



Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

## **Prior-based Bayesian information criterion**

M. J. Bayarri, James O. Berger, Woncheol Jang, Surajit Ray, Luis R. Pericchi & Ingmar Visser

To cite this article: M. J. Bayarri, James O. Berger, Woncheol Jang, Surajit Ray, Luis R. Pericchi & Ingmar Visser (2019) Prior-based Bayesian information criterion, Statistical Theory and Related Fields, 3:1, 2-13, DOI: 10.1080/24754269.2019.1582126

To link to this article: https://doi.org/10.1080/24754269.2019.1582126

Б		П	Т
		П	1
			J

Published online: 14 Mar 2019.



Submit your article to this journal 🗹

Article views: 148



🖸 View related articles 🗹



View Crossmark data 🗹



Citing articles: 3 View citing articles 🕑

## **Prior-based Bayesian information criterion**

M. J. Bayarri<sup>a,†</sup>, James O. Berger<sup>b</sup>, Woncheol Jang<sup>c</sup>, Surajit Ray<sup>d</sup>, Luis R. Pericchi<sup>e</sup> and and Ingmar Visser<sup>f</sup>

<sup>a</sup>Department of Statistics and Operations Research, University of Valencia, Valencia, Spain; <sup>b</sup>Department of Statistical Science, Duke University, Durham, NC, USA; <sup>c</sup>Department of Statistics, Seoul National University, Seoul, Korea; <sup>d</sup>School of Mathematics and Statistics, University of Glasgow, Glasgow, UK; <sup>e</sup>Department of Mathematics, University of Puerto Rico, San Juan, Puerto Rico; <sup>f</sup>Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

#### ABSTRACT

We present a new approach to model selection and Bayes factor determination, based on Laplace expansions (as in BIC), which we call Prior-based Bayes Information Criterion (PBIC). In this approach, the Laplace expansion is only done with the likelihood function, and then a suitable prior distribution is chosen to allow exact computation of the (approximate) marginal likelihood arising from the Laplace approximation and the prior. The result is a closed-form expression similar to BIC, but now involves a term arising from the prior distribution (which BIC ignores) and also incorporates the idea that different parameters can have different effective sample sizes (whereas BIC only allows one overall sample size *n*). We also consider a modification of PBIC which is more favourable to complex models.

### **ARTICLE HISTORY**

Received 24 June 2017 Accepted 10 February 2019

Taylor & Francis

Check for updates

Taylor & Francis Group

### KEYWORDS

Bayes factors; model selection; Cauchy priors; consistency; effective sample size; Fisher information; Laplace expansions; robust priors

### 1. Background

### 1.1. The original BIC (Schwarz, 1978)

Suppose that we observe  $X_i = (X_{i1}, ..., X_{ip}) \sim g(\mathbf{x}_i \mid \boldsymbol{\theta})$  for i = 1, ..., n. Here  $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$  is a unknown vector and, in Schwartz's derivation of BIC,  $g(\mathbf{x} \mid \boldsymbol{\theta})$  is an exponential family. Then the log-likelihood function is

$$l(\boldsymbol{\theta}) = \log f(\mathbf{x} \mid \boldsymbol{\theta}) = \log \left( \prod_{i=1}^{n} g(\mathbf{x}_i \mid \boldsymbol{\theta}) \right),$$

where  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ . The goal of Schwarz (1978) is to find a simple approximation to the marginal density

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

where  $\pi(\theta)$  is a prior density for the unknown  $\theta$ , and use the approximation for model comparison.

**Result 1.1 (Stone, 1979):** Let  $\hat{\theta}$  be the MLE of  $\theta$ . Then, under reasonable conditions and as  $n \to \infty$ ,

BIC 
$$\equiv -2l(\widehat{\theta}) + p \log n = -2 \log m(\mathbf{x}) + c + o(1),$$

where *c* is a constant.

Schwartz then suggested comparing two models  $M_1$  and  $M_2$ , using

$$\Delta BIC = BIC_2 - BIC_1,$$

© East China Normal University 2019

preferring  $M_2$  ( $M_1$ ) as this is negative (positive). Clearly this is equivalent to basing the model comparison on the Bayes factor (odds) of  $M_2$  to  $M_1$ , with the approximation

$$B_{21} \equiv \frac{m_2(\mathbf{x})}{m_1(\mathbf{x})} = \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_2\right)}{\exp\left(-\frac{1}{2}\mathrm{BIC}_1\right)}\exp\left(\frac{1}{2}(c_2 - c_1)\right)$$
$$\times (1 + o(1)) \approx \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_2\right)}{\exp\left(-\frac{1}{2}\mathrm{BIC}_1\right)}.$$
(1)

### 1.2. Problems with general use of BIC

BIC is an excellent tool for the class of problems for which it was developed. Unfortunately, it is today used ubiquitously, for completely different classes of problems. We here outline some of the issues with using BIC inappropriately.

Problem 1. The term  $\exp(\frac{1}{2}(c_2 - c_1))$  in (1) is ignored by BIC.

This could have been a serious problem even with proper use of BIC, except that there happens to be pseudo-prior distributions that yield BIC itself (Raftery, 1999), i.e. for which the term  $\exp(\frac{1}{2}(c_2 - c_1)) = 1$ . These pseudo-priors are not real priors, in that they are centred at the mle's of each model, which is a problematical double use of the data. Nevertheless it is comforting that there is at least some type of prior distribution that yields BIC exactly.

Problem 2. What is n?

- (i) A common mistake in specifying n: Note that, in Schwartz's setup, there are n vector observations of dimension p, so that there are a total of np real observations. It is common to mistakenly use  $n^* = np$  as the sample size in BIC, rather than the correct n.
- (ii) *Different parameters can have different n.*

**Example 1.2 (Group means):** For i = 1, ..., p and l = 1, ..., r, suppose we observe

$$X_{il} = \mu_i + \epsilon_{il},$$

where  $\epsilon_{il} \sim N(0, \sigma^2)$ . If  $\sigma^2$  were known, this would be exactly the setup of Schwartz, and the sample size for  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$  would be *r*. In effect, each  $\mu_i$  has a sample size of *r* associated with it. But, if  $\sigma^2$  is unknown, the parameter is  $\boldsymbol{\theta} = (\mu_1, \dots, \mu_p, \sigma^2)$  and it is not reasonable to also associate the sample size of *r* to  $\sigma^2$ , in that we know there are p(r-1) degrees of freedom associated with the mle of  $\sigma^2$ .

An alternative argument is to note that the observed information matrix  $\widehat{I} = (\widehat{I}_{jk})$ , with (j, k) entry

$$\widehat{I}_{jk} = -\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(\mathbf{x} \mid \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}}$$

is given by

$$\widehat{I} = \begin{pmatrix} \frac{r}{\widehat{\sigma}^2} I_{p \times p} & 0\\ 0 & \frac{pr}{2\widehat{\sigma}^4} \end{pmatrix},$$

where  $\hat{\sigma}^2 = (1/pr) \sum_{i=1}^p \sum_{l=1}^r (X_{il} - \bar{X}_i)^2$ . The information matrix suggests that the effective sample size for each  $\mu_i$  is r, while the effective sample size for  $\sigma^2$  is pr. Whether we use p(r-1) or pr for the sample size associated with  $\sigma^2$  will not typically make much difference, whereas the difference with using r, instead, will be quite large.

## (iii) Different observations can have different observed information content.

**Example 1.3:** Suppose each independent observation,  $X_i$ , i = 1, ..., n, has probability 1/2 of arising from the  $N(\theta, 1)$  distribution and probability 1/2 of arising from the  $N(\theta, 1000)$  distribution. Clearly half the observations are essentially worthless, and the 'effective sample size' is n/2.

**Example 1.4 (Findley's BIC counterexample):** One of the famous counter examples against inappropriate use of BIC is in Findley (1991). Suppose the observations

are

$$X_{i} = \frac{1}{\sqrt{i}} \cdot \theta + \epsilon_{i}, \text{ where } \epsilon_{i} \sim N(0, 1),$$
  
$$i = 1, \dots, n, \tag{2}$$

and we are comparing the models  $H_0: \theta = 0$  and  $H_1: \theta \neq 0$ . It turns out that the mle for  $\theta$  is consistent under  $H_1$  (a necessary condition to apply BIC), but that BIC is inconsistent if  $0 < |\theta| < 1$ , in that BIC will then declare  $H_0$  to be the true model as  $n \to \infty$ . The problem here is that, even though the information about  $\theta$  goes to  $\infty$  as *n* grows, it grows much more slowly than *n* (actually, the information grows at roughly  $\log n$  rate), and BIC erroneously assigns the rate to be *n*.

### Problem 3. What is *p*?

Just as n is often not clearly defined for use in BIC, the parameter dimension p is often not clearly defined (see also Pauler, 1998).

**Example 1.5 (Random effect group means):** Consider hierarchical or random effect versions of the group means problem, where it is assumed that

$$\mu_i \sim N(\xi, \tau^2),$$

with  $\xi$  and  $\tau^2$  being unknown. The number of parameters might appear to be p+3 (the means, along with  $\sigma^2$ ,  $\xi$  and  $\tau^2$ ), but one could, alternatively, integrate out  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$  (since it has a known distribution) obtaining

$$f(\mathbf{x} \mid \sigma^2, \xi, \tau^2) = \int f(\mathbf{x} \mid \boldsymbol{\mu}, \xi, \sigma^2) \pi(\boldsymbol{\mu} \mid \xi, \tau^2) \, \mathrm{d}\boldsymbol{\mu}$$
$$\propto \frac{1}{\sigma^{-p(r-1)}} \exp\left\{\frac{\widehat{\sigma}^2}{2\sigma^2}\right\} \prod_{i=1}^p$$
$$\times \exp\left\{-\frac{(\bar{x}_i - \xi)^2}{2(\frac{\sigma^2}{r} + \tau^2)}\right\}.$$

The marginal likelihood will be the integral of this, with respect to a prior  $\pi(\sigma^2, \xi, \tau^2)$ , so that, if one is really viewing BIC as an approximation to the marginal likelihood, it would be correct to set p = 3.

### Problem 4. What if *p* grows with *n*?

BIC is based on an asymptotic argument with p fixed and n growing, but often p is growing with n; BIC then does not apply. If one were to erroneously apply BIC in such a situation, one could end up with inconsistency, as shone by Stone (1979) for the group means example, with known variance  $\sigma^2 = 1$  for simplicity. Indeed, in comparing models  $H_0: \mu = 0$  and  $H_1: \mu \neq 0$  for the group means problem with r = 2,

$$\Delta BIC = BIC_1 - BIC_0 = -2\sum_{i=1}^p \bar{x}_i^2 + p\log 2,$$

which, under  $H_0$ , behaves like  $p(\log 2 - 1) \rightarrow -\infty$  as p grows, thus incorrectly selecting model  $H_1$ .

### 1.3. Variants of BIC

Noting the limitations of BIC, researchers have proposed a host of generalisations, many of which have performed better than BIC under specific scenarios. Many of these methods arise from the variations in retaining the number of terms in the Laplace approximation of the Bayes Factor (Kass & Raftery, 1995). One variant – called the HBIC – (Haughton, 1988) retains the third term in the Laplace approximation of the Bayes Factor. A simulation study by Haughton, Oud, and Jansen (1997) shows that HBIC performs better in model selection for structural equation models than does the usual BIC. Following HBIC, Bollen, Ray, Zavisca, and Harden (2012) developed a similar criterion, called the Information matrix-based Bayesian Information Criterion (IBIC), which retains more terms in the Bayes Factor approximation and outperforms BIC and HBIC in many scenarios. Bollen et al. (2012) also proposed another criterion, named the scaled unit information prior (SPBIC), which generalises the interpretation of the unit information prior in the context of BIC. For approximation of Bayes factors as the model dimension grows, Berger, Ghosh, and Mukhopadhyay (2003) proposed another approximation, named GBIC. Following Berger et al. (2003), a generalisation of BIC for the general exponential family was proposed by Chakrabarti and Ghosh (2006), and a new BIC for change point analysis was proposed by Shen and Ghosh (2011). Some other extensions of BIC include techniques for comparing graphical models (Foygel & Drton, 2010), singular models (Drton & Plummer, 2017), and sparse models (Zak-Szatkowska & Bogdan, 2011).

### 1.4. Overview of the paper

Section 2 presents a proposal to generalise BIC, in order to overcome the problems mentioned above. It is based on use of a specific (robust) prior distribution in the computation of the approximate marginal likelihood of a model. Section 3 discusses a critical aspect of the definition of PBIC, namely the need to determine the 'effective sample size' corresponding to each parameter in a model. Section 4 presents an alternative called PBIC\*. It employs an empirical Bayes prior in computation of the marginal likelihood approximation, resulting in answers more favourable to complex models. Section 5 illustrates the use of PBIC and PBIC\* in the normal linear model; it is of interest that PBIC and PBIC\* correspond to exact marginal likelihoods here. Illustrations in the section are simple linear regression, testing the equality of normal means with known unequal variances, Findley's counterexample, and the group means problem, where consistency results for PBIC and PBIC\* are established as  $p \to \infty$ .

### 2. The PBIC solution

We propose a solution to these problems that depends only on software that can compute mle's and observed information matrices. The basis of the solution is a modified Laplace approximation to  $m(\mathbf{x})$  for reasonable default priors.

### 2.1. Two important preliminaries

One should analytically integrate out any parameter that has a distribution given other parameters, if it is possible to do so. For example, in the hierarchical group means example, base the analysis on the marginal likelihood  $f(\mathbf{x} | \sigma^2, \xi, \tau^2)$ , rather than the full likelihood.

We will be utilising the Laplace approximation, which is most accurate (Kass & Vaidyanathan, 1992; Tierney, Kass, & Kadane, 1989) if the parameter space is transformed to be all of  $\mathfrak{N}^p$ . Transformation to  $\mathfrak{N}^p$  will also be necessary for the subsequent step of the analysis. As an illustration, in the (non-multilevel) group means example, transform to  $\nu = \log \sigma^2$ . Then  $\boldsymbol{\theta} =$  $(\mu_1, \ldots, \mu_p, \nu) \in \mathfrak{N}^{(p+1)}$ . Note that one then works with the transformed mle  $\log \hat{\sigma}^2$  and the transformed observed information matrix

$$\widehat{\boldsymbol{I}}^*(\boldsymbol{\mu},\boldsymbol{\nu}) = \begin{pmatrix} \frac{r}{\widehat{\sigma}^2} I_{p \times p} & 0\\ 0 & \frac{pr}{2} \end{pmatrix}.$$

In the multilevel group means example, both  $\sigma^2$  and  $\tau^2$  would need to be transformed in this fashion.

### 2.2. PBIC and PBIC\* definitions

Suppose  $\theta = (\theta_{(1)}, \theta_{(2)})$ , where  $\theta_{(2)}$  denotes the parameters that are common to all models under consideration (e.g. an intercept in linear regression). Changing notation, let *p* denote the dimension of  $\theta_{(1)}$  and *q* denote the dimension of  $\theta_{(2)}$ ; note that *p* will typically vary from model to model, while *q* is fixed. Partition the observed information matrix for a model accordingly, as

$$\widehat{I} = \begin{pmatrix} \widehat{I}_{11} & \widehat{I}_{12} \\ \widehat{I}_{21} & \widehat{I}_{22} \end{pmatrix}, \text{ and define}$$
$$\Sigma^{-1} = \widehat{I}_{11} - \widehat{I}_{12}\widehat{I}_{22}^{-1}\widehat{I}_{12}^{t}.$$
(3)

(If there are no common parameters to all models, then  $\Sigma = \widehat{I}^{-1}$ .) Change variables to  $\xi = O\theta_{(1)}$ , where O is an orthogonal matrix such that  $\Sigma = O^t DO$ , with  $D = \text{diag}\{d_i\}$  for i = 1, ..., p, and define  $\widehat{\xi} = O\widehat{\theta}_{(1)}$  (the transformed mle). The choice of O does not affect the definition below. For each transformed parameter  $\xi_j$ , let  $n_j^e$  be the *effective sample size* corresponding to that parameter. This is the most difficult aspect of the construction, but equals the intuitive choices of parameter

sample size discussed in the earlier examples; formal definitions will be presented in Section 3. Then PBIC is defined as

$$PBIC \equiv -2l(\widehat{\theta}) + \log|\widehat{I}_{22}| + \sum_{i=1}^{p} \log(1 + n_i^e)$$
$$-2\sum_{i=1}^{p} \log\frac{(1 - e^{-v_i})}{\sqrt{2}v_i}, \qquad (4)$$

where  $v_i = \hat{\xi}_i^2 / [d_i(1 + n_i^e)]$ . For a certain natural prior distribution, PBIC will be shown to be accurate, as an approximation to  $-2 \log m(\mathbf{x})$ , up to a o(1) term as  $n \to \infty$  (for fixed dimension p). Note that, if there are no common parameters to all models, then

$$PBIC = -2l(\widehat{\theta}) + \sum_{i=1}^{p} \log(1 + n_i^e)$$
$$-2\sum_{i=1}^{p} \log \frac{(1 - e^{-v_i})}{\sqrt{2}v_i}.$$
(5)

In the classic case considered by Schwartz, all  $n_i^e$  would equal a common *n*, and the first two terms in this expression are then BIC (up to a o(1) term); the 'constant' ignored by BIC is the final term in (5).

To summarise results in one place, here is an alternative version of the approximation, one which is more favourable to complex models; its development is given in Section 4:

$$PBIC^* \equiv -2l(\widehat{\theta}) + \log|\widehat{I}_{22}| + \sum_{i=1}^p \log(1+n_i^e) - 2\sum_{i=1}^p \log\frac{(1-e^{-\min\{v_i,1.3\}})}{\sqrt{2v_i \min\{v_i,1.3\}}}.$$
 (6)

Note that, if dealing with only normal mean parameters, PBIC and PBIC<sup>\*</sup> are exact as an approximation to  $-2 \log m(\mathbf{x})$ , as discussed below. This would mean, for instance, that when dealing with  $p \rightarrow \infty$ , there would be no need to worry about the accuracy of the approximations.

Here are the steps in the derivation of PBIC.

### 2.2.1. Laplace approximation

By a Taylor's series expansion of  $l(\theta)$  about the mle  $\hat{\theta}$ , with  $\nabla$  denoting the gradient and  $\hat{I}$  being the observed information matrix as defined earlier,

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$
$$= \int e^{l(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

$$= \int \exp\left[l(\widehat{\theta}) + (\theta - \widehat{\theta})^{t} \nabla l(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^{t} \widehat{I}(\theta - \widehat{\theta})\right] \pi(\theta) \, \mathrm{d}\theta (1 + o(1))$$
$$= e^{l(\widehat{\theta})} \int e^{-(1/2)(\theta - \widehat{\theta})^{t} \widehat{I}(\theta - \widehat{\theta})} \pi(\theta) \, \mathrm{d}\theta (1 + o(1)),$$
(7)

where o(1) denotes a term that goes to zero as the sample size *n* grows. Technical conditions for the validity of this Laplace approximation can be found in, e.g. Tierney et al. (1989), Kass and Vaidyanathan (1992); the key assumption needed is that  $\hat{\theta}$  occurs on the interior of the parameter space, so that  $\nabla l(\hat{\theta}) = 0$ . (If not true, the analysis must proceed as in Dudley & Haughton, 1997; Haughton, 1991, 1993). Also, the presence of o(1) assumes that *p* is fixed as *n* grows. We will nevertheless use this approximation, even as *p* grows with *n*, relying on the considerable evidence that the Laplace approximation is quite generally accurate.

Note that we do not use the more common version of the Laplace expansion which involves  $\pi(\theta)$  in the Taylor's expansion because we will be choosing  $\pi(\theta)$ so that the integral in this expression can be evaluated in closed form. In particular, this means that, if we are dealing with the situation where  $\theta$  is the mean parameter of a normal model, then the computations herein will be entirely closed form, with no approximation being involved (and no need to then worry about *p* growing with *n*).

### 2.2.2. Choosing a good prior $\pi(\theta)$

Assume that the transformations in Section 2.1 have been made.

Step 1. Recall that  $\theta = (\theta_{(1)}, \theta_{(2)})$ , where  $\theta_{(2)}$  denotes the common parameters to all models. We will utilise a prior distribution

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-q} \pi(\boldsymbol{\theta}_{(1)}),$$

where  $\pi(\theta_{(1)})$  is defined later. The key point is that, since  $\theta_{(2)}$  is common to all models, it can be assigned a constant prior density (see, e.g. Bayarri, Berger, Forte, & García-Donato, 2012; Berger, Pericchi, & Varshavsky, 1998); choosing the constant to be  $(2\pi)^{-q}$  is to simplify the resulting expression. With the definitions given in (3), integrating out  $\theta_{(2)}$  results in the expression

$$m(\mathbf{x}) = e^{l(\widehat{\boldsymbol{\theta}})} \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{t}\widehat{\mathbf{I}}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right)$$
$$\times (2\pi)^{-q} d\boldsymbol{\theta}_{(2)}\pi(\boldsymbol{\theta}_{(1)}) d\boldsymbol{\theta}_{(1)}(1 + o(1))$$
$$= e^{l(\widehat{\boldsymbol{\theta}})} |\widehat{\mathbf{I}}|^{-1/2}$$
$$\times \int \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\theta}_{(1)} - \widehat{\boldsymbol{\theta}}_{(1)})^{t} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_{(1)} - \widehat{\boldsymbol{\theta}}_{(1)})\right)}{|\boldsymbol{\Sigma}|^{1/2}}$$
$$\times \pi(\boldsymbol{\theta}_{(1)}) d\boldsymbol{\theta}_{(1)}(1 + o(1)).$$

Step 2. Change variables to  $\boldsymbol{\xi} = \boldsymbol{O}\boldsymbol{\theta}_{(1)}$ , where  $\boldsymbol{O}$  is an orthogonal matrix such that  $\boldsymbol{\Sigma} = \boldsymbol{O}^t \boldsymbol{D} \boldsymbol{O}$ , with  $\boldsymbol{D} = \text{diag}(d_i)$  for  $i = 1, \dots, p$ . (The choice of  $\boldsymbol{O}$  does not matter in the following.) Note that  $\boldsymbol{\hat{\xi}} = \boldsymbol{O}\boldsymbol{\hat{\theta}}_{(1)}$ .

For this model, we will utilise a prior distribution that is independent in the  $\xi_i$ , i.e.  $\pi(\boldsymbol{\xi}) = \prod_{i=1}^p \pi_i(\xi_i)$ . Then we can write

$$m(\mathbf{x}) = e^{l(\widehat{\boldsymbol{\theta}})} |\widehat{\boldsymbol{I}}|^{-1/2} \\ \times \left[ \prod_{i=1}^{p} \int \frac{1}{\sqrt{d_i}} e^{-((\xi_i - \widehat{\xi}_i)^2/2d_i)} \pi_i(\xi_i) \, \mathrm{d}\xi_i \right] \\ \times (1 + o(1)). \tag{8}$$

For  $\pi_i(\xi_i)$ , in a similar situation, Jeffreys (1961) recommended the Cauchy(0,  $b_i$ ) density  $(1/\pi \sqrt{b_i})(1/(1 + \xi_i^2/b_i))$ , where  $b_i$  is chosen to represent *unit information* for  $\xi_i$  (see Kass & Wasserman, 1995; also to be discussed later). A prior that yields almost the same results is

$$\pi_i^R(\xi_i) = \int_0^1 N\left(\xi_i \Big| 0, \frac{1}{2\lambda_i}(d_i + b_i) - d_i\right) \frac{1}{2\sqrt{\lambda_i}} d\lambda_i,\tag{9}$$

which is well-defined if  $b_i \ge d_i$ . Interestingly, this prior is very similar to the Cauchy prior no matter what  $d_i$ happens to be (as shown in the Appendix), so we will interpret this prior (and  $b_i$ ) exactly as we would with the Cauchy prior. The attraction of  $\pi^R$  is that the ensuing computations can be done in closed form. That one can have all the advantages that Jeffreys pointed out are possessed by the Cauchy prior for model selection, while maintaining closed form expressions, is a significant advantage when dealing with large model spaces. This prior was extensively discussed in Berger (1985), as a robust prior (hence the R label) for estimation problems, but its even greater value for model selection was not recognised. (This type of prior was first utilised in Strawderman (1971) in shrinkage estimation.) See also Bayarri et al. (2012), where a multivariate version of this prior is utilised for model selection in normal linear models.

With the prior in (9), the integral in (8) is straightforward to evaluate in closed form (first integrate over  $\xi_i$ , then over  $\lambda_i$ ) yielding

$$m(\mathbf{x}) = e^{l(\theta)} |\widehat{\mathbf{I}}|^{-1/2} \\ \times \left[ \prod_{i=1}^{p} \frac{1}{\sqrt{(d_i + b_i)}} \frac{\left(1 - e^{-\widehat{\xi}_i^2 / (d_i + b_i)}\right)}{\sqrt{2\widehat{\xi}_i^2 / (d_i + b_i)}} \right] \\ \times (1 + o(1)).$$
(10)

*Step 3.* Define the *unit information*,  $b_i$ , by

$$b_{i} = n_{i}^{e} d_{i}, \text{ where}$$

$$n_{i}^{e} = \text{effective sample size for } \xi_{i} \text{ and recall}$$

$$v_{i} = \frac{\widehat{\xi}_{i}^{2}}{d_{i}(1+n_{i}^{e})}. \quad (11)$$

Definitions of the effective sample size will be given in Section 3. It will be the case that  $n_i^e \ge 1$  so that  $b_i \ge d_i$  (the condition mentioned earlier for  $\pi^R$  to be well defined). Then (10) becomes

$$m(\mathbf{x}) = e^{l(\widehat{\theta})} \frac{|\widehat{I}|^{-1/2}}{|\mathbf{D}|^{1/2}} \prod_{i=1}^{p}$$

$$\times \frac{1}{\sqrt{1+n_i^e}} \frac{(1-e^{-v_i})}{\sqrt{2}v_i} (1+o(1)).$$
Since  $|\widehat{I}| = |\mathbf{\Sigma}^{-1}| |\widehat{I}_{22}| = |\mathbf{D}^{-1}| |\widehat{I}_{22}|$ , we thus have that
$$-2\log m(\mathbf{x}) = \text{PBIC} + o(1),$$

with PBIC defined in (4).

# 3. Defining 'effective sample size' $n_j$ , for parameter $\xi_j$

The most difficult aspect of dealing with PBIC turns out to be defining the effective sample size corresponding to a parameter. We first present a solution for linear models, and then suggest a possible solution for the general case.

### 3.1. Effective sample sizes in linear models

Suppose that all models under consideration are linear models of the form

$$Y = X^* \alpha + X \beta + \varepsilon, \text{ where } \varepsilon \sim N(0, \Gamma), \Gamma \text{ known,}$$
(12)  
with dimensions  $Y_{[n \times 1]}, X^*_{[n \times q]}, \alpha_{[q \times 1]}, X_{[n \times p]},$   
 $\beta_{[p \times 1]}, \varepsilon_{[n \times 1]} \text{ and } \Gamma_{[n \times n]}.$  Here  $X^* \alpha$  is a common term  
present in all models (e.g. an intercept in linear regres-  
sion), but  $X\beta$  will differ from model to model. This fits  
into the framework for PBIC by defining  $\theta_{(1)} = \beta$  and  
 $\theta_{(2)} = \alpha.$ 

Since  $\alpha$  will be integrated out in PBIC, only the effective sample size for linear functions of  $\beta$  will be needed. The first step of the process is to orthogonalise the parameters by transforming  $\alpha$  to  $\alpha^* = \alpha + (X^{*t}\Gamma^{-1}X^*)^{-1}X^{*t}\Gamma^{-1}X\beta$  and defining

$$\widetilde{X} = (I - X^* (X^{*t} \Gamma^{-1} X^*)^{-1} X^{*t} \Gamma^{-1}) X.$$
(13)

Since  $X^*\alpha + X\beta = X^*\alpha^* + \tilde{X}\beta$ , the linear part of the model has not changed in this reparameterisation, but now  $\tilde{X}^t \Gamma^{-1} X^* = \mathbf{0}$ , so that  $\alpha^*$  and  $\beta$  are orthogonal. There are two important aspects of this. First, since  $X^*$  has not been altered, the new  $\alpha^*$  can still be considered common parameters in each model, and will be integrated out in PBIC, so that their changed definition is irrelevant. Second,  $\beta$  has not been transformed, crucial because we wish effective sample sizes for linear functions of  $\beta$ 

Write  $\Gamma = \sigma R \sigma$ , with  $\sigma = \text{diag}\{\sigma_1, \dots, \sigma_p\}$ , where R is the correlation matrix, and define  $C_{[p \times p]}$  to be

the diagonal matrix with entries  $c_{ii} = \max_j \{ |\tilde{X}_{ji}| / \sigma_j \}$ . Berger, Bayarri, and Pericchi (2014) gave, as the general definition of the effective sample size (called TESS), for any *scalar* linear transformation  $\xi = v\beta$  (v is  $[1 \times p]$ ) of  $\beta$ ,

$$n^{e} = \frac{|\boldsymbol{v}|^{2}}{\boldsymbol{v}\boldsymbol{C}(\widetilde{\boldsymbol{X}}^{t}\boldsymbol{\Gamma}^{-1}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{C}\boldsymbol{v}^{t}}.$$
 (14)

**Example 3.1 (Group means example):** Assume  $Y_{ij} = \mu_i + \varepsilon_{ij}$  for i = 1, ..., p groups, and  $j = 1, ..., r_i$  replicates in the *i*th group, and that the  $\varepsilon_{ij}$  are i.i.d.  $N(0, \sigma^2)$ . Computation yields that TESS for  $\mu_i$  is  $n_i^e = r_i$ , as is to be expected. Note that  $r_i$  could be 1, which can be seen to be the lower bound on TESS for linear models when  $\Gamma = \sigma^2 I$ .

### Example 3.2 (Orthogonal and related designs):

Assume that *X* has orthogonal columns with entries  $\pm a_i \neq 0$ , and that  $\Gamma = \sigma^2 I$ . Simple computation here shows that  $n_i^e = n$  for each  $\beta_i$ .

Note that the effective sample size here is *n*, in contrast to the group means problem where the effective sample size can be as low as  $r_i = 1$ . Indeed, it can be shown that, when  $\Gamma = \sigma^2 I$ , TESS will always be between 1 and *n*, with both limits attainable.

**Example 3.3 (Heteroscedastic independent observations):** Assume  $Y_i = \mu + \varepsilon_i$ ,  $\varepsilon_i$  independent,  $\varepsilon_i \sim N(0, \sigma_i^2)$ , i = 1, ..., n. Here the effective sample size is

$$n^e = \frac{\sum_{i=1}^n 1/\sigma_i^2}{\max_i \{1/\sigma_i^2\}}$$

Consider the particular case where, for  $i = 1, ..., n_1$ , we have  $Y_i \sim N(\mu, \sigma_1^2)$ , whereas for the remaining  $n_2 = n - n_1$  observations,  $Y_i \sim N(\mu, \sigma_2^2)$ , where  $\sigma_2^2$  is much larger than  $\sigma_1^2$ ; thus, intuitively, only the first  $n_1$ observations count. Then, unless  $n_2$  is large,

$$n^{e} = \frac{n_{1}/\sigma_{1}^{2} + n_{2}/\sigma_{2}^{2}}{1/\sigma_{1}^{2}} = n_{1} + n_{2}\frac{\sigma_{1}^{2}}{\sigma_{2}^{2}} \approx n_{1}.$$

### 3.2. A general definition of effective sample size

Suppose one has independent observations  $(x_1, ..., x_n)$ . A possible general definition for the 'effective sample size' follows from considering the information associated with observation  $x_i$  arising from the single-observation expected information matrix  $I_i^* = O'(I_{i,jk}^*)O$ , where

$$I_{i,jk}^* = -E\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_i(\boldsymbol{x}_i \mid \boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}}$$

Since  $I_{jj}^* = \sum_{i=1}^n I_{i,jj}^*$  is the expected information about  $\xi_j$ , a reasonable way to define the effective sample size,  $n_j^e$ , is

- define information weights  $w_{ij} = I_{i,jj}^* / \sum_{k=1}^n I_{k,jj}^*$ ;
- define the effective sample size for  $\xi_i$  as

$$n_{j}^{e} = \frac{I_{jj}^{*}}{\sum_{i=1}^{n} w_{ij}I_{i,jj}^{*}} = \frac{\left(I_{jj}^{*}\right)^{2}}{\sum_{i=1}^{n} \left(I_{i,jj}^{*}\right)^{2}}$$

Intuitively,  $\sum w_{ij}I_{i,jj}^*$  is a weighted measure of the information 'per observation', and dividing the total information about  $\xi_j$  by this information per case seems plausible as an effective sample size.

Unfortunately, this does not seem to always be an effective definition; for instance, it does not reduce to TESS for all linear models. This should thus be viewed as primarily a starting point for future investigations of effective sample size in nonlinear models.

# 4. PBIC\*: a version more favourable to complex models

Recall, from Raftery (1999), that BIC can be thought of as arising from unit information priors for each model that are centred at the model likelihood. This choice of prior seems highly favourable to more complex models, since the prior gives virtually all of its mass to a modest neighbourhood of the likelihood for each model.

In contrast, PBIC utilises unit information priors that are centred at 0 and, hence, can give little mass to the region of high model likelihood. The fat tails of the prior do result in reasonable answers (cf. Bayarri et al., 2012; Jeffreys, 1961), but it is of interest to investigate an intermediate solution.

The intermediate solution is to keep the prior centred at 0, but choose the scales of the prior,  $b_i$ , so that the prior will extend out to the likelihood. In our setup, this can be implemented by choosing the  $b_i$  so as to maximise  $m(\mathbf{x})$  in (10); thus we are effectively choosing the prior in our class that is most favourable to each model. Clearly this does allow the prior to give more mass to the region of high model likelihood, but does not allow complete concentration of mass in this region.

Since this prior is maximising the marginal likelihood among the given class, it can be viewed as the empirical Bayes prior in the class. It was also a choice popularised in the 'robust Bayes' literature (cf. Berger & Sellke, 1987), and was used in Bollen et al. (2012) to develop a related generalisation of BIC.

The  $b_i$  that maximises (10) can easily be seen to be

$$\widehat{b}_i = \max\{d_i, \frac{\widehat{\xi}_i^2}{w} - d_i\}, \text{ with } w \text{ s.t. } e^w - 1$$
$$= 2w, \text{ or } w \approx 1.3.$$

Unfortunately, the resulting version of BIC has serious problems; in particular it will typically not be consistent as  $n \to \infty$  in that, if  $\xi_i$  is zero, the prior will concentrate about zero at such a fast rate that the models

with and without  $\xi_i$  are essentially equivalent (and one will fail to select the model without  $\xi_i$  with probability approaching 1). This same lack of consistency afflicts the developments in Bollen et al. (2012) and the robust Bayesian choices.

The obvious solution is simply to prevent  $\hat{b}_i$  from becoming too small, and the obvious constraint is to restrict it to the region  $[n_i^e d_i, \infty)$ . This yields the recommended choice

$$b_i^* = \max\left\{n_i^e d_i, \frac{\hat{\xi}_i^2}{1.3} - d_i\right\}.$$
 (15)

This will avoid inconsistency as  $n \to \infty$  in that, as long as  $b_i^* \to c$  with c a non-zero constant, the resulting prior behaves asymptotically when  $\xi_i = 0$  as a fixed prior, and fixed priors will yield consistency as  $n \to \infty$ . (Consistency when the effective sample sizes do not grow is a more delicate issue, discussed in Section 5.5.)

Replacing  $b_i$  with  $b_i^*$ , (10) becomes

$$\begin{split} m(\mathbf{x}) &= e^{l(\widehat{\theta})} |\widehat{\mathbf{I}}|^{-1/2} \left[ \prod_{i=1}^{p} \frac{1}{\sqrt{d_i(1+n_i^e) \max\{1, v_i/1.3\}}} \right] \\ &\times \frac{(1-e^{-\min\{v_i,1.3\}})}{\sqrt{2}\min\{v_i, 1.3\}} \right] (1+o(1)) \\ &= e^{l(\widehat{\theta})} |\widehat{\mathbf{I}}|^{-1/2} \left[ \prod_{i=1}^{p} \frac{1}{\sqrt{d_i(1+n_i^e)}} \right] \\ &\times \frac{(1-e^{-\min\{v_i,1.3\}})}{\sqrt{2v_i \min\{v_i, 1.3\}}} \right] (1+o(1)). \end{split}$$

The resulting approximation to  $-2\log m(\mathbf{x})$  is given in (6).

### 5. PBIC and PBIC\* for the linear model

### 5.1. The expressions

Consider the normal linear model framework in (12) and assume the orthogonalisation discussed there has been carried out. This does not change PBIC, but is more convenient because we can ignore the common orthogonal parameter  $\alpha^*$  (see Bayarri et al., 2012; Berger et al., 1998; Jeffreys, 1961 for justification), and focus only on the other parameters  $\beta$ , with the associated model

$$Y = \widetilde{X}\beta + \varepsilon$$
, where  $\varepsilon \sim N(0, \Gamma)$ ,  $\Gamma$  known,  
(16)

with  $\tilde{X}$  given by (13).

Following the PBIC algorithm, note that  $\Sigma^{-1} = \widetilde{X}' \Gamma^{-1} \widetilde{X}$ . Change variables to  $\xi = O\beta$ , where O is an orthogonal matrix such that  $\Sigma = O^t DO$ , with  $D = \text{diag}(d_i)$  for i = 1, ..., p. Then, for each  $\xi_j = O_j\beta$ , define  $n_j^e$  using (14) with  $v = O_j$ , and let  $\hat{\xi}_j = O_j\hat{\beta}$ , where  $\hat{\beta} = (\widetilde{X}' \Gamma^{-1} \widetilde{X})^{-1} \widetilde{X}' \Gamma^{-1} Y$ . Finally, recalling that

 $v_i = \widehat{\xi}_i^2 / [d_i(1 + n_i^e)]$ , PBIC and PBIC\* are given by

PBIC = 
$$S^2 + C + \sum_{i=1}^{p} \log(1 + n_i^e)$$
  
-  $2\sum_{i=1}^{p} \log \frac{(1 - e^{-v_i})}{\sqrt{2}v_i}$  (17)

PBIC\* = 
$$S^2 + C + \sum_{i=1}^{p} \log(1 + n_i^e)$$
  
-  $2\sum_{i=1}^{p} \log \frac{(1 - e^{-\min\{v_i, 1.3\}})}{\sqrt{2v_i \min\{v_i, 1.3\}}},$  (18)

where  $S^2$  is the usual residual sum of squares corresponding to (12) and

$$C = \log(|\mathbf{\Gamma}|) + n\log(2\pi) = \log(|\mathbf{\Gamma}|) + n\log(2\pi).$$

Note that *C* is the same constant in any model under consideration, and hence it can be ignored in comparing models or determining Bayes factors.

In what follows we describe some important Linear Model examples. There are more, including correlated observations and autoregressive models, in Berger et al. (2014).

### 5.2. Simple linear regression

Let 
$$Y_i = \alpha + X_i \beta + \epsilon_i$$
,  $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ , so that  

$$Y = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$+ \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
, where  $\varepsilon \sim N(0, \sigma^2 I)$ .

Suppose we are considering two models  $M_0: \beta = 0$ and  $M_1: \beta \neq 0$ . Computation under  $M_1$  yields  $\widetilde{X} = (X_1 - \overline{X}, \dots, X_n - \overline{X})'$ , so that  $\Sigma = \sigma^2 / s_x^2 = \sigma^2 / \sum_{i=1}^n (X_i - \overline{X})^2$ . Also, from (14),

$$n^{e} = \frac{\sum_{i=1}^{n} (X_{i} - \overline{X})^{2}}{\max_{i} \{ (X_{i} - \overline{X})^{2} \}} = \frac{s_{x}^{2}}{\max_{i} \{ (X_{i} - \overline{X})^{2} \}}.$$
 (19)

Finally,  $d = \Sigma = \sigma^2 / s_x^2$ ,  $v = \hat{\beta}^2 / [d(1 + n^e)]$ , and

$$S^{2} = \frac{1}{\sigma^{2}} \left( |\mathbf{Y}|^{2} - \frac{\left(\sum_{i=1}^{n} (x_{i} - \bar{x})y_{i}\right)^{2}}{\sum_{i=1}^{n} (x_{i} - \bar{x})^{2}} \right)$$
$$= \frac{1}{\sigma^{2}} (|\mathbf{Y}|^{2} - s_{x}^{2}\widehat{\beta}^{2})$$

complete the terms needed to define PBIC and PBIC\* under  $M_1$ . Under  $M_0$ , we only need  $S^2 = (1/\sigma^2)|\mathbf{Y}|^2$ ; thus, with  $v = \hat{\beta}^2 / [\sigma^2 (s_x^{-2} + (\max_i \{ (X_i - \overline{X})^2 \})^{-1})],$ 

$$\Delta \text{PBIC} = -\frac{s_x^2 \beta^2}{\sigma^2} + \log\left(1 + \frac{s_x^2}{\max_i\{(X_i - \overline{X})^2\}}\right)$$
$$-2\log\frac{(1 - e^{-\nu})}{\sqrt{2\nu}}.$$

 $\triangle$ PBIC<sup>\*</sup> is the obvious modification of this.

## 5.3. Testing equality of two means with unequal variances

Consider comparing two normal means via the test  $H_0: \mu_1 = \mu_2$  versus  $H_1: \mu_1 \neq \mu_2$ , where the associated known variances,  $\sigma_1^2$  and  $\sigma_2^2$  are not equal. The linear model is thus

$$Y = X\mu + \varepsilon = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix},$$
$$\times \varepsilon \sim N(\mathbf{0}, \operatorname{diag}\{\underbrace{\sigma_1^2, \dots, \sigma_1^2}_{n_1}, \underbrace{\sigma_2^2, \dots, \sigma_2^2}_{n_2}\}).$$

Defining  $\alpha = (\mu_1 + \mu_2)/2$  and  $\beta = (\mu_1 - \mu_2)/2$ places this in the linear model comparison framework, where we are comparing  $M_0: \beta = 0$  versus  $M_1: \beta \neq 0$ with the covariate matrix

$$B = X \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}.$$

Under  $M_1$ , computation yields

$$\widetilde{X} = \left(\frac{n_2}{n^* \sigma_2^2}, \dots, \frac{n_2}{n^* \sigma_2^2}, -\frac{n_1}{n^* \sigma_1^2}, \dots, -\frac{n_1}{n^* \sigma_1^2}\right)' \text{ with}$$
$$n^* = \left(\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}\right)^{-1},$$

so that

$$d = \Sigma = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Also, from (14),

$$n^{e} = \frac{\left(\frac{\sigma_{1}^{2}}{n_{1}} + \frac{\sigma_{2}^{2}}{n_{2}}\right)}{\max\left\{\sigma_{1}^{2}/n_{1}^{2}, \sigma_{2}^{2}/n_{2}^{2}\right\}}$$
$$= \min\left\{\frac{n_{1}^{2}}{\sigma_{1}^{2}}, \frac{n_{2}^{2}}{\sigma_{2}^{2}}\right\}\left(\frac{\sigma_{1}^{2}}{n_{1}} + \frac{\sigma_{2}^{2}}{n_{2}}\right),$$

and  $v = \hat{\beta}^2 / [d(1 + n^e)]$ .

A special case is the standard test of equality of means when  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . Then

$$n^{e} = \min\left\{n_{1}\left(1+\frac{n_{1}}{n_{2}}\right), n_{2}\left(1+\frac{n_{2}}{n_{1}}\right)\right\}.$$

While this may look unusual, looking at the extremes indicates why this is reasonable. Indeed, as say  $n_1 \rightarrow \infty$ , note that  $n^e \rightarrow n_2$ . In this scenario, we perfectly learn  $\mu_1$ , so the test of mean equality is really just a test that  $\mu_2$  equals this known mean, based on  $n_2$  observations. Attempting to utilise BIC with an adhoc choice of n, such as  $(n_1 + n_2)/2$ , would clearly be a disaster here.

### 5.4. Findley's counterexample to BIC

For the simple linear model in (2), computation yields that, under  $H_1: \theta \neq 0$ ,

$$d = \Sigma = \left(\sum_{i=1}^{n} \frac{1}{i}\right)^{-1}, \quad n^{e} = \sum_{i=1}^{n} \frac{1}{i},$$
$$S^{2} = |Y|^{2} - \hat{\theta}^{2} \sum_{i=1}^{n} \frac{1}{i}.$$

It follows that

$$\Delta \text{PBIC} = -\widehat{\theta}^2 \sum_{i=1}^n \frac{1}{i} + \log(1 + \sum_{i=1}^n \frac{1}{i})$$
$$-2\log\frac{\left(1 - e^{-v}\right)}{\sqrt{2}v}, \quad v = \frac{\widehat{\theta}^2}{d(1 + n^e)}$$

Since  $\sum_{i=1}^{n} (1/i) = \log n + O(1)$  and  $\widehat{\theta}^2 \to \theta^2$  (because the mle is consistent here),

$$\Delta \text{PBIC} = -\theta^2 (\log n + O(1)) + \log(\log n)$$
$$-2\log \frac{\left(1 - e^{-\theta^2}\right)}{\sqrt{2}\theta^2} + o(1).$$

Under  $H_0: \theta = 0$ ,  $\triangle PBIC = \log(\log n) + \log 2 + o(1)$  $\rightarrow \infty$  and, under  $H_1: \theta \neq 0$ ,  $\triangle PBIC \rightarrow -\infty$ . Thus PBIC is consistent as  $n \rightarrow \infty$ . Essentially the same argument shows that PBIC\* is consistent.

## 5.5. Consistency of PBIC and PBIC\* as $p \to \infty$ in the group means problem

Bayes model selection rules for fixed priors and fixed p are virtually always consistent as the sample size  $n \rightarrow \infty$ . This type of consistency transfers over to rules such as PBIC and PBIC\* because the priors from which they arise converge to fixed priors as  $n \rightarrow \infty$  with p fixed.

There is nothing within Bayesian theory, however, that guarantees consistency of Bayes rules when the dimension *p* also grows. Indeed, it turns out that consistency is then a very delicate property, that can easily be

violated by even standard Bayes rules. The group means problem provides a simple illustration.

**Example 5.1:** Consider the group means problem with known  $\sigma^2 = 1$  and effective sample size  $n_i = r$  fixed, and reduce to the sufficient statistics  $\bar{X}_i \sim N(\mu_i, 1/r)$  for i = 1, ..., p. Consider comparison of the null model  $M_0: \mu_1 = \cdots = \mu_p = 0$  with the full model  $M_1:$  all  $\mu_i$  non-zero. Suppose the  $\mu_i$  are independently assigned  $N(0, \tau_i^2)$  priors. Then it is easy to show that consistency obtains under  $M_1$  as  $p \to \infty$ if and only if  $V \equiv \lim_{p\to\infty} (1/p) \sum_i^p \bar{X}_i^2$  satisfies  $V \ge \lim_{p\to\infty} (1/p) \sum_i^p \tau_i^2$ , assuming the limits exist. (This example was brought to our attention by J. K. Ghosh.)

After reflecting upon this, it might seem surprising that any prior could achieve consistency as  $p \rightarrow \infty$ . However, Berger et al. (2003) computed Laplace approximations to the marginal density, for this problem, that produced consistent Bayes factors when pgrows with n. They used a multivariate Cauchy prior, which does not result in a closed form Bayes factor, as arises with PBIC and PBIC<sup>\*</sup>. The next theorem indicates the situation involving consistency for PBIC and PBIC<sup>\*</sup>.

**Theorem 5.2:** For the group means problem with fixed r, PBIC and PBIC<sup>\*</sup> are consistent under  $M_0$  as  $p \to \infty$ . Under  $M_1$  and assuming that  $\tau^2 = \lim_{p\to\infty} (1/p) \sum_{i}^{p} \mu_i^2$  exists, PBIC and PBIC<sup>\*</sup> are

$$consistent if \tau^{2}$$

$$> \frac{\log 2 + \log(1+r) + 1}{r}; \quad inconsistent if \tau^{2}$$

$$< \frac{\log 2 + \log(1+r) - 1}{r}. \quad (20)$$

**Proof:** We utilise (17) and (18) as the definitions of PBIC and PBIC\*, but will ignore *C* since it is common to all models. Note that the  $n_i^e = r$ ,  $S_1^2 = \sum_{i=1}^p \sum_{j=1}^r (x_{ij} - \bar{x}_i)^2$ ,  $S_0^2 = S_1^2 + r \sum_{i=1}^p \bar{x}_i^2$ ,  $v_i = r\bar{x}_i^2/(r+1)$  under  $M_1$  and  $v_i = 0$  under  $M_0$ . Thus PBIC and PBIC\* become, with subscripts denoting the model,

$$\begin{aligned} \text{PBIC}_0 &= \text{PBIC}_0^* = S_0^2 = S_1^2 + r \sum_{i=1}^p \bar{x}_i^2, \\ \text{PBIC}_1 &= S_1^2 + p \log(r+1) - 2 \sum_{i=1}^p \log \frac{1 - e^{-v_i}}{\sqrt{2} v_i} \\ &= S_1^2 + p \log[2(r+1)] - 2 \sum_{i=1}^p \log \frac{1 - e^{-v_i}}{v_i}, \\ \text{PBIC}_{1}^* &= S_1^2 + p \log(r+1) \\ &- 2 \sum_{i=1}^p \log \frac{\left(1 - e^{-\min\{v_i, 1.3\}}\right)}{\sqrt{2 v_i \min\{v_i, 1.3\}}} \end{aligned}$$

$$= S_1^2 + p \log[2(r+1)] \\ - 2 \sum_{i=1}^p \log \frac{\left(1 - e^{-\min\{v_i, 1.3\}}\right)}{\sqrt{v_i \min\{v_i, 1.3\}}}$$

It is straightforward to show that

$$\frac{1 - e^{-v_i}}{v_i} < 1 \quad \text{and} \quad \frac{\left(1 - e^{-\min\{v_i, 1.3\}}\right)}{\sqrt{v_i \,\min\{v_i, 1.3\}}} < 1,$$

so that  $\triangle PBIC = PBIC_1 - PBIC_0$  and  $\triangle PBIC^* = PBIC^*_1 - PBIC^*_0$  satisfy

$$\Delta PBIC (\Delta PBIC^*)$$

> 
$$p \log[2(r+1)] - r \sum_{i=1}^{p} \bar{x}_i^2 \equiv A(p).$$

Under  $M_0$ ,  $r \sum_{i=1}^p \bar{x}_i^2 \sim \chi_p^2$ , so that

$$A(p) = p \log[2(r+1)]$$
$$- p \left(1 + O\left(\frac{1}{\sqrt{p}}\right)\right) \to \infty \text{ as } p \to \infty,$$

establishing consistency under  $M_0$ .

To show inconsistency under  $M_1$ , note that  $r \sum_{i=1}^{p} \bar{x}_i^2 \sim \chi_p^2(\lambda_p)$ , with non-centrality parameter  $\lambda_p = r \sum_{i=1}^{p} \mu_i^2$ . Thus

$$A(p) = p \log[2(r+1)] - (p+\lambda_p) \left(1 + O\left(\frac{1}{\sqrt{p+\lambda_p}}\right)\right) \to \infty$$

if  $\tau^2 = \lim_{p \to \infty} \lambda_p / [rp] < (\log[2(1+r) - 1])/r$ , establishing the inconsistency result.

To investigate consistency of PBIC and PBIC<sup>\*</sup> under  $M_1$ , note that

$$\frac{\left(1 - e^{-\min\{v, 1.3\}}\right)}{\sqrt{v\,\min\{v, 1.3\}}} \ge \frac{1 - e^{-v}}{v} \ge \frac{1}{1 + v}.$$
 (21)

Also, because of concavity,  $E[\log(1 + v)] \le \log(1 + E[v])$ . Thus,

$$E\left[\log\frac{1-e^{-v_i}}{v_i}\right] \ge -E[\log(1+v_i)]$$
$$\ge -\log(1+E[v_i])$$
$$= -\log\left(1+\frac{1+r\mu_i^2}{1+r}\right)$$

Using this inequality and the fact that  $\prod_i \omega_i^{1/p} \le (\sum \omega_i)/p$ , it follows that

$$\frac{2}{p} E\left[\sum_{i=1}^{p} \log \frac{1-e^{-v_i}}{v_i}\right]$$

$$\geq -2 \log \prod_{i=1}^{p} \left(1 + \frac{1+r\mu_i^2}{1+r}\right)^{1/p}$$

$$\geq -2 \log \left[\frac{1}{p} \sum_{i=1}^{p} \left(1 + \frac{1+r\mu_i^2}{1+r}\right)\right]$$

$$= -2 \log \left(\frac{2+r+\lambda_p/p}{1+r}\right).$$

Hence, by the law of large numbers,

$$\begin{split} \lim_{p \to \infty} \frac{1}{p} \Delta \text{PBIC} \\ &\leq \log[2(r+1)] + 2\log\left(\frac{2+r+r\tau^2}{1+r}\right) \\ &- \lim_{p \to \infty} \left(1 + \frac{\lambda_p}{p}\right) \left(1 + O\left(\frac{1}{\sqrt{p+\lambda}}\right)\right) \\ &= \log[2(r+1)] + 2\log\left(\frac{2+r+r\tau^2}{1+r}\right) \\ &- (1+r\tau^2)(1+o(1)). \end{split}$$

Let  $B(r, \tau^2)$  denote the right hand side above (without the o(1) term). If  $B(r, \tau^2) < 0$ , then  $\Delta$ PBIC goes to  $-\infty$  as  $p \to \infty$ , and we have consistency.

Differentiating with respect to  $\tau^2$  shows that  $B(r, \tau^2)$  is decreasing in  $\tau^2$ , so that, if we can find a value of  $\tau^2$  for which  $B(r, \tau^2) < 0$ , then any larger value of  $\tau^2$  will also work. As a candidate, consider  $\tau_c^2 = [c + \log(1 + r)]/r$ . Then

$$B(r, \tau_c^2) = \log[2(r+1)] + 2\log\left(\frac{2+r+c+\log(1+r)}{1+r}\right) - (1+c+\log(1+r)).$$

Differentiating this with respect to *r* shows that it is decreasing in *r* so that all we need to show is that  $\tau_c^2$  works for *r* = 1. Indeed,

$$B(1, \tau_c^2) = \log[4] + 2\log\left(\frac{3 + c + \log 2}{2}\right) - (1 + c + \log 2) < 0,$$

if c > 1.67. Since  $1 + \log 2 = 1.693 > 1.67$ , the condition for consistency of PBIC in the theorem is established. And because of (21), the same condition ensures that PBIC\* is consistent.

Note that, if *r* is moderately large, PBIC and PBIC\* are consistent under  $M_1$ , unless  $\tau^2$  is extremely close to 0, i.e. unless the non-zero means are extremely close to 0; it is not surprising that it is difficult to distinguish between  $M_1$  and  $M_0$  in this situation.

There is a gap in the theorem between the consistency and inconsistency conditions under  $M_1$ . The gap is quite large for small r, but shrinks as r grows. A more refined analysis would reduce the gap, but the theorem does convey the basic messages about consistency.

More generally,  $M_0$  could be a group means model containing some zero and some non-zero means. If  $M_0$ is nested in  $M_1$  and the number of additional nonzero means in  $M_1$  goes to  $\infty$ , then the theorem still applies, since the common non-zero means will be integrated out at the beginning and will not affect the analysis.

### Acknowledgements

This research was begun under the auspices of the 2004–2005 SAMSI program on Latent Variables in the Social Sciences.

### **Disclosure statement**

No potential conflict of interest was reported by the authors.

### Funding

M. J. Bayarri's research was supported by the Spanish Ministry of Education and Science [grant number MTM2010-19528]; James Berger's research was supported by USA National Science Foundation [grant numbers DMS-1007773 and DMS-1407775]; Woncheol Jang's research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIP), No. 2014R1A4A1007895 and No. 2017R1A2B2012816; Luis Pericchi's research was supported by grant CA096297/CA096300 from the USA National Cancer Institute of the National Institutes of Health.

### References

- Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), 1550–1577.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag.
- Berger, J. O., Bayarri, M. J., & Pericchi, L. R. (2014). The effective sample size. *Econometric Reviews*, 33, 197–217.
- Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations and consistency of bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, 112, 241–258.
- Berger, J. O., Pericchi, L. R., & Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. Sankhya: The Indian Journal of Statistics. Series A, 60, 307–321.
- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, *82*, 112–122.

- Bollen, K. A., Ray, S., Zavisca, J., & Harden, J. J. (2012). A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods and Research*, 41, 294–324.
- Chakrabarti, A., & Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *Journal of Statistical Planning and Inference*, 136(9), 2847–2872.
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 323–380.
- Dudley, R. M., & Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica*, *7*, 265–284.
- Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, 43, 505–514.
- Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In Advances in neural information processing systems (pp. 604–612).
- Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, *16*(1), 342–355.
- Haughton, D. (1991). Consistency of a class of information criteria for model selection in non-linear regression. *Communications in Statistics: Theory and Methods*, 20, 1619–1629.
- Haughton, D. (1993). Consistency of a class of information criteria for model selection in nonlinear regression. *Theory of Probability and its Applications*, *37*, 47–53.
- Haughton, D., Oud, J., & Jansen, R. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics, Part B – Simulation and Computation*, 26(4), 1477–1516.
- Jeffreys, H. (1961). *Theory of probability*. London: Oxford University Press.
- Kass, R. E., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
- Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society*, *54*, 129–144.
- Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.
- Pauler, D. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, *85*, 13–27.
- Raftery, A. E. (1999). Bayes factors and BIC comment on 'A critique of the Bayesian information criterion for model selection'. *Sociological Methods and Research*, 27, 411–427.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shen, G., & Ghosh, J. K. (2011). Developing a new BIC for detecting change-points. *Journal of Statistical Planning and Inference*, 141(4), 1436–1447.
- Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society*, *Series B*, 41, 276–278.
- Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1), 385–388.
- Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances

of nonpositive functions. *Journal of the American Statistical Association*, 84(407), 710–716.

Zak-Szatkowska, M., & Bogdan, M. (2011). Modified versions of Bayesian information criterion for sparse generalized linear models. *Computational Statistics and Data Analysis*, 55, 2908–2924.

### **Appendix**

To see that the prior in (9) is almost the same as  $\pi^{C}$ , the Cauchy(0,b) prior (we drop the *i* subscripts in this appendix), consider the extremes.

**Theorem A.1:** For  $b \ge d$ ,

$$\lim_{|\xi| \to \infty} \frac{\pi^{C}(\xi)}{\pi^{R}(\xi)} = \frac{2\sqrt{b}}{\sqrt{\pi(b+d)}} \in (0.80, 1.13),$$
$$\frac{\pi^{C}(0)}{\pi^{R}(0)} = \frac{2d}{\sqrt{b\pi}(\sqrt{b+d} - \sqrt{b-d})}$$
$$\in (0.80, 1.13).$$

**Proof:** Note that

$$\pi^{R}(0) = \frac{1}{2\sqrt{\pi}} \int_{0}^{1} \frac{1}{\sqrt{d+b-2\lambda d}}$$
$$d\lambda = \frac{\sqrt{b+d} - \sqrt{b-d}}{2d\sqrt{\pi}}.$$

Hence

$$\frac{\tau^C(0)}{\tau^R(0)} = \frac{2d}{\sqrt{b\pi}(\sqrt{b+d} - \sqrt{b-d})}.$$

It is straightforward to show that  $\sqrt{b}(\sqrt{b+d} - \sqrt{b-d})$  is decreasing in  $b \ge d$ , with a maximum of  $\sqrt{2}d$  and minimum of *d*. Thus  $\sqrt{2/\pi} \le \pi^C(0)/\pi^R(0) \le \sqrt{4/\pi}$ , which (to 2 decimal places) is the result above.

To prove the result as  $|\xi| \to \infty$ , separately integrate over  $\Gamma_1 = (0, |\xi|^{-3/2})$  and  $\Gamma_2 = (|\xi|^{-3/2}, 1)$  in (9). For  $\lambda \in \Gamma_1$ , note that  $(d + b - 2\lambda d)^{-1} = (d + b)^{-1} + O(|\xi|^{-3/2})$ , so that

$$(d+b-2\lambda d)^{-1/2} = (d+b)^{-1/2} + O(|\xi|^{-3/2}),$$
  

$$\exp\left(-\frac{\xi^2 \lambda}{d+b-2\lambda d}\right) = \exp\left(-\frac{\xi^2}{d+b}\left[\lambda + O(|\xi|^{-3})\right]\right)$$
  

$$= \exp\left(-\frac{\xi^2}{d+b}\right)\left(1 + O(|\xi|^{-1})\right).$$

Hence the integral over  $\Gamma_1$  is

$$\frac{1}{2\sqrt{\pi}} \int_0^{|\xi|^{-3/2}} \left( \frac{1}{\sqrt{d+b}} + O(|\xi|^{-3/2}) \right)$$
$$\times \exp\left(-\frac{\xi^2 \lambda}{d+b}\right) \left(1 + O(|\xi|^{-1}) d\lambda\right)$$
$$= \frac{\sqrt{d+b}}{2\sqrt{\pi}\xi^2} \left(1 - \exp\left(-\frac{\sqrt{|\xi|}}{d+b}\right)\right) \left(1 + O(|\xi|^{-1})\right).$$

It follows that

Noting that  $\exp(-\xi^2\lambda/[d+b-2\lambda d])$  is decreasing in  $\lambda$ , it is immediate that the integral over  $\Gamma_2$  is bounded above by

$$\frac{\exp\left(-\sqrt{|\xi|}/[d+b]\right)}{2\sqrt{\pi}d} \int_{|\xi|^{-3/2}}^{1} \frac{1}{\sqrt{d+b-2\lambda d}} \, d\lambda$$
$$= \frac{\exp\left(-\sqrt{|\xi|}/[d+b]\right)}{2\sqrt{\pi}d}$$
$$\times \left(\sqrt{d+b-2|\xi|^{-3/2}d} - \sqrt{b-d}\right) = o(|\xi|^{-2}).$$

$$\lim_{|\xi| \to \infty} \frac{\pi^{C}(\xi)}{\pi^{R}(\xi)} = \lim_{|\xi| \to \infty} \frac{2\sqrt{\pi} [\pi^{-1} \xi^{-2} \sqrt{b} (1 + O(|\xi^{-2}))]}{\sqrt{d + b} \xi^{-2} (1 - \exp\left(-\sqrt{|\xi|} / [d + b]\right))} (1 + O(|\xi^{-1})) + o(|\xi|^{-2}) = \frac{2\sqrt{b}}{\sqrt{\pi(b+d)}}.$$

It is straightforward to show that  $\sqrt{2/\pi} \le 2\sqrt{b}/\sqrt{\pi(b+d)} \le \sqrt{4/\pi}$ , yielding (to two decimal places) the conclusion.