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 $p(\boldsymbol{x} \mid \boldsymbol{\omega})$, 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 θ; the posterior distribution, $\pi_\theta(\boldsymbol{\omega} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\omega})\, \pi_\theta(\boldsymbol{\omega})$, thus also depends on θ, and inference for θ is performed using the corresponding marginal reference posterior for $\theta(\boldsymbol{\omega})$, denoted $\pi_\theta(\theta \mid \boldsymbol{x})$. 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

© 2015 International Society for Bayesian Analysis 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 $\mathrm{Mu}(\boldsymbol{x} \,|\, n; \theta_1, \ldots, \theta_m)$, 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 θ_i is of interest, and this is a different prior for each θ_i, as given in the paper. The reference prior for θ_i results in a Beta reference marginal posterior $\mathrm{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$. 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 μ_i are the means, the σ_i are the standard deviations, and ρ 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 $\mathrm{N}(x_i \mid \mu, \sigma)$, the right-Haar prior is $\pi(\mu, \sigma) = \sigma^{-1}$, and this is fine for either μ or σ (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 π_i is fine for μ_i, σ_i and ρ, 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 μ_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 $\mathrm{N}(\mu_i \mid x_i, 1)$ posterior for each individual μ_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 θ 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 θ_T is the true value of θ. Furthermore, the posterior distribution of θ 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 $\mathrm{N}(\mu_i \mid 0, 1000)$ vague proper priors results in a posterior mean for θ 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 $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ 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

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

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 θ. Even for the basic $\mathrm{N}(x_i \mid \mu, \sigma)$ model, the Jeffreys-rule prior is $\pi(\mu, \sigma) = 1/\sigma^2$, which results in posterior inferences for μ and σ 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 ρ 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

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

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 π_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 π_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 π^A is a worse objective prior than either right-Haar prior alone, while π^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, θ_i, for which it was created.

Example 1.6. Multinomial Example (continued): The reference prior derived when the parameter of interest is θ_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 θ_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 θ_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 θ_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 θ_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 θ_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 θ_i, let $\boldsymbol{\theta}_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_k)$ denote the parameters other than θ_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 θ_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 π^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

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

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 $p_i(\boldsymbol{x}_i \mid \theta_i)$; 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

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

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 (θ_1, θ_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 θ_1 or θ_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, $\mathrm{Be}(\theta_i \mid 1/2, 1/2)$ distributions for θ_1 and θ_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 ξ_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 ξ_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:

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

where the $U_i(\cdot)$ are to be specified, $\theta_1 < 0$, $\theta_2 = \mathrm{E}\{U_2(X) \mid (\theta_1, \theta_2)\}$, 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 $h(\theta_1) = -\theta_1 + \theta_1 \log(-\theta_1) + \log(\Gamma(-\theta_1))$.

	$G_1(\theta_1)$	$G_2(\theta_2)$	$U_1(x)$	$U_2(x)$	θ_1	θ_2
Normal (μ, σ)	$-\frac{1}{2}\log(-2\theta_1)$	θ_2^2	x^2	x	$-1/(2\sigma^2)$	μ
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	$-\alpha$	μ
Inverse Gamma	$h(\theta_1)$	$-\log\theta_2$	$\log x$	$1/x$	$-\alpha$	μ

The Fisher information matrix of (θ_1, θ_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 θ_1 or θ_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 α and ψ can be found in Sun and Ye (1996).

For the important special case of the Gamma (α, μ) 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 α and μ 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 η_1 and η_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 θ. 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 θ and ψ 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 θ. Ghosh and Sun (1998) showed that this prior is the second order matching prior for θ 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

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

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

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

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 θ_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 θ_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 μ (or any one-to-one transformation of μ) and σ (or any one-to-one transformation of σ, such as the variance σ^2, or the precision σ^{-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* (μ, σ) in that it can be verified that the corresponding joint credible regions for (μ, σ) 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, $\mathcal{F} = \{\pi(\boldsymbol{\theta} \mid \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$, 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*

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

provided that the integral exists.

The non-negative directed logarithmic divergence $\kappa\{p_0 \mid p\}$ 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 θ_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 θ considerably more weight than the means μ_i. To define the proposed criterion, we will also need to utilize the reference prior predictives for $i = 1, \ldots, m$,

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

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

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

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 θ_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 θ_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 θ_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 θ_i. Moreover, the marginal reference prior for θ_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

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

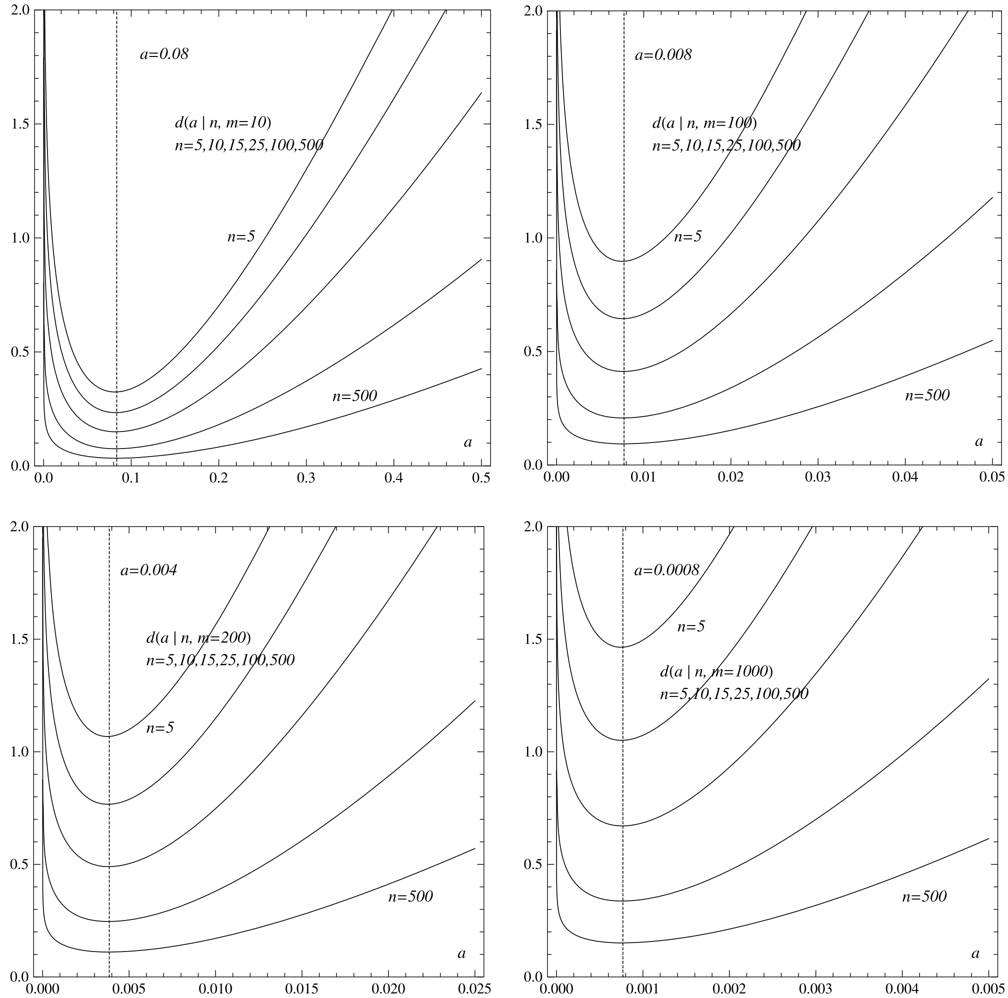

Figure 1: Expected logarithmic losses, when using a Dirichlet prior with parameter $\{a, \ldots, a\}$, in a multinomial model with m cells, for sample sizes $n = 5, 10, 25, 100$ and 500. The panels are for $m = 10, 100, 200$ and 1000. In all cases, the optimal value for all sample sizes is $a^* \approx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

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 $\{1/m, \dots, 1/m\}$ (or $0.8 \times \{1/m, \dots, 1/m\}$ to be slightly more precise), yielding an approximate marginal reference posterior for each of the θ_i's as $\mathrm{Be}(\theta_i \,|\, x_i + 1/m, n - x_i + (m-1)/m)$, having mean and variance

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

The normal model with coefficient of variation

Consider a random sample $\boldsymbol{z} = \{x_1, \dots, x_n\}$ from a normal model $\mathrm{N}(x \,|\, \mu, \sigma)$, with both parameters unknown, and suppose that one is interested in μ and σ, 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 σ/μ).

The joint reference prior when either μ or σ 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

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

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 ϕ 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 ϕ can be shown to be

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

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 ϕ (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 μ or σ 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 μ (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 μ when $a = 1$. The function $d_\mu(a \,|\, n = 10)$ is the lower dotted line in Figure 2. Similarly, the risk associated to σ 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 ϕ cannot be analytically obtained and is numerically computed, using one-dimensional numerical integration over ϕ 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

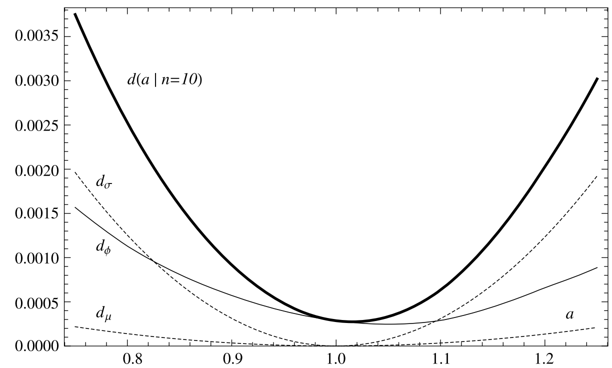

Figure 2: Expected average intrinsic logarithmic losses $d(a \,|\, n, k)$ associated with the use of the joint prior $\pi(\mu, \sigma \,|\, a) = \sigma^{-a}$ rather than the corresponding reference priors when $n = 10$ and $k = 3$.

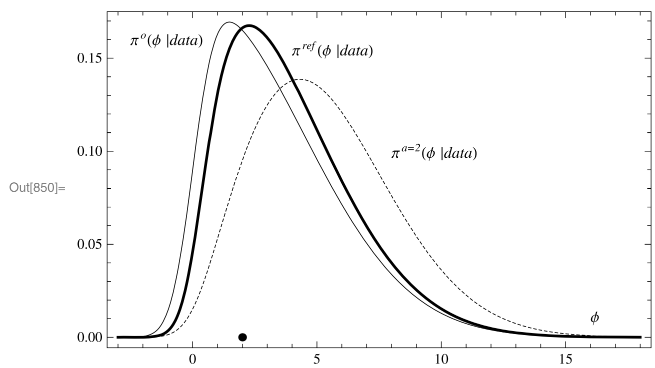

Figure 3: Reference posterior (solid) and marginal overall posterior (black) for ϕ given a minimal random sample of size $n = 2$. The dotted line is the marginal posterior for the prior with $a = 2$, which is the Jeffreys-rule prior.

expected loss, the contribution corresponding to ϕ is the largest, and that corresponding to μ 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 $\mathrm{N}(x \mid 1, \frac{1}{2})$ (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 ϕ 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 ϕ 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 $\pi(\boldsymbol{\theta} \mid a)$, indexed by a single parameter a, can be identified for a given problem, one can compute the marginal likelihood $p(\boldsymbol{x} \mid a)$ (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

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

as the overall prior in the original problem.

4.1 Multinomial problem

The hierarchical prior

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

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

following immediately from the fact that, marginally,

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

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

Proposition 4.1.

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

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

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

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 ∞. 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.

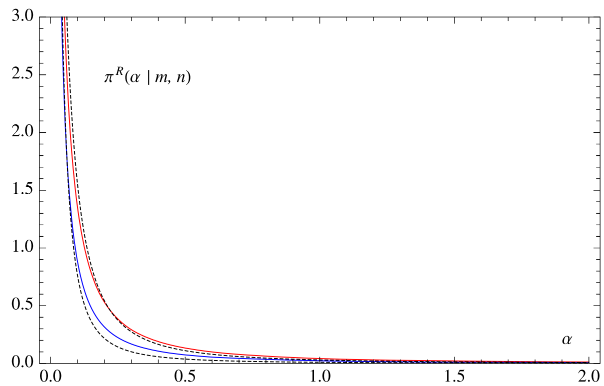

Figure 4: Reference priors $\pi^R(a\,|\,m,n)$ (solid lines) and their approximations (dotted lines) for $(m=150, n=10)$ (upper curve) and for $(m=500, n=10)$ (lower curve).

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 ∞, 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

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

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 ∞. 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

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

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

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

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

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

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

Hence

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

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

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

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

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,

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

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

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 δ 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 δ 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

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

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 δ 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 δ 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

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

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

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

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

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 μ_i and variance 1, for $i = 1 \cdots, m$. We are interested in all the μ_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 τ^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 μ_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 τ^2 being replaced by the rejection step:

Step 1. Propose τ^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 β, 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 β. The temptation here arises because β 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 π^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 β, 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 π^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