*James Berger (Duke University, berger@duke.edu)
Jose Bernardo (University of Valencia, jmbh2008@gmail.com)
Dongchu Sun (University of Nebraska, dsun9@unl.edu)

Objective Bayesian Inference and its Relationship to Frequentism

Contents

0.1 Introduction

In objective Bayesian analysis, the prior distribution is chosen to represent 'minimal information,' the idea being that this allows Bayesian inferences to be made based only on the model (and data). This is thus the part of Bayesian analysis that is compatible with the BFF agenda, since Fiducial and Frequentist analyses are likewise based (primarily) on the model (and data).

There have been many efforts to formally define what it means to be objective and to then develop objective prior distributions. Our belief is that this is generally misguided; there is no unambiguous objectivity. The best one can do is to find prior distributions that represent minimal information, in some sense, and which have desirable properties, such as resulting in statistical procedures that have good frequentist performance. Once found, such priors can be the *conventional priors* to use in objective Bayesian analysis, in the sense that we professionally agree to use them as the objective priors in specified inference problems.

The rest of this section presents background issues that are useful for the development and understanding of objective Bayesian analysis. Section 0.2 reviews some of the most prominent approaches to objective Bayesian analysis that have been developed. Section 0.3 presents our preferred approach, called the *reference prior* approach. The reference prior approach can result in different prior distributions for different unknown parameters in a model, which can be unwieldy in practice. Hence, in Section 0.4, approaches to developing overall objective priors are discussed, with hierarchical normal modeling being highlighted.

0.1.1 Conditioning and Frequentist Performance

Objective Bayesian inference is closely tied with frequentism in various ways, as will be seen throughout the chapter. Here we discuss two general issues needed to understand the relationship.

0.1.1.1 Conditioning

Conditional inference is a crucial concept in statistics, but is often neglected. This is partly because conditioning occurs automatically in Bayesian analysis and is, hence, not taught there. On the other hand, in the frequentist paradigm

conditioning is very difficult and there is no established theory as to how to choose conditional procedures. It is useful to begin with a simple example, taken from [13], which discusses conditioning in general.

Example 0.1 There are two observations, x_1 and x_2, having distribution

$$x_i = \begin{cases} \theta + 1 & \text{with probability } \frac{1}{2} \\ \theta - 1 & \text{with probability } \frac{1}{2} \end{cases} .$$

Consider the confidence set, for the unknown θ, given by

$$C(x_1, x_2) = \begin{cases} \text{the point } \{\frac{1}{2}(x_1 + x_2)\} & \text{if } x_1 \neq x_2 \\ \text{the point } \{x_1 - 1\} & \text{if } x_1 = x_2. \end{cases}$$

The frequentist coverage of this confidence set can be shown to be

$$\Pr(C(x_1, x_2) \text{ contains } \theta \,|\, \theta) = 0.75.$$

This is a very inadequate report, once the data is observed. Indeed, observe that, if $x_1 \neq x_2$, then it is certain that their average is equal to θ, so the confidence should then be given as 100%. Conversely, if $x_1 = x_2$, θ could be the data's common value plus one or the common value minus one, each of which is equally likely to have led to this result.

A common way to obtain sensible frequentist answers here, is to define the conditioning statistic $s = |x_1 - x_2|$ chosen to measure the 'strength of evidence' in the data: $s = 2$ reflects data with maximal evidential content, while $s = 0$ reflects data of minimal evidential content. One then defines frequentist coverage in the usual way, but does so conditional on the strength of evidence s. An easy computation gives this conditional confidence, for the two distinct cases, as

$$\begin{aligned} \Pr(C(x_1, x_2) \text{ contains } \theta \,|\, s = 2, \theta) &= 1 \\ \Pr(C(x_1, x_2) \text{ contains } \theta \,|\, s = 0, \theta) &= \tfrac{1}{2}. \end{aligned}$$

Note that the unconditonal coverage is still 75% (the report of 100% occurs half the time and the report of 50% occurs half the time), so these conditional reports are still proper frequentist reports. But clearly the conditional reports are much more informative.

Finding good conditioning statistics in complex problems is very difficult, so that the conditional frequentist theory of statistics is underdeveloped. In contrast the objective Bayesian approach automatically conditions properly.

Example 0.2 *Example 0.1 continued.* The objective prior for this example is $\pi(\theta) = 1$ (since θ is a location parameter — see Section 0.2.4). Application of Bayes theorem shows that, if $x_1 \neq x_2$, the posterior distribution for θ gives probability one to the point $(x_1 + x_2)/2$ while, if

$x_1 = x_2$, the posterior distribution gives probability 1/2 each to the common value of the data plus 1 and the common value minus 1. Thus the objective Bayesian confidence statements for $C(x_1, x_2)$ are 1 and 0.5 for the two cases, respectively, which is the correct answer.

In the above example, the objective Bayesian analysis produced the same stated confidences as did the optimal conditional frequentist procedure, automatically conditioning correctly. This is not a coincidence. Quite often, objective Bayesian procedures yield results that are (nearly) optimal (conditional) frequentist answers, and often the conditional frequentist results could not be easily obtained in any other way.

0.1.1.2 Frequentist Performance

Objective Bayesian priors are often chosen to achieve (unconditional) frequentist goals. For instance, 'frequentist-matching' priors (see Section 0.2.3) are priors that produce Bayesian credible sets with good frequentist properties. Reference priors will also be seen to yield procedures that have excellent frequentist performance.

Objective Bayesian inference can also overcome some problems that face frequentists. Here is an example.

Example 0.3 *A Variance Components Problem:* For $i = 1, \ldots, m$,

$$x_i \sim \text{Normal}(x_i \mid \mu_i, 1) \quad \text{and} \quad \mu_i \sim \text{Normal}(\mu_i \mid \xi, \tau^2),$$

the μ_i and ξ being means and 1 and τ^2 being variances. The marginal density of x_i, given ξ and τ^2, is

$$m(x_i \,|\, \xi, \tau^2) = \int \text{N}(x_i \,|\, \mu_i, 1)\text{N}(\mu_i \,|\, \xi, \tau^2)\; d\mu_i = \text{N}(x_i \,|\, \xi, 1+\tau^2)\,.$$

The marginal likelihood for the full data $\boldsymbol{x} = (x_1, \ldots, x_m)$, and with $s^2 = \sum (x_i - \bar{x})^2$, is then

$$m(\boldsymbol{x} \mid \xi, \tau^2) = \prod_{i=1}^{m} \text{N}(x_i \,|\, \xi, 1+\tau^2) \propto \frac{1}{(1+\tau^2)^{m/2}} \exp\left\{ -\frac{n(\bar{x}-\xi)^2 + s^2}{2(1+\tau^2)} \right\}. \tag{1}$$

The standard maximum liklihood estimate (mle) $\hat{\xi}_{mle} = \bar{x}$ is fine, but $\hat{\tau}^2_{mle} = \max\{0, \frac{s^2}{m} - 1\}$, the marginal maximum likelihood estimate of τ^2, is problematical. This is particularly so if $s^2/m < 1$, in which case the mle would be $\hat{\tau}^2_{mle} = 0$. Setting τ^2 to 0 is equivalent to declaring that $\mu_1 = \ldots = \mu_p$ exactly (since $\mu_i \sim Normal(\cdot \mid \xi, \tau^2)$); this would be silly.

Indeed, in this situation, there is actually a great deal of uncertainty

about τ^2. This can be seen by looking at the marginal likelihood of τ^2 found by integrating (1) over ξ, resulting in

$$m(\boldsymbol{x} \,|\, \tau^2) \propto (\tau^2+1)^{-(m-1)/2} \, \exp\left\{ -\frac{s^2}{2(\tau^2+1)} \right\}. \tag{2}$$

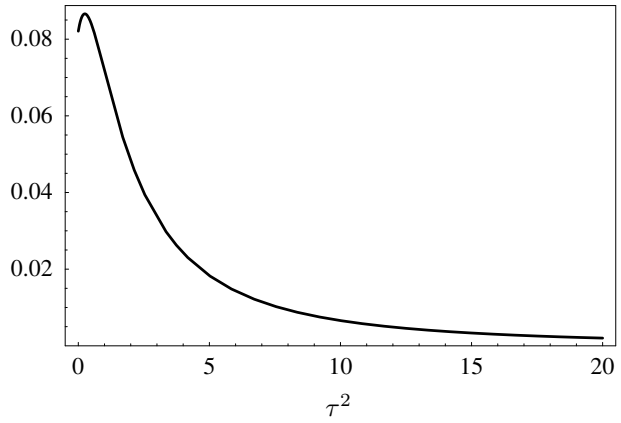

FIGURE 1
Marginal likelihood function of τ^2 when $p = 5$ and $s^2 = 5$ is observed.

For illustration, Figure 1 gives this marginal likelihood when $m = 5$, and $s^2 = 5$. It is mostly decreasing away from 0, but not very quickly, indicating that there is considerable uncertainty about τ^2 even though $\hat{\tau}^2_{mle} = 0$. Frequentist methods have difficulty incorporating the uncertainty in τ^2 when the maximum of the likelihood is achieved at a boundary, while objective Bayesian analysis results in a posterior distribution for τ^2 that correctly reflects the uncertainty in the likelihood.

While a value of $s^2/m \leq 1$ is somewhat unusual here (if, for instance, $m = 5$, $n = 1$ and $\tau^2 = 1$, then $\Pr(s^2/5 < 1) = 0.264$), it is common in variance component models with multiple variances to have at least one mle variance estimate equal to 0.

Another frequentist use of objective Bayesian procedures is as an alternative to asymptotics. Asymptotics is often used to develop statistical procedures, but the performance of the procedures with small or moderate sample sizes may well be uncertain. The objective Bayesian approach is an attractive alternative to asympotics, for two reasons.

First, objective Bayesian procedures typically automatically have the desired asymptotic properties. For instance, the central limit theorem was first developed by [26] as a normal approximation to the posterior distribution; this approximation is identical to the frequentist central limit theorem commonly taught today. Objective Bayesian procedures will typically even have

the correct asymptotics when the problem is irregular (i.e., the usual central limit theorem does not hold); Example 0.17 provides an illustration. Note that objective Bayesians need not do any asymptotic computations; the Bayesian procedures will automatically yield the correct asymptotic answer if the sample size is large enough for the asymptotics to apply. The second motivation for using objective Bayesian analysis is that the objective Bayesian answers are still likely to be sensible when the sample size is too small for asymptotics to apply.

0.1.2 Informal Objective Bayesian Solutions

A variety of *ad hoc* methods have been proposed for choosing a default prior distribution and it is useful to briefly discuss some of them so as to see the problems that can be encountered. Use of a constant prior distribution is the most common objective Bayesian method and will be discussed in Section 0.2.1. While a constant prior is often fine, examples will be seem in which the constant prior is inadequate.

0.1.2.1 Truncation of the Parameter Space

If the parameter space is unbounded, there is often worry about using, say, a constant prior, because the prior then has infinite mass and there is no guarantee that the resulting posterior distribution will be proper. A common 'solution' is to, instead, choose the prior to be constant over some (large) bounded region of the parameter space, since it will then be proper (after normalization). This does not really solve the problem, however, because, if the posterior resulting from the constant prior on the entire parameter space is improper, then inferences arising from the truncated prior will strongly depend on the (arbitrary) choice of the truncation point. Thus, to use this approach one would need to perform careful sensitivity studies to determine the effect of the truncation points.

0.1.2.2 Vague Proper Priors

Use of vague proper priors is popular, the following being a typical example.

Example 0.4 *Normal mean.* Suppose the problem is to estimate a normal mean θ, with known variance. The standard objective prior is $\pi(\theta) = 1$. A vague proper prior alternative is to use the $\mathrm{N}(\theta \,|\, 0, K)$ prior distribution, with a large value of K. In estimation problems this will yield essentially the same answer as the constant prior and is thus reasonable to use but, in hypothesis testing, such vague proper priors can be very bad (see the chapter "Objective Bayesian Testing and Model Uncertainty" in this book).

It is not uncommon to see vague proper priors used that are an approxima-

tion to an inferior default prior. The use of the vague proper prior then is also inferior and its use can hide the problem. Here is a commonly encountered example.

Example 0.5 *Example 0.3 continued.* In the variance components example, consider priors of the form $\pi(\xi, \tau^2) = \pi^*(\tau^2)$ (constant in ξ). The resulting posterior distribution for τ^2 is

$$\begin{aligned} \pi(\tau^2 \mid \mathbf{x}) &\propto \int \frac{1}{(1+\tau^2)^{m/2}} \exp\left\{-\frac{n(\bar{x}-\xi)^2 + s^2}{2(1+\tau^2)}\right\} \pi^*(\tau^2) d\xi \\ &\propto \frac{1}{(1+\tau^2)^{(m-1)/2}} \exp\left\{-\frac{s^2}{2(1+\tau^2)}\right\} \pi^*(\tau^2)\,. \end{aligned}$$

A commonly used prior for a normal variance σ^2 is $1/\sigma^2$, and so it is tempting to choose $\pi^*(\tau^2) = 1/\tau^2$. This results in an improper posterior, however, because there is a non-integrable singularity at zero,

$$\int_0^\epsilon \pi(\tau^2 \mid \mathbf{x}) d\tau^2 \approx \exp\left\{-\frac{s^2}{2}\right\} \int_0^\epsilon \frac{1}{\tau^2} d\tau^2 = \infty\,.$$

A commonly used vague proper alternative is $\pi^*(\tau^2) \propto \tau^{-2(1+\epsilon)} e^{-\epsilon/\tau^2}$, and the resulting posterior will be proper. However, this posterior gives almost all of its mass to a small region near 0, and will essentially ignore the data. Thus the vague proper prior does not fix the problem, and may even cause the problem to be hidden. Instead, one should simply use a good objective prior for the problem. The simplest choice is $\pi^*(\tau^2) = 1/\sqrt{\tau^2}$, which results in an excellent objective posterior distribution.

Another commonly used ad hoc vague proper prior is to choose a flat prior (e.g., a contant) that 'spans the range of the likelihood function.' The rationale for doing this is typically stated to be a desire to reduce the influence of the prior, by choosing a prior that is compatible with the likelihood function. It is not really possible to evaluate this strategy. On the one hand, it may often be better than using an arbitrarily chosen objective prior (e.g. a constant prior), but it corresponds to a disturbing double use of the data.

0.1.3 Justifying Posteriors from Improper Priors

When $\pi(\boldsymbol{\theta})$ is improper, Bayes theorem no longer applies, and so it is not obvious how to justify use of the posterior density

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,. \tag{3}$$

For this expression to define a probability density, it is clearly necessary that

$$p(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty\,, \tag{4}$$

and we take this as a minimal requisite condition for use of the improper prior to make sense (but see [27] for counterarguments).

This section presents a strong justification for use of (3) with improper priors, showing that $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is the limit of proper posteriors arising from proper priors that approximate the improper $\pi(\boldsymbol{\theta})$. To this end, for an improper prior $\pi(\boldsymbol{\theta})$, consider an increasing sequence of subsets of the parameter space $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \ldots\}$, such that

$$\boldsymbol{\Theta}_i \subset \boldsymbol{\Theta}_{i+1}, \int_{\boldsymbol{\Theta}_i} \pi(\boldsymbol{\theta}) < \infty, \ \forall\, i \geq 1, \quad \text{and} \ \lim_{i \to \infty} \boldsymbol{\Theta}_i = \boldsymbol{\Theta}. \tag{5}$$

Then, $\pi(\boldsymbol{\theta})$ can be normalized within each of the $\boldsymbol{\Theta}_i$'s to produce the sequence of *proper* priors $\{\pi_1(\boldsymbol{\theta}), \pi_2(\boldsymbol{\theta}), \ldots\}$ defined by

$$\pi_i(\boldsymbol{\theta}) = \begin{cases} \dfrac{\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}_i} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_i \\ 0 & \text{otherwise.} \end{cases}$$

Since the $\pi_i(\boldsymbol{\theta})$'s are all proper probability distributions, Bayes theorem can be applied, given the data $\boldsymbol{x}$, to obtain the corresponding sequence of posterior densities $\{\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{x}), \pi_2(\boldsymbol{\theta} \,|\, \boldsymbol{x}), \ldots\}$ defined by

$$\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \begin{cases} \dfrac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi_i(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}_i} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}} & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_i \\ 0 & \text{otherwise.} \end{cases}$$

Here is the type of convergence of distributions that we consider.

Definition 0.1 Reverse KL Convergence. *A sequence of probability distributions, defined by their density functions* $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$*, is said to* reverse KL converge *to a probability distribution with density* $p(\boldsymbol{x})$ *whenever*

$$\lim_{i \to \infty} \int_{\boldsymbol{\mathcal{X}}} p_i(\boldsymbol{x}) \log \frac{p_i(\boldsymbol{x})}{p(\boldsymbol{x})}\, d\boldsymbol{x} = 0\,.$$

Theorem 0.1 *Let* $\pi(\boldsymbol{\theta})$ *be an improper prior on* $\boldsymbol{\Theta}$*, having a sequence of subsets satisfying (5). If* $\int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$ *at* $\boldsymbol{x}$*, then* $\{\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})\}$ *is reverse KL convergent to* $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ *at* $\boldsymbol{x}$*.*

This theorem, from [7], is quite powerful in that it says that only the minimal condition (4) is needed for reverse KL convergence; thus if the formal posterior is proper, its use is justified.

0.2 Approaches to Objective Bayesian Analysis

There have been many efforts to develop and implement objective Bayesian analysis. Some of the more prominent approaches are discussed in this section. Discussion of our favorite approach, called *reference analysis*, is delayed until Section 0.3. Two other approaches, the confidence distribution approach and the fiducial approach, can also be considered to be attempts at implementing objective Bayesian analysis, but we do not present them here because they are extensively discussed in other chapters of the handbook.

0.2.1 The Constant Prior and Inverse Probability

The first clear Bayesian was the Reverend Thomas Bayes. His paper (published posthumously) *An essay towards solving a problem in the doctrine of chances* ([1]) was the first unambiguous statement and use of what is now known as Bayes theorem.

The paper was also a significant work in objective Bayesian analysis. In today's language, the paper considered the basic binomial problem of observing x from $p(x \mid \theta) = \text{Bi}(x \mid n, \theta)$, with θ unknown. While inference concerning θ given an observed x (via Bayes theorem) was one focus of the paper, the other focus was that of coming up with a reasonable objective prior distribution, $\pi(\theta)$, for θ. Bayes argued that the marginal distribution of x should be uniform, and showed that $\pi(\theta) = 1$ achieved this; this was thus the recommended objective prior.

The real inventor of objective Bayesian analysis, as a paradigm for statistical analysis, was Simon Laplace. In a series of papers culminating in *Théorie Analytique des Probabilités* ([26]), he invented a statistical paradigm that was to dominate statistics for well over a century. The approach of Laplace to statistical inference was essentially objective Bayesian analysis with utilization of a constant prior density for all unknown parameters of statistical models. Laplace did not do this blindly, but rather argued that one should choose a parameterization of the problem in which different values of the unknown were equally likely.

To distinguish Laplace's method from ordinary probability reasoning (computing probabilities of various data arising from the model) Augustus de Morgan ([20]) called reasoning from the data to the model *inverse probability*. This name reigned until roughly 1950, when it became replaced by 'Bayesian analysis' for rather unclear reasons, as extensively discussed in [22].

Inverse probability (*i.e.*, objective Bayesian analysis using a constant prior density) was arguably the dominant mode of formal statistical inference until the 1920's ([17], [23]). While dominating statistical practice, inverse probability was not without its critics. Much of the criticism was ill-placed, but at

least some of the philosophical criticisms were legitimate. Foremost among these criticisms was the fact that the statistical answers arising from the use of inverse probability will depend on the choice of the parameterization that is used for the statistical model.

Example 0.6 *Inference on a binomial parameter.* Suppose x is distributed according to the binomial $\mathrm{Bi}(x \,|\, n, \theta)$ distribution. Inverse probability says use the objective prior $\pi(\theta) = 1$. Suppose someone chose $\psi(\theta) = \frac{2}{\pi}\mathrm{ArcSin}(\sqrt{\theta})$ as the unknown parameter, rather than θ. Inverse probability then says use the prior $\pi(\psi) = 1$ (note that $0 < \psi < 1$).

To see that this will typically result in different Bayesian answers, make the change of variables from ψ back to θ. With this change of variables, the objective prior $\pi(\psi) = 1$ transforms to

$$\pi(\theta) = 1 \times \psi'(\theta) = \pi^{-1}\, \theta^{-1/2}(1-\theta)^{-1/2},$$

which is the $\mathrm{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$ density. The posterior mean for θ corresponding to the uniform prior can be shown to be $(x+1)/(n+1)$, while that corresponding to the $\mathrm{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$ prior will be $(x+\frac{1}{2})/(n+\frac{1}{2})$, which are clearly different. Of course, these are typically very close; if $x = 5$ and $n = 30$, for instance, the two posterior means are 0.188 and 0.177 respectively, and the difference is negligible from a practical perspective (since the posterior standard errors are on the order of 0.07). Still, it is unappealing to have different answers depending on the rather arbitrary choice of parameterization.

0.2.2 The Jeffreys-rule Prior

In the 1930's, Harold Jeffreys, being aware of the inconsistencies that could arise if one always used the constant prior, sought a 'rule' to specify an objective prior which would give the same results, no matter which parameterization is chosen. In [25] he developed such a rule, for any continuous parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \boldsymbol{\Theta} \subset \mathbb{R}^m$, based on the *Fisher information matrix* $I(\boldsymbol{\theta})$, namely the $m \times m$ matrix having elements (assuming they exist)

$$I(\boldsymbol{\theta})_{ij} = \mathrm{E}^{\boldsymbol{x}\,|\,\theta}\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \right] = \int_{\mathcal{X}} \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \right] p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) d\boldsymbol{x}. \quad (6)$$

It can easily be shown that the prior distribution

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}, \quad (7)$$

called the *Jeffreys-rule* prior, is invariant under one-to-one reparameterization and so will result in coherent answers across different parameterizations.

The Jeffreys-rule prior can be proper (when normalized) or improper. When

improper, it has the surprising property of virtually always resulting in a proper posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, |I(\boldsymbol{\theta})|^{1/2}$, *i.e.*, $\int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) |I(\boldsymbol{\theta})|^{1/2}\, d\boldsymbol{\theta} < \infty$.

When one has n iid observations, $I(\boldsymbol{\theta})$ is just n times the Fisher information for a single observation. Since proportionality constants like n are irrelevant if the prior is improper and are renormalized anyway if the prior is proper, one only needs to compute the Fisher information for one observation.

0.2.2.1 The Single Parameter Case

In one parameter *regular* continuous models, $\{p(x \,|\, \theta), \theta \in \Theta \subset \mathbb{R}, x \in \mathcal{X}\}$, where the sampling space $\mathcal{X}$ does not depend on the parameter θ, and the model probability density is twice continuously differentiable with respect to θ, Jeffreys rule (7) reduces to

$$\pi(\theta) \propto \sqrt{I(\theta)}, \quad I(\theta) = \mathrm{E}^{x \,|\, \theta}\left[- \frac{\partial^2}{(\partial\theta)^2} \log p(x \,|\, \theta) \right]. \tag{8}$$

Example 0.7 *Jeffreys-rule prior for the parameter of a Poisson model.* Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample of size n from a Poisson probability model $\mathrm{Pn}(x \,|\, \theta) = e^{-\theta}\theta^x/x!$. Recalling that, in the iid case, it suffices to compute the Fisher information for a single observation, and noting that $\mathrm{E}[x \,|\, \theta] = \theta$, the Fisher information is

$$I(\theta) \quad = \mathrm{E}^{x \,|\, \theta}\left[- \frac{\partial^2}{(\partial\theta)^2} \log \frac{e^{-\theta}\theta^x}{x!} \right] = \mathrm{E}^{x \,|\, \theta}\left[\frac{x}{\theta^2} \right] = \frac{1}{\theta}\,.$$

Thus, using (8), the Jeffreys-rule prior is the improper prior

$$\pi(\theta) \propto \sqrt{I(\theta)} = \theta^{-1/2}.$$

Example 0.8 *Location models.* Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample from a location model $p(x \,|\, \theta) = h(x - \theta)$, $x \in \mathbb{R}$, $\theta \in \mathbb{R}$. (8) gives

$$\pi(\theta) \propto \left[\int_{-\infty}^{\infty} \frac{h'(x-\theta)^2 - h(x-\theta)h''(x-\theta)}{h(x-\theta)}\, dx \right]^{1/2},$$

and changing the integration variable to $y = x - \theta$ shows that the integral does not depend on θ and, hence, the resulting prior is proportional to some constant.

Example 0.9 *Scale models.* Consider the model $p(x \,|\, \theta) = \theta^{-1} g(x/\theta)$. This is a general scale model and can be transformed to the location model $p(y \,|\, \phi) = g(\exp\{y - \phi\})$, with the change of variables $y = \log x$ and $\phi = \log\theta$. For ϕ, the Jeffreys-rule prior is constant and changing variables back to θ results in $\pi(\theta) \propto \pi(\phi)|d\phi/d\theta| \propto \theta^{-1}$, which is thus the Jeffreys-rule prior for scale models.

As stated earlier, Jeffreys rule cannot be applied to non-regular problems. This includes those situations where the sampling space $\mathcal{X}$ depends on the parameter.

Example 0.10 *Uniform data on* $[0, \theta]$. Let x be a sample from the uniform distribution on the real interval $(0, \theta)$, so that $p(x \,|\, \theta) = \theta^{-1}$, if $0 < x < \theta$, and zero otherwise. The sample space $\mathcal{X} = (0, \theta)$ depends on the parameter θ and, therefore, Jeffreys rule (8) is not applicable. (Blind use of the formula yields a negative information function.) This will be considered, again, in Section 0.3.2, where the reference prior approach will be shown to yield the sensible objective prior $\pi(\theta) = 1/\theta$.

0.2.2.2 The Multi-parameter Case

Although Jeffreys rule (7) does provide a multivariate invariant objective prior, the results are very often unappealing.

Example 0.11 *Inference on a normal mean, with variance unknown.* Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample of size n from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma^2)$, both parameters unknown. Using (6), the corresponding Fisher information matrix for a single observation is

$$I(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/[2\sigma^4] \end{pmatrix}, \tag{9}$$

and, hence, by (7), the Jeffreys-rule prior is

$$\pi(\mu, \sigma^2) \propto |I(\mu, \sigma^2)|^{1/2} \propto \sigma^{-3}. \tag{10}$$

If the prior is of the form $\pi(\mu, \sigma^2) \propto \sigma^{-\alpha}$, for some $\alpha > 0$, Bayes theorem yields a joint posterior

$$\pi(\mu, \sigma^2 \,|\, \boldsymbol{x}, \alpha) \propto \sigma^{-(n+\alpha)} \exp\left\{-\frac{n}{2\sigma^2}[s^2 + (\overline{x} - \mu)^2]\right\}, \tag{11}$$

where $\overline{x}$ is the sample mean and $s^2 = \sum (x_i - \overline{x})^2/n$. Integrating out σ^2, the corresponding marginal posterior for μ is (provided $n + \alpha > 3$), the Student t density

$$\pi(\mu \,|\, \boldsymbol{x}, \alpha) = \mathrm{St}(\mu \,|\, \overline{x},\, s/\sqrt{n + \alpha - 3}, n + \alpha - 3)\,. \tag{12}$$

The use of (10) ($\alpha = 3$) thus leads to $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}(\mu \,|\, \overline{x},\, s/\sqrt{n}, n)$.

Jeffreys was unhappy with this result; he mentions that this is against the 'standard practice' of removing one degree of freedom from the sample size per estimated parameter (which is the result for $\alpha = 2$). This led Jeffreys to recommend the use of $\alpha = 2$, *i.e.*, of $\pi(\mu, \sigma^2) \propto \sigma^{-2}$, even though it contradicted his multivariate rule. His argument to rationalize this choice was to treat μ and σ independently. This Jeffreys-recommended prior has now become known as the *Jeffreys independence prior*.

Example 0.12 *Multinomial distribution.* Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is multinomial $\mathrm{Mu}(\boldsymbol{x} \,|\, n, \theta_1, \ldots, \theta_m)$ (suppressing the $(m+1)$st cell count, $x_{m+1} = n - \sum_{j=1}^{m} x_j$, and probability, $\theta_{m+1} = 1 - \sum_{j=1}^{m} \theta_j$, since they are determined by the others) so that

$$p(\boldsymbol{x} \,|\, n, \theta_1, \ldots, \theta_m) = \frac{n!}{\prod_{j=1}^{m} x_j!(n - \Sigma x_j)!} \prod_{j=1}^{m} \theta_j^{x_j} (1 - \Sigma\theta_j)^{n - \Sigma x_j} \,.$$

Using (6), computation of the Fisher information matrix yields

$$I(\theta_1, \ldots, \theta_m) = \frac{n}{1 - \Sigma\theta_j} \begin{bmatrix} \frac{1+\theta_1-\Sigma\theta_j}{\theta_1} & 1 & \ldots & 1 \\ 1 & \frac{1+\theta_2-\Sigma\theta_j}{\theta_2} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & \frac{1+\theta_m-\Sigma\theta_j}{\theta_m} \end{bmatrix}$$

with

$$|I(\theta_1, \ldots, \theta_m)| = n^m \left[\left(1 - \sum\nolimits_{j=1}^{m} \theta_j\right) \prod\nolimits_{j=1}^{m} \theta_j \right]^{-1} .$$

Thus, the Jeffreys-rule prior, proportional to $|I(\theta_1, \ldots, \theta_m)|^{1/2}$, is the proper Dirichlet prior

$$\pi^J(\theta_1, \ldots, \theta_m) \propto \left(1 - \sum\nolimits_{j=1}^{m} \theta_j\right)^{-1/2} \prod_{j=1}^{m} \theta_j^{-1/2} \,, \tag{13}$$

i.e., the $\mathrm{Di}((\theta_1, \ldots, \theta_m) \,|\, (\frac{1}{2}, \ldots, \frac{1}{2}))$ distribution. Multiplying this by the multinomial likelihood immediately yields that the corresponding posterior distribution is $\mathrm{Di}((\theta_1, \ldots, \theta_m) \,|\, (x_1 + \frac{1}{2}, \ldots, x_m + \frac{1}{2}))$.

Again, (13) does *not* have appropriate behavior for an objective prior. As a numerical illustration of this, consider the case where the sample size n is small relative to the number of classes $m+1$; thus we have a large sparse table. 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 using the Jeffreys-rule prior can be shown to be

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

so that $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] = 2.5/503$, $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] = 1.5/503$, and $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = 0.5/503$. So, cells 240 and 876 only have total posterior probability of $4/503 = 0.008$, 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.

In this section, we have seen two important examples — the normal model with both parameters unknown and the multinomial model — where the multiparameter form of the Jeffreys-rule prior fails to provide appropriate objective posteriors. We actually know of no multiparameter example in which the Jeffreys-rule prior has been verified to be satisfactory. In higher dimensions, that prior always seems to be either 'too diffuse' as in the normal examples, or 'too concentrated' as in the multinomial example.

0.2.3 Frequentist Matching

We have encountered situations where Bayesian credible sets are identical to frequentist confidence sets and have the same stated confidence or coverage. Ensuring that this holds, at least approximately, is the goal of the frequentist matching approach to developing objective priors.

0.2.3.1 Definition of Matching Priors

It is customary to consider one-sided credible sets $(-\infty, \theta_\alpha(\boldsymbol{x})]$, for a real parameter θ, where $\theta_\alpha(\boldsymbol{x})$ denotes the $\alpha-$posterior quantile of θ given $\boldsymbol{x}$, i.e.,

$$\Pr(\theta \leq \theta_\alpha(\boldsymbol{x}) \,|\, \boldsymbol{x}) = \alpha, \tag{14}$$

for $\alpha \in (0,1)$. Now, consider $C(\boldsymbol{x}) \equiv (-\infty, \theta_\alpha(\boldsymbol{x})]$ to be a frequentist confidence set, with $\boldsymbol{x}$ random and arising from $p(\boldsymbol{x} \,|\, \theta)$, for given θ. The frequentist coverage probability of this confidence set is clearly

$$\Pr(\theta \leq \theta_\alpha(\boldsymbol{x}) \,|\, \theta)\,. \tag{15}$$

Note that, in (14), θ is random while $\boldsymbol{x}$ is fixed but, in (15), $\boldsymbol{x}$ is random while θ is fixed. The idea is now to study how close the frequentist coverage probability $\Pr(\theta \leq \theta_\alpha(\boldsymbol{x}) \,|\, \theta)$ is to α, the stated Bayesian coverage.

Definition 0.2 *If the frequentist coverage (15), of the one-sided Bayesian credible sets $(-\infty, \theta_\alpha(\boldsymbol{x})]$, is exactly equal to the Bayesian credible probability α, then the Bayesian credible sets are said to be* exact frequentist matching *(or 'exact matching' for short).*

Example 0.13 *Normal mean with variance known.* Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a random sample of size n from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma^2)$, with μ unknown but σ^2 known. Under the Jeffreys prior, $\pi(\mu) \propto 1$, Bayes theorem yields $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x},\, \sigma^2/n)$. In this case, the posterior α-quantile of μ given $\boldsymbol{x}$ is $\mu_\alpha(\boldsymbol{x}) = \overline{x} + Z_\alpha \sigma/\sqrt{n}$, where Z_α is the α-quantile of $\mathrm{N}(0,1)$. Algebra yields that the frequentist coverage is

$$\Pr(\mu \leq \mu_\alpha(\boldsymbol{x}) \,|\, \mu) = \Pr\left(\frac{\sqrt{n}(\overline{x}-\mu)}{\sigma} \geq -Z_\alpha \;\middle|\; \mu\right) = \Pr\left(\frac{\sqrt{n}(\overline{x}-\mu)}{\sigma} \leq Z_\alpha \;\middle|\; \mu\right) = \alpha,$$

$\forall \mu \in \mathbb{R}$ and $\alpha \in (0,1)$. Consequently, the constant prior for μ is an exact matching prior for μ.

Having exact matching credible sets is actually a fairly common phenomenon. Indeed, this occurs for a wide class of models that are 'invariant' to certain transformations; see Section 0.2.4 for development. But more generally, one must settle for a weaker form of matching.

Definition 0.3 Asymptotic Matching Prior. *A prior* $\pi(\boldsymbol{\theta})$ *is an* i-th order matching *prior for* $\eta = g(\boldsymbol{\theta})$ *if*

$$P(\eta \leq \eta_\alpha(\boldsymbol{x}) \,|\, \boldsymbol{\theta}) = \alpha + o(n^{-i/2}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \alpha \in (0,1). \tag{16}$$

Note that virtually any prior with full support is matching up to an error term of order $O(n^{-1/2})$. What is typically thus sought is an error term of order $O(n^{-1})$, which would be a 'first order' matching prior. Higher order matching is rare (except for the exact matching scenario), but sometimes possible.

0.2.3.2 Single Parameter Case

When there is a single parameter, [33] proved that a one-sided posterior credible interval arising from use of the Jeffreys-rule prior (proportional to the square root of the Fisher information $I(\theta)$), has the desired frequentist coverage probability up to $O(n^{-1})$. In fact, they showed that a prior π is a first order matching prior if, and only if, the differential equation

$$\frac{\partial}{\partial \theta}\Big(\frac{\pi(\theta)}{\sqrt{I(\theta)}}\Big) = 0 \tag{17}$$

is satisfied. Consequently, under the regularity conditions where the Fisher information exists, the Jeffreys prior is the only first order matching prior.

0.2.3.3 Multiparameters

[29] was among first authors to study matching priors for the two parameter case, where one is a parameter of interest and the other is a nuisance parameter. [30] extended the results in [33] and [29] and introduced a method to find a prior that is matching for a parameter to order $O(n^{-1})$, in the presence of nuisance parameters. [32] generalized this method by using a one-to-one transformation of the parameter vector into a parameter of interest and a nuisance parameter vector orthogonal in the sense of [16]. [18] derived a matching prior for a smooth function $\eta = \eta(\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$, their result being stated in the following theorem.

Theorem 0.2 *Let* $I(\boldsymbol{\theta})$ *be the Fisher information for one observation, and define the gradient vector*

$$\nabla\eta = \{\partial\eta/\partial\theta_1, \cdots, \partial\eta/\partial\theta_m\}^t. \tag{18}$$

Also, define

$$\boldsymbol{\xi}(\boldsymbol{\theta}) \equiv (\xi_1, \cdots, \xi_m)^t = \frac{I(\boldsymbol{\theta})^{-1}\nabla\eta}{\sqrt{(\nabla\eta)^t I(\boldsymbol{\theta})^{-1}\nabla\eta}}. \tag{19}$$

Then a first order matching prior $\pi(\boldsymbol{\theta})$ for η must satisfy

$$\sum_{i=1}^{m} \frac{\partial}{\partial\theta_i}[\pi(\boldsymbol{\theta})\boldsymbol{\xi}_i(\boldsymbol{\theta})] = 0. \tag{20}$$

A matching prior is not unique in the multiparameter case. Indeed, the class of solutions of (20) is often quite large, growing with the dimension m. Also, note that the equation to be solved depends on the parameter of interest, so that there could be a different matching prior for each parameter of interest. If the interest is in all the coordinates of $\boldsymbol{\theta}$, one might hope that there is a simultaneous solution to all m matching equations. Sometimes this is so, but not always (cf. [19]).

Example 0.14 *Inference on a normal variance with unknown mean.* Consider again Example 0.11 and define $s_{n-1}^2 = \sum(x_i - \overline{x})^2/(n-1)$. Consider the class of priors $\pi_a(\mu, \sigma^2) \propto 1/\sigma^a$.

When $\eta = \sigma^2$, $\nabla\eta = (0, 1)^t$, and $\boldsymbol{\xi} \equiv (0, \sqrt{2}\sigma^2)^t$. Thus (20) becomes

$$\sqrt{2}\,\frac{\partial\sigma^2\pi(\mu, \sigma^2)}{\partial\sigma^2} = 0.$$

Clearly the independent Jeffreys prior for (μ, σ^2) is a first order matching prior for σ^2, but the Jeffreys-rule prior ($a = 3$) is not. *Thus, contrary to the one-parameter case (under regularity), the Jeffreys-rule prior need not be first order matching in multiparameter problems.*

A comprehensive study of higher order matching priors can be found in [19].

0.2.4 Invariance Priors

The big advance of Jeffreys-rule priors was in providing inferences that were consistent no matter what parameterization was chosen for the model; this is often called 'invariance to parameterization.' Here we discuss an additional type of invariance, namely invariance to other transformations (of both the data and the parameters) that seem to leave the problem unchanged. There are many illustrations of this in the literature but the most extensively studied and reliable invariance theory is 'invariance to a group operation on the model'; here is an illustration.

Example 0.15 Consider a *scale parameter* density

$$p(x \,|\, \theta) = \frac{1}{\theta} f\Big(\frac{x}{\theta}\Big), \quad x > 0, \quad \theta > 0.$$

- Consider a change of scale for the problem (*e.g.*, measure in terms of meters, rather than feet) to $x^* = c\ x > 0$ and $\theta^* = c\ \theta > 0$.
- The density of x^* can be shown to be $p(x^* \mid \theta^*) = (1/\theta^*) f(x^*/\theta^*)$, which is of exactly the same mathematical form as the original density. Another density of exactly the same mathematical form is $(1/\sigma) f(y/\sigma)$.
- Objective priors are determined by the mathematical form of the density so it would seem that the objective prior for θ should be the same function as that for the objective prior for θ^* (or σ).

This logic, when combined with 'invariance to parameterization,' leads to what are called 'invariance priors.'

The development of invariance priors is achieved through the mathematical tools of *group theory*. In the above example, the transformation 'multiplication by positive constants' is called the *scale group* or *multiplication group*. We do not delve deeply into group theory here and, hence, only present invariance priors in simple cases. For an introduction to the use of group theory in developing invariance priors, see [2].

Curiously, the development of invariance priors initially mirrors the success and failure of the Jeffreys-rule prior. For one-parameter problems having a group-invariance structure, the invariance prior is the Jeffreys-rule prior (also the reference prior and first order matching prior), so all is well. In multi-parameter problems, however, the natural invariance prior is what is called the *left Haar prior*[1], and is problematical in much the same ways that the Jeffreys-rule prior is problematical in multi-parameter problems.

Unlike the situation with the Jeffreys-rule prior, however, a 'fix' was developed for the invariance prior in multi-parameter problems; instead of using the left Haar prior, use what is called the *right Haar prior*. This development was actually a frequentist development, finding the prior distribution in invariant problems that is exact frequentist matching; it is the right-Haar prior that achieves this.

Example 0.16 A model of the form $p(x \mid \mu, \sigma) = (1/\sigma) f((x-\mu)/\sigma)$ is called a location-scale model. (Were this the normal model, σ would be the standard deviation, which is differently parameterized than using the variance.) It can be shown for such a model that the left-Haar prior is $\pi^l(\mu, \sigma) = \sigma^{-2}$ (the Jeffreys-rule prior), while the right-Haar prior is $\pi^r(\mu, \sigma) = \sigma^{-1}$ (the independence-Jeffreys prior); it is the latter that results in Bayesian credible sets with correct frequentist coverage (e.g., the left-Haar prior gave the wrong degrees of freedom in the normal case).

As a specific example, suppose the $(x_1, \ldots, x_n)$ are iid observations

[1]The Haar measures, from which the Haar priors are derived, were introduced in the 1930's by the Hungarian mathematician Alfred Haar (*cf.* [31])

from the Cauchy density $p(x \mid \mu, \sigma) = 1/(\pi\sigma[1 + (x-\mu)^2/\sigma^2])$ and that it is desired to find a 90% frequentist confidence set for $\xi = \mu + \sigma$, which is the third quartile of the distribution. The natural Bayesian credible set for ξ is the equal-tailed credible set, given by $C(\boldsymbol{x}) = (l(\boldsymbol{x}), u(\boldsymbol{x}))$, where $\pi^r(\mu, \sigma \mid \boldsymbol{x})$ is the posterior distribution arising from $\pi^r(\mu, \sigma)$,

$$\int_{\{(\mu,\sigma):\mu+\sigma<l(\boldsymbol{x})\}} \pi^r(\mu, \sigma \mid \boldsymbol{x})\, d\mu d\sigma = \int_{\{(\mu,\sigma):\mu+\sigma>u(\boldsymbol{x})\}} \pi^r(\mu, \sigma \mid \boldsymbol{x})\, d\mu d\sigma = 0.05\,.$$

$C(x)$ is guaranteed to be a 90% frequentist confidence set. Furthermore it is a 90% frequentist confidence set conditional on the (ancillary) statistic $s = (\frac{x_2-x_1}{x_1}, \ldots, \frac{x_n-x_1}{x_1})$; as discussed in Section 0.1.1, it is important to find confidence sets that condition properly on relevant statistics, such as s. Trying to develop conditional frequentist confidence procedures directly from frequentist reasoning would be virtually impossible here.

A simple version (without conditioning on $\boldsymbol{s}$) of the exact matching property of the right-Haar prior can be found in [2]. A more general version can be found in [15]. The whole class of such theorems is sometimes called the Hunt-Stein theorem, which was developed in [24] for testing.

0.3 Reference Priors

We have not yet presented a general method for developing good objective priors in multi-parameter problems. The Jeffreys-rule prior in multi-parameter problems simply seems to be unsuitable. Matching priors and invariance priors in multiparameter problems are typically suitable, but they are often unavailable (e.g., the solutions to the system of differential equations may not be available, or the model may not have a suitable group invariance structure). In this section, we describe a method for deriving an objective prior in multi-parameter problems which seems to be essentially always effective.

0.3.1 Introduction to the Reference Prior Approach

In the matching prior approach for multi-parameter models, we saw that different parameters in the model could have different matching priors. This is also true of the reference prior approach. A reference prior will be a prior distribution on the full parameter space, but it will be developed to focus on the parameter of interest. If there is more than one parameter of interest in a problem, there could thus be more than one reference prior.

The information to be expected from one observation from model

$\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \Theta\}$, when the proper prior for $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta})$, is commonly taken to be the Kullback-Liebler divergence between the prior distribution and the posterior distribution, which is given by

$$I\{\pi \,|\, \mathcal{M}\} = \int \int_{\boldsymbol{\mathcal{X}} \times \Theta} p(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{x})}{\pi(\boldsymbol{\theta})} \, d\boldsymbol{x} \, d\boldsymbol{\theta} \,. \tag{21}$$

The sharper the prior, the more information it contains, reducing the information expected from the data; in the extreme, a point mass prior provides complete information about $\boldsymbol{\theta}$, so the data can add no information. The prior that maximizes $I\{\pi \,|\, \mathcal{M}\}$ can thus claim to be that prior which maximizes the amount of information provided by the data and is, hence, a natural candidate for an objective prior.

Unfortunately, the resulting prior turns out to be problematical, in that it will typically be a discrete prior that depends on the sample size, even when $\boldsymbol{\theta}$ is a continuous parameter ([6]). This makes it unappealing in practice.

An alternative, proposed in [14], is to apply this idea asymptotically, considering k independent replications of $\mathcal{M}$, resulting in the sequence of realizations $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ and the ensuing replicated model $\mathcal{M}^k$. Then $I\{\pi \,|\, \mathcal{M}^k\}$ is the amount of information in $\boldsymbol{x}^{(k)}$. As $k \to \infty$, the posterior distribution will typically converge to a point mass, providing complete information about $\boldsymbol{\theta}$. Thus, intuitively, $\lim_{k \to \infty} I\{\pi \,|\, \mathcal{M}^k\}$ provides a measure of the missing information about θ associated to the prior $\pi(\theta)$. Maximizing this missing information is thus another natural way to define an objective prior and is the basis of the *reference prior* approach that we will be discussing.

0.3.2 The Reference Prior for a Real Parameter

Consider first the situation in which the model has only one real continuous parameter θ. The following theorem, from [7] in which the proof can be found, provides an explicit expression for the reference prior for θ (i.e., that prior which maximizes the asymptotic missing information), under mild conditions. Recall that $\boldsymbol{x}$ refers to the entire vector of observations from the model and $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ refers to a vector of (artificial) independent replicates of these vector observations from the model. It is often more convenient (and equivalent) to work with sufficient statistics $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)}) \in \boldsymbol{\mathcal{T}}_k$ for the replicated observations, so the theorem will be stated in terms of such sufficient statistics.

Proposition 0.1 Explicit form of the reference prior. *Consider a standard model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta \subset \mathbb{R}\}$. *Let* $\pi^*(\theta)$ *be a continuous strictly positive function such that the corresponding formal posterior*

$$\pi^*(\theta \,|\, \boldsymbol{t}_k) = \frac{p(\boldsymbol{t}_k \,|\, \theta) \, \pi^*(\theta)}{\int_\Theta p(\boldsymbol{t}_k \,|\, \theta) \, \pi^*(\theta) \, d\theta} \tag{22}$$

is proper and asymptotically consistent and, for any interior point $\theta_0 \in \Theta$*, define*

$$f_k(\theta) = \exp\left\{\int_{\boldsymbol{\mathcal{T}}_k} p(\boldsymbol{t}_k \,|\, \theta)\, \log\left[\pi^*(\theta \,|\, \boldsymbol{t}_k)\right] d\boldsymbol{t}_k\right\} \quad \textit{and} \tag{23}$$

$$f(\theta) = \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)} \ . \tag{24}$$

Suppose $f_k(\theta)$ *is continuous and, for any fixed* θ *and sufficiently large* k*,* $\{f_k^0(\theta)/f_k^0(\theta_0)\}$ *is either monotonic in* k *or bounded above by some* $h(\theta)$ *which is integrable on any compact set. Then* $\pi^R(\theta) = f(\theta)$ *is the* reference prior*.*

The choice of π^* is essentially arbitrary and can be chosen for computational convenience, and the choice of θ_0 is immaterial.

Example 0.17 *(Uniform data, continued.)* Suppose x is one observation from a uniform distribution $\{p(x \,|\, \theta) = \theta^{-1},\, 0 < x < \theta,\, \theta > 0\}$. Here the sampling space is $\mathcal{X} = [0, \theta]$, which depends on the parameter θ, so that Fisher information is not defined and, hence, there is no Jeffreys-rule prior.

The i.i.d. replication data is the vector $\boldsymbol{x}^{(k)} = \{x_1, \ldots, x_k\}$ and using (for mathematical convenience) $\pi^*(\theta) = \theta^{-1}$, the corresponding posterior distribution is the Pareto density

$$\pi^*(\theta \,|\, t_k) = \mathrm{Pa}(\theta \,|\, k, t_k) = k\, \frac{t_k^k}{\theta^{k+1}} \ , \quad \theta > t_k, \tag{25}$$

where $t_k = \max\{x_1, \ldots, x_k\}$ is a sufficient statistic. The sampling distribution of t_k is the inverted pareto density

$$p(t_k \,|\, \theta) = \mathrm{Ip}(t_k \,|\, k, \theta^{-1}) = k\, \frac{t^{k-1}}{\theta^k} \ , \quad 0 < t_k < \theta. \tag{26}$$

Computation yields

$$f_k(\theta) = \exp\left[\int_0^\theta \mathrm{Ip}(t_k \,|\, k, \theta)\, \log\big(\mathrm{Pa}(\theta \,|\, k, t_k)\big)\, dt_k\right] = \frac{c(k)}{\theta} \ , \tag{27}$$

where $c(k) = k\, e^{-(1+1/k)}$. It follows that, for all k (not just in the limit), $f_k^0(\theta)/f_k^0(\theta_0) = \theta_0/\theta$. Thus $\pi^R(\theta) = \theta^{-1}$ is the reference prior.

This is an excellent prior; for instance, it is exact frequentist matching. Indeed, using Bayes theorem, the corresponding reference posterior for an i.i.d. sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ is the Pareto distribution $\mathrm{Pa}(\theta \,|\, n, t_n)$, having density

$$\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t) = \begin{cases} n\, t^n \theta^{-(n+1)} & \text{if } \theta \geq t \\ 0 & \text{otherwise} \end{cases} \ .$$

A one-sided posterior credible set, at level $1-\alpha$, can be shown to be the interval $(t_n, t_n\,\alpha^{-1/n})$ and, using (26), the frequentist coverage of this, as a confidence procedure, can easily be shown to be $1-\alpha$. Interestingly, this interval has width $t_n(\alpha^{-1/n}-1) = t_n(\exp\{-\frac{1}{n}\log(\alpha)\}-1) \approx \frac{-\log(\alpha)}{n}$ for large n, showing that this model has non-standard asymptotics. (Regular models have confidence intervals of width proportional to $C/\sqrt{n}$ for some constant.) This is, thus, an example of how objective Bayesian analysis automatically produces answers with the correct non-regular asymptotics.

0.3.3 Multiple Continuous Parameters

Applying the idea of maximizing missing information works wonderfully if θ is one-dimensional but fails – if utilized directly – when $\boldsymbol{\theta}$ is multi-dimensional; indeed, the result will often be the multivariate Jeffreys-rule prior, which was seen in Section 0.2.2.2 to be highly problematical. The third key component of the reference prior approach is thus to apply the idea of maximizing missing information sequentially.

To see the idea, suppose that θ_1 is the one-dimensional parameter of interest and that the remaining parameters, $\boldsymbol{\theta}^2 = (\theta_2, \ldots, \theta_m)$, are viewed as nuisance parameters. Suppose we were told that the 'correct' prior for the nuisance parameter is $\pi(\boldsymbol{\theta}^2 \,|\, \theta_1)$ (the prior being allowed to depend on θ_1 and assumed here to be proper). Then $\boldsymbol{\theta}^2$ could be integrated out of the model, obtaining the (proper) marginal model

$$\mathcal{M}^* \equiv \{p(\boldsymbol{x} \,|\, \theta_1),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta_1 \in \Theta_1\},\ p(\boldsymbol{x} \,|\, \theta_1) = \int_{\boldsymbol{\Theta}^2} p(\boldsymbol{x} \,|\, \theta_1, \boldsymbol{\theta}^2)\, \pi(\boldsymbol{\theta}^2 \,|\, \theta_1)\, d\boldsymbol{\theta}^2\,.$$

As this marginal model only depends on θ_1, the method of maximizing missing information can be applied (using, say, the implementation in the previous section) to obtain the (marginal) reference prior $\pi^R(\theta_1)$. The final overall reference prior would then simply be $\pi^R(\theta_1, \boldsymbol{\theta}^2) = \pi(\boldsymbol{\theta}^2 \,|\, \theta_1)\pi^R(\theta_1)$.

If the conditional prior $\pi(\boldsymbol{\theta}^2 \,|\, \theta_1)$ is not known, one can condition on θ_1, and then find the reference priors $\pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1)$ for $\boldsymbol{\theta}^2$, given each θ_1. If $\boldsymbol{\theta}^2$ is one-dimensional, these can be determined as in the previous section. In the general case, these can be determined in the regular situation where posterior distributions are asympotically normal; the result is given in Proposition 0.2. If the $\pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1)$ are proper, one can proceed as above to form the marginal model and find the reference prior for θ_1 in the marginal model.

Unfortunately, the conditional reference priors $\pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1)$ are typically not proper, in which case $p(\boldsymbol{x} \,|\, \theta_1)$ is not proper and so cannot be used to determine $\pi^R(\theta_1)$. To overcome this difficulty one must operate, as in Section 0.1.3, on compact subsets of the parameter space which increase to the full space, taking appropriate limits of the constrained reference priors to define the reference prior on the full space. This is illustrated in Proposition 0.2, whose proof can be found in [5].

Proposition 0.2 (Two group reference prior under asymptotic normality.) *The k replications result in model* $\mathcal{M} \equiv \{p(\boldsymbol{x}^{(k)} \,|\, \theta_1, \boldsymbol{\theta}^2), \boldsymbol{x}^{(k)} \in \boldsymbol{\mathcal{X}}^k, \, (\theta_1, \boldsymbol{\theta}^2) \in \Theta_1 \times \boldsymbol{\Theta}^2\}$. *Let* $I(\theta_1, \boldsymbol{\theta}^2)$ *denote the Fisher information for a single* $\boldsymbol{x}$. *Define* $V(\theta_1, \boldsymbol{\theta}^2) = I(\theta_1, \boldsymbol{\theta}^2)^{-1}$ *and, with* $I_{11}(\theta_1, \boldsymbol{\theta}^2)$ *and* $V_{11}(\theta_1, \boldsymbol{\theta}^2)$ *being scalar,*

$$I(\theta_1, \boldsymbol{\theta}^2) = \begin{pmatrix} I_{11}(\theta_1, \boldsymbol{\theta}^2) & I_{12}(\theta_1, \boldsymbol{\theta}^2) \\ I_{21}(\theta_1, \boldsymbol{\theta}^2) & I_{22}(\theta_1, \boldsymbol{\theta}^2) \end{pmatrix}, V(\theta_1, \boldsymbol{\theta}^2) = \begin{pmatrix} V_{11}(\theta_1, \boldsymbol{\theta}^2) & V_{12}(\theta_1, \boldsymbol{\theta}^2) \\ V_{21}(\theta_1, \boldsymbol{\theta}^2) & V_{22}(\theta_1, \boldsymbol{\theta}^2) \end{pmatrix}.$$

Under asymptotic normality, the reference prior when θ_1 *is the parameter of interest is defined as follows:*

Case 1. *If* $c(\theta_1) = \int_{\boldsymbol{\Theta}^2} \sqrt{I_{22}(\theta_1, \boldsymbol{\theta}^2)}\, d\boldsymbol{\theta}^2 < \infty$, *let*

$$\pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1) = \sqrt{I_{22}(\theta_1, \boldsymbol{\theta}^2)}/c(\theta_1), \; \pi^R(\theta_1) \propto \exp\left\{ \int_{\boldsymbol{\Theta}^2} \pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1) \, \log[V_{11}^{-1/2}(\theta_1, \boldsymbol{\theta}^2)] \, d\boldsymbol{\theta}^2 \right\}. \tag{28}$$

Then the reference prior is $\pi^R(\theta_1, \boldsymbol{\theta}^2) = \pi^R(\boldsymbol{\theta}^2 \,|\, \theta_1)\pi^R(\theta_1)$.

Case 2. *If* $c(\theta_1)$ *is infinite, constrain* $\boldsymbol{\theta}^2$ *to compact subsets* $\{\boldsymbol{\Theta}_l^2, \; l = 1, 2, \dots\}$ *of* $\boldsymbol{\Theta}^2$, *such that* $\lim_{l\to\infty} \boldsymbol{\Theta}_l^2 = \boldsymbol{\Theta}^2$, *and define*

$$\begin{aligned} \pi_l(\boldsymbol{\theta}^2 \,|\, \theta_1) &= \frac{1}{c_l(\theta_1)} \sqrt{I_{22}(\theta_1, \boldsymbol{\theta}^2)} 1_{\boldsymbol{\Theta}_l^2}(\boldsymbol{\theta}^2), \; c_l(\theta_1) = \int_{\boldsymbol{\Theta}_l^2} \sqrt{I_{22}(\theta_1, \boldsymbol{\theta}^2)}\, d\boldsymbol{\theta}^2, \\ \pi_l(\theta_1) &\propto \exp\left\{ \int_{\boldsymbol{\Theta}_l^2} \pi_l(\boldsymbol{\theta}^2 \,|\, \theta_1) \, \log[V_{11}^{-1/2}(\theta_1, \boldsymbol{\theta}^2)] \, d\boldsymbol{\theta}^2 \right\}, \\ \pi^R(\theta_1, \boldsymbol{\theta}^2) &= \lim_{l\to\infty} \frac{\pi_l(\boldsymbol{\theta}^2 \,|\, \theta_1)\pi_l(\theta_1)}{\pi_l(\boldsymbol{\theta}_0^2 \,|\, \theta_{10})\pi_l(\theta_{10})}, \end{aligned}$$

for any interior point $(\theta_{10}, \boldsymbol{\theta}_0^2)$. *Then* $\pi^R(\theta_1, \boldsymbol{\theta}^2)$ *is the reference prior.*

Example 0.18 *Multinomial problem (continued).* For the multinomial problem, computations in [4] yield (this is a Case 1 scenario)

$$\pi^R(\theta_1, \boldsymbol{\theta}^2) \propto \left(\prod_{i=1}^m \theta_i^{-1/2} \right) (1 - \theta_1)^{-(m-1)/2} \left(1 - \sum_{i=1}^m \theta_i \right)^{-1/2}. \tag{29}$$

The marginal reference posterior for θ_1 from π^R can be shown to be

$$\pi^R(\theta_1 \mid \boldsymbol{x}) = \mathrm{Be}(\theta_i \,|\, x_i + \tfrac{1}{2}, n - x_i + \tfrac{1}{2}). \tag{30}$$

There is a more general reference prior method called the *one-at-a-time reference prior algorithm* ([4]). In this algorithm one first finds the reference prior $\pi^R(\theta_m \mid \theta_1, \dots, \theta_{m-1})$, forms the marginal model by integrating over θ_m, as was done over $\boldsymbol{\theta}^2$ in (28) (operating on compact sets if necessary), then finds the reference prior $\pi^R(\theta_{m-1} \mid \theta_1, \dots, \theta_{m-2})$, repeating the process until finally obtaining $\pi^R(\theta_1)$, with the appropriate product of all of these

being the reference prior. This is, indeed, our recommended reference prior algorithm, since it is based on one-dimensional information matrices, which all perspectives suggest are optimal.

This can be difficult to implement, however, and the approach discussed above – treating the parameter of interest separately, while grouping all the nuisance parameters together in a single step – has appeal as something much simpler to implement. It is, of interest, in this regard, that application of the one-at-a-time reference prior algorithm to the multinomial problem was shown in [4] to yield a reference prior that differed from (29), but resulted in the same marginal posterior (30) for the parameter of interest; as this is what is needed for inference about the parameter of interest, the cruder method still gave the optimal result.

0.4 Overall Objective Priors

In single parameter problems, the reference prior (usually the Jeffreys-rule prior) is uniquely defined but, in multiparameter models, we have seen that the reference prior often changes, depending on the quantity of interest. There are, however, situations where one is *simultaneously* interested in all the parameters of the model or, at least in several of them; in such situations having a single overall prior for use is appealing. In prediction or decision analysis for instance, all of the parameters of the model are typically relevant and often none are of major interest individually.

Another situation in which having an overall prior would be beneficial is when there is a non-standard quantity of interest (*e.g.*, a non-standard function of the model parameters), and formal derivation of the reference prior is not tenable. Computation can also be a consideration, since having to separately do Bayesian computations with different reference priors for each parameter can be challenging. Finally, when dealing with non-specialists, it may be best to just present them with one overall objective prior, rather than attempting to explain the reasons for preferring different priors for different quantities of interest.

0.4.1 Preliminaries

We have already encountered objective Bayes approaches that produce an overall prior, namely use of the constant prior, use of right-Haar priors in invariant situations, and use of the Jeffreys-rule prior. We extensively discussed the limitations of the constant prior and the Jeffreys-rule prior, and having a suitably invariant model is rather special, so there is more to be done.

The approach taken in [8] to find an overall prior is to find a prior which

yields marginal posteriors for each parameter that are close to the reference prior marginal posteriors for each parameter. To be more precise, assume that $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ is is the set of $m > 1$ parameters that are of interest. The goal is to find a prior, $\pi^o(\boldsymbol{\theta})$, whose corresponding marginal posteriors, $\{\pi^o(\theta_i \,|\, \boldsymbol{x}), i = 1, \ldots m\}$, are close to the reference prior marginal posteriors $\pi^R_{\theta_i}(\theta_i \,|\, \boldsymbol{x})$ (the subscript θ_i indicating that this arose from the reference prior when θ_i is the parameter of interest).

Example 0.19 *Multinomial data.* Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is multinomial $\mathrm{Mu}(\boldsymbol{x} \,|\, n; \theta_1, \ldots, \theta_m)$, with $\sum_{i=1}^m x_i = n$, and $\sum_{i=1}^m \theta_i = 1$. In Example 0.18, it was shown that the reference prior, when the parameter of interest is θ_i, is a different prior for each θ_i, and that the reference prior for θ_i results in a Beta reference marginal posterior $\pi^R_{\theta_i}(\theta_i \,|\, \boldsymbol{x}) = \mathrm{Be}(\theta_i \,|\, x_i + \frac{1}{2}, n - x_i + \frac{1}{2})$.

It is natural here to consider the Dirichlet $\{\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a) : a \in (0, \infty)\}$ class of priors. The choice of a for which the resulting marginal posteriors are closest to the reference marginal posteriors is shown in [8] to be approximately $a = 1/m$. Thus the recommended overall objective prior would be $\pi^o(\boldsymbol{\theta}) = \mathrm{Di}(\boldsymbol{\theta} \,|\, \frac{1}{m}, \ldots, \frac{1}{m})$.

Note that this overall prior overcomes the problem with the Jeffreys-rule prior that was discussed in Example 13; the Jeffreys-rule prior, corresponding to $a = 1/2$, was shown to yield unreasonable posterior means for the θ_i in the situation where $n = 3$, $m = 1000$, $x_{240} = 2$, $x_{876} = 1$, and the other x_i are zero. In particular, the posterior mean for θ_{240} was only 0.005 even though 2/3 of the observations were in that cell. The problem is that the Jeffreys-rule prior effectively adds 1/2 to each cell, so that the cells with $x_i = 0$ overwhelmed the cells with $x_i \neq 0$. In contrast, $\pi^o(\boldsymbol{\theta})$ only adds $1/m = 0.001$ to each cell, 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.

We will discuss various possible approaches to the development of an overall prior in this section. Most of the approaches and examples are taken from either [8] or [12].

0.4.2 When the Reference Prior is Common

If one is able to find a single joint prior $\pi^o(\boldsymbol{\theta})$ whose corresponding marginal posteriors are precisely equal to the reference posteriors for each of the θ_i's, so that, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\pi^o(\theta_i \,|\, \boldsymbol{x}) = \pi^R_{\theta_i}(\theta_i \,|\, \boldsymbol{x})$, $i = 1, \ldots, m$, then it is natural to argue that this should be an appropriate solution to the problem. The

$\mathrm{N}(x_i \mid \mu, \sigma)$ model has already been mentioned as an example, where the reference priors when μ or σ are the parameters of interest are the same. (Their common reference prior is also excellent if μ and σ are jointly of interest, *e.g.*, if a joint credible set for them is sought.)

0.4.3 Hierarchical Reference Approach

The hierarchical reference prior approach consists of the following steps.

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

(ii) Form the marginal density for the hyperparmeter $\boldsymbol{a}$,

$$p(\boldsymbol{x} \,|\, \boldsymbol{a}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{a})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{a})\; d\boldsymbol{\theta}\,,$$

which is proper because the $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{a})$ are proper.

(iii) Find the reference prior, $\pi^R(\boldsymbol{a})$, for $\boldsymbol{a}$ in this marginal model.

(iv) Use, as the overall prior for $\boldsymbol{\theta}$,

$$\pi^o(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} \,|\, \boldsymbol{a})\, \pi^R(\boldsymbol{a})\, d\boldsymbol{a}\,. \tag{31}$$

Note that computation with this prior is often easiest if one keeps $\boldsymbol{a}$ as a random variable, rather than integrating it out and working directly with (31). This is especially true when $\{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}\}$ is a conjugate family to the data distribution, so that Gibbs sampling can be employed.

0.4.3.1 Application to the Multinomial Distribution

The Dirichlet $\{\mathrm{Di}(\boldsymbol{\theta} \,|\, a, \ldots, a) : a \in (0, \infty)\}$ class of hierarchical priors for $\boldsymbol{\theta}$ is natural here, reflecting a desire to treat all the θ_i similarly. The marginal density is

$$\begin{aligned} p(\boldsymbol{x} \,|\, a) &= \int \binom{n}{x_1 \ldots x_m} \left(\prod_{i=1}^m \theta_i^{x_i}\right) \frac{\Gamma(m\,a)}{\Gamma(a)^m} \prod_{i=1}^m \theta_i^{a-1} d\boldsymbol{\theta} \\ &= \binom{n}{x_1 \ldots x_m} \frac{\Gamma(m\,a)}{\Gamma(a)^m} \frac{\prod_{i=1}^m \Gamma(x_i + a)}{\Gamma(n + m\,a)}\,. \end{aligned} \tag{32}$$

This can be shown to be a regular, one-parameter model, so the reference prior, $\pi^R(a)$, for the hyper parameter a, is the Jeffreys prior for this marginal model. In [8], this was shown to be given by

$$\pi^R(a) \propto \left[\sum_{j=0}^{n-1} \left(\frac{Q(j \,|\, a, m, n)}{(a+j)^2} - \frac{m}{(m\,a+j)^2}\right)\right]^{1/2}, \tag{33}$$

where $Q(\cdot \,|\, a, m, n)$, for $j = 0, \ldots, n-1$, is given by

$$Q(j \,|\, a, m, n) = \sum_{l=j+1}^{n} \binom{n}{l} \frac{\Gamma(l+a)\,\Gamma(n-l+(m-1)a)\,\Gamma(m\,a)}{\Gamma(a)\,\Gamma((m-1)a)\,\Gamma(n+m\,a)}\,.$$

Finally, the overall prior for $\boldsymbol{\theta}$ is

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

This prior turns out to be a proper distribution, which is rather unusual for an objective prior on an unbounded space. It turns out that the marginal likelihoood of a does not go to zero as a grows ([8]), so that the prior must be proper for the posterior to exist. It is a rather amazing property of reference/Jeffreys priors that they seem to always be proper when the likelihood does not go to zero.

0.4.4 Hierarchical Normal Models

Hierarchical normal models are the basis of much of applied Bayesian statistics, yet there is uncertainty as to which priors to use for hyperparameters (parameters at higher levels of the hierarchical model). The long history of efforts to develop objective hyperpriors is reviewed in [28] and [12], from which this section is primarily based.

Once the difficulties of using the Jefferys-rule prior in multi-parameter models and hierarchical models became recognized, it became common to use a constant prior for higher level variances or covariances, but the constant prior is much too diffuse, requiring twice as many observations to obtain posterior propriety as is logically needed (see [12], which also showed that the constant prior has markedly poor performance in risk simulations).

[10] and [11] approached the question of choice of hyperpriors in normal hierarchical models by looking at the frequentist notion of admissibility of resulting estimators; an estimator is admissible if it cannot be improved upon in terms of mean squared error and inadmissible if it can be improved. Hyperpriors that are too diffuse result in inadmissible estimators, while hyperpriors that are concentrated enough result in admissible estimators. Hyperpriors that are 'on the boundary of admissibility' thus seem to be sensible choices for objective priors, being as diffuse as possible without resulting in inadmissible procedures. This approach was used in [12] to develop the recommended hyperpriors presented here for any normal hierarchical model.

0.4.4.1 The Hierarchical Model Considered

Consider the following two-stage hierarchical model. Suppose that, independently and for $i = 1, \ldots m$,

$$\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_i \sim \mathrm{N}_k(\cdot \,|\, \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) \quad \text{and} \quad \boldsymbol{\theta}_i \,|\, \boldsymbol{\beta}, \boldsymbol{V} \sim \mathrm{N}_k(\cdot \,|\, \boldsymbol{Z}_i\boldsymbol{\beta}, \boldsymbol{V})\,, \tag{34}$$

where the $\boldsymbol{x}_i$ are the $k \times 1$ observation vectors; the $\boldsymbol{\Sigma}_i$ are *known* $k \times k$ full rank covariance matrices; the $\boldsymbol{\theta}_i$ are $k \times 1$ unknown mean vectors; $\boldsymbol{\beta}$ is an $l \times 1$ unknown "hyper-mean" vector; the $\boldsymbol{Z}_i$ are $k \times l$ known matrices of covariates of full rank; and $\boldsymbol{V}$ is an unknown $k \times k$ "hyper-covariance matrix." If the $\boldsymbol{\Sigma}_i$ are not known, they either need to be estimated from data and plugged in (resulting in a partly empirical Bayes approach) or incorporated into a larger hierarchical Bayesian analysis using, say, the reference prior for a first level covariance matrix, given in [34].

Example 0.20 (Following the path-breaking paper [21]:) Suppose $k = 1$ and, for $i = 1, \ldots, m$, x_i is the batting average, after one month, of baseball player i (after an arcsin transformation of the original binomial random variable to normality), with $\sigma_i^2 = 1/[4n_i]$ and n_i being the number of at-bats the player had in the first month. The goal is to estimate θ_i, the 'true' batting ability of the player, as would be determined (say) by batting average at the end of the year, using the $x_i \sim \mathrm{N}(x_i \mid \theta_i, \sigma_i^2)$.

The (arcsin transformed) batting averages, b_i, of the players are known from the previous year and the θ_i are modeled by regression on these previous batting averages, leading to the model $\theta_i \sim \mathrm{N}(\cdot \mid \beta_1 + b_i\beta_2, V)$. In the above notation, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ (the unknown regression coefficients), $\boldsymbol{Z}_i = (1, b_i)$ are the covariates, and V is the unknown variance of the regression. To proceed, an objective prior distribution is needed for $(\boldsymbol{\beta}, V)$.

Example 0.21 At hospital i, the observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ are the sample averages of the costs of k different medical treatments; $\boldsymbol{\theta}_i$ is the corresponding unknown vector of true mean costs of the treatments at the hospital; and $\boldsymbol{\Sigma}_i$ is the covariance matrix associated with these sample averages.

Consider the j^{th} coordinate, θ_{ij}, of each θ_i; this refers to the cost of the j^{th} medical treatment at the i^{th} hospital. It is natural to model this as depending on p hospital characteristics, such as the number of patients receiving the treatment, the average severity of the condition of the patients for the treatments, the average income of the patients, etc. Denoting these characteristics for the j^{th} treatment at the i^{th} hospital as $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijp})$, a reasonable model would be a regression model

$$\theta_{ij} = \boldsymbol{z}_{ij}\boldsymbol{\alpha}_j + \epsilon_{ij}, \quad \text{for } i = 1, \ldots, m; \quad j = 1, \ldots, k\,,$$

where $\boldsymbol{\alpha}_j$ is a $p \times 1$ column vector of weights (specific to each treatment but assumed to be constant across hospitals) determining the effect of hospital characteristics on the cost of treatment j, and ϵ_{ij} is normal error. There would typically be a separate regression of this form for each treatment j, and stacking these regressions vertically leads to equation (34), where

$l = kp$,

$$\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{z}_{i1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}_{i2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{z}_{ik} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{pmatrix}. \tag{35}$$

If it were thought that each of the treatment cost means were independent, $\boldsymbol{V}$ could be taken as diagonal, but it is far more likely that there is considerable dependence, so that completely unknown $\boldsymbol{V}$ is reasonable.

In this example, there could be a third level of hierarchial modeling,

$$\boldsymbol{\alpha}_i \sim \mathrm{N}_p(\cdot \mid \boldsymbol{\xi}, \boldsymbol{\Omega}) \,. \tag{36}$$

Whether or not this is appropriate depends on the precise context, but we will see that the analysis could still be done for this three-level hierarchical model with the methodology proposed herein.

Remark 0.1 Often the data vectors themselves arise from linear models $\boldsymbol{y}_i \sim N_r(\cdot \mid \boldsymbol{B}_i\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i)$, where the $\boldsymbol{B}_i$ are known design matrices of covariates. In this situation, one can transform to the least squares estimates for $\boldsymbol{\theta}_i$,

$$\boldsymbol{x}_i = (\boldsymbol{B}_i^t\boldsymbol{\Lambda}_i^{-1}\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^t\boldsymbol{\Lambda}_i^{-1}\boldsymbol{y}_i \,,$$

which will be distributed as in (34), with covariance matrix $\boldsymbol{\Sigma}_i = (\boldsymbol{B}_i^t\boldsymbol{\Lambda}_i^{-1}\boldsymbol{B}_i)^{-1}$.

0.4.4.2 The Recommended Prior

In [12], it is recommended to utilize independent priors for the unknown hyperparameters $\boldsymbol{\beta}$ and $\boldsymbol{V}$, i.e., $\pi(\boldsymbol{\beta}, \boldsymbol{V}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{V})$, with the following choices:

$$\pi(\boldsymbol{\beta}) \propto \frac{1}{[1 + ||\boldsymbol{\beta}||^2]^{(l-1)/2}} \,, \quad \pi(\boldsymbol{V}) = \frac{1}{|\boldsymbol{V}|^{1-1/(2k)} \prod_{i<j}(d_i - d_j)} \,, \tag{37}$$

where $||\boldsymbol{\beta}||^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$ and $d_1 > d_2 \ldots > d_k$ are the ordered eigenvalues of $\boldsymbol{V}$. For $k = 1$, the second prior is defined as $\pi(V) = 1\sqrt{V}$. Extensive justification of these choices is given in [12].

One of the most important properties of this prior is that it can be used at any level of a hierarchical model for the mean parameters and the covariance matrix. This is not so for many other proposed priors such as 'marginal Jeffreys priors' (see [12]). That the resulting posterior distributions are guaranteed to be proper is shown in [12].

For the priors in (37) to be reasonable, one must deal with the fact that the covariates (defined by the columns of $\boldsymbol{Z}_i$) often use different units of

measurement, units which can be arbitrary, and objective priors should be invariant to the choice of unit of measurement. The easiest way to achieve this is to first rescale the covariates so they are 'unitless' and then proceed with use of the priors in (37). Here are two common methods of such re-scaling.

Method 1 - Standard Deviation Rescaling: Define c_j to be the sample standard deviation of all non-zero entries in the jth columns of all the $\boldsymbol{Z}_i$, and rescale to obtain new covariates $\boldsymbol{z}^*_{ij} = \boldsymbol{z}_{ij}/c_j$.

Method 2 - Max Coordinate Rescaling: Define c_j to be the maximum absolute value of all entries in the jth columns of all the $\boldsymbol{Z}_i$, and rescale as above. (This puts everything on the interval from -1 to 1.)

Remark 0.2 If covariate j is rescaled by c_j, for all j, the new hyper-mean becomes $\boldsymbol{\beta}^* = (c_1\beta_1, \ldots, c_l\beta_l)'$. Then, if the rescaled covariates are used in Bayesian analysis, the analysis will yield the posterior for $\boldsymbol{\beta}^*$, which must then be transformed back to the posterior for $\boldsymbol{\beta}$, if the latter is desired. If only the posterior distributions of the $\boldsymbol{\theta}_i$ are desired, the analysis carried out in the rescaled space is fine.

0.5 Conclusions

Objective Bayesian inference has been a central part of statistics for over 250 years and this has (necessarily) been a very brief introduction to the area. For a much more extensive review of objective Bayesian inference, see [9] and the over 1500 references cited therein. Study of the interface between the objective Bayesian approach and the frequentist approach is much more recent (necessarily, because the frequentist approach is only about 100 years old). For a more recent study of this interface, see [3] and the accompanying discussions.

0.6 Acknowledgments

James Berger was supported by the U.S. National Science Foundation grant DMS 1821289.

Bibliography

[1] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London*, 53 and 54:370–416 and 370–416. Reprinted in *Biometrika*, **45** (1958), 293–315, with a biographical note by G. A. Barnard. Reproduced in Press (1989), 185–217, 1763.

[2] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer, Second edition, 1985.

[3] James O. Berger. Four types of frequentism and their interplay with Bayesianism (with discussion). *The New England Journal of Statistics in Data Science*, 1, 2022; https://doi.org/10.51387/22-NEJSDS4.

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

[5] James O. Berger and José M. Bernardo. Ordered group reference priors, with applications to multinomial problems. *Biometrika*, 79:25–37, 1992.

[6] James O. Berger, José M. Bernardo, and Manuel Mendoza. On priors that maximize expected information. In Recent Developments in Statistics and their Applications *(J. P. Klein and J. C. Lee, eds)*, pages 1–20. Seoul: Freedom Academy, 1989.

[7] James O. Berger, José M. Bernardo, and Dongchu Sun. The formal definition of reference priors. *Ann. Statist.*, 37:905–938, 2009.

[8] James O. Berger, José M. Bernardo, and Dongchu Sun. Overall objective priors. *Bayesian Analysis*, 10:189–221 (with discussion), 2015.

[9] James O. Berger, Jose M. Bernardo, and Dongchu Sun. *Objective Bayesian Inference*. World Scientific Publishing Company, to appear.

[10] James O. Berger and William E. Strawderman. Choice of hierarchical priors: Admissibility in estimation of normal means. *Ann. Statist.*, 24:931–951, 1996.

[11] James O. Berger, William E. Strawderman, and Dejun Tang. Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Ann. Statist.*, 33:606–646, 2005.

[12] James O. Berger, Dongchu Sun, and Chengyuan Song. An objective prior for hyperparameters in normal hierarchical models. *J. Multivariate Analysis*, 178:104606, 2020.

[13] James O. Berger and Robert L. Wolpert. *The Likelihood Principle*. Hayward, CA: IMS, 2nd edition, 1988.

[14] José M. Bernardo. Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, 41:113–147 (with discussion). Reprinted in *Bayesian Inference* **1** (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 1995, 229–263., 1979.

[15] Ted Chang and Cesareo Villegas. On a theorem of Stein relating Bayesian and classical inferences in group models. *Canad. J. Statist.*, 14:289–296, 1986.

[16] David R. Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B*, 49:1–39 (with discussion), 1987.

[17] Andrew I. Dale. *A History of Inverse Probability: From Thomas Bayes to Karl Pearson*. Berlin: Springer, 1999.

[18] Gauri S. Datta and Jayanta K. Ghosh. On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82:37–45, 1995.

[19] Gauri S. Datta and Rahul Mukerjee. *Probability Matching Priors: Higher Order Asymptotics*. New York: Springer, 2004.

[20] Augustus de Morgan. *An Essay on Probabilities*. London: Longman, 1838.

[21] Bradley Efron and Carl N. Morris. Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, 70:311–319, 1975.

[22] Stephen E. Fienberg. When did Bayesian inference become Bayesian? *Bayesian Analysis*, 1:1–40, 2006.

[23] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Hoboken, NJ: Wiley, 1998.

[24] G. Hunt and C. Stein. Most stringent tests of statistical hypotheses. *Tech. Rep.,* Stanford University, 1946.

[25] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461, 1946.

[26] P. S. Laplace. *Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted in *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars, 1812.

[27] Bo H. Lindqvist and Gunnar Taraldsen. On the proper treatment of improper distributions. *J. Statist. Planning and Inference*, 195:93–104, 2018.

[28] Carl N. Morris and Martin Lysy. Shrinkage estimation in multilevel normal models. *Statistical Sci.*, 27:115–134, 2012.

[29] H. W. Peers. On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B*, 27:9–16, 1965.

[30] Charles Stein. On the coverage probability of confidence sets based on a prior distribution. In Ryszard Zielisnski, editor, Sequential Methods in Statistics, pages 485–514. Varsaw: Polish Scientific Publications, 1985.

[31] Béla Szökefalvi-Nagy. Alfred Haar (1885–1933). *Results in Mathematics*, 8:194–196, 1985.

[32] Robert Tibshirani. Noninformative priors for one parameter of many. *Biometrika*, 76:604–608, 1989.

[33] Bernard L. Welch and H. W. Peers. On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc. B*, 25:318–329, 1963.

[34] Ruoyong Yang and James O. Berger. Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, 22:1195–1211, 1994.