamma A(\boldsymbol{\beta}_l^i)]\} C_l^p(\boldsymbol{\alpha}_l, \gamma) \pi_{1l}(\boldsymbol{\alpha}_l, \gamma)\, \mathrm{d}\boldsymbol{\alpha}_l\, \mathrm{d}\gamma, \tag{41}$$

where $(\boldsymbol{\alpha}_l, \gamma)$ denotes the $l + 1$ hyperparameters of the conjugate prior. Note that, under $M_l$, the prior has support on an $lp$ dimensional subspace of $\mathbb{R}^{pK}$. The exact form of $\log \hat{m}(l)$ and $\text{GBIC}(l)$ can now be obtained using Eqs. (4) and (9) in Section 2. For example, we have

$$\begin{aligned}\text{GBIC}(l) = & -\frac{lp}{2} \log r + \hat{\boldsymbol{\alpha}}_l' \sum_{i=1}^{p} \hat{\boldsymbol{\beta}}_l{}^i + \hat{\gamma} \sum_{i=1}^{p} A(\hat{\boldsymbol{\beta}}_l{}^i) + p \log C_l(\hat{\boldsymbol{\alpha}}_l, \hat{\gamma}) \\ & + \frac{pl}{2} \log 2\pi - \frac{l+1}{2} \log p - \frac{prK}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{p} \hat{\boldsymbol{\beta}}_{l+}^{i}{}' \hat{\boldsymbol{\beta}}_{l+}^{i},\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{l+}^{i}$ is the vector consisting of the last $K - l$ coordinates of $\hat{\boldsymbol{\beta}}_j^i$ for each $j = 1, \ldots, r$, hence is a vector of dimension $(K - l) \times r$.

(b) *Application of variable selection in a multivariate binary data model*: Suppose one has several groups of observations, each group having $r$ observation vectors. Each observation vector is composed of the values of $p$ random variables taking values 0 and 1 only. This kind of multivariate Binary Data has been considered, for example, by Wilbur et al. (2002), where, corresponding to each of 4 agronomic treatments applied to a certain soil (two of which are known to increase yield, but the biological reasons are not fully understood), one has $n_i$ observation vectors, $i = 1, 2, 3, 4$ where each vector indicates the presence or absence of $p$ micro-organisms in a soil sample from the treatment group. The basic goal of that study was to identify those micro-organisms which help differentiate between treatments, and so may have biological significance. Statistically, this is a problem of variable selection.

For simplicity, let us consider the case where one has only 2 groups of observations. Let $X_{jk}^i$ be the $k$th component of the $j$th observation vector in the $i$th group and let $\{P(X_{jk}^i = 1) = \theta_{ik}, k = 1, \ldots, p\}$ be the probabilities for $j = 1, 2, \ldots, r$. One way of knowing which variables help differentiate between the groups is to test hypotheses of the form

$$\mathrm{H}_{k_1, k_2, \ldots, k_l} : \{\theta_{1k_1} = \theta_{2k_1}, \theta_{1k_2} = \theta_{2k_2}, \ldots, \theta_{1k_l} = \theta_{2k_l}\} \tag{42}$$

for $1 \leqslant k_1 < k_2 < \cdots < k_l \leqslant p$ and $1 \leqslant l \leqslant p$. For example, if $\mathrm{H}_{1,2}$ is indeed true, then the probabilities for the first and second variables are invariant under both the treatments and hence these variables are not important in discriminating between the treatments.

For treatment 1, we have $r$ observation vectors, $\boldsymbol{X}_j^1$, $j = 1, 2, \ldots, r$, where $\boldsymbol{X}_j^1 = (X_{j1}^1, X_{j2}^1, \ldots, X_{jp}^1)'$. Similarly for treatment 2, we have $r$ observation vectors, $\boldsymbol{X}_j^2$, $j = 1, 2, \ldots, r$. Note that, $X_{1k}^1, X_{2k}^1, \ldots, X_{rk}^1 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta_{1k})$ for $1 \leqslant k \leqslant p$. Similarly $X_{1k}^2, X_{2k}^2, \ldots, X_{rk}^2 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta_{2k})$, for $1 \leqslant k \leqslant p$. Expressed as an exponential family, we can say, for a given $i$ and $k$, the p.m.f. of $(X_{1k}^i, X_{2k}^i, \ldots, X_{rk}^i)'$ is $f(t_{ik}|\eta_{ik})$, where

$$f(t_{ik}|\eta_{ik}) = \exp[t_{ik}\eta_{ik} - r\log(1 + \mathrm{e}^{\eta_{ik}})]. \tag{43}$$

In the above, $-\infty < \eta_{ik} < \infty$ and $t_{ik} = \sum_{j=1}^r X_{jk}^i$. Let $\boldsymbol{X}_{ik} = (X_{1k}^i, \ldots, X_{rk}^i)'$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})'$ for $i = 1, 2$ and $k = 1, 2, \ldots, p$. Then the likelihood of the whole data can be written as

$$f_1(\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1p}|\boldsymbol{\eta}_1) f_2(\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2p}|\boldsymbol{\eta}_2),$$

where $f_i(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{ip}|\boldsymbol{\eta}_i) = \prod_{k=1}^p f(t_{ik}|\eta_{ik})$.

The prior $\pi(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \pi^1(\boldsymbol{\eta}_1)\pi^2(\boldsymbol{\eta}_2)$ is chosen as follows. We first find the average of the estimated values $\frac{1}{2}\sum_{i=1}^2 \hat{\theta}_{ik}$, for each $k = 1, 2, \ldots, p$, where $\hat{\theta}_{ik} = (1/r)\sum_{j=1}^r X_{jk}^i$. For those $k$'s for which this estimate is between 0.2 and 0.8; we will assume exchangeability of the corresponding $\eta_{ik}$'s for a given treatment $i$. Similarly we assume exchangeability of $\eta_{ik}$'s for those $k$'s having the above estimate less than 0.2 and also for those $\eta_{ik}$'s having the estimate greater than 0.8. In the example from biology, this will mean that the microorganisms with similar probability of appearing in the soil are similar and so may be treated as exchangeable. The middle group has an interesting property, the model variance is stable at about 0.2 as pointed out by Cox (1970). It is worthwhile noting that the same parameters will be considered exchangeable under each treatment, although the priors for the parameters under the two treatments may be different. To some extent, the prior is data dependent in that the variables are grouped on the basis of data.

Under the model with unrestricted parameters, i.e. $l = 0$ in (42), each prior $\pi^i(\boldsymbol{\eta}_i)$ will be a product of 3 parts, each part putting an exchangeable conjugate mixture prior for the parameters inside that group. If $l \geqslant 1$, then the prior under that model will be a product of at least four parts. The coordinates which do not appear in the hypothesis, will be treated as before in terms of exchangeability. The coordinates which appear in the hypothesis will have a common parameter value for each treatment. These (unknown) common values will be given an exchangeable prior if the coordinates belong to the same group (as described before), otherwise coordinates will be separated according to which of the three groups they belong, and they will be given separate exchangeable priors. If the number of coordinates involved in a hypothesis (or the number of components among them which fall inside a particular group) is not large enough, then we may have to resort to direct integration on some or all such parameters. It is worth mentioning here that in this way, under a given hypothesis, observations corresponding to both treatments for the same variable will be treated together, if that variable is included in the hypothesis.

The method of two-stage Laplace approximation and GBIC can be applied here in exactly similar way as in the previous example. For each model/hypothesis considered in (42) (assumed equally likely a priori), we will approximate the corresponding integrated likelihood using our results in Section 2 and the model/hypothesis for which the criterion based on either of them is maximized will be chosen as the correct model/hypothesis and the variables for which the probabilities are the same for both treatments in the corresponding model/hypothesis will be discarded in the further analysis of the problem.

## 4. Inferential implications

This section presents the connection between our approximation method and posterior normality, stochastic complexity and model complexity. Finally, we comment on the penalty of GBIC vis-à-vis that of BIC.

A careful inspection reveals that our approximation to $\log m_2$ also shows that the posterior distribution of the $\theta_i$'s given $t_i$'s is approximately normal with mean vector $(\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ and a diagonal variance–covariance matrix with diagonal entries $\{1/\sqrt{|-rA''(\hat{\theta}_i)|},\ i = 1, \ldots, p\}$. It can be used to calculate the Bayesian measure of model complexity

$$p_D(y, \Theta, \tilde{\theta}(y)) = E_{\theta|y}[-2\log p(y|\theta)] + 2\log p(y|\tilde{\theta}(y)),$$

introduced by Spiegelhalter et al. (2002, Eq. (9)) for high-dimensional $p$. According to the notation used in that paper, $y$ is the data, $\Theta$ the parameter space and $\tilde{\theta}(y)$ is some estimate of $\theta$. As shown by these authors in their Section 3.2, if posterior normality holds (as it holds here approximately), the model complexity is approximately $p$, using $\tilde{\theta}(y)$ as the posterior mean. This in turn implies their interesting, new model selection rule (Eq. (37) in their paper) is equivalent to AIC even in high-dimensional problems.

A related notion is stochastic complexity, as formulated by Rissanen (1987, Eq. (2.1)). In our problem, it reduces to

$$I(x) = -\log[\tfrac{1}{2} f(x|M_1) + \tfrac{1}{2} f(x|M_2)], \tag{44}$$

where $x$ is the data and $f(x|M_i)$ is the marginal density under $M_i$, $i = 1, 2$. Conditionally on $M_2$, the stochastic complexity becomes

$$I(x|M_2) = -\log f(x|M_2). \tag{45}$$

If $x$ comes from $M_2$, then one expects $I(x)$ to be well approximated by $I(x|M_2)$. In any case, GBIC is an approximation to $I(x|M_2)$ (without the negative sign).

Thus, both AIC and GBIC seem to play an important role, at least as approximations to interesting inferential quantities.

We shall now try to give an interpretation of the difference between BIC and GBIC. Let us look at the Normal case first, in which

$$\text{GBIC} = \text{BIC} - \frac{p}{2}\log\left(\frac{1}{p}\bar{y}'\bar{y}\right) - \frac{p}{2} - \frac{\log p}{2}. \tag{46}$$

Note that GBIC is larger than BIC whenever

$$\frac{1}{p}\bar{y}'\bar{y} < \frac{1}{\exp[1 + \log p/p]},$$

which is more likely to happen when $M_1$ is true. Similarly, GBIC is smaller than BIC whenever $(1/p)\bar{y}'\bar{y}$ is bigger than the threshold shown above, which will happen more often under $M_2$. If $\sigma^2 \neq 1$, the threshold will depend on $\sigma^2$. So, as we can see, the GBIC has an adaptive penalty (dependent on the data) which penalizes the maximized log-likelihood in a more conservative way than the BIC in sense that it penalizes more when the maximized log-likelihood is large (which corresponds to large $(1/p)\bar{y}'\bar{y}$) and vice-versa.

In general, we can write

$$\begin{aligned}\mathrm{GBIC} = \mathrm{BIC} &+ \sum_{i=1}^{p} \log\{\pi(\hat{\theta}_i|\hat{\alpha}, \hat{\beta})\} - \frac{1}{2}\sum_{i=1}^{p} \log(|A''(\hat{\theta}_i)|) \\ &+ \frac{pk}{2}\log 2\pi - \frac{k+1}{2}\log p. \end{aligned} \quad (47)$$

Observe that for the Bernoulli case, $\frac{1}{2}\sum_{i=1}^{p}\log(|A''(\hat{\theta}_i)|) - p\log\pi$ happens to be the sum of the logarithm of the conjugate prior $\pi(\theta_i|\alpha, \beta)$ corresponding to $\alpha = \frac{1}{2}$, $\beta = 1$, and $\theta_i = \hat{\theta}_i$ for $i = 1, \ldots, p$. Thus, the first term after BIC in the r.h.s. of Eq. (48) will always dominate this quantity. Hence, GBIC will be bigger than BIC for the Bernoulli case if the difference between the two quantities mentioned above is bigger than $p\log\pi + \log p - (p/2)\log 2\pi$, and smaller than BIC otherwise.

The last two terms in Eq. (48) always add up to something positive, so they reduce the high penalty of BIC to some extent. On the other hand, the first two terms after BIC may be written as

$$\sum_{i=1}^{p} \log \frac{\pi(\hat{\theta}_i|\hat{\alpha}, \hat{\beta})}{\pi_J(\hat{\theta}_i|\hat{\alpha}, \hat{\beta})},$$

where $\pi_J$ is the usual Jeffreys prior. As Stein's example shows, the Jeffreys prior is not suitable for high-dimensional problems. A hierarchical prior of the kind described in this paper is more appropriate (as indicated by the frequent use of the Zellner–Siow (1980) prior for normal problems). It tends to zero faster than the Jeffreys prior at the tails. The above quantity may be interpreted as the information in the hierarchical prior relative to the Jeffreys prior. This difference has to be interpreted as the evidence in favor of the alternative. If this difference is bigger than the remaining terms in the adjustment of BIC, we adjust BIC upwards in favour of the more complex model. Hopefully, our representation of GBIC will ultimately be useful in better understanding what constitutes a good objective prior in high-dimensional problems.

## 5. Numerical study

This section presents the results of our simulation study for Bernoulli, normal, exponential and Poisson distributions. As mentioned in the introduction, we think that $-(p/2)\log(r)$

Table 1

| $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | BIC | GBIC | $\log m_1$ |
|---|---|---|---|---|---|---|
| 10 | 1000 | −4874.098 | −4874.220 | −4872.829 | −4873.095 | −6931.472 |
| 25 | 200 | −3220.397 | −3220.378 | −3191.230 | −3220.858 | −3465.736 |
| 25 | 500 | −6629.866 | −6629.968 | −6584.883 | −6628.231 | −8664.340 |
| 50 | 200 | −4018.026 | −4018.072 | −4052.757 | −4018.587 | −6931.472 |

Table 2

| $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | BIC | GBIC | $\log m_1$ | $c_p$ |
|---|---|---|---|---|---|---|---|
| 10 | 1000 | −14238.110 | −14238.088 | −14226.334 | −14237.735 | −28477.869 | 2.857 |
| 25 | 200 | −7302.068 | −7302.061 | −7280.659 | −7301.840 | −11616.155 | 1.761 |
| 25 | 500 | −17756.563 | −17756.557 | −17720.352 | −17755.954 | −52524.523 | 5.581 |
| 50 | 200 | −14261.908 | −14261.909 | −14247.653 | −14261.676 | −17095.678 | 0.596 |

Table 3

| $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | BIC | GBIC | $\log m_1$ |
|---|---|---|---|---|---|---|
| 10 | 1000 | −45248.556 | −45248.472 | −45228.558 | −45245.564 | $-2.742 \times 10^7$ |
| 25 | 200 | −9323.693 | −9323.817 | −9319.807 | −9323.621 | −13700.164 |
| 25 | 500 | −6845.628 | −6845.688 | −6847.959 | −6845.021 | −8969.687 |
| 50 | 200 | −22186.199 | −22186.100 | −22178.759 | −22186.117 | −44146.644 |

Table 4

| $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | BIC | GBIC | $\log m_1$ |
|---|---|---|---|---|---|---|
| 10 | 1000 | −16581.437 | −16581.458 | −16577.887 | −16580.974 | −23024.404 |
| 25 | 200 | −5724.295 | −5724.374 | −5725.977 | −5724.219 | −6415.674 |
| 25 | 500 | −14511.767 | −14511.781 | −14517.795 | −14512.137 | −15648.107 |
| 50 | 200 | −15704.585 | −15704.618 | −15713.139 | −15705.010 | −18518.298 |

should be used as the penalty term in the definition of BIC in our setup. If we use $-(p/2)\log(n)$ instead, the performance of BIC becomes terrible. In the tables, $-(p/2)\log(r)$ is used as the penalty in BIC. For Tables 1, 2 and 4, we use $\theta_0 = 0$ for calculating $\log m_1$, while for Table 3 we use $\theta_0 = 1$. Tables 1–4 list simulation results for sampling from the Bernoulli, normal, exponential and Poisson distributions respectively, the priors used therein being listed below.

*Bernoulli distribution*: Prior used: $\pi(\boldsymbol{\theta}) = \int\int C^p(\alpha, \beta)\exp\{\alpha\sum_{i=1}^p \theta_i - \beta\sum_{i=1}^p \log(1 + \mathrm{e}^{\theta_i})\}\pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta$, where $\pi_1(\alpha, \beta) = 1/(75 - 5\alpha)$, $0 < \alpha < 5$, $\alpha < \beta < 15$.

*Normal distribution*: Prior used: Zellner–Siow prior.

*Exponential distribution*: Prior used: $\pi(\boldsymbol{\theta}) = \int\int C^p(\alpha, \beta)\exp\{\alpha\sum_{i=1}^p \theta_i + \beta\sum_{i=1}^p \log(-\theta_i)\}\pi_1(\alpha, \beta)\mathrm{d}\alpha\,\mathrm{d}\beta$, where $\pi_1(\alpha, \beta) = \frac{1}{99}$, $(\alpha, \beta) \in [1, 10] \times [-1, 10]$.

Table 5

| Distribution | $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | BIC | GBIC | $\log m_1$ |
|---|---|---|---|---|---|---|---|
| Normal | 50 | 2 | −165.322 | −165.717 | −125.445 | −165.506 | −192.571 |
| Bernoulli | 50 | 10 | −327.45 | −327.684 | −349.577 | −327.863 | −346.574 |
| Exponential | 50 | 10 | −662.526 | −661.979 | −640.320 | −660.384 | −1026.271 |
| Poisson | 50 | 10 | −671.504 | −670.775 | −687.383 | −671.374 | −669.929 |

*Poisson distribution*: Prior used: $\pi(\boldsymbol{\theta}) = \int\int C^p(\alpha, \beta)\exp\{\alpha\sum_{i=1}^{p}\theta_i - \beta\sum_{i=1}^{p} \mathrm{e}^{\theta_i}\}\pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta$, where $\pi_1(\alpha, \beta) = \frac{1}{67.21866}$, $(\alpha, \beta) \in [1, 10] \times [1, 8.46874]$.

In all of the above simulations presented in Tables 1–4, either both $p$ and $r$ are large or $p$ is moderate and $r$ is large. For all these cases, $\log \hat{m}_2$ approximates $\log m_2$ extremely well, and so does GBIC. Another point worth mentioning here is that the performance of BIC is much worse than GBIC as an approximation to $\log m_2$, even though both are only supposed to approximate $\log m_2$ up to O(1). For moderate values of $p$ ($\leqslant 10$) and large $r$, the performance of BIC is reasonable. But, for large values of $p$, the difference between $\log m_2$ and BIC is so large as to cast doubt on whether it actually approximates $\log m_2$ even up to O(1) when $p \to \infty$ and $r \to \infty$. This phenomenon is observed for all the four distributions we considered in our simulation study, and becomes even more severe for the normal distribution. It is clear that the BIC is a poor measure of evidence in high-dimensional problems.

We also made some limited numerical studies to see how our approximation works for moderate/small $r$ and very large $p$. The results are presented in Table 5. The priors used were as before.

For the normal example, the $c_p$ value was 1.6891.

We see that for these cases also, our approximation works pretty well while BIC performs terribly. But comparing the result for the normal with the corresponding approximation in BGM, we see that their approximation is more precise for small values of $r$. The reason for this is explained in Section 3.

For all the simulations above, data are generated assuming $M_2$ is true. We did some limited simulation studies by sampling data when model $M_1$ is true. As expected (see e.g. Theorem 4.1 of BGM), the approximation of $\log m_2$ is less accurate in this case but $\log \hat{m}_2$ appears to be within $\mathrm{o}(\log \hat{m}_2)$ of the correct value. In some cases, the $(\hat{\alpha}, \hat{\beta})$ values fall outside the support of the second-stage prior $\pi_1(\alpha, \beta)$. In these cases, one has to use numerical integration to get a good approximation.

The error rates actually observed are much better than the worst error rates indicated in Section 3. The error estimates appearing in the proof of Theorem 1 are based on worst-case scenarios. Better heuristic error estimates can be obtained by keeping higher-order terms in the Taylor expansion before integration and retaining the sign of the error in approximating $I_i$ by $\hat{I}_i$. These two factors improve the error estimates substantially. We performed one simulation with the Poisson distribution and a conjugate Gamma prior with $p = 30$ and $r = 400$. We estimated $\prod_{i=1}^{p} I_i(\hat{\alpha}, \hat{\beta}) = \int\int\cdots\int\prod_{i=1}^{p} f(\mathbf{y}_i|\theta_i)\pi(\theta_i|\hat{\alpha}, \hat{\beta})\,\mathrm{d}\theta_i$, by $\prod_{i=1}^{p} \hat{I}_i(\hat{\alpha}, \hat{\beta})$, where symbols have meanings as before. The actual ratio of the integral $I(\hat{\alpha}, \hat{\beta}) = \prod_{i=1}^{p} I_i(\hat{\alpha}, \hat{\beta})$ and $\hat{I}(\hat{\alpha}, \hat{\beta}) = \prod_{i=1}^{p} \hat{I}_i(\hat{\alpha}, \hat{\beta})$ is 0.819534, i.e. the actual relative error in approximation $[\hat{I}(\hat{\alpha}, \hat{\beta}) - I(\hat{\alpha}, \hat{\beta})]/I(\hat{\alpha}, \hat{\beta}) = 0.2202$. The estimate of this relative error in approximation, using the heuristic method described above, turns out to be $\frac{1-0.821518}{0.821518} = 0.2172$.

## 6. Heuristic approximations for more general examples

Now consider a case where the $y_{ij}$'s are not necessarily arising out of a general exponential family. Consider $y_{ij} \sim f_i(y_{ij}|\boldsymbol{\theta}_i)$, $j = 1, 2, \ldots, r_i$, where $\boldsymbol{\theta}_i$ is of dimension $k_i$, $k_i \geqslant 1$, $i = 1, 2, \ldots, p$. Assume that the $\boldsymbol{\theta}_i$'s follow $\pi_i(\boldsymbol{\theta}_i|\alpha_1, \ldots, \alpha_s)$, $i = 1, 2, \ldots, p$, where $\alpha_1, \ldots, \alpha_s$ is a common set of hyperparameters for all the $\pi_i(.|)$'s. We then use a second-stage prior $\pi_H(\alpha_1, \ldots, \alpha_s)$ on the hyperparameters. We need to approximate

$$m_2 = \int \cdots \int \left\{ \prod_{i=1}^{p} f_i(\mathbf{y}_i|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|\alpha_1, \ldots, \alpha_s) \right\} \pi_H(\alpha_1, \ldots, \alpha_s)\, \mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_p \mathrm{d}\alpha_1 \ldots \mathrm{d}\alpha_s, \tag{48}$$

where $f_i(\mathbf{y}_i|\boldsymbol{\theta}_i) = \prod_{j=1}^{r_i} f_i(y_{ij}|\boldsymbol{\theta}_i)$.

Using approximations as in Section 2, we get

$$\begin{aligned}
\log \hat{m}_2 = {} & \sum_{i=1}^{p} \sum_{j=1}^{r_i} \log f_i(x_{ij}|\hat{\boldsymbol{\theta}}_i) \\
& - \frac{1}{2} p \left\{ \frac{1}{p} \sum_{i=1}^{p} \log \left( \left| -\frac{\partial^2}{\partial \boldsymbol{\theta}_i^2} \left( \sum_{j=1}^{r_i} \log f_i(y_{ij}|\boldsymbol{\theta}_i) \right) \right|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i} \right) \right\} \\
& + p \left\{ \frac{1}{p} \sum_{i=1}^{p} \log \pi_i(\hat{\boldsymbol{\theta}}_i|\hat{\alpha}_1, \ldots, \hat{\alpha}_s) + \frac{1}{p} \left( \sum_{i=1}^{p} \frac{k_i}{2} \right) \log 2\pi \right\} \\
& - \frac{1}{2} \log \left( \left| -\frac{\partial^2}{\partial \boldsymbol{\alpha}^2} \left\{ \sum_{i=1}^{p} \log \pi_i(\hat{\boldsymbol{\theta}}_i|\alpha_1, \alpha_2, \ldots, \alpha_s) \right\} \right|_{\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_s} \right) \\
& + \frac{s}{2} \log 2\pi + \log \pi_H(\hat{\alpha}_1, \ldots, \hat{\alpha}_s)
\end{aligned} \tag{49}$$

and

$$\mathrm{GBIC} = \log \hat{m}_2 - \frac{s}{2} \log 2\pi - \log \pi_H(\hat{\alpha}_1, \ldots, \hat{\alpha}_s). \tag{50}$$

It would be extremely difficult to try to prove this approximation rigorously, but the results of simulations are promising. We draw samples from Cauchy distribution with a location parameter $\mu_i$. We use $k_i = 1$ and $r_i = r$, for $i = 1, \ldots, p$, a normal prior on the location parameter $\mu_i$, with $\mu_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(t, 1)$ given $t$ and another prior on $t \sim \mathrm{N}(1, 1)$. The density of

Table 6

| $p$ | $r$ | $\log m_2$ | $\log \hat{m}_2$ | GBIC |
|---|---|---|---|---|
| 25 | 100 | −6449.28345 | −6449.31246 | −6450.429192 |
| 15 | 200 | −7737.03941 | −7737.05933 | −7737.019305 |

the $i$th sample is

$$\prod_{j=1}^{r_i} \frac{1}{\pi\{1+(x_{ij}-\mu_i)^2\}}.$$

Considering the kind of generality we are looking at this is an extremely simplistic situation, but this will give us an inkling of how general the structures (4) and (9) are.

Table 6 indicates that even for a non-exponential family situation, our approximation gives pretty accurate results. In principle, it can also be applied to the ecological example (Example 5) of Ghosh and Samanta (2001), treated via Bayes factors. This example appeared earlier in Burnham and Anderson (1998). It can also be applied in principle to the similar ecological examples discussed in Brooks et al. (2000).

## Acknowledgements

The authors would like to thank the associate editor and a referee for their valuable comments and suggestions which led to substantial improvement in the content and presentation of the paper.

## Appendix

We present here one simple example (normal case with $\sigma^2 = 1$) to show how one can get an upper bound for the expected proportion of co-ordinates $i$ for which one might have $\hat{\theta}_i \notin S$. Fix any arbitrary $\varepsilon > 0$. Then

$$\begin{aligned} P\{\hat{\theta}_i \notin S\} &= P\{\hat{\theta}_i \notin S, |\hat{\theta}_i - \theta_i| \leqslant \varepsilon\} + P\{\hat{\theta}_i \notin S, |\hat{\theta}_i - \theta_i| > \varepsilon\} \\ &\leqslant P\{\theta_i \in S'\} + P\{|\hat{\theta}_i - \theta_i| > \varepsilon\}, \end{aligned}$$

where $S'$ is determined from $S$ and the value of $\varepsilon$. First note that,

$$\begin{aligned} P\{|\hat{\theta}_i - \theta_i| > \varepsilon\} &= \int P\{|\hat{\theta}_i - \theta_i| > \varepsilon | \theta_i\} \pi(\theta_i)\, \mathrm{d}\theta_i \\ &\leqslant \frac{1}{\varepsilon^2 r}, \end{aligned} \tag{51}$$

since $\hat{\theta}_i$ is the sample mean and $E_{\theta_i}(\theta_i - \hat{\theta}_i)^2 = 1/r$. As mentioned in Section 3, we need $|\hat{\theta}_i| \leqslant K(\log r)^\gamma$ for some constant $K > 0$ and $0 < \gamma < 1$, for all large $r$, in order for $\hat{\theta}_i$ to fall inside $S$. So in this case we can take

$$S' = \{K(\log r)^\gamma - \varepsilon, \infty\} \cup \{-\infty, -K(\log r)^\gamma + \varepsilon\}.$$

Then, look at

$$\begin{aligned} &P\{\theta_i > K(\log r)^{\gamma} - \varepsilon\} \\ &\quad \leqslant \int\int P\{\theta_i > K(\log r)^{\gamma} - \varepsilon|\alpha, \beta\}\pi_1(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta. \end{aligned} \tag{52}$$

For the conjugate prior we are considering, one has, in general, $\pi(\theta_i|\alpha, \beta) \sim \mathrm{N}(\alpha/\beta, 1/\sqrt{\beta})$. Assuming that $\alpha$ is bounded and $\beta$ is bounded away from zero and infinity, one has

$$P\{\theta_i > K(\log r)^{\gamma} - \varepsilon|\alpha, \beta\} = \mathrm{O}\left(\frac{1}{(\log r)^{\gamma}}\right), \tag{53}$$

as $r \to \infty$, uniformly in $\alpha$ and $\beta$. So, one also has, for large $r$,

$$P\{\theta_i > K(\log r)^{\gamma} - \varepsilon\} = \mathrm{O}\left(\frac{1}{(\log r)^{\gamma}}\right). \tag{54}$$

Similar estimate can be obtained for $P\{\theta_i < -K(\log r)^{\gamma} + \varepsilon\}$; hence, the expected proportion of $i$'s for which $\hat{\theta}_i \notin S$ when $r$ is large is seen to be $\mathrm{O}(1/(\log r)^{\gamma})$, which is negligible for sufficiently large values of $r$.

In general, if one has $\pi(.|\alpha, \beta)$ and $\pi_1(\alpha, \beta)$ such that the mean and variance of $\theta$ are both finite (with non-zero variance), the same estimate for the expected proportion of coordinates $i$ for which $\hat{\theta}_i \notin S$ as obtained above will be true, i.e. this expected proportion will be bounded above by a quantity of magnitude $\mathrm{O}(1/(\log r)^{\gamma})$. The condition that $\alpha$ and $\beta$ be bounded and $\beta > 0$ will not be needed in that case.

## References

Berger, J.O., Ghosh, J.K., Mukhopadhyay, N., 2003. Approximations and consistency of Bayes factors as model dimension grows. J. Statist. Plann. Inference 112, 241–258.

Brooks, S.P., Catchpole, E.P., Morgan, B.J.T., 2000. Bayesian animal survival estimation. Statist. Sci. 15, 357–376.

Brown, L.D., 1986. In: Gupta, S.S. (Ed.), Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory, Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol. 9. Institute of Mathematical Statistics, Hayward, CA.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference—A Practical Information-Theoretic Approach. Springer, Berlin.

Cox, D.R., 1970. Analysis of Binary Data. Chapman & Hall, London.

Ghosh, J.K., Ramamoorti, R.V., 2002. Bayesian Nonparametrics. Springer, New York.

Ghosh, J.K., Samanta, T.S., 2001. Model selection—an overview. Current Sci. 80, 1135–1144.

Ghosh, J.K., Samanta, T.S., 2002. Towards a Non-subjective Bayesian paradigm. In: Misra, J.C. (Ed.), Uncertainty and Optimality. World Scientific Publishing Company, Singapore.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Kass, R.E., Vaidyanathan, S., 1992. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. J. Roy. Statist. Soc. Ser. B 54, 129–144.

Kass, R.E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J. Amer. Statist. Assoc. 90, 928–934.

Kass, R.E., Tierney, L., Kadane, J.B., 1990. The validity of posterior asymptotic expansions based on Laplace's method. In: S. Geisser, et al. (Eds.), Bayesian and Likelihood Methods in Statistics and Econometrics. North-Holland, New York, pp. 473–488.

Pauler, D.K., 1998. The Schwarz criterion and related methods for normal linear models. Biometrika 85, 13–27.

Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Biometrika 83, 251–266.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14, 465–471.

Rissanen, J., 1983. A universal prior for integers and estimation by minimum description length. Ann. Statist. 11, 416–431.

Rissanen, J., 1987. Stochastic complexity. J. Roy. Statist. Soc. Ser. B 49, 223–239.

Schwarz, G., 1978. Estimating the dimension of the model. Ann. Statist. 6, 461–464.

Stone, M., 1979. Comments on model selection criteria of Akaike and Schwarz. J. Roy. Statist. Soc. Ser. B 41, 276–278.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. Roy. Statist. Soc. Ser. B 64, 583–639.

Volinsky, C.T., Raftery, A.E., 2000. Bayesian information criterion for censored survival models. Biometrics 56, 256–262.

Wilbur, J.D., Ghosh, J.K., Nakatsu, C.H., Brouder, S.M., Doerge, R.W., 2002. Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints. Biometrics. 58, 378–386.

Zellner, A., Siow, A., 1980. Posterior odds ratios for selected regression hypothesis. In: Bernardo, J.M., DeGroot, M.H., Lindley, D., Smith, A.F.M. (Eds.), Bayesian Statistics: Proceedings of the First International Meeting held in Valencia. University of Valencia Press,