Bayesian Analysis (2015) **10**, Number 1, pp. 189–221

# Overall Objective Priors*

James O. Berger†, Jose M. Bernardo‡, and Dongchu Sun§

**Abstract.** In multi-parameter models, reference priors typically depend on the parameter or quantity of interest, and it is well known that this is necessary to produce objective posterior distributions with optimal properties. There are, however, many situations where one is simultaneously interested in all the parameters of the model or, more realistically, in functions of them that include aspects such as prediction, and it would then be useful to have a single objective prior that could safely be used to produce reasonable posterior inferences for all the quantities of interest. In this paper, we consider three methods for selecting a single objective prior and study, in a variety of problems including the multinomial problem, whether or not the resulting prior is a reasonable overall prior.

**Keywords:** Joint Reference Prior, Logarithmic Divergence, Multinomial Model, Objective Priors, Reference Analysis.

## 1 Introduction

### 1.1 The problem

Objective Bayesian methods, where the formal prior distribution is derived from the assumed model rather than assessed from expert opinions, have a long history (see *e.g.*, Bernardo and Smith, 1994; Kass and Wasserman, 1996, and references therein). Reference priors (Bernardo, 1979, 2005; Berger and Bernardo, 1989, 1992a,b, Berger, Bernardo and Sun, 2009, 2012) are a popular choice of objective prior. Other interesting developments involving objective priors include Clarke and Barron (1994), Clarke and Yuan (2004), Consonni, Veronese and Gutiérrez-Peña (2004), De Santis et al. (2001), De Santis (2006), Datta and Ghosh (1995a; 1995b), Datta and Ghosh (1996), Datta et al. (2000), Ghosh (2011), Ghosh, Mergel and Liu (2011), Ghosh and Ramamoorthi (2003), Liseo (1993), Liseo and Loperfido (2006), Sivaganesan (1994), Sivaganesan, Laud and Mueller (2011) and Walker and Gutiérrez-Peña (2011).

In single parameter problems, the reference prior is uniquely defined and is invariant under reparameterization. However, in multiparameter models, the reference prior depends on the quantity of interest, *e.g.*, the parameter concerning which inference is being performed. Thus, if data $\boldsymbol{x}$ are assumed to have been generated from $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, with $\boldsymbol{\omega} \in \boldsymbol{\Omega} \subset \Re^k$, and one is interested in $\theta(\boldsymbol{\omega}) \in \Theta \subset \Re$, the reference prior $\pi_\theta(\boldsymbol{\omega})$, will typically depend on $\theta$; the posterior distribution, $\pi_\theta(\boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi_\theta(\boldsymbol{\omega})$, thus also depends on $\theta$, and inference for $\theta$ is performed using the corresponding marginal reference posterior for $\theta(\boldsymbol{\omega})$, denoted $\pi_\theta(\theta \,|\, \boldsymbol{x})$. The dependence of the reference prior

*Related articles: DOI: 10.1214/14-BA935, DOI: 10.1214/14-BA936, DOI: 10.1214/14-BA937, DOI: 10.1214/14-BA938; rejoinder at DOI: 10.1214/15-BA943.

†Duke University, USA and King Abdulaziz University, Saudi Arabia, berger@stat.duke.edu
‡Universitat de València, Spain, jose.m.bernardo@uv.es
§University of Missouri-Columbia, USA, sund@missouri.edu

 DOI: 10.1214/14-BA915

on the quantity of interest has proved necessary to obtain objective posteriors with appropriate properties – in particular, to have good frequentist coverageproperties (when attainable) and to avoid marginalization paradoxes and strong inconsistencies.

There are however many situations where one is *simultaneously* interested in all the parameters of the model or perhaps in several functions of them. Also, in prediction and decision analysis, parameters are not themselves the object of direct interest and yet an overall prior is needed to carry out the analysis. Another situation in which having an overall prior would be beneficial is when a user is interested in a non-standard quantity of interest (*e.g.*, a non-standard function of the model parameters), and is not willing or able to formally derive the reference prior for this quantity of interest. Computation can also be a consideration; having to separately do Bayesian computations with a different reference prior for each parameter can be onerous. Finally, when dealing with non-specialists it may be best pedagogically to just present them with one overall objective prior, rather than attempting to explain the technical reasons for preferring different reference priors for different quantities of interest.

To proceed, let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) = \{\theta_1(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ be the set of $m > 1$ functions of interest. Our goal is to find a joint prior $\pi(\boldsymbol{\omega})$ whose corresponding marginal posteriors, $\{\pi(\theta_i \,|\, \boldsymbol{x})\}_{i=1}^m$, are sensible from a reference prior perspective. This is not a well-defined goal, and so we will explore various possible approaches to the problem.

**Example 1.1. Multinomial Example:** Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is multinomial $\text{Mu}(\boldsymbol{x} \,|\, n; \theta_1, \ldots, \theta_m)$, where $\sum_{i=1}^m x_i = n$, and $\sum_{i=1}^m \theta_i = 1$. In Berger and Bernardo (1992b), the reference prior is derived when the parameter $\theta_i$ is of interest, and this is a different prior for each $\theta_i$, as given in the paper. The reference prior for $\theta_i$ results in a Beta reference marginal posterior $\text{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$. We would like to identify a single joint prior for $\boldsymbol{\theta}$ whose marginal posteriors could be expected to be close to each of these reference marginal posteriors, in some average sense.

## 1.2 Background

It is useful to begin by recalling earlier efforts at obtaining an overall reference prior. There have certainly been analyses that can be interpreted as informal efforts at obtaining an overall reference prior. One example is given in Berger and Sun (2008) for the five parameter bivariate normal model. Priors for all the quantities of interest that had previously been considered for the bivariate normal model (21 in all) were studied from a variety of perspectives. One such perspective was that of finding a good overall prior, defined as one which yielded reasonable frequentist coverage properties when used for at least the most important quantities of interest. The conclusion was that the prior $\pi^o(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_1\sigma_2(1-\rho^2)]$, where the $\mu_i$ are the means, the $\sigma_i$ are the standard deviations, and $\rho$ is the correlation in the bivariate normal model, was a good choice for the overall prior.

We now turn to some of the more formal efforts to create an overall objective prior.

### Invariance-based priors

If $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$ has a group invariance structure, then the recommended objective prior is typically the right-Haar prior. Often this will work well for all parameters that define the

invariance structure. For instance, if the sampling model is $\mathrm{N}(x_i \,|\, \mu, \sigma)$, the right-Haar prior is $\pi(\mu, \sigma) = \sigma^{-1}$, and this is fine for either $\mu$ or $\sigma$ (yielding the usual objective posteriors). Such a nice situation does not always obtain, however.

**Example 1.2. Bivariate Normal Distribution:** The right-Haar prior is not unique for the bivariate normal problem. For instance, two possible right-Haar priors are $\pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_1^2(1-\rho^2)]$ and $\pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = 1/[\sigma_2^2(1-\rho^2)]$. In Berger and Sun (2008) it is shown that $\pi_i$ is fine for $\mu_i$, $\sigma_i$ and $\rho$, but leads to problematical posteriors for the other mean and standard deviation.

The situation can be even worse if the right-Haar prior is used for other parameters that can be considered.

**Example 1.3. Multi-Normal Means:** Let $x_i$ be independent normal with mean $\mu_i$ and variance 1, for $i = 1, \cdots, m$. The right-Haar prior for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ is just a constant, which is fine for each of the individual normal means, resulting in a sensible $\mathrm{N}(\mu_i \,|\, x_i, 1)$ posterior for each individual $\mu_i$. But this prior is bad for overall quantities such as $\theta = \frac{1}{m}|\boldsymbol{\mu}|^2 = \frac{1}{m}\sum_{i=1}^m \mu_i^2$, as discussed in Stein (1959) and Bernardo and Smith (1994, p. 365). For instance, the resulting posterior mean of $\theta$ is $[1 + \frac{1}{m}\sum_{i=1}^m x_i^2]$, which is inconsistent as $m \to \infty$ (assuming $\frac{1}{m}\sum_{i=1}^m \mu_i^2$ has a limit); indeed, it is easy to show that then $[1 + \frac{1}{m}\sum_{i=1}^m x_i^2] \to [\theta_T + 2]$, where $\theta_T$ is the true value of $\theta$. Furthermore, the posterior distribution of $\theta$ concentrates sharply around this incorrect value.

### Constant and vague proper priors

Laplace (1812) advocated use of a constant prior as the overall objective prior and this approach, eventually named *inverse probability*, dominated statistical practice for over 100 years. But the problems of a constant prior are well-documented, including the following:

(i) Lack of invariance to transformation, the main criticism directed at Laplace's approach.

(ii) Frequent posterior impropriety.

(iii) Possible terrible performance, as in the earlier multi-normal mean example.

Vague proper priors (such as a constant prior over a large compact set) are perceived by many as being adequate as an overall objective prior, but they too have well-understood problems. Indeed, they are, at best, equivalent to use of a constant prior, and so inherit most of the flaws of a constant prior. In the multi-normal mean example, for instance, use of $\mathrm{N}(\mu_i \,|\, 0, 1000)$ vague proper priors results in a posterior mean for $\theta$ that is virtually identical to the inconsistent posterior mean from the constant prior.

There is a common misperception that vague proper priors are safer than a constant prior, since a proper posterior is guaranteed with a vague proper prior but not for a constant prior. But this actually makes vague proper priors more dangerous than a

constant prior. When the constant prior results in an improper posterior distribution, the vague proper prior will yield an essentially arbitrary posterior, depending on the degree of vagueness that is chosen for the prior. And to detect that the answer is arbitrary, one has to conduct a sensitivity study concerning the degree of vagueness, something that can be difficult in complex problems when several or high-dimensional vague proper priors are used. With the constant prior on the other hand, the impropriety of the posterior will usually show up in the computation—the Markov Chain Monte Carlo (MCMC) will not converge—and hence can be recognized.

### Jeffreys-rule prior

The Jeffreys-rule prior (Jeffreys, 1946, 1961) is the same for all parameters in a model, and is, hence, an obvious candidate for an overall prior. If the data model density is $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ the Jeffeys-rule prior for the unknown $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ has the form

$$\pi(\theta_1, \ldots, \theta_m) = |I(\boldsymbol{\theta})|^{1/2},$$

where $I(\boldsymbol{\theta})$ is the $m \times m$ Fisher information matrix with $(i, j)$ element

$$I(\boldsymbol{\theta})_{ij} = \mathrm{E}_{\boldsymbol{x} \,|\, \boldsymbol{\theta}} \left[ -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \right].$$

This is the optimal objective prior (from many perspectives) for *regular one-parameter* models, but has problems for multi-parameter models. For instance, the right-Haar prior in the earlier multi-normal mean problem is also the Jeffreys-rule prior there, and was seen to result in an inconsistent estimator of $\theta$. Even for the basic $\mathrm{N}(x_i \,|\, \mu, \sigma)$ model, the Jeffreys-rule prior is $\pi(\mu, \sigma) = 1/\sigma^2$, which results in posterior inferences for $\mu$ and $\sigma$ that have the wrong 'degrees of freedom.'

For the bivariate normal example, the Jeffreys-rule prior is $1/[\sigma_1^2 \sigma_2^2 (1 - \rho^2)^2]$; this yields the natural marginal posteriors for the means and standard deviations, but results in quite inferior objective posteriors for $\rho$ and various derived parameters, as shown in Berger and Sun (2008). More, generally, the Jeffreys-rule prior for a covariance matrix is studied in Yang and Berger (1994), and shown to yield a decidedly inferior posterior.

There have been efforts to improve upon the Jeffreys-rule prior, such as consideration of the "independence Jeffreys-rule prior," but a general alternative definition has not resulted.

Finally, consider the following well-known example, which suggests problems with the Jeffreys-rule prior even when it is proper.

**Example 1.4. Multinomial Distribution (continued):** Consider the multinomial example where the sample size $n$ is small relative to the number of classes $m$; thus we have a large sparse table. The Jeffreys-rule prior is the *proper* prior, $\pi(\theta_1, \ldots, \theta_m) \propto \prod_{i=1}^m \theta_i^{-1/2}$ , but is not a good candidate for the overall prior. For instance, suppose $n = 3$ and $m = 1000$, with $x_{240} = 2$, $x_{876} = 1$, and all the other $x_i = 0$. The posterior means resulting from use of the Jeffreys-rule prior are

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/2}{\sum_{i=1}^m (x_i + 1/2)} = \frac{x_i + 1/2}{n + m/2} = \frac{x_i + 1/2}{503},$$

so $\mathrm{E}[\theta_{240} \mid \boldsymbol{x}] = \frac{2.5}{503}$, $\mathrm{E}[\theta_{876} \mid \boldsymbol{x}] = \frac{1.5}{503}$, $\mathrm{E}[\theta_i \mid \boldsymbol{x}] = \frac{0.5}{503}$ otherwise. So, cells 240 and 876 only have total posterior probability of $\frac{4}{503} = 0.008$ even though all 3 observations are in these cells. The problem is that the Jeffreys-rule prior effectively added 1/2 to the 998 zero cells, making them more important than the cells with data! That the Jeffreys-rule prior can encode much more information than is contained in the data is hardly desirable for an objective analysis.

An alternative overall prior that is sometimes considered is the uniform prior on the simplex, but this is even worse than the Jeffreys prior, adding 1 to each cell. The prior that adds 0 to each cell is $\prod_i \theta_i^{-1}$, but this results in an improper posterior if any cell has a zero entry, a virtual certainty for very large tables.

We actually know of no multivariable example in which we would recommend the Jeffreys-rule prior. In higher dimensions, the prior always seems to be either 'too diffuse' as in the multinormal means example, or 'too concentrated' as in the multinomial example.

### Prior averaging approach

Starting with a collection of reference (or other) priors $\{\pi_i(\boldsymbol{\theta}), i = 1, \ldots, m\}$ for differing parameters or quantities of interest, a rather natural approach is to use an average of the priors. Two natural averages to consider are the arithmetic mean

$$\pi^A(\boldsymbol{\theta}) = \frac{1}{m} \sum\nolimits_{i=1}^{m} \pi_i(\boldsymbol{\theta}) \,,$$

and the geometric mean

$$\pi^G(\boldsymbol{\theta}) = \prod\nolimits_{i=1}^{m} \pi_i(\boldsymbol{\theta})^{1/m} \,.$$

While the arithmetic average might seem most natural, arising from the hierarchical reasoning of assigning each $\pi_i$ probability $1/m$ of being correct, geometric averaging arises naturally in the definition of reference priors (Berger, Bernardo and Sun, 2009), and also is the optimal prior if one is trying to choose a single prior to minimize the average of the Kullback-Leibler (KL) divergences of the prior from the $\pi_i$'s (a fact of which we were reminded by Gauri Datta). Furthermore, the weights in arithmetic averaging of improper priors are rather arbitrary because the priors have no normalizing constants, whereas geometric averaging is unaffected by normalizing constants.

**Example 1.5. Bivariate Normal Distribution (continued):** Faced with the two right-Haar priors in this problem,

$$\pi_1(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \sigma_1^{-2}(1-\rho^2)^{-1}, \qquad \pi_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \sigma_2^{-2}(1-\rho^2)^{-1},$$

the two average priors are

$$\pi^A(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\sigma_1^2(1-\rho^2)} + \frac{1}{2\sigma_2^2(1-\rho^2)} \,, \tag{1}$$

$$\pi^G(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1\sigma_2(1-\rho^2)} \,. \tag{2}$$

Interestingly, Sun and Berger (2007) show that $\pi^A$ is a worse objective prior than either right-Haar prior alone, while $\pi^G$ is the overall recommended objective prior.

One problem with the averaging approach is that each of the reference priors can depend on all of the other parameters, and not just the parameter of interest, $\theta_i$, for which it was created.

**Example 1.6. Multinomial Example (continued):** The reference prior derived when the parameter of interest is $\theta_i$ actually depends on the sequential ordering chosen for all the parameters (e.g. $\{\theta_i, \theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m\}$) in the reference prior derivation; there are thus $(m-1)!$ different reference priors for each parameter of interest. Each of these reference priors will result in the same marginal reference posterior for $\theta_i$,

$$\pi_{\theta_i}(\theta_i \,|\, \boldsymbol{x}) = \mathrm{Be}(\theta_i \,|\, x_i + \tfrac{1}{2}, n - x_i + \tfrac{1}{2}),$$

but the full reference prior and the full posterior, $\pi_{\theta_i}(\theta \,|\, \boldsymbol{x})$, do depend on the ordering of the other parameters. There are thus a total of $m!$ such full reference priors to be averaged, leading to an often-prohibitive computation.

In general, the quality of reference priors as overall priors is unclear, so there is no obvious sense in which an average of them will make a good overall reference prior. The prior averaging approach is thus best viewed as a method of generating interesting possible priors for further study, and so will not be considered further herein.

## 1.3 Three approaches to construction of the overall prior

### Common reference prior

If the reference prior that is computed for any parameter of the model (when declared to be the parameter of interest) is the same, then this common reference prior is the natural choice for the overall prior. This is illustrated extensively in Section 2; indeed, the section attempts to catalogue the situations in which this is known to be the case, so that these are the situations with a ready-made overall prior.

### Reference distance approach

In this approach, one seeks a prior that will yield marginal posteriors, for each $\theta_i$ of interest, that are close to the set of reference posteriors $\{\pi(\theta_i \,|\, \boldsymbol{x})\}_{i=1}^m$ (yielded by the set of reference priors $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^m$), in an average sense over both posteriors and data $\boldsymbol{x} \in \mathcal{X}$.

**Example 1.7. Multinomial Example (continued):** In Example 1.4 consider, as an overall prior, the Dirichlet $\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ distribution, having density proportional to $\prod_i \theta_i^{a-1}$, leading to $\mathrm{Be}(\theta_i \,|\, x_i + a, n - x_i + (m-1)a)$ as the marginal posterior for $\theta_i$. In Section 3.2, we will study which choice of $a$ yields marginal posteriors that are as close as possible to the reference marginal posteriors $\mathrm{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$, arising when $\theta_i$ is the parameter of interest. Roughly, the recommended choice is

$a = 1/m$, resulting in the overall prior $\pi^o(\theta_1, \ldots, \theta_m) \propto \prod_{i=1}^m \theta_i^{(1-m)/m}$. Note that this distribution adds only $1/m = 0.001$ to each cell in the earlier example, so that

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/m}{\sum_{i=1}^m (x_i + 1/m)} = \frac{x_i + 1/m}{n+1} = \frac{x_i + 0.001}{4}\,.$$

Thus $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] \approx 0.5$, $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] \approx 0.25$, and $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] \approx \frac{1}{4000}$ otherwise, all sensible results.

### Hierarchical approach

Utilize hierarchical modeling to transfer the reference prior problem to a 'higher level' of the model (following the advice of I. J. Good). In this approach one

(i) Chooses a class of *proper* priors $\pi(\boldsymbol{\theta} \,|\, a)$ reflecting the desired structure of the problem.

(ii) Forms the marginal likelihood $p(\boldsymbol{x} \,|\, a) = \int p(\boldsymbol{x} \,|\, a)\pi(\boldsymbol{\theta} \,|\, a)\; d\boldsymbol{\theta}$.

(iii) Finds the reference prior, $\pi^R(a)$, for $a$ in this marginal model.

Thus the overall prior becomes

$$\pi^o(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} \,|\, a)\, \pi^R(a)\, da\,,$$

although computation is typically easier by utilizing both $\boldsymbol{\theta}$ and $a$ in the computation rather than formally integrating out $a$.

**Example 1.8. Multinomial (continued)** The Dirichlet $\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ class of priors is natural here, reflecting the desire to treat all the $\theta_i$ similarly. We thus need only to find the reference prior for $a$ in the marginal model,

$$\begin{aligned} p(\boldsymbol{x} \,|\, a) &= \int \binom{n}{x_1 \ldots x_m} \left( \prod_{i=1}^m \theta_i^{x_i} \right) \frac{\Gamma(m\,a)}{\Gamma(a)^m} \prod_{i=1}^m \theta_i^{a-1} d\boldsymbol{\theta} \\ &= \binom{n}{x_1 \ldots x_m} \frac{\Gamma(m\,a)}{\Gamma(a)^m} \frac{\prod_{i=1}^m \Gamma(x_i + a)}{\Gamma(n + m\,a)}\,. \end{aligned} \tag{3}$$

The reference prior for $\pi^R(a)$ would just be the Jeffreys-rule prior for this marginal model; this is computed in Section 4. The implied prior for $\boldsymbol{\theta}$ is, of course

$$\pi(\boldsymbol{\theta}) = \int \mathrm{Di}(\boldsymbol{\theta} \,|\, a)\, \pi^R(a)\, da\,.$$

Interestingly, $\pi^R(a)$ turns out to be a proper prior, necessary because the marginal likelihood is bounded away from zero as $a \to \infty$.

As computations in this hierarchical setting are more complex, one might alternatively simply choose the Type-II maximum likelihood estimate—*i.e.*, the value

of $a$ that maximizes (3)—at least when $m$ is large enough so that the empirical Bayes procedure can be expected to be close to the full Bayes procedure. For the data given in the earlier example (one cell having two counts, another one count, and the rest zero counts), this marginal likelihood is proportional to $[a(a+1)]/[(m\,a+1)(m\,a+2)]$, which is maximized at roughly $a = \sqrt{2}/m$. In Section 4 we will see that it is actually considerably better to maximize the reference posterior for $a$, namely $\pi^R(a \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid a)\,\pi^R(a)$, as it can be seen that the marginal likelihood does not go to zero as $a \to \infty$ and the mode may not even exist.

### 1.4 Outline of the paper

Section 2 presents known situations in which the reference priors for any parameter (of interest) in the model are identical. This section is thus the beginnings of a catalogue of good overall objective priors. Section 3 formalizes the reference distance approach and applies it to two models—the multinomial model and the normal model where the coefficient of variation is also a parameter of interest. In Section 4 we consider the hierarchical prior modeling approach, applying it to three models—the multinomial model, a hypergeometric model, and the multinormal model—and misapplying it to the bivariate normal model. Section 5 presents conclusions.

## 2 Common reference prior for all parameters

In this section we discuss situations where the reference prior is unique, in the sense that it is the same no matter which of the specified model parameters is taken to be of interest and which of the possible possible parameter orderings is used in the derivation. (In general, a reference prior will depend on the parameter ordering used in its derivation.) This unique reference prior is typically an excellent choice for the overall prior.

### 2.1 Structured diagonal Fisher information matrix

Consider a parametric family $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ with unknown parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$. For any parameter $\theta_i$, let $\boldsymbol{\theta}_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_k)$ denote the parameters other than $\theta_i$. The following theorem encompasses a number of important situations in which there is a common reference prior for all parameters.

**Theorem 2.1.** *Suppose that the Fisher information matrix of $\boldsymbol{\theta}$ is of the form,*

$$\boldsymbol{I}(\boldsymbol{\theta}) = diag(f_1(\theta_1)g_1(\boldsymbol{\theta}_{-1}), f_2(\theta_2)g_2(\boldsymbol{\theta}_{-2}), \cdots, f_k(\theta_m)g_k(\boldsymbol{\theta}_{-k})), \tag{4}$$

*where $f_i$ is a positive function of $\theta_i$ and $g_i$ is a positive function of $\boldsymbol{\theta}_{-i}$, for $i = 1, \cdots, k$. Then the one-at-a-time reference prior, for any chosen parameter of interest and any ordering of the nuisance parameters in the derivation, is given by*

$$\pi^R(\boldsymbol{\theta}) \propto \sqrt{f_1(\theta_1)f_2(\theta_2)\cdots f_k(\theta_k)}. \tag{5}$$

*Proof.* The result follows from Datta and Ghosh (1996). □

This prior is also what was called the *independent reference prior* in Sun and Berger (1998), and is the most natural definition of an *independence Jeffreys prior* under condition (4). Note that being the common reference prior for all of the original parameters of interest in the model does not guarantee that $\pi^R$ will be the reference prior for every potential parameter of interest (see Section 3.1) but, for the scenarios in which an overall prior is desired, this unique reference prior for all natural parameters is arguably optimal.

A simple case in which (4) is satisfied is when the density is of the form

$$p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{k} p_i(\boldsymbol{x}_i \,|\, \theta_i)\,, \tag{6}$$

with $\boldsymbol{x}$ decomposable as $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. In this case the (common to all parameters) reference prior is simply the product of the reference priors for each of the separate models $p_i(\boldsymbol{x}_i \,|\, \theta_i)$; this is also the Jeffreys-rule prior.

#### Bivariate binomial distribution

Crowder and Sweeting (1989) consider the following bivariate binomial distribution, whose probability density is given by

$$p(r, s \,|\, \theta_1, \theta_2) = \binom{m}{r} \theta_1^r (1-\theta_1)^{m-r} \binom{r}{s} \theta_2^s (1-\theta_2)^{r-s},$$

where $0 < \theta_1, \theta_2 < 1$, and $s$ and $r$ are nonnegative integers satisfying $0 \le s \le r \le n$. The Fisher information matrix for $(\theta_1, \theta_2)$ is given by

$$\boldsymbol{I}(\theta_1, \theta_2) = n \,\mathrm{diag}[\{\theta_1(1-\theta_1)\}^{-1},\ \theta_1\{\theta_2(1-\theta_2)\}^{-1}]\,,$$

which is of the form (4). (Note that this density is not of the form (6).) Hence the reference prior, when either $\theta_1$ or $\theta_2$ are the parameter of interest, is

$$\pi^R(\theta_1, \theta_2) \propto \{\theta_1(1-\theta_1)\theta_2(1-\theta_2)\}^{-\frac{1}{2}}\,,$$

*i.e.*, independent Beta, $\mathrm{Be}(\theta_i \,|\, 1/2, 1/2)$ distributions for $\theta_1$ and $\theta_2$; this reference prior was first formally derived for this model by Polson and Wasserman (1990). This is thus the overall recommended prior for this model.

#### Multinomial distribution for directional data

While we have already seen that determining an overall reference prior for the multinomial distribution is challenging, there is a special case of the distribution where doing so is possible. This happens when the cells are ordered or directional. For example, the cells could be grades for a class such as A, B, C, D, and F; outcomes from an attitude survey such as strongly agree, agree, neutral, disagree, and strongly disagree; or discrete survival times. Following Example 1.1 (multinomial example), with this cell

ordering, there is a natural reparameterization of the multinomial probabilities into the conditional probabilities

$$\xi_j = \frac{\theta_j}{\theta_j + \cdots + \theta_m}, \text{ for } j = 1, \cdots, m-1\,. \tag{7}$$

Here $\xi_j$ is the conditional probability of an observation being in cell $j$ given that the observation is in cells $j$ to $m$. The Fisher information matrix of $(\xi_1, \cdots, \xi_{m-1})$ is

$$\boldsymbol{I}^*(\xi_1, \xi_2 \cdots, \xi_{m-1}) = n \text{ diag}(\eta_1, \cdots, \eta_{m-1}), \tag{8}$$

where

$$\delta_j \quad = \quad \frac{1}{\xi_j(1-\xi_j)} \prod_{i=1}^{j-1}(1-\xi_i),$$

for $j = 1, \cdots, m-1$. Clearly (8) is of the form (4), from which it immediately follows that the one-at-a-time reference prior for any of the parameters $(\xi_1, \xi_2 \cdots, \xi_{m-1})$ (and any ordering of them in the derivation) is the product of independent Beta $(1/2, 1/2)$ distributions for the $\xi_j$ for $j = 1, \ldots m-1$. This is the same as Berger and Bernardo's (1992b) reference prior for this specific ordering of cells.

### A two-parameter exponential family

Bar-Lev and Reiser (1982) considered the following two-parameter exponential family density:

$$p(x \,|\, \theta_1, \theta_2) \quad = \quad a(x) \exp\{\theta_1 U_1(x) - \theta_1 G_2'(\theta_2) U_2(x) - \psi(\theta_1, \theta_2)\}, \tag{9}$$

where the $U_i(\cdot)$ are to be specified, $\theta_1 < 0$, $\theta_2 = \mathrm{E}\{U_2(X) \,|\, (\theta_1, \theta_2)\}$, the $G_i(\cdot)$'s, are infinitely differentiable functions with $G_i'' > 0$, and $\psi(\theta_1, \theta_2) = -\theta_1\{\theta_2 G_2'(\theta_2) - G_2(\theta_2)\} + G_1(\theta_2)$. This is a large class of distributions, which includes, for suitable choices of $G_1$, $G_2$, $U_1$ and $U_2$, many popular statistical models such as the normal, inverse normal, gamma, and inverse gamma. Table 1, reproduced from Sun (1994), indicates how each distribution arises.

Table 1. Special cases of Bar-Lev and Reiser's (1982) two parameter exponential family, where $h(\theta_1) = -\theta_1 + \theta_1 \log(-\theta_1) + \log(\Gamma(-\theta_1))$.

| | $G_1(\theta_1)$ | $G_2(\theta_2)$ | $U_1(x)$ | $U_2(x)$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|---|---|---|
| Normal $(\mu, \sigma)$ | $-\frac{1}{2}\log(-2\theta_1)$ | $\theta_2^2$ | $x^2$ | $x$ | $-1/(2\sigma^2)$ | $\mu$ |
| Inverse Gaussian | $-\frac{1}{2}\log(-2\theta_1)$ | $1/\theta_2$ | $1/x$ | $x$ | $-\alpha/2$ | $\sqrt{\alpha/\mu}$ |
| Gamma | $h(\theta_1)$ | $-\log\theta_2$ | $-\log x$ | $x$ | $-\alpha$ | $\mu$ |
| Inverse Gamma | $h(\theta_1)$ | $-\log\theta_2$ | $\log x$ | $1/x$ | $-\alpha$ | $\mu$ |

The Fisher information matrix of $(\theta_1, \theta_2)$ based on (10) is

$$\boldsymbol{I}(\theta_1, \theta_2) = \begin{pmatrix} G_1''(\theta_1) & 0 \\ 0 & -\theta_1 G_2''(\theta_2) \end{pmatrix},$$

which is of the form (4). Thus, when either $\theta_1$ or $\theta_2$ is the parameter of interest, the one-at-a-time reference prior (first shown in Sun and Ye (1996)) is

$$\pi^R(\theta_1, \theta_2) = \sqrt{G_1''(\theta_1) G_2''(\theta_2)}. \tag{10}$$

For the important special case of the Inverse Gaussian density,

$$p(x \mid \alpha, \psi) = (\alpha/2\pi x^3)^{1/2} \exp\Big\{-\frac{1}{2}\alpha x(1/x - \psi)^2\Big\},\ x > 0 \tag{11}$$

where $\alpha > 0, \psi > 0$, the common reference prior (and overall recommended prior) is

$$\pi^R(\alpha, \psi) \propto \frac{1}{\alpha\sqrt{\psi}}. \tag{12}$$

The resulting marginal posteriors of $\alpha$ and $\psi$ can be found in Sun and Ye (1996).

For the important special case of the Gamma $(\alpha, \mu)$ density,

$$p(x \mid \alpha, \mu) = \alpha^\alpha x^{\alpha-1} \exp(-\alpha x/\mu)/\{\Gamma(\alpha)\mu^\alpha\}, \tag{13}$$

the common reference prior (and overall recommended prior) is

$$\pi^R(\alpha, \mu) \propto \frac{\sqrt{\alpha\xi(\alpha) - 1}}{\sqrt{\alpha}\mu}, \tag{14}$$

where $\xi(\alpha) = (\partial^2/\partial\alpha^2)\log\{\Gamma(\alpha)\}$ is the polygamma function. The resulting marginal posteriors of $\alpha$ and $\mu$ can be found in Sun and Ye (1996).

### A stress-strength model

Consider the following stress-strength system, where $Y$, the strength of the system, is subject to stress $X$. The system fails at any moment the applied stress (or load) is greater than the strength (or resistance). The reliability of the system is then given by

$$\theta = P(X \le Y). \tag{15}$$

An important instance of this situation was described in Enis and Geisser (1971), where $X_1, \cdots, X_m$, and $Y_1, \cdots, Y_n$ are independent random samples from exponential distributions with unknown means $\eta_1$ and $\eta_2$, in which case

$$\theta = \eta_1/(\eta_1 + \eta_2). \tag{16}$$

As the data density is of the form (6), the (common to all parameters) reference prior is easily seen to be $\pi^R(\eta_1, \eta_2) = 1/(\eta_1\eta_2)$, which is also the Jeffreys prior as noted in Enis and Geisser (1971). Our interest, however, is primarily in $\theta$. Defining the nuisance parameter to be $\psi = \eta_1^{(m+n)/n}\eta_2^{(m+n)/m}$, the resulting Fisher information matrix is

$$\boldsymbol{I}(\theta, \psi) = \text{diag}\left(\frac{mn}{(m+n)\theta^2(1-\theta)^2}, \frac{m^2n^2}{(m+n)^3\psi^2}\right),$$

again of the form (4). So the Jeffreys prior and the one-at-a-time reference prior of any ordering for $\theta$ and $\psi$ is $\pi^R(\theta,\psi) = 1/\{\theta(1-\theta)\psi\}$, which can be seen to be the transformed version of $\pi^R(\eta_1,\eta_2) = 1/(\eta_1\eta_2)$. So the Jeffreys prior is also the one-at-a-time reference prior for $\theta$. Ghosh and Sun (1998) showed that this prior is the second order matching prior for $\theta$ when $m/n \to a > 0$.

## 2.2 Other scenarios with a common reference prior

A common reference prior can exist in scenarios not covered by Theorem 2.1. Two such situations are considered here, the first which leads to a fine overall prior and the second which does not.

### The location-scale family

Consider the location-scale family having density

$$p(x \,|\, \mu,\sigma) = \frac{1}{\sigma} g\Big(\frac{x-\mu}{\sigma}\Big),$$

where $g$ is a specified density function and $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown. The Fisher information of $(\mu,\sigma)$ is

$$\boldsymbol{I}(\mu,\sigma) = \frac{1}{\sigma^2} \begin{pmatrix} \int \frac{[g'(y)]^2}{g(y)} dy & \int \{y\frac{[g'(y)]^2}{g(y)} + g'(y)\} dy \\ \int \{y\frac{[g'(y)]^2}{g(y)} + g'(y)\} dy & \int \frac{[yg'(y)+g(y)]^2}{g(y)} dy \end{pmatrix}. \tag{17}$$

Although this is not of the form (4), it is easy to see that the one-at-a-time reference prior for either $\mu$ or $\sigma$ is $\pi^R(\mu,\sigma) = 1/\sigma$. This prior is also the right-Haar prior for the location-scale group, and known to result in Bayes procedures with optimal frequentist properties. Hence it is clearly the recommended overall prior.

### Unnatural parameterizations

A rather unnatural parameterization for the bivariate normal model arises by defining $\psi_1 = 1/\sigma_1, \psi_2 = 1/\sqrt{\sigma_2^2(1-\rho^2)}$, and $\psi_3 = -\rho\sigma_2/\sigma_1$. From Berger and Sun (2008), the Fisher information matrix for the parameterization $(\psi_1,\psi_2,\psi_3,\mu_1,\mu_2)$ is

$$\boldsymbol{I} = \mathrm{diag}\Big(\frac{2}{\psi_1^2}, \frac{2}{\psi_2^2}, \frac{\psi_2^2}{\psi_1^2}, \boldsymbol{\Sigma}^{-1}\Big), \tag{18}$$

where $\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \psi_1^2 + \psi_2^2\psi_3^2 & \psi_2^2\psi_3 \\ \psi_2^2\psi_3 & \psi_2^2 \end{pmatrix}$. While this is not of the form (4), direct computation shows that the one-at-a-time reference prior for any of these five parameters and under any ordering is

$$\pi^R(\psi_1,\psi_2,\psi_3,\mu_1,\mu_2) = \frac{1}{\psi_1\psi_2}. \tag{19}$$

Unfortunately, this is equivalent to the right Haar prior, $\pi^H(\sigma_1,\sigma_2,\rho,\mu_1,\mu_2) = \frac{1}{\sigma_1^2(1-\rho^2)}$, which we have argued is not a good overall prior. This suggests that the parameters

used in this 'common reference prior' approach need to be natural, in some sense, to result in a good overall prior.

# 3 Reference distance approach

Recall that the goal is to identify a single overall prior $\pi(\boldsymbol{\omega})$ that can be systematically used for all the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) = \{\theta_1(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ of interest. The idea of the reference distance approach is to find a $\pi(\boldsymbol{\omega})$ whose corresponding marginal posteriors, $\{\pi(\theta_i \,|\, \boldsymbol{x})\}_{i=1}^m$ are close, in an average sense, to the reference posteriors $\{\pi_i(\theta_i \,|\, \boldsymbol{x})\}_{i=1}^m$ arising from the separate reference priors $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^m$ derived under the assumption that each of the $\theta_i$'s is of interest. (In situations where reference priors are not unique for a parameter of interest, we assume other considerations have been employed to select a preferred reference prior.) In the remainder of the paper, $\boldsymbol{\theta}$ will equal $\boldsymbol{\omega}$, so we will drop $\boldsymbol{\omega}$ from the notation.

We first consider the situation where the problem has an exact solution.

## 3.1 Exact solution

If one is able to find a single joint prior $\pi(\boldsymbol{\theta})$ whose corresponding marginal posteriors are precisely equal to the reference posteriors for each of the $\theta_i$'s, so that, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\pi(\theta_i \,|\, \boldsymbol{x}) = \pi_i(\theta_i \,|\, \boldsymbol{x}), \quad i = 1, \ldots, m\,, \tag{20}$$

then it is natural to argue that this should be an appropriate solution to the problem. The most important situation in which this will happen is when there is a common reference prior for each of the parameters, as discussed in Section 2. It is conceivable that there could be more than one overall prior that would satisfy (20); if this were to happen it is not clear how to proceed.

**Example 3.1.** *Univariate normal data.* Consider data $\boldsymbol{x}$ which consist of a random sample of normal observations, so that $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \mu, \sigma) = \prod_{i=1}^n \mathrm{N}(x_i \,|\, \mu, \sigma)$, and suppose that one is equally interested in $\mu$ (or any one-to-one transformation of $\mu$) and $\sigma$ (or any one-to-one transformation of $\sigma$, such as the variance $\sigma^2$, or the precision $\sigma^{-2}$.) The common reference prior when any of these is the quantity of interest is known to be the right Haar prior $\pi_\mu(\mu, \sigma) = \pi_\sigma(\mu, \sigma) = \sigma^{-1}$, and this is thus an exact solution to the overall prior problem under the reference distance approach (as is also clear from Section 2.2, since this is a location-scale family).

Interestingly, this prior also works well for making *joint inferences on* $(\mu, \sigma)$ in that it can be verified that the corresponding joint credible regions for $(\mu, \sigma)$ have appropriate coverage properties. This does not mean, of course, that the overall prior is necessarily good for any function of the two parameters. For instance, if the quantity of interest is the centrality parameter $\theta = \mu/\sigma$, the reference prior is easily found to be $\pi_\theta(\theta, \sigma) = (1 + \frac{1}{2}\theta^2)^{-1/2}\sigma^{-1}$ (Bernardo, 1979), which is not the earlier overall reference prior. Finding a good overall prior by the reference distance situation when this is added to the list of parameters of interest is considered in Section 3.2.

## 3.2 Reference distance solution

When an exact solution is not possible, it is natural to consider a family of candidate prior distributions, $\mathcal{F} = \{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$, and choose, as the overall prior, the distribution from this class which yields marginal posteriors that are closest, in an average sense, to the marginal reference posteriors.

### Directed logarithmic divergence

It is first necessary to decide how to measure the distance between two distributions. We will actually use a divergence, not a distance, namely the directed logarithmic or Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) given in the following definition.

**Definition 1.** *Let $p(\boldsymbol{\psi})$ be the probability density of a random vector $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, and consider an approximation $p_0(\boldsymbol{\psi})$ with the same or larger support. The* directed logarithmic divergence *of $p_0$ from $p$ is*

$$\kappa\{p_0 \,|\, p\} = \int_{\boldsymbol{\Psi}} p(\boldsymbol{\psi}) \log \frac{p(\boldsymbol{\psi})}{p_0(\boldsymbol{\psi})} \, d\boldsymbol{\psi} \,,$$

*provided that the integral exists.*

The non-negative directed logarithmic divergence $\kappa\{p_0 \,|\, p\}$ is the expected log-density ratio of the true density over its approximation; it is invariant under one-to-one transformations of the random vector $\boldsymbol{\psi}$; and it has an operative interpretation as the amount of information (in natural information units or *nits*) which may be expected to be required to recover $p$ from $p_0$. It was first proposed by Stein (1964) as a loss function and, in a decision-theoretic context, it is often referred to as the *entropy loss*.

### Weighted logarithmic loss

Suppose the relative importance of the $\theta_i$ is given by a set of weights $\{w_1, \ldots, w_m\}$, with $0 < w_i < 1$ and $\sum_i w_i = 1$. A natural default value for these is obviously $w_i = 1/m$, but there are many situations where this choice may not be appropriate; in Example 1.3 for instance, one might give $\theta$ considerably more weight than the means $\mu_i$. To define the proposed criterion, we will also need to utilize the reference prior predictives for $i = 1, \ldots, m$,

$$p_{\theta_i}(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \, \pi_{\theta_i}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \,.$$

**Definition 2.** *The best overall prior $\pi^o(\boldsymbol{\theta})$ within the family $\mathcal{F} = \{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$ is defined as that—assuming it exists and is unique—which minimizes the* weighted average expected logarithmic loss*, so that*

$$\begin{aligned} \pi^o(\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta} \,|\, \boldsymbol{a}^*), \quad \boldsymbol{a}^* = \arg \inf_{\boldsymbol{a} \in \mathcal{A}} d(\boldsymbol{a}), \\ d(\boldsymbol{a}) &= \sum_{i=1}^{m} w_i \int_{\mathcal{X}} \kappa\{\pi_{\boldsymbol{\theta}_i}(\cdot \,|\, \boldsymbol{x}, \boldsymbol{a}) \,|\, \pi_{\boldsymbol{\theta}_i}(\cdot \,|\, \boldsymbol{x})\} \, p_{\theta_i}(\boldsymbol{x}) \, d\boldsymbol{x}, \quad \boldsymbol{a} \in \mathcal{A} \,. \end{aligned}$$

*This can be rewritten, in terms of the sum of expected risks, as*

$$d(\boldsymbol{a}) = \sum_{i=1}^{m} w_i \int_{\boldsymbol{\Theta}} \rho_i(\boldsymbol{a} \,|\, \boldsymbol{\theta})\, \pi_{\theta_i}(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad \boldsymbol{a} \in \mathcal{A}\,,$$

*where*

$$\rho_i(\boldsymbol{a} \,|\, \boldsymbol{\theta}) = \int_{\mathcal{X}} \kappa\{\pi_{\boldsymbol{\theta}_i}(\cdot \,|\, \boldsymbol{x}, \boldsymbol{a}) \,|\, \pi_{\boldsymbol{\theta}_i}(\cdot \,|\, \boldsymbol{x})\}\, p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, d\boldsymbol{x}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Note that there is no assurance that $d(\boldsymbol{a})$ will be finite if the reference priors are improper. Indeed, in cases we have investigated with improper reference priors, $d(\boldsymbol{a})$ has failed to be finite and hence the reference distance approach cannot be directly used. However, as in the construction of reference priors, one can consider an approximating sequence of proper priors $\{\pi_{\theta_i}(\boldsymbol{\theta} \,|\, k), k = 1, 2 \ldots\}$ on increasing compact sets. For each of the $\pi_{\theta_i}(\boldsymbol{\theta} \,|\, k)$, one can minimize the expected risk

$$d(\boldsymbol{a} \,|\, k) = \sum_{i=1}^{m} w_i \int_{\boldsymbol{\Theta}} \rho_i(\boldsymbol{a} \,|\, \boldsymbol{\theta})\, \pi_{\theta_i}(\boldsymbol{\theta} \,|\, k)\, d\boldsymbol{\theta},$$

obtaining $\boldsymbol{a}_k^* = \arg\inf_{\boldsymbol{a} \in \mathcal{A}} d(\boldsymbol{a} \,|\, k)$. Then, if $\boldsymbol{a}^* = \lim_{k \to \infty} \boldsymbol{a}_k^*$ exists, one can declare this to be the solution.

### Multinomial model

In the multinomial model with $m$ cells and parameters $\{\theta_1, \ldots, \theta_m\}$, with $\sum_{i=1}^{m} \theta_i = 1$, the reference posterior for each of the $\theta_i$'s is $\pi_i(\theta_i \,|\, \boldsymbol{x}) = \mathrm{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$, while the marginal posterior distribution of $\theta_i$ resulting from the joint prior $\mathrm{Di}(\theta_1, \ldots, \theta_{m-1} \,|\, a)$ is $\mathrm{Be}(\theta_i \,|\, x_i + a, n - x_i + (m-1)a)$. The directed logarithmic discrepancy of the posterior $\mathrm{Be}(\theta_i \,|\, x_i + a, n - x_i + (m-1)a)$ from the reference posterior $\mathrm{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$ is

$$\kappa_i\{a \,|\, \boldsymbol{x}, m, n\} = \kappa_i\{a \,|\, x_i, m, n\} = \kappa_{\mathrm{Be}}\{x_i + a, n - x_i + (m-1)a \,|\, x_i + \tfrac{1}{2}, n - x_i + \tfrac{1}{2}\}$$

where

$$\begin{aligned}
\kappa_{\mathrm{Be}}\{\alpha_0, \beta_0 \,|\, \alpha, \beta\} &= \int_0^1 \mathrm{Be}(\theta_i \,|\, \alpha, \beta) \log\left[\frac{\mathrm{Be}(\theta_i \,|\, \alpha, \beta)}{\mathrm{Be}(\theta_i \,|\, \alpha_0, \beta_0)}\right] d\theta_i \\
&= \log\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha_0+\beta_0)}\, \frac{\Gamma(\alpha_0)}{\Gamma(\alpha)}\, \frac{\Gamma(\beta_0)}{\Gamma(\beta)}\right] \\
&\quad + (\alpha - \alpha_0)\psi(\alpha) + (\beta - \beta_0)\psi(\beta) - ((\alpha+\beta) - (\alpha_0+\beta_0))\psi(\alpha+\beta),
\end{aligned}$$

and $\psi(\cdot)$ is the digamma function.

The divergence $\kappa_i\{a \,|\, x_i, m, n\}$ between the two posteriors of $\theta_i$ depends on the data only through $x_i$ and the sampling distribution of $x_i$ is Binomial $\mathrm{Bi}(x_i \,|\, n, \theta_i)$, which only depends of $\theta_i$. Moreover, the marginal reference prior for $\theta_i$ is $\pi_{\theta_i}(\theta_i) = \mathrm{Be}(\theta_i \,|\, 1/2, 1/2)$ and, therefore, the corresponding reference predictive for $x_i$ is

$$p(x_i \,|\, n) = \int_0^1 \mathrm{Bi}(x_i \,|\, n, \theta_i)\, \mathrm{Be}(\theta_i \,|\, 1/2, 1/2)\, d\theta_i = \frac{1}{\pi}\, \frac{\Gamma(x_i + \frac{1}{2})\, \Gamma(n - x_i + \frac{1}{2})}{\Gamma(x_i + 1)\, \Gamma(n - x_i + 1)}\,.$$

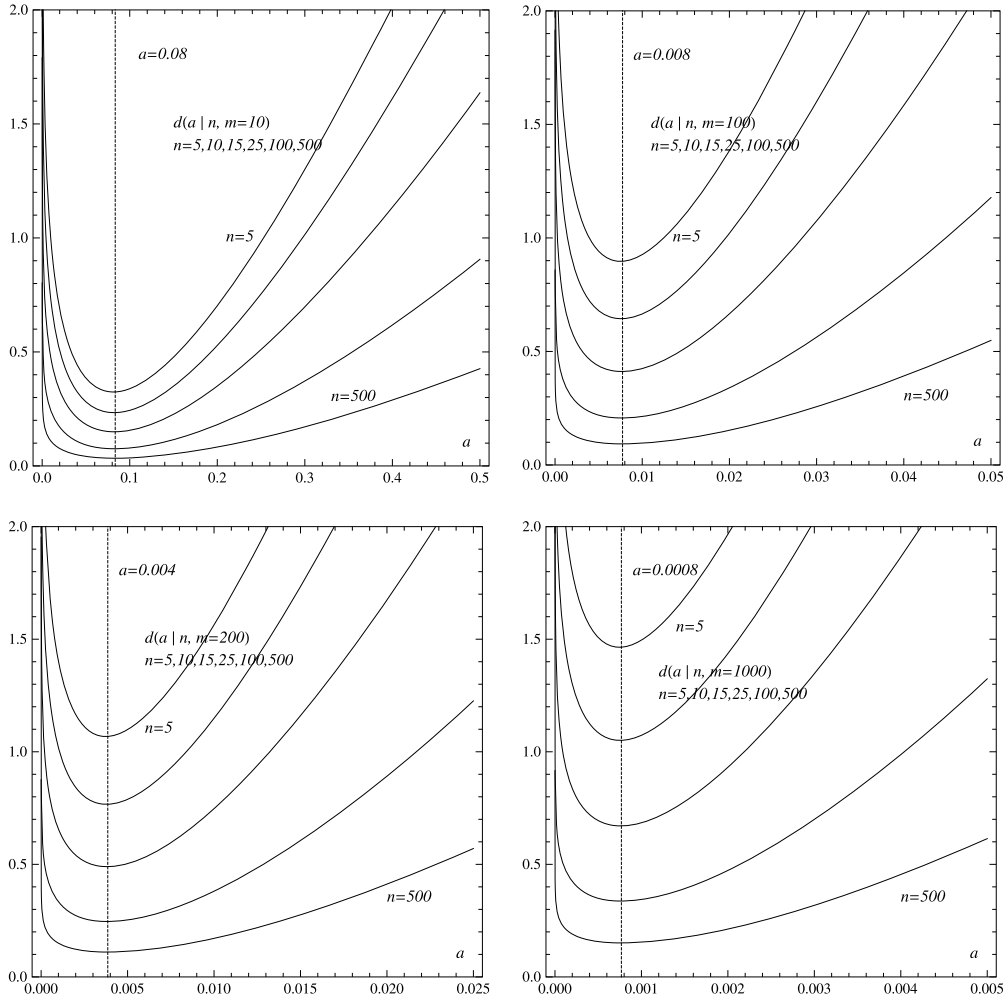


Figure 1: Expected logarithmic losses, when using a Dirichlet prior with parameter $\{a, \ldots, a\}$, in a multinomial model with $m$ cells, for sample sizes $n = 5, 10, 25, 100$ and 500. The panels are for $m = 10, 100, 200$ and 1000. In all cases, the optimal value for all sample sizes is $a^* \approx 0.8/m$.

Hence, using Definition 2 with uniform weights, the average expected logarithmic loss of using a joint Dirichlet prior with parameter $a$ with a sample of size $n$ is simply

$$d(a \,|\, m, n) = \sum_{x=0}^{n} \kappa\{a \,|\, x, m, n\}\, p(x \,|\, n)$$

since, by the symmetry of the problem, the $m$ parameters $\{\theta_1, \ldots, \theta_m\}$ yield the same expected loss.

The function $d(a \,|\, m = 10, n)$ is graphed in the upper left panel of Figure 1 for several values of $n$. The expected loss decreases with $n$ and, for any $n$, the function $d(a \,|\, m, n)$ is concave, with a unique minimum numerically found to be at $a^* \approx= 0.8/m = 0.08$. The approximation is rather precise. For instance, the minimum is achieved at 0.083 for $n = 100$.

Similarly, the function $d(a \,|\, m = 1000, n)$ is graphed in the lower right panel of Figure 1 for the same values of $n$ and with the same vertical scale, yielding qualitatively similar results although, as one may expect, the expected losses are now larger than those obtained with $m = 10$. Once more, the function $d(a \,|\, m = 1000, n)$ is concave, with a unique minimum numerically found to be at $a^* \approx 0.8/m = 0.0008$, with the exact value very close. For instance, for $n = 100$, the minimum is achieved at 0.00076.

If can be concluded that, for all practical purposes when using the reference distance approach, the best global Dirichlet prior, when one is interested in all the parameters of a multinomial model, is that with parameter vector $\{1/m, \ldots, 1/m\}$ (or $0.8 \times \{1/m, \ldots, 1/m\}$ to be slightly more precise), yielding an approximate marginal reference posterior for each of the $\theta_i$'s as $\text{Be}(\theta_i \,|\, x_i + 1/m, n - x_i + (m-1)/m)$, having mean and variance

$$\text{E}[\theta_i \,|\, x_i, n] = \hat{\theta}_i = (x_i + 1/m)/(n+1), \quad \text{Var}[\theta_i \,|\, x_i, n] = \hat{\theta}_i(1 - \hat{\theta}_i)/(n+2).$$

### The normal model with coefficient of variation

Consider a random sample $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ from a normal model $\text{N}(x \,|\, \mu, \sigma)$, with both parameters unknown, and suppose that one is interested in $\mu$ and $\sigma$, but also in the standardized mean $\phi = \mu/\sigma$ (and/or any one-to-one function of them such as $\log \sigma$, or the coefficient of variation $\sigma/\mu$).

The joint reference prior when either $\mu$ or $\sigma$ are the quantities of interest is

$$\pi_\mu(\mu, \sigma) = \pi_\sigma(\mu, \sigma) = \sigma^{-1} \tag{21}$$

and this is known to lead to the Student and squared root Gamma reference posteriors

$$\pi^{ref}_\mu(\mu \,|\, \boldsymbol{z}) = \text{St}(\mu \,|\, \overline{x}, s/\sqrt{n-1}, n-1)\,, \quad \pi^{ref}_\sigma(\sigma \,|\, \boldsymbol{z}) = \text{Ga}^{-1/2}(\sigma \,|\, (n-1)/2, ns^2/2),$$

with $n\overline{x} = \sum_{i=1}^n x_i$ and $ns^2 = \sum_{i=1}^n (x_i - \overline{x})^2$, which are proper if $n \geq 2$, and have the correct probability matching properties. However, the reference prior if $\phi$ is the parameter of interest is $\pi_\phi(\phi, \sigma) = (2 + \phi^2)^{-1/2}\sigma^{-1}$ (Bernardo, 1979), and the corresponding reference posterior distribution for $\phi$ can be shown to be

$$\pi^{ref}_\phi(\phi \,|\, \boldsymbol{z}) = \pi^{ref}_\phi(\phi \,|\, t) \propto (2 + \phi^2)^{-1/2} p(t \,|\, \phi)\,,$$

where $t = (\sum_{i=1}^n x_i)/(\sum_{i=1}^n x_i^2)^{1/2}$ has a sampling distribution $p(t \,|\, \phi)$ depending only on $\phi$ (see Stone and Dawid, 1972). Note that all posteriors can be written in terms of the sufficient statistics $\overline{x}$ and $s^2$ and the sample size $n$, which we will henceforth use.

A natural choice for the family of joint priors to be considered as candidates for an overall prior is the class of *relatively invariant* priors (Hartigan, 1964),

$$\mathcal{F} = \{\pi(\mu, \sigma \,|\, a) = \sigma^{-a},\, a > 0\}$$

which contains, for $a = 1$, the joint reference prior (21) when either $\mu$ or $\sigma$ are the parameters of interest, and the Jeffreys-rule prior, for $a = 2$. Since these priors are improper, a compact approximation procedure, as described at the end of Section 3.2, is needed. The usual compactification for location-scale parameters considers the sets

$$\mathcal{C}_k = \{\mu \in (-k, k),\ \sigma \in (e^{-k}, e^k)\}, \quad k = 1, 2, \ldots.$$

One must therefore derive

$$d(a \,|\, n, k) = d_\mu(a \,|\, n, k) + d_\sigma(a \,|\, n, k) + d_\phi(a \,|\, n, k),$$

where each of the $d_i$'s is found by integrating the corresponding risk with the appropriately renormalized joint reference prior. Thus,

$$\begin{aligned}
d_\mu(a \,|\, n, k) &= \int_{\mathcal{C}_k} \left[ \int_{\mathcal{T}} \kappa\{\pi_\mu(\cdot \,|\, n, \boldsymbol{t}, a) \,|\, \pi_\mu^{ref}(\cdot \,|\, n, \boldsymbol{t})\}\, p(\boldsymbol{t} \,|\, n, \mu, \sigma)\, d\boldsymbol{t} \right] \pi_\mu(\mu, \sigma \,|\, k)\, d\mu\, d\sigma, \\
d_\sigma(a \,|\, n, k) &= \int_{\mathcal{C}_k} \left[ \int_{\mathcal{T}} \kappa\{\pi_\sigma(\cdot \,|\, n, \boldsymbol{t}, a) \,|\, \pi_\sigma^{ref}(\cdot \,|\, n, \boldsymbol{t})\}\, p(\boldsymbol{t} \,|\, n, \mu, \sigma)\, d\boldsymbol{t} \right] \pi_\sigma(\mu, \sigma \,|\, k)\, d\mu\, d\sigma, \\
d_\phi(a \,|\, n, k) &= \int_{\mathcal{C}_k} \left[ \int_{\mathcal{T}} \kappa\{\pi_\phi(\cdot \,|\, n, \boldsymbol{t}, a) \,|\, \pi_\phi^{ref}(\cdot \,|\, n, \boldsymbol{t})\}\, p(\boldsymbol{t} \,|\, n, \mu, \sigma)\, d\boldsymbol{t} \right] \pi_\phi(\mu, \sigma \,|\, k)\, d\mu\, d\sigma,
\end{aligned}$$

where $\boldsymbol{t} = (\overline{x}, s)$, and the $\pi_i(\mu, \sigma \,|\, k)$'s are the joint *proper* prior reference densities of each of the parameter functions obtained by truncation and renormalization in the $\mathcal{C}_k$'s.

It is found that the risk associated to $\mu$ (the expected KL divergence of $\pi_\mu(\cdot \,|\, n, \boldsymbol{t}, a)$ from $\pi_\mu^{ref}(\cdot \,|\, n, \boldsymbol{t})$ under sampling) does *not* depend on the parameters, so integration with the joint prior is not required, and one obtains

$$d_\mu(a \,|\, n) = \log\left[ \frac{\Gamma[n/2]\,\Gamma[(a+n)/2 - 1]}{\Gamma[(n-1)/2]\,\Gamma[(a+n-1)/2]} \right] - \frac{a-1}{2} \left( \psi[\frac{n-1}{2}] - \psi[\frac{n}{2}] \right),$$

where $\psi[\cdot]$ is the digamma function. This is a concave function with a unique minimum $d_1(1 \,|\, n) = 0$ at $a = 1$, as one would expect from the fact that the target family $\mathcal{F}$ contains the reference prior for $\mu$ when $a = 1$. The function $d_\mu(a \,|\, n = 10)$ is the lower dotted line in Figure 2. Similarly, the risk associated to $\sigma$ does not depend either of the parameters, and one obtains

$$d_\sigma(a \,|\, n, k) = d_\sigma(a \,|\, n) = \log\left[ \frac{\Gamma[(a+n)/2 - 1]}{\Gamma[(n-1)/2]} \right] - \frac{a-1}{2}\, \psi[\frac{n-1}{2}],$$

another concave function with a unique minimum $d_2(1 \,|\, n) = 0$, at $a = 1$. The function $d_\sigma(a \,|\, n = 10)$ is the upper dotted line in Figure 2.

The risk associated with $\phi$ cannot be analytically obtained and is numerically computed, using one-dimensional numerical integration over $\phi$ to obtain the KL divergence, and Monte Carlo sampling to obtain its expected value with the truncated and renormalized reference prior $\pi_\phi(\mu, \sigma \,|\, k)$. The function $d_\phi(a \,|\, n = 10, k = 3)$ is represented by the black line in Figure 2. It may be appreciated that, of the three components of the

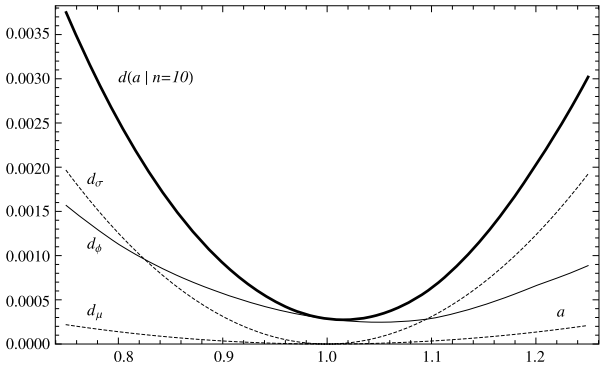


Figure 2: Expected average intrinsic logarithmic losses $d(a \,|\, n, k)$ associated with the use of the joint prior $\pi(\mu, \sigma \,|\, a) = \sigma^{-a}$ rather than the corresponding reference priors when $n = 10$ and $k = 3$.

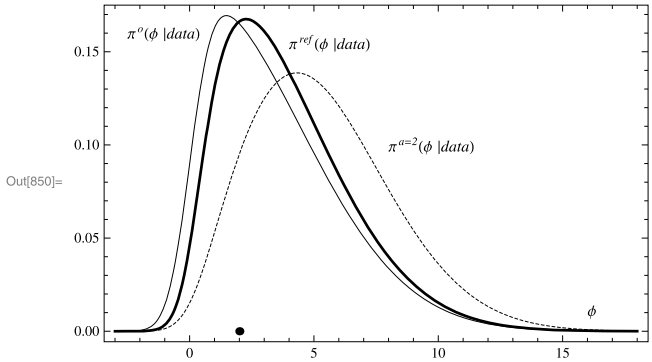


Figure 3: Reference posterior (solid) and marginal overall posterior (black) for $\phi$ given a minimal random sample of size $n = 2$. The dotted line is the marginal posterior for the prior with $a = 2$, which is the Jeffreys-rule prior.

expected loss, the contribution corresponding to $\phi$ is the largest, and that corresponding to $\mu$ is the smallest, in the neighborhood of the optimal choice of $a$. The sum of the three is the expected loss to be minimized, $d(a \,|\, n, k)$. The function $d(a \,|\, n = 10, k = 3)$ is represented by the solid line in Figure 2, and has a minimum at $a_3^* = 1.016$. The sequence of numerically computed optimum values is $\{a_k^*\} = \{1.139, 1.077, 1.016, \ldots\}$ quickly converging to some value $a^*$ larger than 1 and smaller than 1.016, so that, pragmatically, the overall objective prior may be taken to be the usual objective prior for the normal model,

$$\pi^o(\mu, \sigma) = \sigma^{-1}.$$

It is of interest to study the difference in use of this overall prior when compared with the reference prior for $\phi = \mu/\sigma$. The difference is greater for smaller samples, and

the minimum sample size here is $n = 2$. A random sample of two observations from $\mathrm{N}(x \,|\, 1, \frac{1}{2})$ (so that the true value of the standardized mean is $\phi = 2$) was simulated yielding $\{x_1, x_2\} = \{0.959, 1.341\}$. The corresponding reference posterior for $\phi$ is the solid line in Figure 3. The posterior that corresponds to the recommended overall prior $a = 1$ is the black line in the figure. For comparison, the posterior corresponding to the prior with $a = 2$, which is Jeffreys-rule prior, is also given, as the dotted line. Thus, even with a minimum sample size, the overall prior yields a marginal posterior for $\phi$ which is quite close to that for the reference posterior. (This was true for essentially all samples of size $n = 2$ that we tried.) For sample sizes beyond $n = 4$ the differences are visually inappreciable.

# 4 Hierarchical approach with hyperpriors

If a natural family of proper priors $\pi(\boldsymbol{\theta} \,|\, a)$, indexed by a single parameter $a$, can be identified for a given problem, one can compute the marginal likelihood $p(\boldsymbol{x} \,|\, a)$ (necessarily a proper density), and find the reference prior $\pi^R(a)$ for $a$ for this marginal likelihood. This hierarchical prior specification is clearly equivalent to use of

$$\pi^o(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} \,|\, a)\, \pi^R(a)\, da$$

as the overall prior in the original problem.

## 4.1 Multinomial problem

### The hierarchical prior

For the multinomial problem with the $\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a)$ prior, the marginal density of any of the $x_i$'s is

$$p(x_i \,|\, a, m, n) = \binom{n}{x_i} \frac{\Gamma(x_i + a)\, \Gamma(n - x_i + (m-1)a)\, \Gamma(m\,a)}{\Gamma(a)\, \Gamma((m-1)a)\, \Gamma(n + m\,a)},$$

following immediately from the fact that, marginally,

$$p(x_i \,|\, \theta_i) = \mathrm{Bi}(x_i \,|\, n, \theta_i) \quad \pi(\theta_i \,|\, a) = \mathrm{Be}(\theta_i \,|\, a, (m-1)a).$$

Then $\pi^R(a)$, the reference (Jeffreys) prior for the integrated model $p(\boldsymbol{x} \,|\, a)$ in (3), is given in the following proposition:

**Proposition 4.1.**

$$\pi^R(a \,|\, m, n) \propto \left[ \sum_{j=0}^{n-1} \left( \frac{Q(j \,|\, a, m, n)}{(a+j)^2} - \frac{m}{(m\,a + j)^2} \right) \right]^{1/2}, \tag{22}$$

*where $Q(\cdot \,|\, a, m, n)$ is the right tail of the distribution of $p(x \,|\, a, m, n)$, namely*

$$Q(j \,|\, a, m, n) = \sum_{l=j+1}^{n} p(l \,|\, a, m, n), \quad j = 0, \ldots, n-1.$$

*Proof.* Computation yields that

$$\mathrm{E}\left[-\frac{d^2}{da^2}\log p(\boldsymbol{x}\,|\,a)\right] = -\sum_{j=0}^{n-1}\frac{m^2}{(m\,a+j)^2} + E\left[\sum_{i=1}^{m}\sum_{j=0}^{x_i-1}\frac{1}{(a+j)^2}\right], \tag{23}$$

where $\sum_{j=0}^{-1} \equiv 0$. Since the $x_i$ are exchangeable, this equals

$$-\sum_{j=0}^{n-1}\frac{m^2}{(m\,a+j)^2} + mE^{X_1}\left[\sum_{j=0}^{X_1-1}\frac{1}{(a+j)^2}\right],$$

and the result follows by rearranging terms. □

**Proposition 4.2.** *$\pi^R(a)$ is a proper prior.*

*Proof.* The prior is clearly continuous in $a$, so we only need show that it is integrable at 0 and at $\infty$. Consider first the situation as $a \to \infty$. Then

$$\begin{aligned} p(0\,|\,a,m,n) &= \frac{\Gamma(a)\Gamma(n+[m-1]a)\Gamma(m\,a)}{\Gamma(a)\Gamma([m-1]a)\Gamma(n+m\,a)} \\ &= \frac{(m-1)a[(m-1)a+1]\cdots[(m-1)a+n-1]}{m\,a(m\,a+1)\cdots(m\,a+n-1)} \\ &= \frac{(m-1)}{m}(1-c_n a+O(a^2)), \end{aligned}$$

where $c_n = 1+1/2+\cdots+1/(n-1)$. Thus the first term of the sum in (22) is

$$\frac{1-p(0\,|\,a,m,n)}{a^2} - \frac{1}{m\,a^2} = \frac{(m-1)c_n}{m\,a} + O(1).$$

All of the other terms of the sum in (22) are clearly $O(1)$, so that

$$\pi^R(a) = \frac{\sqrt{(m-1)c_n/m}}{\sqrt{a}} + O(\sqrt{a}),$$

as $a \to 0$, which is integrable at zero (although unbounded).

To study propriety as $a \to \infty$, a laborious application of Stirling's approximation yields

$$p(x_1\,|\,a,m,n) = \mathrm{Bi}(x_1\,|\,n,1/m)(1+O(a^{-1})),$$

as $a \to \infty$. Thus

$$\begin{aligned} \pi^R(a,m,n) &= \left[\sum_{j=0}^{n-1}\left(\frac{\sum_{l=j+1}^{n}\mathrm{Bi}(l\,|\,n,1/m)}{a^2} - \frac{1}{m\,a^2}\right) + O(a^{-3})\right]^{1/2} \\ &= \left[\left(\frac{\sum_{l=1}^{n} l\mathrm{Bi}(l\,|\,n,1/m)}{a^2} - \frac{n}{m\,a^2}\right) + O(a^{-3})\right]^{1/2} = O(a^{-3/2}), \end{aligned}$$

which is integrable at infinity, completing the proof. □

As suggested by the proof above, the reference prior $\pi^R(a \,|\, m, n)$ behaves as $O(a^{-1/2})$ near $a = 0$ and behaves as $O(a^{-2})$ for large $a$ values. Using series expansions, it is found that, for sparse tables where $m/n$ is relatively large, the reference prior is well approximated by the proper prior

$$\pi^*(a \,|\, m, n) = \frac{1}{2}\,\frac{n}{m}\,a^{-1/2}\Big(a + \frac{n}{m}\Big)^{-3/2}, \tag{24}$$

which only depends on the ratio $m/n$, and has the behavior at the extremes described above. This can be restated as saying that $\phi(a) = a/(a+(n/m))$ has a Beta distribution $\mathrm{Be}(\phi \,|\, \frac{1}{2}, 1)$. Figure 4 gives the exact form of $\pi^R(a \,|\, m, n)$ for various $(m, n)$ values, and the corresponding approximation given by (24). The approximate reference prior $\pi^*(a \,|\, m, n)$ appears to be a good approximation to the actual reference prior, and hence can be recommended for use with large sparse contingency tables.

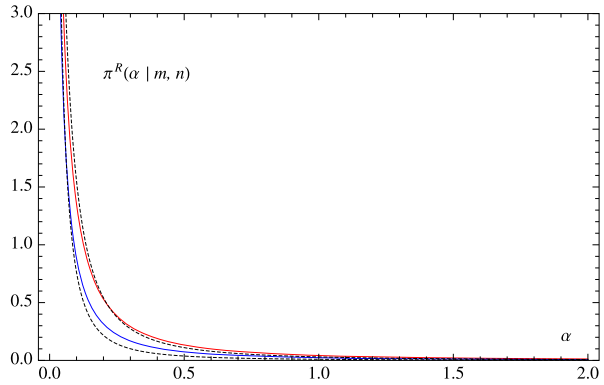


Figure 4: Reference priors $\pi^R(a \,|\, m, n)$ (solid lines) and their approximations (dotted lines) for $(m = 150, n = 10)$ (upper curve) and for $(m = 500, n = 10)$ (lower curve).

It is always a surprise when a reference prior turns out to be proper, and this seems to happen when the likelihood does not go to zero at a limit. Indeed, it is straightforward to show that

$$p(\boldsymbol{x} \,|\, a) = \begin{cases} O(a^{r_0 - 1}), & \text{as } a \to 0, \\ \binom{n}{\boldsymbol{x}}\, m^{-n}, & \text{as } a \to \infty, \end{cases}$$

where $r_0$ is the number of nonzero $x_i$. Thus, indeed, the likelihood is constant at $\infty$, so that the prior must be proper at infinity for the posterior to exist.

### Computation with the hierarchical reference prior

If a full Bayesian analysis is desired, the obvious MCMC sampler is as follows:

*Step 1.* Use a Metropolis Hastings move to sample from the marginal posterior $\pi^R(a \,|\, \boldsymbol{x}) \propto \pi^R(a)\, p(\boldsymbol{x} \,|\, a)$.

*Step 2.* Given $a$, sample from the usual beta posterior $\pi(\theta \,|\, a, \boldsymbol{x})$.

This will be highly efficient if a good proposal distribution for Step 1 can be found. As it is only a one-dimensional distribution, standard techniques should work well. Even simpler computationally is the use of the approximate reference prior $\pi^*(a \mid m, n)$ in (24), because of the following result.

**Proposition 4.3.** *Under the approximate reference prior (24), and provided there are at least three nonempty cells, the marginal posterior distribution of $a$ is log-concave.*

*Proof.* It follows from (23) that

$$\frac{d^2}{da^2} \log[p(\boldsymbol{x} \mid a)\pi^*(a \mid m, n)] = \sum_{j=0}^{n-1} \frac{m^2}{(m\,a + j)^2} - \sum_{i=1}^{m} \sum_{j=0}^{x_i - 1} \frac{1}{(a + j)^2} + \frac{1}{2a^2} + \frac{3}{2(a + n/m)^2}\,.$$

Without loss of generality, we assume that $x_i > 0$, for $i = 1, 2, 3$. Then

$$\frac{d^2}{da^2} \log[p(\boldsymbol{x} \mid a)p^*(a \mid m, n)] < -\sum_{i=2}^{3} \sum_{j=0}^{x_i - 1} \frac{1}{(a + j)^2} + \frac{1}{2a^2} + \frac{3}{2a^2} < 0.$$

□

Thus adaptive rejection sampling (Gilks and Wild, 1992) can be used to sample from the posterior of $a$.

Alternatively, one might consider the empirical Bayes solution of fixing $a$ at its posterior mode $\widehat{a}^R$. The one caveat is that, when $r_0 = 1$, it follows from (25) that the likelihood is constant at zero, while $\pi^R(a)$ is unbounded at zero; hence the posterior mode will be $a = 0$, which cannot be used. When $r_0 \geq 2$, it is easy to see that $\pi^R(a)p(\boldsymbol{x} \mid a)$ goes to zero as $a \to 0$, so there will be no problem.

It will typically be considerably better to utilize the posterior mode than the maximum of $p(\boldsymbol{x} \mid a)$ alone, given the fact that the likelihood does not go to zero at $\infty$. For instance, if all $x_i = 1$, it can be shown that $p(\boldsymbol{x} \mid a)$ has a likelihood increasing in $a$, so that there is no mode. (Even when $r_0 = 1$, use of the mode of $p(\boldsymbol{x} \mid a)$ is not superior, in that the likelihood is also maximized at 0 in that case.)

### Posterior behavior as $m \to \infty$

Since we are contemplating the "large sparse" contingency table scenario, it is of considerable interest to study the behavior of the posterior distribution as $m \to \infty$. It is easiest to state the result in terms of the transformed variable $v = m\,a$. Let $\pi_m^R(v \mid \boldsymbol{x})$ denote the transformed reference posterior.

**Proposition 4.4.**

$$\Psi(v) = \lim_{m \to \infty} \pi_m^R(v \mid \boldsymbol{x}) = \frac{\Gamma(v + 1)}{\Gamma(v + n)}\, v^{(r_0 - \frac{3}{2})} \left[\sum_{i=1}^{n-1} \frac{i}{(v + i)^2}\right]^{1/2}. \tag{25}$$

*Proof.* Note that

$$\pi^R(a \mid \boldsymbol{x}) \quad \propto \quad m(\boldsymbol{x} \mid a)\, \pi^R(a)$$

$$\begin{aligned}
&\propto \frac{\Gamma(m\,a)}{\Gamma(m\,a+n)}\left[\prod_{i=1}^{m}\frac{\Gamma(a+x_i)}{\Gamma(a)}\right]\pi^R(a)\\
&\propto \frac{\Gamma(m\,a)}{\Gamma(m\,a+n)}\left[\prod_{i:x_i\neq 0} a(a+1)\;s(a+x_i-1)\right]\pi^R(a)\\
&\propto \frac{\Gamma(m\,a)}{\Gamma(m\,a+n)}\left[\prod_{j=0}^{n-1}(a+j)^{r_j}\right]\pi^R(a)\,,
\end{aligned}$$

where $r_j = \{\#x_i > j\}$. Change of variables to $v = m\,a$ yields

$$\begin{aligned}
\pi_m^R(v\,|\,\boldsymbol{x}) &\propto \frac{\Gamma(v)}{\Gamma(v+n)}\left[\prod_{j=0}^{n-1}\Big(\frac{v}{m}+j\Big)^{r_j}\right]\pi^R\Big(\frac{v}{m}\Big)\\
&\propto \frac{\Gamma(v)\;v^{r_0}}{\Gamma(v+n)}\left[C+\sum_{i=1}^{n-r_0}K_i\Big(\frac{v}{m}\Big)^i\right]\pi^R\Big(\frac{v}{m}\Big)\,, \qquad (26)
\end{aligned}$$

where $C = \prod_{j=2}^{n-1} j^{r_j}$ and the $K_i$ are constants.

Next we study the behavior of $\pi^R(v/m)$ for large $m$. Note first that, in terms of $v$, the marginal density of $x_1 = 0$ is

$$\begin{aligned}
p(0\,|\,v) &= \frac{\Gamma(\frac{(m-1)}{m}v+n)}{\Gamma(\frac{(m-1)}{m}v)}\;\frac{\Gamma(v)}{\Gamma(v+n)}\\
&= \frac{\frac{(m-1)}{m}v[\frac{(m-1)}{m}v+1]\cdots[\frac{(m-1)}{m}v+n-1]}{v(v+1)\cdots(v+n-1)}\\
&= \frac{(m-1)}{m}\left(1-\frac{v}{m[v+1]}\right)\cdots\left(1-\frac{v}{m[v+n-1]}\right)\\
&= \frac{(m-1)}{m}\left(1-\frac{v}{m}\sum_{i=1}^{n-1}\frac{1}{v+i}+O\left(\frac{v^2}{m^2(v+1)^2}\right)\right).
\end{aligned}$$

Hence

$$\begin{aligned}
Q(0\,|\,a) &= 1-p(0\,|\,v)\\
&= \frac{1}{m}+\frac{v(m-1)}{m^2}\sum_{i=1}^{n-1}\frac{1}{v+i}+O\left(\frac{v^2}{m^2(v+1)^2}\right)=O\left(\frac{1}{m}\right)\ \text{(uniformly in } v)\,.
\end{aligned}$$

It follows that all $Q(i\,|\,a) \le O(1/m)$, so that $\pi^R\left(\frac{v}{m}\right)$ is proportional to

$$\left[\Big(\frac{m}{v}\Big)^2\left(\frac{1}{m}+\frac{v(m-1)}{m^2}\sum_{i=1}^{n-1}\frac{1}{v+i}\right)+O(1)+\sum_{j=1}^{n-1}\frac{1}{(\frac{v}{m}+j)^2}O\left(\frac{1}{m}\right)-\sum_{i=0}^{n-1}\frac{m}{(v+i)^2}\right]^{1/2}$$

$$= \left[\sum_{i=1}^{n-1} \frac{1}{v+i}\left(\frac{(m-1)}{v} - \frac{m}{(v+i)}\right) + O(1)\right]^{1/2}$$
$$= \left[\frac{(m-1)}{v}\sum_{i=1}^{n-1} \frac{i}{(v+i)^2} + O(1)\right]^{1/2}$$
$$= \sqrt{m-1}\left[\frac{1}{v}\sum_{i=1}^{n-1} \frac{i}{(v+i)^2} + O\left(\frac{1}{m}\right)\right]^{1/2}.$$

Combining this with (26), noting that $v\Gamma(v) = \Gamma(v+1)$, and letting $m \to \infty$, yields the result. □

It follows, of course, that $a$ behaves like $v/m$ for large $m$, where $v$ has the distribution in (25). It is very interesting that this "large $m$" behavior of the posterior depends on the data only through $r_0$, the number of nonzero cell observations.

If, in addition, $n$ is moderately large (but much smaller than $m$), we can explicitly study the behavior of the posterior mode of $a$.

**Proposition 4.5.** *Suppose $m \to \infty$, $n \to \infty$, and $n/m \to 0$. Then (25) has mode*

$$\hat{v} \approx \begin{cases} \frac{(r_0-1.5)}{\log(1+n/r_0)} & \text{if } \frac{r_0}{n} \to 0, \\ c^* n & \text{if } \frac{r_0}{n} \to c < 1, \end{cases}$$

*where $r_0$ is the number of nonzero $x_i$, $c^*$ is the solution to $c^* \log(1 + \frac{1}{c^*}) = c$, and $f(n,m) \approx g(n,m)$ means $f(n,m)/g(n,m) \to 1$. The corresponding mode of the reference posterior for $a$ is $\hat{a}^R = \hat{v}/m$.*

*Proof.* Taking the log of (25) and differentiating with respect to $v$ results in

$$\Psi'(v) = \frac{(r_0 - 1.5)}{v} - \sum_{i=1}^{n-1} \frac{1}{v+i} - \frac{\sum_{i=1}^{n-1} \frac{i}{(v+i)^3}}{\sum_{i=1}^{n-1} \frac{i}{(v+i)^2}}.$$

Note first that, as $n$ grows, and if $v$ also grows (no faster than $n$), then

$$\sum_{i=1}^{n-1} \frac{1}{v+i} = \int_1^n \frac{1}{v+x}\,dx + O\left(\frac{1}{v+1}\right) + O\left(\frac{1}{n}\right) = \log\left(\frac{v+n}{v+1}\right) + O\left(\frac{1}{v+1}\right) + O\left(\frac{1}{n}\right).$$

Next,

$$\sum_{i=1}^{n-1} \frac{i}{(v+i)^3} = \int_1^n \frac{x}{(v+x)^3}\,dx + O\left(\frac{1}{(v+1)^2}\right) + O\left(\frac{1}{n^2}\right)$$
$$= \frac{1}{2}\left[\frac{(v+2)}{(v+1)^2} - \frac{(v+2n)}{(v+n)^2}\right] + O\left(\frac{1}{(v+1)^2} + \frac{1}{n^2}\right) = O\left(\frac{1}{v+1}\right) + O\left(\frac{1}{n}\right),$$

$$\sum_{i=1}^{n-1} \frac{i}{(v+i)^2} = \int_1^n \frac{x}{(v+x)^2}\, dx + O\left(\frac{1}{(v+1)}\right) + O\left(\frac{1}{n}\right)$$
$$= \frac{v(1+n)}{(v+1)(v+n)} + \log\left(\frac{v+n}{v+1}\right) + O\left(\frac{1}{v+1} + \frac{1}{n}\right) \geq \log 2,$$

again using that $v$ will not grow faster than $n$. Putting these together we have that

$$\Psi'(v) = \frac{(r_0 - 1.5)}{v} - \log\left(\frac{v+n}{v+1}\right) + O\left(\frac{1}{v+1}\right) + O\left(\frac{1}{n}\right).$$

*Case 1.* $\frac{r_0}{n} \to c$, for $0 < c < 1$. For this case, write $v = c^* n/(1+\delta)$ for $\delta$ small, and note that then

$$\Psi'(v) = \frac{c}{c^*}(1 + o(1))(1+\delta) - \log\left(\frac{(c^*+1)}{c^*}\right) + o(1).$$

Since $\frac{c}{c^*} - \log\left(\frac{(c^*+1)}{c^*}\right) = 0$, it is clear that $\delta$ can be appropriately chosen as $o(1)$ to make the derivative zero.

*Case 2.* $\frac{r_0}{n} \to 0$. Now choose $v = \frac{(r_0 - 1.5)}{(1+\delta)\log(1+n/r_0)}$ and note that $\frac{v}{n} \to 0$. It follows that

$$\log\left(1 + \frac{n}{r_0}\right) = (\log n - \log r_0 + o(1))(1+\delta) \quad \text{and}$$
$$\log\left(\frac{v+n}{v+1}\right) = [\log n - \log(v+1)](1 + o(1)).$$

Consider first the case $v \to \infty$. Then

$$\log(v+1) = (1+o(1))(\log r_0 - \log\log(1 + n/r_0)) = (1+o(1))\log r_0,$$

so that

$$\Psi'(v) = (\log n - \log r_0 + o(1))(1+\delta) - (\log n - \log r_0)(1+o(1)) + o(1),$$

and it is clear that $\delta$ can again be chosen $o(1)$ to make this zero. Lastly, if $v \leq K < \infty$, then $(\log r_0)/(\log n) = o(1)$, so that $\Psi'(v) = (\log n)(1+o(1))(1+\delta) - (\log n)(1+o(1)) + o(1)$, and $\delta$ can again be chosen $o(1)$ to make this zero, completing the proof. □

Table 1 gives the limiting behavior of $\hat{v}$ for various behaviors of the number of nonzero cells, $r_0$. Only when $r_0 = \log n$ does the posterior mode of $a$ (i.e., $v/m$) equal $1/m$, the value selected by the reference distance method. Of course, this is not surprising; empirical Bayes is using a fit to the data to help select $a$ whereas the reference distance method is pre-experimental.

| $r_0$ | $cn$ $(0 < c < 1)$ | $n^b$ $(0 < b < 1)$ | $(\log n)^b$ | $\log n$ | $O(1)$ |
|---|---|---|---|---|---|
| $\hat{v}$ | $c^* n$ | $\frac{n^b}{(1-b)\log n}$ | $(\log n)^{(b-1)}$ | 1 | $O(1/\log n)$ |

Table 1: The limiting behavior of $\hat{v}$ as $n \to \infty$, for various limiting behaviors of $r_0$, the number of non-zero cells.

## 4.2 Multivariate hypergeometric model

Let $\mathcal{N}_+$ be the set of all nonnegative integers. Consider a multivariate hypergeometric distribution $\mathrm{Hy}_k(\boldsymbol{r}_k \,|\, n, \boldsymbol{R}_k, N)$ with the probability mass function

$$\mathrm{Hy}_k(\boldsymbol{r}_k \,|\, n, \boldsymbol{R}_k, N) \;=\; \frac{\binom{R_1}{r_1} \cdots \binom{R_k}{r_k}\binom{R_{k+1}}{r_{k+1}}}{\binom{N}{n}}\,, \quad \boldsymbol{r}_k \in \mathcal{R}_{k,n}, \tag{27}$$

$$\mathcal{R}_{k,n} = \{\boldsymbol{r}_k = (r_1, \cdots, r_k); \quad r_j \in \mathcal{N}_+, \quad r_1 + \cdots + r_k \le n\},$$

where the $k$ unknown parameters $\boldsymbol{R}_k = (R_1, \cdots, R_k)$ are in the parameter space $\mathcal{R}_{k,N}$. Here and in the following, $R_{k+1} = N - (R_1 + \cdots + R_k)$. Notice that the univariate hypergeometric distribution is the special case when $k = 1$.

A natural hierarchical model for the unknown $\boldsymbol{R}_k$ is to assume that it is multinomial $\mathrm{Mu}_k(\boldsymbol{R}_k \,|\, N, \boldsymbol{p}_k)$, with $\boldsymbol{p}_k \in \mathcal{P}_k \equiv \{\boldsymbol{p}_k = (p_1, \cdots, p_k)\}$, $0 \le p_j \le 1$, and $p_1 + \cdots + p_k \le 1$. The probability mass function of $\boldsymbol{R}_k$ is then

$$\mathrm{Mu}_k(\boldsymbol{R}_k \,|\, N, \boldsymbol{p}_k) = \frac{N!}{\prod_{j=1}^{k+1} R_j!} \prod_{j=1}^{k+1} p_j^{R_j}.$$

Berger, Bernardo and Sun (2012) prove that the marginal likelihood of $\boldsymbol{r}_k$ depends only on $(n, \boldsymbol{p}_k)$ and it is given by

$$\begin{aligned} p(\boldsymbol{r}_k \,|\, \boldsymbol{p}_k, n, N) \;&=\; p(\boldsymbol{r}_k \,|\, \boldsymbol{p}_k, n) = \sum_{\boldsymbol{R}_k \in \mathcal{N}_{k,N}} \mathrm{Hy}_k(\boldsymbol{r}_k \,|\, n, \boldsymbol{R}_k, N)\, \mathrm{Mu}_k(\boldsymbol{R}_k \,|\, N, \boldsymbol{p}_k) \\ &=\; \mathrm{Mu}_k(\boldsymbol{r}_k \,|\, n, \boldsymbol{p}_k), \;\; \boldsymbol{r}_k \in \mathcal{R}_{k,n}. \end{aligned} \tag{28}$$

This reduces to the multinomial problem. Hence, the overall (approximate) reference prior for $(\boldsymbol{R}_k \,|\, N, \boldsymbol{p}_k)$ would be Multinomial-Dirichlet $\mathrm{Di}(\boldsymbol{R}_k \,|\, 1/k, \cdots, 1/k)$.

## 4.3 Multi-normal means

Let $x_i$ be independent normal with mean $\mu_i$ and variance 1, for $i = 1 \cdots, m$. We are interested in all the $\mu_i$ and in $|\boldsymbol{\mu}|^2 = \mu_1^2 + \cdots + \mu_m^2$.

The natural hierarchical prior modeling approach is to assume that $\mu_i \overset{iid}{\sim} \mathrm{N}(\mu_i \,|\, 0, \tau)$. Then, marginally, the $x_i$ are iid $\mathrm{N}(x_1 \,|\, 0, \sqrt{1+\tau^2})$ and the reference (Jeffreys) prior for $\tau^2$ in this marginal model is

$$\pi^R(\tau^2) \propto (1+\tau^2)^{-1}.$$

The hierarchical prior for $\boldsymbol{\mu}$ (and recommended overall prior) is then

$$\pi^o(\boldsymbol{\mu}) = \int_0^\infty \frac{1}{(2\pi\tau^2)^{m/2}} \exp\left(-\frac{|\boldsymbol{\mu}|^2}{2\tau^2}\right) \frac{1}{1+\tau^2}\, d\tau^2 \,. \tag{29}$$

This prior is arguably reasonable from a marginal reference prior perspective. For the individual $\mu_i$, it is a shrinkage prior known to result in Stein-like shrinkage estimates of the form

$$\hat{\mu}_i = \left(1 - \frac{r(|\boldsymbol{x}|)}{|\boldsymbol{x}|^2}\right) x_i \,,$$

with $r(\cdot) \approx p$ for large arguments. Such shrinkage estimates are often viewed as actually being superior to the reference posterior mean, which is just $x_i$ itself. The reference prior when $|\mu|$ is the parameter of interest is

$$\pi_{|\boldsymbol{\mu}|}(\boldsymbol{\mu}) \propto \frac{1}{|\boldsymbol{\mu}|^{m-1}} \propto \int_0^\infty \frac{1}{(2\pi\tau^2)^{m/2}} \exp\left(-\frac{|\boldsymbol{\mu}|^2}{2\tau^2}\right) \frac{1}{\tau}\, d\tau^2 \,, \tag{30}$$

which is similar to (29) in that, for large values of $|\boldsymbol{\mu}|$, the tails differ by only one power. Thus the hierarchical prior appears to be quite satisfactory in terms of its marginal posterior behavior for any of the parameters of interest. Of course, the same could be said for the single reference prior in (30); thus here is a case where one of the reference priors would be fine for all parameters of interest, and averaging among reference priors would not work.

Computation with the reference prior in (30) can be done by a simple Gibbs sampler. Computation with the hierarchical prior in (29) is almost as simple, with the Gibbs step for $\tau^2$ being replaced by the rejection step:

*Step 1.* Propose $\tau^2$ from the inverse gamma density proportional to

$$\frac{1}{(\tau^2)^{(1+m/2)}} \exp\left(-\frac{|\boldsymbol{\mu}|^2}{2\tau^2}\right),$$

*Step 2.* Accept the result with probability $\tau^2/(\tau^2+1)$ (or else propose again).

## 4.4 Bivariate normal problem

Earlier for the bivariate normal problem, we only considered the two right-Haar priors. More generally, there is a continuum of right-Haar priors given as follows. Define an orthogonal matrix by

$$\boldsymbol{\Gamma} = \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}$$

where $-\pi/2 < \beta \le \pi/2$. Then it is straightforward to see that the right-Haar prior based on the transformed data $\boldsymbol{\Gamma X}$ is

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \,|\, \beta) = \frac{\sin^2(\beta)\,\sigma_1^2 + \cos^2(\beta)\,\sigma_2^2 + 2\sin(\beta)\cos(\beta)\,\rho\,\sigma_1\,\sigma_2}{\sigma_1^2\,\sigma_2^2\,(1-\rho^2)}.$$

We thus have a class of priors indexed by a hyperparameter $\beta$, and it might be tempting to try the hierarchical approach even though the class of priors is not a class of proper priors and hence there is no proper marginal distribution to utilize in finding the hyperprior for $\beta$. The temptation here arises because $\beta$ is in a compact set and it seems natural to use the (proper) uniform distribution (being uniform over the set of rotations is natural.) The resulting joint prior is

$$\pi^o(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \,|\, \beta)\, d\beta\,,$$

which equals the prior $\pi^A$ in (1), since

$$\int_{-\pi/2}^{\pi/2} \sin(\beta)\cos(\beta) d\beta = 0, \quad \int_{-\pi/2}^{\pi/2} \sin^2(\beta) d\beta = \int_{-\pi/2}^{\pi/2} \cos^2(\beta) d\beta = \text{constant}\,.$$

Thus the overall prior obtained by the hierarchical approach is the same prior as obtained by just averaging the two reference priors. It was stated there that this prior is inferior as an overall prior to either reference prior individually, so the attempt to apply the hierarchical approach to a class of improper priors has failed.

**Empirical hierarchical approach:** Instead of integrating out over $\beta$, one could find the empirical Bayes estimate $\hat{\beta}$ and use $\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \,|\, \hat{\beta})$ as the overall prior. This was shown in Sun and Berger (2007) to result in a terrible overall prior, much worse than either the individual reference priors, or even $\pi^A$ in (1).

## 5 Discussion

When every parameter of a model has the same reference prior, this prior is very natural to use as the overall prior. A number of such scenarios were catalogued in Section 2. This common reference prior can depend on the parameterization chosen for the models (although it will be invariant to coordinatewise one-to-one-transformations). Indeed, an example was given in which a strange choice of model parameterization resulted in an inadequate common reference prior.

The reference distance approach to developing an overall prior is natural, and seems to work well when the reference priors themselves are proper. It also appears to be possible to implement the approach in the case where the reference priors are improper, by operating on suitable large compact sets and showing that the result is not sensitive to the choice of compact set. Of course, the approach is dependent on the parameterization used for the model and on having accepted reference priors available for all the parameters in the model; it would have been more satisfying if the overall prior depended only on the model itself. The answer will also typically depend on weights used for the individual reference priors, although this can be viewed as a positive in allowing more important parameters to have more influence. The implementation considered in this paper also utilized a class of candidate priors, with the purpose of finding the candidate which minimized the expected risk. The result will thus depend on the choice of the

candidate class although, in principle, one could consider the class of all priors as the candidate class; the resulting minimization problem would be formidable, however.

The hierarchical approach seems excellent (as usual), and can certainly be recommended if one can find a natural hierarchical structure based on a class of proper priors. Such hierarchical structures naturally occur in settings where parameters can be viewed as exchangeable random variables but may not be available otherwise. In the particular examples considered, the overall prior obtained for the multi-normal mean problem seems fine, and the recommended hierarchical prior for the contingency table situation is very interesting, and seems to have interesting adaptations to sparsity; the same can be said for its empirical Bayes implementation. In contrast, the attempted application of the hierarchical and empirical Bayes idea to the bivariate normal problem using the class of right-Haar priors was highly unsatisfactory, even though the hyperprior was proper. This is a clear warning that the hierarchical or empirical Bayes approach should be based on an initial class of proper priors.

The failure of arithmetic prior averaging in the bivariate normal problem was also dramatic; the initial averaging of two right-Haar priors gave an inferior result, which was duplicated by the continuous average over all right-Haar priors. Curiously in this example, the geometric average of the two right-Haar improper priors seems to be reasonable, suggesting that, if averaging of improper priors is to be done, the geometric average should be used.

The 'common reference prior' and 'reference distance' approaches will give the same answer when a common reference prior exists. However, the reference distance and hierarchical approaches will rarely give the same answer because, even if the initial class of candidate priors is the same, the reference distance approach will fix the hyperparameter $a$, while the hierarchical approach will assign it a reference prior; and, even if the empirical Bayes version of the hierarchical approach is used, the resulting estimate of $a$ can be different than that obtained from the reference distance approach, as indicated in the multinomial example at the end of Section 4.1.

The 'common reference prior' and hierarchical approaches will mostly have different domains of applicability and are the recommended approaches when they can be applied. The reference distance approach will be of primary utility in situations such as the coefficient of variation example in Section 3.2, where there is no natural hierarchical structure to utilize nor common reference prior available.

#### Acknowledgments

Berger's work was supported by NSF Grants DMS-0757549-001 and DMS-1007773, and by Grant 53-130-35-HiCi from King Abdulaziz University. Sun's work was supported by NSF grants DMS-1007874 and SES-1260806. The research is also supported by Chinese 111 Project B14019.

## References

Bar-Lev, S. K. and Reiser, B. (1982). An exponential subfamily which admits UMPU tests based on a single test statistic. *The Annals of Statistics* **10**, 979–989. 198

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association* **84**, 200–207. 189

Berger, J. O. and Bernardo, J. M. (1992a). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (with discussion). 189

Berger, J. O. and Bernardo, J. M. (1992b). Ordered group reference priors, with applications to multinomial problems. *Biometrika* **79**, 25–37. 190, 198

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics* **37**, 905–938. 189, 193

Berger, J. O., Bernardo, J. M. and Sun, D. (2012). Objective priors for discrete parameter spaces. *Journal of the American Statistical Association* **107**, 636-648. 189, 215

Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics* **36**, 963–982. 190, 191, 192, 200

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B* **41**, 113–147 (with discussion). 189, 201, 205

Bernardo, J. M. (2005). Reference analysis. *Bayesian Thinking: Modeling and Computation, Handbook of Statistics* **25** (D. K. Dey and C. R Rao, eds). Amsterdam: Elsevier, 17–90. 189

Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 1–68 (with discussion).

Bernardo, J. M. (2006). Intrinsic point estimation of the normal variance. *Bayesian Statistics and its Applications*. (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) New Delhi: Anamaya Pub, 110–121.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. 189

Clarke, B. and Barron, A. (1994). Jeffreys' prior is the reference prior under entropy loss. *Journal of Statistical Planning and Inference* **41**, 37–60. 189

Clarke, B. and Yuan A. (2004). Partial information reference priors: derivation and interpretations. *Journal of Statistical Planning and Inference* **123**, 313–345. 189

Consonni, G., Veronese, P. and Gutiérrez-Peña E. (2004). Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Analysis* **88**, 335–364. 189

Crowder, M. and Sweeting, T. (1989). Bayesian inference for a bivariate binomial distribution. *Biometrika* **76**, 599–603. 197

Datta, G. S. and Ghosh, J. K. (1995a). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45. 189

Datta, G. S. and Ghosh, J. K. (1995b). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114. 189

Datta, G. S. and Ghosh, M. (1996). On the invariance of noninformative priors. *The Annals of Statistics* **24**, 141–159. 189, 196

Datta, G. S., Mukerjee, R., Ghosh, M. and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity. *The Annals of Statistics* **28**, 1414–1426. 189

De Santis, F., Mortea, J. and Nardi, A. (2001). Jeffreys priors for survival models with censored data. *Journal of Statistical Planning and Inference* **99**, 193–209. 189

De Santis, F. (2006). Power priors and their use in clinical trials. *The American Statistician* **60**, 122–129. 189

Enis, P. and Geisser, S. (1971). Estimation of the probability that $Y < X$. *Journal of the American Statistical Association* **66**, 162–168. 199

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* New York: Springer 189

Ghosh, M., Mergel, V., and Liu, R. (2011). A general divergence criterion for prior selection. *Annals of the Institute of Statistical Mathematics* **60**, 43–58. 189

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science* **26**, 187–202. 189

Ghosh, M. and Sun, D. (1998). Recent developments of Bayesian inference for stress-strength models. *Frontiers in Reliability.* Indian Association for Productivity Quality and Reliability (IAPQR), 143-158. 200

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348. 211

Hartigan, J. A. (1964). Invariant prior distributions. *Annals of Mathematical Statistics* **35**, 836–845. 205

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society, Series A* **186**, 453–461. 192

Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford: Oxford University Press. 192

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370. 189

Kullback, S. and R. A. Leibler, R .A. (1951). On information and suffiency. *Annals of Mathematical Statistics* **22**, 79–86. 202

Laplace, P. S. (1812). *Théorie Analytique des Probabilités.* Paris: Courcier. Reprinted as *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars. 191

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295-304. 189

Liseo, B, and Loperfido, N, (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference* **136**, 373–389. 189

Polson, N. and Wasserman, L. (1990). Prior distributions for the bivariate binomial. *Biometrika* **77**, 901–904. 197

Sivaganesan, S. (1994). Discussion to "An Overview of Bayesian Robustness" by J. Berger. *Test* **3**, 116–120. 189

Sivaganesan, S., Laud, P., Mueller, P. (2011). A Bayesian subgroup analysis using zero-inflated Polya-urn scheme. *Sociological Methodology* **30**, 312–323. 189

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Annals of mathematical Statistics* **30**, 877–880. 191

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics* **16**, 155–160. 202

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika* **59**, 269–375. 205

Sun, D. (1994). Integrable expansions for posterior distributions for a two-parameter exponential family. *The Annals of Statistics* **22**, 1808-1830. 198

Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83**, 55-65. 199

Sun, D. and Berger, J. O. (1998). Reference priors under partial information. *Biometrika* **85**, 55–71. 197

Sun, D. and Berger, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 525–562 (with discussion). 194, 217

Walker, S. G. and Gutiérrez-Peña, E. (2011). A decision-theoretical view of default priors *Theory and Decision* **70**, 1–11. 189

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics* **22**, 1195–2111. 192

Bayesian Analysis (2015) **10**, Number 1, pp. 223–226

# Comment on Article by Berger, Bernardo, and Sun*

Siva Sivaganesan†

Congratulations to the authors on this important paper that leads the way in selecting an objective overall prior for estimation. The paper is very enjoyable to read.

The authors provide three possible approaches one could use to find an overall objective prior suitable for use when there is interest in simultaneous estimation of several parameters. They illustrate the approaches in several examples, and give a comprehensive evaluation of the resulting priors. The proposed new approaches are very carefully thought out, and hold much promise for the development of a single overall objective prior in many more models. This is a very interesting paper and is likely to, and hopefully will, spur increased research in this new development to find overall objective priors for estimation.

Selection of good objective priors is very important in the practice of Bayesian analysis since, often, there is little or no prior information available for at least some of the parameters, especially in complex models with large number of parameters. Use of diffuse priors is not always good or optimal. The reference prior approach has been very successful in providing a way to get objective priors for estimation in numerous standard and non-standard models. It was introduced in Bernardo (1979) to derive a non-informative prior for estimation of a scalar parameter. In simple terms, the reference prior is the prior that maximizes, in an asymptotic sense, the missing information in a prior measured by the Kullback–Leibler distance between the prior and the posterior distribution. The approach gave good priors in the one-parameter case, but did not easily extend to multi-parameter cases. A series of influential articles beginning with Berger and Bernardo (1989, 1992), and later by Berger, Bernardo and Sun extended the reference prior approach to multi-parameter problems. and formalized the approach, e.g., see Berger et al. (2009, 2012), Sun and Berger (1998), and Berger and Sun (2008). It is reasonable to say that the reference prior approach is the best formal approach to obtain an objective prior for estimation.

The literature is now filled with reference priors for several standard and non-standard models, ready for use when objective Bayesian estimation is desired. The reference prior approach has often been found to have the virtue of giving good priors when the conventional choices fail, for example, due to the behavior of the likelihood in the tail. One case in point is in spatial modeling, see Berger et al. (2001). How this is achieved seems to be a mystery to me. In this paper too, for the multlinomial example using the hierarchical approach, the reference prior for the hyper parameter turned out to be a proper prior to compensate for the slow decay of likelihood in the tail. However, one runs into difficulty in implementing the reference prior approach when there are

*Main article DOI: 10.1214/14-BA915.
†University of Cincinnati, sivagas@ucmail.uc.edu

 DOI: 10.1214/14-BA935

more than one parameter of interest. Given a model, there are many reference priors; one prior for each parameter of interest, or even a set of priors for each parameter of interest based on different ordering of the rest of the parameters. These priors can be different for different parameters, requiring a user to switch priors depending which parameter(s) one is interested in estimating. This is not convenient to explain or appealing to use in practice, when one is interested in estimating more than one parameter and the corresponding reference priors are different for these parameters. Having a single objective prior for a given model, that works well for most natural parameters of interest is desirable. In this paper, the authors have taken up this important task and have given three possible approaches to get a single common "Overall Objective Prior" for simultaneously estimating several parameters of interest.

First, the authors set out to identify models for which there is a unique common reference prior for each of the natural parameters in the model under different orderings of the rest of the parameters. The authors give a condition on the the Fisher information matrix for such a single reference prior to exist, and provide examples which show that such a common reference prior can exist for the natural parameters of many different models.

The other two approaches provided in the paper constitute interesting novel ideas and developments, and include the Reference Distance approach and the Hierarchical Prior approach.

Hierarchical prior approach assumes a priori that the parameters of interest, $\theta_i$'s, conditionally on a hyper-parameter $a$, have a joint proper prior, leaving a prior for $a$ to be determined. When this conditional prior is in a convenient form in relation to the likelihood such as a conjugate prior so that the marginal likelihood for $a$ can be computed in closed form, one can obtain the reference prior for $a$, which is the Jeffreys prior based on the marginal likelihood. Then the overall objective prior for $\theta_i$'s is the marginal prior obtained by integrating the conditional prior for $\theta_i$'s with respect to the reference prior for $a$.

The reference distance approach is relatively more involved. Suppose that for each of the parameters of interest $\theta_i$, $i = 1, \ldots, n$, one can choose a reference prior. Then the reference distance approach first postulates a joint parametric family of priors for $(\theta_1, \ldots, \theta_n)$, not necessarily proper priors, indexed by a hyper-parameter $a$. Then the overall prior is that prior in the family whose marginal posterior distributions of $\theta_i$'s is closest on average, in terms of expected Kullback–Leibler distance, to the marginal reference posteriors of $\theta_i$'s.

The two approaches hold much promise in achieving the goal of finding overall objective priors for various models and parameters of interest. The hierarchical approach is particularly appealing, because the resulting prior itself is a reference prior, and it may also be relatively easy to derive, which can be a big advantage. However, the assumption of a convenient hierarchical or exchangeable structure for the joint prior of the parameters of interest is not always tenable. In comparison, the derivation of the reference distance approach requires computation of reference priors for each parameter of interest and a not-always-easy computation to find the optimal value of $a$, and the resulting overall objective prior is not necessarily a reference prior.

But, the reference distance approach holds an advantage – it seems in most cases one can write down a joint prior for the parameters of interest, indexed by a suitable hyper-parameter $a$ by inspecting the reference priors associated with each parameter. As always with the reference prior, once the hard work is done, it is readily available for use by everyone. It is a pleasant surprise that the reference distance approach for the normal model (Section 3.2.4) gives a prior that is the reference prior for the natural parameters. However, in general the reference distance approach may yield a prior that is different from any of the reference priors used in the derivation. Such an an overall objective prior may also turn out not to have good posterior behavior for some of the parameters of interest. In some instances, there may be more than one choice for the parametric class, each leading to different overall objective prior, and one has to make a determination which one to use. Would the authors comment on this and whether they have encountered such scenarios?

In light of these comments, the recommendation by the authors to use the common reference prior or the hierarchical approach first, and if not successful, to try the reference distance approach is noteworthy.

It is surprising that the reference prior for $a$ in the hierarchical approach to the multinomial example turns out to be a proper prior, making up for the behavior of the marginal likelihood being bounded away from 0 at infinity. As indicated before, the phenomenon that the reference prior distributes its mass selectively compensating for the the likelihood's slow decays in some tail regions is indeed amazing. Perhaps, the authors can give some general insight into this phenomenon. Both approaches have been illustrated for the multinomial example, yielding different overall objective priors. The reference distance approach sets $a = 1/m$, and the reference prior for $a$ in the hierarchical approach also seems to favor small values for $a$ for large $m$. However, for moderate values of $m$, the uncertainty in $a$ induced by the hierarchical prior approach would have an influence on the estimation of the parameters of interest, may be of an adaptive nature, unlike in the reference distance approach. Can the authors comment on this and how one may choose between the two choices?

While the hierarchical prior approach has its advantages, it appears that there may be more than one choice for the joint distribution for the parameters of interest, $\theta_i$'s, in terms of the second stage parameter $a$. In such cases, one would have to determine what would be the best choice. For example, in the multi-normal example in Section 4.3, one may alternatively use $\mu_i \overset{iid}{\sim} N(\mu_0, \tau^2)$ with known $\mu_0$, or $N(\mu, \tau_0^2)$ with known $\tau_0$, or more generally $N(\mu, \tau^2)$. In the case of $N(\mu_0, \tau^2)$, it appears that the resulting estimates for individual $\mu_i$'s would shrink towards $\mu_0$. Is there any particular justification for the choice of $\mu_0 = 0$?

## References

Berger, J. and Bernardo, J. M. (1989). "Estimating a product of means: Bayesian analysis with reference priors." *Journal of the American Statistical Association*, 84: 200–207. MR0999679. 223

— (1992). "On the development of reference priors." *Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press*, 35–60. MR1380266. 223

Berger, J., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics*, 37: 905–938. MR2502655. doi: http://dx.doi.org/10.1214/07-AOS587. 223

— (2012). "Objective priors for discrete parameter spaces." *Journal of the American Statistical Association*, 107: 636–648. MR2980073. doi: http://dx.doi.org/10.1080/01621459.2012.682538. 223

Berger, J., De Oliveira, V., and Sansó, B. (2001). "Bayesian Analysis of Spatially Correlated Data." *Journal of the American Statistical Association*, 96, 456: 1361–1374. MR1946582. doi: http://dx.doi.org/10.1198/016214501753382282. 223

Berger, J. and Sun, D. (2008). "Objective priors for a bivariate normal model with multivariate generalizations." *The Annals of Statistics*, 36: 963–982. MR2396821. doi: http://dx.doi.org/10.1214/07-AOS501. 223

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 41: 113–147 (with discussion). MR0547240. 223

Sun, D. and Berger, J. (1998). "Reference priors under partial information." *Biometrika*, 85(1): 55–71. MR1627242. doi: http://dx.doi.org/10.1093/biomet/85.1.55. 223

Bayesian Analysis (2015) **10**, Number 1, pp. 227–231

# Comment on Article by Berger, Bernardo, and Sun*

Manuel Mendoza† and Eduardo Gutiérrez-Peña‡

## 1 Introduction

The search of a prior distribution $p(\boldsymbol{\omega})$ to be used as part of an objective Bayesian analysis of a model $p(x|\boldsymbol{\omega})$ has proved to be a formidable endeavour. This is an area where we do not have a definitive answer yet, and any contribution to the understanding of the subject must be welcome. The authors of this paper are among the most prominent contributors to this field, and reading the manuscript has been very stimulating.

Research on the problem has mainly dealt with three issues: first, a definition of what a *non-informative, reference* or *objective* prior $p(\boldsymbol{\omega})$ must be; second, an operational algorithm to calculate such priors; third, the evaluation of the resulting prior(s) in accordance to certain criteria such as invariance, the avoidance of paradoxes, or desirable frequentist properties.

To us, and this is a subjective judgment, the most convincing approach to produce this sort of priors is reference analysis (Bernardo, 1979; Berger and Bernardo, 1992a,b; Bernardo, 2005; Berger et al., 2009). This procedure: (i) defines the reference prior as the prior maximizing the expected gain of information provided by a sample; (ii) includes a general (although potentially involved) algorithm to calculate the prior; and (iii) avoids a number of paradoxes. Moreover, it generalizes the Jeffreys prior and exhibits its limitations. Among its most remarkable results, it shows that the form of the reference prior $p(\boldsymbol{\omega})$ may depend on the function of the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ which is considered by the researcher to be of main interest.

Since its inception, the algorithm to obtain reference priors has evolved. This is the case specifically in the multiparameter setting. The most recent version (Berger and Bernardo, 1992a,b) requires all scalar components of the parameter to be *strictly* ordered in terms of their inferential interest. Thus, in principle, the current approach does not offer any solution if the researcher is simultaneously interested in two or more scalar parameters (or functions thereof). Interestingly, the original algorithm of Bernardo (1979) did cover this situation, although the solution was the multivariate Jeffreys prior which leads to unsettling paradoxes in some cases.

In this paper, the authors explore some ideas to extend the reference analysis to this yet unsolved case. They also seem to be considering a more general version of the problem by assuming that the number of scalar parameters (or functions of the parameters) of interest may be greater than the number of parameters in the model.

*Main article DOI: 10.1214/14-BA915.
†Dept. of Statistics, ITAM, Mexico, mendoza@itam.mx
‡Dept. of Probability and Statistics, UNAM, Mexico, eduardo@sigma.iimas.unam.mx

 DOI: 10.1214/14-BA936

So, the question is: What should the objective prior $\pi^R(\boldsymbol{\omega})$ ($\boldsymbol{\omega} \in \mathbb{R}^k$) be if there are $m$ functions $(\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega}))$ which are of simultaneous interest, where $m$ is not constrained to be less than or equal to $k$? Three methods to produce the required prior distribution are discussed: (i) the common reference prior; (ii) the reference distance approach; and (iii) the hierarchical approach.

## 2 Common reference prior

This is not really a method. If the reference priors corresponding to $\theta_i(\boldsymbol{\omega})$ as the parameter of interest ($i = 1, \ldots, m$) are the same for any ordering of the remaining parameters, then the posed problem simply vanishes. It is interesting to see some examples illustrating particular cases where the common prior exists, but it is desirable – and would be much more useful – to have general results characterizing sampling models where, for example, Theorem 2.1 applies and hence a common reference prior may be found. In this regard, results such as those in Gutiérrez-Peña and Rueda (2003) and Consonni et al. (2004) could provide a good starting point. These authors find reference priors for wide classes of exponential families that include the family discussed in Section 2.1.3 of the present paper as a particular case.

It must be pointed out that this section relies on the analysis of the information matrix $\boldsymbol{I}(\boldsymbol{\theta})$, so all reviewed scenarios assume $m \leq k$. Also, a somewhat disquieting result is that of Section 2.2.2, where the authors show that $\pi^R(\psi_1, \psi_2, \psi_3, \mu_1, \mu_2) \propto (\psi_1\psi_2)^{-1}$ is the one-at-a-time reference prior for any of these parameters and any possible ordering. In particular, it is the reference prior for the case where $\mu_2$ is the parameter of main interest. It so happens, however, that this prior is equivalent to the right-Haar prior which leads to a problematic posterior precisely for $\mu_2$. This result would imply that, in general, reference analysis might produce inadequate posteriors *for the parameter of interest*, depending on the specific accompanying parameters.

## 3 Reference distance method

In order to introduce this method, the authors explicitly assume that $\boldsymbol{\theta} = \boldsymbol{\omega}$, hence $m = k$. The idea is to find an overall prior $\pi(\boldsymbol{\theta})$ such that each of its marginal posteriors $\pi(\theta_i|\boldsymbol{x})$ is close to the corresponding marginal posterior $\pi_i(\theta_i|\boldsymbol{x})$ obtained when $\theta_i$ is the parameter of interest ($i = 1, \ldots, m$). As a measure of approximation the authors propose a weighted average of expected logarithmic divergences, although other measures could in principle be used. Also, the search for the overall prior is restricted to a specific parametric family $\mathcal{F} = \{\pi(\boldsymbol{\theta}|\boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$. Apart from the fact (acknowledged by the authors) that the existence of an optimal $\boldsymbol{a}$ is not guaranteed, a rather unappealing feature of this proposal is its dependence on the family $\mathcal{F}$. The authors offer no guidance on how to choose $\mathcal{F}$ *in general*. If the aim is to produce an objective approach, it seems desirable that $\mathcal{F}$ be somehow *intrinsic* to the sampling model. The examples in the paper suggest that perhaps this could be achieved through some kind of conjugacy.

Incidentally, the reference distance method bears some resemblance to the mean-field approach to variational inference, which is relatively straightforward in the case of

exponential families with conjugate priors; see, for example, Bishop (2006, Chapter 10). What is the authors' take on this?

We would like now to comment on Example 3.2.4. There, the normal model $N(x|\mu, \sigma)$ is considered, and the parameters of interest are $\mu$, $\sigma$ and $\phi = \mu/\sigma$. (Note that, despite the authors' remark at the beginning of Section 3, here $\boldsymbol{\theta} \neq \boldsymbol{\omega}$ and $m > k$.) In any case, the authors remind us that the reference prior when $\mu$ or $\sigma$ is the parameter of interest is $\pi(\mu, \sigma) = \sigma^{-1}$, whereas the reference prior for $\phi = \mu/\sigma$ is given by $\pi_\phi(\mu, \sigma) = (2\sigma^2 + \mu^2)^{-1/2}\sigma^{-1}$. They then propose, as a "natural" choice, the class of relatively invariant priors $\mathcal{F} = \{\pi(\mu, \sigma) = \sigma^{-a}; a > 0\}$. For this family, they show that the overall prior for $(\mu, \sigma, \phi)$ can be approximated by $\pi^o(\mu, \sigma) = \sigma^{-1}$, so that inclusion of $\phi$ as an additional parameter of interest makes no difference. We find this rather disappointing. From an algorithmic point of view, this outcome is not surprising given the choice of $\mathcal{F}$ and the form of the reference priors for $\mu$ and $\sigma$. Only a large weight on the divergence corresponding to $\phi$ could lead to a different result. An idea that springs to mind is to try another (arguably more "natural" family) such as $\mathcal{G} = \{\pi(\mu, \sigma|a_1, a_2) = (2\sigma^2 + \mu^2)^{-a_1}\sigma^{-a_2}; a_1 > 0, a_2 > 0\}$, which includes all three reference priors for $\mu$, $\sigma$ and $\phi$. On the other hand, since $\pi_\mu(\mu, \sigma)$ and $\pi_\sigma(\mu, \sigma)$ are equal in this case, the authors could alternatively have minimized the sum of the two divergences corresponding to the marginal posterior of $\phi$ and the *joint* posterior of $(\mu, \sigma)$). We wonder how these alternative ideas compare with that proposed in the paper for this example.

## 4 Hierarchical approach

The idea of this approach is, first, to find a "natural" parametric family of proper priors $\pi(\boldsymbol{\theta}|a)$ such that $a \in \mathbb{R}$ and the integrated likelihood results in a proper density $p(x|a)$. Then, the univariate reference prior for $a$, $\pi^R(a)$, is obtained for this latter model. Finally, the overall prior $\pi^o(\boldsymbol{\theta})$ is defined as the expectation of $\pi(\boldsymbol{\theta}|a)$ with respect to $\pi^R(a)$. This is an intuitive and seemingly reasonable idea. However, it is not clear how to make explicit that $\boldsymbol{\theta}$ is the parameter of interest even though the model is originally indexed by $\boldsymbol{\omega}$, especially when the dimension of $\boldsymbol{\theta}$ is larger than that of $\boldsymbol{\omega}$. (See the comment below concerning the multi-normal means example.) We wonder if the authors can provide some advice on how this could be achieved in general. On the other hand, as in the reference distance case, dependence upon a specific family of priors introduces no small amount of arbitrariness in the method. Here, again, a proper objective method would use an *intrinsic* family entirely determined by the sampling model. One possibility, particularly suitable for the case of hierarchical models, would be to elaborate on the idea of conjugate likelihood distributions (George et al., 1993), although a suitable restriction should be imposed on the corresponding conjugate family in order to get a one-dimensional hyperparameter. Concerning the implementation of the method, the authors suggest that integration to get the overall prior can be avoided by using $\pi^o(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\hat{a})$ instead, where $\hat{a}$ is the mode of the posterior $p(a|\boldsymbol{x})$. This proposal may be efficient from a computational point of view, but it is both surprising and disappointing since it essentially reduces the hierarchical approach to a standard empirical Bayes procedure and leads to a data-dependent prior.

The example in Section 4.2 concerning the multivariate hypergeometric model is confusing and does not quite illustrate the method described above. First, the parameters of the sampling model are given a multinomial prior (which does not depend on a single scalar parameter $a$, but on a vector of probabilities $\boldsymbol{p}_k$); then, the likelihood is integrated and shown to yield a multinomial distribution. In the process, the $k$ original parameters $R_1, R_2, \ldots, R_k$ are replaced by the parameters $p_1, p_2, \ldots, p_k$, so the idea of reducing the problem to the determination of the reference prior for a scalar parameter is abandoned. Next, in the multinomial model, the approximate overall prior obtained *using the reference distance method* is adopted for the hyperparameters $\boldsymbol{p}_k$. Finally, the corresponding integrated Multinomial–Dirichlet distribution is declared as the overall prior for $R_1, R_2, \ldots, R_k$. We find this *ad hoc* combination of methods difficult to justify as a general procedure.

An alternative formulation could be based on the idea of super-populations (quite common in the field of survey sampling) as follows. Let us assume that a random sample of size $N$ is obtained from a multinomial distribution $Mu_k(\boldsymbol{Y}_k|1, \boldsymbol{p}_k)$. As a result we get a vector $R_1, R_2, \ldots, R_k$ describing the number of sampled units in each category. Now imagine that we then get a subsample of size $n$, *without replacement*, from the sample of size $N$. In this setting, the multinomial distribution describes an infinite super-population, the sample of size $N$ is the finite population of interest and the subsample of size $n$ is the actual sample we observe. Given the sample, the likelihood based on the subsample corresponds to that of a hypergeometric distribution. However, with respect to the super-population, the subsample is just a sample of the original multinomial population whose parameters are given by the vector $\boldsymbol{p}_k$. Within this framework, $R_1, R_2, \ldots, R_k$ are observables and any inference regarding these quantities must be produced through the corresponding posterior predictive distribution. This argument shows that the hypergeometric problem can be viewed as a multinomial one where the interest is not really on the parameters but on observables, and the relevant overall prior is that for $\boldsymbol{p}_k$, no matter which method we use.

The example on the multi-normal means (Section 4.3) deserves a few words as well. Here, the parameters of interest are, using the same notation as the authors, $\mu_i$; $i = 1, \ldots, m$ and $|\boldsymbol{\mu}|^2 = \mu_1^2 + \cdots + \mu_m^2$. First, we note that throughout the paper $k$ refers to the dimension of $\boldsymbol{\omega}$ and $m$ is the number of parameters of interest (the dimension of $\boldsymbol{\theta}$), so in this example we have $m = k + 1$. It must be pointed out, however, that the hierarchical method, as defined, cannot be applied when $m > k$ since the distribution $\pi(\boldsymbol{\theta}|a)$ would then be defined over a space of functionally related components of $\boldsymbol{\theta}$ and would be singular. This fact is implicitly recognized by the authors when they propose a prior for $(\mu_1, \ldots, \mu_m)$ only, ignoring the last parameter of interest, $|\boldsymbol{\mu}|^2$. They then argue that the resulting overall prior is reasonable not only for each mean $\mu_i$ but also for $|\boldsymbol{\mu}|^2$. The key issue here is the convenient choice of $\pi(\boldsymbol{\mu}|a)$ as the product of the normals $N(\mu_i|0, a)$. So, strictly speaking, this problem is not actually solved by using the hierarchical approach as proposed in the paper but by an *ad hoc* choice of $\pi(\boldsymbol{\mu}|a)$.

## 5 Final remarks

This paper contains many interesting ideas and examples. However, it offers more of a brainstorming than a systematic treatment and a general solution to the problem. It is somewhat disappointing that the methods proposed in the paper bear little resemblance with the original reference prior approach, where the problem is clearly stated, the criterion used is sensible, and one can typically obtain *unique* and reasonable solutions. The approaches proposed here are still far from becoming operational algorithms since they require a number of arbitrary inputs. Hopefully, at least one of these methods will evolve into an *overall objective procedure* to find overall objective priors. We believe the reference distance method to be the most promising in this regard.

## References

Berger, J. O. and Bernardo, J. M. (1992a). "On the development of reference priors." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 35–60. Oxford: University Press (with discussion). MR1380269. 227

Berger, J. O. and Bernardo, J. M. (1992b). "Ordered group reference priors, with applications to multinomial problems." *Biometrika*, 79: 25–37. MR1158515. doi: http://dx.doi.org/10.1093/biomet/79.1.25. 227

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *Annals of Statistics*, 37: 905–938. MR2502655. doi: http://dx.doi.org/10.1214/07-AOS587. 227

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society, Series B*, 41: 113–147 (with discussion). MR0547240. 227

Bernardo, J. M. (2005). "Reference analysis." In Dey, D. K. and Rao, C. R. (eds.), *Bayesian Thinking: Modeling and Computation*, *Handbook of Statistics 25*, 17–90. Amsterdam: Elsevier. MR2490522. doi: http://dx.doi.org/10.1016/S0169-7161(05)25002-2. 227

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. MR2247587. doi: http://dx.doi.org/10.1007/978-0-387-45528-0. 229

Consonni, G., Veronese, P., and Gutiérrez-Peña, E. (2004). "Order-invariant group reference priors for natural exponential families having a simple quadratic variance function." *Journal of Multivariate Analysis*, 88: 335–364. MR2025617. doi: http://dx.doi.org/10.1016/S0047-259X(03)00095-2. 228

George, E. I., Makov, U. E., and Smith, A. F. M. (1993). "Conjugate Likelihood Distributions." *Scandinavian Journal of Statistics*, 20: 147–156. MR1229290. 229

Gutiérrez-Peña, E. and Rueda, R. (2003). "Reference priors for exponential families." *Journal of Statistical Planning and Inference*, 110: 35–54. MR1944632. doi: http://dx.doi.org/10.1016/S0378-3758(01)00281-6. 228

Bayesian Analysis (2015) **10**, Number 1, pp. 233–236

# Comment on Article by Berger, Bernardo, and Sun*

Judith Rousseau†

In this paper, the authors undertake to expose an encompassing principle to handle objective priors in competition, their difficulties, their contemners, and their multiplicity! Great target, for which we congratulate them. However, it may be a doomed attempt if they mean to achieve the ultimate reference prior, since this quest has been going on for centuries, including the contributions of the French Polytechnicians Émile Lhoste and Maurice Dumas in the 1920s (Broemeling and Broemeling, 2003), with no indication that we are near reaching an agreement. The authors thus aim for a less ambitious construction.

Let us point out why we think this is an important problem. That we would have to change priors by changing parameters of interest is disturbing and somehow goes against the use of Bayesian methodologies. Ideally, one would want a single prior and various loss functions. Interestingly, this difficulty associated to the construction of noninformative priors – in the sense that it needs to be targeted on the parameter of interest – is amplified in large or infinite dimensional models. In finite dimensional regular models, the prior has an impact – at least asymptotically – to second order only. In infinite dimensional models, the influence of the prior does not completely vanish asymptotically, although some aspects of the prior may have influence only to second order. It has been noted recently that in a nonparametric problem, such as density or regression function estimation, nonparametric prior models may lead to well behaved posterior distributions under global loss functions such as the Hellinger distance for the density or the $L_2$-norm for the regression function while have pathological behaviour for some specific functionals of the parameter; see, for instance, (Rivoirard and Rousseau, 2012; Castillo, 2012; Castillo and Rousseau, 2013). This means that one needs to target the prior to specific parameters of interest, or that somehow it is asking too much of a prior to be able to give satisfactory answers for every aspects of the parameter. The larger the model, the more crucial the problem.

Obviously, it is of interest to derive priors which are *well behaved* for a large range of parameters of interest. The problem is then to define what well behaved means. This does not seem to be really defined in the present paper. Is it possible to derive a general notion of *well behaved* in the case of multiple parameters of interest without referring to a specific task or, in other words, to a specific loss function or family of loss functions?

The authors consider three possibilities: (1) a common reference prior existing for various parameters of interest which then should be used, (2) choosing the prior belonging to some parametric family of priors closest to the set of reference priors associated

*Main article DOI: 10.1214/14-BA915.
†Université Paris-Dauphine, CEREMADE and CREST – ENSAE, rousseau@ceremade.dauphine.fr

 DOI: 10.1214/14-BA937

to the various parameters of interest, (3) using a hierarchical model based on a parametric family of the prior where the hyperparameter is itself given a reference prior. The authors consider a series of examples and discuss the merits of the various approaches on each of these examples.

With regards to (1), the authors propose conditions such that marginal references are common for various parameters of interest; it is interesting but once again challenging. First, it implies that there are not more parameters of interest than there are parameters in the model, and second, even in that case it does not always exist. However, given that *all models are wrong but some are useful*, would that indicate that we should change the point of view entirely and, given a set of parameters of interest, define a model which would allow for *good* (whatever that means) inference on them; for instance, that would lead to a common reference prior for all of them? In particular, in this respect, how do reference priors behave under model misspecification?

Given the limitations of the first case, the authors propose to relax the notion of reference priors in methods (2) and (3).

We believe that the distance approach is a very interesting idea to obtain a global consensus between the different reference priors, however, there are a number of issues that they raise.

## 1 Some issues with the distance approach

One of the advantages of the idea behind the distance approach is that it can deal with more parameters of interest than the actual dimension of the parameter and leads to tractable posterior distributions. One of its disadvantages is that it depends on the sample size.

• **Dependence on the sample size** The construction of the reference priors is based on a limiting argument, assuming that infinite information (infinite sample size) is available. Why cannot we use the same perspective here? For instance, in the case of regular models using the Laplace approximation to second order, the integrated Kullback–Leibler divergence between $\pi_{\theta_i}(\cdot|\mathbf{x})$ and $\pi_a(\cdot|\mathbf{x})$ (or the directed logarithmic divergence from $\pi_a(\cdot|\mathbf{x})$ to $\pi_{\theta_i}(\cdot|\mathbf{x})$ as termed in the paper) is approximately

$$K_i = \frac{1}{n} \int (\nabla \log \pi_{\theta_i} - \nabla \log \pi_a)^t \, I^{-1}(\theta) \, (\nabla \log \pi_{\theta_i} - \nabla \log \pi_a) \, \pi_{\theta_i}(\theta) d\theta$$

where $b_3(\theta)$ corresponds to the third order derivative of the log-likelihood and $I$ is the Fisher information matrix. Hence asymptotically minimizing the sums of the distances corresponds to minimizing

$$\sum_i w_i \int (\nabla \log \pi_{\theta_i} - \nabla \log \pi_a)^t \, I^{-1}(\theta) \, (\nabla \log \pi_{\theta_i} - \nabla \log \pi_a) \, \pi_{\theta_i}(\theta) d\theta.$$

• **An alternative idea with the same flavour** On a general basis, and following Simpson et al. (2014), the choice of minimising a distance in (2) could be replaced in a

more Bayesian manner by a prior on the distance as, e.g.

$$\pi(a) = \exp\left\{-\sum_i \lambda_i d_i(a)\right\}$$

where $d_i(a)$ is derived as in the paper. This offers several advantages from dealing with partial information settings to defining a baseline model.

In addition, a neophyte reader could also ask what is so essential with reference priors that one has to seek recovering them at the marginal level.

## 2 On the hierarchical approach

Both the hierarchical and the distance approaches have been considered in the paper with univariate hyperparameters. It is not clear if, in the case of the distance approach, this is a key issue, but it certainly is in the hierarchical construction since a reference prior needs to be constructed on this hyperparameter. This restricts the flexibility of the prior.

In the immense variety of encompassing models where recovering the reference marginals is the goal, what about copulas?! There are many varieties of copulas and a prior could be set on any of those, with once again non-informative features.

Finally, although the authors have considered examples renown to be difficult for constructing objective priors, such as the multinomial model, they do not cover the more realistic framework of complex and partly-defined sampling models. In Simpson et al. (2014), the authors advocate the construction of priors within sub-models of a more complex model, without taking into account the larger model. This contradicts the nature of the reference prior, at the same time these sub-models might be the only ones where the reference prior construction may be feasible. Would the ideas considered by the authors here be useful in combining the local construction (within a sub-model) of the reference prior with the larger model?

Once again, I would like to thank the authors for a thought-provoking paper on an important issue.

## References

Broemeling, L. and Broemeling, A. (2003). "Studies in the history of probability and statistics XLVIII The Bayesian contributions of Ernest Lhoste." *Biometrika*, 90(3): 728–731. MR2006848. doi: http://dx.doi.org/10.1093/biomet/90.3.728. 233

Castillo, I. (2012). "Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors." *Sankhya A*, 74(2): 194–221. MR3021557. doi: http://dx.doi.org/10.1007/s13171-012-0008-6. 233

Castillo, I. and Rousseau, J. (2013). "A General Bernstein–von Mises Theorem in semiparametric models." Technical report. 233

Rivoirard, V. and Rousseau, J. (2012). "Bernstein–von Mises theorem for linear functionals of the density." *The Annals of Statistics*, 40: 1489–1523. MR3015033. doi: http://dx.doi.org/10.1214/12-AOS1004. 233

Simpson, D. P., Martins, T. G., Riebler, A., Fuglstad, G.-A., Rue, H., and Sørbye, S. H. (2014). "Penalising model component complexity: A principled, practical approach to constructing priors." arXiv:1403.4630v3. MR3277029. 234, 235

**Acknowledgments**

The author wishes to thank Christian Robert for fruitful discussions.

Bayesian Analysis (2015) **10**, Number 1, pp. 237–241

# Comment on Article by Berger, Bernardo, and Sun*

Gauri Sankar Datta† and Brunero Liseo‡

It is our distinct pleasure to comment on a very thought provoking paper, and we first congratulate the Authors for this new masterly contribution in the field of objective priors.

The main goal of the paper is to find a multi-purpose objective prior for a model that should be used by different researchers with varying goals, with the consequence that no single parameter or parametric function can be identified as a parameter of interest. In this situation, the most popular approaches either fail or, as in the case of the reference prior algorithm, they cannot be used.

Three general methods are discussed by the Authors. The first one is limited to a number of particular situations where the reference prior is the same for all quantities of interest: this case is not of much concern since a natural solution exists. The second method is based on the reference prior approach: one looks for the prior which produces the marginal posteriors for the quantities of interest which are closer – in some sense – to the marginal reference posteriors. Whereas this method is perfectly reasonable, the final result will depend on the particular set of the quantities of interest considered and it cannot be considered as the "overall" objective prior. The third method is based on a hierarchical representation of the model, when it is available. It shifts the problem of determining an objective prior to an upper level of the hierarchy, where the impact of the prior might be less serious.

We believe that the latter method is superior to the others because

- it is compatible with a predictive approach where all the parameters are nuisance parameters and there is no particular quantity of interest; however, one should be careful here: if the quantity of interest is, for example, the posterior predictive mean
$$E(X_{n+1} \mid X_1, \ldots, X_n)$$
of a future observation – and not the entire predictive density – then a parameter of interest actually does exist!
- it is clearly superior to Method 2, especially when the model is used repeatedly by different people which are interested in different sets of parameters.

In terms of prediction, it would be worth discussing the proposal of Datta et al. (2000).

*Main article DOI: 10.1214/14-BA915.
†University of Georgia, Athens, GA (USA), gaurisdatta@gmail.com
‡Sapienza Università di Roma, Italy, brunero.liseo@uniroma1.it

 DOI: 10.1214/14-BA938

In this contribution, we will briefly consider the multinomial example, and provide some comments on the concept of prior averaging.

## 1 The multinomial model in the sparse case

This is a very interesting problem. Jeffreys' prior allocates a weight of 1/2 to each original component of the vector $(\theta_1, \theta_2, \ldots, \theta_m)$. This is too much when $m$ is large compared to the sample size $n$ and the distribution is very sparse. This suggests that the prior mass should be adequately spread on the parameter space in such a way that each cell has a negligible prior mean, especially when compared with the weight of the data.

In the multinomial case, the prior weight (expressed as the sum of the hyper-parameters of the Dirichlet prior) is equal to $m/2$ for the Jeffreys' prior, while in the hierarchical approach, arising from a Dirichlet$(a, a, \ldots, a)$ hyper-prior, it is a random quantity $v = ma$ with density given by expression (25) of the paper, at least in the case of an infinite $m$. Several numerical computations, with different values of $n$ and $r_0$ (i.e., the number of non-empty cells), show that the mode and the median of $v$ are rarely larger than 2, so the hierarchical approach automatically accounts for the sparsity and the corresponding marginal posteriors are dramatically different from those arising from the use of Jeffreys' prior.

There are many ways in which this problem can be handled. If we transform it to a multiple testing problem, that is, for each cell $i$ we test

$$H_0 : \theta_i = 0 \qquad \text{vs.} \qquad H_1 : \theta_i \neq 0,$$

the problem can be rephrased as that of finding an ad-hoc prior, just like in the sparse normal problem, which is well studied in literature, see, for example, Scott and Berger (2010). The two problems are similar but not identical: here we do not necessarily observe data for each cell, and the difficulties associated with this discrete version of the problem are even greater since the values of the $\theta_i$'s will affect the standard deviation of the cells, not only the means.

From a testing perspective there is also another interesting connection: the Authors propose to add – as a prior weight – something close to $1/m$ to each cell. So the total weight of the prior will be approximately one. This reminds us of the unit prior information of Kass and Wasserman (1996).

The sparse multinomial case is also of theoretical interest because it represents a bridge between parametric and non-parametric models, when the number of cells goes to infinity.

Our personal view of the example is close to that of the Authors, although it is not of great surprise that the Jeffreys' prior does not clearly discriminate between observed and non-observed cells, when $n$ is so small compared to $m$. In other words, this is too much to ask of the prior. When $n$ is as small as 3, and the number of parameters is about 1000, it is hopeless to find a good automatic objective prior and some external

guidance (in this case, the choice of a "proper" prior within the Dirichlet class) seems unavoidable.

More interesting is the fact that the hierarchical prior depends on $m$ and $n$ only through their ratio: this is actually what one would expect.

We have also considered a variant of the multinomial example. In particular, we have considered the case when the multinomial likelihood can be rephrased as one arising from a sample of $m$ independent Poisson random variables with mean vector $(\psi_1, \ldots, \psi_m)$ and then setting $\theta_j = \psi_j / \sum_i \psi_i$. Doing the usual reference prior calculations here, we ended up with the same conclusions as if we have used the standard Jeffreys' Beta prior $(1/2, 1/2)$ for the $\theta_i$'s. We wonder how to get the same result (weights $\approx m^{-1}$ for the cells) in this alternative perspective. It is very likely that this can be obtained by assuming independent gamma priors with shape parameter $a$ and scale parameter $\beta$ for the $\psi_j$. If the "nuisance" scale parameter $\beta$ is eliminated by conditioning on the total counts, we end up with the same conclusion. However, the rationale behind this last choice is – again – only pragmatic.

A related issue is the ordered multinomial example in Section 2.1.2. Here the overall prior for any of the parameters $(\xi_1, \ldots, \xi_m)$ is the product of independent Beta$(1/2, 1/2)$: what happens for large $m$? Is the overall prior still a sensible prior or should we take into account this problem?

## 2 A comment on geometric average of priors

Consider the following divergence function

$$d(\eta) = \sum_{i=1}^{m} \alpha_i \int \eta(\theta) \log \frac{\eta(\theta)}{\pi_i(\theta)} d\theta,$$

where $\alpha_1, \ldots, \alpha_m \geq 0$ are suitable constants adding to 1, and $\pi_i(\theta)$ may be a suitable objective prior when one is interested in one of a given set of $m$ parametric functions. The above function is a weighted average Kullback–Leibler divergence between a global prior and the marginal priors we would like to use in the case we were interested in a single parametric function $t_i(\theta)$, $i = 1, \ldots, m$. Note that

$$\begin{aligned} d(\eta) &= \int \eta(\theta) \log \eta(\theta) d\theta - \sum_{i=1}^{m} \int \eta(\theta) \log \pi_i^{\alpha_i}(\theta) d\theta \\ &= \int \eta(\theta) \log \frac{\eta(\theta)}{\prod_{i=1}^{m} \pi_i^{\alpha_i}(\theta)} d\theta. \end{aligned}$$

By Jensen's inequality, $d(\eta)$ will be minimized with respect to $\eta$ if $\eta(\theta) / \prod_{i=1}^{m} \pi_i^{\alpha_i}(\theta)$ is a degenerate function. This leads to the geometric mean prior

$$\pi_G(\theta) \propto \prod_{i=1}^{m} \pi_i^{\alpha_i}(\theta).$$

Usually, the component priors $\pi_i(\theta)$'s are improper, which in turn may also make $\pi_G(\theta)$ an improper prior. The authors indicated that the geometric mean prior is preferable to the arithmetic mean prior since one or more of the component priors $\pi_i$ may be improper, and the arithmetic mean posterior may be highly influenced by one or a few component posteriors. Indeed, for any arbitrary positive constant $c_i$, $c_i\pi_i(\theta)$ is as much an objective prior as $\pi_i(\theta)$ is. While the posterior propriety of the arithmetic mean prior is an immediate consequence of the propriety of the component posteriors, the same is not so obvious for the geometric mean prior. However, the following lemma shows that the posterior corresponding to $\pi_G(\theta)$ will be proper provided that each component prior $\pi_i(\theta)$ generates a proper posterior.

**Lemma 1.** *For two prior densities $\mu(\theta)$ and $\nu(\theta)$, if*

$$\int \mu(\theta)L(\theta;\mathbf{x})d\theta < \infty, \quad \text{and} \quad \int \nu(\theta)L(\theta;\mathbf{x})d\theta < \infty,$$

*then, for any $\alpha \in (0,1)$,*

$$\int \mu^{\alpha}(\theta)\nu^{1-\alpha}(\theta)L(\theta;\mathbf{x})d\theta < \infty,$$

*where $L(\theta;\mathbf{x})$ denotes the joint density of data $\mathbf{x}$ corresponding to the parameter value $\theta$.*

*Proof.* By Hölder's inequality, it follows that

$$\begin{aligned}\int \mu^{\alpha}(\theta)\nu^{1-\alpha}(\theta)L(\theta;\mathbf{x})d\theta &= \int [\mu(\theta)L(\theta;\mathbf{x})]^{\alpha}\,[\nu(\theta)L(\theta;\mathbf{x})]^{1-\alpha}\,d\theta \\ &\le \left[\int \mu(\theta)L(\theta;\mathbf{x})d\theta\right]^{\alpha}\left[\int \nu(\theta)L(\theta;\mathbf{x})d\theta\right]^{1-\alpha}.\end{aligned}$$

Thus $\mu(\theta)^{\alpha}\nu(\theta)^{1-\alpha}$ generates a proper posterior density for the given data $\mathbf{x}$. □

By repeated use of this lemma, the propriety of the posterior based on the geometric prior $\pi_G(\theta)$ easily follows.

## 3 An anecdote

While preparing the present comments one of the authors attended a seminar on applied probability where the following situation was presented. In a small village, there is a chief and several shepherds. Each shepherd runs a flock of sheeps. The chief knows that the ground of their village is going to become parched so the shepherds have to move away. All the roads starting from the village – but one – are full of hungry wolves. The chief has his own probability distribution about which is the safe road. If the chief communicates his/her information to the shepherds, it is very likely that all of them would choose the same road. This implies that either all the sheeps or none will survive. If the chief does not communicate his/her information, it is likely that the shepherds will randomly choose the road.

The question is: should the chief share this information with the shepherds or not? If so, (s)he is playing a risky (all or nothing) strategy. If not, (s)he is taking a minimax strategy where it is more likely that some of the flocks will survive. Is there a way to calibrate the amount of information to be shared?

There are several interesting similarities between this story and the main issue of the paper. Is there a way to find a compromise between the general goal and a single objective? Is it possible to find a prior – or a strategy – which is not so bad for any of the problems at hand?

Our view is that, if the answer is "yes", this prior should not depend on the particular list of problems. In other words, it would be great to have just "one" overall prior. In this respect, the hierarchical approach seems to be more promising.

## References

Datta, G., Mukerjee, R., Ghosh, M., and Sweeting, T. (2000). "Bayesian Prediction with Approximate Frequentist Validity." *The Annals of Statistics*, 28: 1414–1426. MR1805790. doi: http://dx.doi.org/10.1214/aos/1015957400. 237

Kass, R. and Wasserman, L. (1996). "The selection of prior distribution via formal rules." *Journal of the American Statistical Association*, 91: 1343–1370. MR1478684. doi: http://dx.doi.org/10.1214/lnms/1215453065. 238

Scott, J. and Berger, J. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: http://dx.doi.org/10.1214/10-AOS792. 238

Bayesian Analysis (2015) **10**, Number 1, pp. 243–246

# Rejoinder*

James O. Berger[†], Jose M. Bernardo[‡], and Dongchu Sun[§]

Our thanks to all the discussants for their insightful observations and comments. We respond to their discussions in turn.

## 1 Response to Datta and Liseo

We agree that Method 3 is preferable to Method 2, in that it is not dependent on the specification of a collection of quantities of interest and, hence, need only be determined once (and not separately for each potential user of a model). It is because a hierarchical embedding is not always available that we introduce the other methods as possible solutions.

We found the discussion of the multinomial example interesting, with numerous additional insights being provided. Likewise the additional material on the geometric averaging approach was enlightening, especially the nice lemma showing that, if a collection of priors all yield proper posteriors, then their geometric average also yields a proper posterior. This certainly strengthens the argument that geometric averaging is superior to arithmetic averaging in the search for an overall prior.

The moral of the amusing anecdote is indeed sound, and can be attempted to be implemented even when there is no hierarchical embedding available. For instance, Berger and Sun (2008) considered 21 different derived parameters for the five-parameter bivariate normal distribution, seeking a prior that was good 'on average' for the 21 parameters.

## 2 Response to Mendoza and Gutiérrez-Peña

The discussants highlight the importance of cataloguing those situations in which there is a common reference prior for all the parameters of a model and give useful references that could be a starting point for identifying additional such situations. But they then, interestingly, question whether this is sufficient, especially when the number, $m$, of quantities of interest exceeds the number, $k$, of parameters in the model.

Section 3.1 highlights one such situation: there is a common reference prior for $\mu$ and $\sigma$ from the normal model but this cannot necessarily be claimed to be the overall objective prior because the reference prior for $\mu/\sigma$ is different. This simple example suggests that one can probably never have an overall objective prior that is optimal for everything and that just having it be reasonable for everything (of interest) might be the

*Main article DOI: 10.1214/14-BA915.
[†]Duke University, USA and King Abdulaziz University, Saudi Arabia, berger@stat.duke.edu
[‡]Universitat de València, Spain, jose.m.bernardo@uv.es
[§]University of Missouri-Columbia, USA, sund@missouri.edu

 DOI: 10.1214/15-BA943

best we can hope for. It is interesting, in this regard, that Section 3.2.4 shows that the common reference prior for $\mu$ and $\sigma$ is also nearly optimal when $\mu/\sigma$ is also considered, although the discussants are correct that this result was probably inevitable once we restricted the candidate priors to be only of the form $\sigma^{-a}$. Utilizing the alternative and more general class that they suggest might well have given a different result (but the computation would have been much more formidable).

It would be nice if one could show, in general, that, if there is a common reference prior for all of original parameters, then that prior will be reasonable for other derived parameters or quantities. Our experience strongly supports this claim, but it is difficult to see how to formally approach verification of the claim especially, as the discussants note, because of the disquieting result in Section 2.2.2.

We agree that it would be nice if the family of candidate priors considered in both the reference distance method and the hierarchical method could somehow be intrinsically identified from the model itself; this would make the label 'objective' more compelling. We have not tried to do so ourselves, but the discussants give several potentially useful starting points for such an endeavor. Computational considerations are central here so, as the discussants note, we always chose the candidate class of priors to be a conjugate class (or as close to conjugate as possible).

Thanks for pointing out the possible relationship of the reference distance method with the mean field approach to variational inference. The approximation tools being used in each case are clearly related, but it is not clear to us that this can be usefully exploited.

We agree with the discussants concerning Section 4.2. It is hard to know how to deal with the hypergeometric parameters directly, so we used the common technique of 'transferring' them into uncertain multinomial parameters that we can deal with. But this is, indeed, a somewhat *ad hoc* addition to the proposed methodology. In this light, the suggested reformulation of the discussants (which ends up in the same place) will be a more appealing justification to many.

## 3 Response to Rousseau

Rousseau makes the important observation that we are considering the 'simple' parametric case, where there is some hope of having an overall objective prior that is at least reasonable for likely quantities of interest. This hope could well be impossible in nonparametric situations, where it can be a challenge to even find a prior that is satisfactory for a single given quantity of interest.

Rousseau observes that maybe the search for an overall prior should be considered together with choice of the model. This is an intriguing idea, but we have no idea how to approach the issue.

Rousseau observes that, for the reference distance method, the solution depends on the sample size. It is not appealing, in general, to have objective priors depend on the sample size, but there are situations (hierarchical models) where it seems correct

and inevitable. Here, however, the numerical evidence in the examples indicates that there is only a very slight dependence on the sample size, so Rousseau observes that one can simply try to implement the approach asymptotically, avoiding the sample size dependence and – more importantly – perhaps considerably simplifying the derivation of the overall prior. This is an idea definitely worth pursuing!

In her final comments, Rousseau addresses scenarios considerably more complex than any we consider, and outlines issues in finding good (objective) priors for those scenarios. In our own statistical practice, we encounter these problems all the time. There is little or no theory to guide us, so it is perhaps most useful to simply say what we do. A complex model is usually made up of simpler subcomponents, and we may well know a good overall objective prior for a subcomponent. We will use it, even though there is no assurance that it is a good overall objective prior in the context of the full model; the alternative of using a prior that we know is suboptimal for the component does not appeal.

This is the but the tip of an iceberg, however, in that many complex models are hierarchical in nature, and it is well known that standard objective priors for a model can be terrible if that model appears at a higher level in a hierarchy. See Berger, Strawderman, and Tang (2005) for discussion of this.

## 4 Response to Sivaganesan

We enjoyed Sivaganesan's comment that "How ... [reference priors]... work seems to be a mystery ...," because it is also a mystery to us. But their consistently astonishing properties explains why the approaches we suggest for developing an overall prior all center around some application of reference prior theory.

Sivaganesan points out that the choice of the candidate priors will surely affect the answers, and asks if we have tried alternative classes of candidate priors. He is certainly right that the class will likely have some effect, but our experience with Bayesian robustness in other contexts suggests that the class may not be that important when, as here, we are optimizing over the class. But this an important topic for future study.

We appreciated Sivaganesan's comment that "It is surprising that the reference prior for $a$ in the hierarchical approach to the multinomial example turns out to be a proper prior, making up for the behavior of the marginal [likelihood] being bounded away from 0 at infinity." The history was that we first – and to our surprise – discovered that the reference prior for $a$ was proper; we then went back to look at the likelihood, and discovered that, indeed, it was not integrable at infinity, 'explaining' why the reference prior decided to be proper. (This is part of the mystery of reference priors referred to above.)

## 5 Closing comments

Mendoza and Gutiérrez-Peña comment about the paper "... it offers more of a brainstorming than a systematic treatment and a general solution to the problem [of obtaining an overall objective prior]." We couldn't agree more. We have been working on this for

more years than we care to reveal and finally admitted to ourselves that we were not going to find the general solution to the problem. So the paper is simply a reflection of what we encountered in attempting to find a general solution.

Sivaganesan asks us to comment on which of the three approaches to an overall objective prior we would recommend. Details are given in the final section of the paper, but it is useful to highlight the main points (with the caveat of the comment above):

- If all (natural) parameters of the model have the same reference prior, use it as the overall objective prior.
- If one can find a natural and computationally feasible hierarchical structure for the model parameters, use that, along with finding the reference prior for the parameters in the hierarchical structure.
- If the above are not implementable,
  - Try the reference distance approach; the suggestion of Rousseau to do so asymptotically is perhaps the first thing to try here.
  - Try the geometric average of parameter reference priors, supported by the results of Liseo and Datta.

## References

Berger, J. O. and Sun, D. (2008). "Objective priors for the bivariate normal model." *The Annals of Statistics*, 36: 963–982. MR2396821. doi: http://dx.doi.org/10.1214/07-AOS501. 243

Berger, J., Strawderman, J., and Tang, D. (2005). "Posterior propriety and admissibility of hyperpriors in normal hierarchical models." *The Annals of Statistics*, 33: 606–646. MR2163154. doi: http://dx.doi.org/10.1214/009053605000000075. 245