

Divergence Based Priors for Bayesian Hypothesis testing

M.J. Bayarri
University of Valencia

G. García-Donato
University of Castilla-La Mancha

November, 2006

Abstract

Maybe the main difficulty for objective Bayesian hypothesis testing (and model selection in general), is that usual objective improper priors can not be used for parameters not occurring in all of the models. In this paper we introduce (objective) proper prior distributions for hypothesis testing and model selection based on measures of divergence between the competing models; we call them *divergence based* (DB) priors. DB priors have simple forms and desirable properties, like information (finite sample) consistency; often, they are similar to other existing proposals like the intrinsic priors; moreover, in normal linear models scenarios, they exactly reproduce Jeffreys-Zellner-Siow priors. Most importantly, in challenging scenarios such as irregular models and mixture models, the DB priors are well defined and very reasonable, while alternative proposals are not. We derive approximations to the DB priors as well as MCMC and asymptotic expressions for the associated Bayes factors, which also reveals interesting connections with other proposals (like the unit information priors).

1 Introduction

For the data \mathbf{y} , with density $\{f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu})\}$, we consider the hypothesis testing problem:

$$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \text{ vs. } H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \quad (1)$$

where $\boldsymbol{\theta}_0 \in \Theta$ is a known value. This is equivalent to the model selection problem of choosing between models:

$$M_1 : f_1(\mathbf{y} | \boldsymbol{\nu}_1) = f(\mathbf{y} | \boldsymbol{\theta}_0, \boldsymbol{\nu}_1) \text{ vs. } M_2 : f_2(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}_2) = f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}_2), \quad (2)$$

where the notation reflects the fact that often $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ represent different quantities in each model. In Jeffreys' scenarios (Jeffreys, 1961), $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ had the same meaning; he called $\boldsymbol{\theta}$ the

new parameter, and $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, the *common parameters* (also known as nuisance parameters). We revisit this issue in Section 4.

We aim for an *objective Bayes* solution to this model selection problem; that is, no ‘external’ (subjective) information is assumed, other than the data, \mathbf{y} , and the information implicitly needed to pose the problem, choose the competing models, etc. An excellent exposition of the advantages of Bayesian methods, specially objective Bayes methods, for problems with model uncertainty is Berger and Pericchi (2001).

Usual Bayesian solutions (for 0- k_i loss functions) to (1) (or, equivalently, to (2)) are based in the posterior odds:

$$\frac{\Pr(H_1 | \mathbf{y})}{\Pr(H_2 | \mathbf{y})} = \frac{\Pr(H_1)}{\Pr(H_2)} \times B_{12},$$

where $\Pr(H_i)$, $i = 1, 2$ are the prior probabilities of the hypotheses, and B_{12} is *Bayes Factor* for H_1 against H_2 :

$$B_{12} = \frac{m_1(\mathbf{y})}{m_2(\mathbf{y})} = \frac{\int f_1(\mathbf{y} | \boldsymbol{\nu}_1) \pi_1(\boldsymbol{\nu}_1) d\boldsymbol{\nu}_1}{\int f_2(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}_2) \pi_2(\boldsymbol{\theta}, \boldsymbol{\nu}_2) d\boldsymbol{\theta} d\boldsymbol{\nu}_2}, \quad (3)$$

where $\pi_1(\boldsymbol{\nu}_1)$ is the prior under H_1 and $\pi_2(\boldsymbol{\theta}, \boldsymbol{\nu}_2)$ the prior under H_2 . That is, B_{12} is the ratio of the marginal (averaged) likelihoods of the models.

It is common practice in objective Bayes approaches to concentrate on derivations of the Bayes factors, letting the ultimate choice (whether objective or subjective) of the prior model probabilities (and the derivations of the posterior odds) to the user. Bayes factors were extensively used by Jeffreys (1961) as a measure of evidence in favor of a model (see also Berger, 1985; Berger and Delampady, 1987, and Berger and Sellke, 1987); Kass and Raftery (1995) is a good reference for review and applications. Bayes factors are also crucial ingredients of model averaging approaches (see Clyde, 1999; Hoeting et al, 1999). In the rest of the paper, we concentrate on the derivation of objective priors to compute Bayes factors.

A main issue for deriving objective Bayes factors is appropriate choice of $\pi_1(\boldsymbol{\nu}_1)$ and $\pi_2(\boldsymbol{\theta}, \boldsymbol{\nu}_2)$ for use in (3). It is well known that familiar improper objective priors (or non-informative priors) for estimation problems (under a fixed model) are usually seriously inadequate in the presence of model uncertainty, generally producing arbitrary answers. (Interesting exceptions are studied in Berger, Pericchi and Varshavsky, 1998.) Of course, when improper priors can not be used, use of arbitrarily vague (but proper) priors is not a cure, and generally it is even worse. Another bad solution often encountered in practice is use of an apparently ‘innocuous’, harmless, but yet arbitrary, proper prior, since it can severely dominate the likelihood in ways that are not anticipated (and can not be investigated for high dimensional problems).

There are two basic approaches to compute Bayes factors when there is not enough information available for trustworthy subjective assessment of $\pi_1(\boldsymbol{\nu}_1)$ and $\pi_2(\boldsymbol{\theta}, \boldsymbol{\nu}_2)$. A very successful one is to directly derive the objective Bayes factors themselves, usually by ‘training’ and calibrating in several ways the non-appropriate Bayes factors obtained from usual objective improper

priors (see Berger and Pericchi, 2001 for reviews and references). However, all these objective Bayes factors should ultimately be checked to correspond (approximately) to a genuine Bayes factor derived from a sensible prior. The alternative approach is to look for ‘formal rules’ for constructing ‘objective’ but proper priors that have nice properties and are appropriate for using in model selection; Bayes factors are then just computed from these objective proper priors. Whether these Bayes factors are appropriate can then be directly judged from the adequacy of the priors used.

Choice of prior distributions in scenarios of model uncertainty is still largely an open question, and only partial answers are known. Several methods have been proposed for use in general scenarios, like the arithmetic intrinsic (AI) priors (Berger and Pericchi, 1996; Moreno, Bertolino and Racugno, 1998); the fractional intrinsic (FI) priors (De Santis and Spezaferrri, 1999; Berger and Mortera, 1999); the expected posterior (EP) priors (Pérez and Berger, 2002); the unit information priors (Kass and Wasserman, 1995) and the predictively matched priors (Ibrahim and Laud, 1994; Laud and Ibrahim, 1995; Berger, Pericchi and Varshavsky, 1998; Berger and Pericchi, 2001). In the specific context of linear models, widely used prior with nice properties are Jeffreys-Zellner-Siow (JZS) priors (Jeffreys, 1961; Zellner and Siow, 1980,1984; Bayarri and García-Donato, 2007). A very interesting generalization is the mixtures of g -priors (Liang et al., 2007).

In this paper we generalize an earlier suggestion by Jeffreys (1961), and use divergence measures between the competing models to derive the required (proper) priors. We call these priors *divergence based* (DB) priors. The main motivation was to generalize the useful JZS priors for use in scenarios other than the normal linear model. We will show that indeed the DB priors are the JZS priors in linear model contexts; also, they are as easy to derive (often easier) than other popular proposals (AI, FI or EP priors), being quite similar to them in many instances; most interestingly, they are well defined in certain scenarios where the other proposals fail.

For clarity of exposition, we consider first the case when there are no nuisance parameters. Development for the general case is delayed till Section 4, once the basic ideas have been introduced, and the behavior of DB priors studied in this considerably simpler scenario.

2 DB priors

Assume first the problem without nuisance parameters:

$$M_1 : f_1(\mathbf{y}) = f(\mathbf{y} \mid \boldsymbol{\theta}_0) \text{ vs. } M_2 : f_2(\mathbf{y} \mid \boldsymbol{\theta}) = f(\mathbf{y} \mid \boldsymbol{\theta}). \quad (4)$$

That is, the simpler model (M_1) involve no unknown parameters; hence only the prior for $\boldsymbol{\theta}$ under M_2 is needed. We drop the subindex in the previous section and denote such prior simply by $\pi(\boldsymbol{\theta})$; clearly $\pi(\boldsymbol{\theta})$ has to be proper.

Our proposal for DB priors for $\boldsymbol{\theta}$ will be in terms of divergence measures between the competing models $f(\mathbf{y} \mid \boldsymbol{\theta}_0)$ and $f(\mathbf{y} \mid \boldsymbol{\theta})$, based on Kullback-Leibler directed divergences

$$KL[\boldsymbol{\theta}_0 : \boldsymbol{\theta}] = \int [\log f(\mathbf{y} \mid \boldsymbol{\theta}) - \log f(\mathbf{y} \mid \boldsymbol{\theta}_0)] f(\mathbf{y} \mid \boldsymbol{\theta}) d\mathbf{y}, \quad (5)$$

(assuming continuous \mathbf{y} for simplicity). KL is a measure of the information in \mathbf{y} to discriminate between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$; it is designed to measure how far apart the two competing densities are in the sense of the likelihood (Schervish, 1995).

KL is not symmetric with respect to its arguments, but symmetric measures of divergence can be derived by taking sums or minimums of KL divergences. We define:

$$D^S[\boldsymbol{\theta}, \boldsymbol{\theta}_0] = KL[\boldsymbol{\theta} : \boldsymbol{\theta}_0] + KL[\boldsymbol{\theta}_0 : \boldsymbol{\theta}], \quad (6)$$

and

$$D^M[\boldsymbol{\theta}, \boldsymbol{\theta}_0] = 2 \times \min\{KL[\boldsymbol{\theta} : \boldsymbol{\theta}_0], KL[\boldsymbol{\theta}_0 : \boldsymbol{\theta}]\}. \quad (7)$$

Generalizations of KL , D^S and D^M to include marginal parameters are discussed in Section 4. Notice that D^M is well defined even when one of directed KL divergences is not, as when the competing models have different support. Except for these irregular scenarios, D^S is well defined and it is considerably easier to derive than D^M . Most of the derivations and properties to follow are common to both D^S and D^M . To avoid tedious repetitions, we then simply use D to refer to anyone of them. We use the superindex S or M only when necessary.

It is well known that $D \geq 0$ with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, although it is not a metric (the triangle inequality does not hold). Our proposal, is based on *unitary measures of divergence*, \bar{D} , which we take to be D divided by the *effective sample size* n^* , $\bar{D} = D/n^*$. In simple univariate i.i.d. data the effective sample equals the number of scalar data points, but it does not need to be so in general. Indeed, in complex situations, it can be a difficult concept; although there have been several attempts in the literature to formalize it (see e.g. Pauler, 1998; Pauler, Wakefield and Kass, 1999; Berger et al. 2007), no general agreed definition seems to exist. In all the examples of this paper, it is quite clear what n^* should be, so we rely for now in simple, intuitive interpretations.

2.1 Motivation: scalar location parameters

Suppose \mathbf{y} is a random sample from a univariate location family:

$$f(\mathbf{y} \mid \theta) = \prod_{i=1}^n f(y_i \mid \theta) = \prod_{i=1}^n g(y_i - \theta), \quad \theta \in \mathcal{R}.$$

It has been argued (Berger and Delampady 1987; Berger and Sellke 1987) that in symmetric problems with $\Theta = \mathcal{R}$, objective testing priors $\pi(\theta)$ under $H_2 : \theta \neq \theta_0$ should be unimodal and symmetric about θ_0 ; these priors prevent introducing excessive bias toward H_2 . Accordingly, we look for a proper $\pi(\theta)$ that has these desirable characteristics in this simple scenario and that is easily generalizable to other situations.

As before, let \bar{D} be a *unitary* symmetrized divergence. We consider use of a function, h of \bar{D} as a testing prior under H_2 ; that is $\pi(\theta) \propto h(\bar{D}[\theta, \theta_0])$. Since π has to be proper, $h(t)$ has to be a decreasing (no-increasing) function for $t > 0$. A first possibility could be to take $h(t) = \exp\{-qt\}$ for some $q > 0$, but this results in priors with short tails. Short-tailed priors are usually not adequate for model selection, since they tend to exhibit undesirable (finite sample) inconsistent behavior (see Liang et al 2007).

We explore instead use of the functions $h_q(t) = (1+t)^{-q}$, where $q > 0$ controls thickness of the tails of $\pi(\theta)$. Let

$$c(q) = \int h_q(\bar{D}[\theta, \theta_0]) d\theta = \int \left(1 + \frac{D[\theta, \theta_0]}{n^*}\right)^{-q} d\theta,$$

and define

$$\underline{q} = \inf\{q \geq 0 : c(q) < \infty\}, \quad q_* = \underline{q} + 1/2.$$

For finite \underline{q} , our specific proposal for a DB prior in this location problem is

$$\pi^D(\theta) = c(q_*)^{-1} \left(1 + \frac{D[\theta, \theta_0]}{n^*}\right)^{-q_*} \propto h_{q_*}(\bar{D}[\theta, \theta_0]). \quad (8)$$

Generalization to vector valued θ is trivial.

Of course, if \underline{q} is finite, any $q = \underline{q} + \delta$, with $\delta > 0$ results in proper priors, and hence could have been used to define a DB prior. Our specific proposal, $\delta = 1/2$ was chosen to reproduce the well known Jeffreys-Zellner-Siow prior in the Normal context; in general this choice results in densities with heavy tails. Moreover, we have found that in general $0 < \delta < 1$ produces priors without moments, which in normal scenarios is needed to avoid undesirable behavior of conjugate g priors (Liang et al, 2007). The following lemma establishes the desired symmetry and unimodality of the DB prior. The proof follows easily from properties of D in these location problems and is avoided.

Lemma 2.1. *Assume $\underline{q} < \infty$; then $\pi^D(\theta)$ is unimodal and symmetric around θ_0 .*

Definition of DB priors for scale parameters is also direct. Indeed assume that θ is a scale parameter for a positive random variable X ; then, $\xi = \log \theta$ is a location parameter for $Y = \log X$, with density $f^*(y | \xi)$. Applying the definition in (8), the DB prior for ξ is:

$$\pi^D(\xi) \propto h_{q_*}(\bar{D}^*[\xi, \xi_0]), \quad (9)$$

where $\xi_0 = \log(\theta_0)$ and $\bar{D}^*[\xi, \xi_0]$ is the unitary measure of divergence between $f^*(\mathbf{y} \mid \xi_0)$ and $f^*(\mathbf{y} \mid \xi)$. Therefore, in the original parameterization:

$$\pi^D(\theta) \propto h_{q_*}(\bar{D}^*[\log \theta, \log \theta_0]) \frac{1}{\theta} = h_{q_*}(\bar{D}[\theta, \theta_0]) \pi^N(\theta), \quad (10)$$

where, because of invariance of \bar{D} under reparameterizations, $\bar{D}^*[\log \theta, \log \theta_0] = \bar{D}[\theta, \theta_0]$, and $\pi^N(\theta) = 1/\theta$ is the non informative prior (right Haar invariant prior) for θ . Definition of DB priors for general parameters, formalized in next section, will basically be a generalization of (10).

2.2 General parameters

Assume the more general problem (4) and let $\pi^N(\boldsymbol{\theta})$ be an objective (usually improper) ‘estimation’ prior (reference, invariant, Jeffreys, Uniform, ... prior) for $\boldsymbol{\theta}$, and let $\boldsymbol{\xi}$ be a transformation such that $\pi^N(\boldsymbol{\xi}) = 1$ for $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$. We can then derive a DB prior for $\boldsymbol{\theta}$ by considering $\boldsymbol{\xi}$ as a ‘location parameter’, applying the definition (8), and transforming back to $\boldsymbol{\theta}$. This transformation was first proposed by Jeffreys (1961). Bernardo (2005) uses it with a reference prior π^N for a scalar θ , and notes that ξ asymptotically behaves as a location parameter.

Giving $\boldsymbol{\xi}$ a DB prior for location parameters results in:

$$\pi^D(\boldsymbol{\xi}) \propto h_{q_*}(\bar{D}^*[\boldsymbol{\xi}, \boldsymbol{\xi}_0]), \quad (11)$$

where, as before, $\bar{D}^*[\boldsymbol{\xi}, \boldsymbol{\xi}_0]$ denotes ‘unit’ (symmetrized) discrepancy between $f^*(\mathbf{y} \mid \boldsymbol{\xi})$ and $f^*(\mathbf{y} \mid \boldsymbol{\xi}_0)$, and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(\boldsymbol{\theta}_0)$. Hence, the corresponding (DB) prior for $\boldsymbol{\theta}$ is

$$\pi^D(\boldsymbol{\theta}) \propto h_{q_*}(\bar{D}^*[\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\xi}(\boldsymbol{\theta}_0)]) |\mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| \propto h_{q_*}(\bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]) \pi^N(\boldsymbol{\theta}), \quad (12)$$

as long as π^N is invariant under transformations; $\mathcal{J}(\boldsymbol{\theta})$ is the jacobian of the transformation. It should be noted from (12) that the explicit transformation to $\boldsymbol{\xi}$ is not needed in order to derive the prior π^D . We can now formally define a DB prior.

Definition 2.1. (General DB priors) *For the model selection problem (4), let $\bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]$ be a unitary measure of divergence between $f(\mathbf{y} \mid \boldsymbol{\theta})$ and $f(\mathbf{y} \mid \boldsymbol{\theta}_0)$. Also let $\pi^N(\boldsymbol{\theta})$ be an objective (possibly improper) prior for $\boldsymbol{\theta}$ under the complex model, M_2 , and $h_q(\cdot)$ be a decreasing function. Define:*

$$\underline{q} = \inf\{q \geq 0 : c(q) < \infty\}, \quad q_* = \underline{q} + 1/2,$$

where $c(q) = \int h_q(\bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]) \pi^N(\boldsymbol{\theta}) d\boldsymbol{\theta}$. If $q_* < \infty$, then a divergence based prior under M_2 is defined as

$$\pi^D(\boldsymbol{\theta}) = c(q_*)^{-1} h_{q_*}(\bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]) \pi^N(\boldsymbol{\theta}). \quad (13)$$

Note that, by definition, the DB priors either do not exist, or they are proper (hence they do not involve arbitrary constants).

Specific Proposals. Definition 2.1 is very general, in that several definitions of \bar{D} , h_q and π^N could be explored (as well as different choices of $0 < \delta < 1$ in $q_* = \underline{q} + \delta$). We give specific choices which, in part, are based on previous explorations and desired properties of the resulting π^D ; however our specific choices are mainly intended to reproduce JZS priors in normal scenarios, so that our proposals for DB priors can be best contemplated as extensions of JZS priors to non-normal scenarios.

In what follows, we take D to be either D^S in (6) or D^M in (7), and $h_q(t) = (1+t)^{-q}$. Since we will explore both, we need different notations:

Definition 2.2. (Sum and Minimum DB priors) *The sum DB prior π^S and the minimum DB prior π^M are the DB priors given in definition 2.1 with $h_q(t) = (1+t)^{-q}$ and D being respectively D^S (see (6)) and D^M (see (7)). When needed, we refer to their corresponding c 's and q 's as $c_S, \underline{q}^S, q_*^S$, and $c_M, \underline{q}^M, q_*^M$, respectively.*

It can easily be shown that $c_S(q) \leq c_M(q)$, so that, for regular problems (in which $\bar{D}^S < \infty$), $q_*^M < \infty$ implies $q_*^S < \infty$, and therefore, in these problems, existence of π^M implies existence of π^S .

It should be noted that, although we are not explicitly assuming a specific objective prior π^N in the definition of DB priors, properties of π^N are inherited by the DB prior π^D ; some properties will be crucial for sensible DB priors, and hence appropriate choice of π^N becomes very important.

We now explore some appealing properties of DB priors. Since these are common to both proposals in Definition 2.2, we drop unneeded super and sub indexes and refer to the prior simply as π^D . This convention will be kept through the paper; distinction between π^S and π^M will only be done when needed.

Local behavior of DB priors. The next result shows that, when $\pi^N(\boldsymbol{\theta}) = 1$ (as when $\boldsymbol{\theta}$ is a location parameter), the DB priors behave approximately, as a Student density.

Lemma 2.2. *Assume $\Theta \subset \mathcal{R}^k$. If $\pi^N(\boldsymbol{\theta}) = 1$, then the mode of π^D is $\boldsymbol{\theta}_0$ (so π^D is 'centered' at the simplest model). Furthermore, if $\boldsymbol{\theta}$ is in a neighborhood of $\boldsymbol{\theta}_0$ in Θ , then*

$$\pi^D(\boldsymbol{\theta}) \approx St_k(\boldsymbol{\theta}_0, n^* \mathbf{J}(\boldsymbol{\theta}_0)^{-1}/d, d),$$

where $d = 2q - k + 1$.

Proof. See Appendix. □

That is, in a neighborhood of $\boldsymbol{\theta}_0$, the DB priors behave approximately as k multivariate Student distributions, centered at $\boldsymbol{\theta}_0$, with d degrees of freedom, and scaled by Fisher information matrix under the simpler model. Moreover, by definition of q_*^i , d above is generally close to 1, and then the DB priors would approximately be Cauchy.

As highlighted in Section 4.3.2, the approximation in Lemma 2.2 exactly holds in Normal scenarios with $d = 1$, and hence the DB priors reproduce precisely the proposals of Jeffreys-Zellner-Siow.

Invariance under one-to-one transformations An important question is whether the DB priors are invariant under reparameterizations of the problem. Suppose that $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$ is a one-to-one monotone mapping $\boldsymbol{\xi} : \Theta \rightarrow \Theta_\xi$. The model selection problem (4) now becomes:

$$M_1^* : f_1^*(\mathbf{y}) = f^*(\mathbf{y} \mid \boldsymbol{\xi}_0) \quad \text{vs.} \quad M_2^* : f_2^*(\mathbf{y} \mid \boldsymbol{\xi}) = f^*(\mathbf{y} \mid \boldsymbol{\xi}), \quad (14)$$

where $f^*(\mathbf{y} \mid \boldsymbol{\xi}(\boldsymbol{\theta})) = f(\mathbf{y} \mid \boldsymbol{\theta})$ and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(\boldsymbol{\theta}_0)$. The next result shows that, if π^N is invariant under the reparameterization $\boldsymbol{\xi}(\boldsymbol{\theta})$ then so are the DB priors.

Proposition 1. *Let $\pi_\theta^D(\boldsymbol{\theta})$ and $\pi_\xi^D(\boldsymbol{\xi})$ denote the DB priors for the original (4), and reparameterized (14) problems respectively. If $\pi_\theta^N(\boldsymbol{\theta}) \propto \pi_\xi^N(\boldsymbol{\xi}(\boldsymbol{\theta}))|\mathcal{J}_\xi(\boldsymbol{\theta})|$, where \mathcal{J}_ξ is the Jacobian of the transformation then*

$$\pi_\theta^D(\boldsymbol{\theta}) = \pi_\xi^D(\boldsymbol{\xi}(\boldsymbol{\theta}))|\mathcal{J}_\xi(\boldsymbol{\theta})|.$$

Proof. See Appendix. □

Under the conditions of Proposition 1, Bayes factors computed from DB priors are not affected by reparameterizations. It is important to note that invariance of DB priors is a direct consequence of both the invariance of the divergence measure used and the invariance of π^N . Some objective priors π^N invariant under reparameterizations are Jeffreys' priors and (partially) the reference priors.

Compatibility with sufficient statistics. DB priors are sometimes compatible with reduction of the data via sufficient statistics. This attractive property is not shared by other objective Bayesian methods, as intrinsic Bayes factors.

Proposition 2. *Let $\mathbf{t} = \mathbf{t}(\mathbf{y})$ be a sufficient statistic for $\boldsymbol{\theta}$ in $f(\mathbf{y} \mid \boldsymbol{\theta})$ with distribution $f^*(\mathbf{t} \mid \boldsymbol{\theta})$. Assume that π^N and n^* remain the same in the problem defined by f^* , then the DB prior π^D*

for the original problem (4) and the reduced problem

$$M_1^* : f_1^*(\mathbf{t}) = f^*(\mathbf{t} \mid \boldsymbol{\theta}_0) \quad \text{vs.} \quad M_2^* : f_2^*(\mathbf{t} \mid \boldsymbol{\theta}) = f^*(\mathbf{t} \mid \boldsymbol{\theta}), \quad (15)$$

are the same.

Proof. See Appendix. □

DB priors and Jeffreys' general rule. Jeffreys (1961) tried to derive objective proper priors for testing situations other than the normal mean. Specifically, when \mathbf{y} is a random sample of size n , and for univariate θ he proposed the following model testing prior:

$$\pi^J(\theta) = \frac{1}{\pi} \frac{d}{d\theta} \tan^{-1} \bar{D}^S[\theta, \theta_0]^{1/2}. \quad (16)$$

This reduces to Jeffreys Cauchy proposal when θ is a normal mean. Also, when $|\theta - \theta_0|$ is small, $\pi^J(\theta)$ can be approximated by

$$\pi^J(\theta) \approx \frac{1}{\pi} (1 + \bar{D}^S[\theta, \theta_0])^{-1} \pi^{NJ}(\theta), \quad (17)$$

where $\pi^{NJ}(\theta)$ is Jeffreys' (estimation) prior (i.e. the squared root of the expected Fisher information number).

Note that π^J can lead to improper priors and at least in principle can not be applied for multivariate parameters. However, the approximation (17) was a main inspiration for the definition of DB priors, with clear similarities between them.

3 Comparative examples: simple null

In the spirit of Berger and Pericchi (2001) we investigate in this section the performance of DB priors in a series of situations chosen to be somehow representative of wider classes of statistical problems. We also explicitly derive well established, alternative proposals for objective priors in Bayesian hypothesis testing and compare their performance with that of DB priors. We show that in simple standard situations, DB priors produce similar results to these alternative proposals. More interestingly, in more sophisticated situations where these proposals fail (models with irregular asymptotics or improper likelihoods), the DB priors are well defined and very sensible.

We will compute and compare Bayes factors derived with DB priors, with those derived with two of the most popular objective priors for objective Bayes model selection, namely:

1. Arithmetic intrinsic prior:

$$\pi^A(\boldsymbol{\theta}) = \pi^N(\boldsymbol{\theta}) E_{\boldsymbol{\theta}}^{M_2} (B_{12}^N(\mathbf{y}^*)),$$

where the Bayes factor B^N is computed with the objective estimation prior π^N , and \mathbf{y}^* is an imaginary sample of minimum size such that $0 < m_2^N(\mathbf{y}^*) < \infty$.

2. Fractional intrinsic prior:

$$\pi^F(\boldsymbol{\theta}) = \pi^N(\boldsymbol{\theta}) \frac{\exp\{m E_{\boldsymbol{\theta}}^{M_2} \log f(y | \boldsymbol{\theta}_0)\}}{\int \exp\{m E_{\tilde{\boldsymbol{\theta}}}^{M_2} \log f(y | \tilde{\boldsymbol{\theta}})\} \pi^N(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}}.$$

In the iid case and asymptotically, π^A produces the *arithmetic intrinsic Bayes factor* (Berger and Pericchi, 1996), and π^F the *fractional Bayes factor* (O'Hagan, 1995) if the exponent of the likelihood is $b = m/n$ for a fixed m (see De Santis and Spezaferrri, 1999). Following the recommendation of Berger and Pericchi (2001) we take m to be the size of the minimal training sample \mathbf{y}^* .

In the examples of this Section, \mathbf{y} is an iid sample of size n from $f(y | \theta)$, and unless otherwise specified, $n^* = n$. We let B_{12}^S denote the Bayes factor in favor of H_1 computed with π^S (see Definition 2.2); B_{12}^M, B_{12}^A and B_{12}^F are defined similarly.

3.1 Bounded parameter space (Example 1)

We begin with a simple example, in which data is a random sample from a Bernoulli distribution, that is

$$f(y | \theta) = \theta^y (1 - \theta)^{1-y}, \quad y \in \{0, 1\}, \quad \theta \in \Theta = [0, 1].$$

The usual estimation objective prior (both reference and Jeffreys) in this problem is the beta density $\pi^N(\theta) = Be(\theta | 1/2, 1/2) \propto \theta^{-1/2} (1 - \theta)^{-1/2}$.

The DB prior for the sum-symmetrized divergence can be computed to be

$$\pi^S(\theta) \propto \left[1 + (\theta - \theta_0) \log \frac{\theta(1 - \theta_0)}{\theta_0(1 - \theta)} \right]^{-1/2} \pi^N(\theta),$$

and the DB prior for the min-symmetrized divergence

$$\pi^M(\theta) \propto \left(1 + \bar{D}^M[\theta, \theta_0] \right)^{-1/2} \pi^N(\theta),$$

where

$$\bar{D}^M[\theta, \theta_0] = \begin{cases} 2 KL[\theta : \theta_0] & \text{if } \min\{\theta_0, 1 - \theta_0\} < \theta < \max\{\theta_0, 1 - \theta_0\} \\ 2 KL[\theta_0 : \theta] & \text{otherwise,} \end{cases}$$

and $KL[\theta : \theta_0] = \theta_0 \log \frac{\theta_0}{\theta} + (1 - \theta_0) \log \frac{1 - \theta_0}{1 - \theta}$.

The intrinsic priors are derived in the next result. The proof is straightforward and hence it is omitted.

Lemma 3.1. *The arithmetic intrinsic prior is*

$$\pi^A(\theta) = \left(\frac{2}{\pi} (1 - \theta_0)(1 - \theta) + \theta_0\theta \right) \pi^N(\theta)$$

and the fractional intrinsic prior is

$$\pi^F(\theta) = \left(\frac{\theta_0^\theta (1 - \theta_0)^{1 - \theta}}{\Gamma(\theta + 1/2)\Gamma(3/2 - \theta)} \right) \pi^N(\theta).$$

By construction, π^S and π^M are proper priors; π^A is proper but π^F is not. For instance, for $\theta_0 = 1/2$, π^F integrates to 1.28 and for $\theta_0 = 3/4$, π^F integrates to 1.18. This implies a small bias in the Bayes factor in favor of M_2 . In Figure 1 we display π^S, π^M, π^A and π^F for $\theta_0 = 1/2$ and $\theta_0 = 3/4$. They can be seen to be very similar. When $\theta_0 = 1/2$ they are also similar to the objective estimation prior $Be(\theta | 1/2, 1/2)$, but not for other values of θ_0 .

We also compute the Bayes factors for the four different priors, when $\theta_0 = 1/2$, for two different sample sizes, $n = 10$ and $n = 100$, and for different values of the MLE, $\hat{\theta} = \sum_{i=1}^{10} y_i/n$ (see Table 1). All the results are quite similar. As expected, B_{12}^F gives the most support to M_2 ; B_{12}^A gives the least. Both DB priors produce similar results, being slightly closer to B_{12}^A than to B_{12}^F .

Finally, we consider application to real data taken from Conover (1971). Under the hypothesis of simple Mendelian inheritance, a cross between two particular plants produces, in a proportion of $\theta = 3/4$ a specie called ‘giant’. To determine whether this assumption is true, Conover (1971) crossed $n = 925$ pair of plants, getting $T = 682$ giant plants. The Bayes factors in favor of the Mendelian inheritance hypothesis (simplest model) are also given in Table 1 for the four different priors. Again the results are very similar, the fractional intrinsic prior providing the least support to M_1 .

3.2 Scale parameter (Example 2)

We next consider another simple example of testing a scale parameter. Specifically, we consider that data come from the one parameter exponential model with mean μ , that is,

$$f(y | \mu) = Exp\left(y \mid \frac{1}{\mu}\right) = \frac{1}{\mu} \exp\left\{-\frac{y}{\mu}\right\}, \quad y > 0, \quad \mu > 0,$$

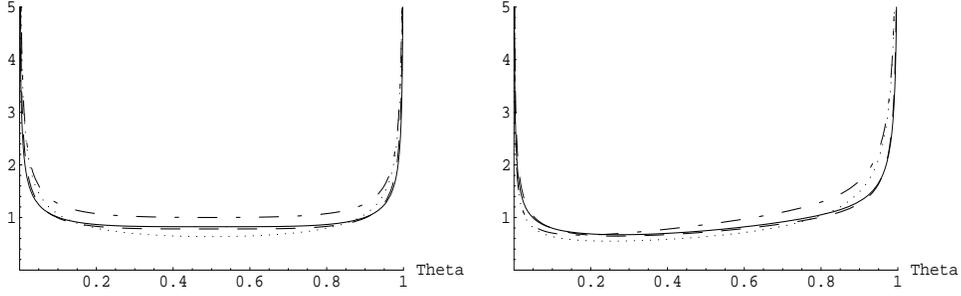


Figure 1: In Bernoulli example: π^S (Solid line), π^M (Dot-dashed line), π^A (Dots) and π^F (Dashed line), for the case $\theta_0 = 1/2$ (left) and $\theta_0 = 3/4$ (right).

$n = 10$	$\hat{\theta}$	B_{12}^S	B_{12}^M	B_{12}^A	B_{12}^F
	0.50	3.26	3.44	4.06	2.68
	0.65	2.14	2.24	2.58	1.75
	0.80	0.55	0.57	0.60	0.44
$n = 100$					
	0.50	9.74	10.28	12.56	8.03
	0.55	5.93	6.26	7.61	4.89
	0.60	1.33	1.40	1.68	1.09
Conover		19.38	20.20	20.79	16.02

Table 1: Bayes factors in favor of M_1 for Bernoulli testing of $\theta_0 = 1/2$, for different values of the MLE and $n = 10$, $n = 100$. Also, Bayes factors for Conover data.

and that it is desired to test $H_1 : \mu = \mu_0$ vs. $H_2 : \mu \neq \mu_0$. Here $\pi^N(\mu) = \mu^{-1}$, and the DB priors are computed to be:

$$\pi^S(\mu) \propto \left[1 + \frac{(\mu - \mu_0)^2}{\mu\mu_0}\right]^{-1/2} \mu^{-1}, \quad \pi^M(\mu) \propto (1 + \bar{D}^M[\mu, \mu_0])^{-3/2} \mu^{-1},$$

where

$$\bar{D}^M[\mu, \mu_0] = \begin{cases} 2 KL[\mu_0 : \mu] & \text{if } \mu > \mu_0 \\ 2 KL[\mu : \mu_0] & \text{if } \mu \leq \mu_0, \end{cases}$$

and $KL[\mu : \mu_0] = \log(\mu_0/\mu) - (\mu_0 - \mu)/\mu_0$. The intrinsic priors are given in the next lemma (the proof is straightforward and is omitted):

Lemma 3.2. *The arithmetic and fractional intrinsic priors are*

$$\pi^A(\mu) = \mu_0^{-1} \left(1 + \frac{\mu}{\mu_0}\right)^{-2}, \quad \pi^F(\mu) = \mu_0^{-1} \exp\left\{-\frac{\mu}{\mu_0}\right\} = \text{Exp}\left\{\mu \mid \frac{1}{\mu_0}\right\}.$$

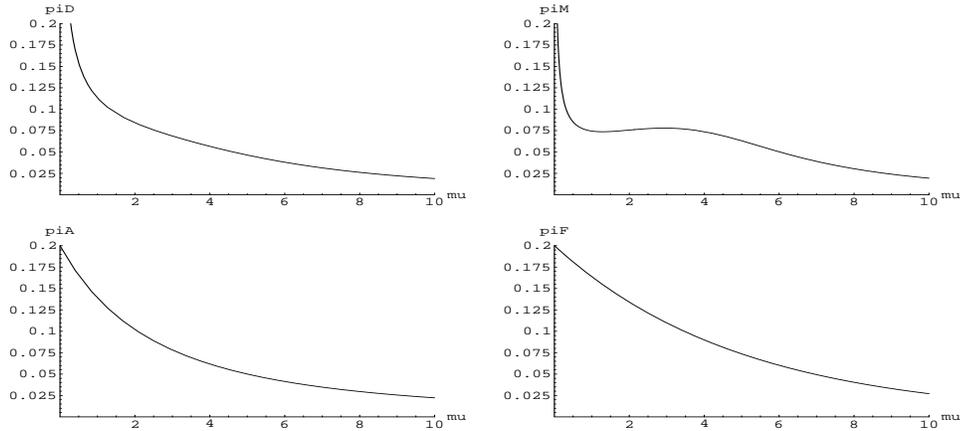


Figure 2: π^S (upper left), π^M (upper right), π^A (lower left) and π^F (lower right) for the Exponential testing of $\mu_0 = 5$.

The four priors are shown in Figure 2 when testing $\mu_0 = 5$. They all have similar shapes, although that of π^M is somehow unusual; they have some interesting properties:

1. In the log scale, both π^M and π^S are symmetric around $\log \mu_0$; this is in accordance to JZS proposal, since $\log(\mu)$ is a location parameter.
2. All four priors are proper.
3. Neither the arithmetic intrinsic nor the DB priors have moments; the arithmetic fractional has all the moments.
4. π^M has the heaviest tails, and π^F the thinnest. π^S has heavier tails than π^A
5. All four priors are ‘centered’ at the null value μ_0 ; indeed, μ_0 is the median of the DB priors and of π^A , and it is the mean of π^F .

The four Bayes factors B_{12} in favour of $M_1 : \mu = 5$ appear in Table 2, for two values of n ($n = 10$ and $n = 100$) and some few values of the MLE $\hat{\mu} = \sum_{i=1}^n y_i/n \in \{5, 7.5, 2.5\}$. We again find very similar results for the different priors, with B_{12}^S and B_{12}^A providing slightly more support to M_1 than B_{12}^M and B_{12}^F when data is compatible with M_1 .

We next investigate a desirable property of Bayes factors which often fails when they are computed using conjugate priors (see Berger and Pericchi, 2001). It is natural to expect that, for any given sample size, $B_{12} \rightarrow 0$ as the evidence against the simpler model M_1 becomes overwhelming. When this property holds, we say that the Bayes factor is *evidence consistent* (or *finite sample consistent*). It is easy to show that, if $\bar{y} \rightarrow \infty$ then $B_{12} \rightarrow 0 \forall n$, no matter what prior is used to obtain the Bayes factor. The following lemma provides sufficient conditions for $B_{12} \rightarrow 0$ as $\bar{y} \rightarrow 0$.

$n = 10$	$\hat{\mu}$	B_{12}^S	B_{12}^M	B_{12}^A	B_{12}^F
	5	5.65	4.43	5.13	3.59
	7.5	2.36	2.02	2.09	1.58
	2.5	0.95	0.88	0.82	0.59
$n = 100$					
	5	17.28	12.81	15.98	10.89
	7.5	14.6×10^{-4}	12.2×10^{-4}	13×10^{-4}	9.4×10^{-4}
	2.5	0.86×10^{-7}	0.83×10^{-7}	0.73×10^{-7}	0.54×10^{-7}

Table 2: Bayes factors for the exponential testing with $\mu_0 = 5$ for different values of the MLE and $n = 10$, $n = 100$.

Lemma 3.3. *Let B_{12}^π be the Bayes factor computed with $\pi(\mu)$. $B_{12}^\pi \rightarrow 0$ as $\bar{y} \rightarrow 0$, for all $n \geq k > 0$ if and only if*

$$\int_0^1 \mu^{-k} \pi(\mu) d\mu = \infty. \quad (18)$$

Proof. See Appendix. □

It follows that all four priors considered produce *evidence consistent* Bayes factors for all $n \geq 1$. Evidence consistency provides further insight into the behaviour of the DB priors. Indeed, we recall that in the general definition of DB priors we used the power $\underline{q} + \delta$, and then we recommended the specific choice $\delta = .5$. Interestingly, if $\delta > 1$ is used instead, then π^S would not be evidence consistent as $\bar{y} \rightarrow 0$.

Last, we study the behavior of B_{12} as the evidence in favor of M_1 grows (that is, as $\bar{y} \rightarrow \mu_0$). For this example, it is easy to show that, when $\bar{y} \rightarrow \mu_0$, B_{12} grows to a constant, $B_{12}^0(n, \pi)$ say, that depends only on n and the prior used. Of course, it then follows from the dominated convergence theorem that $B_{12}^0(n, \pi) \rightarrow \infty$ with n , but this also follows from general consistency of Bayes factors (for proper, fix priors), so it is not very interesting. Of more interest for our comparison is to study how fast $B_{12}^0(n, \pi_2)$ goes to ∞ . In Figure 3 we show $B_{12}^0(n, \pi)$ for the four priors considered. It can be seen that π^S is the one producing the largest values of B_{12}^0 for all values of n , with those for π^A following very closely.

3.3 Location-scale (Example 3)

DB priors are defined in general for vector parameters $\boldsymbol{\theta}$. As an illustration, we next consider a most popular example, namely the normal distribution; here the 2-dimensional $\boldsymbol{\theta}$ has two components of different nature (location and scale). Specifically, assume that

$$f(y | \mu, \sigma) = N(y | \mu, \sigma^2),$$

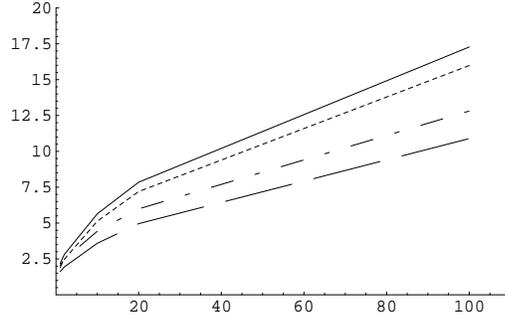


Figure 3: Upper bounds $B_{12}^0(n, \pi)$ of Bayes factors as a function of n for the priors π^S (Solid line), π^M (Dot-dashed line), π^F (Dashed Line), and π^A (Dots).

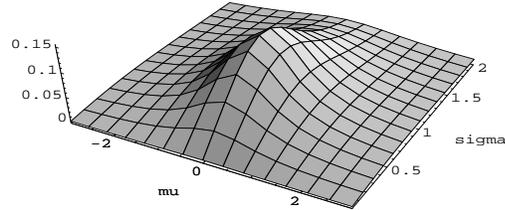


Figure 4: $\pi_2^{D_s}$ for the Normal problem, with $\mu_0 = 0, \sigma_0 = 1$.

and that we want to test $M_1 : (\mu, \sigma) = (\mu_0, \sigma_0)$ versus $M_2 : (\mu, \sigma) \neq (\mu_0, \sigma_0)$. This hypothesis testing problem occurs often in *statistical process control*, where a production process is considered ‘in control’ if its production outputs have a specified mean and standard deviation (the so called *nominal values*); the question of interest is whether the process is in control, that is, whether the mean and variance are equal to the nominal values.

To compute the DB priors we use the reference prior $\pi^N(\mu, \sigma) = \sigma^{-1}$; for the sum-DB prior we get:

$$\pi^S(\mu, \sigma) = \pi^S(\sigma) \pi^S(\mu | \sigma), \quad \pi^S(\sigma) \propto \frac{\sigma}{(\sigma_0^4 + \sigma^4)^{1/2} (\sigma_0^2 + \sigma^2)^{1/2}},$$

and

$$\pi^S(\mu | \sigma) = \text{Ca}(\mu | \mu_0, \Sigma), \quad \Sigma = \frac{\sigma_0^4 + \sigma^4}{\sigma_0^2 + \sigma^2},$$

where Ca represents the Cauchy density. In this example, the minimum-DB prior π^M does not exist, since $q^M = \infty$. It can be checked that $\pi^S(\mu | \sigma)$ is symmetric around μ_0 , which is a location parameter in $\pi^S(\mu | \sigma)$; σ_0 is a scale parameter in $\pi^S(\sigma)$. The joint density π^S is shown in Figure 4.

The intrinsic priors, which have simpler forms and thinner tails, are derived next (the proof

	$\bar{x} = 0$			$\bar{x} = 1$			$\bar{x} = 2$		
	B_{12}^s	B_{12}^A	B_{12}^F	B_{12}^s	B_{12}^A	B_{12}^F	B_{12}^s	B_{12}^A	B_{12}^F
$S = 0.5$	2.30	1.35	0.70	0.03	0.02	0.01	$3 \cdot 10^{-8}$	$4 \cdot 10^{-8}$	$6 \cdot 10^{-8}$
$S = 1$	18.67	18.55	11.72	0.21	0.19	0.18	$1 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$6 \cdot 10^{-7}$
$S = 2$	0.006	0.006	0.017	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$21 \cdot 10^{-5}$	$2 \cdot 10^{-11}$	$2 \cdot 10^{-11}$	$41 \cdot 10^{-11}$

Table 3: For multidimensional parameter problem ($\mu_0 = 0, \sigma_0 = 1$), values of B_{12} for different values of (\bar{x}, S) with $n = 10$.

is omitted):

Lemma 3.4. *The arithmetic intrinsic prior is*

$$\pi^A(\mu, \sigma) = \pi^A(\sigma) \pi^A(\mu | \sigma), \quad \pi^A(\sigma) = \frac{2}{\pi} \frac{\sigma_0}{\sigma^2 + \sigma_0^2}, \quad \pi^A(\mu | \sigma) = N(\mu | \mu_0, \frac{\sigma^2 + \sigma_0^2}{2}),$$

and the fractional intrinsic prior is

$$\pi^F(\mu, \sigma) = N^+(\sigma | 0, \frac{\sigma_0^2}{2}) N(\mu | \mu_0, \frac{\sigma_0^2}{2}),$$

where N^+ stands for the normal density truncated to the positive real line.

The intrinsic priors are proper; also, as with the sum-DB prior, μ_0 and σ_0 are location and scale parameters for $\mu | \sigma$ and σ respectively. Under the fractional intrinsic prior π^F , μ and σ are independent a priori.

Values of B_{12} for all three priors and differer values of the sufficient statistic (\bar{x}, S) are given in Table 3 when $(\mu_0, \sigma_0) = (0, 1)$. The Bayes factors corresponding to the different priors can be seen to be quite similar, specially B_{12}^S and B_{12}^A .

For the three priors, we display in Figure 5 the marginal distributions of σ and in Figure 6, the conditional distributions of μ given σ . It can clearly be seen that $\pi^F(\sigma)$ has thinner tails than π_2^A and π_2^S (recall, thicker tails seem to perform better for testing). Also, all conditional priors for μ are symmetric around their mode μ_0 , with $\pi^S(\mu | \sigma)$ having the heaviest tails.

Finally we compare the behavior of the three priors in a real example taken from Montgomery (2001). The example refers to controlling the piston ring for an automotive engine production process. The process was considered to be *in control* if the mean and the standard deviation of the inside diameter (in millimeters) of the pistons were $\mu_0 = 74.001$ and $\sigma_0 = 0.0099$. At some specific time, the following sample was taken from the process:

$$74.035, 74.010, 74.012, 74.015, 74.026,$$

and it had to be checked whether the process was in control. Bayes factors are given in Table 4. B_{12}^F provides about twice more support to M_1 than B_{12}^S and B_{12}^A , which are very similar to each

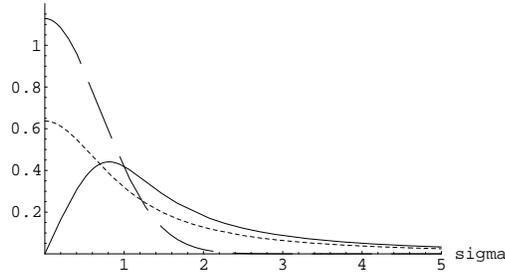


Figure 5: Marginal distributions of σ when $(\mu_0, \sigma_0) = (0, 1)$; $\pi_2^S(\sigma)$ (solid line), $\pi_2^A(\sigma)$ (dots), and $\pi_2^F(\sigma)$ (dashed line). The pair (mode, median) for these priors are $(0.81, 1.56)$ for π^D , $(0, 1)$ for π^A , and $(0, 0.48)$ for π^F .

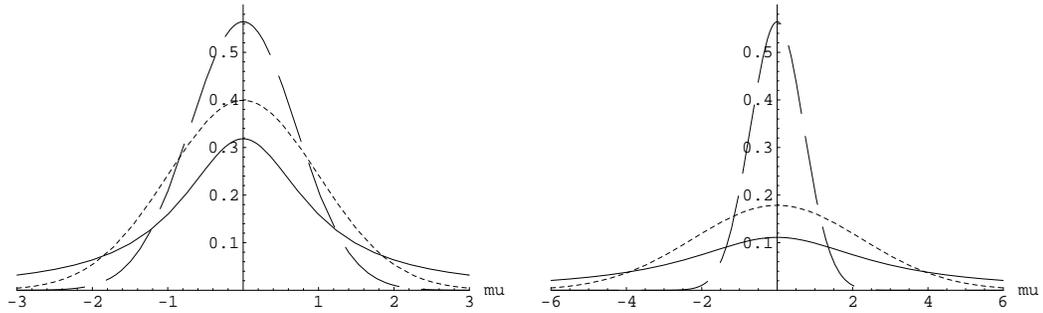


Figure 6: Conditional distributions of μ given $\sigma = 1$ (left) and $\sigma = 3$ (right) when $(\mu_0, \sigma_0) = (0, 1)$; π^S (solid), π^A (dots), and π^F (dashed).

other.

B_{12}^S	B_{12}^A	B_{12}^F
0.004	0.005	0.011

Table 4: Bayes factors B_{12} for Montgomery (2001) example.

3.4 Irregular models (Example 4)

There is an important class of models for which the parameter space is constrained by the data. These models do not have regular asymptotics and hence solutions based on asymptotic theory (like the Bayesian information criteria, BIC) do not apply. Moreover, these models are very challenging for the intrinsic approach; indeed, as discussed in Berger and Pericchi (2001), the fractional Bayes factor is completely unreasonable (and hence the fractional intrinsic prior is useless), and the arithmetic intrinsic prior (which was only derived for the one side problem) is “something of a conjecture” (authors’ verbatim). We take here the simplest such models,

namely an exponential distribution with unknown location. Accordingly, assume that

$$f(y | \theta) = \exp\{-(y - \theta)\}, \quad y > \theta,$$

and that it is wanted to test $H_1 : \theta = \theta_0$ vs. $H_2 : \theta \neq \theta_0$. To the best of our knowledge, no objective priors have been proposed for this testing problem in the literature.

In these situations, the sum-symmetrized kulback-Leibler divergence $D^S[\theta, \theta_0]$ is ∞ , so we have to use the minimum. It can be checked that $\bar{D}^M[\theta, \theta_0] = 2|\theta - \theta_0|$, a well defined divergence. Also, $\pi^N(\theta) = 1$ since θ is a location parameter. The Minimum DB prior is then given by

$$\pi^M(\theta) = \frac{1}{2}(1 + 2|\theta - \theta_0|)^{-3/2}, \quad \theta \in \mathcal{R},$$

which is symmetric with respect to θ_0 (as expected, since θ is a location parameter); also, π^M has no moments. Figure 7 (left) shows $\pi^M(\theta)$ when $\theta_0 = 0$.

We next investigate the *evidence* consistency for any n . The sufficient statistic is $T = \min\{y_1, \dots, y_n\}$. It is trivially true that $B_{12} \rightarrow 0$, as $T \rightarrow -\infty$ for any (proper) prior. The next lemma provides a sufficient condition on the prior to produce evidence consistency $\forall n$, as $T \rightarrow \infty$.

Lemma 3.5. *Let $\pi(\theta)$ be any proper prior (on M_2) and B_{12}^π be the corresponding Bayes factor. If for some integer $k > 0$*

$$\int_{\theta_0}^{\infty} e^{k\theta} \pi(\theta) d\theta = \infty, \quad (19)$$

then $B_{12}^\pi \rightarrow 0$ as $T \rightarrow \infty \forall n \geq k$.

Proof. See Appendix. □

It follows from the previous lemma that π^M produces *evidence consistent* Bayes factors $\forall n \geq 1$. We next investigate the situation for increasing evidence *in favor* of M_1 , that is, as $T \rightarrow \theta_0^+$. Let

$$B_{12}^0(n) = \lim_{T \rightarrow \theta_0^+} B_{12}^{\pi^D}.$$

$B_{12}^0(n)$ is an upper bound of B_{12} when the evidence in favor of M_1 is largest. It can be seen in Figure 7 (right) that $B_{12}^0(n)$ is nearly linear. Of course $B_{12}^0(n) \rightarrow \infty$ when $n \rightarrow \infty$.

As mention before, there does not seem to be any other proposals in the literature for the two-side testing problem. However, Berger and Pericchi (2001), do consider the ‘one side testing’ version, namely testing $M_1 : \theta = \theta_0$ vs $M_2 : \theta > \theta_0$; they conjecture that the arithmetic intrinsic prior for this problem is the proper density

$$\pi_2^A(\theta) = (-e^{\theta - \theta_0} \log(1 - e^{\theta_0 - \theta}) - 1), \quad \theta > \theta_0,$$

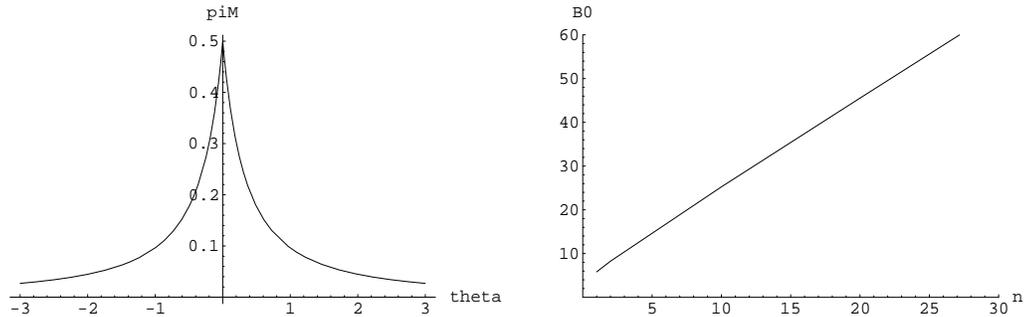


Figure 7: Irregular example, two-side testing of $M_1 : \theta = 0$. Left: the DB prior π^M ; Right: $B_{12}^0(n)$ as a function of n .

which is a decreasing and unbounded function of θ . We next compare the (minimum) DB prior for this problem with Berger and Pericchi proposal.

Although our original formulation appears to be in terms of two side testing (see (1)) in reality it suffices to define Θ appropriately to cover other testing situations. For instance, in our one-side testing, we take $\Theta = [\theta_0, \infty)$. The (minimum) DB prior is

$$\pi^M(\theta) = (1 + 2(\theta - \theta_0))^{-3/2}, \quad \theta > \theta_0.$$

It can be checked, that π^A meets condition (19) for $k = 1$ and hence π^A produces evidence consistent Bayes factors $\forall n \geq 1$. The priors π^A and π^M are displayed in Figure 8. We find that also in this example π^M has thicker tails.

In this one side testing scenario (in sharp contrast to the behavior in the two-side testing) the Bayes factor in favor of M_1 for every $n > 0$ does grow to ∞ as the evidence in favor of M_1 grows. Indeed, the Bayes factor B_{12} is

$$\left(\int_{\theta_0}^T \exp\{n(\theta - \theta_0)\} \pi(\theta) d\theta \right)^{-1},$$

so that, $B_{12} \rightarrow \infty$ when $T \rightarrow \theta_0^+$, $\forall n > 0$, no matter what prior is used. Note that here θ_0 is in the boundary of the parameter space.

In Table 5, we produce the Bayes factors computed with π^A and π^M when $\theta_0 = 0$ for various values of $T = \min\{y_1, \dots, y_n\}$, and for $n = 10$ and $n = 20$. For small values of T ($T < 0.20$), when evidence supports M_1 , B_{12}^M is considerably larger than B_{12}^A , thus giving more support to M_1 . For larger values of T (that is, when data contradict M_1) both priors result in very similar Bayes factors.

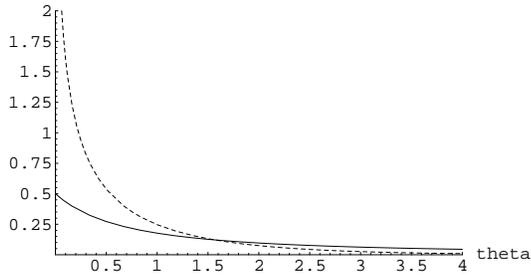


Figure 8: Irregular, one side testing problem: π^D (solid) and π^A (dots) for the case $\theta_0 = 0$.

	T					
	0.02	0.05	0.10	0.20	0.50	1.00
	$n = 10$					
B_{12}^M	46.56	16.66	6.83	2.19	0.16	0.002
B_{12}^A	11.54	5.16	2.57	1.02	0.10	0.001
	$n = 20$					
B_{12}^M	41.96	12.65	3.75	0.55	0.002	$2 \cdot 10^{-7}$
B_{12}^A	10.52	4.04	1.50	0.28	0.002	$2 \cdot 10^{-7}$

Table 5: Irregular models, one side testing. Values of B_{12} for different values of T , n and for the two priors π^A, π^M , when testing $\theta_0 = 0$.

3.5 Mixture models (Example 5)

Mixture models are among the most challenging scenarios for objective Bayesian methodology. These models have *improper likelihoods*, i.e., likelihoods for which no improper prior yields a finite marginal density (integrated likelihood). Recently, Pérez and Berger (2001), have used *expected posterior priors* (see Pérez and Berger, 2002) to derive objective estimation priors, but basically no general method seems to exist for deriving objective priors for testing with these models.

However, the divergence measures are well defined (although the integrals are now more involved) providing a reasonable DB prior to be used in model selection. We consider a simple illustration. Assume

$$f(y | \mu, p) = p N(y | 0, 1) + (1 - p) N(y | \mu, 1),$$

and the testing of $H_1 : \mu = 0$, vs. $H_2 : \mu \neq 0$, where $p < 1$ is known (if $p = 1$, both hypotheses define the same model). As Berger and Pericchi (2001) point out, there is no minimal training sample for this problem and hence the intrinsic Bayes factor cannot be defined. The fractional Bayes factor does not exist either. The only prior we know for this problem is the recommendation in Berger and Pericchi (2001) of using $\pi^{BP}(\mu) = Ca(\mu|0, 1)$.

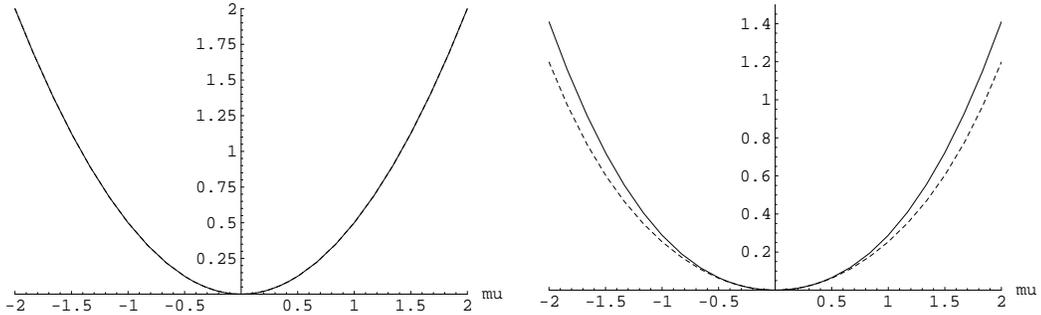


Figure 9: $G(p, \mu, \mu) - G(p, \mu, 0)$ (solid) and its Laplace approximation $G^L(p, \mu, \mu) - G^L(p, \mu, 0)$ (dots). Left: $p = 0.50$. Right: $p = 0.75$.

Although there is no formal $\pi^N(\mu)$ here, $\pi^N(\mu) = 1$ is usually assumed (see for instance Pérez and Berger, 2002). It can be shown that $\underline{q}^M = \infty$, and hence, π^M does not exist. Let

$$G(p, \mu, \mu^*) = \int_{-\infty}^{\infty} \log \left[1 + \frac{1-p}{p} e^{y\mu - \mu^2/2} \right] N(y \mid \mu^*, 1) dy. \quad (20)$$

Then

$$D^S[\mu, \mu_0] = n(1-p)(G(p, \mu, \mu) - G(p, \mu, 0)).$$

It can be shown that $\underline{q}^S < \infty$, and hence that the sum DB prior π^S exists. The normalizing constant, however, can not be derived in closed form. Numerical procedures could be used to exactly derive the sum-DB prior. We use instead a Laplace approximation (see Tanner 1996) to (20) to get an approximate DB prior. Specifically

$$G(p, \mu, \mu^*) \approx \log \left[1 + \frac{1-p}{p} e^{\mu^* \mu - \mu^2/2} \right] = G^L(p, \mu, \mu^*). \quad (21)$$

Figure 9 shows $G(p, \mu, \mu^*) - G(p, \mu, 0)$ and its approximation $G^L(p, \mu, \mu^*) - G^L(p, \mu, 0)$ for $p = .5$ and $p = .75$. The approximation is very good as long as p is not too extreme.

We can now use this approximation to derive the DB prior. Note that the effective sample size here is $n^* = n(1-p)$, so that the unitary sum-symmetrized divergence is

$$\bar{D}^S[\mu, \mu_0] = \frac{D^S(\mu, \mu_0)}{n(1-p)} \approx \log \frac{1 + \frac{1-p}{p} e^{\mu^2/2}}{1 + \frac{1-p}{p} e^{-\mu^2/2}} = \bar{D}^{SL}[\mu, \mu_0].$$

This approximation is specially appealing because it also keeps essential properties of the divergence measures. In particular, $\bar{D}^{SL}(\mu, \mu_0) \geq \bar{D}^{SL}(\mu_0, \mu_0) = 0$, so that the approximate DB prior

$$\pi^{SL}(\mu) \propto (1 + \bar{D}^{SL}(\mu, \mu_0))^{-q_*^s},$$

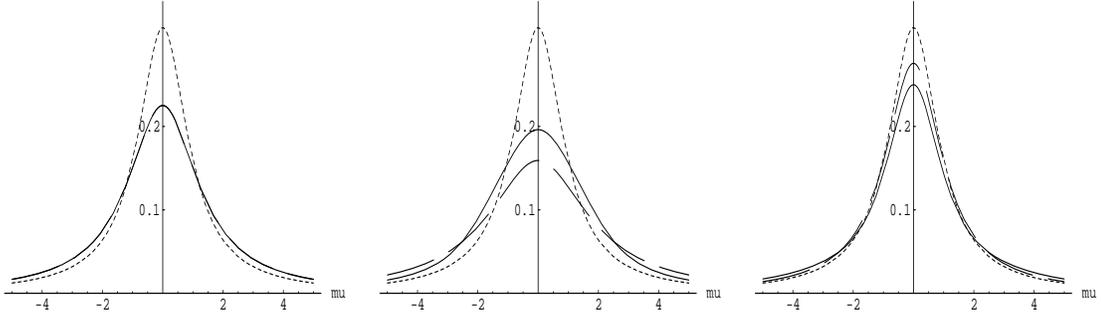


Figure 10: π^{SL} (solid line), $Ca(0, 1/1 - p)$ (dashed line) and $\pi^{BP}(\mu) = Ca(\mu|0, 1)$ (dots) for $p = 0.50$ (left), $p = 0.75$ (middle) and $p = 0.25$ (right).

has a mode at zero. Since $\underline{q}^s = 1/2$, we finally get

$$\pi^{SL}(\mu) \propto (1 + \bar{D}^{LS}(\mu, \mu_0))^{-1}.$$

Interestingly, the prior π^{SL} is close to a Cauchy density, which was Berger and Pericchi proposal, although the scale differs. Indeed a Taylor expansion of order 3, around $\mu = 0$ gives

$$\bar{D}^{SL}(\mu, \mu_0) \approx (1 - p)\mu^2, \quad (22)$$

so that, unless p is very close to 1, π^{SL} behaves around 0 as a $Ca(\mu | 0, 1/(1 - p))$; the approximation is excellent when p is close to 0.5. In the tails, on the other hand, we have that, as $|\mu| \rightarrow \infty$

$$\bar{D}^{SL}(\mu, \mu_0) \approx \frac{\mu^2}{2}, \quad (23)$$

independently of p . Hence, the tails of π^{SL} are close to those of a $Ca(\mu | 0, 2)$ density. Note that both approximations (22) and (23) coincide for $p = 0.5$.

The scale of the $Ca(\mu | 0, 1/(1 - p))$ makes intuitive sense. Indeed, the larger p , the less observations providing information about μ we get, and the DB prior adjust to a less informative likelihood by inflating its scale. Figure 10 displays π^{SL} , its $Ca(\mu | 0, 1/(1 - p))$ approximation, and the proposal of Berger and Pericchi (2001) for different values of p . Notice that, for values of p close to 0, π^{SL} (and its approximation $Ca(0, 1/1 - p)$) approximately behaves as a $Ca(0, 1)$, the Berger and Pericchi proposal (see Figure 10, right). This has an interesting interpretation since, as $p \rightarrow 0$ the testing problem in this example essentially coincides with that of testing $H_1 : \mu = 0$ vs. $H_2 : \mu \neq 0$, when μ is the mean of a normal density, for which the $Ca(\mu | 0, 1)$ is perhaps the most popular prior to be used as prior distribution for μ under H_2 .

In this example, the DB prior (as well as Berger and Pericchi proposal) again produces evidence consistent Bayes factors for all n . Indeed, it can be shown that if one of the y'_i 's tends

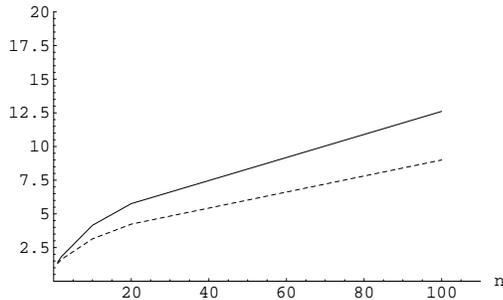


Figure 11: B_{12}^0 for π^{SL} (solid line) and π^{BP} (dots) as a function of n , for $p = 0.5$.

μ	$p = 0.25$			$p = 0.5$			$p = 0.75$		
	B_{12}^{SL}	B_{12}^{ap}	B_{12}^{BP}	B_{12}^{SL}	B_{12}^{ap}	B_{12}^{BP}	B_{12}^{SL}	B_{12}^{ap}	B_{12}^{BP}
0	5.49	4.97	4.39	2.56	2.56	2.01	2.37	2.90	1.87
0.5	1.82	1.65	1.49	0.36	0.36	0.33	1.69	2.06	1.42
1	0.07	0.06	0.06	0.04	0.04	0.04	0.01	0.01	0.01

Table 6: Bayes factors B_{12} for simulated samples of size $n = 20$ from the mixture model with various values of p and μ and the priors π^{SL} , its approximation $Ca(\mu | 0, 1/(1-p))$ and $\pi^{BP}(\mu) = Ca(\mu | 0, 1)$.

to ∞ or $-\infty$, then the corresponding Bayes factor tends to 0 no matter what prior is used. On the other hand, as the evidence for H_1 increases, we get a finite upper bound on B_{12} for every fixed sample size n :

$$B_{12}^0(n, p, \pi) = \lim_{y_i \rightarrow 0, \forall i} B_{12}.$$

In Figure 11 we show B_{12}^0 for $\pi = \pi^{SL}$ and $\pi = Ca(\mu | 0, 1)$ as a function of n for $p = 0.5$. As in the previous examples, it is an immediate consequence that $B_{12}^0(n, p, \pi) \rightarrow \infty$ as $n \rightarrow \infty$ for both priors, but the support for H_1 is larger when π^{SL} is used for every n .

In Table 6 we show the Bayes factors B_{12}^{SL} , B_{12}^{ap} and B_{12}^{BP} computed respectively with the priors π^{SL} , its $Ca(\mu | 0, 1/(1-p))$ approximation and the $Ca(\mu | 0, 1)$ proposed by Berger and Pericchi. Since reduction by sufficient statistic is not possible, the Bayes factors are computed for simulated samples of size $n = 20$, with mean $\mu \in \{0, 0.5, 1\}$, and $p \in \{0.25, .5, 0.75\}$. B_{12}^{SL} and its approximation B_{12}^{ap} are very close, demonstrating that the approximation is very good for the considered range of p . B_{12}^{SL} and B_{12}^{BP} are also very similar.

4 Nuisance parameters

In this section we deal with more realistic problems in which the distribution of the data is not fully specified under the null (simplest model), but depends on some nuisance parameter.

Assume that y_i , $i = 1, \dots, n$ are independent (not necessarily i.i.d.) and that $\mathbf{y} = (y_1, \dots, y_n) \sim \{f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\nu} \in \Upsilon\}$. We want to test $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Equivalently we want to solve the model selection problem (2) where it is carefully acknowledged that $\boldsymbol{\nu}$ can have a different meanings in each model.

However, from now on we assume, after suitable reparameterization if needed, that $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ are *orthogonal* (that is, that Fisher information matrix is block diagonal). It is then customary to assume that $\boldsymbol{\nu}$ has the same meaning under both models (see Berger and Pericchi, 1996, for an asymptotic justification). This will be needed for the divergence measures to have intuitive meaning, and also to justify assessment of the same (possibly improper) prior for $\boldsymbol{\nu}$ under both models thus considerably simplifying the assessment task. The suitability of orthogonal parameters in the presence of model uncertainty was first exploited by Jeffreys (1961) and it has been successfully used by many others (see for example Zellner and Siow, 1980, 1984, and Clyde, DeSimone and Parmigiani, 1996). For univariate θ , Cox and Reid (1987) explicitly provide an orthogonal reparameterization.

Accordingly, we assume that the hypothesis testing problem above is equivalent to that of choosing between the competing models:

$$M_1 : f_1(\mathbf{y} | \boldsymbol{\nu}) = f(\mathbf{y} | \boldsymbol{\theta}_0, \boldsymbol{\nu}) \quad \text{vs.} \quad M_2 : f_2(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}) = f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}), \quad (24)$$

where $\boldsymbol{\theta}_0 \in \Theta$ is a specified value, and $\boldsymbol{\nu}$ (the *old parameter* in Jeffrey's terminology) is assumed to be common to both models, which only differ by the different value of the *new parameter* $\boldsymbol{\theta}$ under M_2 .

4.1 Divergence Measures

The basic measure of discrepancy between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ is again Kullback-Leibler directed divergence (5) where $\boldsymbol{\nu}$ is taken to be the same in both models:

$$KL[(\boldsymbol{\theta}_0, \boldsymbol{\nu}) : (\boldsymbol{\theta}, \boldsymbol{\nu})] = \int_{\mathbf{y}} (\log f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}) - \log f(\mathbf{y} | \boldsymbol{\theta}_0, \boldsymbol{\nu})) f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}) d\mathbf{y}.$$

Note that using the same $\boldsymbol{\nu}$ only makes intuitive sense if $\boldsymbol{\nu}$ has the same meaning under both models, and hence can be considered common. Actually, Pérez (2005) using geometrical arguments, shows that under orthogonality $KL[(\boldsymbol{\theta}_0, \boldsymbol{\nu}) : (\boldsymbol{\theta}, \boldsymbol{\nu})]$ can be interpreted as a measure of divergence between f_1 and f_2 due solely to the parameter of interest $\boldsymbol{\theta}$. This interpretation does not hold for other divergence measures, as the intrinsic loss divergence defined in Bernardo and Rueda (2002).

Similarly to Section 2 we symmetrize Kullback-Leibler directed divergence by adding or taking the minimum of them, resulting in the sum-divergence and min-divergence measures

between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ for a given $\boldsymbol{\nu}$

$$D^S[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}] = KL[(\boldsymbol{\theta}, \boldsymbol{\nu}) : (\boldsymbol{\theta}_0, \boldsymbol{\nu})] + KL[(\boldsymbol{\theta}_0, \boldsymbol{\nu}) : (\boldsymbol{\theta}, \boldsymbol{\nu})], \quad (25)$$

and

$$D^M[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}] = 2 \times \min\{KL[(\boldsymbol{\theta}, \boldsymbol{\nu}) : (\boldsymbol{\theta}_0, \boldsymbol{\nu})], KL[(\boldsymbol{\theta}_0, \boldsymbol{\nu}) : (\boldsymbol{\theta}, \boldsymbol{\nu})]\}. \quad (26)$$

D^M is used by Pérez (2005) to define what he calls the “orthogonal intrinsic loss”.

In what follows, many of the definitions and properties apply to both D^S and D^M , in which case we again generically use D to denote any of them. Their basic properties were discussed in Section 2. As before, the building block of the DB prior is the *unitary measure of divergence* $\bar{D} = D/n^*$, where n^* is the equivalent sample size for $\boldsymbol{\theta}$.

4.2 DB priors in the presence of nuisance parameters

For testing $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, or equivalently choosing between models M_1 and M_2 in (24), we need priors $\pi_1(\boldsymbol{\nu})$ under M_1 and $\pi_2(\boldsymbol{\nu}, \boldsymbol{\theta})$ under M_2 .

In the spirit of Jeffreys (and many others after him) we take (under each of the models) the *same* objective (possibly improper) prior for the common parameter $\boldsymbol{\nu}$ and a proper prior for the conditional distribution of the new parameter $\boldsymbol{\theta} | \boldsymbol{\nu}$ under M_2 , which will be derived similarly to the DB priors in Section 2.2. Note that since $\boldsymbol{\nu}$ occurs in the two models, if we take the same $\pi^N(\boldsymbol{\nu})$ in both, then the (common) arbitrary constants cancel when computing the Bayes factor; however $\boldsymbol{\theta}$ which only occurs in M_2 has to have a proper prior. A common prior for the old parameter only makes sense when $\boldsymbol{\nu}$ has the same meaning in both models (another reason to take $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ orthogonal). Moreover, it is well known that under orthogonality, the specific *common* prior for $\boldsymbol{\nu}$ has little impact on the resulting Bayes factor (see Jeffreys 1961; Kass and Vaidyanathan 1992), thus supporting use of objective priors for common parameters.

Let $\pi^N(\boldsymbol{\nu})$ be an objective (usually either Jeffreys or reference) prior for model f_1 and $\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu})$ the corresponding one for model f_2 ($\boldsymbol{\theta}$ is of interest if the reference prior is used). We define $\pi^N(\boldsymbol{\theta} | \boldsymbol{\nu})$ such that

$$\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu}) = \pi^N(\boldsymbol{\theta} | \boldsymbol{\nu}) \pi^N(\boldsymbol{\nu}).$$

To define the DB priors, let D any of (25) or (26) (other appropriate divergence measures could also be explored). Then we define:

Definition 4.1. (DB priors) Let $c(q, \boldsymbol{\nu}) = \int (1 + \bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}])^{-q} \pi^N(\boldsymbol{\theta} | \boldsymbol{\nu}) d\boldsymbol{\theta}$, and

$$\underline{q} = \inf\{q \geq 0 : c(q, \boldsymbol{\nu}) < \infty\}, \text{ a.e. } \boldsymbol{\nu} \in \Upsilon, \quad q_* = \underline{q} + 1/2$$

If $\underline{q} < \infty$, the D -divergence based prior under M_1 is $\pi_1^D(\boldsymbol{\nu}) = \pi^N(\boldsymbol{\nu})$, and under M_2 is $\pi_2^D(\boldsymbol{\theta}, \boldsymbol{\nu}) =$

$\pi^D(\boldsymbol{\theta} \mid \boldsymbol{\nu}) \pi^N(\boldsymbol{\nu})$, where the (proper) $\pi^D(\boldsymbol{\theta} \mid \boldsymbol{\nu})$ is

$$\pi^D(\boldsymbol{\theta} \mid \boldsymbol{\nu}) = c(q_*, \boldsymbol{\nu})^{-1} (1 + \bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \mid \boldsymbol{\nu}])^{-q_*} \pi^N(\boldsymbol{\theta} \mid \boldsymbol{\nu}) .$$

In this definition we are implicitly using the recommended non-increasing function $h_q(t) = (1+t)^{-q}$, but again other non-increasing functions on $t \in [0, \infty)$ could be explored.

Definition 4.2. (Sum and Minimum DB priors) *The sum DB prior π^S and the minimum DB prior π^M are the DB priors given in definition 4.1 with D being respectively D^S (see (25)) and D^M (see (26)). When needed, we refer to their corresponding c 's and q 's as $c_S, \underline{q}^S, q_*^S$, and $c_M, \underline{q}^M, q_*^M$, respectively.*

We next investigate whether the DB priors are invariant under reparameterizations. Suppose that $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$ and $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\nu})$ are, respectively one-to-one monotone mappings $\boldsymbol{\xi} : \Theta \rightarrow \Theta_\xi$, $\boldsymbol{\eta} : \Upsilon \rightarrow \Upsilon_\eta$. Clearly, the reparameterization $(\boldsymbol{\xi}, \boldsymbol{\eta})$ preserves orthogonality.

The original problem (24) in this parameterization becomes:

$$M_1^* : f_1^*(\mathbf{y} \mid \boldsymbol{\eta}) = f^*(\mathbf{y} \mid \boldsymbol{\xi}_0, \boldsymbol{\eta}) \quad \text{vs.} \quad M_2^* : f_2^*(\mathbf{y} \mid \boldsymbol{\xi}, \boldsymbol{\eta}) = f^*(\mathbf{y} \mid \boldsymbol{\xi}, \boldsymbol{\eta}), \quad (27)$$

where $f^*(\mathbf{y} \mid \boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\nu})) = f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\nu})$ and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(\boldsymbol{\theta}_0)$. We next show that if $\pi^N(\boldsymbol{\nu})$ and $\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu})$ are invariant under these reparameterizations, so are the DB priors. (See Datta and Ghosh, 1995 for a detailed analysis about the invariance of several non informative priors in the presence of nuisance parameters.)

Theorem 1. (Invariance under one-to-one transformations.) *Let $\pi_\nu^D(\boldsymbol{\nu})$ and $\pi_\eta^D(\boldsymbol{\eta})$ be either the sum or the minimum DB priors under M_1 for the original (24), and reparameterized (27) problems, respectively, and similar notation for $\pi_{\boldsymbol{\theta}, \boldsymbol{\nu}}^D(\boldsymbol{\theta}, \boldsymbol{\nu})$ and $\pi_{\boldsymbol{\xi}, \boldsymbol{\eta}}^D(\boldsymbol{\xi}, \boldsymbol{\eta})$, under M_2 . If $\pi_\nu^N(\boldsymbol{\nu}) = \kappa \pi_\eta^N(\boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_\eta(\boldsymbol{\nu})|$, where κ is a constant, and $\pi_{\boldsymbol{\theta}, \boldsymbol{\nu}}^N(\boldsymbol{\theta}, \boldsymbol{\nu}) \propto \pi_{\boldsymbol{\xi}, \boldsymbol{\eta}}^N(\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_{\boldsymbol{\xi}, \boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\nu})|$, then*

$$\pi_\nu^D(\boldsymbol{\nu}) = \kappa \pi_\eta^D(\boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_\eta(\boldsymbol{\nu})|, \quad \pi_{\boldsymbol{\theta}, \boldsymbol{\nu}}^D(\boldsymbol{\theta}, \boldsymbol{\nu}) = \kappa \pi_{\boldsymbol{\xi}, \boldsymbol{\eta}}^D(\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_{\boldsymbol{\xi}, \boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\nu})|.$$

Proof. See Appendix. □

As a consequence, DB Bayes factors are not affected by reparameterizations of the type considered. These are the most natural and interesting reparameterizations of the problem (and indeed other reparameterizations seem questionable). Also, the DB priors are compatible with reduction by sufficiency in the same spirit as in Proposition 2.

4.3 Examples

We next demonstrate the behavior of DB priors and corresponding Bayes factors in a couple of examples. The first is testing the mean of a gamma model, a difficult problem in general. The

second discusses linear models.

4.3.1 Gamma model (Example 6)

Let $\mathbf{y} = (y_1, \dots, y_n)$ be an iid sample from a Gamma model with mean μ , and shape parameter α , that is, from

$$f(y \mid \alpha, \mu) = \left(\frac{\alpha}{\mu}\right)^\alpha \Gamma(\alpha)^{-1} y^{\alpha-1} e^{-y\alpha/\mu}.$$

It is desired to test $H_1 : \mu = \mu_0$ vs. $H_2 : \mu \neq \mu_0$. It is easy to show that μ is orthogonal to α .

The objective (reference) priors are $\pi^N(\alpha) = (\psi^{(1)}(\alpha) - 1/\alpha)^{1/2}$ and $\pi^N(\mu, \alpha) = \mu^{-1}(\psi^{(1)}(\alpha) - 1/\alpha)^{1/2}$, where $\psi^{(1)}$ represents the digamma function. Hence $\pi_2^N(\mu \mid \alpha) = \mu^{-1}$.

The DB priors are $\pi^D(\alpha) = \pi^N(\alpha)$, under both hypotheses and for D either the sum or min divergence. Under H_2 , the conditional sum-DB prior for μ is

$$\pi^S(\mu \mid \alpha) = c_s^{-1}(\alpha) \left[1 + \alpha \frac{(\mu - \mu_0)^2}{\mu\mu_0}\right]^{-1/2} \frac{1}{\mu}$$

where $c_s(\alpha)$ is the proportionality constant

$$c_s(\alpha) = \int_0^\infty \left(1 + \alpha \frac{(t-1)^2}{t}\right)^{-1/2} \frac{1}{t} dt.$$

The conditional min-DB prior is

$$\pi^M(\mu \mid \alpha) = c_m(\alpha)^{-1} \left(1 + \bar{D}^M[(\mu, \mu_0) \mid \alpha]\right)^{-3/2} \frac{1}{\mu}$$

where

$$\bar{D}^M[(\mu, \mu_0) \mid \alpha] = \begin{cases} 2\alpha(\log \frac{\mu}{\mu_0} - 1 + \frac{\mu_0}{\mu}) & \text{if } \mu > \mu_0 \\ 2\alpha(\log \frac{\mu_0}{\mu} - 1 + \frac{\mu}{\mu_0}) & \text{if } \mu \leq \mu_0, \end{cases}$$

and

$$c_m(\alpha) = 2 \int_0^\infty (1 + 2\alpha(t - 1 + e^{-t}))^{-3/2} dt.$$

In Table 7 we show the corresponding Bayes factors B_{12}^S and B_{12}^M for $n = 10$; the null value is $\mu_0 = 10$, and we have considered several combinations of $(\hat{\mu}, \hat{\sigma})$, the maximum likelihood estimates of the mean and standard deviation. When $\hat{\mu} = 12$ (casting doubt on the null), both Bayes factors are very similar, and increasing with $\hat{\sigma}$, an intuitive behavior. When the data shows the most support for the null, that is, when $\hat{\mu} = 10$, the Bayes factors differ, with the sum-DB prior giving the most support to the null.

In contrast with DB priors, it is not possible to derive relatively simple expressions for the intrinsic priors. Hence, in this example, we compare the DB Bayes factors with the intrinsic

	$\hat{\mu} = 10$		$\hat{\mu} = 11$		$\hat{\mu} = 12$	
	B_{12}^S	B_{12}^M	B_{12}^S	B_{12}^M	B_{12}^S	B_{12}^M
$\hat{\sigma} = 0.5$	12.94	2.83	0.005	0.004	$1 \cdot 10^{-5}$	$3 \cdot 10^{-5}$
$\hat{\sigma} = 1$	11.27	2.92	0.353	0.150	0.003	0.003
$\hat{\sigma} = 2$	9.49	3.06	3.102	1.136	0.22	0.12

Table 7: Values of B_{12} for gamma mean testing with $\mu_0 = 10$; we use $n = 10$, and different values of $(\hat{\mu}, \hat{\sigma})$.

	$\mu = 10$			$\mu = 11$			$\mu = 12$		
	B_{12}^S	B_{12}^M	IB_{12}^A	B_{12}^S	B_{12}^M	IB_{12}^A	B_{12}^S	B_{12}^M	IB_{12}^A
$\sigma = 0.5$	13.17	2.93	0.08	0.004	0.003	0.001	$1.4 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$0.1 \cdot 10^{-5}$
$\sigma = 1$	11.15	2.88	0.55	0.33	0.14	0.07	0.003	0.003	0.001
$\sigma = 2$	9.57	3.08	3.71	3.07	1.12	1.23	0.22	0.12	0.07

Table 8: For Gamma model problem, and test $H_1 : \mu = 10$ vs. $H_2 : \mu \neq 10$. In each cell, values of B_{12} and arithmetic intrinsic Bayes factor IB_{12}^A , associated with a sample of size $n = 10$, from a Gamma model with mean μ and standard deviation σ .

arithmetic Bayes factor IB_{12}^A (see Berger and Pericchi 1996). Although IB_{12}^A does not exactly correspond to a Bayes factor derived from a specific prior, it does asymptotically correspond to a Bayes factor derived with the intrinsic arithmetic prior. Since IB_{12}^A is not defined with reduction by sufficiency, the comparison are carried out for (specific) simulated samples with the given parameters. In Table 8 we show the arithmetic intrinsic and DB Bayes factors for testing $H_1 : \mu = 10$, with $n = 10$ and samples generated from Gamma distributions with $\mu \in \{10, 11, 12\}$ and $\sigma \in \{0.5, 1.0, 2.0\}$. The resulting MLEs $(\hat{\mu}, \hat{\sigma})$ in lexicographical order are: $\{(10.02, 0.52), (9.98, 0.99), (9.98, 1.97), (11.01, 0.48), (11.00, 0.99), (10.98, 1.99), (11.99, 0.51), (11.98, 0.99), (12.01, 1.99)\}$. When H_2 is true ($\mu = 11$ or $\mu = 12$), the three measures are rather close. Similar values are also obtained when the ‘null’ model H_1 is true and $\sigma = 2$. In all these cases, the three measures provide support to the true model. Nevertheless, when H_1 is true and the variance is small, the DB Bayes factors are very sensible (with B_{12}^S giving the largest support to the null) but the IB_{12}^A is not, giving support to H_2 . This behavior of IB_{12}^A is likely due to the well known instability of IB_{12}^A when the sample size is small (worsened in this case because the variance is small).

4.3.2 Linear models.

We briefly show next two standard applications of DB priors in the context of linear models. More elaborated examples can be found in Bayarri and García-Donato (2006). Derivations of DB priors for random effects are given in García-Donato and Sun (2005).

Variable selection (Example 7). Consider the full rank General Linear Model $\{N_n(\mathbf{y} \mid \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_e\boldsymbol{\beta}_e, \sigma^2\mathbf{I}_n)\}$ and the problem of testing $H_1 : \boldsymbol{\beta}_e = \mathbf{0}$. After the usual orthogonal reparameterization (see e.g. Zellner and Siow 1984) and taking $n^* = n$ and $\pi^N(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1}$, the DB priors are

$$\pi_1^D(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1}, \quad \pi_2^D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1}Ca_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\mathbf{V}^t\mathbf{V})^{-1}),$$

where k_e is the dimension of $\boldsymbol{\beta}_e$ and

$$\mathbf{V} = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_e, \quad \mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t.$$

Both DB priors coincide and also coincide with the priors proposed by Jeffreys (1961) and Zellner and Siow (1980, 1984). This ‘coincidence’ was the original motivation for the specific choice $\underline{q} + 1/2$ in the definition of DB priors (see García-Donato, 2003 for details). Zellner-Siow priors have been extensively studied in Bayarri and García-Donato (2007) for general null models given by linear restrictions of the regression parameters, and for not necessarily full rank design matrices.

Note that the exact matching of JZS and DB priors only occur if the effective sample size is $n^* = n$. However, n^* might well depend on the design matrix (or covariates). For example, in the linear model $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$, with $\mathbf{X} : n \times 1$ and θ scalar, it is intuitively clear that if $\mathbf{X} = (1, \dots, 1)^t$ then n^* should be n , but if $\mathbf{X} = (1, \varepsilon, \dots, \varepsilon)^t$ with ε very small, then n^* should be 1. The effective sample size defined in Berger et al. (2007) satisfies this requirement but other definitions might not. Extended investigation of this issue is beyond the scope of this paper and will be pursued elsewhere.

Since comparison among existing objective Bayesian testing procedures for the Linear model have extensively been given in the literature, including Bayes factors derived with JZS priors, we skip them here (see Berger, Ghosh and Mukhopadhyay, 2003; Liang et al., 2007; Bayarri and García-Donato, 2007).

ANOVA models (Example 8). We next consider testing equality of group means in a classical ANOVA setting. Let

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, p; \quad j = 1, \dots, r,$$

where ϵ_{ij} are i.i.d. $\sim N(0, \sigma^2)$. We want to test $H_1 : \alpha = \dots = \alpha_p$ vs. H_2 : the α_i are not equal. With the usual restriction $\sum_{i=1}^p \alpha_i = 0$, α_p is a function of $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{p-1})$, and H_1 reduces to $\boldsymbol{\alpha} = \mathbf{0}$. Here, the common parameters μ, σ are orthogonal to $\boldsymbol{\alpha}$. An easy derivation shows

that both symmetrized divergence measures are

$$\bar{D}[(\boldsymbol{\alpha}, \mathbf{0}) | \sigma, \mu] = \frac{\boldsymbol{\alpha}^t \mathbf{A} \boldsymbol{\alpha}}{p \sigma^2},$$

where $\mathbf{A} = \mathbf{I}_{p-1} + \mathbf{1}_{p-1} \mathbf{1}_{p-1}^t$. Hence, the DB priors are $\pi_1^D(\mu, \sigma) = \sigma^{-1}$ and

$$\pi_2^D(\boldsymbol{\alpha}, \mu, \sigma) = \sigma^{-1} C a_{p-1}(\boldsymbol{\alpha} | \mathbf{0}, p \sigma^2 \mathbf{A}^{-1}).$$

5 Approximations and computation

In this Section, we derive simple approximations to DB priors and show their connections with already existing proposal. We also exploit the connection between DB Bayes factors and a corrected Bayes factor computed with usual (possibly improper) non-informative priors to propose easy MCMC computation of DB Bayes factors.

5.1 Approximated DB priors

It is well known (see Kullback 1968; Schervish 1995) that the Kullback-Leibler divergence measures can be approximated up to second order using the expected Fisher information, so that:

$$D^S[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}] \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t J_\theta(\boldsymbol{\theta}_0, \boldsymbol{\nu}) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \approx D^M[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}],$$

where $J_\theta(\boldsymbol{\theta}_0, \boldsymbol{\nu})$ is the block in Fisher information matrix corresponding to $\boldsymbol{\theta}$, evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\nu})$. Hence, for the problem (24) (recall that $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ are orthogonal), the DB priors π^D (either π^S or π^M) can be approximated by $\pi_1^D(\boldsymbol{\nu}) = \pi^N(\boldsymbol{\nu})$ and

$$\pi^D(\boldsymbol{\theta} | \boldsymbol{\nu}) = c(q_*, \boldsymbol{\nu})^{-1} h_{q_*} \left((\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t \frac{J_\theta(\boldsymbol{\theta}_0, \boldsymbol{\nu})}{n^*} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right) \pi^N(\boldsymbol{\theta} | \boldsymbol{\nu}), \quad (28)$$

where now $q_* = \underline{q} + 1/2$, and \underline{q} is the infimum of q values for which the conditional defined in (28) (in terms of Fisher information) is proper.

The cases when $\pi^N(\boldsymbol{\theta} | \boldsymbol{\nu})$ does not depend on $\boldsymbol{\theta}$ (so $\boldsymbol{\theta}$ behaves asymptotically as a location parameter) are specially interesting. It is easy to show that then $\underline{q} = k/2$, where k is the dimension of $\boldsymbol{\theta}$ and hence

$$\pi^D(\boldsymbol{\theta} | \boldsymbol{\nu}) \approx C a_k(\boldsymbol{\theta} | \boldsymbol{\theta}_0, n^* J_\theta^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\nu})), \quad (29)$$

The conditional prior (29) has been interpreted by many authors (see for instance Kass and Wasserman 1995) as the generalization of Jeffreys' ideas to multivariate problems.

Moreover, if $h_q(t) = e^{-qt}$ is used instead, then π^D would essentially be the normal unit information priors, as defined by Kass and Wasserman (1995) and further studied by Raftery

(1998). Note that we have shown that this proposals can be interpreted as approximated DB priors only when $\boldsymbol{\theta}$ is asymptotically a location parameter.

5.2 Computation of Bayes factor

Interestingly enough, and similarly to other objective Bayesian proposals (like the intrinsic and fractional Bayes factors), it can be shown that Bayes factors computed with DB priors, B_{21}^D , can be expressed as an (invalid) Bayes factor computed with non-informative (usually improper) priors, B_{21}^N , multiplied by a correction factor. This expression also allows for easy computation of DB Bayes factors when B_{21}^N is easy to compute.

Lemma 5.1. *For problem (24) (with $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ orthogonal), let B_{12}^N denote the Bayes factor computed using $\pi_1^N(\boldsymbol{\nu})$ and $\pi_2^N(\boldsymbol{\theta}, \boldsymbol{\nu})$, then for both the sum and min DB-priors*

$$B_{21}^D = B_{21}^N \times E^{\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu} | \mathbf{y})} \left(c(q_*, \boldsymbol{\nu})^{-1} h_{q_*}(\bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}]) \right). \quad (30)$$

Proof. See Appendix. □

Computation of B_{21}^N is often simpler than computation of proper Bayes factors. Then a sample (usually MCMC) from the posterior distribution $\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu} | \mathbf{y})$ can be used to evaluate the expectation in (30), thus considerably simplifying computation of B_{12}^S or B_{12}^M . This is actually how we computed the Bayes factors for Example 6 in Section 4.3.1.

Moreover, if n is large (relative to the dimension of $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\nu})$, assumed fixed) we can approximate (30) using asymptotic expressions to posterior distribution along with the approximated DB priors given in (28).

We illustrate the approach in a simple setting. First we assume that the asymptotic posterior distribution is given by (see conditions in e.g. Berger 1985),

$$\pi^N(\boldsymbol{\theta}, \boldsymbol{\nu} | \mathbf{y}) \approx N(\hat{\boldsymbol{\phi}}, \mathbf{J}^{-1}(\hat{\boldsymbol{\phi}})),$$

where $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})$ is the (assumed to exist) maximum likelihood estimate of $(\boldsymbol{\theta}, \boldsymbol{\nu})$ and $\mathbf{J} = \mathbf{J}_\theta \oplus \mathbf{J}_\nu$ is the (block diagonal) expected Fisher information matrix of $f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu})$.

Next we assume that $\pi^N(\boldsymbol{\theta} | \boldsymbol{\nu})$ does not depend on $\boldsymbol{\theta}$, so the approximating (conditional) DB prior is the Cauchy prior in (29). As a notational device, it will be convenient to then write $\pi^N(\boldsymbol{\theta} | \boldsymbol{\nu})$ as $\pi^N(\boldsymbol{\theta}_0 | \boldsymbol{\nu})$. Expressing the Cauchy density (29) in the usual way as a scale mixture of a Normal and an inverse gamma, and using the asymptotic posterior, the DB Bayes factors, as given in (30), can be approximated by

$$B_{21}^D \approx B_{21}^N \int \int \frac{1}{\pi^N(\boldsymbol{\theta}_0 | \boldsymbol{\phi})} N_k(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}_0, \Sigma(\boldsymbol{\nu}, t)) N_p(\boldsymbol{\nu} | \hat{\boldsymbol{\nu}}, \mathbf{J}_\nu(\hat{\boldsymbol{\phi}})) d\boldsymbol{\nu} IGa(t | \frac{1}{2}, \frac{1}{2}) dt,$$

where p is the dimension of $\boldsymbol{\nu}$ and $\Sigma(\boldsymbol{\nu}, t) = tn\mathbf{J}_\theta^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\nu}) + \mathbf{J}_\theta^{-1}(\hat{\boldsymbol{\phi}})$. A similar asymptotic approximation to B_{12}^N , finally gives the desired asymptotic approximation to the DB Bayes factor:

$$B_{21}^D \approx \frac{p(\mathbf{y} | \hat{\boldsymbol{\phi}})}{p(\mathbf{y} | \boldsymbol{\theta}_0, \hat{\boldsymbol{\nu}})} (2\pi)^{k/2} \frac{1}{\det \mathbf{J}_\theta(\hat{\boldsymbol{\phi}})^{1/2}} \\ \times \int \int \frac{\pi^N(\hat{\boldsymbol{\theta}} | \hat{\boldsymbol{\nu}})}{\pi^N(\boldsymbol{\theta}_0 | \boldsymbol{\nu})} N_k(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}_0, \Sigma(\boldsymbol{\nu}, t)) N_p(\boldsymbol{\nu} | \hat{\boldsymbol{\nu}}, \mathbf{J}_\nu(\hat{\boldsymbol{\phi}})) IGa(t | \frac{1}{2}, \frac{1}{2}) d\boldsymbol{\nu} dt,$$

which is very easy to evaluate by simple Monte Carlo. Note that arbitrary constants in the possibly improper $\pi^N(\boldsymbol{\theta} | \boldsymbol{\nu})$ cancel out in the expression above.

6 Summary and conclusions

We propose a new class of priors for objective Bayes hypothesis testing based on Divergence measures, which we call ‘Divergence Based’ (DB) priors. For divergence measures, we propose use of symmetrized versions (sum and the minimum) of Kullback Liebler divergences. The resulting DB priors are usually easy to compute and have a number of desirable properties as invariance under reparameterizations, evidence consistency and compatibility with sufficient statistics. We explore DB priors in a series of estudy examples, in which they show to be intuitively sound and to produce sensible Bayes factors. This is so even for irregular models and improper likelihoods, which are extremely challenging scenarios for other objective Bayes testing methodologies. We recommend use of the sum-DB prior when it exists because it is considerably easier to compute than the min-DB prior and seems to exhibit a nicer behavior.

The DB priors seem to behave similarly to the arithmetic intrinsic prior (when defined). Also, in normal scenarios, they exactly reproduce the proposals of Jeffreys (1961) and Zellner and Siow (1980, 1984), so that they can be considered an extension of these classical proposals to non-normal situations. Approximations to DB priors are also shown to be connected with other proposals as the unit information priors. Finally, we also provide asymptotic approximations to DB Bayes factors for large sample size.

The definition of DB priors are based on particular choices of both 1) an ‘objective prior’ π^N for estimation problems and 2) an equivalent sample size n^* . Of course, there is no general agreement in the literature about a single definition for any of these concepts (and there might never be). We think that any sensible proposals would produce nice results, but this in an issue that needs to be further investigated. We recommend, when possible, use of the *reference prior* (Berger and Bernardo, 1992) and of the equivalent sample size in Berger et al. (2007).

Other apparently arbitrary choices that we made were those of h_q and of q_* , however they were based on some compelling arguments

- Choice of $h_q(t) = (1+t)^{-q}$ was specifically chosen to reproduce in the normal case Jeffreys-Zellner-Siow priors, but there are other reasons for it. A compelling reason is that it is a simple function resulting in Bayes factors with nice properties; another simple function to use could be the exponential, but this results in normal priors that are not *evidence consistent*. Also, h_q results in priors with very heavy tails, which is important so as not to ‘knock-out’ the likelihood when data is not well explained by the null model. However, we do not rule out that other choices of functions $h(t)$ which are decreasing for $t \in [0, \infty)$, with maximum at zero, and producing proper DB-type priors could work better in specific scenarios.
- Choice of $q^* = \underline{q} + 1/2$. In principle, any $\underline{q} + \delta$ could be used. As a matter of fact, we do not expect that the specific choice of δ matters much as long as $\delta \in (0, 1)$ (needed to produce priors with heavy tails and no moments), but this again needs further investigation. We recommend use of $\delta = 1/2$ because it is the value reproducing Jeffreys proposal.

7 Acknowledgements

This research was supported in part by the Spanish Ministry of Science and Education, under grant MTM2004-03290.

References

- [1] Bayarri, M.J. and García-Donato, G. (2007), “Extending Conventional priors for Testing General Hypotheses in Linear Models,” *Biometrika*, in press.
- [2] Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis (2nd ed.)*, New York: Springer-Verlag.
- [3] Berger, J. O. and Bernardo, J. M. (1992), “On the development of the reference prior method.”. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 35-60. Oxford: Oxford University Press.
- [4] Berger, J.O. and Delampady, M. (1987), “Testing precise hypotheses,” *Statistical Science*, **3**, 317-352.
- [5] Berger, J.O. and Mortera, J. (1999), “Default Bayes Factors for Nonnested Hypothesis Testing,” *Journal of the American Statistical Association*, **94**, 542-554.
- [6] Berger, J.O., Ghosh, J.K. and Mukhopadhyay, N. (2003), “Approximations to the Bayes factor in model selection problems and consistency issues,” *Journal of Statistical Planning and Inference*, **112**, 241-58.

- [7] Berger, J. O. and Pericchi, L. R. (1996), “The intrinsic Bayes factor for model selection and prediction,” *Journal of the American Statistical Association*, **91**, 109-22.
- [8] Berger, J. O., and Pericchi, R. L. (2001), “Objective Bayesian methods for model selection: introduction and comparison (with discussion)”. In *Model Selection* (ed P. Lahiri), pp. 135-207. Institute of Mathematical Statistics Lecture Notes-Monograph Series, volume 38. Beachwood Ohio.
- [9] Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998), “Bayes factors and marginal distributions in invariant situations,” *Sankhya A*, **60**, 307-21.
- [10] Berger, J.O. and Sellke, T. (1987), “Testing a point null hypothesis: the irreconcilability of P-values and evidence,” *Journal of the American Statistical Association*, **82**, 112-122.
- [11] Berger, J. at all. (2007). “Extensions and generalizations of BIC”, ISDS Working paper, in preparation.
- [12] Bernardo, J.M. and Rueda, R. (2002), “Bayesian hypothesis testing: A reference approach,” *International Statistical Review*, **70**, 351-372.
- [13] Bernardo, J.M. (2005), “Intrinsic credible regions: An objective Bayesian approach to interval estimation,” *Test*, **14**, 317-384.
- [14] Clyde, M. (1999), “Bayesian Model Averaging and Model Search Strategies (with discussion)”. In *Bayesian Statistics 6* (eds J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith), pp. 157-185. Oxford: Oxford University Press.
- [15] Clyde, M., DeSimone, H. and Parmigiani, G. (1996), “Prediction via Orthogonalized Model Mixing,” *Journal of the American Statistical Association*, **91**, 1197-1208.
- [16] Conover, W. J. (1971), *Practical nonparametric statistics*, New York: John Wiley and Sons.
- [17] Cox, D. R. and Reid, N. (1987), “Parameter orthogonality and approximate conditional inference,” *Journal of the Royal Statistical Society B*, **49**, 1-39.
- [18] Datta, G.S. and Ghosh, M. (1995), “On the invariance of noninformative priors”, *Annals of Statistics*, **24**, 141-159.
- [19] De Santis, F. and Spezzaferrri, F. (1999), “Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach,” *International Statistics Review*, **67**, 267-286.
- [20] García-Donato, G. (2003), *Factores Bayes Factores Bayes Convencionales: Algunos Aspectos Relevantes*, Unpublished PhD Thesis, Department of Statistics, University of Valencia.

- [21] García-Donato, G. and Sun, D. (2005), “Objective Priors for Model Selection in One-Way Random Effects Models.” Submitted to *The Canadian journal of Statistics*.
- [22] Hoeting, J.A, Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999), “Bayesian Model Averaging: A Tutorial,” *Statistical Science*, **14**, 382-417.
- [23] Ibrahim, J. and Laud, P. (1994), “A Predictive Approach to the Analysis of Designed Experiments,” *Journal of the American Statistical Association*, **89**, 309-319
- [24] Jeffreys, H. (1961). *Theory of Probability*, 3rd edn. London: Oxford University Press.
- [25] Kass, R. E. and Raftery, A. E. (1995), “Bayes factors,” *Journal of the American Statistical Association*, **90**, 773-95.
- [26] Kass, R. E. and Vaidyanathan, S. (1992), “Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions,” *Journal of the Royal Statistical Society B*, **54**, 129-44.
- [27] Kass, R. E. and Wasserman, L. (1995), “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *Journal of the American Statistical Association*, **90**, 928-34.
- [28] Kullback, S. (1968), *Information Theory and Statistics*, New York: Dover Publications, Inc.
- [29] Laud, P.W. and Ibrahim, J. (1995), “Predictive Model Selection,” *Journal of the Royal Statistical Society B*, **57**, 247-262.
- [30] Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. O. (2007), “Mixtures of g -priors for Bayesian Variable Selection,” *Journal of the American Statistical Society*, in press.
- [31] Montgomery, D. (2001), *Introduction to Statistical Quality Control*, 4th edn. John Wiley and Sons, Inc.
- [32] Moreno, E., Bertolino, F. and Racugno, W. (1998), “An intrinsic limiting procedure for model selection and hypotheses testing,” *Journal of the American Statistical Association*, **93**, 1451-60.
- [33] O’Hagan, A. (1995), “Fractional Bayes factors for model comparison (with discussion),” *Journal of the Royal Statistical Society, B*, **57**, 99-138.
- [34] Pauler, D. (1998), “The Schwarz Criterion and Related Methods for Normal Linear Models,” *Biometrika*, **85**, 13-27.

- [35] Pauler, D.K., Wakefield, J.C. and Kass, R.E. (1999), “Bayes factors and approximations for variance component models,” *Journal of the American Statistical Association*, **94**, 1242-1253.
- [36] Pérez, J.M. and Berger, J. (2001), “Analysis of mixture models using expected posterior priors, with application to classification of gamma ray bursts.” In *Bayesian Methods, with applications to science, policy and official statistics*, (eds E. George and P. Nanopoulos), pp. 401-410. Official Publications of the European Communities, Luxembourg.
- [37] Pérez, J. M. and Berger, J. O. (2002), “Expected posterior prior distributions for model selection,” *Biometrika*, **89**, 491-512.
- [38] Pérez, S. (2005), *Métodos Bayesianos objetivos de comparación de medias*, Unpublished PhD Thesis, Department of Statistics, University of Valencia.
- [39] Raftery, A.E. (1998), “Bayes factor and BIC: comment on Weakliem,” Technical Report 347, Department of Statistics, University of Washington.
- [40] Schervisch, M.J. (1995), *Theory of Statistics*. New York: Springer-Verlag.
- [41] Tanner, M.A. (1996), *Tools for Statistical Inference. Methods for the exploration of Posterior Distributions and Likelihood Functions*. 3rd edn. New York: Springer Verlag.
- [42] Zellner, A. and Siow, A. (1980), “Posterior odds ratio for selected regression hypotheses”. In *Bayesian Statistics 1* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 585-603. Valencia: University Press.
- [43] Zellner, A. and Siow, A. (1984). *Basic Issues in Econometrics*. Chicago: University of Chicago Press.

Appendix. Proofs.

Proof of Lemma 2.2. That the mode of π^D is θ_0 is an immediate consequence of the property $\bar{D}[\theta, \theta_0] \geq \bar{D}[\theta_0, \theta_0]$. For the second part of the result, note that (see Kullback, 1968)

$$KL[\theta_0, \theta] \approx \frac{1}{2}(\theta - \theta_0)^t \mathbf{J}(\theta_0)(\theta - \theta_0),$$

when θ is in a neighborhood of θ_0 . Then

$$\bar{D}[\theta, \theta_0] \approx (\theta - \theta_0)^t \frac{\mathbf{J}(\theta_0)}{n^*}(\theta - \theta_0),$$

and the result follows.

Proof of Proposition 1. Let $\bar{D}^*[\boldsymbol{\xi}, \boldsymbol{\xi}_0]$ be the unitary measure of divergence between $f_1^*(\mathbf{y})$ and $f_2^*(\mathbf{y} \mid \boldsymbol{\xi})$ in (14). It is well known that KL remains the same under one-to-one reparameterizations, and clearly $\bar{D}^*[\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\xi}(\boldsymbol{\theta}_0)] = \bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]$. Now, by definition of DB priors, and using the relation between π_θ^N and π_ξ^N , it follows that

$$\begin{aligned}\pi_\theta^D(\boldsymbol{\theta}) &= c(q_*) h_{q_*}(\bar{D}[\boldsymbol{\theta}, \boldsymbol{\theta}_0]) \pi_\theta^N(\boldsymbol{\theta}) = c(q_*) h_{q_*}(\bar{D}^*[\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\xi}(\boldsymbol{\theta}_0)]) \pi_\xi^N(\boldsymbol{\xi}(\boldsymbol{\theta})) |\mathcal{J}_\xi(\boldsymbol{\theta})| \\ &= \pi_\xi^D(\boldsymbol{\xi}(\boldsymbol{\theta})) |\mathcal{J}_\xi(\boldsymbol{\theta})|.\end{aligned}$$

Proof of Proposition 2. Let $D^*[\boldsymbol{\theta}, \boldsymbol{\theta}_0]$ be the symmetric divergence between $f_1^*(\mathbf{t})$ and $f_2^*(\mathbf{t} \mid \boldsymbol{\theta})$ in (15), and hence $D^*[\boldsymbol{\theta}, \boldsymbol{\theta}_0] = D[\boldsymbol{\theta}, \boldsymbol{\theta}_0]$. The result now follows from the assumption that neither π^N nor n^* change when the problem is formulated in terms of sufficient statistics.

Proof of Lemma 3.3. First we show that (18) implies that $B_{12}^\pi \rightarrow 0$ as $\bar{y} \rightarrow 0$. Assume $\int_0^1 \mu^{-k} \pi(\mu) d\mu = \infty$. Then

$$\lim_{\bar{y} \rightarrow 0} m_2(\mathbf{y}) = \lim_{\bar{y} \rightarrow 0} \int_0^\infty \mu^{-n} e^{-n\bar{y}/\mu} \pi(\mu) d\mu \geq \int_0^1 \mu^{-k} \pi(\mu) d\mu = \infty,$$

and the result follows. To show the converse, note that, since $\pi(\mu)$ is proper,

$$\lim_{\bar{y} \rightarrow 0} \int_1^\infty \mu^{-n} e^{-n\bar{y}/\mu} \pi(\mu) d\mu < \infty. \quad (31)$$

Now, by contradiction suppose that for $n \geq k$, $\int_0^1 \mu^{-k} \pi(\mu) d\mu < \infty$, so in particular $\int_0^1 \mu^{-n} \pi(\mu) d\mu < \infty$, and hence the limiting function $g(\mu) = \mu^{-n} \pi(\mu)$ is integrable; now, the Dominated Convergence Theorem gives

$$\lim_{\bar{y} \rightarrow 0} \int_0^1 \mu^{-n} e^{-n\bar{y}/\mu} \pi(\mu) d\mu = \int_0^1 \mu^{-n} \pi(\mu) d\mu < \infty,$$

which jointly with (31) contradicts the assumption of $B_{12}^\pi \rightarrow 0$ as $\bar{y} \rightarrow 0$, proving the result.

Proof of Lemma 3.5. It can easily be seen that, as $T \rightarrow \infty$

$$B_{21}^\pi \rightarrow e^{-n\theta_0} \int_{-\infty}^\infty e^{n\theta} \pi(\theta) d\theta,$$

Now, $\forall n \geq k$, it follows that

$$\int_{-\infty}^\infty e^{n\theta} \pi(\theta) d\theta \geq \int_{-\infty}^\infty e^{k\theta} \pi(\theta) d\theta \geq \int_{\theta_0}^\infty e^{k\theta} \pi(\theta) d\theta,$$

proving the lemma.

Proof of Theorem 1. By definition, the DB priors for the reparameterized problem are $\pi_{\nu}^D(\boldsymbol{\nu}) = \pi_{\nu}^N(\boldsymbol{\nu})$ and (recall $h_q(t) = (1+t)^{-q}$)

$$\pi_{\xi, \eta}^D(\boldsymbol{\xi}, \boldsymbol{\eta}) = c^*(q_*, \boldsymbol{\eta})^{-1} h_{q_*}(\bar{D}^*[(\boldsymbol{\xi}, \boldsymbol{\xi}_0) | \boldsymbol{\eta}]) \pi_{\xi|\eta}^N(\boldsymbol{\xi} | \boldsymbol{\eta}) \pi_{\eta}^N(\boldsymbol{\eta}),$$

where $\bar{D}^*[(\boldsymbol{\xi}, \boldsymbol{\xi}_0) | \boldsymbol{\eta}]$ is the corresponding unitary measure of divergence between the competing models f_1^* and f_2^* in (27) and

$$c^*(q_*, \boldsymbol{\eta}) = \int h_{q_*}(\bar{D}^*[(\boldsymbol{\xi}, \boldsymbol{\xi}_0) | \boldsymbol{\eta}]) \pi_{\xi|\eta}^N(\boldsymbol{\xi} | \boldsymbol{\eta}) d\boldsymbol{\xi}.$$

It can be easily shown that $\bar{D}^*[(\boldsymbol{\xi}, \boldsymbol{\xi}_0) | \boldsymbol{\eta}] = \bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}]$. Also, under the assumptions of the theorem, $\pi_{\theta, \nu}^N(\boldsymbol{\theta}, \boldsymbol{\nu}) = \kappa_2 \pi_{\xi, \eta}^N(\boldsymbol{\xi}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_{\xi, \eta}(\boldsymbol{\theta}, \boldsymbol{\nu})|$, where κ_2 is a constant. Then

$$\pi_{\theta|\nu}^N(\boldsymbol{\theta}, \boldsymbol{\nu}) = \frac{\kappa_2}{\kappa} \pi_{\xi|\eta}^N(\boldsymbol{\xi}(\boldsymbol{\theta}) | \boldsymbol{\eta}(\boldsymbol{\nu})) |\mathcal{J}_{\xi}(\boldsymbol{\theta})|,$$

and hence

$$c(q_*, \boldsymbol{\nu}) = \frac{\kappa_2}{\kappa} c^*(q_*, \boldsymbol{\eta}(\boldsymbol{\nu})),$$

and the result follows.

Proof of Lemma 5.1. For $i = 1, 2$, let $m_i^D(\mathbf{y})$ and $m_i^N(\mathbf{y})$ denote the prior predictive marginals obtained with π_i^D and π_i^N , respectively. By definition of DB priors, $m_i^N(\mathbf{y}) = m_i^D(\mathbf{y})$, and hence

$$B_{21}^D = \frac{m_2^D(\mathbf{y})}{m_1^D(\mathbf{y})} = \frac{m_2^N(\mathbf{y})}{m_1^N(\mathbf{y})} \frac{m_2^D(\mathbf{y})}{m_2^N(\mathbf{y})} = B_{21}^D \frac{m_2^D(\mathbf{y})}{m_2^N(\mathbf{y})}.$$

Finally

$$\begin{aligned} m_2^D(\mathbf{y}) &= \int f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}) \pi^D(\boldsymbol{\theta}, \boldsymbol{\nu}) d\boldsymbol{\theta} d\boldsymbol{\nu} \\ &= \int f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\nu}) c(q_*, \boldsymbol{\nu})^{-1} h_{q_*}(\bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}]) \pi^N(\boldsymbol{\theta}, \boldsymbol{\nu}) d\boldsymbol{\theta} d\boldsymbol{\nu} \\ &= m_2^N(\mathbf{y}) \int c(q_*, \boldsymbol{\nu})^{-1} h_{q_*}(\bar{D}[(\boldsymbol{\theta}, \boldsymbol{\theta}_0) | \boldsymbol{\nu}]) \pi^N(\boldsymbol{\theta}, \boldsymbol{\nu} | \mathbf{y}) d\boldsymbol{\theta} d\boldsymbol{\nu}, \end{aligned}$$

and the result holds.