

# Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography

Robert E. Kass and Larry Wasserman \*

June 22, 1994

## Abstract

Subjectivism has become the dominant philosophical foundation for Bayesian inference. Yet, in practice, most Bayesian analyses are performed with so-called “noninformative” priors, that is, priors constructed by some formal rule. We review the plethora of techniques for constructing such priors, and discuss some of the practical and philosophical issues that arise when they are used. We give special emphasis to Jeffreys’s rules and discuss the evolution of his point of view about the interpretation of priors, away from unique representation of ignorance toward the notion that they should be chosen by convention. We conclude that the problems raised by the research on priors chosen by formal rules are serious and may not be dismissed lightly; when sample sizes are small (relative to the number of parameters being estimated) it is dangerous to put faith in any “default” solution; but when asymptotics take over, Jeffreys’s rules and their variants remain reasonable choices. We also provide an annotated bibliography.

*Key words and phrases:* Bayes factors, coherence, data-translated likelihoods, Entropy, Fisher information, Haar measure, improper priors, insufficient reason, Jeffreys’s prior, marginalization paradoxes, noninformative priors, nuisance parameters, reference priors, sensitivity analysis.

---

\*Robert E. Kass is Professor and Larry Wasserman is Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-2717. The work of both authors was supported by NSF grant DMS-9005858 and NIH grant R01-CA54852-01. The authors thank Nick Polson for helping with a few annotations, and Jim Berger, Teddy Seidenfeld and Arnold Zellner for useful comments and discussion.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Jeffreys's Methods</b>	<b>4</b>
2.1	Philosophy . . . . .	4
2.2	Rules for priors in problems of estimation . . . . .	7
2.3	Bayes factors . . . . .	9
<b>3</b>	<b>Methods For Constructing Reference Priors</b>	<b>12</b>
3.1	Laplace and the Principle of Insufficient Reason . . . . .	12
3.2	Invariance . . . . .	14
3.3	Data-translated likelihoods . . . . .	17
3.4	Maximum Entropy . . . . .	18
3.5	The Berger-Bernardo Method . . . . .	20
3.5.1	Missing Information . . . . .	21
3.5.2	Nuisance Parameters . . . . .	23
3.5.3	Related Work . . . . .	24
3.6	Geometry . . . . .	25
3.7	Coverage Matching Methods . . . . .	26
3.8	Zellner's Method . . . . .	29
3.9	Decision-Theoretic methods . . . . .	30
3.10	Rissanen's Method . . . . .	31
3.11	Other Methods . . . . .	33
<b>4</b>	<b>Issues</b>	<b>34</b>
4.1	Interpretation of reference priors . . . . .	34
4.2	Impropriety . . . . .	35
4.2.1	Incoherence, Strong Inconsistencies and Non-conglomerability . . . . .	35
4.2.2	The Dominating Effect of the Prior . . . . .	39
4.2.3	Inadmissibility . . . . .	40
4.2.4	Marginalization Paradoxes . . . . .	41
4.2.5	Improper Posteriors . . . . .	42
4.3	Sample Space Dependence . . . . .	43
4.4	Sensitivity Analysis . . . . .	44
<b>5</b>	<b>Discussion</b>	<b>46</b>
5.1	Local uniformity . . . . .	46
5.2	Reference priors with large samples . . . . .	47
5.3	Open problems . . . . .	49
	<b>References</b>	<b>51</b>
	<b>Annotated Bibliography</b>	<b>57</b>

# 1 Introduction

Since Bayes (1763), and especially since Fisher (1922; see Zabell, 1992), the scope and merit of Bayesian inference have been debated. Critics find arbitrariness in the choice of prior an overwhelming difficulty, while proponents are attracted to the logical consistency, simplicity, and flexibility of the Bayesian approach and tend to view determination of a prior as an important but manageable technical detail. These days most Bayesians rely on the subjectivist foundation articulated by De Finetti (1937, 1972, 1974, 1975) and Savage (1954, 1972). This has led to suggestions for personal prior “elicitation” (Savage 1954, Lindley, Tversky and Brown 1979, Kadane, Dickey, Winkler, Smith and Peters 1980) but these inherently problem-specific methods have not been developed extensively and have had relatively little impact on statistical practice. Thus, as increased computing power has widened interest in Bayesian techniques, new applications continue to raise the question of how priors are to be chosen.

The alternative to elicitation is to try to find structural rules that determine priors. From time to time, especially during the 1960’s and 1970’s, and again in the past several years, various such schemes have been investigated and there is now a substantial body of work on this topic. Feeling the urgency of the problem, and recognizing the diversity of the articles on this subject, we undertook to review the literature and appraise the many methods that have been proposed for selecting priors by formal rules. This paper is the result of our efforts.

Since the fundamental ideas and methods originate with Jeffreys, we begin, in Section 2, with an overview of his work. We discuss Jeffreys’s philosophy and we explain the techniques he used to construct priors in estimation and testing problems. An essential observation is that Jeffreys’s point of view evolved toward seeing priors as chosen by convention, rather than as unique representations of ignorance. Section 3 is a list of methods for constructing prior distributions. In Section 4 we discuss some of the philosophical and practical issues that arise when choosing priors conventionally, by formal rules. We draw conclusions from our study and provide our own interpretations in Section 5. This is followed by an annotated bibliography.

Since our discussion is fairly abstract it is worth keeping in mind some concrete examples.

One important class, which is useful for this purpose, is that of the multivariate Normal distributions, with mean  $\mu$  and variance matrix  $\Sigma$ . There are many special cases of interest. For instance,  $\mu$  and/or  $\Sigma$  may depend on some lower-dimensional parameter vector  $\theta$ ; when  $\mu = \mu(\theta)$  with  $\Sigma = \sigma^2 \cdot I$  we obtain the standard nonlinear regression models, and the structure  $\Sigma = \Sigma(\theta)$  includes “components of variance”, hierarchical, and time-series models.

We take for granted the fundamental difficulty in uniquely specifying what “non-informative” should mean. Thus, we prefer to call the priors we discuss *reference priors*. Because Bernardo (1980) used the term “reference prior” for a prior chosen by a particular formal rule (described in Section 3.5, below), we have struggled with alternative labels such as “conventional prior”, “default prior”, or “generic prior”. In the end, however, we have returned to the terminology of Box and Tiao (1973, pp. 22-23), who followed Jeffreys (1955), because we feel it is the best word for the purpose. Our reasons should become clear in the next section.

## 2 Jeffreys’s Methods

The concept of selecting a prior by convention, as a “standard of reference,” analogous to choosing a standard of reference in other scientific settings, is due to Jeffreys. Subsequent efforts to formulate rules for selecting priors may often be seen as modifications of Jeffreys’s scheme. Thus, we devote a section to a description of his methods. We begin with some philosophical background, then move on to specific rules. Jeffreys was careful to distinguish estimation and testing problems. We review his methods for choosing priors in testing problems in Section 2.3.

### 2.1 Philosophy

As is true of methods generally, Jeffreys’s should be understood in conjunction with the philosophy that generated them and, in turn, was defined by them.

Jeffreys has been considered by many to have been an “objectivist” or “necessarist”. Certainly, there is a sense in which this label is accurate, and it was useful for Savage (1962a, 1962b) to distinguish Jeffreys’s from his own subjectivist point of view. But there is

a subtlety in the opinions voiced by Jeffreys, as they evolved over time, that is fundamental and advances the discussion beyond the plateau Savage surveyed. As we document below, Jeffreys believed in the existence of states of ignorance, and he subscribed to the “Principle of Insufficient Reason”, neither of which play a part in subjectivist theory. But in his reliance on convention he allowed ignorance to remain a vague concept, that is, one that may be made definite in many ways, rather than requiring a unique definition. This provided a more flexible, vibrant framework that could support modern practice.

Savage (1962a, 1962b) labeled “necessarist” the position that “there is one and only one opinion justified by any body of evidence, so that probability is an objective logical relationship between an event  $A$  and the evidence  $B$ .” Jeffreys’s point of view in the first edition of *Scientific Inference* (1931, p. 10) puts him in this category.

... logical demonstration is right or wrong as a matter of the logic itself, and is not a matter for personal judgment. We say the same about probability. On a given set of data  $p$  we say that a proposition  $q$  has in relation to these data one and only one probability. If any person assigns a different probability, he is simply wrong, and for the same reasons as we assign in the case of logical judgments.

A similar passage may be found in the first edition of *Theory of Probability* (1939, p. 36).

The historical basis for Savage’s categorization is already clear but there is a further reason for identifying Jeffreys as a “necessarist”. This comes from considering the case in which there are only finitely many events (or values of a parameter, or hypotheses). One test for adherence to the necessarist point of view is whether, in this case, a uniform distribution is advocated, according to what has been called (after Laplace, 1820; see Section 3.1, below) the “Principle of Insufficient Reason”. This principle requires the distribution on the finitely many events to be uniform unless there is some definite reason to consider one event more probable than another. The contentious point is whether it is meaningful to speak of a “definite reason” that does not involve subjective judgment.

According to this test, Jeffreys continued to be a necessarist. He believed in the existence of an “initial” stage of knowledge, and thought it was important to be able to make inferences

based on data collected at this stage. In the case of a particular hypothesis being considered, he described this stage (1961, p. 33) as one at which an investigator has “no opinion” about whether the hypothesis is true. He went on, “If there is no reason to believe one hypothesis rather than another, the probabilities are equal . . . if we do not take the prior probabilities equal we are expressing confidence in one rather than another before the data are available . . . and this must be done only from definite reason.” Jeffreys added that the Principle of Insufficient Reason is “merely a formal way of expressing ignorance.”

Note that a subjectivist would agree that assigning unequal probabilities to two hypotheses would be “expressing confidence in one rather than another.” A subjectivist, however, would not accept any restriction on, nor require any special justification for, the belief. To a subjectivist, the probability assessment is in just this sense supposed to be “subjective.” Thus, a subjectivist has no pressing need for a “way of expressing ignorance.”

Despite his belief in an “initial” stage at which an investigator is ignorant, and his application of Insufficient Reason at this stage, we have in his later writings what might be regarded as Jeffreys’s attempt to sidestep the major obstacle in the necessarist construction. In the second edition of *Scientific Inference* the passage cited above, concerning probability as a uniquely determined logical relation, is absent. Instead, Jeffreys took reasonable degree of belief as a primitive concept, and said simply (1957, p. 22), “If we like, there is no harm in saying that probability expresses a degree of reasonable belief.” The choice of an initial assignment of probability then became a matter of convention, in the same way that the correspondence between a real-world object and a primitive concept in any axiom system is outside the formal system and must rely on some external rule for its application. Thus, Jeffreys maintained that his approach did not assume that only one prior was logically correct. In explaining his position (1955, p. 277), he wrote:

It may still turn out that there are many equally good methods. . .if this happens there need be no great difficulty. Once the alternatives are stated clearly a decision can be made by international agreement, just as it has been in the choice of units of measurement and many other standards of reference.

Meanwhile, the section cited above from the first edition of *Theory of Probability* is

altered in the second and third editions (1948, pp. 36-37; 1961, pp. 36-37), and says, “. . . in a different world, the matter would be one for decision by the International Research Council.” Thus priors, like weights and measures, are defined by convention. As long as we agree on these conventions, the particular choice is not crucial.

It is clear from these passages that Jeffreys did not insist on unique representations of ignorance, so that statements such as, “According to Jeffreys’s conception there is only one right distribution” (Hacking, 1976, p. 203) are inaccurate. When Savage (1962b, p. 21) remarked that, “It has proved impossible to give a precise definition of the tempting expression ‘know nothing’ ” Jeffreys responded (1963), “But who needs a definition?” by which we interpret him to mean that conventional rules suffice without incorporation of a formal definition into his axiomatic framework. On the other hand, although he did not claim that logic demanded a particular prior to represent ignorance, Jeffreys did work to find “the best” rule in each of many cases. His principles for doing so were supposed to provide “a guide”, but in some cases he thought these would “indicate a unique choice” (1961, p. 37). Ideally, that is, “in a different world”, there could be agreement on a single prior for use under ignorance in each problem.

The net effect of this re-examination is to make Jeffreys’s approach seem somewhat less rigid, and to recognize the importance of convention in his scheme. We have based our remarks on Kass (1982), which responded to Zellner (1982).

## **2.2 Rules for priors in problems of estimation**

Jeffreys considered several scenarios in formulating his rules, and treated each separately. The simplest is the case of a finite parameter space, in which, as we said in Section 2.1, he adhered to the Principle of Insufficient Reason in advocating the assignment of equal probabilities to each of the parameter values. Jeffreys then considered the cases in which the parameter space was a finite interval, the interval  $(-\infty, \infty)$ , or the interval  $(0, \infty)$ . In the first two cases Jeffreys took the prior density to be constant over the interval. In the second case this entails, of course, that the prior be improper, i.e., that it not integrate. He did not consider this to raise any fundamental difficulties. For the third case, most commonly

associated with an unknown standard deviation  $\sigma$ , he used the prior  $\pi_\sigma(\sigma) = 1/\sigma$ . His chief justification for this choice was its invariance under power transformations of the parameter: if  $\gamma = \sigma^a$  and the change-of-variables formula is applied to  $\pi_\sigma$  one obtains  $\pi_\gamma(\gamma) = 1/\gamma$ ; thus, applications of the rule to  $\sigma$  and  $\gamma$  lead to the same formal prior.

In a 1946 paper, Jeffreys proposed his “general rule.” Writing the Fisher information matrix as  $I(\theta)$ , where  $I(\theta)_{ij} = E(-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j})$ , the rule is to take the prior to be

$$\pi_\theta(\theta) \propto \det(I(\theta))^{1/2}. \tag{1}$$

(Here and throughout we use  $\det(\cdot)$  to denote the determinant.) It is applicable as long as  $I(\theta)$  is defined and positive-definite. As is easily checked, this rule has the invariance property that for any other parameterization  $\gamma$  for which it is applicable,

$$\pi_\theta(\theta) = \pi_\gamma(\gamma(\theta)) \cdot \det\left(\frac{\partial \gamma}{\partial \theta}\right)$$

i.e., the priors defined by the rule on  $\gamma$  and  $\theta$  transform according to the change-of-variables formula. Thus, it does not require the selection of any specific parameterization, which could in many problems be rather arbitrary; in this sense it is quite general. Additional discussion of the rule is given in Section 3.1. (There are other priors that are parameterization invariant; see Hartigan, 1964.)

Jeffreys noted that this rule may conflict with the rules previously stated, which depend on the interval in which a parameter lies. In particular, in the case of data that follow a  $N(\mu, \sigma^2)$  distribution, the previous rule gives  $\pi(\mu, \sigma) = 1/\sigma$  while the general rule gives  $\pi(\mu, \sigma) = 1/\sigma^2$ . The latter he found unacceptable (because if extended to the case of spherical Normal data it would imply that the marginal posterior distribution of each component of the mean would have a  $t$  distribution with degrees of freedom no longer depending on the dimensionality of the mean vector). He solved this problem by stating that  $\mu$  and  $\sigma$  ought to be judged independent *a priori* and so should be treated separately. When the general rule is applied while holding  $\sigma$  fixed it gives the uniform prior on  $\mu$ , and when it is applied while holding  $\mu$  fixed it gives the prior  $\pi(\sigma) \propto 1/\sigma$ . Thus, with this modification, the general rule

becomes consistent with his previous rules.

Jeffreys went further, and suggested this modification for general location-scale problems. He also proposed that priors in problems involving parameters in addition to location and scale parameters be taken by treating the location parameters separately from the rest (1961, pp. 182-183). That is, if there are parameters  $\mu_1, \dots, \mu_k$ , and  $\theta$ , where  $\theta$  is multidimensional, then the prior he recommended becomes

$$\pi(\mu_1, \dots, \mu_k, \theta) \propto \det(I(\theta))^{1/2}, \quad (2)$$

where  $I(\theta)$  is calculated holding  $\mu_1, \dots, \mu_k$  fixed. When there are also scale parameters involved, these become part of  $\theta$  and (2) is applied. The prior in (2) may then also be written in the form  $\pi(\mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k, \theta) \propto \det(I(\theta))^{1/2} \prod_{i=1}^k \sigma_i^{-1}$ , where  $I(\theta)$  is now calculated holding all of  $\mu_1, \dots, \mu_k$  and  $\sigma_1, \dots, \sigma_k$  fixed.

**DEFINITION.** We will call (1) and (2) *the prior determined by Jeffreys's general rule*, letting the context distinguish these two cases. To contrast (2) with the prior obtained by applying (1) when there are location parameters, we will refer to (1) as *the prior obtained from Jeffreys's non-location rule*. Thus, what we call Jeffreys's non-location rule is a rule Jeffreys recommended *not* be applied to families having location parameters.

Though the calculations are sometimes somewhat involved, it is straightforward to apply (2) to the class of multivariate Normal models mentioned in the Introduction. When either  $\mu$  or  $\Sigma$  depend on a parameter vector  $\theta$ , the information matrix on  $\theta$  may be obtained via the chain rule from that on  $(\mu, \Sigma)$  in the unrestricted case.

We note Jeffreys also suggested (1961, p. 185) that in the case of mixtures, the mixing parameters should be treated independently from the other parameters.

### 2.3 Bayes factors

Jeffreys emphasized the distinction between problems of estimation and problems of testing. Importantly, in testing he did not advocate the use of the rules discussed in Section 2.2, above, but instead recommended a different method.

Suppose  $Y = (Y_1, \dots, Y_n)$  follow a distribution in a family parameterized by  $(\beta, \psi) \in B \times \Psi$  having a density  $p(y | \beta, \psi)$ , and the hypothesis  $H_0 : \psi = \psi_0$  is to be tested against the unrestricted alternative  $H_A : \psi \in \Psi$ . Jeffreys's method is based on what is now usually called the "Bayes factor "

$$B = \frac{\int p(y | \beta, \psi_0) \pi_0(\beta) d\beta}{\int \int p(y | \beta, \psi) \pi(\beta, \psi) d\beta d\psi} \quad (3)$$

where  $\pi_0(\beta)$  and  $\pi(\beta, \psi)$  are priors under  $H_0$  and  $H_A$ . The Bayes factor may be interpreted as the posterior odds of  $H_0$  when the prior odds are 1 : 1. More generally, it is the ratio of posterior odds to prior odds, regardless of the prior odds on  $H_0$ . For an extensive review of modern methodology using Bayes factors, see Kass and Raftery (1993).

Jeffreys's proposals for priors  $\pi_0$  and  $\pi$  appear in Sections 5.02, 5.1-5.3, and 6.2 of *Theory of Probability*. Generally, he used his estimation reference priors on the nuisance parameter  $\beta$ . As he showed, and Kass and Vaidyanathan (1992) elaborated upon, when  $\psi$  and  $\beta$  are assumed orthogonal and *a priori* independent the value of the Bayes factor is not very sensitive to the choice of  $\pi_0$ . The prior on  $\psi$ , on the other hand, remains important.

When  $\psi$  was a probability, as in a Binomial problem, Jeffreys (1961, Section 5.1) used a flat prior on  $(0, 1)$ . For the Normal location problem, in which  $\beta$  is the Normal standard deviation and the null hypothesis on the mean  $\psi$  becomes  $H_0 : \psi = 0$ , Jeffreys (1961, Section 5.2) took the prior on  $\psi$  to be Cauchy. He argued that, as a limiting case, the Bayes factor should become 0 if the observed standard deviation were zero, since this would say that the location parameter was in fact equal to the observed value of the observations. This requires that the moments of the prior do not exist and, he said, the simplest distributional form satisfying this condition is the Cauchy. Furthermore, he liked this form because he felt it offered a reasonable representation of "systematic errors" in observations (as opposed to "random errors"): a non-zero location parameter would be treated as if arising from one among many such, corresponding to one series of observations among many series made under differing conditions.

Jeffreys treated the general case, in which  $\beta$  and  $\psi$  were one-dimensional but the distribution for the data was arbitrary, by assuming the parameters were orthogonal and then,

drawing an analogy with the Normal location problem, taking the prior on  $\psi$  to be Cauchy in terms of the symmetrized Kullback-Leibler number (Jeffreys, 1961, pp. 275 and 277). He then used an asymptotic approximation to obtain a simple computable form.

Kass and Wasserman (1993) have shown how Jeffreys's method may be generalized to arbitrarily many dimensions by replacing Jeffreys's requirement of parameter orthogonality (that the information matrix be block diagonal for all parameter values) with "null-orthogonality" (that the information matrix be block diagonal when  $\psi = \psi_0$ ). The log of resulting approximation has the form  $S + c$  where  $S$  is the Schwarz criterion and  $c$  is a constant. In addition, they note the disappearance of the constant  $c$  when a Normal prior is used, and they point out the interpretation of such a prior is that "the amount of information in the prior on  $\psi$  is equal to the amount of information about  $\psi$  contained in one observation." They find this a reasonable prior to use and conclude that there is good motivation for using the Schwarz criterion (or some minor modification of it) as a large-sample testing procedure. Their results generalize some given previously, for the special case of linear regression, by Smith and Spiegelhalter (1980) and Zellner and Siow (1980).

I.J. Good has written extensively on Bayes factors. In Good (1967) he followed Jeffreys in suggesting a Cauchy prior for the parameter of interest, in that case the log of the concentration parameter for a Dirichlet distribution. He suggested subjectively determining the choice of Cauchy location and scale parameters, but in his tabulations (p. 414) used the standard Cauchy as a reference prior.

In most cases, Jeffreys assumed the initial probabilities of the two hypotheses were equal, which is a reference choice determined by "insufficient reason" (Section 2.1, above). Alternatives have more recently been proposed: Pericchi (1984), following on earlier work by Bernardo (1980), discussed maximizing expected information gain as a method of selecting between competing linear regression models. Here, both parameters appearing within the models and the probabilities assigned to them are considered quantities about which an experiment provides information. The design matrices introduce an interesting complication to the problem, generally leading to unequal probabilities.

### 3 Methods For Constructing Reference Priors

Many methods have been proposed for constructing reference priors. In this section we describe most of these methods. Whenever possible, we avoid technical details and present the arguments in their simplest forms. Often, different arguments lead back to Jeffreys's prior or some modification of it. Sometimes the parameter  $\theta$  can be written in the form  $\theta = (\omega, \lambda)$  where  $\omega$  is a parameter of interest and  $\lambda$  is a nuisance parameter. In this case, reference priors that are considered satisfactory for making inferences about  $\theta$  may not be satisfactory for making inferences about  $\omega$ . Much of the recent research on reference priors has been inspired by this latter observation. This work is highlighted in sections 3.5 and 3.7 and in the end of 3.2.

#### 3.1 Laplace and the Principle of Insufficient Reason

If the parameter space is finite, Laplace's rule, or the Principle of Insufficient Reason is to use a uniform prior which assigns equal probability to each point in the parameter space. The use of uniform probabilities on finite sets dates back to the origins of probability in gambling problems. The terminology come from references by Laplace to a lack of sufficient reason to suppose an alternative (e.g., Laplace, 1820; Howson and Urbach, 1989, p. 40, attribute its statement as a "Principle" to von Kries, 1886).

This rule is appealing but is subject to a partitioning paradox: it is inconsistent to apply the rule to all coarsenings and refinings of the parameter space simultaneously. Shafer (1976, pp 23-24) gives a simple example. Let  $\Theta = \{\theta_1, \theta_2\}$  where  $\theta_1$  denotes the event that there is life in orbit about the star Sirius and  $\theta_2$  denotes the event that there is not. Laplace's rule gives  $P(\{\theta_1\}) = P(\{\theta_2\}) = 1/2$ . But now let  $\Omega = \{\omega_1, \omega_2, \omega_3\}$  where  $\omega_1$  denotes the event that there is life around Sirius,  $\omega_2$  denotes the event that there are planets but no life and  $\omega_3$  denotes the event that there are no planets. Then Laplace's rule gives  $P(\{\omega_1\}) = P(\{\omega_2\}) = P(\{\omega_3\}) = 1/3$ . The paradox is that the probability of life is  $P(\{\theta_1\}) = 1/2$  if we adopt the first formulation but it is  $P(\{\omega_1\}) = 1/3$  if we adopt the second.

In practice, the partitioning paradox is not such a serious problem. One uses scientific judgment to choose a particular level of refinement that is meaningful for the problem at

hand. The fact that the space could, in principle, be refined further, is not usually of great practical concern. Indeed, according to Stigler (1986, p. 103), Laplace assumed that the problem at hand had already been specified in such a way that the outcomes were equally likely. And one could argue that in a decision problem, the structure of the problem determines the level of partition that is relevant (Chernoff 1954).

For a continuous parameter space, the natural generalization of the principle of insufficient reason is to use a flat prior. A problem with this rule is that it is not parameterization invariant. For example, if  $\theta$  is given a uniform distribution then,  $\phi = e^\theta$  will not have a uniform distribution. Conversely, if we start with a uniform distribution for  $\phi$  then  $\theta = \log \phi$  will not have a uniform distribution. To avoid a paradox we need a way to determine a privileged parameterization.

Perhaps the oldest and most famous use of a uniform prior on an infinite set is Bayes (1763) who used a uniform prior for estimating the parameter of a binomial distribution. Stigler (1982) argues that Bayes' paper has largely been misunderstood. According to Stigler, the thrust of Bayes' argument was that  $X_n$ , the number of successes in  $n$  trials, should be uniform, for every  $n \geq 1$ . This entails that  $\theta$  must have a uniform prior. This argument is compelling because it is based on observable quantities, although the uniform distribution on  $X_n$  is still subject to refining paradoxes.

The partitioning paradox on finite sets and the lack of parameterization invariance are closely related. In both cases we have two spaces  $\Theta$  and  $\Omega$  and a mapping  $g : \Omega \rightarrow \Theta$ . We then have the choice of adopting a uniform prior on  $\Theta$  or adopting a uniform prior  $\mu$  on  $\Omega$  which then induces a prior  $\pi$  on  $\Theta$ , where  $\pi$  is defined by  $\pi(A) = \mu(g^{-1}(A))$ . In general,  $\pi$  will not be uniform. In the continuous case, the mapping  $g$  corresponds to some reparameterization. In the finite case,  $\Omega$  is a refinement of  $\Theta$  and  $g$  relates the original space  $\Theta$  to its refinement. In the "life on Sirius" example,  $g$  is defined by  $g(\omega_1) = \theta_1$ ,  $g(\omega_2) = \theta_2$ ,  $g(\omega_3) = \theta_2$ .

## 3.2 Invariance

Invariance theory has played a major role in the history of reference priors. Indeed, Laplace's principle of insufficient reason is an application of an invariance argument. In this section, we review the key aspects of this approach to the selection of priors. Good descriptions of the role of invariance are given by Dawid (1983), Hartigan (1964), and Jaynes (1968).

The simplest example of invariance is the permutation group on a finite set. Let  $\Theta = \{\theta_1, \dots, \theta_n\}$  and let  $\mathcal{G}$  be the set of permutations of the integers  $\{1, \dots, n\}$ . Write  $B = gA$  if  $I_B = I_A \circ g$  where  $I_A$  is the indicator function for  $A$  and  $g \in \mathcal{G}$ . If we have little prior information then it seems natural to demand that the prior should be invariant under permutations, i.e.,  $P(A) = P(gA)$  for every  $A$  and every  $g \in \mathcal{G}$ . This implies that  $P$  is the uniform probability, and could be viewed as a formal expression of the Principle of Insufficient Reason discussed in Section 3.1.

When the parameter space is infinite, the invariance arguments are more complicated. We begin with the Normal location model. Suppose a statistician,  $S_1$  records a quantity  $X$  that has a  $N(\theta, 1)$  distribution, and has a prior  $\pi_1(\theta)$ . A second statistician  $S_2$  records the quantity  $Y = X + a$ , with  $a$  being a fixed constant. Then  $Y$  has a  $N(\phi, 1)$  distribution, where  $\phi = \theta + a$  and let this statistician's prior be  $\pi_2(\phi)$ . Since both statisticians are dealing with the same formal model – a Normal location model – their reference priors should be the same. Thus, we require  $\pi_1 = \pi_2$ . On the other hand, since  $\phi = \theta + a$ ,  $\pi_1$  and  $\pi_2$  can be related by the usual change of variables formula. The relationships between  $\pi_1$  and  $\pi_2$  should hold for every  $a$  and this implies that they must both be uniform distributions.

This Normal location model may be re-expressed in terms of group invariance. Each real number  $a$  determines a transformation  $h_a : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $h_a(x) = a + x$ . The set of all such transformations  $H = \{h_a; a \in \mathbb{R}\}$  forms a group if we define  $h_a h_b = h_{a+b}$ . We say that the model is *invariant under the action of the group* since  $X \sim N(\theta, 1)$  and  $Y = h_a(X)$  implies that  $Y \sim N(h_a(\theta), 1)$ . The uniform prior  $\mu$  is the unique prior (unique up to an additive constant) that is invariant under the action of the group, that is,  $\mu(h_a A) = \mu(A)$  for every  $A$  and every  $a$ , where  $h_a A = \{h_a(\theta) : \theta \in A\}$ .

Now suppose that  $X \sim N(\theta, \sigma^2)$ . Let  $H = \{h_{a,b}; a \in \mathbb{R}, b \in \mathbb{R}^+\}$  where  $h_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$

is defined by  $h_{a,b}(x) = a + bx$ . Again,  $H$  is a group. Define another group  $G = \{g_{a,b}; a \in \mathbb{R}, b \in \mathbb{R}^+\}$  where  $g_{a,b} : \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R} \times \mathbb{R}^+$  is defined by  $g_{a,b}(\theta, \sigma) = (a + b\theta, b\sigma)$ . Note that the group  $G$  is formally identical to the parameter space for this problem. Thus, every pair  $(\theta, \sigma) \in \mathbb{R} \times \mathbb{R}^+$  identifies both an element of the Normal family and a transformation in  $G$ . Now, as before, the model is invariant under the action of the group in the sense that if  $X \sim N(\theta, \sigma^2)$  and  $Y = h_{a,b}(X)$  then  $Y \sim N(\mu, \lambda^2)$  where  $(\mu, \lambda) = g_{a,b}(\theta, \sigma)$ . The prior  $P$  that is invariant to left multiplication, i.e.,  $P(g_{a,b}A) = P(A)$  for all  $A$  and all  $(a, b) \in \mathbb{R} \times \mathbb{R}^+$ , has density  $p(\mu, \sigma) \propto 1/\sigma^2$ . This is the same prior we would get by using (1) but, as we discussed in section 2, Jeffreys preferred the prior  $Q$  with density  $q(\mu, \sigma) \propto 1/\sigma$ . It turns out that  $Q$  is invariant to right-multiplication, meaning that  $Q(Ag_{a,b}) = Q(A)$  for all  $A$  and all  $(a, b) \in \mathbb{R} \times \mathbb{R}^+$ , where  $Ag_{a,b} = \{g_{\theta, \sigma}g_{a,b}; (\theta, \sigma) \in A\}$ . The priors  $P$  and  $Q$  are called, respectively, *left Haar measure* and *right Haar measure*.

The preceding arguments can be applied to more general group transformation models in which the parameter space is identified with the group  $G$ . In the simplest case,  $G$  is transitive (for every  $\theta_1, \theta_2 \in \Theta$  there exists  $g \in G$  such that  $\theta_2 = g\theta_1$ ) and acts freely ( $g\theta = \theta$  for some  $\theta \in \Theta$  only if  $g$  is the identity) on both  $\Theta$  and the sample space, with  $X \sim P_\theta$  if and only if  $gX \sim P_{g\theta}$ . In this case the left and right Haar measures on  $G$  provide distributions on  $\Theta$  that are again unique (up to a multiplicative constant). Somewhat more complicated cases occur when the group action on the sample space is either non-transitive or non-free. Here, the sample space  $\mathcal{X}$  may be identified with the product  $G \times \mathcal{X}/G$  where  $\mathcal{X}/G$  is the “coset space”. See, for instance, Chang and Villegas (1986). In all of these situations, if the group is non-compact and non-commutative, the left and right Haar measures may be distinct. (See Nachbin, 1965, for details on Haar measures.) Jeffreys’s non-location prior (1) is the left Haar measure (see, e.g., Dawid, 1983; this also follows from its derivation as a volume element determined by a Riemannian metric, see, e.g., Kass, 1989).

Villegas (1981) made the following argument for the right Haar measure in the case in which  $G$  is transitive and acts freely. Let  $\lambda$  be a measure on the group  $G$ . Choose a reference point  $a \in \Theta$ . This defines a mapping  $\phi_a : G \rightarrow \Theta$  by  $\phi_a g = ga$  which induces a measure  $\pi_a = \lambda\phi_a^{-1}$  on  $\Theta$ . If we insist that the measure  $\pi_a$  not depend on the choice of reference point  $a$  then  $\pi$  must be the right Haar measure. The argument generalizes to the case in

which the sample space  $\mathcal{X}$  may be identified with the product  $G \times \mathcal{X}/G$  and Chang and Eaves (1990, Proposition 4) show that different possible such decompositions lead to the same right invariant prior.

Another argument in favor of right Haar priors comes from the demonstration by Stone (1965, 1970) that a necessary and sufficient condition for an invariant posterior to be obtained as a limit, in probability, of posteriors based on proper priors, is (under the assumption that the group is amenable) that the prior is right Haar measure. (See section 4.2.1 for more discussion on probability limits of proper priors.) Also, we note that posteriors based on right Haar measure arise formally in a type of conditional inference called structural inference, developed by Fraser (1968). Furthermore, the right Haar measure gives the best invariant decision rule in invariant decision problems (Berger 1985 section 6.6.2)

Related to this is a result proved by Chang and Eaves (1968) that repeated-sampling coverage probabilities and posterior probabilities agree when the prior on the group is right Haar measure. (See Section 3.7.)

The invariance arguments may be replaced by weaker *relative invariance* arguments that require proportionality rather than equality for statements of invariance. In particular, if we want  $\pi(A|X = x) = \pi(g^{-1}(A)|g^{-1}(X) = g^{-1}(x))$  say, when  $\Theta$  and  $\Theta'$  are related by a transformation  $g$ , then we only need that  $\pi'(A) \propto \pi(g^{-1}(A))$ . The class of relatively invariant priors is much larger than the class of invariant priors; see Hartigan (1964).

Sometimes the group action is not itself of interest but instead group elements correspond to nuisance parameters, i.e., the full parameter vector is  $\theta = (\omega, g)$  where  $g \in G$  and  $\omega$  is the parameter of interest. Assuming  $\omega$  is an index for the orbits of the group (the orbit of  $x$  is  $\{gx; g \in G\}$ ), Chang and Eaves (1990) recommend the prior  $\pi(\omega)\pi(g|\omega)$  where  $\pi(g|\omega)$  is right Haar measure and

$$\pi(\omega) = \lim_{n \rightarrow \infty} \sqrt{\det(I_n(\omega))/n}.$$

Here,  $I_n(\omega)$  is the information matrix for  $y_n$ , the maximal invariant of the  $G$ -action. This is similar to the Berger-Bernardo approach except that Berger and Bernardo would use the non-location Jeffreys prior (and hence left Haar measure) for  $\pi(g|\omega)$ .

### 3.3 Data-translated likelihoods

Box and Tiao (1973, Section 1.3) introduced the notion of “data-translated likelihood” to motivate the use of uniform priors. Let  $y$  be a vector of observations and let  $L_y(\cdot)$  be a likelihood function on a real one-dimensional parameter space  $\Phi$ . According to Box and Tiao (1973, eqn (1.3.13)), the likelihood function is data translated if it may be written in the form

$$L_y(\phi) = f\{\phi - t(y)\} \quad (4)$$

for some real-valued functions  $f(\cdot)$  and  $t(\cdot)$ , with the definition of  $f(\cdot)$  not depending on  $y$ . When (4) is satisfied, Box and Tiao recommend the use of the uniform prior on  $\Phi$  because two different samples  $y$  and  $y^*$  will then produce posteriors that differ only with respect to location. That is, the uniform prior does not produce posterior densities with different shapes for different samples. This feature of the uniform prior is, for Box and Tiao, what makes it “noninformative.”

They then introduced “approximate data-translated likelihood” to motivate Jeffreys’s general rule. For a likelihood to be approximately data translated, Box and Tiao require it to be “nearly independent of the data  $y$  except for its location.” Operationally, they discuss samples of size  $n$  consisting of independent and identically distributed observations and begin with the Normal approximation to the likelihood

$$L_y(\theta) \simeq n(\theta; \hat{\theta}, \hat{\sigma}_y^2), \quad (5)$$

where  $n(x; \mu, \sigma^2)$  is the Normal density with argument  $x$ , mean  $\mu$  and variance  $\sigma^2$ , and  $\hat{\sigma}_y^2 = \{ni(\hat{\theta})\}^{-1}$ , the inverse of the expected Fisher information evaluated at the maximum likelihood estimate  $\hat{\theta}$ . They then take  $\phi$  to be a variance-stabilizing parameterization, that is,  $i(\phi) = c$  for some constant  $c$ , so that

$$L_y(\phi) \simeq n(\phi; \hat{\phi}, c/n). \quad (6)$$

The Normal approximate likelihood of (6) has the form (4), so that the likelihood itself is, in a sense Box and Tiao do not make explicit, approximately data translated. Based on the analogy with (4), they recommend the use of a prior that is uniform on  $\phi$ , and they note that this prior is the one determined by Jeffreys's general rule.

To see more clearly what Box and Tiao's approach entails, notice that from (4) the likelihood functions based on alternative data  $y$  and  $y^*$  are translated images of one another in the sense that

$$L_y(\phi) = L_{y^*}(\phi^*) \tag{7}$$

for  $\phi^* = \phi + \{t(y^*) - t(y)\}$ . Clearly, if (7) holds, the translation group may be defined on  $\Phi$  and on the image of  $t(\cdot)$  so that the likelihood function is invariant under its action. Kass (1990) noted that, once seen from this group-theoretic perspective, the definition is revealed to be very restrictive (if  $\Phi$  is the whole real line and the support of the distribution is independent of  $\phi$  then only the Normal and gamma families yield exactly data-translated likelihoods). The concept is easily modified by requiring the likelihood to be data-translated only for each fixed value of an ancillary statistic. When this is done, the definition extends to general transformation models. Kass then showed that in one dimension likelihoods become approximately data-translated to order  $O(n^{-1})$ , which is stronger than the order  $O(n^{-1/2})$  implied by the data-translatedness of the limiting Normal distributions. A somewhat weak extension of the result was given for the multidimensional case: likelihoods may be considered approximately data-translated along information-metric geodesics in any given direction, but it is not in general possible to find a parameterization in which they become jointly approximately data-translated. (This is related to the inability to directly extend work of Welch and Peers (1963) as discussed in Stein (1985); see Section 3.7.)

### 3.4 Maximum Entropy

If  $\Theta = \{\theta_1, \dots, \theta_n\}$  is finite and  $\pi$  is a probability function of  $\Theta$ , the entropy of  $\pi$ , which is meant to capture the amount of uncertainty implied by  $\pi$ , is defined by  $\mathcal{E}(\pi) = -\sum \pi(i)\log\pi(i)$ . Entropy is a fundamental concept in statistical thermodynamics and information theory

(Shannon 1948, Wiener 1948, Ash 1965). The functional  $\mathcal{E}(\pi)$  can be justified as a measure of uncertainty by appealing to three axioms (Shannon 1948). Priors with larger entropy are regarded as being less informative and the method of maximum entropy is to select the prior that maximizes  $\mathcal{E}(\pi)$ . If no further constraints are imposed on the problem then the prior with maximum entropy is the uniform prior. Suppose now that partial information is available in the form of specified expectations for a set of random variables:  $\{E(X_1) = m_1, \dots, E(X_r) = m_r\}$ . Maximum entropy prescribes choosing the prior that maximizes entropy subject to the given moment constraints. The solution is the prior

$$\pi(\theta_i) \propto \exp\left\{\sum_j \lambda_j X_j(\theta_j)\right\}.$$

Jaynes (1957, 1968, 1980, 1982, 1983) has been the main developer of entropy based methods. The method of maximum entropy has been very successful in many problems including, for example, spectral analysis and image processing. A recent review of entropy based methods may be found in Zellner (1991). See also Zellner (1993), Zellner and Min (1992) and Press (1993). There are, however, some problems with the theory. Seidenfeld (1987) gives an excellent review and critique of maximum entropy. Here, we review the main points discussed in Seidenfeld's paper.

First, there is a conflict between the maximum entropy paradigm and Bayesian updating. Consider a six sided die and suppose we have the information that  $E(X) = 3.5$  where  $X$  is the number of dots on the uppermost face of the die. Following Seidenfeld, it is convenient to list the constraint set:  $C_0 = \{E(X) = 3.5\}$ . The probability that maximizes the entropy subject to this constraint is  $P_0$  with values  $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ . Let  $A$  be the event that the die comes up odd and suppose we learn that  $A$  has occurred. There are two ways to include this information. We can condition  $P_0$  to obtain  $P_0(\cdot|A)$  which has values  $(1/3, 0, 1/3, 0, 1/3, 0)$ . On the other hand, we can regard the occurrence of  $A$  as another constraint, namely,  $E(A) = 1$ . The probability  $Q$  that maximizes the entropy subject to the constraint set  $C_1 = \{E(X) = 3.5, E(A) = 1\}$  has values  $(.22, 0, .32, 0, .47, 0)$  which conflicts with  $P_0(\cdot|A)$ . One might conjecture that it is possible to refine the space under consideration so that a constraint expressed as an expectation on a random variable may be re-expressed

as an event. Perhaps, in this larger space, the conflict will disappear. But Friedman and Shimony (1971) and Shimony (1973) have shown that, in general, there is no such possible extension except in a trivial sense. They show that an extended space for which the constraint is represented as an event and for which conditionalization is consistent with maximum entropy, must be such that the constraint is given prior probability one. Seidenfeld shows that the Friedman-Shimony result applies not only to entropy, but to minimum Kullback-Leibler shifts from any given base measure; maximum entropy is obtained by taking the base measure to be uniform.

The second problem is that maximum entropy is subject to the same partitioning paradox that afflicts the principle of insufficient reason. Thus, in the die example, we can record, not just the value of the upper face, but also whether the sum of all visible spots on side faces of the die is less than, equal to, or greater than the value showing. For example, the outcome  $(3, Less)$  means the top face shows 3 and the sum of visible side faces is less than 3. There are 14 outcomes. Maximum entropy leads to a probability  $Q$  that assigns probability  $1/14$  to each outcome. The marginal of  $Q$  for the six original outcomes is not  $P_0$ . The problem is then, which probability should we use,  $Q$  or  $P_0$ ?

Entropy methods can be extended to the continuous case by measuring entropy relative to a base density  $\mu$ . Thus, the entropy of a density  $\pi$  with respect to  $\mu$  is  $-\int \pi \log \pi d\mu$ . Unfortunately, having to choose a base measure is no different than having to choose a prior so that this solution is rather circular. Indeed, in the finite case, a uniform measure has implicitly been chosen as a base measure. Jaynes (1968) suggests using base measures based on invariance arguments.

### 3.5 The Berger-Bernardo Method

Bernardo (1979a) suggested a method for constructing priors that involved two innovations. The first was to define a notion of missing information and the second was to develop a stepwise procedure for handling nuisance parameters. Since Bernardo's original paper, there has been a series of papers, mostly by Berger and Bernardo, refining the method and applying it to various problems. For this reason we refer to this method as the Berger-Bernardo

method.

When there are no nuisance parameters and certain regularity conditions are satisfied, Bernardo’s prior turns out to be (1). When there is a partitioning of the parameter into “parameters of interest” and “nuisance parameters”, this method will often produce priors that are distinct from (1). We shall discuss the notion of missing information first and then the stepwise procedure.

### 3.5.1 Missing Information

Let  $X_1^n = (X_1, \dots, X_n)$  be  $n$  iid random variables and let  $K_n(p(\theta|x_1^n), p(\theta))$  be the Kullback-Leibler distance between the posterior density and the prior density:  $K_n(p(\theta|x_1^n), p(\theta)) = \int p(\theta|x_1^n) \log(p(\theta|x_1^n)/p(\theta)) d\theta$ . Loosely, this is the gain in information provided by the experiment. Let  $K_n^\pi = E(K_n(p(\theta|x_1^n), p(\theta)))$  be the expected gain in information where the expectation is with respect to the marginal density  $m(x_1^n) = \int f(x_1^n|\theta)\pi(\theta)d\theta$ . experiment. Bernardo’s (1979a) idea was to think of  $K_n^\pi$  for large  $n$  as a measure of the missing information in the experiment. Bernardo (1979a) suggested finding the prior that maximizes  $K_\infty^\pi = \lim_{n \rightarrow \infty} K_n^\pi$  and called the result “the” reference prior. Since the term “reference prior” had already been used by Box and Tiao (1973) following Jeffreys, we prefer to use it in its more general sense. we shall stick to the name Berger-Bernardo prior. Hartigan (1983, Section 5.2) uses the term *maximal learning prior*. The reason for not performing the above optimization for finite  $n$  is that the priors turn out to have finite support (Berger, Bernardo and Mendoza 1989).

Now a technical problem arises, namely,  $K_\infty^\pi$  is usually infinite. (In fact, the infinities can occur for finite  $n$ ; see Hartigan’s discussion of Bernardo 1979a). To circumvent this problem, Bernardo finds the prior  $\pi_n$  that maximizes  $K_n^\pi$ . He then finds the limit of the corresponding sequence of posteriors and finally defines the reference prior as the prior that produces the limiting reference posterior via Bayes theorem. With sufficient regularity, this prior turns out to be (1) for continuous parameter spaces and the uniform prior for finite parameter spaces.

Another way around the infinities is simply to standardize  $K_n$ . Using asymptotic Nor-

mality we have  $K_n^\pi = (d/2)\log(n/2\pi\epsilon) + \int \pi(\theta)\log(\sqrt{\det(I)}/\pi(\theta))d\theta + o(1)$  as  $n \rightarrow \infty$  where  $d$  is the dimension of  $\theta$ . See Ibragimov and H'asminsky (1973) and Clarke and Barron (1990b). Define the standardized expected distance  $\widetilde{K}_n^\pi = K_n^\pi - (d/2)\log(n/2\pi\epsilon)$  and the *standardized missing information* by  $\widetilde{K}_\infty^\pi = \lim_{n \rightarrow \infty} \widetilde{K}_n^\pi = \int \pi(\theta)\log(\sqrt{\det(I)}/\pi(\theta))d\theta$ . A calculus of variations argument shows that the standardized missing information is maximized by (1). (More precisely, it is maximized by (1) if the space is truncated to an appropriate compact set.)

When the data are not i.i.d. there is some question about how to do the asymptotics. A recent discussion of this point is given in Berger and Yang (1992). They consider the AR(1) process:  $X_t = \rho X_{t-1} + \epsilon_t$  where  $\epsilon_t \sim N(0, 1)$ . There are two ways to do the asymptotics. One can consider  $n$  vectors  $X^1, \dots, X^n$  where each  $X^i = (X_1^i, \dots, X_T^i)$  is a single run of  $T$  observations from the process. Maximizing missing information and letting  $n$  go to infinity gives the Jeffreys's prior. This prior depends on  $T$  and so has strong sample space dependence. Also, Jeffreys's prior seems to put too much weight in the region of the parameter space that corresponds to non-stationarity. If asymptotic missing information is maximized instead for  $T \rightarrow \infty$  the prior is  $\pi(\rho) \propto (1 - \rho^2)^{-1/2}$  if the problem is restricted to  $\rho \in [-1, 1]$ . If the parameter space is  $[a, b]$  with  $a < -1$  or  $b > 1$  the prior is instead a discrete prior with mass at the endpoints. An alternative prior, called the symmetrized reference prior is also considered. This is defined by

$$\pi(\rho) = \begin{cases} \{2\pi\sqrt{1-\rho^2}\}^{-1} & \text{if } |\rho| < 1 \\ \{2\pi|\rho|\sqrt{1-\rho^2}\}^{-1} & \text{if } |\rho| > 1. \end{cases}$$

Clearly, for  $\rho \in [-1, 1]$  this is the Berger-Bernardo prior and the prior outside this range is obtained by the mapping  $\rho \rightarrow 1/\rho$ . Berger and Yang compared the sampling properties of the point and interval estimates based on these priors and found that the symmetrized reference prior performed better in mean-squared error and reasonably well in terms of coverage. More importantly, this is an interesting example showing that the prior can depend on how the asymptotics are carried out. This example has generated much debate among econometricians. Phillips (1991) argues in favor of the Jeffreys's prior. His article is followed

by a series of papers in which several authors discuss the merits of various approaches.

### 3.5.2 Nuisance Parameters

Suppose that  $\theta = (\omega, \lambda)$  where  $\omega$  is the parameter of interest and  $\lambda$  is a nuisance parameter. In this case, Bernardo suggests modifying his procedure. Ignoring some technical problems, the method is as follows. First, define  $\pi(\lambda|\omega)$  to be the Berger-Bernardo prior for  $\lambda$  with  $\omega$  fixed. Next, find the marginal model  $f(x|\omega) = \int f(x|\omega, \lambda)\pi(\lambda|\omega)d\lambda$ . (The technical problem is that the integral may diverge necessitating restriction to a compact set or a sequence of compact sets.) Now take  $\pi(\omega)$  to be the Berger-Bernardo prior based on the marginal model  $f(x|\omega)$ . The recommended prior is then  $\pi(\omega)\pi(\lambda|\omega)$ .

Assuming some regularity conditions, it can be shown that the Berger-Bernardo prior is

$$\pi(\omega, \lambda) \propto j_\omega(\lambda) \exp\left\{\int j_\omega(\lambda) \log S(\omega, \lambda) d\lambda\right\}$$

where  $j_\omega(\lambda)$  is the non-location Jeffreys prior for  $\lambda$  when  $\omega$  is fixed (not to be confused with  $j(\lambda|\omega)$ , the conditional of the non-location Jeffreys prior) and  $S = \sqrt{|I|/|I_{22}|}$ . Here,  $I$  is the Fisher information matrix and  $I_{22}$  is the portion of the  $I$  corresponding to the nuisance parameters.

As an example, we consider the Neyman-Scott problem discussed in Berger and Bernardo (1992b). The data consist of  $n$  pairs of observations:  $X_{ij} \sim N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, j = 1, 2$ . The non-location Jeffreys prior is  $\pi(\mu_1, \dots, \mu_n, \sigma) \propto \sigma^{-(n+1)}$ . Then  $E(\sigma^2|x) = s^2/(2n - 2)$  where  $s^2 = \sum_{i=1}^n \sum_{j=1}^2 (x_{ij} - \bar{x}_i)^2$  and  $\bar{x}_i = (x_{i1} + x_{i2})/2$ . Now  $E(\sigma^2|x) = s^2/(2n - 2)$  is inconsistent since  $s^2/n$  converges to  $\sigma^2$ . By treating  $\sigma$  as the parameter of interest, the Berger-Bernardo method leads to the prior  $\pi(\mu_1, \dots, \mu_n, \sigma) \propto \sigma^{-1}$  in accordance with with Jeffreys's general rule (2); this gives a posterior mean of  $s^2/(n - 2)$  which is consistent. There are other Bayesian ways to handle this problem. One might, for instance, introduce a hierarchical model by putting a distribution on the  $\mu'_i$ s and then apply Jeffreys's general rule to the hyperparameters, based on the marginal distribution of the data. But this is an example in which the Berger-Bernardo method yields a prior that seems reasonable when judged by the long-run sampling behavior the posterior; see Berger and Bernardo (1992b).

The Berger-Bernardo method has now been applied to many examples including exponential regression (Ye and Berger 1991) multinomial models (Berger and Bernardo 1992a), AR(1) models (Berger and Yang 1992) and the product of Normal means problem (Berger and Bernardo 1989) to name just a few. Typically, this method leads to priors that are

There are now many papers with examples

In all the above discussion, we have lumped the parameters into two groups: parameter of interest and nuisance parameters. Berger and Bernardo (1991, 1992a, 1992b) and Ye and Berger (1991) have extended the method to deal with parameters that have been lumped into any number of ordered groups. The ordering is supposed to reflect the degree of importance of the different groups. Generally, different orderings produce different priors. One way to assess the sensitivity of the posterior to the prior is to consider the priors arising from various orderings of the parameters. If the posterior is similar for all these priors then we have some evidence that the posterior is robust to the choice of prior.

### 3.5.3 Related Work

Ghosh and Mukerjee (1992a) and Clarke and Wasserman (1992, 1993) proposed alternatives to the Berger-Bernardo method that use Bernardo's missing information idea in a different way. Specifically, they work directly with  $\widetilde{K}_\infty^\pi(\omega)$ , the standardized missing information for  $\omega$ , i.e., the asymptotic expected Kullback distance between the marginal prior  $\pi(\omega)$  and the marginal posterior  $\pi(\omega|X_1^n)$  minus a standardizing constant:

$$\widetilde{K}_\infty^\pi(\omega) = \int \int p(\omega, \lambda) \log \frac{S}{p(\omega)} d\omega d\lambda$$

where  $S = \{|I| |I_{22}|^{-1}\}^{1/2}$ ,  $I$  is the Fisher information matrix and  $I_{22}$  is the part of the Fisher information matrix corresponding to  $\lambda$ .

Ghosh and Mukerjee (1992a) showed that maximizing  $K_\infty^\pi(\omega)$  subject to the condition that  $\pi(\lambda|\omega) = j_\omega(\lambda)$  gives the Berger-Bernardo prior. Thus the Berger-Bernardo prior maximizes the missing information for  $\omega$  subject to the condition that given  $\omega$ , the missing information for  $\lambda$  is maximized. But it seems reasonable to examine priors that maximize  $\widetilde{K}_\infty^\pi(\omega)$ .

Ghosh and Mukerjee conjectured, and Clarke and Wasserman showed, that priors that maximize  $\widetilde{K}_\infty^\pi(\omega)$  typically are degenerate. Clarke and Wasserman proposed a tradeoff prior  $\pi_\alpha$  that maximizes  $\widetilde{K}_\infty^\pi(\omega) - \alpha K(j, \pi)$  where the latter term is a penalty term that measures distance from a prior  $j$  where  $j$  is usually taken to be the Jeffreys prior or the non-location Jeffreys prior. The interpretation is that we are trying to make the distance between the prior for  $\omega$  and the posterior for  $\omega$  far apart but we add a penalty term to ensure that the prior does not depart too far from  $j$ . Without the penalty term, degenerate priors can result. Generally,  $\pi_\alpha$  cannot be written in closed form but Clarke and Wasserman (1993) gave an algorithm for computing it. Ghosh and Mukerjee suggested shrinking towards a uniform prior. Later, Clarke and Wasserman (1992) proposed maximizing  $\widetilde{K}_\infty^\pi(\omega) - \alpha K(\pi, j)$ . The solution is  $\pi_\alpha \propto hH^{-1/(\alpha+1)}$  where  $h = S^{1/\alpha}j(\omega, \lambda)$ ,  $H = \int h d\lambda$ , and, as before,  $S = \sqrt{|I|/|I_{22}|}$ . This reduces to  $j$  when  $\alpha \rightarrow \infty$  and, if  $S$  is a function of  $\omega$  only then it reduces to the Berger-Bernardo prior when  $\alpha = 0$ . More generally,  $\pi_\alpha$  converges to a degenerate distribution when  $\alpha \downarrow 0$  but, strangely, may still agree with the Berger-Bernardo prior when  $\alpha = -1$ .

The Berger-Bernardo program involves maximizing missing information for  $\lambda$  given  $\omega$ , then forming the marginal model and maximizing missing information for  $\omega$ . If  $\omega$  is the parameter of interest then perhaps we should maximize missing information for  $\omega$  given  $\lambda$ . That would ensure that missing information is maximized for  $\omega$  whatever the value of the nuisance parameter. Berger (1992) notes that such a scheme may give results that are similar to the coverage matching methods (section 3.7). Unfortunately, the prior will then depend on the parameterization of the nuisance parameter.

### 3.6 Geometry

The straightforward verification of invariance of Jeffreys's general rule hides its origin. In outline, Jeffreys (1946, 1961) noted that the Kullback-Leibler number behaves locally like the square of a distance function determined by a Riemannian metric; the natural volume element of this metric is  $\det(I(\theta))^{1/2}$ ; and natural volume elements of Riemannian metrics are automatically invariant to reparameterization. See Kass (1989, sections 2.1.2 and 2.1.3) for explication of this argument in the case of multinomial distributions.

Jeffreys treated the procedure formally, but Kass (1989, section 2.3) elaborated, arguing that natural volume elements provide appropriate generalizations of Lebesgue measure by capturing intuition favoring “flat” priors; while the information metric may be motivated by statistical considerations. Thus, the suggestion is that Jeffreys’s rule is based on an appealing heuristic. The key idea here is that natural volume elements generate “uniform” measures on manifolds, in the sense that equal mass is assigned to regions having equal volumes, and this uniformity seems to be what is appealing about Lebesgue measure. Since Fisher information is central in asymptotic theory, it seems a natural choice for defining a metric to generate a distribution that would serve as a pragmatic substitute for a more precise representation of *a priori* knowledge.

It is also possible to use this geometrical derivation to generate alternative priors by beginning with some discrepancy measure other than the Kullback-Leibler number, and defining a Riemannian metric and then a natural volume element. Specification of this idea was given in unpublished manuscripts by Kass (1981) and George and McCulloch (1989). It was also mentioned by Good (1969).

### 3.7 Coverage Matching Methods

One way to try to characterize “noninformative” priors is through the notion that they ought to “let the data speak for themselves.” A lingering feeling among many statisticians is that frequentist properties may play a role in giving meaning to this appealing phrase. From this point of view it is considered desirable to have posterior probabilities agree with sampling probabilities.

To be specific, suppose that  $\theta$  is a scalar parameter and that  $\ell(x)$  and  $u(x)$  are such that  $Pr(\ell(x) \leq \theta \leq u(x)|x) = 1 - \alpha$  so that  $A_x = [\ell(x), u(x)]$  is set with posterior probability content  $1 - \alpha$ . One can also consider the frequency properties of  $A_x$  under repeated sampling. In general, the coverage of  $A_x$  will not be  $1 - \alpha$ . There are, however, some examples where coverage and posterior probability do agree. For example, if  $X \sim N(\theta, 1)$  and  $\theta$  is given a uniform prior then  $A_x = [x - n^{-1/2}z_{\alpha/2}, x + n^{-1/2}z_{\alpha/2}]$  has posterior probability  $1 - \alpha$  and also has coverage  $1 - \alpha$ , where  $Pr(Z > z_c) = c$  if  $Z \sim N(0, 1)$ . Jeffreys (1961) noted

the agreement between his methods and Fisher's in many Normal-theory problems; see also Box and Tiao (1973). Lindley (1958) showed that for a scalar parameter and a model that admits a real-valued sufficient statistic, the fiducial based confidence intervals agree with some posterior if and only if the problem is a location family (or can be transformed into such a form). A very general result for group transformation models, essentially due to Stein (1965) but proved elegantly by Chang and Villegas (1986), is that repeated-sampling coverage probabilities and posterior probabilities agree when the prior on the group is right Haar measure. (See Section 3.2.) In multiparameter problems, it may be the case that priors which lead to frequentist regions jointly, do not do so for each individual component simultaneously. This point is discussed in the context of the multivariate Normal problem by Geisser and Cornfield (1963).

We emphasize that some authors see the good frequentist properties of certain posterior intervals as providing a vehicle for justifying certain non-Bayesian methods, but do not argue that such properties in any sense justify the choice of a prior. Jeffreys (1961) is certainly in this group, as are Box and Tiao (1973) and Zellner (1971). Others, however, such as Berger and Bernardo (1989), Berger and Yang (1992, 1993) use coverage properties to discriminate among alternative candidate prior distributions.

Sometimes it is not possible to get exact agreement (see Bartholomew 1965) and instead we might seek approximate agreement. Let  $B_\alpha$  be a one-sided posterior region with posterior probability content  $1 - \alpha$ . Welch and Peers (1963) showed that, under certain regularity conditions, the confidence coverage of  $B_\alpha$  is  $1 - \alpha + O(n^{-1/2})$ . However, if (1) is used then the region has coverage  $1 - \alpha + O(n^{-1})$ . Hence, another justification for (1) is that it produces accurate confidence intervals.

This work was further examined and extended by Welch (1965), Peers (1965), Peers (1968) and Stein (1985). Recently, there has been interest in extending the Welch-Peers results when the parameter  $\theta$  has been partitioned into a parameter of interest  $\omega$  and nuisance parameters  $\lambda = (\lambda_1, \dots, \lambda_k)$ . Some progress was made on this in Peers (1965) and Stein (1985). Based on the Stein paper, Tibshirani (1989) showed that a prior that leads to accurate confidence intervals for  $\omega$  can be obtained as follows. Let  $I$  denote the Fisher

information matrix and let  $\ell$  be the log-likelihood function. Write

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

where  $I_{11} = -E \left( \frac{\partial^2 \ell}{\partial \omega^2} \right)$ ,  $I_{22}$  is the  $k \times k$  matrix with  $ij^{th}$  entry  $-E \left( \frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_j} \right)$ ,  $I_{12}$  is the  $k \times 1$  matrix with  $j^{th}$  entry  $-E \left( \frac{\partial^2 \ell}{\partial \omega \partial \lambda_j} \right)$  and  $I_{21}$  is the  $1 \times k$  matrix with  $i^{th}$  entry  $-E \left( \frac{\partial^2 \ell}{\partial \lambda_i \partial \omega} \right)$ . Now, reparameterize the model as  $(\omega, \gamma)$  where  $\gamma = (\gamma_1, \dots, \gamma_k)$  is orthogonal to  $\omega$ . Here  $\gamma_i \equiv \gamma(\omega, \lambda_1, \dots, \lambda_k)$ . Orthogonality means that  $I_{12} = I_{21} = 0$ ; see Cox and Reid (1987). Tibshirani suggests that the prior  $\pi(\omega, \gamma) \propto g(\lambda) I_{11}^{1/2}$  produces accurate confidence intervals for  $\omega$ , where  $g(\lambda)$  is an arbitrary, positive function of  $\lambda$ . This result was made rigorous by Nicolaou (1993). For comparison, note that (1) is  $\pi(\omega, \gamma) \propto I_{11}^{1/2} I_{22}^{1/2}$  and the Berger-Bernardo prior (section 3.5) is  $\pi(\omega, \gamma) \propto f(\omega) I_{22}^{1/2}$  for some function  $f(\omega)$ . It is interesting that these confidence based methods seem to produce priors of the form that would be obtained from the Berger-Bernardo scheme if roles of the parameter of interest and nuisance parameter were switched; Berger (1992) comments on this fact.

Ghosh and Mukerjee (1992a) suggest requiring that

$$\int P_\theta(\omega \leq \omega_\alpha(X)) \pi(\lambda|\omega) d\lambda = 1 - \alpha + O(n^{-1})$$

where  $\omega_\alpha$  is such that  $P(\omega \leq \omega_\alpha(X)|X) = 1 - \alpha + O(n^{-1})$ . This leads to the condition

$$\pi(\omega) \propto \left( \int \frac{\pi(\lambda|\omega)}{I_{11}^{1/2}} d\lambda \right)^{-1}.$$

Mukerjee and Dey (1992) found priors that match frequentist coverage to order  $o(n^{-1})$  and they give a differential equation that must be solved in order to find the prior. Tibshirani's method generally has solutions that leave part of the prior unspecified but in many cases, the Mukerjee-Dey method completely specifies the prior up to a constant. Ghosh and Mukerjee (1993) find priors such that  $P(\mathbf{W} \leq \mathbf{t}|\mathbf{X}) = P(\mathbf{W} \leq \mathbf{t}|\theta) + o(n^{-1/2})$  for all  $\theta$  and  $\mathbf{t} = (t_1, \dots, t_p)'$  where  $\mathbf{W} = (W_1, \dots, W_n)'$ ,  $W_1$  is an appropriately standardized version of  $\sqrt{n}(\theta_1 - \hat{\theta}_1)$  and  $W_i$  is a type of standardized regression residual of  $\sqrt{n}(\theta_i - \hat{\theta}_i)$

on  $\sqrt{n}(\theta_1 - \hat{\theta}_1), \dots, \sqrt{n}(\theta_{i-1} - \hat{\theta}_{i-1})$ . The priors are characterized as having to satisfy a certain differential equation. The idea is that  $\mathbf{W}$  is attempt to list the parameters in order of importance in the spirit of the work by Berger and Bernardo. Severini (1991) shows that under certain circumstances some priors will give HPD regions which agree with their nominal frequentist coverage to order  $n^{-3/2}$ . Similar calculations, but for which there is a scalar nuisance parameter, are considered in Ghosh and Mukerjee (1992b). DiCiccio and Stern (1992) find conditions on the prior so that coverage and posterior probability content agree to order  $n^{-2}$  when both the parameter of interest and the nuisance parameter are vectors. Connections between the Welch-Peers approach and frequentist approaches based on the signed square root of the likelihood ratio statistic are made in DiCiccio and Martin (1993). On a related topic, Severini (1993) shows how to choose intervals for which Bayesian posterior probability content and frequentist coverage agree to order  $n^{-3/2}$  for a fixed prior. Also, connections can be made between priors that produce good frequentist intervals and priors for which Bayesian and frequentist Bartlett corrections to the likelihood ratio statistic are  $o(1)$ ; see Ghosh and Mukerjee (1992b).

### 3.8 Zellner's Method

Let  $I(\theta) = \int f(x|\theta) \log f(x|\theta) dx$  be the information about  $X$  in the sampling density. Zellner (1971, 1977, 1993) and Zellner and Min (1992) suggest choosing the prior  $\pi$  that maximizes the difference  $G = \int I(\theta)\pi(\theta)d\theta - \int \pi(\theta) \log(\pi(\theta))d\theta$ . (Note that the negative entropy of the joint density of  $x$  and  $\theta$  is  $\int I(\theta)\pi(\theta)d\theta + \int \pi(\theta) \log(\pi(\theta))d\theta$ . Also note that  $G = \int \int \pi(\theta|x) \log[f(x|\theta)/\pi(\theta)]m(x)d\theta dx$  where  $m(x) = \int f(x|\theta)\pi(\theta)d\theta$ .) The solution is  $\pi(\theta) \propto \exp\{I(\theta)\}$ . He calls this prior, the maximal data information prior (MDIP). This leads to some interesting priors. In location scale problems, it leads to right-Haar measure. In the binomial  $(n, \theta)$  model it leads to the prior  $\pi(\theta) \propto \theta^\theta (1 - \theta)^{1-\theta}$  which has tail behavior in between that of (1) which in this case is  $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$ , and the uniform prior. MDIP priors for the Weibull are found in Sinha and Zellner (1990). Recently, Moulton (1993) obtained MDIP priors for the  $t$  family and the power exponential family.

Zellner's method is not parameterization invariant. However, Zellner (1991) points out

that invariance under specific classes of reparameterizations can be obtained by adding the appropriate constraints. For example, if we are interested in the transformations  $\eta_i = h_i(\theta)$ ,  $i = 1, \dots, m$  then he suggests maximizing

$$G = \int \pi(\theta)I(\theta)d\theta - \int \pi(\theta) \log \pi(\theta)d\theta + \sum_{i=1}^m [\int \pi_i(\eta_i)I(\eta_i)d\eta_i - \int \pi_i(\eta_i) \log \pi_i(\eta_i)d\eta_i]$$

subject to  $\pi(\theta)d\theta = \pi_i(\eta_i)d\eta_i$ . The solution is

$$\pi(\theta) \propto \exp\{I(\theta) + \sum_{i=1}^m \log |h'_i(\theta)|/(m + 1)\}.$$

The resulting prior then has the desired invariance properties over the given transformations. Other side conditions such as moment constraints can be added too. Zellner's prior can be related to (1) in the following way (Zellner, personal communication): maximize Zellner's functional subject to the condition that the expected value of the log square root of the Fisher information equals a constant. This leads to a prior proportional to  $j^\lambda(\theta) \exp\{I(\theta)\}$  where  $\lambda$  is a constant and  $j$  is from (1).

### 3.9 Decision-Theoretic methods

Several authors have used decision theoretic arguments to select priors. Chernoff (1954) derives the uniform prior on finite sets by way of eight postulates for rational decision making. Partitioning paradoxes are avoided since his argument is restricted to sets with fixed, given number of outcomes. Good (1969) takes a different approach. He defines  $U(G|F)$  to be "the utility of asserting that a distribution is  $G$  when, in fact, it is  $F$ ." He shows that if  $U$  takes on a particular form then (1) is the least favorable prior distribution. Good also relates this idea to Jeffreys's geometrical argument; see Section 3.6. See also Clarke and Barron (1990).

Hartigan (1965) calls a decision  $d(x)$  unbiased for the loss function  $L$  if

$$E_{\theta_0}(L(d(x), \theta)|\theta_0) \geq E_{\theta_0}(L(d(x), \theta_0)|\theta_0)$$

for all  $\theta, \theta_0$ . Hartigan shows that, if  $\theta$  is one-dimensional, a prior density  $h$  is asymptotically

unbiased if and only if

$$h(\theta) = E(\partial/\partial\theta \log f(x|\theta))^2 / (\partial^2/\partial\phi^2 L(\theta, \phi))_{\theta=\phi}^{1/2}.$$

If the loss function is Hellinger distance, this gives (1). Hartigan also extends this to higher dimensions.

Gatsonis (1984) considers estimating the posterior distribution as a decision problem using  $L_2$  distance as a loss function. The best invariant estimator of the posterior in a location problem is the posterior obtained from a uniform prior. He also shows that this estimate is inadmissible for dimension greater than 3.

Bernardo's method (section 3.5) may also be given a decision theoretic interpretation. Specifically, the Kullback-Leibler distance can be justified by viewing the problem of reporting a prior and posterior as a decision problem. Bernardo (1979b) shows that Kullback-Leibler divergence is the unique loss function satisfying certain desiderata. Polson (1988) also discusses this approach.

Kashyap (1971) considers the selection of a prior as a 2-person zero sum game against nature. Using the average divergence between the data density and the predictive density as a loss function, he shows that the minimax solution is the prior  $\pi(\theta)$  that minimizes  $E \log p(y|\theta)/\pi(\theta)$  where the expectation is with respect to the joint measure on  $y$  and  $\theta$ . Asymptotically, this leads to (1). This is very similar to Bernardo's (1979a) approach.

### 3.10 Rissanen's Method

Consider the problem of finding a reference prior for  $\Theta = \{1, 2, \dots\}$ . Many familiar techniques, like maximum entropy (3.4) do not give meaningful answers for finding a prior on  $\Theta$ . Jeffreys (1961, p. 238) suggested  $Q(n) \propto 1/n$  though he did not derive it from any formal argument.

Rissanen (1983) used the following coding theory motivation for a prior. Suppose you have to construct a code for the integers, that is you must assign a binary string to each integer. We assume that your code is a prefix code, which means that no codeword is allowed to be a prefix of another another codeword. This condition ensures that a decoder can detect

the beginning and end of each codeword. Let  $L = (L(1), L(2), \dots)$  be the codeword lengths. An adversary will choose an integer from a distribution  $P$ . Your task is to assign the codes so that the code lengths are as short as possible. More formally, you must try to minimize the inverse of the code efficiency which is defined to be the ratio of the mean code length to the entropy. This optimization problem can be expressed as

$$\minsup_L \sup_P \lim_{N \rightarrow \infty} \frac{\sum_{i=1}^N P(i)L(i)}{-\sum_{i=1}^N P(i) \log P(i)}.$$

The optimization is carried out subject to

- (i)  $P(i) < 1$  for all  $i$  and the sequence  $P(1), P(2), \dots$  is eventually decreasing,
- (ii)  $-\sum P(i) \log P(i) = \infty$
- (iii)  $0 < L(i) \leq L(i + 1)$  for all  $i$  and
- (iv)  $\sum 2^{-L(i)} \leq 1$ .

The last condition is called the Kraft inequality and is necessarily satisfied by a prefix code. Rissanen shows that there is a code with code lengths  $L_0(n) = \log^*(n) + \log c$  where  $\log^*(n) = \log x + \log \log x + \dots$  where only the finitely many terms of the sum that are positive are included and  $c \approx 2.865064$ . Furthermore, any optimal length  $L$  satisfies  $\log n < L(n) < \log n + r(n)$  where  $r(n)/\log n \rightarrow 0$  and  $r(n) \rightarrow \infty$  as  $n \rightarrow \infty$ . Rissanen then suggests we adopt  $Q(n) = 2^{-L_0(n)}$  as a universal prior for the integers. The Kraft inequality implies that the prior is proper. Since  $Q(n) \propto (1/n) \times (1/\log n) \times (1/\log \log n) \cdots$  we see that this will be close to the improper prior suggested by Jeffreys.

Rissanen's prior is interesting and might well be useful in some problems. There are some problems with the prior, however. First, there does not seem to be any convincing argument for turning the code length  $L$  into a prior. Second, since the prior is proper, we can find a constant  $n_0$  such that  $Q(\{1, \dots, n_0\}) \approx 1$  and in certain problems this will not be appropriate. Finally, note that any prior of the form  $R(n) \propto (1/\sigma)Q(n/\sigma)$  has the same tail behavior as Rissanen's prior and could equally well be used in place of  $Q(n)$ .

### 3.11 Other Methods

Novick and Hall (1965) define an “indifference prior” by identifying a conjugate class of priors and then selecting a prior from this class that satisfies two properties: first, the prior should be improper; second, a “minimum necessary sample” should induce a proper posterior. In a binomial problem for example, with the class of Beta priors, they obtain the prior  $\{p(1-p)\}^{-1}$  as an indifference prior. This prior is improper, but a single success and a single failure induce a proper posterior. Novick (1969) considers extensions to multiparameter problems.

Hartigan (1971, 1983 section 5.5) defines the similarity of events  $E$  and  $F$  by  $S(E, F) = P(E \cap F)/(P(E)P(F))$ . For random variables  $X$  and  $Y$  with joint density  $f_{X,Y}$  and marginal densities  $f_X$  and  $f_Y$  the definition is  $s(x, y) = f_{X,Y}(x, y)/(f_X(x)f_Y(y))$  whenever the ratio is well-defined. Then (1) can be justified in two ways using this approach: it makes present and future observations have constant similarity, asymptotically and it maximizes the asymptotic similarity between the observations and the parameter.

Piccinato (1978) considers the following method. A point  $\xi_0$  is a *representative point of the probability  $P$*  if  $\phi(\xi, P)$  is minimized by  $\xi_0$  where  $\phi$  is some discrepancy measure; an example is  $\phi(\xi, P) = \int |\xi - x|^k dP$ . A predictive distribution  $f(y|x)$  is *conservative* if the data point is always a typical point. The prior is called noninformative if it produces a conservative prediction. In a binomial problem with conjugate priors, and using the mean as a typical point, we the prior  $\{\theta(1 - \theta)\}^{-1}$ . A Normal with a Normal-gamma prior gives  $\pi(\mu, \sigma) \propto \sigma^{-3}$ .

Using finitely additive priors for an exponential model, Cifarelli and Regazzini (1987) show that a large class of priors give perfect association between future and past observations in the sense that there are functions  $\phi_n : \mathbb{R}^n \rightarrow \mathbb{R}$  such that

$$P(X_N \leq x, \phi_n(X_1, \dots, X_n) \leq x) = P(X_N \leq x) = P(\phi_n(X_1, \dots, X_n) \leq x)$$

for all  $N > n$ ,  $n = 1, 2, \dots$  and  $x \in \mathbb{R}$ . Under certain conditions, they show that the only prior that gives  $E(X_N|X_1, \dots, X_n) = \bar{X}_n$  is the usual improper uniform prior. In a related paper, (Cifarelli and Regazzini 1983) these authors show that the usual conjugate priors for the exponential family are the unique priors that maximize the correlation between  $X_N$  and

$\bar{X}_n$  subject to fixed values of  $Var(E(X_n|\theta))/Var(X_n)$ .

Spall and Hill (1990) define a least informative prior by finding the prior that maximizes expected gain in Shannon information. They approximate this by only looking over convex combinations of a set of base priors. As shown in Berger, Bernardo and Mendoza (1989), maximizing this measure can lead to discrete priors. Indeed, this is why Berger and Bernardo maximize this quantity asymptotically.

## 4 Issues

In this section we discuss four general issues beginning, in Section 4.1, with the interpretation of reference priors, where we argue that it is not necessary to regard a reference prior as being noninformative for it to be useful. Reference priors are often improper and may depend on the experimental design. We discuss consequences of these characteristics in Sections 4.2 and 4.3, respectively. In Section 4.4 we consider the possibility of performing sensitivity analysis in conjunction with the use of reference priors.

### 4.1 Interpretation of reference priors

At the risk of over-simplification, it seems useful to identify two interpretations of reference priors. The first asserts that reference priors are formal representations of ignorance; the second asserts that there is no objective, unique prior that represents ignorance. Instead, reference priors are chosen by public agreement, much like units of length and weight. In this interpretation, reference priors are akin to a default option in a computer package. We fall back to the default when there is insufficient information to otherwise define the prior.

Let us pursue the second interpretation a bit further. In principle, we could construct a systematic catalogue of reference priors for a variety of models. The priors in the catalogue do not represent ignorance. Still, the priors are useful in problems where it is impractical to elicit a subjective prior. The statistician may feel that the reference prior is, for all practical purposes, a good approximation to any reasonable subjective prior for that problem.

The first interpretation was, at one time, the dominant interpretation and much effort

was spent trying to justify one prior or another as being noninformative (see section 2). For the most part, the mood has shifted towards the second interpretation. In the recent literature, it is rare that anyone makes any claim that a particular prior can logically be defended as being truly noninformative. Instead, the focus is on investigating various priors and comparing them to see if any have advantages in any practical sense. For example, Berger and Bernardo (1989) consider several priors for estimating the product of two Normal means. Rather than defending any particular prior on logical grounds, they instead compare the frequency properties of the credible regions generated by the priors. This is an example of using an ad-hoc but practically motivated basis for defending a reference prior instead of a formal logical argument.

A slight variant on the second interpretation is that, although the priors themselves do not formally represent ignorance, our willingness to use a reference prior does represent our ignorance – or at least it is acting *as if* we were ignorant. That is, according to this interpretation, when we decide to use a reference prior, the decision itself may be regarded as an admission of ignorance in so far as we are apparently unable (or we act as if we were unable) to determine the prior subjectively.

## 4.2 Impropriety

Many reference priors are improper, that is, they do not integrate to a finite number. In this section we discuss five problems caused by improper priors: (i) incoherence and strong inconsistencies, (ii) the dominating effect of the prior, (iii) inadmissibility, (iv) marginalization paradoxes and (v) impropriety of the posterior.

### 4.2.1 Incoherence, Strong Inconsistencies and Non-conglomerability

An example from Stone (1976, 1982) nicely illustrates potential inconsistencies in using improper priors. Suppose we flip a four sided die many times. The four faces of the die are marked with with the symbols  $\{a, b, a^{-1}, b^{-1}\}$ , respectively. Each time we toss the die we record the symbol on the lowermost face of the die – there is no uppermost face on a four-sided die. The tosses result in a string of letters. Any time the symbols  $a$  and  $a^{-1}$

are juxtaposed in our list, they “annihilate” each other, that is, they cancel each other out. Similarly for  $b$  and  $b^{-1}$ . For example, if we tossed the die four times and obtained  $(a b b^{-1} a)$ , then the resulting string is  $(a a)$  since  $b$  and  $b^{-1}$  annihilate each other. Denote the resulting string by  $\theta$ . (To avoid annoying edge effects, we will assume that the length of  $\theta$  is large so that the possibility of a null string is eliminated.) Now we suppose that one additional toss of the die is made and the resulting symbol is added to  $\theta$ . The annihilation rule is applied, if appropriate, resulting in a new string  $x$ . The problem is to infer  $\theta$  from  $x$ .

Having seen  $x$  we note that there are four possible values for  $\theta$ , each with equal likelihood. For example, suppose  $x = (a a)$ . The extra symbol added by the last toss was either  $a$ ,  $a^{-1}$ ,  $b$  or  $b^{-1}$  each with probability  $1/4$ . So,  $\theta$  is one of  $(a)$ ,  $(a a a)$ ,  $(a a b^{-1})$  or  $(a a b)$  each having likelihood  $1/4$ . If we adopt a flat prior on  $\theta$  and formally apply Bayes rule the posterior will give probability  $1/4$  to each of these points and will have zero probability elsewhere. Denote the mass function of this posterior by  $\pi(\theta|x)$ . Let  $A$  be the event that the last symbol selected resulted in an annihilation. We see that  $P(A|x) = 3/4$  for every  $x$ . On the other hand, for fixed  $\theta$ , a new symbol results in annihilation with probability  $1/4$ , i.e.  $P(A|\theta) = 1/4$  for every  $\theta$ . These two probability statements are contradictory. Since  $P(A|x) = 3/4$  for every  $x$  it seems we should conclude that  $P(A) = 3/4$ . But since  $P(A|\theta) = 1/4$  for every  $\theta$  it seems we should conclude that  $P(A) = 1/4$ . Stone called such a phenomenon a *strong inconsistency*. It is also an example of a super-relevant betting procedure (Robinson 1979a, 1979b) and is related to a consistency principle in Bondar (1977).

To see what went wrong, let us think about the improper prior as a limit of proper priors. Let  $\pi_p$  be uniform on all strings of length  $p$ . It can be shown that, for fixed  $x$ ,  $\pi_p(A|x)$  tends to  $3/4$  as  $p \rightarrow \infty$ . It is tempting to argue that the posterior is valid since it approximates the posterior using the proper prior  $\pi_p$ . But  $\pi_p$  induces a marginal probability  $m_p$  on  $x$   $m_p(x) = \sum_{\theta} f(x|\theta)\pi_p(\theta)$ . Let  $X_p$  be the set of  $x$ 's of length  $p$  or  $p+1$ . When  $x \in X_p$ ,  $\pi_p(\theta|x)$  is concentrated on a single point and so  $\pi(\theta|x)$  is a terrible approximation to  $\pi_p(\theta|x)$ . Recall that  $\pi(\theta|x)$  gives equal mass to four points. The total variation distance between  $\pi(\cdot|x)$  and  $\pi_p(\cdot|x)$  is thus  $3/4$  for  $x \in X_p$ . Stone showed that  $m_p(X_p)$  tends to  $2/3$ ; this is the essence of the problem. Although  $\pi_p(\cdot|x)$  converges to  $\pi(\cdot|x)$  for fixed  $x$ , it does not follow that the two are close with increasingly high probability. This led Stone to suggest that we should seek

posteriors with the property that the total variation distance between the formal posterior based on an improper prior and the posterior from a proper prior should tend in probability to 0 for some sequence of proper priors; see Stone (1963, 1965, 1970).

It turns out that strong inconsistencies and Stone's proposal for avoiding them, are closely tied to the notion of coherence developed in a series of papers by Heath, Lane and Sudderth (HLS) (Heath and Sudderth 1978, 1989, Lane and Sudderth 1983). (Their notion of coherence is slightly stronger than the notion of coherence introduced by de Finetti (1937, 1972, 1974, 1975)). In their framework, probabilities are allowed to be finitely, rather than countably additive. To see the difference between finitely additive priors and improper priors let  $P_n$  be the uniform measure on  $[-n, n]$  and define  $P$  by  $P(A) = \lim_{n \rightarrow \infty} P_n(A)$  for all  $A$  for which the limit exists.  $P$  is an example of a finitely additive prior on the real that is diffuse in the sense that it gives zero probability to every compact set. On the other hand,  $P$  is proper since  $P(\mathbb{R}) = 1$ . Compare this to Lebesgue measure  $\mu$  which gives positive measure to many compact sets but which is improper since  $\mu(\mathbb{R}) = \infty$ . One way to connect these two concepts in practice is to start with an improper prior and, as in the example just considered, generate a finitely additive prior by way of a limit of truncated proper priors.

Formally, the HLS approach, which is inspired by Freedman and Purves (1969), begins with a sample space  $\mathcal{X}$  and a parameter space  $\Theta$ . Let  $\mathcal{B}(X)$  and  $\mathcal{B}(\Theta)$  be  $\sigma$ -fields on these spaces. A model is a collection of probabilities  $\{p_\theta; \theta \in \Theta\}$  on  $\mathcal{B}(X)$ . An inference is a collection of probabilities  $\{q_x; x \in \mathcal{X}\}$  on  $\mathcal{B}(\Theta)$ . For a bounded function  $\phi$  and a probability  $P$  write  $P(\phi) = \int \phi dP$ .

A prior  $\pi$  on  $\Theta$  defines a marginal  $m$  on the sample space  $\mathcal{X}$  by way of the equation  $m(\phi) = \int p_\theta(\phi) \pi(d\theta)$  for all bounded  $\phi : X \rightarrow \mathbb{R}$ . An inference is *coherent* if it is not possible to place a finite number of bets, using odds based on  $q_x$ , to guarantee an expected payoff that is greater than a positive constant, for every  $\theta$ . Heath and Sudderth (1978) show that an inference  $\{q_x; x \in \mathcal{X}\}$  is coherent if and only if there exists a prior  $\pi$  such that

$$\int \int \phi(\theta, x) p_\theta(dx) \pi(d\theta) = \int \int \phi(\theta, x) q_x(d\theta) m(dx)$$

for all bounded  $\phi : \Theta \times X \rightarrow \mathbb{R}$  that are measurable with respect to  $\mathcal{B}(\Theta) \times \mathcal{B}(X)$ , where

$m$  is the marginal induced by the prior  $\pi$ . This means that the joint measure can be disintegrated with respect to the  $\theta$  partition or the  $x$  partition without contradiction. We call  $q_x$  a posterior of  $\pi$ . Heath and Sudderth (1989, Theorem 3.1) prove that an inference  $\{\tilde{q}_x; x \in \mathcal{X}\}$  is coherent if and only if it can be approximated by proper priors in the sense that  $\inf f \|q_x - \tilde{q}_x\| m(dx) = 0$  where the infimum is over all (proper but possibly finitely additive) priors  $\pi$  where  $q$  is the posterior of  $\pi$ ,  $m$  is the induced marginal and  $\|\cdot\|$  is total variation norm. This is Stone's proposed condition except that HLS allow for finitely additive distributions. Coherence, in the HLS sense, is essentially the same as requiring that there be no strong inconsistency; see Lane and Sudderth (1983). It is worth noting that incoherence can arise in standard statistical models. For example, Eaton and Sudderth (1993a) recently showed that the right Haar prior for MANOVA models gives an incoherent posterior. Another example of incoherence for commonly used priors is given in Eaton and Sudderth (1993b).

In fact, incoherence and strong inconsistencies are manifestations of a phenomenon called non-conglomerability which plagues every probability measure that is finitely but not countably additive. A probability  $P$  is conglomerable with respect to a partition  $\mathcal{B}$  if for every event  $A$ ,  $k_1 \leq P(A|B) \leq k_2$  for all  $B \in \mathcal{B}$  implies that  $k_1 \leq P(A) \leq k_2$ . The Stone example exhibits non-conglomerability for the following reason. Since  $P(A|x) = 3/4$  for all  $x$ , conglomerability would imply  $P(A) = 3/4$ . Similarly,  $P(A|\theta) = 1/4$  for all  $\theta$  implies  $P(A) = 1/4$ . This contradiction implies that either the  $x$  partition or the  $\theta$  partition or both must display non-conglomerability. The import of HLS coherence is to rule out non-conglomerability in the  $\theta$  and  $x$  margins. But we should not be sanguine just because conglomerability holds in these two margins. For one thing, HLS coherence is not always preserved under conditioning or under convex combinations (Kadane, Schervish and Seidenfeld 1986). Furthermore, HLS coherence only guarantees protection from nonconglomerability in the  $\theta$  and  $x$  partitions of the joint space  $\Theta \times X$ . There is no guarantee that other strong inconsistencies cannot occur in other margins. In fact, every finitely additive probability that is not countably additive displays non-conglomerability in at least on margin (Schervish, Seidenfeld and Kadane, 1984; Hill and Lane, 1986).

The HLS approach is only one among many ways to strengthening De Finetti's notion

of coherence. Other related ideas have been considered by many authors, among them are: Akaike (1980), Berti, Regazzini and Rigo (1991), Buehler (1959), Buehler and Feddersen (1963), Bondar (1977), Brunk (1991), Dawid and Stone (1972, 1973), Hartigan (1983), Pierce (1973), Regazzini (1987), Robinson (1978, 1979a,b), Seidenfeld (1981) and Wallace (1959). One particular alternative that is worth mentioning is the notion using uniform approximations. For example, Mukhopadhyay and Das Gupta (1993) showed the following: consider a location families that possess a moment generating function. Let  $\pi^x$  be the posterior using a flat prior. For every  $\epsilon > 0$  there exists a proper, countably additive prior  $q$  with posterior  $q^x$  such that  $d(\pi^x, q^x) < \epsilon$  for all  $x$ . (This implies HLS coherence). It remains an open question how far this approach can be taken.

#### 4.2.2 The Dominating Effect of the Prior

Sometimes, reference priors can overwhelm the data even though the posterior is HLS coherent. A famous example of this is the many Normal means problem. Let  $X_i \sim N(\theta_i, 1)$  independently, where  $i = 1, \dots, n$  and consider the problem of estimating  $\xi = \sum \theta_i^2$ . If we adopt a flat prior on  $\theta = (\theta_1, \dots, \theta_n)'$  then the posterior for  $\theta$  is multivariate Normal with mean  $X = (X_1, \dots, X_n)'$  and covariance equal to the identity matrix  $I$ . This posterior is coherent in the sense described in section 4.2.1. The posterior  $Q(d\xi|x)$  for  $\xi$  is a non-central  $\chi^2$  with  $n$  degrees of freedom and non-centrality parameter  $Y = \sum_i X_i^2$ ; we denote this by  $\xi|x \sim \chi_n^2(Y)$ . Hence,  $\hat{\xi} = E(\xi|X_1, \dots, X_n) = Y + n$ . There are reasons for thinking that  $\hat{\xi}$  is too large, as we now discuss.

Let  $\theta$  have a  $N(0, aI)$  prior. The posterior  $Q_a(d\xi|x)$  for  $\xi$  is such that  $\xi \sim [a/(a+1)] \cdot \chi_n^2(aY/(a+1))$ . The posterior  $Q$  approximates  $Q_a$  when  $a$  is large but the means of  $Q$  and  $Q_a$  are quite different. In fact, the expected value of  $\hat{\xi}_a = E_{Q_a}(\xi)$  with respect to the marginal  $m_a$  for  $x$  induced by the  $N(0, aI)$  prior is  $\hat{\xi}_0 = Y - n$ . (This is also the U.M.V.U.E. for this problem.) This suggests that we can expect  $\hat{\xi}_a$  to be close to  $\hat{\xi}_0$ . Perlman and Rasmussen (1975) confirm this intuition by showing  $|\hat{\xi}_0 - \hat{\xi}_a| = o_p(\sqrt{n})$  and  $|\hat{\xi} - \hat{\xi}_a| = o_p(\sqrt{n}) + 2n$ . In summary,  $Q(d\xi|x)$  and  $Q_a(d\xi|x)$  tend to be close in distributional distance but their means are not close. (There is no contradiction between these two statements: if  $Z_1 \sim N(0, a^2)$  and  $Z_2 \sim N(1, a^2)$  then  $E(Z_1) - E(Z_2) = 1$  for all  $a$  but the total variation distance between

the two distributions tends to 0 as  $a \rightarrow \infty$ .) This shows that closeness in distributional distance, which is what coherence is all about, may not be strong enough to avoid undesirable properties.

Similar problems occur with interval estimation for  $\xi$ . Under the posterior  $Q$ , a one-sided  $\alpha$ -level credible region for  $\xi$  is  $[\Phi_{\alpha,n}(Y), \infty)$  where  $P(\chi_n^2(Y) > \Phi_{\alpha,n}(Y)) = \alpha$ . Stein (1959) shows that the coverage probability of this interval tends to 0 as  $n \rightarrow \infty$ . The strong disagreement with the confidence level suggests something is amiss. (In his proof, Stein assumes that  $\xi = o(n^2)$  which, it might be argued, is implicitly assuming some prior information; indeed, Pinkham (1966) shows that if instead  $\xi = Mn^h + o(1)$  where  $M > 0$  and  $h > 2$  then the coverage and posterior probability agree asymptotically.)

What are we to make of this? The problem is that a posterior  $Q$  based on an improper prior may have moments quite different from a posterior  $Q_a$  based on a proper prior even though  $Q$  and  $Q_a$  may be close in distributional distance. Generally, this problem is not serious unless the dimension of the parameter space is large. The message from this and similar examples is that improper priors must be used with care when the dimension of the parameter space is large. Of course, that does not imply that subjective priors are necessarily any better in these problems. As long as the dimension is large and the data set is small, all priors must be used with care.

### 4.2.3 Inadmissibility

Under certain conditions, Bayes estimators based on proper priors lead to admissible estimators but that improper priors can lead to inadmissible Bayes estimators. Consider the many Normal means problem from the previous subsection. Stein (1956) showed that the posterior mean using a flat prior is an admissible estimator of  $\theta$  under squared error loss if  $n \geq 3$ . Thus, if  $L(\theta, \delta) = \sum(\theta_i - \delta_i)^2$  then the Bayes estimator arising from the flat prior, namely,  $X = (X_1, \dots, X_n)'$  is such that there exists another estimator  $\gamma = (\gamma_1, \dots, \gamma_n)'$  with the property that  $E_\theta L(\theta, \gamma) \leq E_\theta L(\theta, X)$  for every  $\theta$ , with strict inequality for at least one  $\theta$ . (In fact, one can construct estimates that uniformly beat  $X$ .)

Although,  $X$  is inadmissible in the many Normal means problem, it is extended admissible (Heath and Sudderth 1978). This means that there does not exist an  $\epsilon > 0$  and an estimator

$\delta_0$  such that  $E_\theta L(\theta, \delta_0) < E_\theta L(\theta, X) - \epsilon$  for all  $\theta$ . Thus, there is no estimator that beats  $X$  uniformly. In general, every Bayes rule is extended admissible (even if the prior is only finitely additive). If the loss function is bounded and the set of decision rules is convex then every extended admissible rule is Bayes (Heath and Sudderth 1978, Theorem 2). But, as we have seen, this does not guarantee admissibility.

Eaton (1992) gave conditions under which the Bayes rule from an improper prior produces admissible decision rules for a class of decision problems called “quadratically regular decision problems.” He showed that these conditions are equivalent to the recurrence of a Markov chain with transition function  $R(d\theta|\eta) = \int_{\mathcal{X}} Q(d\theta|x)P(dx|\eta)$  where  $\mathcal{X}$  is the sample space,  $Q(d\theta|x)$  is the posterior and  $P(dx|\theta)$  is the sampling model. He shows that some prediction problems are included in this class of decision problems.

Another approach to choosing priors is to look for priors that are on the “boundary between admissibility and inadmissibility.” This approach is considered in Berger and Strawderman (1993).

#### 4.2.4 Marginalization Paradoxes

Suppose we have a model  $f(x|\alpha, \beta)$  and prior  $\pi(\alpha, \beta)$  and the marginal posterior  $\pi(\alpha|x)$  satisfies  $\pi(\alpha|x) = \pi(\alpha|z(x))$  for some function  $z(x)$ . If  $f(z|\alpha, \beta) = f(z|\alpha)$  but  $f(z|\alpha)\pi(\alpha)$  is not proportional to  $\pi(\alpha|z(x))$  for any  $\pi(\alpha)$ , then we have marginalization paradox since it seems we should be able to recover  $\pi(\alpha|x)$  from  $f(z|\alpha)$  and some prior  $p(\alpha)$ . Dawid, Stone and Zidek (1973) present many examples. Here, we consider example 1 from that paper.

$X_1, \dots, X_n$  are independent exponential random variables. The first  $\xi$  have mean  $1/\eta$  and the rest have mean  $1/(c\eta)$ , with  $c \neq 1$  known and  $\xi \in \{1, \dots, n-1\}$ . The prior for  $\eta$  is taken to be uniform. Let  $z_i = x_i/x_1$ ,  $i = 1, \dots, n$ . It turns out that the posterior is a function of  $z = (z_1, \dots, z_n)$  only. The probability density for  $z$  is

$$f(z|\eta, \xi) = f(z|\xi) \propto \left( \sum_1^\xi z_i + c \sum_{\xi+1}^n z_i \right) c^{-\xi}$$

which is a function of  $\xi$  only. But there is no choice of prior  $\pi(\xi)$  that makes  $f(z|\xi)\pi(\xi)$  proportional to  $\pi(\xi|x)$ .

This contradiction can happen only if the prior is improper. An analysis of the problem is contained in Dawid, Stone and Zidek (1973) and the ensuing discussion; also, see Hartigan (1983, page 28-29). Of course, the problem is that we cannot expect the rules of probability to hold when the measure has infinite mass. Sudderth (1980) shows that the marginalization paradox cannot happen if we treat improper priors as finitely additive priors and manipulate the probabilities carefully, according to the rules of finitely additive probability. An interesting debate about the meaning of this paradox is contained in Jaynes (1980) and the ensuing discussion by Dawid, Stone and Zidek.

#### 4.2.5 Improper Posteriors

Sometimes, improper priors lead to improper posteriors. Consider the following hierarchical model:

$$\begin{aligned} Y_i | \mu_i, \sigma &\sim N(\mu_i, \sigma^2) \\ \mu_i | \tau &\sim N(\mu, \tau^2) \end{aligned}$$

for  $i = 1, \dots, n$  where  $\sigma^2$  is known. A seemingly natural choice for a prior is  $\pi(\mu, \tau) \propto \{\tau\}^{-1}$  but this leads to an improper posterior (Berger 1985, p. 187).

In this problem application of Jeffreys's general rule, based on the marginal distribution of the data, i.e.,  $Y_i \sim N(\mu, \sigma^2 + \tau^2)$ , leads to a proper posterior (cf. the discussion of one-way ANOVA in Box and Tiao, 1973). It does so in many other problems as well, but there are counterexamples (in which Jeffreys's general rule leads to an improper posterior) and there are as yet no simple general conditions to ensure propriety. Ibrahim and Laud (1991) give conditions that guarantee proper posteriors for generalized linear models. Dey, Gelfand and Peng (1993) extend this work for some overdispersed generalized linear models. Results that apply in greater generality have not been discovered. For the most part, characterizing improper priors that give proper posteriors remains an open problem.

Sometimes, improper posteriors will reveal themselves by creating obvious numerical problems but this is not always the case. Because of increased computing power, analysts

use models of ever greater complexity which in turn makes it more difficult to check whether the posterior is proper. It would be helpful to have a diagnostic for detecting impropriety.

One way to avoid improper posteriors is to use proper priors. But this may not solve the problem. In situations where intuitively reasonable priors give rise to improper posteriors, it is often a sign that the likelihood is not highly informative. A proper prior might formally produce a proper posterior but it is likely that the posterior will be very sensitive to the choice of prior. Thus, situations in which improper posteriors arise from familiar reference priors must be treated with care.

### 4.3 Sample Space Dependence

Another problem with reference priors is that they are often dependent on the sample space, sometimes called “design dependent” or “experiment dependent”. For example, if we obtain several replications of a Bernoulli experiment, then (1) will depend on whether we used binomial sampling or negative binomial sampling. This is not only odd from the subjectivist point of view but is generally considered undesirable since it violates the likelihood principle, which states that two experiments that produce proportional likelihoods should produce the same inferences (Berger and Wolpert 1988). It could be argued that the choice of design is informative and so the prior should depend on the design. Nonetheless, design dependence leads to some problems.

Aside from violating the likelihood principle, sample space dependent priors lead to situations where the posterior depends on what order we receive the data. Yet, for a fixed prior, we get the same posterior no matter what order the data are processed, assuming independence. Suppose  $X_1$  is the number of successes in  $n$  tosses of a biased coin with success probability  $p$ . Then (1) gives  $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$  and the posterior is  $\pi_1(p|X_1) \propto p^{X_1-1/2}(1-p)^{n-X_1-1/2}$ . Now suppose we flip the coin until another head appears and suppose this takes  $r$  tosses. Using  $\pi_1$  as a prior and updating to include the new information we get the posterior  $\pi_2(p|X_1, r) \propto p^{X_1+1-1/2}(1-p)^{n-X_1+r-1-1/2}$ . On the other hand, if we did the experiment in reverse order, we would begin with (1) for the negative binomial, namely,  $\pi(p) \propto p^{-1}(1-p)^{-1/2}$ . Updating sequentially on  $X_2$  then  $X_1$  gives the posterior

$\pi_2(p|X_1, r) \propto p^{X_1+1-1}(1-p)^{n-X_1+r-1-1/2}$  so we get a different posterior depending on what order we process the data.

Another type of sample space dependence is illustrated by right Haar priors (section 3.2). Consider the following example from McCullagh (1992). Let  $x_1, \dots, x_n$  have a Cauchy  $(\mu, \sigma)$  distribution. The right Haar prior is  $\pi(\mu, \sigma) \propto 1/\sigma$ . Now, let  $y_i = 1/x_i, i = 1, \dots, n$ . Then, the  $y_i$ 's are distributed as Cauchy  $(\nu, \tau)$  where  $\nu = \mu/(\mu^2 + \sigma^2)$  and  $\tau = \sigma/(\mu^2 + \sigma^2)$ . Right Haar measure for  $(\nu, \tau)$  is  $\pi(\nu, \tau) \propto 1/\tau$ . Transforming to  $(\mu, \sigma)$  we get  $\pi(\mu, \sigma) \propto 1/(\sigma(\mu^2 + \sigma^2))$  which differs from the first prior. Thus, our choice of prior will depend on how we choose to represent the sample space. Put another way, we can get different right Haar priors depending on how we label the sample space.

Zellner (1993) has pointed out that his method (3.8) can explicitly handle design dependence by maximizing average information in a set of experiments simultaneously. This results in the geometric mean of the Zellner priors from each experiment.

## 4.4 Sensitivity Analysis

There now exists a substantial literature on sensitivity analysis in Bayesian inference. Extensive references are contained in Berger (1984, 1990), Walley (1991) and Wasserman (1992). Most of this work is directed at quantifying the sensitivity of the posterior to the choice of prior and assumes that prior is a proper, subjectively elicited prior or that at least some features of the prior have been subjectively elicited. There is virtually no work on sensitivity analysis with respect to reference priors.

Sensitivity analysis often proceeds by embedding the prior  $\pi$  in a large class of similar priors  $\Gamma$ . The simplest and class of priors is the  $\epsilon$  contaminated class defined by

$$\Gamma_\epsilon(\pi) = \{(1 - \epsilon)\pi + \epsilon Q; Q \in \mathcal{P}\}$$

where  $\mathcal{P}$  is the set of all priors and  $\epsilon \in [0, 1]$  represents the uncertainty in the prior. If  $g(\theta)$

be some function of interest it is straightforward to compute

$$\underline{E}_\epsilon(g|y) = \inf_{P \in \Gamma_\epsilon(\pi)} E_P(g|x) \quad \text{and} \quad \overline{E}_\epsilon(g|y) = \sup_{P \in \Gamma_\epsilon(\pi)} E_P(g|x).$$

These bounds may be plotted by  $\epsilon$  so we can assess the sensitivity to the prior. Now consider a  $N(\theta, 1)$  model with  $\pi(\theta) \propto c$ . An obvious way to use existing sensitivity techniques is to regard the posterior to be the limit of the posteriors obtained from the sequence of priors  $\pi_a$  as  $a \rightarrow \infty$  where  $\pi_a$  is uniform on  $[-a, a]$ . As noted in section 4.2.1 this notion can be made rigorous by using probability limits of posteriors though we shall not worry about that here. It turns out that  $\underline{E}_\epsilon(\theta|y) = -\infty$  and  $\overline{E}_\epsilon(\theta|y) = \infty$  if we define

$$\overline{E}_\epsilon(\theta|y) = \lim_{a \rightarrow \infty} \sup_{P \in \Gamma_\epsilon(\pi_a)} E_P(\theta|y).$$

Apparently, this is not a useful way to proceed.

This does not rule out the possibility of finding some other neighborhood structure that produces finite bounds for improper priors. DeRobertis and Hartigan (1981) found such a class defined in the following way: let  $\Gamma_k$  be all prior densities  $p$  such that

$$\frac{p(\theta)\pi(\phi)}{p(\phi)\pi(\theta)} \leq k$$

for almost all  $\theta, \phi$  where  $k$  varies from 1 to  $\infty$ . We call this a *density ratio class*. (They considered a more general class but we shall confine our attention to this special case.) Again it is easy to compute upper and lower bounds on posterior expectations. Even when  $\pi$  is improper, the bounds are usually finite and are easy to calculate. But this class achieves this pleasant behavior at the cost of being unrealistically small. For example, a  $\Gamma_k$  neighborhood of a  $N(0, 1)$  will never contain a  $N(a, 1)$  density if  $a \neq 0$ , no matter how large  $k$  is.

All this leads to the following question: Is there a class that is larger than the density ratio class and that gives non-trivial bounds on posterior expectations if we interpret the posterior as a limit of posteriors from proper priors? The answer is no. Wasserman (1992) showed that, subject to certain regularity conditions, any class that gives finite bounds for improper priors is contained in a density ratio class. Current work with C. Srinivasan is aimed at

building classes of priors by defining norms directly on the space of improper priors. It is too early to know how successful these techniques will be.

## 5 Discussion

Reference priors are a part of Bayesian statistical practice. Often, a data analyst chooses some parameterization and uses a uniform prior on it. Logical difficulties with this procedure prompted Jeffreys’s search for alternatives, which led to the developments we surveyed here.

Jeffreys’s notion was that a prior could be chosen “by convention” as a “standard of reference”. (We did not wish to imply an interchangeability of alternatives, and thus avoided the term “conventional prior”; for a philosophical discussion of the notion of conventionality see Sklar, 1976 p. 88-112.) The term “reference prior” is intended to connote standardization. There is a sense in which these priors serve as “defaults”, that is, choices that may be made automatically without any contemplation of their suitability in a particular problem. Indeed, it is entirely possible that in future Bayesian software such default selections will be available. This should not undermine or replace inherently subjective judgment, but rather acknowledges the convenience that standardization provides.

As we have seen, there are situations in which reference priors have undesirable properties and consequences. These include incoherence, inadmissibility, marginalization paradoxes, sample space dependence, impropriety of the posterior, and unsuspected marginal effects in high-dimensional problems. In practice, the most serious and worrisome of these are probably the latter two, though the others have collectively sent a strong signal of caution.

### 5.1 Local uniformity

One response to the worries about reference priors, in applications, has been to use a proper prior that is quite diffuse. Box and Tiao (1973, p. 23) call such a prior *locally uniform*, meaning that its density is slowly varying over the region in which the likelihood function is concentrated. One might, for instance, truncate an improper reference prior so that its domain is compact and it becomes proper. An alternative is to use a probability distribution,

such as a Normal, that has a very large spread.

As a practical device, this approach will work fine in many problems. It does not, however, have any fundamental ability to avoid the difficulties that arise in using reference priors. To specify the meaning “quite diffuse” one must, for instance, determine the size of the compact set defining the domain in the truncation case or pick the spread when using a distribution such as a Normal. It is certainly possible to make a choice so that the resulting proper prior  $\pi^*(\theta)$  succeeds in approximating the “uniformity” of a reference improper prior  $\pi(\theta)$  (e.g., when  $\theta$  is one-dimensional, taking the Normal standard deviation to be  $10^{10}$  times the largest imaginable value of  $\theta$ ); but then the posterior based on  $\pi^*(\theta)$  will also approximate the formal posterior that would be obtained from  $\pi(\theta)$ . While it is true, mathematically, that the posterior based on  $\pi^*(\theta)$  will be proper, computationally the two posteriors will behave in much the same way and, thus, any serious analytical difficulties present with the original posterior will remain with its modification. As we said, we do believe it is often possible to choose the spread to be suitably large and still obtain reasonable results. Our point is that the method is not necessarily easy or automatic: when difficulties with reference priors arise in a problem, it should serve as a warning about *the problem* that care will be needed with proper priors as well. We have found this an important practical matter, and thus do not accept facile arguments implying that difficulties may be safely ignored by using proper priors.

## 5.2 Reference priors with large samples

A more positive side to the point of view articulated by Box and Tiao (1973) appears when we consider the “data-dominated” cases, in which they assumed a reference prior would be likely to succeed. These could also be called large-sample cases, since they involve situations in which the posterior is dominated by a peaked likelihood function. Jeffreys, too, focused on these cases (e.g., in Jeffreys, 1963, and also in his 1961 book, on page 212, where he finds approximate posterior for the median of a distribution). Here, the difficulties associated with reference priors will be greatly diminished and results using any of the various possible choices for them will not be much different.

Let us carry this observation a step further by considering the case in which a reference prior leads to an improper posterior yet it is not hard to find a suitable proper prior that leads to sensible results. An example occurs in the one-dimensional Normal hierarchical model, with the prior on the second-stage parameters  $\pi(\mu, \sigma) = \sigma^{-1}$ . This is not the prior determined by Jeffreys’s general rule but it illustrates the point we wish to make. This prior leads to an improper posterior (e.g., Berger 1985 p. 187) yet, in practice, with reasonably large sample sizes and a non-negligible second-stage variance, a data-analyst who uses it together with asymptotic approximation and, perhaps, other numerical methods, will rarely run into trouble. The reason is that the likelihood function will tend to have a peak away from the boundary  $\sigma = 0$ , so that if one ignores the region near the boundary the posterior is integrable and well-behaved. This amounts to substituting for the improper prior a proper version obtained by truncation to a compact set. In principle the choice of compact set could be very influential on the results, but often, in practice, the likelihood peak is sufficiently far from the boundary that there is much leeway in the choice; the impropriety of the posterior in such cases becomes a mere technicality that may be ignored.

The situation just described is what Box and Tiao called “data-dominated”. The difficulty with the argument that one may always substitute a suitable proper prior for an improper one is simply that it may not be obvious whether or not a particular posterior is data-dominated. To summarize, we see a dichotomy between large-sample and small-sample problems. Again, by “large-sample” situations we mean those in which the posterior is dominated by a single peak. We would confine the discussion of “default” methods to problems of the former kind while considering the latter to require much more serious attention, beyond what reference analysis can yield.

With this large-sample motivation in mind, we note that several of the methods we discussed specifically rely on asymptotic theory. For example, Jeffreys’s general rule and its geometrical interpretation, the Berger-Bernardo rule, coverage matching methods, and methods based on data-translated likelihoods are all built from asymptotic arguments. Importantly, these all lead to Jeffreys’s general rule or some modification of it. Thus, we believe Jeffreys’s general rule, together with its variants (such as the Berger-Bernardo rule for parameter subsets), remains an acceptable standard or, to repeat a phrase used previously, it

is “the default among the defaults.”

### 5.3 Open problems

If we regard Jeffreys’s general rule as a reasonable standard, two problems present themselves: (i) computation of it and (ii) verification that it leads to a proper posterior. For some models, such as the Normal families mentioned in the Introduction, it is not difficult to compute the prior of Jeffreys’s general rule. But for others, such as in many non-Normal hierarchical models, it may not be clear how the prior may be efficiently computed.

Although we pointed out, above, that results based on improper posteriors are sometimes quite sensible they will remain worrisome unless the data analyst has good reason to think the posterior is data-dominated (and away from troublesome boundaries). Thus, it would be very helpful to know whether Jeffreys’s general rule, and related methods, lead to proper posteriors for particular models. Some work along these lines was cited in Section 4.2.5 but more general results are needed.

Finally, we come to the biggest issue: How is one to know whether a particular posterior is data-dominated and, thus, whether a reference analysis is acceptable? If this could somehow be determined by following a reasonably straightforward procedure, Bayesian statistical practice would advance substantially.

One simple idea is to use two alternative reference methods and check the results for agreement, but this is at best rather indirect and, furthermore, may be more informative about the two alternative priors than about the data. A useful partial answer ought to involve asymptotics, since we would be trying to determine whether the sample size is sufficiently large, and for this, one might check whether the posterior is approximately Normal as suggested by Kass and Slate (1992, 1994). Once again, however, the latter approach fails to assess directly how much the posterior would change if an appropriate informative prior were to replace the reference prior. The negative results of Wasserman (1993) also indicate the difficulty of this problem. Ultimately, there seems to be no way around the exercise of some subjective judgment: the only completely reliable way to assess the effect of using an appropriate informative prior is to do so. Nonetheless, we believe this aspect of judgment

may be improved by statistical research and experience as are the many other data-analytic judgments statistical scientists must make.

We hope that our classification, summary, and discussion will help others understand better this diverse literature, and that the outstanding problems we have noted will receive further examination.

## REFERENCES

- Ash, Robert B. (1965). *Information Theory*. Dover Publications: New York.
- Beale, E.M.L. (1960). Confidence regions in non-linear estimation (with discussion). *J. Roy. Statist. Soc. B.*, **22**, 41-88.
- Berger, J. (1992). Comment on Ghosh and Mukerjee. In *Bayesian Statistics 4: Proceedings of the Fourth International Meeting*. p 205-206. Clarendon Press: Oxford.
- Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics*. (J. Kadane ed.): North-Holland, Amsterdam.
- Berger, James, O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York.
- Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference* **25** 303-328.
- Berger, J.O. and Strawderman, W. (1993). Manuscript in preparation.
- Berger, James O. and Wolpert, Robert, L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Volume 6.
- Bondar, J.V. (1977). A conditional confidence principle. *Ann. Statist.* **5**, 881-891.
- Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* **26**, 211-252.
- Buehler, R.J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30**, 845-863.
- Buehler, R.J. and A.P. Feddersen (1963). Note on a conditional property of Student's t. *Ann. Math. Stat.* **34**, 1098-1100.
- Clarke, Bertrand S. and Barron, Andrew R. (1990b). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*. **36**, 453-471.
- Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B.*, **49**, 1-18.
- de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. Translated and reprinted in *Studies in Subjective Probability*, H. Kyburg and H. Smokler (EDs.), 1964, 93-158. Wiley: New York.

- de Finetti, B. (1972). *Probability, Induction, and Statistics*. Wiley: New York.
- de Finetti, B. (1974, 1975). *Theory of Probability, Vols. 1 and 2*. Wiley: New York.
- DeRobertis, Lorraine and Hartigan, J.A. (1981). Bayesian inference with intervals of measures. *Ann. Statist.* **9**, 235-244.
- Dickey, J. M.(1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680-689.
- Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Rev.* **70**, 193-242.
- Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London (Ser. A)* **222**, 309-368.
- Fraser, D.A.S. (1968). *The Structure of Inference*. Krieger Publishing Company, Huntington. N.Y.
- Freedman, D. and Purves, R. (1969). Bayes' method for bookies. *Ann. Math. Statist.* **40**, 1177-1186.
- Friedman, K. and Shimony, A. (1971). Jayne's maximum entropy prescription and probability theory. *J. Statist. Physics*, **3**, 381-384.
- Good, I.J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Royal Statist. Soc., B*, **29**, 399-431.
- Hacking, I. (1976). *Logic of Statistical Inference*, paperback ed., original published in 1965. Cambridge University Press.
- Hartigan, J.A. (1983). *Bayes Theory*. Springer-Verlag. New York.
- Hill, B.M. (1980). On some statistical paradoxes and nonconglomerability. In *Bayesian Statistics* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.). University Press, Valencia. Am. Stat.
- Hill, B.M. and Lane, D. (1985). Conglomerability and countable additivity. *Sankhyā Ser. A* **47** 366-379.
- Hill, B.M. and Lane, D. (1986). Conglomerability and countable additivity. In *Bayesian Inference and Decision Techniques*, (P. Goel and A. Zellner eds). Elsevier Science Publishers.
- Howson, Colin and Urbach, Peter. (1989). *Scientific Reasoning: The Bayesian Approach*.

Open Court: La Salle, Illinois.

Ibragimov, I.A. and H'asminsky, R.Z. (1973). On the information contained in a sample about a parameter. *2nd Int. Symp. on Info. Theory.* 295-309.

Jaynes, E.T. (1957). Information theory and statistical mechanics. I, II. *Physical Review*, **106**, 620-630; **108**, 171-190.

Jeffreys, H. (1955). The present position in probability theory. *Brit. J. Philos. Sci.* **5**, 275-289.

Jeffreys, H. (1957). *Scientific Inference* (second edition,) (1st ed. 1931). Cambridge University Press, Cambridge.

Jeffreys, H. (1963). Review of *The Foundations of Statistical Inference* by L.J. Savage and others. *Technometrics* **5**, 407-410.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. and Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75**, 845-854.

Kass, R.E. (1981). The geometry of asymptotic inference. Technical report, Department of Statistics, Carnegie Mellon University.

Kass, R.E. (1982). A comment on "Is Jeffreys a 'necessarist'?" *Amer. Statist.* **36**, 390-391.

Kass, R.E. and Raftery, A.E. (1992). Bayes factors and model uncertainty. Technical report 571, Department of Statistics, Carnegie Mellon University.

Kass, R.E. and Slate, E.H. (1992). Reparameterization and diagnostics of posterior non-Normality (with discussion). In *Bayesian Statistics 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith ed.). 289-306. Clarendon Press, Oxford.

Kass, R.E. and Slate, E.H. (1994). Some diagnostics of maximum likelihood and posterior Normality. *Ann. Statist.*, to appear.

Kass, R.E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. Royal Statist. Soc., B*, **54**, 129-144.

Kass, R.E. and Wasserman, L. (1992). The surprising accuracy of the Schwarz criterion as an approximation to the log Bayes factor. Technical report 567, Department of Statistics, Carnegie Mellon University.

- Keynes, J.M. (1921). *A Treatise on Probability*. London: Macmillan.
- Kries, J. von (1886). *Die Principien der Wahrscheinlichkeitsrechnung. Eine Logische Untersuchung*. Freiburg.
- Lindley, Dennis V. (1990). The 1988 Wald Memorial Lectures: The present position in Bayesian statistics (with discussion). *Stat. Sci.* **5** 44-9.
- Lindley, D.V., Tversky, A., and Brown, R.V. (1979). On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. Ser. A* **142**, 146-180.
- McCullagh, P. (1992). Conditional inference and Cauchy models. *Biometrika*, **79**, 247-259.
- Nachbin, L. (1965). *The Haar Integral*. van Nostrand: New York.
- Peisakoff, M.P. (1950). Transformation parameters. Ph.D. thesis. Princeton University.
- Pierce, Donald A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1**, 241-250.
- Polson, N. G. (1988). Bayesian perspectives on statistical modeling. Ph.D. Dissertation, Department of Mathematics, University of Nottingham.
- Robinson, G.K. (1978). On the necessity of Bayesian inference and the construction of measures of nearness to Bayesian form. *Biometrika*, **65**, 49-52.
- Robinson, G.K. (1979a). Conditional properties of statistical procedures. *Ann. Statist.* **7**, 742-755.
- Robinson, G.K. (1979b). Conditional properties of statistical procedures for location and scale parameters. *Ann. Statist.* **7**, 756-771.
- Rosenkrantz, R.D. (1977). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. Reidel, Boston.
- Savage, L.J. (1962a). Bayesian statistics, in *Recent Developments in Information and Decision Theory*, eds. R.F. Machol and P. Gray, New York: Macmillan.
- Savage, L.J. et al. (1962b). *The foundations of Statistical Inference*. London: Methuen.
- Savage, Leonard J. (1972). *The foundations of Statistics*. (2nd ed.), (1st ed. 1954). Dover Publications: New York.

Schervish, M.J., Seidenfeld, T. and Kadane, J. (1984). The extent of non-conglomerability of finitely additive probabilities. *Z. Wahr. v. Gebiete* **66** 205-226.

Seidenfeld, T. (1981). Paradoxes of conglomerability and fiducial inference. In *Proceedings of the 6th International Congress on Logic Methodology and Philosophy of Science*. (J. Los and H. Pfeiffer, eds.) North Holland: Amsterdam.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press: Princeton.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Systems Tech. J.* **27**, 379-423, 623-656.

Shimony, A. (1973). Comment on the interpretation of inductive probabilities. *Journal of Statistical Physics*. **9**, 187-191.

Skala, H.J. (1988). On  $\sigma$ -additive priors,  $\sigma$ -coherence, and the existence of posteriors. In *Risk, Decision and Rationality*. (B.R. Munier, ed.) 563-574. Reidel, Dordrecht.

Sklar, Lawrence (1976). *Space, Time, and Spacetime*. University of California Press: Berkeley.

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. *Bernoulli-Bayes-Laplace Anniversary Volume: Proceedings of an International Research Seminar Statistical Laboratory*. 217-240. Springer-Verlag, New York.

Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press: Cambridge, Massachusetts.

Wallace, David L. (1959). Conditional confidence level properties. *Ann. Math. Statist.* **30**, 864-876.

Walley, P. (1991). *Statistical Reasoning With Imprecise Probabilities*. Chapman and Hall: London.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference (with discussion). In *Bayesian Statistics 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith ed.). 583-502. Clarendon Press, Oxford.

Wiener, N. (1948). *Cybernetics*. Wiley: New York.

Zabell, S.L. (1992). R.A. Fisher and the fiducial argument. *Stat. Sci.* **7**, 369-387.

Zellner, A. (1982). Is Jeffreys a "necessarist"? *American Statistician*. **36**, 28-30.

Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. In *Maximum Entropy and Bayesian Methods*. 17-31. W.T. Grandy, Jr. L.H. Schick (eds). Kluwer Academic Publishers. The Netherlands.

## ANNOTATED BIBLIOGRAPHY

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 53-59.

Defines a prior to be impartial if it is uniform in a homogeneous parameterization. In general, a globally homogeneous parameterization cannot be found. A locally homogeneous parameterization can be found and this leads to (1). This is close to Jeffreys's original argument.

Akaike, Hirotugu (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J.R. Statist. Soc. B* **42**, 46-52.

Suggests that improper priors be regarded as limits of data dependent proper priors. The author considers an example of a strong inconsistency (section 4.2.1) and an example of a marginalization paradox (section 4.2.4) and, in each case, argues that the paradoxes are best resolved by using a sequence of proper priors that depends on the data.

Bartholomew, D.J. (1965). A comparison of some Bayesian and frequentist inferences. *Biometrika*, **52**, 19-35.

Investigates the discrepancy between Bayesian posterior probability and frequentist coverage. It is noted that, among other things, better agreement can sometimes be reached in sequential experiments.

Bayes, T.R. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53**, 370-418. Reprinted in *Biometrika* **45** (1958), 243-315.

The paper where a uniform prior for the binomial problem was first used. There has been some debate over exactly what Bayes had in mind when he used a flat prior; see Stigler (1982). Other interesting information about Bayes is contained in Stigler (1986).

Berger, J.O. (1992). *Objective Bayesian Analysis: Development of Reference Noninformative Priors*. Unpublished lecture notes.

A lucid and informative review of reference priors with emphasis on the methods developed by Berger and Bernardo.

Berger, J. and Bernardo, J. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200-207.

The method of reference priors is applied to the problem of estimating the product of means of two normal distributions. This is one of the first examples to show that Bernardo's (1979a) method cannot be applied as originally presented because of technical problems relating to the nonintegrability of the reference prior conditional on the parameter of interest. It also shows that the method depends on how improper priors are approximated by proper priors.

Berger, J. and Bernardo, J. (1991). Reference priors in a variance components problem. In *Bayesian Inference in Statistics and Econometrics*. (P. Goel and N.S. Iyengar, eds.) Springer-Verlag, NY.

The Berger-Bernardo method is applied to balanced variance components problems. Various priors are derived depending on how the parameters are grouped.

Berger, J. and Bernardo, J. (1992a). Ordered group reference priors with application to the multinomial problem. *Biometrika*, **25**, 25-37.

The Berger-Bernardo stepwise method can produce different priors, depending on how the parameters are grouped. This issue is discussed and illustrated with the multinomial problem.

Berger, J. and Bernardo, J. (1992b). On the development of the reference prior method. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith eds.) 35-60. Clarendon Press: Oxford.

Synthesizes much recent work by these two authors on the stepwise approach to constructing priors. Attention is given to several practical matters including the choice of partitioning the parameter and the use of sequences of compact sets that are used to deal with impropriety. Also, there is some discussion of non-regular cases.

Berger, J.O., Bernardo, J.M. and Mendoza, M. (1989). On priors that maximize expected information. In *Recent Developments in Statistics and Their Applications*. J. Klein and J. Lee, editors. Freedom Academy: Seoul.

Some technical matters related to the Berger-Bernardo approach are dealt with. These include the existence of maximizing measures, discreteness of solutions for finite experiments, and questions about limits, both in terms of sample size and in terms of sequences of compact subsets of the parameter space.

Berger, James O. and Ruo-yong Yang. (1992). Noninformative priors and Bayesian testing for the AR(1) model. Technical report 92-45C, Department of Statistics, Purdue University.

Considers several possible reference priors for the AR(1) process:  $X_t = \rho X_{t-1} + \epsilon_t$  where  $\epsilon_t \sim N(0, 1)$ . The priors considered are the flat prior, Jeffreys's prior and two version of the Berger-Bernardo prior. An alternative prior, called the symmetrized reference prior is also considered. This is defined by

$$\pi(\rho) = \begin{cases} \{2\pi\sqrt{1-\rho^2}\}^{-1} & \text{if } |\rho| < 1 \\ \{2\pi|\rho|\sqrt{1-\rho^2}\}^{-1} & \text{if } |\rho| > 1. \end{cases}$$

The priors are compared in simulation studies for the coverage properties and mean-squared errors of the Bayes estimators. The symmetrized reference prior performed better in mean-squared error and reasonably well in terms of coverage and the authors recommend this prior as the reference prior. The authors also consider the problem of testing for a unit root.

Berger, James O. and Ruo-yong Yang. (1993). Estimation of a covariance matrix using the reference prior. Technical report 93-13C, Department of Statistics, Purdue University.

The problem is to estimate the covariance matrix  $\Sigma$  in a  $N(0, \Sigma)$  model. The authors argue that Jeffreys's prior does not "appropriately shrink the eigenvalues." They decompose  $\Sigma$  as  $\Sigma = O'DO$  where  $O$  is an orthogonal matrix and  $D$  is diagonal with decreasing elements. Then the method of Berger and Bernardo (1992b) is applied treating the parameters as being ordered in importance, with the elements of  $D$  being the most important. The authors discuss methods for computing the posterior and they evaluate the accuracy of the Bayes estimator by simulation.

Bernardo, J.M. (1979a). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.* **41**, 113-147.

Two ideas emerge in this paper. The first is to define reference priors as priors that maximize the asymptotic missing information relative to a given experiment. In continuous parameter spaces this is typically the prior given by (1). In finite spaces it is the uniform prior. The second idea is to decompose parameters into a parameter of interest and a nuisance parameter. Then a stepwise argument is applied, by finding the reference prior for the nuisance parameter conditional on the parameter of interest, finding the marginal model for the parameter of interest and then finding the marginal reference prior for the parameter of interest. Many anomalies of noninformative priors are apparently solved this way. See section 3.5.

Bernardo, J.M. (1979b). Expected information as expected utility. *Ann. Statist.* **7** 686-690.

This paper views the task of reporting a posterior distribution as a decision problem. Suppose that  $u(p^*, \theta)$  is the utility of reporting a distribution  $p^*$  when  $\theta$  is the true value of the parameter. If  $x$  is observed, the expected utility is  $\int u(p^*, \theta)p(\theta|x)d\theta$ . If this is maximized by reporting one's true posterior, then  $u$  is said to be proper. The function  $u$  is local if  $u(p^*, \theta) = u(p^*(\theta), \theta)$  for all  $\theta$ . This means that the utility only depends on the value of the density at the true value. Bernardo shows that if  $u$  is smooth, proper and local, then  $u(p^*, \theta) = A \log p^* + B(\theta)$  for some constant  $A$  and some function  $B$ .

Bernardo, J.M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds., 605-647.

Applies the method in Bernardo (1979a) to the problem of hypothesis testing. First, the Berger-Bernardo prior for the prior probability of the null is obtained for a fixed prior on the parameter, conditional on the alternative. Then, he suggests using Jeffreys's rule (possibly on a truncated space) for the parameter under the alternative. In the case where the data are Normal with known variance

$\sigma^2$ , and we are testing  $\mu = \mu_0$ , the posterior odds using this method turn out to be

$$\frac{\pi(H_1|\text{Data})}{\pi(H_0|\text{Data})} = \exp\{(1/2)(\gamma_x^2 - 1)\}$$

where  $\gamma_x = \sqrt{n}(\bar{x} - \mu_0)/\sigma$ .

Berti, Patrizia, Regazzini, Eugenio and Rigo, Pietro (1991). Coherent statistical inference and Bayes theorem. *Ann. Statist.*, **19**, 366-381.

When dealing with finitely additive probabilities, the formal application Bayes theorem need not generate a coherent posterior. Similarly, a coherent posterior need not be generated by Bayes theorem. This paper investigates conditions for which posteriors from Bayes theorem are coherent.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts.

They argue that it is not possible to model complete ignorance, but that it is possible to model ignorance relative to a given experiment (page 25). They suggest using a flat prior in a parameterization that makes the likelihood depend on the data only through its location. Jeffreys's prior accomplishes this, at least approximately. Kass (1990) shows that these ideas can be made more precise and can be extended. See section 3.3.

Brunk, H.D. (1991). Fully coherent inference. *Ann. Statist.* **19**, 830-849.

Investigates coherence in the spirit of Dawid and Stone (1973, 1973), Heath and Sudderth (1978) and Lane and Sudderth (1983). Notes that coherent inferences may have some unpleasant properties; for example, the posterior might put mass in places where the prior does not. The author introduces a notion of compatibility between the prior and the posterior to rule out such behavior.

Chang, T. and Eaves, D. (1990). Reference priors for the orbit in a group model. *Ann. Statist.* **18** 1595-1614.

Suppose that the parameter  $\omega = (\theta, g)$ , where  $g \in G$ ,  $G$  is a group, and  $\theta$ , the parameter of interest, indexes the orbits of the group. The authors propose the prior  $p(\theta)p(g|\theta)$  where  $p(g|\theta)$  is right Haar measure,

$$p(\theta) = \lim_{n \rightarrow \infty} \sqrt{\det(I_n(\theta))/n}$$

and  $I_n(\theta)$  is the information matrix for  $y_n$ , the maximal invariant of the  $G$ -action. They show that this is a reference prior in the sense of Bernardo (1979). Further, they show that the decomposition  $\omega = (\theta, g)$  need not be found explicitly to find the prior. Examples based on the multivariate Normal are given. See section 3.2.

Chang, T. and Villegas, C. (1986). On a theorem of Stein relating Bayesian and classical inferences in group models. *Canad. J. Statist.* **14** 289-296.

Gives a new proof of Stein's (1965) theorem that equivariant posterior regions correspond to confidence intervals in group models when right Haar measure is used as a prior. The proof avoids the need for an equivariant factorization of the sample space. Some applications to the multivariate normal are considered.

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica*, **22** 422-443.

Derives the Principle of Insufficient Reason for finite spaces based on eight postulates of rational decision making. He avoids partitioning paradoxes by restricting the theory to sets with a given number of outcomes.

Cifarelli, D.M. and Regazzini, E. (1987). Priors for exponential families which maximize the association between past and future observations. *Probability and Bayesian Statistics*. (Viertl, R. ed.) Plenum Press: New York. 83-95.

Using finitely additive priors for an exponential model, the authors show that a large class of priors give perfect association between future and past observations in the sense that there are functions  $\phi_n : \mathbb{R}^n \rightarrow \mathbb{R}$  such that

$$P(X_N \leq x, \phi_n(X_1, \dots, X_n) \leq x) = P(X_N \leq x) = P(\phi_n(X_1, \dots, X_n) \leq x)$$

for all  $N > n$ ,  $n = 1, 2, \dots$  and  $x \in \mathbb{R}$ . Under certain conditions, they show that the only prior that gives  $E(X_N | X_1, \dots, X_n) = \bar{X}_n$  is the usual improper uniform prior.

Cifarelli, Donato Michele and Regazzini, Eugenio (1983). Qualche osservazione sull'uso di distribuzioni iniziali coniugate alla famiglia esponenziale. *Statistica* 43:415.

Shows that the conjugate priors for the exponential family are the unique priors that maximize the correlation between  $X_N$  and  $\bar{X}_n$  subject to fixed values of  $Var(E(X_n | \theta)) / Var(X_n)$ .

Clarke, B. and Barron, A. (1990a). Bayes and minimax asymptotics of entropy risk. Technical report, Purdue.

Shows that Jeffreys's prior is the unique, continuous prior that achieves the asymptotic minimax risk when the loss function is the Kullback-Leibler distance between the true density and the predictive density. See also Good (1969) and Kashyap (1971).

Clarke, B. and Wasserman, L. (1993). Noninformative priors and nuisance parameters. *J. Amer. Statist. Soc.*, to appear.

A functional that measures missing information for a parameter of interest minus a penalty term that measures the implied information for other parameters is defined. The prior that maximizes the functional is called a trade-off prior.

Clarke, B. and Wasserman, L. (1992). Information tradeoff. Technical report 558, Department of Statistics, Carnegie Mellon University.

Further investigates the trade-off priors in Clarke and Wasserman (1993). A closed-form expression for the trade-off prior is obtained and the relationship with the Berger-Bernardo prior is derived.

Clarke, B. and Dong Chu Sun. (1992). References priors under the Chi-squared distance. Technical report, Department of Statistics, Purdue University.

Noting that Jeffreys's prior can be obtained by maximizing expected Kullback-Leibler distance between prior and posterior, the authors consider instead maximizing expected Chi-squared distance. Within a certain class of priors, the maximizing prior turns out to be proportional to the inverse of Jeffreys's prior squared.

Consonni, G. and Veronese, P. (1987). Coherent distributions and Lindley's paradox. In *Probability and Bayesian Statistics*. (R. Viertl, ed.) 111-120. Plenum, New York.

The Jeffreys's-Lindley paradox is that there may be sharp disagreement between the classical and Bayesian tests of a sharp null hypothesis. In its extreme form, where an improper prior is used, this leads to a situation where the Bayes factor for the simpler model is infinite. The authors discuss the latter version in the context of finitely additive probability theory. In particular, by assigning mass adherent to the null (loosely, probability arbitrarily close to but not at the null) then the paradox is avoided.

Consonni, G. and Veronese, Piero. (1988). A note on coherent invariant distributions as non-informative priors for exponential and location-scale families. *Studi Statistici* n. 19, Università L. Bocconi, Milano.

Dawid's notion of context invariance (Dawid 1983) is used to derive non-informative priors for exponential and location-scale families. Improper priors are interpreted as finitely additive priors and are obtained by taking limits of priors on expanding sequences of compact sets. The conclusion is that these methods lead to only a class of priors. The prior from Jeffreys's general rule is not generally in this class but Hartigan's ALI prior (Hartigan 1964) appears to be in the class.

Mukhopadhyay, Saurabh and DasGupta, Anirban (1993). Uniform approximation of Bayes solutions and posteriors: frequently valid Bayes inference. Technical report 93-12C, Department of Statistics, Purdue University.

Suppose that  $f(x - \theta)$  is a location family such that the moment generating function for  $f(x)$  exists. Let  $\pi^x$  be the posterior using a flat prior. The authors show (among other things) that, for every  $\epsilon > 0$ , there exists a proper, countably additive prior  $q$  with posterior  $q^x$  such that  $d(\pi^x, q^x) < \epsilon$  for all  $x$ .

Dawid, A. P. (1983). Invariant prior distributions. *Encyclopedia of Statistical Sciences*. (Kotz, S. and Johnson, N. L. eds.), vol. 4, 228-236.

Excellent review of invariant priors. Explains the principles of parameter invariance, data invariance and context invariance.

Dawid, A.P. and Stone, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika*, **59**, 486-489.

Investigates “expectation consistency” which means, loosely, that functions with zero posterior mean for every data point should not have positive expected value with respect to every parameter value. Inferences from Bayesian posteriors are shown to be expectation consistent. If the model gives positive probability to all data points, then an expectation-consistent inference is a posterior with respect to some prior.

Dawid, A.P. and Stone, M. (1973). Expectation consistency and generalized Bayes inference. *Ann. Statist.* **1**, 478-485.

Extends work in Dawid and Stone (1972). The assumption that the model gives positive probability to all data points is dropped. Priors that produce a given expectation consistent posterior are characterized.

Dawid, A.P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. B* **35**, 189-233.

This paper discusses a paradox that can arise when using improper priors. Essentially, the problem is that the marginal of the posterior may depend on only a function of the data. Then, it is found that the distribution of this function cannot be combined with any prior to reproduce the marginal posterior. The paper is now a classic in this area. It is filled with examples, has a detailed analysis of the group theoretic case and also considers Fraser’s theory of structural inference. There is a long and interesting discussion. See section 4.2.4.

Dey, Dipak K., Gelfand, Alan E. and Peng, Fengchun. (1993). Overdispersed generalized linear models. Technical report, Department of Statistics, University of Connecticut.

Gives conditions for the propriety of the posterior in some overdispersed generalized linear models. See also Ibrahim and Laud (1991).

DiCiccio, Thomas J. and Martin, Michael, M. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. B*. **55**, 305-316.

Shows that the approximate  $1 - \alpha$  confidence limit obtained by using the approach of Welch and Peers (1963) and Peers (1965) differs by order  $O(n^{-3/2})$  from a conditional confidence limit using the signed square root likelihood ratio statistics.

DiCiccio, Thomas J. and Stern, Steven E. (1992). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihood. Technical report 404, Department of Statistics, Stanford University.

Characterizes priors for which highest posterior density regions and likelihood regions with content  $1 - \alpha$  have coverage  $1 - \alpha + O(n^{-2})$ . This generalizes results in Ghosh and Mukerjee (1992b) and Severini (1991).

Eaton, M. (1992). A Statistical Diptych: Admissible inferences – Recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147-1179.

Finds a sufficient condition so that the formal Bayes rules for all quadratically regular decision problems are admissible. The condition is related to the recurrence of a Markov chain on the parameter space generated by the model and the prior.

Eaton, Morris L. and William D. Sudderth (1993a). The formal posterior of a standard flat prior in MANOVA is incoherent. Unpublished manuscript, Department of Statistics, University of Minnesota.

The authors show that the right Haar prior in a MANOVA model produces an incoherent posterior in the sense that it is possible to devise a finite system of bets that is guaranteed to have expected payoff greater than a positive constant. Coherence is discussed in Heath and Sudderth (1978, 1989) and Lane and Sudderth (1983).

Eaton, Morris L. and William D. Sudderth (1993b). Prediction in a multivariate normal setting: coherence and incoherence. Unpublished manuscript, Department of Statistics, University of Minnesota.

Shows that a common, invariant prior for the multivariate normal leads to incoherent predictions.

Eaves, D.M. (1983a). On Bayesian non-linear regression with an enzyme example. *Biometrika* **70**, 373-379.

Notes the form of Jeffreys's rule in this setting and points out that it can be derived by the method of Bernardo (1979). This prior was also mentioned by Beale (1960).

Eaves, D.M. (1983b). Minimally informative prior analysis of a non-linear model. *The Statist.*, **32**: 117.

This one-page article describes work applying the scheme of Bernardo (1979a) to partially nonlinear models. See Eaves (1983a).

Eaves, D.M. (1985). On maximizing missing information about a hypothesis. *J. Roy. Stat. Soc. Ser. B*, **47**, 263-266.

Suppose that  $X$  is  $N(\mu, \sigma^2)$ ,  $\sigma^2$  known, and consider the hypothesis  $H_0 : \mu = \mu_0$ . Let the prior for  $\mu$  conditional on  $H_0^c$  be  $N(\mu_0, \tau^2)$  and let the prior for odds for  $H_0$  be  $\omega$ . This paper shows that, for fixed  $\tau^2$ , the expected information for  $H_0$  is maximized by some  $\omega > 1$ . Over all  $\tau^2$  the maximum occurs at  $\omega = 1$  and  $\tau^2 = \infty$  leading to the Jeffreys-Lindley paradox ( $H_0$  is accepted always). Maximizing the expected information for  $H_0$  plus the information for  $\mu$  leads to  $\omega = 0$  and  $\tau^2 = \infty$ . Other approaches to choosing priors for testing are given by Bernardo (1980) and Pericchi (1984).

Efron, B. (1973). Discussion of Dawid, Stone and Zidek (1973). *J. R. Statist. Soc. B* **35**, 219.

Discusses “noninformative” priors in multiparameter situations. In particular, let  $\theta_1, \dots, \theta_{100}$  be parameters and  $x_1, x_2, \dots, x_{100}$  data, where  $x_i \sim N(\theta_i, 1)$  independently given the  $\{\theta_i\}$ . Considers the “noninformative” prior  $\theta_i \sim N(0, A)$ ,  $A$  large and shows that, if the parameter of interest is  $\xi = \sum_1^{100} \theta_i^2$ , this prior can overwhelm the data. This can be overcome by assuming  $A$  itself has a diffuse prior, say proportional to  $(A + 1)^{-2}$ . But then if the parameter of interest is  $\xi = \max \theta_i$ , this appears informative.

Efron, B. (1986). Why isn't everyone a Bayesian? (with discussion). *Am. Statist.* **40**, 1-11.

Suggests several reasons why the Bayesian paradigm has not been widely accepted among practicing statisticians, including the difficulty in defining “objective” Bayesian inference. Some of the discussion takes up this point as well.

Gatsonis, C.A. (1984). Deriving posterior distributions for a location parameter: a decision-theoretic approach. *Ann. Statist.* **12**, 958-970.

Shows that the best invariant estimator of the posterior distribution for a location parameter using  $L_2$  distance as a loss function is the posterior arising from a uniform prior. Also, shows that this estimator is inadmissible for dimension greater than 3 and suggests alternative estimators.

Geisser, S. (1984). On prior distributions for binary trials. *Am. Statist.* **38**, 244-251.

In estimating the success probability  $\theta$  from a Binomial or negative Binomial sample, it is argued that the interval  $(0, 1)$  of possible values of  $\theta$  is a convenient representation of the finitely many values of  $\theta$  that are actually possible [e.g., according to machine precision in a computer]. When there are finitely many values, a uniform prior is generally taken to be appropriate (according to the Principle of Insufficient Reason, see Section 3.1). Thus, a uniform prior on  $\theta$  should be used for Binomial or negative Binomial sampling. Predictive distribution calculations are given as a way of formalizing this argument. See also Stigler (1982).

Geisser, S. and Cornfield, J. (1963). Posterior distributions for the multivariate normal distribution. *J. R. Statist. Soc. B*, **25** 368-376.

Contrasts posterior distributions with Fiducial and confidence. The motivation is the discrepancy between joint confidence regions for a multivariate normal based on Hotelling's  $T^2$  and regions based on a fiducial distribution. A class of priors indexed by a parameter  $\nu$  is proposed. The fiducial answer corresponds to  $\nu = 2$  and Hotelling's answer corresponds to  $\nu = p + 1$  where  $p$  is the dimension of the problem. Further, there is no value of  $\nu$  that gives the usual Student intervals for a single mean and Hotelling's regions for the joint problem. See Stone (1964) for a criticism of this prior, namely, the prior is not a probability limit of proper priors.

George, E.I. and McCulloch, R. (1989). On obtaining invariant prior distributions. Technical Report # 73. Graduate School of Business, University of Chicago.

Motivated by Jeffreys, the authors define a prior in terms of a discrepancy measure  $\psi(\cdot, \cdot)$  on a family of distributions. The prior is defined by

$$\pi(\theta) \propto \det(\nabla \nabla \psi(\theta, \theta))^{1/2}.$$

Variance discrepancies are considered. The priors are parameterization invariant. Requiring sample space invariance as well leads to (1). Left invariant discrepancies produce left invariant Haar measure. Similar invariance arguments are also considered in Hartigan (1964), Kass (1989), and Good (1969).

Ghosh J.K. and Mukerjee, R. (1992a). Non-informative priors. In *Bayesian Statistics 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith ed.). 195-210. Clarendon Press, Oxford.

Examines the Berger-Bernardo prior and suggests using the marginal missing information; see also Clarke and Wasserman (1992, 1993) for this approach. Then, priors that match posterior probability and frequentist coverage are considered. For this, Bartlett corrections to the posterior distribution of the likelihood ratio are used. Finally, some results on finding least favorable priors are given.

Ghosh, J.K. and Mukerjee, Rahul. (1992b). Bayesian and frequentist Bartlett corrections for likelihood ratio and conditional likelihood ratio tests. *J. Roy. Statist. Soc. B* **54**, 867-875.

Characterizes priors for which Bayesian and frequentist Bartlett corrections for the likelihood ratio statistic differ by  $o(1)$ . Posterior regions based on the Bartlett corrected likelihood ratio statistic have the same frequentist nominal coverage to order  $o(n^{-1})$ . See section 3.7.

Ghosh, J.K. and Mukerjee, Rahul. (1993). On priors that match posterior and frequentist distribution functions. *Canad. J. Stat.* **21**, 89-96.

Characterizes priors, by way of a differential equation, that make  $P(W \leq t|X) = P(W \leq t|\theta) + o(n^{-1/2})$  for all  $\theta$  and all  $t$ , in multiparameter models. Here,  $W = n^{1/2}C^*(\theta - \hat{\theta})$  where the matrix  $C^*$  is chosen in a certain way so that  $W$  reflects an ordering of the parameters in terms of importance. Thus,  $W_1$  is a scaled version of  $n^{1/2}(\theta_1 - \hat{\theta}_1)$ ,  $W_2$  is a standardized regression residual of  $n^{1/2}(\theta_2 - \hat{\theta}_2)$  on  $n^{1/2}(\theta_1 - \hat{\theta}_1)$  and so on.

Good, I.J. (1969). What is the use of a distribution? *Multivariate Analysis* (Krishnaiah, ed.), **II**, 183-203. New York: Academic Press.

This paper defines  $U(G|F)$  to be “the utility of asserting that a distribution is  $G$  when, in fact, it is  $F$ ”. Various functional form for  $U$  are studied. For a particular form of  $U$ , a minimax argument establishes (1) as the least favourable distribution.

Haldane, J.B.S. (1948). The precision of observed values of small frequencies. *Biometrika* **35**, 297-303.

Suggests the prior  $p^{-1}(1-p)^{-1}$  for a binomial parameter  $p$  when the event is expected to be rare.

Hartigan, J.A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836-845.

Defines a prior  $h$  to be relatively invariant if  $h(z\theta)(dz\theta/d\theta) = ch(\theta)$  for some  $c$ , whenever  $z$  is a 1-1 differentiable transformation satisfying  $f(zx|z\theta)(dzx/dx) = f(x|\theta)$  for all  $x$  and  $\theta$ . An asymptotic version leads to an asymptotically locally invariant (ALI) prior defined, in the one dimensional case, by

$$\left(\frac{\partial}{\partial\theta}\right)\log h(\theta) = -E(f_1f_2)/E(f_2)$$

where  $f_1 = [\partial/\partial\theta \log f(x|\theta)]_0$  and  $f_2 = [\partial^2/\partial\theta^2 \log f(x|\theta)]_0$ . Some unusual priors are obtained this way. For example, in the normal  $(\mu, \sigma^2)$  model we get  $\pi(\mu, \sigma) = \sigma^{-5}$ .

Hartigan, J.A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.* **36**, 1137-1152.

The author approaches non-informative priors with a decision theoretic motivation. A decision  $d(x)$  is unbiased for loss function  $L$  if

$$E_{\theta_0}(L(d(x), \theta)|\theta_0) \geq E_{\theta_0}(L(d(x), \theta_0)|\theta_0)$$

for all  $\theta, \theta_0$ . If the parameter space is one dimensional, then the Bayes' estimator is asymptotically unbiased if and only if the prior density  $h$  satisfies

$$h(\theta) = E(\partial/\partial\theta \log f(x|\theta))^2 / (\partial^2/\partial\phi^2 L(\theta, \phi))_{\theta=\phi}^{1/2}$$

Jeffreys's rule can be obtained by using Hellinger distance as a loss function. Extensions to higher dimensions lead to possibly intractable differential equations. The prior  $\pi(\mu, \sigma) = 1/\sigma$  for location-scale problems is apparently not obtainable from this approach.

Hartigan, J.A. (1966). Note on the confidence-prior of Welch and Peers. *J. Roy. Statist. Soc. B* **28**, 55-56.

In this note, the author shows that a two-sided Bayesian  $1 - \alpha$  credible region has confidence size  $1 - \alpha + O(n^{-1})$  for every prior. This is in contrast to the result of Welch and Peers (1963) where, for one-sided intervals, the prior from Jeffreys's rule was shown to have confidence  $1 - \alpha + O(n^{-1})$  compared to other priors that have confidence  $1 - \alpha + O(\frac{1}{\sqrt{n}})$ .

Hartigan, J.A. (1971). Similarity and probability. *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, (Eds.), Holt, Rinehart and Winston, Toronto.

Defines the similarity of events  $E$  and  $F$  by  $S(E, F) = P(E \cap F)/(P(E)P(F))$ . Shows that (1) makes present and future observations have constant similarity, asymptotically. Also, (1) maximizes (asymptotically) the similarity between the observations and the parameter.

Heath, David and Sudderth, William (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333-345.

Shows that inferences for a parameter given an observation are coherent (in a certain sense) if and only if the inferences are the posterior for some prior. The development takes place using finitely additive probabilities. The coherence condition essentially boils down to conglomerability in the parameter margin and the data margin.

Heath, D. and Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.* **17**, 907-919.

Conditions are given so that the formal posterior obtained from an improper prior are coherent in the sense of Heath and Sudderth (1978).

Hills, S. (1987). Reference priors and identifiability problems in non-linear models. *The Statistician*. **36**, 235-240.

Argues that the contours of the Jeffreys's prior give clues about regions of the parameter space that are nearly non-identifiable.

Ibrahim, Joseph G. and Laud, Purushottan W. (1991). On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior. *J. Amer. Statist. Assoc.* **86**, 981-986.

Sufficient conditions are given for the propriety of the posterior and the existence of moments for generalized linear models. In particular, they show that Jeffreys's prior leads to proper posteriors for many models.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, 227-241.

Takes the position that objective priors exist, and can often be found from the method of maximum entropy. A connection is made between maximum entropy and frequency distributions. When a parameter is continuous, a base measure is needed. The author recommends using group invariant measures for this purpose when they are available. A critique of this approach is given in Seidenfeld (1987).

Jaynes, E.T. (1980). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (Ed.), North Holland, Amsterdam.

A rebuttal to the Dawid, Stone and Zidek (1973) paper. He claims that the marginalization paradoxes are illusory and occur only because relevant information is ignored in the analysis. Specifically, the two conflicting posteriors in the marginalization paradox are based on different background information  $I_1$  and  $I_2$ , say. Jaynes' thesis is that if we are more careful about notation and write  $p(A|x, I_i)$  instead of  $p(A|x)$  the paradox disappears. Further, he proposes that priors that are immune to the illusion of marginalization paradoxes are interesting in their own right. A rejoinder by Dawid, Stone and Zidek follows.

Jaynes, E.T. (1982). On the rationale of maximum entropy methods. *Proc. of IEEE*, **70**, 939-952.

A discussion of maximum entropy methods for spectral analysis. Much attention is given to the observation that "most" sample paths give relative frequencies concentrated near the maximum entropy estimate.

Jaynes, E.T. (1983). *Papers on Probability, Statistics and Statistical Physics*. (R. Rosenkrantz ed.) Dordrecht: D. Reidel.

A collection of some of Jaynes most influential papers. Includes commentary by Jaynes.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London A* **186**, 453-461.

Proposes his prior. The material in this paper is essentially contained in Jeffreys (1961).

Jeffreys, H. (1961). *Theory of Probability*, (3rd Edition). (1st ed. 1939, 2nd ed. 1947). Oxford University Press, London.

This extremely influential text lays the foundation for much of Bayesian theory as it is practiced today. Jeffreys's rule is defined and hypothesis testing is studied in great detail. See section 2.

Kadane, Joseph B., Mark J. Schervish and Teddy Seidenfeld (1986). Statistical implications of finitely additive probability. In *Bayesian Inference and Decision Techniques*. (P. Goel and A. Zellner eds.), p 59-76, Elsevier Science Publishers.

Discusses various paradoxes that occur with finitely additive probabilities. The authors argue that these paradoxes do not undermine the utility of finitely additive probabilities. Furthermore, they critically examine the Heath, Lane, Sudderth approach to coherence (section 4.2.1) and suggest that their notion of coherence is too strong. They discuss several famous statistical paradoxes in the framework in a finitely additive framework.

Kashyap, R.L. (1971). Prior probability and uncertainty. *IEEE Trans. Information Theory* **1T-14**, 641-650.

Views the selection of a prior as a 2-person zero sum game against nature. The minimax solution, using the average divergence between the data density and the predictive density as a loss function, is that prior  $\pi(\theta)$  that minimizes  $E \log p(y | \theta) / \pi(\theta)$  where expectation is with respect to the joint measure on  $y$  and  $\theta$  (this is the Berger/ Bernardo solution). Asymptotically, he derives (1). He also considers the ergodic, but non-independent case.

Kass, R.E. (1989). The geometry of asymptotic inference. *Statistical Science*. **4** 188-234.

Discusses a geometric interpretation of Jeffreys's prior, based on its derivation as the volume element of the Riemannian metric determined by Fisher information. See section 3.6.

Kass, R.E. (1990). Data-translated likelihood and Jeffreys's rule. *Biometrika*. **77**, 107-114.

Provides an explanation and elaboration of Box and Tiao's concept of data-translated likelihood. Also shows that it may be extended by conditioning on an ancillary statistic, and then interprets the concept as essentially group-theoretic. Box and Tiao's concept of approximately data-translated likelihood is similarly discussed. (See Box and Tiao, 1973, and Section 3.3.)

Laplace, P.S. (1820). *Essai philosophique sur les probabilités*, English translation: *Philosophical Essays on Probabilities*, 1951. New York: Dover.

For extensive discussion of this and other early works involving "inverse probability" (i.e., Bayesian inference) see Stigler (1986, Chapter 3).

Lane, David A. and Sudderth, William D. (1983). Coherent and Continuous Inference. *Ann. Statist.* **11**, 114-120.

Establishes that if either the sample space or parameter space is compact, then, assuming some weak regularity conditions, an inference is coherent if and only if the posterior arises from a proper, countably additive priors.

Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. B.*, **20**, 102-107.

Shows that, for a scalar parameter and a model that admits a real-valued sufficient statistic, the fiducial based confidence intervals agree with some posterior if and only if the problem is a location family (or can be transformed into such a form).

Mitchell, Ann F.S. (1967). Comment on: "A Bayesian Significance Test for Multinomial Distributions," by I.J. Good. *J. Roy. Statist. Assoc.* **29**, 423.

Points out that for the exponential regression model  $Ey_x = \alpha + \beta\rho^x$  the uniform prior on  $\alpha$ ,  $\beta$ ,  $\log \sigma$  and  $\rho$  yields an improper posterior. Says that the non-location Jeffreys prior is unsatisfactory "on common-sense grounds," and proposes an alternative class of priors. See also Ye and Berger (1991).

Moulton, Brent R. (1993). Bayesian analysis of some generalized error distributions for the linear model. Unpublished manuscript, Division of Price and Index Number Research, Bureau of Labor Statistics.

Obtains Zellner's MDIP prior for the  $t$  family and the power exponential family.

Mukerjee, R. and Dey, D.K. (1992). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: higher order asymptotics. Technical report 92-07, Dept. of Statistics, The University of Connecticut.

Finds priors to match frequentist coverage to order  $o(n^{-1})$ . It is assumed that  $\theta = (\omega, \lambda)$  where the parameter of interest  $\omega$  and the nuisance parameter  $\lambda$  are one-dimensional.

Nicolaou, Anna. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. B.* **55**, 377-390.

Finds priors that produce Bayesian intervals that also have accurate frequentist coverage to order  $O(n^{-1})$ . The emphasis is on extending the work of Welch and Peers (1963) to the case where there are nuisance parameters. See section 3.7.

Novick, M.R. and Hall, W.J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104-1117.

Defines an indifference prior by first identifying a class a conjugate priors, and then requiring (i) that the prior be improper and (ii) that a *minimum necessary sample* induces a proper posterior. The identification of a minimum necessary sample and an initial parameterization in which to define the conjugate class, varies from problem to problem.

Novick, M.R. (1969). Multiparameter Bayesian indifference procedures. *J. R. Statist. Soc. B* **31**, 29-64 (with discussion).

Extends the procedure in Novick and Hall (1963) to multiparameter settings. Requires a consistency condition between conditionals of posteriors based on the multiparameter approach and the posterior from the single parameter approach. The prior for a bivariate normal depends on whether we cast the problem as a correlation problem or a regression problem.

Peers, H.W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B* **27**, 9-16.

Considers the problem of finding a prior that will give one sided  $\alpha$ -level posterior intervals that have frequentist coverage  $\alpha + O(1/\sqrt{n})$  in multiparameter models. This extends work of Welch and Peers (1963).

Peers, H.W. (1968). Confidence properties of Bayesian interval estimates. *J. Roy. Statist. Soc. B.*, **30**, 535-544.

Finds priors to make various two-sided intervals – equal-tailed regions, likelihood regions and HPD regions – have posterior probability content and frequentist coverage match to order  $1/n$ .

Pericchi, L.R. (1981). A Bayesian approach to transformations to normality. *Biometrika* **68**, 35-43.

Considers the problem of choosing priors for a Normal problem when Box-Cox transformations are used. The goal is to avoid the data-dependent prior that was used by Box and Cox (1964). The resulting priors lead to inferences that mimic the maximum likelihood analysis.

Pericchi, L. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika*, **71**, 575-586.

In choosing between models  $M_1, \dots, M_k$ , the author argues that the posterior tends to favor models for which the expected gain in information is low. This is an explanation for the Jeffreys-Lindley paradox. To avoid this, he suggests weighting the prior probabilities of the models appropriately.

Perks, W. (1947). Some observations on inverse probability; including a new indifference rule. *J. Inst. Actuaries* **73**, 285-334.

Suggests taking the prior to be inversely proportional to the asymptotic standard error of the estimator being used. When the estimator is sufficient, this amounts to Jeffreys's rule; Perks was not aware of Jeffreys's 1946 paper. Perks shows this rule to be invariant to differentiable transformations and treats the Binomial

case. In his motivational remarks Perks seems to be groping for the concept of an asymptotic pivotal quantity. There is extensive philosophical discussion in the paper, and in contributions from discussants. Perks notes that when there is no sufficient estimator his rule is not explicit, and that Jeffreys's paper, then in press, solved this problem.

Perlman, M.D. and Rasmussen, U.A. (1975). Some remarks on estimating a noncentrality parameter. *Commun. Statist.*, 4: 455-468.

Let  $Y \sim N_k(\mu, I)$ , a  $k$ -dimensional spherical normal vector, and let  $X = \|Y\|^2$  and  $\delta = \|\mu\|^2$ . Three estimators of  $\delta$  are compared: the UMVUE, which is  $X - k$ , the posterior mean of  $\delta$  under the flat prior on  $\mu$ , which is  $X + k$ , and the posterior mean of  $\delta$  under the conjugate normal prior  $\mu \sim N(0, \gamma \cdot I)$  for  $\gamma > 0$ . Citing Savage as the source of their argument, the authors show that under the marginal distribution of  $X$  based on the conjugate prior, the latter posterior mean is likely to be much closer to  $X - k$  than  $X + k$ , no matter how large  $\gamma$  is (with probability arbitrarily close to 1 as  $k \rightarrow \infty$ ). See section 4.2.2.

Phillips, P.C.B. (1991). To criticize the critics: an objective Bayesian analysis of stochastic trends. *J. Applied Econometrics*, **6**, 333-364.

Argues vigorously for (1) for some time series models. The problem with Jeffreys's prior for time series models is, among other things, that it depends on sample size making it unclear how to update the posterior when there is a new observation. Also, Jeffreys's prior can be obtained from minimizing asymptotic missing information where the asymptotics are done using many replications each consisting of a fixed number of observations from the process. It can be argued that one should instead consider one process and let the total time of observation tend to infinity. This leads to a very different prior; see Berger and Yang (1992). Choosing priors in these problems is a controversial issue. This paper is followed by a lengthy discussion where the virtues of various priors are debated.

Piccinato, L. (1973). Un Metodo per determinare distribuzioni iniziali relativamente non-informative. *Metron* **31**, 1-13.

Derives priors that yield, for any experimental result, posteriors concentrated on an empirical estimate of the parameter.

Piccinato, L. (1978). Predictive distributions and non-informative priors. *Trans. 7th Prague Conf. Information Theory*.

A predictive distribution is conservative if the data are a typical point with respect to the distribution. Here, a typical point means a point that minimizes some functional, such as squared error. Priors that yield conservative predictive distributions are derived.

Pinkham, R.S. (1966). On a fiducial example of C. Stein. *J. Roy. Statist. Soc. B.* **37**, 53-54.

Responds to Stein's (1959) proof that a one-sided  $\alpha$  level Bayes credible region has frequentist coverage tending to zero as  $n \rightarrow \infty$  for the problem of estimating  $\sum_i \xi_i^2$  using a flat prior when  $X_i \sim N(\xi_i, 1)$ ,  $i = 1, \dots, n$ . Stein's proof assumes that  $\xi^2 = o(n^2)$ . The author shows that if  $\xi^2 = Mn^h + o(1)$  where  $M > 0$  and  $h > 2$  then the posterior probability content and frequentist coverage agree asymptotically. See section 4.2.2.

Press, S. James (1993). The de Finetti transform. Technical report 201, Department of Statistics, University of California, Riverside.

Considers finding priors and models that produce exchangeable sequences of random variables such that the marginal distribution of the data has maximum entropy, possibly subject to moment constraints.

Regazzini, E. (1987). De Finetti's coherence and statistical inference. *Ann. Statist.* **15**, 845-864.

Investigates conditions that guarantee that a posterior be coherent in the sense of de Finetti. This notion of coherence is weaker than that developed by Heath and Sudderth (1978, 1989) and Lane and Sudderth (1983).

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. of Statist.* **11**, 416-431.

Uses ideas from coding theory to simultaneously estimate parameters and choose models. He defines a universal prior for the integers that approximates Jeffreys's prior but is proper.

Seidenfeld, T. (1979). Why I am not an objective Bayesian: some reflections prompted by Rosenkrantz. *Theory and Decision.* **11**, 413-440.

Critique of Rosenkrantz (1977) and, more generally, of objective Bayesian inference. Emphasis is placed on inconsistencies that arise from invariance arguments and from entropy methods based on partial information.

Seidenfeld, T. (1987). Entropy and uncertainty. In *Foundations of Statistical Inference.* I.B. MacNeill and G.J. Umphrey (eds.) 259-287, Reidel.

Critique of the method of maximum entropy. Discusses the disagreement between maximum entropy and conditioning. Then goes on to discuss the Freidman-Shimony (1971) result that it is not possible to extend the algebra to fix this problem, except by extending in a degenerate way. Shows that there is a conflict between maximum entropy and exchangeability. Also critiques the supposed connections between frequencies and maximum entropy.

Severini, Thomas, A. (1991). On the relationship between Bayesian and Non-Bayesian interval estimates. *J. Roy. Statist. Soc. B.* **53**, 611-618.

Shows that in some cases some priors give HPD regions that agree with nominal frequentist coverage to order  $n^{-3/2}$ .

Severini, Thomas, A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. B.* **55**, 533-540.

Shows how to choose intervals so that posterior probability content and frequentist coverage agree to order  $n^{-3/2}$  for a fixed prior.

Sinha, S.K. and Zellner, Arnold. (1990). A note on the prior distributions of Weibull parameters. *SCIMA*, **19**, 5-13.

Examines Jeffreys's prior, Zellner's prior and Hartigan's (1964) asymptotically locally invariant prior for the Weibull.

Smith, A.F.M. and Spiegelhalter, D.J. (1982). Bayes factors for linear and log-linear models with vague prior information. *J.R.S.S. B* **44**, 377-387.

Priors for computing Bayes factors are obtained by using an imaginary prior sample. This sample is the smallest sample that would just favor the null hypothesis.

Spall, J.C. and Hill, S.D. (1990). Least-informative Bayesian prior distributions for finite samples based on information theory. *IEEE Trans. Aut. Control.* **35** 580-583.

Considers the least informative prior to be that which maximizes the expected gain in Shannon information. (Asymptotically, this would be (1).) Approximates this prior by considering a finite set of base priors, especially finite sets of normals, and maximizing the expected gain. See Berger, Bernardo and Mendoza (1989) for a discussion on some problems with maximizing the non-asymptotic version of the gain in information.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 197-206. University of California Press, Berkeley.

Establishes the now famous result that the maximum likelihood estimator (and hence the Bayes estimator using a flat prior) of the mean for a multivariate normal is inadmissible for dimensions greater than or equal to 2.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877-880.

Suppose  $X_i \sim N(\xi_i, 1)$ ,  $i = 1, \dots, n$  independently. The author shows that a one-sided  $\alpha$  level Bayes credible region for  $\sum_i \xi_i^2$  using a flat prior for  $(\xi_1, \dots, \xi_n)$  has frequentist coverage tending to zero as  $n \rightarrow \infty$ . The proof assumes that  $\xi^2 = o(n^2)$ . This assumption is crucial; see Pinkham (1966).

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, Banach center publications, **16**. Warsaw: PWN-Polish Scientific Publishers.

Examines the argument in Welch and Peers (1963) which shows that one sided  $\alpha$  level posterior Bayesian intervals based on (1) have coverage  $\alpha + O(1/n)$ . A different proof is given and then an extension is made for the case where the parameter space is multi-dimensional and there is one parameter of interest. This is the basis of Tibshirani (1989).

Stigler, Stephen M. (1982). Thomas Bayes's' Bayesian inference. *J. Roy. Statist. Soc. A.* **145**, 250-258.

Argues that Bayes's use of a uniform prior for the parameter  $\theta$  of a binomial was not based on the principle of insufficient reason applied to  $\theta$  but rather to  $X_n$ , the number of successes in  $n$  trials. Requiring this for each  $n$  implies a uniform prior for  $\theta$ .

Stone, M. (1963). The posterior  $t$  distribution. *Ann. Math. Statist.* **34**, 568-573.

Shows that the prior  $\pi(\mu, \sigma) \propto \sigma^{-1}$  may be justified because the posterior is the probability limit of a sequence of proper priors. Similar results, of much greater generality are proved in Stone (1965, 1970) and are related to the notion of coherence (section 4.2.1).

Stone, M. (1964). Comments on a posterior distribution of Geisser and Cornfield. *J. Roy. Statist. Soc. B*, **26**, 274-276.

Establishes that a prior recommended by Geisser and Cornfield (1963) for inference in the multivariate normal model cannot be justified as the probability limit of a sequence of proper priors. See section 4.2.1.

Stone, M. (1965). Right Haar measures for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.*, **36**, 440-453.

Shows that the right Haar measure is the only relatively invariant measure such that there exists a sequence of proper priors for which the posteriors converge in probability to the posterior based on the invariant prior, for all  $\theta \in \Theta$ . This type of convergence, a prospective asymptotic justification, is in contrast to the retrospective justification that uses a sequence of proper priors that depends on the observed data case in stable estimation. See section 3.2.

Stone, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **41**, 1349-1353.

Simplifies and generalizes Stone (1965). Shows that an invariant posterior can be obtained as a probability limit of proper priors if and only if the prior is right Haar measure and there exists an asymptotically right-invariant sequence of proper priors. Introduces two examples of non-amenable groups that appear later (Stone 1976) as examples of strong inconsistencies.

Stone, M. (1976). Strong inconsistency from uniform priors (with discussion). *J. Amer. Statist. Assoc.* **71**, 114-125.

Presents two examples of strong inconsistencies in which  $P(A|x) = a$  for all  $x$  but  $P(A|\theta) = b$  for all  $\theta$  where  $a \neq b$ . One, the famous flatland example is based on the free group with two generators (a non-amenable group). The second is the general linear group. These inconsistencies can be viewed as examples of the non-conglomerability of finitely additive priors; see Stone (1982).

Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. In *Logic, Methodology and Philosophy of Science VI. Proc. of the Sixth International Congress of Logic, Methodology and Philosophy of Science*, Hanover 1979, 413-426. North-Holland Publishing Co.

Reviews some problems with improper priors. The first is an example of nonconglomerability of finitely additive priors. The second is a marginalization paradox. He argues that justifying improper priors by claiming they are limits of sequences of proper priors can be misleading.

Stone, M. and Dawid, A.P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika* **59**, 369-375.

Investigates two marginalization paradoxes arising from improper priors. The first involves estimating the ratio of two exponential means. The second involves estimating the coefficient of variation of a normal. More examples are considered in Dawid, Stone and Zidek (1973).

Stone, M. and Springer, B.G.F. (1965). A paradox involving quasi prior distributions. *Biometrika* **52**, 623-627.

Considers some anomalies in a one-way random effects model using improper priors. For example, a Bayesian who uses only a marginal likelihood for inference about the mean and marginal variance ends up with a more concentrated posterior for  $\mu$  than a Bayesian who uses the whole likelihood. See Box and Tiao (1973, page 303-304) for a comment on this paper.

Sudderth, W.D. (1980). Finitely additive priors, coherence and the marginalization paradox. *J. Roy. Statist. Soc. B.* **42** 339-341.

Shows that the marginalization paradox does not occur if finitely additive distributions are used and the posterior is appropriately defined.

Sun, Dongchu and Ye, Keying (1993). Reference prior Bayesian analysis for Normal mean products. Unpublished manuscript.

Extends the work of Berger and Bernardo (1989) for estimating the product of Normal means. Here, the number of means is  $n > 2$ . There is discussion of computation and frequentist coverage.

Sweeting, Trevor J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model.

Argues that Pericchi's (1981) prior for the Normal model with Box-Cox transformations is inappropriate. Instead, he derives a prior based on invariance arguments.

Sweeting, Trevor J. (1985). Consistent prior distributions for transformed models. In *Bayesian Statistics 2*, (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds.) 755-762, Elsevier Science Publishers, North Holland.

Constructs priors for models that are transformations of standard parametric models. This generalizes the work in Sweeting (1984) on Box-Cox transformations. The goal is to use priors that satisfy certain invariance requirements while avoiding priors that cause marginalization paradoxes.

Thatcher, A.R. (1964). Relationships between Bayesian and confidence limits for predictions. *J. R. Statist. Soc. B* **26**, 176-210.

Considers the problem of setting confidence limits on the future number of successes in a binomial experiment. Shows that the upper limits using the prior  $\pi(p) \propto 1/(1-p)$  and the lower limits using the prior  $\pi(p) \propto 1/p$  agree exactly with a frequentist solution.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*. **76** 604-608.

Constructs priors to produce accurate confidence intervals for a parameter of interest in the presence of nuisance parameters. The method is based on results of Stein (1985) and leads to differential equations that can be solved if the parameters are orthogonal. See section 3.7.

Villegas, C. (1971). On Haar priors. *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (Eds.), 409-414. Toronto: Holt, Rinehart & Winston.

Argues for the right Haar measure when the parameter space is the group of non-singular linear transformations. He then derives the marginal distribution for the covariance matrix. Also, the marginal distribution for the subgroup of upper triangular matrices is shown to be right invariant. See section 3.2.

Villegas, C. (1972). Bayes inference in linear relations. *Ann. Math. Stat.* **43** 1767-91.

Suppose we observe vectors  $y_1, \dots, y_n$  with unknown means  $x_1, \dots, x_n$  lying on a  $m$ -dimensional affine subspace. The model is  $\Gamma(y_i - x_i) = v_i$  where  $v_1, \dots, v_n$  are a random sample from a standard Gaussian and  $\Gamma$  is a positive upper triangular matrix with positive diagonal elements. Using the theory in Villegas (1971) a prior is derived which turns out to be a product of several Haar measures.

Villegas, C. (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453-458.

Argues for the  $\pi(\mu, \sigma) \propto \sigma^{-2}$  in the location-scale problem based on invariance. Also shows that the profile likelihood region for  $\mu$  has posterior probability that is a weighted average of conditional confidence levels. Argues that the prior  $\pi(\mu, \sigma) \propto \sigma^{-1}$  requires the “external” judgment of independence.

Villegas, C. (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72**, 651-654.

Two problems are considered. A scale invariance argument is used to justify the prior  $\pi(\lambda) \propto 1/\lambda$  for a Poisson model. In a multinomial model, the prior  $\pi(p_1, \dots, p_k) \propto \prod_i p_i^{-1}$  is justified by requiring permutation invariance and consistency with respect to the collapsing of categories. For example, inferences on  $p_1$  may be made by collapsing the other categories and treating it like a binomial or by finding the marginal of  $p_1$  from the joint posterior. The consistency condition requires these to be the same.

Villegas, C. (1981). Inner statistical inference, II. *Ann. Statist.* **9**, 768-776.

Derives two priors, the inner and outer prior, for group invariant model. The inner prior is left Haar measure and the outer prior is right Haar measure. Shows that, for left Haar measure, the posterior probability of the likelihood set is the posterior expected value of the conditional confidence level. The scale multivariate normal is considered.

Wasserman, Larry (1992). The conflict between improper priors and robustness. Technical report 559, Department of Statistics, Carnegie Mellon University.

Shows that any sequence of neighborhoods around a sequence of increasingly diffuse priors will lead to finite bounds on posterior expectations if and only if the neighborhood is contained in a density ratio neighborhood. This implies that the neighborhood must have limited tail behavior. A proposal is made to replace the improper prior with certain sequences of data-dependent priors.

Welch, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. R. Statist. Soc. B*, **27**, 1-8.

Compares Bayesian intervals based on (1) to some other asymptotically accurate confidence intervals; see also Welch and Peers (1963).

Welch, B.L. and Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318-329.

Considers “Lindley’s problem” of giving conditions under which Bayesian and confidence inference regions are identical. They treat the asymptotic version of the problem showing that, in the one-dimensional case, when Jeffreys’s general rule is used the resulting posterior distribution provides correct confidence coverage probabilities with error of order  $O_p(n^{-1})$ . They show that higher-order agreement is not generally possible. They also extend Lindley’s analysis of the location problem by conditioning on an ancillary statistic. (A general group-theoretic treatment of the latter problem was given by Chang and Villegas, 1986.) See section 3.7.

Ye, Keying. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88**, 360-363.

The author points out that Jeffreys’s rule depends on the stopping rule and that, if this is ignored, the coverage properties of the credible regions can be poor. Also considers the Berger-Bernardo prior for sequential experiments.

Ye, Keying. (1992). Bayesian reference prior analysis on the ratio of variances for the one-way random effects model. Technical report, Department of Statistics, Virginia Polytechnic Institute and State University.

Uses the Berger-Bernardo method for finding priors in the one-way random effects model when the ratio of variance components are of interest. Different groupings of the parameters give different models. These priors are compared.

Ye, Ke-Ying and Berger, James. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika*, **78**, 645-656.

For the exponential model  $Y_{ij} \sim N(\alpha + \beta\rho^{x+x_i^a}, \sigma^2)$ , the prior  $\pi(\alpha, \beta, \sigma, \rho) \propto \sigma^{-1}$  yields an improper posterior. The authors believe that Jeffreys’s prior has undesirable features, citing Mitchell (1967). They consider the Berger-Berger prior for this problem and they study the frequentist coverage properties of the resulting intervals.

Zellner, A. (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods*, A. Aykac and C. Brumat, (Eds.), 201-215. North Holland, Amsterdam.

Defines a maximal data information prior (MDIP) to be that prior which maximizes the difference between the expected Shannon information of the sampling density and the Shannon information of the prior. Several standard priors are obtained this way. For the binomial model, the unusual prior  $\pi(\theta) = c\theta^\theta(1-\theta)^{1-\theta}$  is obtained. An earlier version of these ideas appears in Zellner (1971). See section 3.8.

Zellner, Arnold. (1982). Is Jeffreys a “necessarist”? *Amer. Statist.*, **36**, 28-30.

Argues that Jeffreys should not be considered a necessarist, as he had been classified by Savage. This point was elaborated upon by Kass (1982) along the lines of Section 2.1, here.

Zellner, Arnold. (1993). Models, prior information and Bayesian analysis. Technical report, Graduate School of Business, University of Chicago.

Considers using entropy methods, not just for finding priors but for constructing models as well.

Zellner, Arnold and Min, Chung-ki (1992). Bayesian analysis, model selection and prediction. Invited paper presented at the Symposium in Honor of E.T. Jaynes, University of Wyoming, Laramie.

Considers several problems. First, there is a discussion of maximal data information priors (3.8) with applications to some time series models. Next follows a discussion on model selection and prediction.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds., University of Valencia Press: Valencia.

Jeffreys’s approach to hypothesis testing is extended to deal with the normal linear multiple regression model.