Journal of Statistical Planning and
Inference 112 (2003) 241–258

Approximations and consistency of Bayes factors as model dimension grows

James O. Berger[a], Jayanta K. Ghosh[b,c], Nitai Mukhopadhyay[c,*]

[a] *Duke University, USA*
[b] *Indian Statistical Institute, India*
[c] *Purdue University, Eli Lilly and Co. LCC, Indianapolis, IN 46285, USA*

Abstract

Stone (J. Roy. Statist. Soc. Ser. B 41 (1979) 276) showed that BIC can fail to be asymptotically consistent. Note, however, that BIC was developed as an asymptotic approximation to Bayes factors between models, and that the approximation is valid only under certain conditions. The counterexample of Stone arises in situations in which BIC is not an adequate approximation. We develop some new approximations to Bayes factors, that are valid for the situation considered in Stone (1979) and discuss related issues of consistency.

Keywords: Model selection; Bayes factor; BIC; Asymptotic consistency

1. Introduction

Stone (1979) had observed that BIC can be inconsistent when the dimension of the parameter goes to infinity. Our first objective here is to address Stone's counterexample, showing that Bayes factors under reasonable priors are consistent; the problem lies in the inappropriateness of BIC as an approximation to Bayes factors in this situation. A new approximation to Bayes factors, for the situation considered by Stone, is then introduced and its accuracy and consistency are considered.

The study is performed with respect to two particular priors that have been proposed for model comparison in normal linear models. The first is the multivariate Cauchy prior, used in Zellner and Siow (1980), and the second is the *Smooth Cauchy* prior, introduced in Berger and Pericchi (1997). When the parameter is high dimensional, it is shown that the choice of prior can make a substantial difference, not only in terms of the answer, but also in terms of consistency in model choice.

Assume a simple ANOVA model where M_1 and M_2 are two nested linear models for n independent normal random variables with known variance σ^2. Under M_1, all n random variables have the same mean, while, under M_2, each block of r random variables has a different mean. Formally, the independent observations y_{ij} are assumed to arise from the linear model:

$$y_{ij} = \delta + \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2),$$

$$i = 1, \ldots, p, \quad j = 1, \ldots, r, \; n = pr.$$

The two models being compared are $M_1 : \mu_i = 0$ for all i against $M_2 : \underset{\sim}{\mu} \in \mathbb{R}^p$.

The BIC criterion (see Schwarz, 1978) selects M_1 (M_2) as ΔBIC is negative (positive), where

$$\begin{aligned}\Delta\mathrm{BIC} &= \mathrm{BIC}_2 - \mathrm{BIC}_1 \\ &= \Delta \text{ Maximized log likelihood} - \frac{(p-1)}{2}\log n \\ &= \frac{r}{2\sigma^2}\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2 - \frac{(p-1)}{2}\log n. \end{aligned} \tag{1}$$

Clearly $\Delta\mathrm{BIC} < 0$ if and only if $r\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2/[(p-1)\sigma^2] < \log n$. Stone assumes that, as $n \to \infty$, $(p/n)\log n \to \infty$ and

$$\sum_{i=1}^{p}\frac{(\mu_i - \bar{\mu})^2}{(p-1)} \to \tau^2 > 0. \tag{2}$$

This condition implies that $r\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2/[(p-1)(\sigma^2 + r\tau^2)] \to 1$ in probability as $n \to \infty$. So, for large n and if M_2 is true, BIC selects M_1 if $(\sigma^2 + r\tau^2)/\sigma^2 < \log n$, which happens with probability 1 since $r/\log n \to 0$. Thus, under any alternative, BIC selects the null model asymptotically, demonstrating its inconsistency.

The first condition of Stone holds, in particular, when r is fixed and $p \to \infty$, and this will henceforth be assumed throughout the paper. Also, without loss of generality, we will set the known variance to one, i.e., assume $\sigma^2 = 1$. Finally, the intercept term δ does not affect consistency of the Bayes factor or BIC, so we will consider only the simplified model

$$\underset{\sim}{y}_{ij} = \mu_i + \varepsilon_{ij}, \; i = 1, \ldots, p, \; j = 1, \ldots, r \text{ and } \varepsilon_{ij} \text{ are iid } \mathrm{N}(0, 1). \tag{3}$$

The two competing models are now $M_1 : \underset{\sim}{\mu} = 0$ and $M_2 : \underset{\sim}{\mu} \in \mathbb{R}^p$.

In Section 2, we discuss the usual Laplace approximation to Bayes factors and its relationship to BIC. In Section 3, we study the consistency of the Bayes factors developed under the multivariate Cauchy and Smooth Cauchy priors. Section 4 develops improved approximations to Bayes factors for this situation of increasing dimension, including one, called generalized Bayes information criterion (GBIC), that does not depend on the prior. Numerical studies of the quality of the various approximations are carried out in Section 5 and conclusions given in Section 6.

2. The Laplace approximation and BIC

In this section, we consider the full Laplace approximation to Bayes factors as a possible improvement over BIC. For the situation in (3), it clearly suffices to consider only the sufficient statistic $\bar{\underset{\sim}{y}} = (\bar{\underset{\sim}{y}}_1, \ldots, \bar{\underset{\sim}{y}}_p) \sim N_p(\underset{\sim}{\mu}, (1/r)I)$. Then, the likelihoods under M_2 and M_1 are, respectively,

$$\mathscr{L}_2 = \frac{r^{p/2}}{(2\pi)^{p/2}} e^{-(r/2)(\bar{\underset{\sim}{y}} - \underset{\sim}{\mu})'(\bar{\underset{\sim}{y}} - \underset{\sim}{\mu})},$$

$$\mathscr{L}_1 = \frac{r^{p/2}}{(2\pi)^{p/2}} e^{-(r/2)\bar{\underset{\sim}{y}}'\bar{\underset{\sim}{y}}}.$$

If the prior $\pi(\underset{\sim}{\mu})$ is specified for $\underset{\sim}{\mu}$ under M_2, then the Bayes factor of M_2 to M_1 is given by

$$\text{BF}_{21} = \frac{m_2(\bar{\underset{\sim}{y}})}{\mathscr{L}_1} = \frac{\int \mathscr{L}_2 \pi(\underset{\sim}{\mu})\, d\underset{\sim}{\mu}}{\mathscr{L}_1}.$$

The Laplace approximation to the Bayes factor, for fixed p and large r, can be obtained from Pauler (1998) and Kass and Wasserman (1995). The penalty term in BIC, namely $(p/2)\log n$, arises from the Laplace approximation to the marginal likelihood when the log determinant of the sample information matrix is $O(n)$. However, the sample information matrix in our context is rI, so the corresponding Laplace approximation to the log of the Bayes factor of Model 2 to Model 1 leads to

$$\log \text{BF}_{21} \approx \log \mathscr{L}_2(\hat{\underset{\sim}{\mu}}) - \log \mathscr{L}_1 - \frac{p}{2}\log r + \frac{p}{2}\log(2\pi) + \log \pi(\hat{\underset{\sim}{\mu}}) \tag{4}$$

$$= \frac{r}{2\sigma^2}\bar{\underset{\sim}{y}}'\bar{\underset{\sim}{y}} - \frac{p}{2}\log r + \frac{p}{2}\log(2\pi) + \log \pi(\hat{\underset{\sim}{\mu}}), \tag{5}$$

where $\pi(\hat{\underset{\sim}{\mu}})$ is the prior density evaluated at the mle of $\underset{\sim}{\mu}$. Approximation (5) suggests that $\log r$ may be a better penalty than $\log n$ when $p \to \infty$. Indeed, the approximation based on the first two terms in (5) can easily be shown to be consistent, as $r \to \infty$, for any $\tau^2 > 0$.

The necessity of properly defining the sample size appearing in BIC was also discussed in Kass and Wasserman (1995). In the context of mixed models, a new version of sample size is suggested in Pauler (1998), and leads to r when applied to the above example. We will refer to (5) as the KWP approximation.

The motivation for the Laplace approximation is based on considering fixed p while $r \to \infty$. When p can also go to ∞, the Laplace approximation need no longer be valid. (A 'valid' Laplace approximation here is one for which the error goes to zero as $r \to \infty$.) For instance, if one chooses a multivariate normal prior for $\underset{\sim}{\mu}$, a straightforward computation shows that the difference between the Bayes factor and the Laplace approximation is $O(p/r)$, which actually goes to ∞ if p grows faster than r (which is the case considered in this paper). It would certainly be of interest to study, in general, when the Laplace approximation is valid, but such a study is beyond the scope of this paper. (Some related results, for the case when p grows much more slowly than r, are available from Ghosal, 1999.) We do, however, develop a better approximation to the considered Bayes factors, for increasing p, in Section 4.

3. Consistency and inconsistency of the Bayes factors

For the Stone example, we consider two specific priors that have been suggested in the literature, the multivariate Cauchy prior, recommended by Zellner and Siow (1980), and the *Smooth Cauchy* prior introduced in Berger and Pericchi (1997). It was suggested in the latter paper that Cauchy tails are appropriate for a default prior when selecting from among linear models (see also Jeffreys, 1961), but that the multivariate Cauchy prior has too sharp a spike at zero; hence, the Smooth Cauchy prior may be more reasonable. The results below show that this is not so when $p \to \infty$; the likely explanation for the phenomenon is discussed in Section 6.

Multivariate Cauchy prior:

$$\pi_c(\underset{\sim}{\mu}) = \frac{\Gamma((p+1)/2)}{\pi^{(p+1)/2}}(1 + \underset{\sim}{\mu}'\underset{\sim}{\mu})^{-(p+1)/2}$$
$$= \int_0^{\infty} \frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\underset{\sim}{\mu}'\underset{\sim}{\mu}} \frac{1}{\sqrt{2\pi}} e^{-t/2} t^{-1/2}\, dt. \qquad (6)$$

To define the Smooth Cauchy prior, let $M(p,q,\lambda)$ denote the hypergeometric 1F1 function (Abramowitz and Stegun, 1970)

$$M(p,q,\lambda) = \frac{\Gamma(q)}{\Gamma(p)} \sum_{j=0}^{\infty} \frac{\Gamma(p+j)}{\Gamma(q+j)} \frac{\lambda^j}{j!}.$$

Then the Smooth Cauchy prior is

$$\pi_{sc}(\underset{\sim}{\mu}) = \frac{\Gamma((p+1)/2)}{\Gamma((p+2)/2)\Gamma(\frac{1}{2})(2\pi)^{p/2}} \exp\left\{-\frac{\underset{\sim}{\mu}'\underset{\sim}{\mu}}{2}\right\} M\left(\frac{1}{2}, \frac{p+2}{2}, \frac{\underset{\sim}{\mu}'\underset{\sim}{\mu}}{2}\right) \qquad (7)$$

$$= \int_0^{1} \frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\underset{\sim}{\mu}'\underset{\sim}{\mu}} \frac{1}{\pi\sqrt{t(1-t)}}\, dt \qquad (8)$$

$$= \int_0^{\infty} \frac{(u/(1+u))^{p/2}}{(2\pi)^{p/2}} e^{-(\underset{\sim}{\mu}'\underset{\sim}{\mu}/2)(u/(1+u))} \frac{1}{\pi\sqrt{u}(1+u)}\, du. \qquad (9)$$

Eqs. (6) and (8) express the prior density as a scale mixture of normals, which will be key in deriving the new approximations to the Bayes factor. In the following, denote the Gamma($\frac{1}{2}, \frac{1}{2}$) density by $g(t, \frac{1}{2}, \frac{1}{2})$. Then, the marginal densities under M_2 with the above priors are as follows:

$$m_2^c(\underset{\sim}{y}) = \int_0^{\infty} \phi\left(\bar{\underset{\sim}{y}}; \underset{\sim}{0}, \left(\frac{1}{r} + \frac{1}{t}\right) I_p\right) g(t, 1/2, 1/2)\, dt$$
$$= \int_0^{\infty} \mathscr{L}_2^c(t) g(t, 1/2, 1/2)\, dt, \qquad (10)$$

$$m_2^{sc}(\underset{\sim}{y}) = \int_0^\infty \phi\left(\bar{\underset{\sim}{y}};\ \underset{\sim}{0}, \left(\frac{1}{r}+\frac{1}{u}+1\right) I_p\right) F_{1,1}(u)\,\mathrm{d}u$$

$$= \int_0^\infty \mathscr{L}_2^{sc}(u) F_{1,1}(u)\,\mathrm{d}u, \tag{11}$$

where $F_{1,1}(u)$ is the F density with (1,1) degrees of freedom and

$$\mathscr{L}_2^{c} = \phi\left(\bar{\underset{\sim}{y}};\ \underset{\sim}{0}, \left(\frac{1}{r}+\frac{1}{t}\right) I_p\right), \tag{12}$$

$$\mathscr{L}_2^{sc} = \phi\left(\bar{\underset{\sim}{y}};\ \underset{\sim}{0}, \left(\frac{1}{r}+\frac{1}{u}+1\right) I_p\right). \tag{13}$$

The marginal density under M_1 is

$$\mathscr{L}_1(\underset{\sim}{y}) = \frac{r^{p/2}}{(2\pi)^{p/2}}\,\mathrm{e}^{-(r/2)\,\bar{\underset{\sim}{y}}'\,\bar{\underset{\sim}{y}}}.$$

No surprises concerning consistency are encountered for the Cauchy prior. Indeed, the following general theorem, whose proof is given in Section 7.2, covers this case.

Theorem 3.1. *For any prior of the form*

$$\pi_g(\underset{\sim}{\mu}) = \int_0^\infty \frac{t^{p/2}}{(2\pi)^{p/2}}\,\mathrm{e}^{-(t/2)\,\underset{\sim}{\mu}'\underset{\sim}{\mu}}\,g(t)\,\mathrm{d}t \tag{14}$$

with $g(t)$ having support equal to $(0,\infty)$, the Bayes factor is consistent under M_1. Consistency under M_2 holds if

$$\tau^2 = \liminf_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p}\mu_i^2 > 0.$$

The condition under M_2 can be relaxed to the situation in which the μ_i are exchangeable. The Bayes factor is then consistent provided $E(\mu_1^2|A_\infty)$ has no mass at 0, where A_∞ is the symmetric σ field for which the μ_i's are conditionally iid.

The situation for the Smooth Cauchy prior turns out to be quite surprising. Indeed, consistency depends on the value of $\tau_2 = \lim_p \sum \mu_i^2/p$ (assuming the limit exists). For simplicity, we state results for $r=1$. The corresponding Bayes Factor, BF_{21}^{sc}, can be shown to be consistent if $\tau^2 > 2\log 2 - 1$, but for smaller τ^2 the following surprising result holds.

Theorem 3.2. *When $r=1$, the Bayes Factor BF_{21}^{sc} is inconsistent under M_2 for $0<\tau^2<2\log 2-1$. Indeed, for any prior of form (14), with $g(t)$ being supported on a finite interval $[0,T]$ (note that the Smooth Cauchy prior has $T=1$), there will be such an inconsistency region.*

Proof. Using Eq. (8) to compute the marginal m_2^{sc}, we have

$$\mathrm{BF}_{21}^{sc} = \int_{t=0}^{1} \mathrm{e}^{(p/2)(c_p/(t+1) - \log(1+1/t))} \pi(\mathrm{d}t)$$
$$= \int_{t=0}^{1} \mathrm{e}^{(p/2)h(t)} \pi(\mathrm{d}t),$$

where π is the Beta$(\frac{1}{2}, \frac{1}{2})$ density. Note that $h'(t) = (1 - t(c_p - 1))t^{-1}(1+t)^{-2}$. Thus $h'(t) > 0$ for $t < 1/(c_p - 1)$ and $c_p > 1$. If $c_p < 2$ then $h(t)$ is strictly increasing on $(0, 1)$. Thus $h(1) = c_p/2 - \log 2$ implies that $h(t) < 0$ on $(0, 1)$ if $1 < c_p < 2\log 2$, i.e. $\mathrm{BF}_{21}^{HDsc} < 1$. By the SLLN, $0 < \tau^2 < 2\log 2 - 1 \Rightarrow 1 < c_p < 2\log 2$ for large p.

To prove the result for priors of form (14), note that the function $h(t)$ is increasing on $[0, (c_p - 1)^{-1}]$. Thus, for small values of c_p, $h(t)$ increases on $[0, T]$ to the maximum value $h(T) = c_p(T+1)^{-1} - \log(1 + 1/T) = c_p(T+1)^{-1} + \log(1 - 1/(1+T))$. This is less than zero for values of c_p sufficiently close to 1, i.e. for sufficiently small values of τ^2 close to 0. □

The above theorem provides a strong reason for preferring a $g(t)$ that has full support and, in particular, for preferring the Cauchy prior to the Smooth Cauchy as a default prior.

4. Approximations to Bayes factors as $p \to \infty$

The new approximation to the Bayes factor is calculated using the Laplace approximation to the one-dimensional integrals in (10) or (11). As the approximation is deemed to be useful for high-dimensional problems, and is different from the Laplace approximation discussed earlier, we denote the approximation by the suffix HD. For this Laplace approximation, define

$$t_0 = \operatorname*{argmax}_t \mathscr{L}_2^c(t) g(t),$$
$$u_0 = \operatorname*{argmax}_u \mathscr{L}_2^{sc}(u) F_{1,1}(u).$$

The desired Laplace approximation to the log Bayes factor under the multivariate Cauchy prior, denoted $\log \mathrm{BF}_{21}^c$, is then given by

$$\begin{aligned}\log \mathrm{BF}_{21}^{HDc} &= \log m_2^{HDc} - \log \mathscr{L}_1(y) \\ &= \log(\mathscr{L}_2^c(t_0) g(t_0)) + \frac{1}{2}\left(\log(2\pi) - \log\left(-\frac{\mathrm{d}^2 \log(\mathscr{L}_2^c g)}{\mathrm{d}t^2}(t_0)\right)\right) \\ &\quad - \log \mathscr{L}_1. \end{aligned} \tag{15}$$

Similarly, $\log \mathrm{BF}_{21}^{\mathrm{sc}}$ is approximated by

$$\begin{aligned}
&\log \mathrm{BF}_{21}^{\mathrm{HDsc}} \\
&\quad = \log m_2^{\mathrm{HDsc}} - \log \mathscr{L}_1(\underset{\sim}{y}) \\
&\quad = \log(\mathscr{L}_2^{\mathrm{sc}}(u_0) F_{1,1}(u_0)) + \frac{1}{2}\left(\log(2\pi) - \log\left(-\frac{\mathrm{d}^2 \log(\mathscr{L}_2^{\mathrm{sc}} F_{1,1})}{\mathrm{d}u^2}(u_0)\right)\right) \\
&\qquad - \log \mathscr{L}_1.
\end{aligned} \tag{16}$$

While t_0 and u_0 do not have a convenient closed form, they can be easily computed numerically. To obtain insight into the behavior of (15) and (16), however, it is useful to find approximations for t_0 and u_0, by maximizing $\mathscr{L}_2^{\mathrm{c}}$ and $\mathscr{L}_2^{\mathrm{sc}}$, respectively. The result for t_0 is simplest. Indeed, under the assumption that

$$c_p = \frac{1}{p} \sum \bar{y}_i^2 > \frac{1}{r} + \varepsilon$$

for some $\varepsilon > 0$, as $p \to \infty$, it is easy to see that $\mathscr{L}_2^{\mathrm{c}}$ is maximized at $t_1 = (c_p - 1/r)^{-1}$. Also, $t_1 = t_0 + \mathrm{O}(1/p)$. Hence, if one performs the Laplace approximation about t_1 instead of t_0, the additional error of approximation is o(1). After some algebraic simplification, this alternative approximation reduces to

$$\log \mathrm{BF}_{21}^{\mathrm{HDc}} = \frac{r}{2} \bar{y}' \bar{y} - \frac{p}{2} \log(r c_p) - \frac{p}{2} - \frac{\log p}{2} + C + \mathrm{o}(1), \tag{17}$$

where $C = (-t_0 + \log(2 c_p^2 t_0^3))/2$.

The analogous approximation for the Smooth Cauchy Bayes factor is slightly more complicated. Indeed, defining

$$u_1 = \left(c_p - \frac{1}{r} - 1\right)^{-1} \text{ and } \psi(t) = c_p \left(\frac{1}{r} + \frac{1}{t}\right)^{-1} + \log\left(\frac{1}{r} + \frac{1}{t}\right),$$

it can be shown that

$$\begin{aligned}
&\log \mathrm{BF}_{21}^{\mathrm{HDsc}} \\
&\quad = \begin{cases}
\frac{r}{2} \bar{y}' \bar{y} - \frac{p}{2} \log(r c_p) - \frac{p}{2} - \frac{\log p}{2} + C + \mathrm{o}(1) & \text{if } c_p > 1 + \frac{1}{r}, \\
\frac{r}{2} \bar{y}' \bar{y} - \frac{p}{2} \log(r) - \frac{p}{2} \psi(1) + \mathrm{o}(1) & \text{if } c_p = 1 + \frac{1}{r}, \\
\frac{r}{2} \bar{y}' \bar{y} - \frac{p}{2} \log(r) - \frac{p}{2} \psi(1) & \text{if } \frac{1}{r} < c_p < 1 + \frac{1}{r}, \\
-\frac{1}{2} \log(-\frac{p}{2} \pi \psi'(1)) + \mathrm{o}(1) &
\end{cases}
\end{aligned} \tag{18}$$

where $C = \log(\sqrt{2}/\sqrt{\pi}(1 + u_1)) + \frac{1}{2} \log(2 c_p^2 u_1^3)$. The proof of the above is given in Section 7.3. That the approximation to the Bayes factor in (15) is accurate follows from the following theorem, whose proof is delayed until Section 7.1.

Theorem 4.1. *Under M_2 and with the multivariate Cauchy prior, the relative error in approximating $\mathrm{BF}_{21}^{\mathrm{c}}$ by $\mathrm{BF}_{21}^{\mathrm{HDc}}$ goes to 0, i.e.,*

$$\log \mathrm{BF}_{21}^{\mathrm{c}} - \log \mathrm{BF}_{21}^{\mathrm{HDc}} = \mathrm{o}(1). \tag{19}$$

Under M_1 and with the multivariate Cauchy prior, the relative error in approximating $\log(\mathrm{BF}_{21}^{c})$ *by* $\log(\mathrm{BF}_{21}^{\mathrm{HDc}})$ *goes to* 0, *i.e.*,

$$\log \mathrm{BF}_{21}^{c} - \log \mathrm{BF}_{21}^{\mathrm{HDc}} = \log \mathrm{BF}_{21}^{\mathrm{HDc}} \mathrm{o}(1). \tag{20}$$

The convergence in (19) of the alternative Laplace approximation under M_2 can be shown to hold for any prior of the form (14), for which the second stage density, $g(t)$, is continuous and has $(0,\infty)$ as its support. Note, from the representation in (8), that the Smooth Cauchy prior is not of this form; the support of its $g(t)$ is only $(0,1)$. This makes it difficult to obtain an analogue of Theorem 4.1 for the Smooth Cauchy prior.

It is interesting to note that, for larger c_p, (17) and (18) are equal, up to a constant. It is thus tempting to use this common dominant term to develop a generalization of BIC. Indeed, with a slight adjustment for smaller c_p, we will consider, as the difference between the generalized BIC for the comparison of M_2 and M_1,

$$\Delta\mathrm{GBIC} = \left(\frac{r}{2}\,\bar{y}'\bar{y} - \frac{p}{2}\log(rc_p) - \frac{p}{2}\right)^{+} - \frac{\log p}{2}, \tag{21}$$

where '+' refers to the positive part. It can, indeed, be shown that, as $p \to \infty$ and when $c_p > \epsilon + 1/r$, this approximation to the log Bayes factor holds for any prior of the form (14) with $g(t)$ having support equal to $(0,\infty)$.

A final observation of interest is that ΔGBIC provides a valid approximation to $\log \mathrm{BF}_{21}$, up to O(1), for the case of fixed p as well as when $p \to \infty$ (when $c_p > \epsilon + 1/r$). The performance of ΔGBIC is studied numerically in Section 5. Also, it is interesting to observe that ΔGBIC is a consistent model selection criterion. Consistency under M_2 follows from the consistency of the Bayes factor (Theorem 3.1) and the error convergence theorem (Theorem 4.1). To prove consistency under M_1, note that

$$\begin{aligned}
\Delta\mathrm{GBIC} &= \frac{p}{2}\left(r\,\frac{\bar{y}'\bar{y}}{p} - \log(rc_p) - 1\right)^{+} - \frac{1}{2}\log p \\
&= \frac{p}{2}((rc_p - 1) - \log(1 + (rc_p - 1)))^{+} - \frac{\log p}{2} \\
&\to -\infty,
\end{aligned}$$

since $\log(1+(rc_p-1)) = (rc_p-1) - \frac{1}{2}(rc_p-1)^2 + \mathrm{o}_p((rc_p-1)^2)$ and $rc_p - 1 = \mathrm{O}_p(1/\sqrt{p})$.

5. A numerical comparison of the different approximations to the Bayes factor

We present a small numerical comparison of the performance of the different approximations to the Bayes factors for the Cauchy and Smooth Cauchy priors. The approximations considered are that based on ordinary BIC, as was used by Stone, the ordinary Laplace approximation, as described in Pauler (1998) and denoted by $\mathrm{Lap}_{\mathrm{KWP}}$, GBIC described in Section 4, and the mixture Laplace approximation $\log \mathrm{BF}^{\mathrm{HD}}$. The actual Bayes factors are computed using one-dimensional numerical integration in (10) and (11).

Table 1
Log (Bayes factor) and its approximations under the Cauchy prior

c_p	$\log \mathrm{BF}^c$	BIC	GBIC	$\mathrm{Lap}^c_{\mathrm{KWP}}$	$\log \mathrm{BF}^{\mathrm{HDc}}$
0.1	−8.5348	−110.129	−1.956	15.1269	−8.5776
0.5	−3.8251	−90.129	−1.956	−2.2647	−3.9083
1.0	6.0388	−65.129	5.715	5.555	5.9236
1.5	20.8203	−40.129	20.579	20.3831	20.7564
2.0	38.4814	−15.129	38.387	38.1312	38.4408
2.1	42.235	−10.129	42.167	41.8991	42.1971
10.0	397.369	384.871	398.151	397.293	397.369

Table 2
Log (Bayes factor) and its approximations under the Smooth Cauchy prior

c_p	$\log \mathrm{BF}^{sc}$	BIC	GBIC	$\mathrm{Lap}^{sc}_{\mathrm{KWP}}$	$\log \mathrm{BF}^{\mathrm{HDsc}}$
0.1	−26.093	−110.129	−1.956	−16.9651	−27.2466
0.5	−12.6056	−90.129	−1.956	−6.6981	−13.7322
1.0	4.3422	−65.129	5.715	6.4891	3.2007
1.5	21.5136	−40.129	20.579	21.1541	20.4648
2.0	39.2827	−15.129	38.387	38.5356	38.6141
2.1	42.9655	−10.129	42.167	42.2631	42.3613
10.0	397.255	384.871	398.151	397.172	397.234

The previous discussions suggested that $\mathrm{Lap}_{\mathrm{KWP}}$ will be considerably more accurate than BIC, which ignores the order of the information matrix, whereas $\mathrm{BF}^{\mathrm{HD}}$ attempts to improve upon $\mathrm{Lap}_{\mathrm{KWP}}$ by capturing the additional correction needed when p grows with n. The main goal of the numerical study is to indicate whether these suggestions are reflected in practice. Theory has less to say concerning the accuracy of the comparatively adhoc GBIC, so the numerical evidence will be especially important in its validation.

The Bayes factors depend on the data only through the statistic $\bar{y}'\bar{y} = pc_p$, so results are given only for various values of c_p. We have taken $p = 50$ and $n = 100$, so there are only $r = 2$ replicates in each group. The different values of c_p are taken to be 0.1, 0.5, 1, 1.5, 2, 2.1, 10. Note that $c_p \approx \tau^2 + 1/r$, where $\tau^2 = (1/p)\sum \mu_i^2$, so that values near 0.5 are expected under M_1, with larger values under M_2.

Tables 1 and 2 support the suggestions from theory. The terrible performance of BIC, through its ignoring the correct order of the information matrix, is stunning. The excellent performance of the new Laplace approximation, $\log \mathrm{BF}^{\mathrm{HD}}$, is gratifying. The standard Laplace approximation seems fine for the larger values of c_p (i.e., when M_2 is true), but does not do well for smaller c_p (i.e., when M_1 is true). GBIC is similar, in doing well for larger c_p but not for smaller c_p; interestingly the performance of the simple GBIC is comparable to that of the much more complicated $\mathrm{Lap}_{\mathrm{KWP}}$ (which requires knowledge of the prior).

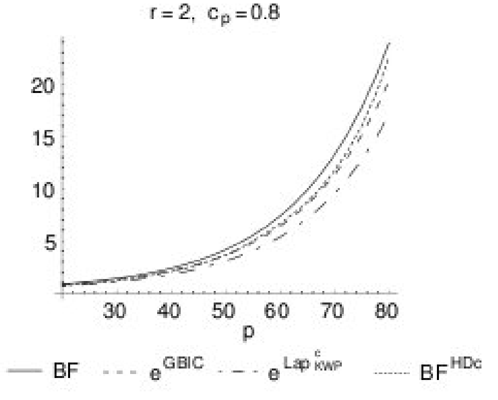

Fig. 1. Performance of different approximations to the Bayes factor under a Cauchy prior as $p \to \infty$, when $c_p = 0.8$ and $r = 2$.

In Fig. 1, the performance of various approximations to the Bayes factor are studied as $p \to \infty$. For the figure, we used $r = 2$, $c_p = 0.8$ and the Cauchy prior. Both $\mathrm{BF}^{\mathrm{HDc}}$ and GBIC perform well for increasing p. The Laplace approximation is also reasonably good but its error does seem to significantly increase with p. Since it also requires specification of the prior, it would seem that GBIC is clearly preferable for the situation considered in this paper. BIC, in the form used in Stone's paper, is inconsistent and so cannot be graphed with the others.

This is, of course, only a limited numerical study, but it confirms the two main messages of the paper, that (as noted in Kass and Wasserman, 1995; Pauler, 1998) appropriate definition of sample size is needed in defining approximations to Bayes factors, and that it can be important to correct the effects of increasing dimension when that occurs as the sample size increase.

6. Discussion

BIC is often equated with 'the Bayesian answer' in model selection, and counterexamples to BIC, such as that of Stone (1979), are sometimes misinterpreted as being counterexamples to Bayesian model selection. We demonstrated that suitable Bayes factors will be consistent, that suitable approximations to them will also be consistent, but that BIC can be a terrible approximation to Bayes factors.

A further aspect of Stone (1979) was the demonstration that AIC is consistent for the situation considered in this paper. Comparison of Bayes factors and AIC is a very complex undertaking, a beginning towards which was made in Mukhopadhyay (2000). AIC is primarily designed to minimize prediction error and is well known to be potentially inconsistent for model selection in even simple problems (such as the normal linear model with fixed parameter dimension, sample size going to infinity, and the less complex model being true). The consistency of AIC here, when the parameter dimension grows to infinity, is thus somewhat unusual.

That the Smooth Cauchy prior results in an inconsistent Bayes factor is surprising in two respects. First, it is rare that actual proper Bayes factors fail to be consistent in model selection and so it is important to understand how this can happen. Second, the Smooth Cauchy prior was derived as the *intrinsic prior* for default model selection in Berger and Pericchi (1997), and lauded therein for its desirable properties. It is important to note, however, that the derivation in that paper was based on fixed p with $r \to \infty$, and that the intrinsic prior for the scenario of fixed r and $p \to \infty$ could be very different.

Finally, a beginning has been made in defining a generalization of BIC that does not depend on the prior distribution and that is effective as the model dimension grows with the sample size. Whether GBIC can itself be generalized to include situations in which p and r both grow is a question of great importance and very substantial difficulty.

7. Proofs

7.1. Error convergence

We will use the following notations:

$$D_i(t) = \frac{\mathrm{d}^i \log(\mathscr{L}_2^c)}{\mathrm{d}t^i} \quad \text{and} \quad D_i \log g(t) = \frac{\mathrm{d}^i \log g}{\mathrm{d}t^i}$$

Proof of Theorem 4.1. *Case* I: M_2 is true. In this case, as $p \to \infty$, using the assumption $c_p > 1/r + \varepsilon$, for some $\varepsilon > 0$, t_0 is almost surely bounded away from zero and infinity. Using notations from Section 4,

$$\mathscr{L}_2^c = (2\pi)^{-p/2} \mathrm{e}^{-(p/2)\psi(t)} \quad \text{and} \quad m_2^c(y) = \int_0^\infty \mathscr{L}_2^c g(t)\,\mathrm{d}t.$$

Denote $\log m_2^{\mathrm{HDc}*}(y) = \log \mathscr{L}_2^c g(t_1) - \frac{1}{2}\log(-D_2(t_1)) + \frac{1}{2}\log(2\pi)$. It follows from continuity of $\psi(t)$ and the fact that $|t_0 - t_1| = \mathrm{O}(1/p)$ and $D_2(t_0) = D_2(t_1)(1 + \mathrm{o}(1)) = \mathrm{O}(p)$ that

$$\begin{aligned}
&|\log m_2^{\mathrm{HDc}}(y) - \log m_2^{\mathrm{HDc}*}(y)| \\
&\quad = |\log \mathscr{L}_2^c g(t_0) - \log \mathscr{L}_2^c g(t_1) - \tfrac{1}{2}\log(-D_2(t_0) \\
&\qquad - D_2 \log g(t_0)) + \tfrac{1}{2}\log(-D_2(t_1))| \\
&\quad = |\log \mathscr{L}_2^c g(t_0) - \log \mathscr{L}_2^c g(t_1) - \tfrac{1}{2}\log(-D_2(t_0)/p \\
&\qquad - D_2 \log g(t_0)/p) + \tfrac{1}{2}\log(-D_2(t_1)/p)| = \mathrm{o}(1).
\end{aligned}$$

We will show that $|m_2^c(y) - m_2^{\mathrm{HDc}*}(y)| = m_2^{\mathrm{HDc}*}(y)\mathrm{o}(1)$. As $|\log m_2^{\mathrm{HDc}*}(y) - \log m_2^{\mathrm{HDc}}(y)| = \mathrm{o}(1)$, the result follows. For a constant $b > 0$, denote $A = \{t\colon |t - t_1| \leqslant b \log p/\sqrt{D_2}(t_1)\}$. The argument t_1 in $D_2(t_1)$ is dropped in further calculations for notational simplicity:

$$\begin{aligned}|m_2^c(y) - m_2^{HDc*}(y)| &= \left|\int e^{\log(\mathscr{L}_2^c g)(t)}\,dt - \int e^{\log(\mathscr{L}_2^c g)(t_1)+\frac{1}{2}D_2(t-t_1)^2}\,dt\right| \\ &= \left|\int (\psi_1(t) - \psi_2(t))\,dt\right| \\ &= \int_A |\psi_1 - \psi_2|\,dt + \int_{A^c} |\psi_1 - \psi_2|\,dt \\ &\leqslant \int_A |\psi_1 - \psi_2|\,dt + \int_{A^c} |\psi_1|\,dt + \int_{A^c} |\psi_2|\,dt \\ &= I_1 + I_2 + I_3.\end{aligned}$$

Part I: Denote $R_p(t) = D_1 \log g(t_1)(t - t_1) + D_2 \log g(t_1)(t - t_1)^2 + D_3(t)(t - t_1)^3/3!$. Then computations show that $D_1 \log g(t_1) = O(1)$, $D_2 \log g(t_1) = O(1)$, $D_2(t_1) = O(p)$ and $D_3(t_1) = O(p)$. Therefore,

$$\begin{aligned}\sup_A |R_p(t)| &= \left(\frac{b \log p}{\sqrt{D_2}}\right)^3 O(p) + \left(\frac{b \log p}{\sqrt{D_2}}\right)^2 O(1) + \frac{b \log p}{\sqrt{D_2}} O(1) \\ &= O\left(\frac{(\log p)^3}{\sqrt{p}}\right)(1 + o(1)).\end{aligned}$$

Thus

$$\begin{aligned}I_1 &= \int_A |\psi_1 - \psi_2|\,dt \\ &= \mathscr{L}_2^c(t_1)g(t_1) \int_A e^{(1/2)D_2(t-t_1)^2} |e^{R_p(t')} - 1|\,dt \ [t' \text{ is between } t \text{ and } t_1] \\ &\leqslant \mathscr{L}_2^c(t_1)g(t_1) \int_A e^{(1/2)D_2(t-t_1)^2} |R_p(t')| e^{|R_p(t')|}\,dt \\ &\leqslant \mathscr{L}_2^c(t_1)g(t_1) \sup_A |R_p(t)| e^{|R_p(t)|} \int e^{-(1/2)D_2(t-t_1)^2}\,dt \\ &\leqslant m_2^{HDc*} O\left(\frac{(\log p)^3}{\sqrt{p}}\right) e^{O((\log p)^3/\sqrt{p})} \\ &= m_2^{HDc*} o(1).\end{aligned}$$

Part II: Define $f(x) = (1/x)e^{x-1} - 1$. Then, $f(1) = f'(1) = 0$ and $f''(1) = 1$ imply $f(x) = (x-1)^2/2 + o((x-1)^2)$. Defining $s = 1/r + (t_1 + (b \log p)/\sqrt{D_2})^{-1}$,

$$\frac{\mathscr{L}_2^c(t_1 + (b\log p)/\sqrt{D_2})}{\mathscr{L}_2^c(t_1)} = \left(\frac{s}{c_p} e^{c_p/s-1}\right)^{-p/2}$$

$$= \left(1 + \frac{s}{c_p} e^{c_p/s-1} - 1\right)^{-p/2}$$

$$= \left(1 + \frac{1}{2}\left(\frac{c_p}{s} - 1\right)^2 + o\left(\left(\frac{c_p}{s} - 1\right)^2\right)\right)^{-p/2}.$$

Also,

$$\frac{c_p}{s} - 1 = \frac{rb}{t_1(r+t_1)} \frac{\log p}{\sqrt{D_2}} + o\left(\frac{\log p}{\sqrt{D_2}}\right).$$

Similarly, denoting $q = 1/r + (t_1 - (b \log p)/\sqrt{D_2})^{-1}$, we have

$$\frac{\mathscr{L}_2^c(t_1 - (b\log p)/\sqrt{D_2})}{\mathscr{L}_2^c(t_1)} = \left(1 + \frac{1}{2}\left(\frac{c_p}{q} - 1\right)^2 + o\left(\left(\frac{c_p}{q} - 1\right)^2\right)\right)^{-p/2},$$

$$\frac{c_p}{q} - 1 = -\frac{rb}{t_1(r+t_1)} \frac{\log p}{\sqrt{D_2}} + o\left(\frac{\log p}{\sqrt{p}}\right).$$

Therefore,

$$\begin{aligned} I_2 &= \int_{A^c} \mathscr{L}_2^c(t) g(t)\,dt \\ &\leqslant \mathscr{L}_2^c(t_1) \sup_{A^c} \frac{\mathscr{L}_2^c(t)}{\mathscr{L}_2^c(t_1)} P_g(A^c) \\ &\leqslant \mathscr{L}_2^c(t_1) \left(1 + \frac{1}{2} \frac{r^2 b^2}{t_1^2 (r+t_1)^2} \frac{(\log p)^2}{p} + o\left(\frac{(\log p)^2}{p}\right)\right)^{-p/2}. \\ &= \mathscr{L}_2^c(t_1) O(e^{-(1/4)(r^2b^2/t_1^2(r+t_1)^2)(\log p)^2}) \\ &= m_2^{\mathrm{HDc}*} \sqrt{p} O(p^{-(1/4)(r^2b^2 \log p)/t_1^2(r+t_1)^2}) \text{ [as } D_2 = O(p)] \\ &= m_2^{\mathrm{HDc}*} o(1). \end{aligned}$$

Part III:

$$\int_{A^c} |\psi_2|\,dt = \mathscr{L}_2^c(t_1) g(t_1) \int_{A^c} e^{(1/2)D_2(t-t_1)^2}\,dt$$

$$= m_2^{\mathrm{HDc}*} \int_{|u| > b \log p} e^{-(1/2)u^2}\,du$$

$$= m_2^{\mathrm{HDc}*}\mathrm{O}\left(\frac{1}{b\log p}\mathrm{e}^{-(b^2/2)\log p}\right)$$

$$= m_2^{\mathrm{HDc}*}\mathrm{o}(1).$$

Taking all three parts together,

$$|m_2^{\mathrm{c}}(y) - m_2^{\mathrm{HDc}*}(y)| = m_2^{\mathrm{HDc}*}\mathrm{o}(1),$$

which directly yields

$$|\mathrm{BF}_{21}^{\mathrm{c}} - \mathrm{BF}_{21}^{\mathrm{HDc}}| = \mathrm{BF}_{21}^{\mathrm{HDc}}\mathrm{o}(1).$$

Case II: M_1 *is true*. Note that $t_0 = \mathrm{argmax}\,\mathscr{L}_2^{\mathrm{c}}(t)g(t)$ is the solution of

$$\frac{\mathrm{d}\log g(t)}{\mathrm{d}t} + \frac{\mathrm{d}\log \mathscr{L}_2^{\mathrm{c}}}{\mathrm{d}t} = 0$$

$$\Rightarrow \frac{p}{2t^2}\left(\frac{1}{r}+\frac{1}{t}\right)^{-2}\left(\frac{1}{r}+\frac{1}{t}-c_p\right) - \frac{1}{2} - \frac{1}{2t} = 0$$

$$\Rightarrow p\left(\frac{1}{r}+\frac{1}{t}-c_p\right) = t(t+1)\left(\frac{1}{r}+\frac{1}{t}\right)^2$$

$$\Rightarrow \frac{p}{t} + \mathrm{O}(\sqrt{p}) = t(t+1)\left(\frac{1}{r}+\frac{1}{t}\right)^2$$

$$\Rightarrow t_0 = \mathrm{O}(p^{1/3}) \to \infty.$$

Thus $D_2(t_0) = \mathrm{O}(1)$. Choose $\bar{t} > t_0$ such that $t_0 = \mathrm{o}(\bar{t})$ and $(\mathrm{d}/\mathrm{d}t)[\log(\mathscr{L}_2^{\mathrm{c}}g)] = (p/2t^2)$ $(1/r+1/t)^{-2}(1/r+1/t-c_p) - 1/2t - 1/2 < -1/4$ for $t > \bar{t}$. So, under M_1,

$$\log \mathrm{BF}_{21}^{\mathrm{c}} - \log \mathrm{BF}_{21}^{\mathrm{HDc}}$$

$$\leqslant \left|\log\left\{\int_0^\infty \mathrm{e}^{\log \mathscr{L}_2^{\mathrm{c}}(t)g(t) - \log \mathscr{L}_2^{\mathrm{c}}(t_0)g(t_0)}\,\mathrm{d}t\right\}\right| \text{const.}$$

$$\leqslant \left|\log\left\{\int_0^{\bar{t}} \mathrm{d}t + \int_{\bar{t}}^\infty \exp\left\{\int_{\bar{t}}^t \frac{\mathrm{d}}{\mathrm{d}t}\log \mathscr{L}_2^{\mathrm{c}} g\,\mathrm{d}t\right\}\,\mathrm{d}t\right| + \text{const.}$$

$$\leqslant \left|\log\left\{\bar{t} + \int_{\bar{t}}^\infty \mathrm{e}^{(t-\bar{t})/4}\,\mathrm{d}t\right\}\right| + \text{const.}$$

$$= \mathrm{o}(\log \mathrm{BF}_{21}^{\mathrm{HDc}}),$$

completing the proof. □

7.2. Consistency theorem

Lemma 7.1. $\liminf rc_p = 1 + r\tau^2$.

Proof. Clear $\bar{y}_i = \mu_i + (1/\sqrt{r})Z_i$ for iid standard normal $Z_i s$. Therefore,

$$\frac{r}{p}\sum_1^p(\bar{y}_i^2 - \mu_i^2) = 2\sqrt{r}\frac{\sum Z_i\mu_i}{p} + \frac{\sum Z_i^2}{p}.$$

By SLLN, $\sum Z_i^2/p \to 1$ a.s. Also, $X_p = \sum Z_i\mu_i \sim \mathrm{N}(0, \sum \mu_i^2)$. So, using the inequality $P(|Z| > a) < (2/a)\phi(a)$, it follows that

$$P\left(|X_p| > a\left(\log p \sum \mu_i^2\right)^{1/2}\right) < \frac{2}{a\sqrt{2\pi \log p}}\mathrm{e}^{-(a^2/2)\log p}.$$

The sum of the RHS over p is finite if $a > \sqrt{2}$ and, hence, by the Borel–Cantelli lemma, $P(|X_p| > a\sqrt{\log p \sum \mu_i^2} \text{ i.o.}) = 0$. This implies that

$$\overline{\lim}\frac{1}{p}X_p = \mathrm{o}\left(\frac{1}{pr}\sum_1^p \mu_i^2\right) \text{ a.s.}$$

$$\Rightarrow rc_p - \frac{r\sum\mu_i^2}{p} - 1 = \mathrm{o}\left(\frac{1}{p}\sum_1^p \mu_i^2\right) \text{ a.s.}$$

$$\Rightarrow \liminf rc_p = 1 + \liminf r\sum\frac{\mu_i^2}{p}$$

$$= 1 + r\tau^2,$$

completing the proof. □

Proof of Theorem 3.1. *Case* I: Consistency under M_2.

Using notations from Section 4, the Bayes factor is

$$\mathrm{BF}_{21}^c = \int_{t=0}^{\infty} \mathrm{e}^{(p/2)[rc_p(1-(1+r/t)^{-1})+\log(1+r/t)^{-1}]}\pi(t)\,\mathrm{d}t.$$

Denote $x = 1 - (1 + r/t)^{-1}$ and $f(x) = rc_p x + \log(1 - x)$. Then

$$\mathrm{BF}_{21}^c = \int_{x=0}^{1} \mathrm{e}^{(p/2)f(x)}\tilde{\pi}(\mathrm{d}x), \tag{22}$$

where $\tilde{\pi}(\mathrm{d}x)$ is the measure induced by the change of variables. Now $\liminf rc_p = 1 + r\tau^2$ and $\tau^2 > 0$ imply almost surely that $rc_p > 1 + \delta$ for some $\delta > 0$, for large p. Thus we can choose an interval (a, b) such that on (a, b), $f(x) > (1 + \delta)x + \log(1 - x) > \varepsilon$ for some $\varepsilon > 0$. That yields (almost surely)

$$\mathrm{BF}_{21}^c \geqslant \int_a^b \mathrm{e}^{(p/2)\varepsilon}\tilde{\pi}(\mathrm{d}x)$$

$$\to \infty,$$

completing the proof. □

Case II: *Consistency under* M_1. Under M_1, we have $rc_p = 1 + \mathrm{O}_p(1/\sqrt{p})$. Denote $S_\varepsilon = \{x\colon f(x) > -\varepsilon\}$. From the concavity of $f(x)$, S is an interval of the form $[0, x_0]$.

A simple limit argument shows that $x_0 \downarrow 0$ as $rc_p \downarrow 1$. Also $f_{\max} = O_p(1/p)$. Therefore, from Eq. (22)

$$\begin{aligned}
\mathrm{BF}_{21}^{c} &\leqslant \int_0^{x_0} e^{(p/2)f_{\max}} \tilde{\pi}(dx) + \int_{x_0}^1 e^{(p/2)\varepsilon} \tilde{\pi}(dx) \\
&= K\tilde{\pi}(0, x_0) + e^{-p\varepsilon/2} \tilde{\pi}(x_0, 1) \\
&\xrightarrow{P} 0,
\end{aligned}$$

completing the proof. □

7.3. Laplace approximation to the Bayes factor with Smooth Cauchy prior

Proof of Eq. (18). For the case $c_p > 1 + 1/r$, $\mathscr{L}_2^c(t)$ has a peak in $(0,1)$, and the Laplace approximation to $m_2^{sc}(\underset{\sim}{y})$ is done by Taylor approximation around that peak. When $1/r < c_p \leqslant 1 + 1/r$, conventional Laplace approximation is impossible to implement as the peak occurs at the boundary of the support of the mixing distribution. Define $\psi(t) = c_p(1/r + 1/t)^{-1} + \log(1/r + 1/t)$. Then

$$m_2^{sc}(\underset{\sim}{y}) = \int_0^1 e^{-(p/2)\psi(t)} \frac{dt}{\pi\sqrt{t(1-t)}}.$$

So keeping the first term in the Taylor expansion after $\psi(1)$,

$$\begin{aligned}
m_2^{sc}(\underset{\sim}{y}) &\approx \int_0^1 e^{-(p/2)(\psi(1)+(t-1)\psi'(1))} \frac{dt}{\pi\sqrt{t(1-t)}} \\
&= \frac{1}{\pi} e^{-(p/2)\psi(1)} \int_0^1 e^{p\psi'(1)t/2} \frac{dt}{\sqrt{t(1-t)}} \\
&= e^{-(p/2)\psi(1)} \text{ if } \psi'(1) = 0.
\end{aligned}$$

If $\psi'(1) \neq 0$, then the above integral can be exactly evaluated as

$$\frac{1}{\pi} e^{-(p/2)\psi(1)} \int_0^1 e^{p\psi'(1)t/2} \frac{dt}{\sqrt{t(1-t)}} = e^{(p/2)(\psi'(1)/2 - \psi(1))} \mathrm{Bessel}I(0, \tfrac{p\psi'(1)}{4}).$$

But we will further approximate it as follows. Suppose $\psi'(1) = -\alpha$ for some $\alpha > 0$. Then, for some $\delta = p^{-\eta}$ where $0 < \eta < 1$, we have

$$\begin{aligned}
m_2^{sc}(\underset{\sim}{y}) &= \frac{1}{\pi} e^{-(p/2)\psi(1)} \int_0^1 \frac{e^{-p\alpha t/2}\, dt}{\sqrt{t(1-t)}} \\
&= \frac{1}{\pi} e^{-(p/2)\psi(1)} \left\{ \int_0^{\delta} \frac{e^{-p\alpha t/2}\, dt}{\sqrt{t(1-t)}} + \int_1^{\delta} \frac{e^{-p\alpha t/2}\, dt}{\sqrt{t(1-t)}} \right\}.
\end{aligned}$$

Clearly,

$$\int_\delta^1 \frac{e^{-p\alpha t/2}\,dt}{\sqrt{t(1-t)}} \leqslant e^{-p\alpha\delta/2}\int_0^1 \frac{dt}{\sqrt{t(1-t)}} \sim e^{-p\alpha\delta/2}.$$

For the other part, note that

$$\frac{1}{\sqrt{1-\delta}}\int_0^\delta \frac{e^{-p\alpha t/2}\,dt}{\sqrt{t}} \leqslant \int_0^\delta \frac{e^{-p\alpha t/2}\,dt}{\sqrt{t(1-t)}} \leqslant \int_0^\delta \frac{e^{-p\alpha t/2}\,dt}{\sqrt{t}}.$$

Since δ goes to zero as $p \to \infty$, the middle integral above can be approximated by $\int_0^\delta e^{-p\alpha t/2} t^{-1/2}\,dt$. This can be further approximated by $\int_0^\delta e^{-p\alpha t/2} t^{-1/2}\,dt$, since the error term is given by

$$\begin{aligned}\int_\delta^\infty \frac{e^{-p\alpha t/2}\,dt}{\sqrt{t}} &= \int_{-p\alpha\delta/2}^\infty \frac{e^{-s}\,ds}{\sqrt{s}} \\ &= e^{-p\alpha\delta/2}\int_0^\delta \frac{e^{-s}\,ds}{\sqrt{s+p\alpha\delta/2}} \leqslant e^{-p\alpha\delta/2}\int_0^\infty \frac{e^{-s}\,ds}{\sqrt{s}} \\ &= o\left(\int_0^\infty \frac{e^{-s}\,ds}{\sqrt{s}}\right).\end{aligned}$$

The above calculation justifies the approximation

$$\begin{aligned} m_2^{sc}(\underset{\sim}{y}) &= \frac{1}{\pi} e^{-(p/2)\psi(1)} \int_0^\delta e^{p\psi'(1)t/2}\frac{dt}{\sqrt{t}}(1+o(1)) \\ &\approx \frac{1}{\pi} e^{-(p/2)\psi(1)} \frac{\Gamma(\frac{1}{2})}{\sqrt{-p\psi'(1)/2}} \\ &= \frac{e^{-(p/2)\psi(1)}}{\sqrt{-p\pi\psi'(1)/2}}, \end{aligned}$$

which gives

$$\begin{aligned} \log \mathrm{BF}_{21}^{sc} &\approx \log m_2^{sc}(\underset{\sim}{y}) - \log \mathscr{L}_1(\underset{\sim}{y}) \\ &= \begin{cases} \frac{r}{2}\bar{y}'\bar{y} - \frac{p}{2}\log(r) - \frac{p}{2}\psi(1) - \frac{1}{2}\log(-\frac{p}{2}\pi\psi'(1)) & \text{if } \psi'(1) \neq 0, \\ \frac{r}{2}\bar{y}'\bar{y} - \frac{p}{2}\log(r) - \frac{p}{2}\psi(1) & \text{if } \psi'(1) = 0. \end{cases} \end{aligned}$$

Uncited reference

Berger and Pericchi, 1996.

Acknowledgements

We thank two referees for their many helpful comments. This research was supported by The National Science Foundation (USA), Grants DMS-9802261 and DMS-0103265.

References

Abramowitz, M., Stegun, I., 1970. Handbook of Mathematical Functions, Vol. 55. National Bureau of Standards Applied Mathematics.

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. 91 (433), 109–122.

Berger, J., Pericchi, L., 1997. On the justification of default and intrinsic Bayes factors. In: Lee, J.C., et al. (Ed.), Modeling and Prediction. Springer, New York, pp. 276–293.

Ghosal, S., 1999. Asymptotic normality of posterior distributions in high dimensional linear models. Bernoulli 5, 315–331.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, London.

Kass, R.E., Wasserman, L., 1995. A reference Bayesian test for nested hypothesis and its relationship to the Schwarz criterion. J. Amer. Statist. Assoc. 90 (431), 928–934.

Mukhopadhyay, N.D., 2000. Bayesian model selection for high dimensional models with prediction error loss and 0–1 loss. Ph.D. Thesis, Purdue University.

Pauler, D., 1998. The Schwarz criterion and related methods for normal linear methods. Biometrika 85 (1), 13–27.

Schwarz, G., 1978. Estimating the dimension of a 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.

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.