The Selection of Prior Distributions by Formal Rules

Robert E. KASS and Larry WASSERMAN

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 viewpoint 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: Bayes factors; Entropy; Haar measure; Improper priors; Jeffreys's prior; Marginalization paradoxes; Noninformative priors; Reference priors.

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, whereas 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" (Kadane, Dickey, Winkler, Smith, and Peters 1980; Lindley, Tversky, and Brown 1979; Savage 1954), but these inherently problem-specific methods have not been extensively developed 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 1960s and 1970s 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 article is the result of our efforts.

Because 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 explain the techniques he used to construct priors in estimation and testing problems. An essential observation is that Jeffreys's viewpoint evolved toward seeing priors as chosen by convention, rather than as unique representations of ignorance. In Section 3 we list methods for constructing prior distributions. In reviewing these, we observe that various different arguments lead to the priors suggested by Jeffreys or to modified versions of Jeffreys's priors.

In Section 4 we discuss some of the philosophical and practical issues that arise when choosing priors conventionally, by formal rules. Many of these issues are raised only when the priors derived by formal rules are improper. In Section 5.1, however, we argue that impropriety per se is not the practically important source of difficulties. When improper priors lead to badly behaved posteriors, it is a signal that the problem itself may be hard; in this situation diffuse proper priors are likely to lead to similar difficulties. In section 5.2 we add our opinion that reference priors are primarily useful with large samples but may also be helpful when the data analyst is unsure whether a sample is "large." In Section 5.3 we highlight some important outstanding problems. This is followed by an annotated bibliography.

Because our discussion is fairly abstract, it may be 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 μ and variance matrix Σ. There are many special cases of interest. For instance, μ and Σ may depend on some lower-dimensional parameter vector θ; when $\mu = \mu(\theta)$ with $\Sigma = \sigma^2 \cdot \mathbf{I}$, 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 "noninformative" should mean. Thus we prefer to call the priors that we discuss *reference priors.* Because Bernardo (1979a) used the term "reference prior" for a prior chosen by a particular formal rule (as we describe in Sec. 3.5), we have struggled with alterna-

Robert E. Kass and Larry Wasserman are Professors, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. The work of both authors was supported by National Science Foundation (NSF) Grants DMS-9005858 and DMS-9303557 and National Institutes of Health Grant R01-CA54852-01. Wasserman was also supported by NSF grant DMS-9357646. The authors thank Nick Polson for helping with a few annotations; Jim Berger, Ted Chang, Andrew Gelman, Seymour Geisser, Ken Lang, Agostino Nobile, Nancy Paul, Donna Pauler, Teddy Seidenfeld, and Arnold Zellner for useful comments and discussion; and the referees and Associate Editor for their detailed comments on the submitted manuscript.

Journal of the American Statistical Association
September 1996, Vol. 91, No. 435, Review Paper

tive 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 viewpoint from his own subjectivist viewpoint. 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 that Savage surveyed. As we document later, Jeffreys believed in the existence of states of ignorance and 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 (1962, p. 168) 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 viewpoint 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 viewpoint is whether in this case a uniform distribution is advocated, according to what has been called (after Laplace 1820; see Sec. 3.1) 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 to state that "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." But a subjectivist 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 Jeffreys's 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 earlier, 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) that "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.

This section from the first edition of *Theory of Probability* was altered in the second and third editions (1948, pp. 36–37; 1961, pp. 36–37), stating "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, 1965, p. 203) are inaccurate. When Savage remarked (Savage et al., 1962, p. 21) 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 reexamination 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 those of Kass (1982), who was responding 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 bounded interval, the interval $(-\infty, \infty)$, or the interval $(0, \infty)$. For bounded intervals or for the whole real line, Jeffreys took the prior density to be constant. In the second case this of course entails that the prior be improper; that is, 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 σ, 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 $\gamma = \sigma^2$ and the change-of-variables formula is applied to π_σ, then one obtains $\pi_\gamma(\gamma) = 1/\gamma$; thus applications of the rule to σ and γ lead to the same formal prior.

In a 1946 paper, Jeffreys proposed his "general rule." Writing the Fisher information matrix as $\mathbf{I}(\theta)$, where

$$\mathbf{I}(\theta)_{ij} = E\left(-\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right)$$

and l is the log-likelihood, the rule is to take the prior to be

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

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

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

that is, the priors defined by the rule on γ and θ 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 $\pi(\mu, \sigma) = 1/\sigma$, whereas the general rule gives $\pi(\mu, \sigma) = 1/\sigma^2$. The latter he found unacceptable (Jeffreys 1961, p. 182), because when extended to the case of k unknown means μ_i and common variance σ^2, the resulting degrees of freedom in the marginal posterior t distributions of each μ_i depend only on the total number of observations, regardless of the value of k. (Thus, for instance, for a given sample mean and pooled variance from 30 observations, there would be no greater uncertainty about μ_1 with 10 means being estimated than with only 1 mean estimated.) He solved this problem by stating that μ and σ should be judged independent a priori and so should be treated separately, which leads back to the more desirable $\pi(\mu, \sigma) = 1/\sigma$. When the general rule is applied while holding σ fixed, it gives the uniform prior on μ, and when it is applied while holding μ fixed, it gives the prior $\pi(\sigma) \propto 1/\sigma$.

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 location parameters $\mu_1, \ldots, \mu_k$, and an additional multidimensional parameter θ, then the prior he recommended becomes

$$\pi(\mu_1, \ldots, \mu_k, \theta) \propto \det(\mathbf{I}(\theta))^{1/2}, \tag{2}$$

where $\mathbf{I}(\theta)$ is calculated holding $\mu_1, \ldots, \mu_k$ fixed. When there are also scale parameters involved, these become part of θ, and (2) is applied.

Definition. We call (1) and (2) the *prior determined by Jeffreys's general rule,* letting the context distinguish between these two cases. To contrast (2) with the prior obtained by applying (1) when there are location parameters, we refer to (1) as the *prior obtained from Jeffreys's nonlocation rule.* Thus what we call Jeffreys's nonlocation 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 Section 1. When either μ or Σ depend on a parameter vector θ, the information matrix on θ may be obtained via the chain rule from that on (μ, Σ) in the unrestricted case.

We note that 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, but instead recommended a different method.

Suppose that $Y = (Y_1, \ldots, Y_n)$ follow a distribution in a family parameterized by (β, ψ) 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,"

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

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 1995.)

Jeffreys's proposals for priors π_0 and π appear in secs. 5.02, 5.1–5.3, and 6.2 of *Theory of Probability.* Generally, he used his estimation reference priors on the nuisance parameter β. As he showed, and Kass and Vaidyanathan (1992) elaborated on, when ψ and β are assumed orthogonal and a priori independent, the value of the Bayes factor is not very sensitive to the choice of π_0. The prior on ψ, on the other hand, remains important.

When ψ was a probability, as in a binomial problem, Jeffreys (1961, sec. 5.1) used a flat prior on (0, 1). For the normal location problem, in which β is the normal standard deviation and the null hypothesis on the mean ψ becomes H_0: $\psi = 0$, Jeffreys (1961, sec. 5.2) took the prior on ψ to be Cauchy. He argued that as a limiting case, the Bayes factor should become zero if the observed standard deviation were zero, because 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 nonzero 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 β and ψ are one-dimensional but the distribution for the data is arbitrary, by assuming that the parameters are orthogonal and then drawing an analogy with the normal location problem, taking the prior on ψ 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 (1995) have shown how Jeffreys's method may be generalized to arbitrarily many dimensions by replacing Jeffreys's requirement of parameter orthogonality (i.e., that the information matrix be block diagonal for all parameter values) with "null-orthogonality" (i.e., that the information matrix be block diagonal when $\psi = \psi_0$). The log of the resulting approximation has the form $\mathrm{S} + c$, where c is a constant and S is the Schwarz criterion (Schwarz 1978),

$$\mathrm{S} = l_0(\hat{\beta}_0) - l(\hat{\beta}, \hat{\psi}) + \frac{1}{2}\ (m - m_0)\log n,$$

where $\hat{\beta}_0$ maximizes the null-hypothetical log-likelihood $l_0(\beta) = \log p(y|\beta, \psi_0), (\hat{\beta}, \hat{\psi})$ maximizes the unrestricted log-likelihood $l(\beta, \psi)$, and $m - m_0 = \dim(\psi)$. In addition, Kass and Wasserman noted the disappearance of the constant c when a normal prior is used and pointed out the interpretation of such a prior is that "the amount of information in the prior on ψ is equal to the amount of information about ψ contained in one observation." They deemed this a reasonable prior to use and concluded 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), Spiegelhalter and Smith (1982), and Zellner and Siow (1980). Raftery (1995) has proposed a heuristic that is different but similar in spirit, to produce a class of proper reference priors when considering alternative generalized linear models.

I. J. Good has written extensively on Bayes factors. 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 (Good 1967). He suggested subjectively determining the choice of Cauchy location and scale parameters, but in his tabulations (Good 1967, p. 414) used the standard Cauchy as a reference prior.

In most cases, Jeffreys assumed that the initial probabilities of the two hypotheses were equal, which is a reference choice determined by "insufficient reason" (see Sec. 2.1). Alternatives have 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

In this section we describe most of the many methods that have been proposed for constructing reference priors. Whenever possible, we avoid technical details and present the arguments in their simplest forms. As our summary shows, various alternative arguments lead back to Jeffreys's prior or some modification of it. Sometimes the parameter θ can be written in the form $\theta = (\omega, \lambda)$, where ω is a parameter of interest and λ is a nuisance parameter. In this case reference priors that are considered satisfactory for making inferences about θ may not be satisfactory for making inferences about ω. Recent research on reference priors inspired by this latter observation is highlighted in Sections 3.5 and 3.7 and at the end of Section 3.2.

3.1 Laplace and the Principle of Insufficient Reason

If the parameter space is finite, then Laplace's rule, or the principle of insufficient reason, is to use a uniform prior that assigns equal probability to each point in the parameter space. Use of uniform probabilities on finite sets dates back to the origins of probability in gambling problems. The terminology comes from references by Laplace to a lack of sufficient reason for assuming nonuniform probabilities (e.g., Laplace, 1820). Howson and Urbach (1989, p. 40) attributed 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) gave a simple example. Let $\Theta = \{\theta_1, \theta_2\}$, where θ_1 denotes the event that there is life in orbit about the star Sirius and θ_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 ω_1 denotes the event that there is life around Sirius, ω_2 denotes the event that there are planets but no life, and ω_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 is $P(\{\omega_1\}) = 1/3$ if we adopt the second formulation.

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. One could also argue that in a decision problem, the structure of the problem determines the level of partition that is relevant (Chernoff 1954).

A natural generalization is to apply the principle of insufficient reason when the parameter space is continuous, and thereby obtain a flat prior, that is, a prior that is equal to a positive constant. A problem with this rule is that it is not parameterization invariant. For example, if θ is given a uniform distribution, then $\phi = e^\theta$ will not have a uniform distribution. Conversely, if we start with a uniform distribution for ϕ, 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 application of a uniform prior on an infinite set is that of Bayes (1763) who used a uniform prior for estimating the parameter of a binomial distribution. Stigler (1982) argued 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 $n \geq 1$, which entails θ having a uniform prior. This argument is supposed to be more 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, Θ and Ω, and a mapping, $g\colon \Omega \rightarrow \Theta$. We then have the choice of adopting a uniform prior on Θ or adopting a uniform prior μ on Ω, which then induces a prior π on Θ, where π is defined by $\pi(A) = \mu(g^{-1}(A))$. In general, π will not be uniform. In the continuous case, the mapping g corresponds to some reparameterization. In the finite case, Ω is a refinement of Θ, and g relates the original space Θ to its refinement. In the "life on Sirius" example, g is defined by $g(\omega_1) = \theta_1, g(\omega_2) = \theta_2$, and $g(\omega_3) = \theta_2$. In essence, the partitioning paradox is the finite-set version of the lack of parameterization invariance.

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 have been given by Dawid (1983), Hartigan (1964), and Jaynes (1968).

The simplest example of invariance involves the permutation group on a finite set. It is clear that the uniform probability distribution is the only distribution that is invariant under permutations of a finite set. When the parameter space is infinite, the invariance arguments are more complicated. We begin with the normal location model. Suppose that 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 $\pi_2(\phi)$. Because both statisticians are dealing with the same formal model—a normal location model—their reference priors should be the same. Thus we require that $\pi_1 = \pi_2$. On the other hand, because $\phi = \theta + a, \pi_1$ and π_2 can be related by the usual change-of-variables formula. The relationships between π_1 and π_2 should hold for every a, and this implies that they must both be uniform distributions.

This normal location model may be reexpressed in terms of group invariance. Each real number a determines a transformation $h_a\colon \mathbb{R} \rightarrow \mathbb{R}$ defined by $h_a(x) = x + a$. 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,* because $X \sim N(\theta, 1)$ and $Y = h_a(X)$ implies that $Y \sim N(h_a(\theta), 1)$. The uniform prior μ 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 $h_{a,b}: \mathbb{R} \to \mathbb{R}$ 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 $g_{a,b}: \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+$ 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, that is, $P(g_{a,b}A) = P(A)$ for all A and all $(a, b) \in \mathbb{R} \times \mathbb{R}^+$, has density $\pi(\mu, \sigma) \propto 1/\sigma^2$. This is the same prior that 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 *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 previous case we had two groups: one acting on the sample space and one acting on the parameter space. In many cases, it is convenient to think of these as the same group that happens to act differently on the sample space and on the parameter space. For example, in the normal case we have $g_{a,b}(x) = a + bx$ on the sample space but $g_{a,b}(\theta, \sigma) = (a + b\theta, b\sigma)$ on the parameter space. Assume first that G is transitive (i.e., for every $\theta_1, \theta_2 \in \Theta$, there exists $g \in G$ such that $\theta_2 = g\theta_1$) and acts freely (i.e., $g\theta = \theta$ for some $\theta \in \Theta$ only if g is the identity) on both Θ 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 Θ that are again unique (up to a multiplicative constant). Somewhat more complicated cases occur when the group action on the sample space is either nontransitive or nonfree. 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 noncompact and noncommutative, then the left and right Haar measures may be distinct. (See Nachbin 1965 for details on Haar measures.) There are various arguments in favor of one over the other. If we carry out the argument given earlier for the normal location model more generally, then we are led to left Haar measure. Furthermore, Jeffreys's nonlocation 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). Generally, however, right Haar measure is preferred in practice. We now review some arguments that lead to this choice.

Villegas (1981) made the following argument for the right Haar measure in the case in which G is transitive and acts freely. Let λ be a measure on the group G. Choose a reference point $a \in \Theta$. This defines a mapping $\phi_a: G \to \Theta$ by $\phi_a g = ga$, which induces a measure $\pi_a = \lambda\phi_a^{-1}$ on Θ. In other words, we can relate the elements of the group to the elements of the parameter space; a prior on the group induces a prior on the parameter space. If we insist that the measure π_a not depend on the choice of reference point a, then π 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, prop. 4) showed 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 be right Haar measure. (See Sec. 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, Sec. 6.6.2). Related to this is a result proved by Chang and Villegas (1986) that repeated-sampling coverage probabilities and posterior probabilities agree when the prior on the group is right Haar measure (see Sec. 3.7).

These 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 Θ and Θ' are related by a transformation g, then we need only 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 of interest itself, but instead group elements correspond to nuisance parameters; that is, the full parameter vector is $\theta = (\omega, g)$, where $g \in G$ and ω is the parameter of interest. Assuming that ω is an index for the orbits of the group (i.e., the orbit of x is $\{gx; g \in G\}$), Chang and Eaves (1990) recommended the prior $\pi(\omega)\pi(g|\omega)$, where $\pi(g|\omega)$ is right Haar measure and

$$\pi(\omega) = \lim_{n\to\infty} \sqrt{\det(\mathbf{I}_n(\omega))/n}.$$

Here, $\mathbf{I}_n(\omega)$ is the information matrix for y_n, the maximal invariant of the G action. This is similar to the Berger–Bernardo approach (Sec. 3.5), except that Berger and Bernardo would use the nonlocation Jeffreys prior (and hence left Haar measure) for $\pi(g|\omega)$. Datta and Ghosh (1995d) gave a careful description of the relationship between the Chang–Eaves prior and the Berger-Bernardo prior. They also gave a thorough account of the properties of these priors. A recent discussion of the invariance properties of several other priors was provided by Datta and Ghosh (1994).

3.3 Data-Translated Likelihoods

Box and Tiao (1973, sec. 1.3) introduced the notion of "data-translated likelihood" to motivate the use of uniform priors. Let $\mathbf{y}$ be a vector of observations and let $L_{\mathbf{y}}(\cdot)$ be a likelihood function on a real one-dimensional parameter space Φ. According to Box and Tiao (1973, Eq. 1.3.13), the likelihood function is data-translated if it may be written in the form

$$L_{\mathbf{y}}(\phi) = f\{\phi - t(\mathbf{y})\} \tag{4}$$

for some real-valued functions $f(\cdot)$ and $t(\cdot)$, with the definition of $f(\cdot)$ not depending on $\mathbf{y}$. When (4) is satisfied, Box and Tiao recommended using the uniform prior on Φ, because two different samples $\mathbf{y}$ and $\mathbf{y}^*$ will then produce posteriors that differ only with respect to location. That is, the uniform prior produces posterior densities with the same shape for different samples. This feature of the uniform prior is, for Box and Tiao, what makes it "noninformative."

Box and Tiao (1973) then introduced "approximate data-translated likelihood" to motivate Jeffreys's general rule. For a likelihood to be approximately data translated, Box and Tiao required it to be "nearly independent of the data $\mathbf{y}$ except for its location." Operationally, they discussed samples of size n consisting of (iid) observations and began with the normal approximation to the likelihood

$$L_{\mathbf{y}}(\theta) \simeq n(\theta; \hat{\theta}, \hat{\sigma}_{\mathbf{y}}^2), \tag{5}$$

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

$$L_{\mathbf{y}}(\phi) \simeq n(\phi; \hat{\phi}, c/n). \tag{6}$$

The normal approximate likelihood of (6) has the form (4), so that the likelihood itself is, in a sense that Box and Tiao did not make explicit, approximately data translated. Based on the analogy with (4), they recommended using a prior that is uniform on ϕ and noted 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 $\mathbf{y}$ and $\mathbf{y}^*$ are translated images of one another in the sense that

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

for $\phi^* = \phi + \{t(\mathbf{y}^*) - t(\mathbf{y})\}$. Clearly, if (7) holds, then the translation group may be defined on Φ 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 Φ is the whole real line and the support of the distribution is independent of ϕ, 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 generally is not 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 Sec. 3.7.)

3.4 Maximum Entropy

If $\Theta = \{\theta_1, \ldots, \theta_n\}$ is finite and π is a probability mass function on Θ, then the entropy of π, which is meant to capture the amount of uncertainty implied by π, is defined by $\mathcal{E}(\pi) = -\sum \pi(i)\log \pi(i)$. Entropy is a fundamental concept in statistical thermodynamics and information theory (Ash 1965; Shannon 1948; Wiener 1948). 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, $\{E(X_1) = m_1, \ldots, E(X_r) = m_r\}$. Maximum entropy prescribes choosing the prior that maximizes entropy subject to the given moment constraints. The solution is the prior

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

Jaynes (1957, 1968, 1980, 1982, 1983) has been the main developer of entropy-based methods. The maximum entropy method has been used very successfully in many problems including, for example, spectral analysis and image processing. Furthermore, Jaynes has used entropy-based methods for constructing models as well as priors. (A recent review of entropy based methods may be found in Zellner 1991; see also Press 1995, Zellner 1995, and Zellner and Min 1993.) There are, however, some problems with the theory. Seidenfeld (1987) gave 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 that 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), or we can regard the occurrence of A as another constraint; that is, $E(I_A) = 1$, where I_A is the indicator function for the event A. The probability Q maximizes the entropy subject to the constraint set $C_1 = \{E(X) = 3.5, E(I_A) = 1\}$ has values (.22, 0, .32, 0, .47, 0), which conflict 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 reexpressed as an event. In this larger space, perhaps the conflict will disappear. But Friedman and Shimony (1971) and Shimony (1973) have shown that generally there is no such possible extension except in a trivial sense. They showed 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 1. Seidenfeld showed that the Friedman–Shimony result applies not only to entropy, but also 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. Consider again the die example. After rolling a die, we typically can see two or three visible surfaces. That is, in addition to the uppermost side of the die, we can see one or two side faces depending on the orientation of the die. Thus 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. There are 14 such possible outcomes. For example, the outcome (3, equal) means the top face shows 3 and the sum of visible side faces equals 3. The original sample space can now be viewed as a partition of this larger sample space. 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 we should use, Q or P_0.

Entropy methods can be extended to the continuous case by measuring entropy relative to a base density μ. Thus the entropy of a density π with respect to μ is $-\int \pi \log \pi \, d\mu$. Unfortunately, having to choose a base measure is almost as hard as choosing 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) suggested 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 first discuss the notion of missing information, then discuss the stepwise procedure.

3.5.1 Missing Information. Let $X_1^n = (X_1, \ldots, X_n)$ be n iid random variables and let $K_n(\pi(\theta|x_1^n), \pi(\theta))$ be the Kullback–Leibler distance between the posterior density and the prior density, $K_n(\pi(\theta|x_1^n), \pi(\theta)) = \int \pi(\theta|x_1^n) \log(\pi(\theta|x_1^n)/\pi(\theta)) \, d\theta$. Loosely, this is the gain in information provided by the experiment. Let $K_n^\pi = E(K_n(\pi(\theta|x_1^n), \pi(\theta)))$ be the expected gain in information, where the expectation is with respect to the marginal density $m(x_1^n) = \int p(x_1^n|\theta)\pi(\theta) \, d\theta$. Bernardo's (1979a) idea was to think of K_n^π for large n as a measure of the missing information in the experiment, an idea that has its roots in work of Good (1960, 1966) and Lindley (1956). Bernardo (1979a) suggested finding the prior that maximizes $K_\infty^\pi = \lim_{n\to\infty} K_n^\pi$ and called the result "the" reference prior. Because 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 and stick to the term Berger–Bernardo prior. Hartigan (1983, sec. 5.2) used the term *maximal learning prior.* The reason for not performing the foregoing 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, that K_∞^π is usually infinite. (In fact, the infinities can occur for finite n; see Hartigan 1979.) To circumvent this problem, Bernardo found the prior π_n that maximizes K_n^π. He then found the limit of the corresponding sequence of posteriors and finally defined the Berger–Bernardo prior (he used the term "reference prior") as the prior that produces the limiting 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 normality, we have $K_n^\pi = (d/2)\log(n/2\pi e) + \int \pi(\theta) \log(\sqrt{\det(\mathbf{I})}/\pi(\theta)) \, d\theta + o(1)$ as $n \to \infty$, where d is the dimension of θ (see Clarke and Barron 1990 and Ibrigamov and H'asminsky 1973). Define the standardized expected distance $\tilde{K}_n^\pi = K_n^\pi - (d/2)\log(n/2\pi e)$ and the *standardized missing information* by $\tilde{K}_\infty^\pi = \lim_{n\to\infty} \tilde{K}_n^\pi = \int \pi(\theta) \log(\sqrt{\det(\mathbf{I})}/\pi(\theta)) \, d\theta$. It is easy to show 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 iid, there is some question about how to do the asymptotics. An example is the AR(1) process where $X_t = \rho X_{t-1} + \varepsilon_t$ and $\varepsilon_t \sim N(0, 1)$. This example has generated much debate among econometricians.

Phillips (1991) argued in favor of the Jeffreys's prior. His article was followed by a series of papers in which several authors discussed the merits of various approaches. A recent discussion of this example was given by Berger and Yang (1994a). There are two ways to do the asymptotics. One can consider n vectors $X^1, \ldots, X^n$, where each $X^i = (X^i_1, \ldots, X^i_T)$ is a single run of T observations from the process. Maximizing missing information and letting n go to infinity gives the prior determined by Jeffreys's general rule. 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 nonstationarity. If asymptotic missing information is maximized instead for $T \to \infty$, then the prior is either $\pi(\rho) \propto \{\sqrt{1-\rho^2}\}^{-1}$ when the parameter space is restricted to $\rho \in \{-1, 1\}$ or is discrete, with mass at the endpoints if the parameter space is $[a, b]$ with $a < -1$ or $b > 1$. Berger and Yang also considered an alternative prior, which they called the symmetrized reference prior. This is defined by

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

For $\rho \in [-1, 1]$ this is the Berger–Bernardo prior, and the prior outside this range is obtained by the mapping $\rho \to 1/\rho$. Berger and Yang (1994a) 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 (see Section 3.7). More importantly, this is an interesting example showing that the prior can depend on how the asymptotics are carried out.

3.5.2 Nuisance Parameters. Suppose that $\theta = (\omega, \lambda)$, where ω is the parameter of interest and λ is a nuisance parameter. In this case Bernardo suggested modifying his procedure. Ignoring some technical problems, the method is as follows. First, define $\pi(\lambda|\omega)$ to be the Berger–Bernardo prior for λ with ω fixed. Next, find the marginal model $p(x|\omega) = \int p(x|\omega, \lambda)\pi(\lambda|\omega)\, d\lambda$. (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 $p(x|\omega)$. 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 nonlocation Jeffreys prior for λ when ω is fixed (not to be confused with $j(\lambda|\omega)$, the conditional of the nonlocation Jeffreys prior) and $S = \sqrt{|\mathbf{I}|/|\mathbf{I}_{22}|}$. Here $\mathbf{I}$ is the Fisher information matrix and $\mathbf{I}_{22}$ is the portion of the $\mathbf{I}$ corresponding to the nuisance parameters.

As an example, we consider the Neyman–Scott (1948) problem discussed by Berger and Bernardo (1992b), Datta and Ghosh (1995c), and Ghosh (1994). The data consist of n pairs of observations: $X_{ij} \sim N(\mu_i, \sigma^2), i = 1, \ldots, n\ j = 1, 2$. The nonlocation Jeffreys prior is $\pi(\mu_1, \ldots, \mu_n, \sigma) \propto \sigma^{-(n+1)}$. 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, because s^2/n converges to σ^2. By treating σ as the parameter of interest, the Berger–Bernardo method leads to the prior $\pi(\mu_1, \ldots, \mu_n, \sigma) \propto \sigma^{-1}$ in accordance 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. For instance, one might introduce a hierarchical model by putting a distribution on the μ_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 of the posterior (see Berger and Bernardo 1992b). A detailed discussion of a large class of priors for this problem (including minimaxity properties and coverage matching properties) was given by Datta and Ghosh (1995c).

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 1994a), and the product of normal means problem (Berger and Bernardo 1989; Sun and Ye 1994a, 1995), to name just a few. Wolfinger and Kass (1996), use the Berger–Bernardo prior for variance components, which becomes the prior of Jeffreys's general rule applied to the REML likelihood function.

In the foregoing 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.

3.5.3 Related Work. Ghosh and Mukerjee (1992a) and Clarke and Wasserman (1993, 1995) proposed other priors based on Bernardo's missing information idea. Specifically, they worked directly with $\tilde{K}^\pi_\infty(\omega)$, the standardized missing information for ω; that is, the asymptotic expected Kullback distance between the marginal prior $\pi(\omega)$ and the marginal posterior $\pi(\omega|X^n_1)$ minus a standardizing constant,

$$\tilde{K}^\pi_\infty(\omega) = \int\int \pi(\omega, \lambda)\log\frac{S}{\pi(\omega)}\, d\omega\, d\lambda,$$

where $S = \{|\mathbf{I}||\mathbf{I}_{22}|^{-1}\}^{1/2}, \mathbf{I}$ is the Fisher information matrix and $\mathbf{I}_{22}$ is the part of the Fisher information matrix corresponding to λ. Ghosh and Mukerjee (1992a) showed that maximizing $\tilde{K}^\pi_\infty(\omega)$ 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 ω subject to the condition that given ω, the missing information for λ is maximized. But it seems reasonable to examine priors that maximize $\tilde{K}^\pi_\infty(\omega)$.

Ghosh and Mukerjee conjectured, and Clarke and Wasserman showed, that priors that maximize $\tilde{K}_{\infty}^{\pi}(\omega)$ typically are degenerate. Clarke and Wasserman proposed a trade-off prior π_α that maximizes $\tilde{K}_{\infty}^{\pi}(\omega) - \alpha K(j,\pi)$, where the latter term is a penalty term measuring distance from a prior j, where j is usually taken to be the Jeffreys prior or the nonlocation Jeffreys prior. (Recall that $K(j,\pi) = \int j \log(j/\pi)$.) The interpretation is that we are trying to make the distance between the prior for ω and the posterior for ω 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, π_α cannot be written in closed form, but Clarke and Wasserman (1993) gave an algorithm for computing it. Ghosh and Mukerjee suggested shrinking the conditional prior $\pi(\lambda|\omega)$ toward a uniform prior. Later, Clarke and Wasserman (1995) proposed maximizing $\tilde{K}_{\infty}^{\pi}(\omega) - \alpha K(\pi,j)$, thus switching $K(j,\pi)$ to $K(\pi,j)$. 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, $S = \sqrt{|\mathbf{I}|/|\mathbf{I}_{22}|}$. This reduces to j when $\alpha \to \infty$; if S is a function of ω only, then it reduces to the Berger–Bernardo prior when $\alpha = 0$. More generally, π_α 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 λ given ω, then forming the marginal model and maximizing missing information for ω. If ω is the parameter of interest, then perhaps we should maximize missing information for ω given λ. This would ensure that missing information is maximized for ω whatever the value of the nuisance parameter. This might be called a reverse Berger–Bernardo prior. Berger (1992) noted that such a scheme may give results that are similar to the coverage matching methods (see Sec. 3.7). Unfortunately, the prior will then depend on the parameterization of the nuisance parameter. The relationships between the Berger–Bernardo prior and the reverse Berger–Bernardo prior have been studied by Datta and Ghosh (1994, 1995b).

3.6 Geometry

The straightforward verification of invariance of Jeffreys's general rule hides its origin. In outline, Jeffreys (1956, 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 $\det(\mathbf{I}(\theta))^{1/2}$, and natural volume elements of Riemannian metrics are automatically invariant to reparameterization. (See Kass 1989, secs. 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, sec. 2.3) elaborated, arguing that natural volume elements provide appropriate generalizations of Lebesgue measure by capturing intuition favoring "flat" priors and that the information metric may be motivated by statistical considerations. Thus 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. Because 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 by George and McCulloch (1993) and Kass (1981). 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 viewpoint, it may be considered desirable to have posterior probabilities agree with sampling probabilities. Indeed, some statisticians argue that frequency calculations are an important part of applied Bayesian statistics (see Rubin 1984, for example).

To be specific, suppose that θ is a scalar parameter and $l(x)$ and $u(x)$ satisfy $\Pr(l(x) \leq \theta \leq u(x)|x) = 1 - \alpha$, so that $A_x = [l(x), u(x)]$ is a set with posterior probability content $1 - \alpha$. One can also consider the frequency properties of A_x (in the sense of confidence intervals) under repeated sampling given θ. In general, the frequentist coverage probability of A_x will not be $1 - \alpha$. But there are some examples where coverage and posterior probability do agree. For example, if $X \sim N(\theta, 1)$ and θ 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 methods 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 generalization of this result (eliminating the need for a one-dimensional sufficient statistic) was obtained by Welch and Peers (1963) by conditioning on an ancillary. A very general result for group transformation models, essentially due to Stein (1965) and 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 Sec. 3.2).

Some authors seem to applaud the agreement between certain frequentist and Bayesian inference regions, but refrain from justifying a particular prior on the basis of its production of correct frequentist coverage probabilities. Jeffreys (1961) is in this group, as are Box and Tiao (1973) and Zellner (1971). Others, however, such as Berger and Bernardo (1989) and Berger and Yang (1994a, 1994b) used 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_α be a one-sided posterior region for a scalar parameter with posterior probability content $1-\alpha$. Welch and Peers (1963) showed that under certain regularity conditions, the confidence coverage of B_α is $1-\alpha+O(n^{-1/2})$. But 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 Peers (1965, 1968), Stein (1985), and Welch (1965). Recently, there has been interest in extending the Welch–Peers results to multiparameter problems when the parameter θ has been partitioned into a parameter of interest ω and nuisance parameters $\lambda=(\lambda_1,\ldots,\lambda_k)$. Some progress was made on this by Peers (1965) and Stein (1985). Based on Stein's paper, Tibshirani (1989) showed that a prior that leads to accurate confidence intervals for ω can be obtained as follows. Let $\mathbf{I}$ denote the Fisher information matrix and let l be the log-likelihood function. Write

$$\mathbf{I}=\begin{bmatrix}\mathbf{I}_{11} & \mathbf{I}_{12}\\ \mathbf{I}_{21} & \mathbf{I}_{22}\end{bmatrix},$$

where $\mathbf{I}_{11}=-E(\partial^2 l/\partial\omega^2), \mathbf{I}_{22}$ is the $k\times k$ matrix with ijth entry $-E(\partial^2 l/\partial\lambda_i\partial\lambda_j), \mathbf{I}_{21}$ is the $k\times 1$ matrix with jth entry $-E(\partial^2 l/\partial\omega\partial\lambda_j)$, and $\mathbf{I}_{12}$ is the $1\times k$ matrix with ith entry $-E(\partial^2 l/\partial\lambda_i\partial\omega)$. Now, reparameterize the model as (ω,γ), where $\gamma=(\gamma_1,\ldots,\gamma_k)$ is orthogonal to ω. Here $\gamma_i\equiv\gamma(\omega,\lambda_1,\ldots,\lambda_k)$. Orthogonality means that $\mathbf{I}_{12}=\mathbf{I}_{21}=0$ (see Cox and Reid 1987). Tibshirani suggested that the prior $\pi(\omega,\gamma)\propto g(\lambda)\mathbf{I}_{11}^{1/2}$ produces accurate confidence intervals for ω, where $g(\lambda)$ is an arbitrary, positive function of λ. The resulting intervals were rigorously shown to be correct to order $O(n^{-1})$ by Nicolau (1993). For comparison, note that (1) is $\pi(\omega,\gamma)\propto\mathbf{I}_{11}^{1/2}\mathbf{I}_{22}^{1/2}$, and the Berger–Bernardo prior (Sec. 3.5) is $\pi(\omega,\gamma)\propto f(\omega)\mathbf{I}_{22}^{1/2}$ 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) commented on this fact.

Ghosh and Mukerjee (1992a) suggested requiring that

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

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

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

Mukerjee and Dey (1993) found priors that match frequentist coverage to order $o(n^{-1})$ and gave a differential equation that must be solved 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) found priors such that $P(W\le t|X)=P(W\le t|\theta)+o(n^{-1/2})$ for all θ and $t=(t_1,\ldots,t_p)'$ where $W=(W_1,\ldots,W_n)', W_1$ 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 $\sqrt{n}(\theta_1-\hat\theta_1),\ldots,\sqrt{n}(\theta_{i-1}-\hat\theta_{i-1})$. The priors are characterized as having to satisfy a certain differential equation. The idea is that W is an attempt to list the parameters in order of importance in the spirit of the work by Berger and Bernardo. (Ghosh and Mukerjee [1993] reported that in the balanced case of the three-parameter "variance components" model, which we discuss in Section 4.2.5, the Berger–Bernardo priors satisfy their asymptotic coverage matching criterion for some particular orderings of the parameters, but not for others.) Datta and Ghosh (1995a) derived a differential equation characterizing priors for coverage matching up to order $O(n^{-1})$ for a single parameter of interest.

Severini (1991) showed that under certain circumstances, some priors will give HPD regions that agree with their nominal frequentist coverage to order $n^{-3/2}$. Similar calculations, but for which there is a scalar nuisance parameter, were considered by Ghosh and Mukerjee (1992b). DiCiccio and Stern (1994) found 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 have been made by DiCiccio and Martin (1993). On a related topic, Severini (1993) showed 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). Coverage matching methods were also studied by Datta and Ghosh (1995b, 1995c). There have also been attempts to match frequentist and Bayesian procedures in testing problems. We do not attempt a review here (see, for example, DeGroot 1973 and Hodges 1992).

3.8 Zellner's Method

Let $Z(\theta)=-\int p(x|\theta)\log p(x|\theta)\,dx$ be the information about X in the sampling density. (Zellner called this quantity $\mathbf{I}(\theta)$.) Zellner (1971, 1977, 1995, 1996) and Zellner and Min (1993) suggested choosing the prior π that maximizes the difference $G=\int Z(\theta)\pi(\theta)\,d\theta-\int\pi(\theta)\log(\pi(\theta))\,d\theta$. (Note that the negative entropy of the joint density of x and θ is $\int Z(\theta)\pi(\theta)\,d\theta+\int\pi(\theta)\log(\pi(\theta))\,d\theta$. Also note that $G=\int\int\pi(\theta|x)\log[p(x|\theta)/\pi(\theta)]m(x)\,d\theta\,dx$, where $m(x)=\int p(x|\theta)\pi(\theta)\,d\theta$.) The solution is $\pi(\theta)\propto\exp\{Z(\theta)\}$. Zellner called 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, θ) model, it leads to the prior $\pi(\theta) \propto \theta^{\theta}(1-\theta)^{1-\theta}$, which has tail behavior 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 were provided by 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. But Zellner (1991) pointed 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 $\eta_i = h_i(\theta), i = 1, \ldots, m$, then he suggested maximizing

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

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

$$\pi(\theta) \propto \exp\left\{ Z(\theta) + \sum_{i=1}^{m} \log|h_i'(\theta)|/(m+1) \right\}.$$

The resulting prior then has the desired invariance properties over the given transformations. Other side conditions such as moment constraints can be added as well. 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 determinant of the Fisher information equals a constant. This leads to a prior proportional to $j^{\lambda}(\theta)\exp\{Z(\theta)\}$, where λ is a constant and j is Jeffreys's nonlocation rule.

3.9 Decision-Theoretic Methods

Several authors have used decision theoretic arguments to select priors. Chernoff (1954) derived the uniform prior on finite sets by way of eight postulates for rational decision making. Partitioning paradoxes are avoided, because his argument is restricted to sets with fixed, given number of outcomes. Good (1969, 186) took a different approach. He defined $U(G|F)$ to be "the utility of asserting that a distribution is G when, in fact, it is F." He showed that if U takes on a particular form, then (1) is the least favorable prior distribution. Clarke and Barron (1994) showed that (1) achieves the asymptotic minimax cumulative risk when the loss function is Kullback–Leibler distance. Good also related these ideas to Jeffreys's geometrical argument; see Section 3.6.

Hartigan (1965) called 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 θ, θ_0. Hartigan showed that if θ is one-dimensional, then a prior density π is asymptotically unbiased if and only if

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

If the loss function is Hellinger distance, then this gives (1). Hartigan also extended this to higher dimensions. (A referee has provided us with an interesting historical footnote. Apparently, Hellinger did not propose the distance that we now call "Hellinger distance." It was introduced by Bhattacharyya [1943] and independently by Kakutani [1948], who called it Hellinger distance. See also Chentsov 1990.)

Gatsonis (1984) considered 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 showed that this estimate is inadmissible for dimension greater than 3.

Bernardo's method (Sec. 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) showed that Kullback–Leibler divergence is the unique loss function satisfying certain desiderata. Polson (1988, 1992a, 1992b) also discussed this approach.

Kashyap (1971) considered the selection of a prior as a two-person zero-sum game against nature. Using the average divergence between the data density and the predictive density as a loss function, he showed 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 θ. Asymptotically, this leads to (1) and is very similar to Bernardo's (1979a) approach.

3.10 Rissanen's Method

Consider the problem of finding a reference prior for $\Theta = \{1, 2, \ldots\}$. Many familiar techniques, like maximum entropy (see Sec. 3.4) do not give meaningful answers for finding a prior on Θ. 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. We warn the reader that the motivation for the argument that follows is of a much different nature than the other methods considered here. (The interested reader is encouraged to refer to Rissanen 1983 for more details.)

Suppose that you have to construct a code; that is, you must assign a binary string to each integer. We assume that your code is a prefix code, which means that no code word is allowed to be a prefix of another code word. This condition ensures that a decoder can detect the beginning and the end of each code word. Let $L = (L(1), L(2), \ldots)$ be the code word 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 as the ratio of the mean code length to the entropy. This optimization problem can be expressed as

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

The optimization is carried out subject to certain regularity conditions. Rissanen (1983) showed that there is a code that satisfies the minimization condition with code lengths $L_0(n) = \log_2^*(n) + \log_2 c$, where $\log_2^*(n) = \log_2 n + \log_2\log_2 n + \cdots$, where only the finitely many terms of the sum that are positive are included and $c \approx 2.865064$. Rissanen then suggested adopting $Q(n) = 2^{-L_0(n)}$ as a universal prior for the integers. It can be shown that this prior is proper. Because $Q(n) \propto (1/n) \times (1/\log_2 n) \times (1/\log_2\log_2 n)\cdots$, 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 cases. There are some difficulties with the prior, however. First, the motivation for turning the code length L into a prior is not clear. Second, because the prior is proper, we can find a constant n_0 such that $Q(\{1,\ldots,n_0\}) \approx 1$. (This is true for any proper prior.) In certain problems it will not be appropriate to assign a high probability to this particular set. Indeed, one reason for using improper priors is to avoid assigning high probability to any finite set.

3.11 Other Methods

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

Hartigan (1971; 1983, sec. 5.5) defined 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) considered the following method. A point ξ_0 is a *representative point of the probability* P if $\phi(\xi, P)$ is minimized by ξ_0, where ϕ is some discrepancy measure; an example is $\phi(\xi,P) = \int |\xi - x|^2\, dP$. 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 get 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 family, Cifarelli and Regazzini (1987) showed that a large class of priors give perfect association between future and past observations in the sense that there are functions $\phi_n : \mathbb{R}^n \to \mathbb{R}$ such that

$$P(X_N \le x, \phi_n(X_1,\ldots,X_n) \le x) = P(X_N \le x) = P(\phi_n(X_1,\ldots,X_n) \le x)$$

for all $N > n, n = 1,2,\ldots$ and $x \in \mathbb{R}$. These might be regarded as reference priors. Under certain conditions, they showed that the only prior that gives $E(X_N|X_1,\ldots,X_n) = \bar{X}_n$ is the uniform prior on the natural parameter. In a related paper (Cifarelli and Regazzini 1983), these authors showed 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 $\mathrm{var}(E(X_n|\theta))/\mathrm{var}(X_n)$.

Spall and Hill (1990) defined a least informative prior by finding the prior that maximizes expected gain in Shannon information. They approximated this by only looking over convex combinations of a set of base priors. As shown by Berger et al. (1989), maximizing this measure can lead to discrete priors; indeed, this is why Berger and Bernardo maximized 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. Finally, we consider the possibility of performing sensitivity analysis in conjunction with the use of reference priors in Section 4.4.

4.1 Interpretation of Reference Priors

At the risk of oversimplification, it seems useful to identify two interpretations of reference priors. The first interpretation 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 catalog of reference priors for a variety of models. The priors in the catalog do not represent ignorance, but are useful in problems where it is too difficult to elicit an appropriate 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 Sec. 2). For the most part, the mood has shifted toward the second interpretation. In the recent literature, it is rare for anyone to make 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 some practical sense. For example, Berger and Bernardo (1989) considered several priors for estimating the product of two normal means. Rather than defending any particular prior on logical grounds, they instead compared 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: incoherence and strong inconsistencies, the dominating effect of the prior, inadmissibility, marginalization paradoxes and impropriety of the posterior.

4.2.1 Incoherence, Strong Inconsistencies, and Nonconglomerability. An example from Stone (1976, 1982) nicely illustrates potential inconsistencies in using improper priors. Suppose that we flip a four-sided die (a triangular pyramid) many times. The four faces of the die are marked with the symbols $\{a, b, a^{-1}, b^{-1}\}$. Each time that 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 that the symbols a and a^{-1} are juxtaposed in our list, they "annihilate" each other; that is, they cancel each other out. This also occurs 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 $(a\,a)$, because b and b^{-1} annihilate each other. Denote the resulting string by θ. (To avoid annoying edge effects, we assume that the length of θ 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 θ. The annihilation rule is applied if appropriate, resulting in a new string x. The problem is to infer θ from x.

Having seen x, we note that there are four possible values for θ, each with equal likelihood. For example, suppose that $x = (a\,a)$. The extra symbol added by the last toss was either a, a^{-1}, b, or b^{-1}, each with probability 1/4. So θ 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 θ and formally apply the Bayes rule, then the posterior will give probability 1/4 to each of these points and 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 θ, a new symbol results in annihilation with probability 1/4; that is, $P(A|\theta) = 1/4$ for every θ. These two probability statements are contradictory. Because $P(A|x) = 3/4$ for every x, it seems we should conclude that $P(A) = 3/4$. But because $P(A|\theta) = 1/4$ for every θ, 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 was related to a consistency principle by Bondar (1977).

To see what went wrong, let us think about the improper prior as a limit of proper priors. Let π_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 $p \to \infty$. It is tempting to argue that the posterior is valid, because it approximates the posterior using the proper prior π_p. But π_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 $x \in X_p$. (Recall that the total variation distance between probability distributions P and Q is $\sup_A|P(A) - Q(A)|$, the sup being taken over all measurable sets.) 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 zero 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 the HLS 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 $P(A) = \lim_{n\to\infty} P_n(A)$ for all A for which the limit exists. P is an example of a finitely additive prior on the real line that is diffuse in the sense that it gives zero probability to every compact set. On the other hand, P is proper, because $P(\mathbb{R}) = 1$. Compare this to Lebesgue measure μ, which gives positive measure to many compact sets but is improper because $\mu(\mathbb{R}) = \infty$. 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 Θ. Let $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$ be σ fields on these spaces. A model is a collection of probabilities

$\{p_\theta; \theta \in \Theta\}$ on $\mathcal{B}(\mathcal{X})$. An inference is a collection of probabilities $\{q_x; x \in \mathcal{X}\}$ on $\mathcal{B}(\Theta)$. For a bounded function ϕ and a probability P, write $P(\phi) = \int \phi \, dP$.

A prior π on Θ 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 ϕ: $X \to \mathbb{R}$. 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 θ. Heath and Sudderth (1978) showed that an inference $\{q_x; x \in \mathcal{X}\}$ is coherent if and only if there exists a prior π 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 ϕ: $\Theta \times X \to \mathbb{R}$ that are measurable with respect to $\mathcal{B}(\Theta) \times \mathcal{B}(X)$, where m is the marginal induced by the prior π. This means that the joint measure can be disintegrated with respect to the θ partition or the x partition without contradiction. We call q_x a posterior of π. Heath and Sudderth (1989, thm. 3.1) proved 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 $\inf \int \|q_x - \tilde{q}_x\| m(dx) = 0$, where the infimum is over all (proper but possibly finitely additive) priors π, where q_x is the posterior of π, 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 recently showed that the right Haar prior for MANOVA models gives an incoherent posterior (1993a) and gave another example of incoherence for commonly used priors (1993b).

In fact, incoherence and strong inconsistencies are manifestations of a phenomenon called nonconglomerability, 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 nonconglomerability for the following reason. Because $P(A|x) = 3/4$ for all x, conglomerability would imply $P(A) = 3/4$. Similarly, $P(A|\theta) = 1/4$ for all θ implies $P(A) = 1/4$. This contradiction implies that the x partition, the θ partition, or both partitions must display nonconglomerability. The import of HLS coherence is to rule out nonconglomerability in the θ 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 guarantees protection from nonconglomerability only in the θ 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 nonconglomerability in at least one margin (Hill and Lane 1986; Schervish, Seidenfeld, and Kadane 1984).

The HLS approach is only one among many ways of strengthening De Finetti's notion of coherence. Other related ideas have been considered by many authors, among them 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 worth mentioning is the notion using uniform approximations. For example, Mukhopadhyay and Das Gupta (1995) showed the following. Consider a location family that possess a moment-generating function. Let π^x be the posterior using a flat prior. For every $\varepsilon > 0$, there exists a proper, countably additive prior q with posterior q^x such that $d(\pi^x, q^x) < \varepsilon$ for all x. (This implies HLS coherence.) A similar result was given by Mukhopadhyay and Ghosh (1995) where the existence of a moment-generating function is not assumed. 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. This point was made forcefully by Efron (1970, 1973) in his examination of the many normal means problem, which we now describe. Our description closely follows work of Perlman and Rasmussen (1975). Let $X_i \sim N(\theta_i, 1)$ independently, where $i = 1, \ldots, n$, and consider the problem of estimating $\xi = \sum \theta_i^2$. If we adopt a flat prior on $\theta = (\theta_1, \ldots, \theta_n)'$, then the posterior for θ is multivariate normal with mean $X = (X_1, \ldots, X_n)'$ and covariance equal to the identity matrix $\mathbf{I}$. This posterior is coherent in the sense described in Section 4.2.1. The posterior $Q(d\xi|x)$ for ξ is a noncentral χ^2 with n degrees of freedom and noncentrality parameter $Y = \sum_i X_i^2$; we denote this by $\xi|x \sim \chi_n^2(Y)$. Hence $\hat{\xi} = E(\xi|X_1, \ldots, X_n) = Y + n$. There are reasons for thinking that $\hat{\xi}$ is too large, as we now discuss.

Let θ have a $N(0, a\mathbf{I})$ prior. The posterior $Q_a(d\xi|x)$ for ξ 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 in the sense that the total variation distance between Q and Q_a is small when a is large but the means of Q and Q_a are quite different. To see this, note that the expected value of $\hat{\xi}_a = E_{Q_a}(\xi)$ with respect to the marginal m_a for x induced by the $N(0, a\mathbf{I})$ prior satisfies $E(\hat{\xi}_a) = E(\hat{\xi}_0)$, where $\hat{\xi}_0 = Y - n$ is the uniformly minimum variance unbiased estimator (UMVUE) for this problem. This suggests that we can expect $\hat{\xi}_a$ to be close to $\hat{\xi}_0$. Perlman and Rasmussen (1975) confirmed 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 $Z_1 \sim N(1, a^2)$ and $Z_2 \sim N(0, a^2)$, then $E(Z_1) - E(Z_2) = 1$ for all a but the total variation distance between the two distributions tends to 0 as $a \to \infty$.) This shows that closeness in distributional distance, which is what coherence is all about, may not be strong enough to avoid undesirable properties. Efron (1973) emphasized the difficulty in the following terms: Even when a is large, the prior can overwhelm the data. Furthermore, if we adopted a different prior that did not overwhelm the data, there still might be a different function of the parameters, $\xi = \max \theta_i$, for example, where the posterior once again might be driven by the prior rather than by the data.

Similar problems occur with interval estimation for ξ. Under the posterior Q, a one-sided α-level credible region for ξ is $[\Phi_{\alpha,n}(Y), \infty)$, where $P(\chi_n^2(Y) > \Phi_{\alpha,n}(Y)) = \alpha$. Stein (1959) showed that the coverage probability of this interval tends to zero as $n \to \infty$. The strong disagreement with the confidence level suggests that something is amiss. (In his proof, Stein made the reasonable assumption that $\xi = o(n^2)$; Pinkham [1966] showed that if instead $\xi = Mn^h + o(1)$, where $M > 0$ and $h > 2$, then the coverage and posterior probability agree asymptotically.)

These results are disquieting. The difficulty is that in high-dimensional problems, the effects of the prior may be subtle; it may have little influence on some functions of the parameters but may have an overwhelming effect on others. 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 proper priors are necessarily any better in these problems. Indeed, the remarks by Efron (1973) were made to emphasize the practical difficulties with diffuse proper priors that may accompany theoretical difficulties found with improper priors. We return to this point in Section 5.1.

4.2.3 Inadmissibility. Under certain conditions, Bayes estimators based on proper priors lead to admissible estimators, but 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 inadmissible estimator of θ 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 $X = (X_1, \ldots, X_n)'$, is such that there exists another estimator $\gamma = (\gamma_1, \ldots, \gamma_n)'$ with the property that $E_\theta L(\theta, \gamma) \leq E_\theta L(\theta, X)$ for every θ, with strict inequality for at least one θ. (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 do not exist an $\varepsilon > 0$ and an estimator δ_0 such that $E_\theta L(\theta, \delta_0) < E_\theta L(\theta, X) - \varepsilon$ for all θ. (This follows from the fact that X is minimax.) 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, thm. 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 showed 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 was considered by Berger and Strawderman (1993).

4.2.4 Marginalization Paradoxes. Suppose that we have a model $p(x|\alpha, \beta)$ and prior $\pi(\alpha, \beta)$ and that the marginal posterior $\pi(\alpha|x)$ satisfies $\pi(\alpha|x) = \pi(\alpha|z(x))$ for some function $z(x)$. Further suppose that $f(z|\alpha, \beta) = f(z|\alpha)$. It seems that we should be able to recover $\pi(\alpha|x)$ from $p(z|\alpha)$ and some prior $\pi(\alpha)$. Indeed, if $\pi(\alpha, \beta)$ is proper, then this will be the case, as we show. On the other hand, in some situations using improper priors, one obtains $p(z|\alpha, \beta) = p(z|\alpha)$, but $p(z|\alpha)\pi(\alpha)$ is not proportional to $\pi(\alpha|z(x))$ for any $\pi(\alpha)$, in violation of that seemingly desirable recoverability condition. Dawid, Stone, and Zidek (1973) called this a "marginalization paradox" and presented many examples. Here we consider their example 1.

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

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

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

$$\pi(\xi|x) \propto \pi(\xi) \left(\sum_1^\xi z_i + c \sum_{\xi+1}^n z_i \right)^{-n-1} c^{-\xi}.$$

This contradiction can happen only if the prior is improper. To see this, we reproduce the proof from Dawid et al. (1973) that proper priors are immune to this paradox. Let the data be $x = (y, z)$. By assumption, $\pi(\alpha|x)$ is a function of the data through z only, so we can write $\pi(\alpha|x) = a(z, \alpha)$ where

$$a(z, \alpha) = \frac{\int p(y, z|\alpha, \beta)\pi(\alpha, \beta)\, d\beta}{\int \int p(y, z|\alpha, \beta)\pi(\alpha, \beta)\, d\beta\, d\alpha}. \tag{8}$$

Now $p(y, z|\alpha, \beta) = p(z|\alpha, \beta)p(y|z, \alpha, \beta) = p(z|\alpha)p(y|z, \alpha, \beta)$. Substitute this into (8) to conclude that

$$p(z|\alpha) \int p(y|z, \alpha, \beta)\pi(\alpha, \beta)\, d\beta = a(z, \alpha) \int \int p(y, z|\alpha, \beta)\pi(\alpha, \beta)\, d\beta\, d\alpha. \tag{9}$$

Because the prior is proper, we may integrate both sides of (9) with respect to y to get

$$p(z|\alpha)\pi(\alpha) = a(z,\alpha) \int p(z|\alpha)\pi(\alpha)\, d\alpha.$$

Thus $p(z|\alpha)\pi(\alpha)/\int p(z|\alpha)\pi(\alpha)\, d\alpha = a(z,\alpha) = \pi(\alpha|x)$, so the marginal can be recovered from $p(z|\alpha)$ and $\pi(\alpha)$.

An analysis of the problem was presented by Dawid et al. (1973) and the ensuing discussion (see also Hartigan 1983, pp. 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) showed that the marginalization paradox cannot happen with finitely additive priors. An interesting debate about the meaning of this paradox was presented by Jaynes (1980) and discussed by Dawid et al. (1973).

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 $i = 1,\ldots n$, where σ^2 is known. A seemingly natural choice for a prior is $\pi(\mu,\tau) \propto 1/\tau$, but this leads to an improper posterior (see, e.g., Berger 1985, p. 187).

In this problem application of Jeffreys's general rule, based on the marginal distribution of the data, that is, $Y_i \sim N(\mu,\sigma^2+\tau^2)$, 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) gave conditions that guarantee proper posteriors from Jeffreys's general rule for generalized linear models. Dey, Gelfand, and Peng (1993) extended this work for some overdispersed generalized linear models. (Related results were given by Natarajan and McCulloch, 1995.) Berger and Strawderman (1993) gave conditions in the problem of estimating many normal means, and Yang and Chen (1995) provided useful conditions for certain hierarchical normal 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.

Improper posteriors will sometimes 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.

In principle one may avoid improper posteriors by using diffuse proper priors, but in practice this may not really solve the problem. In situations where intuitively reasonable priors produce improper posteriors, unless the likelihood function is highly peaked there may be extreme posterior sensitivity to the choice of the proper prior. We discuss this phenomenon further in Section 5.1. On the other hand, we argue in Section 5.2 that when the likelihood function *is* highly peaked, an improper posterior need not be very worrisome.

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 viewpoint but is generally considered undesirable, because it violates the likelihood principle, which states that two experiments that produce proportional likelihoods should produce the same inferences (Berger and Wolpert 1988). Indeed, reference prior analyses generally violate the likelihood principle, because the definition of a reference prior usually involves an expectation over the sample space. Jeffreys's rule involves the expected information, for example. 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 the data are received. Yet for a fixed prior, we get the same posterior no matter what order the data are processed, assuming independence. Suppose that 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 that we flip the coin until another head appears and suppose that this takes r tosses. Using π_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, then 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 that we process the data.

Another type of sample space dependence is illustrated by right Haar priors (Sec. 3.2). Consider the following example from McCullagh (1992). Let $x_1,\ldots,x_n$ have a Cauchy(μ,σ) distribution. The right Haar prior is $\pi(\mu,\sigma) \propto 1/\sigma$. Now, let $y_i = 1/x_i, i = 1,\ldots,n$. Then the y_i's are distributed as Cauchy(ν,τ), where $\nu = \mu/(\mu^2+\sigma^2)$ and $\tau = \sigma/(\mu^2+\sigma^2)$. Right Haar measure for (ν,τ) is $\pi(\nu,\tau) \propto 1/\tau$. Transforming to (μ,σ), 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.

4.4 Sensitivity Analysis

There now exists a substantial literature on sensitivity analysis in Bayesian inference. Recent accounts with extensive references include work of Berger (1984, 1990, 1994), 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 π in a large class of similar priors Γ. The simplest class of priors is the ε-contaminated class defined by

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

where $\mathcal{P}$ is the set of all priors and $\varepsilon \in [0,1]$ represents the uncertainty in the prior. Of course, this class is familiar in non-Bayesian robustness too (see Huber 1981 and Tukey 1960, for example). If $g(\theta)$ is some function of interest, then it is straightforward to compute

$$\underline{E}_\varepsilon(g|y) = \inf_{P\in\Gamma_\varepsilon(\pi)} E_P(g|x)$$

and

$$\bar{E}_\varepsilon(g|y) = \sup_{P\in\Gamma_\varepsilon(\pi)} E_P(g|x).$$

These bounds may be plotted by ε 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 would be to regard the posterior to be the limit of the posteriors obtained from the sequence of priors π_a as $a \to \infty$, where π_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 will not worry about that here. If we define $\bar{E}_\varepsilon(\theta|y)$ by

$$\bar{E}_\varepsilon(\theta|y) = \lim_{a\to\infty} \sup_{P\in\Gamma_\varepsilon(\pi_a)} E_P(\theta|y)$$

and define $\underline{E}_\varepsilon(\theta|y)$ analogously, then it turns out that $\underline{E}_\varepsilon(\theta|y) = -\infty$ and $\bar{E}_\varepsilon(\theta|y) = \infty$. Because the bounds are always infinite, the ε-contaminated class cannot be used to assess sensitivity when starting with this uniform improper prior.

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 Γ_k be the set of all prior densities p such that

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

for almost all θ, ϕ, where k varies from 1 to ∞. We call this a *density ratio class.* (They considered a more general class, but we confine our attention to this special case.) Again it is easy to compute upper and lower bounds on posterior expectations. Even when π 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 Γ_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 nontrivial bounds on posterior expectations if we interpret the posterior as a limit of posteriors from proper priors? The answer is no. Wasserman (1995) showed that subject to certain regularity conditions, any class that gives finite bounds for improper priors is contained in a density ratio class. Because density ratio classes are already too small, this implies that there is no sufficiently large class that gives nontrivial bounds. Thus current methods for performing formal sensitivity analysis cannot be directly applied to improper reference priors.

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. This is a particular choice of reference prior, however, and thus begs the questions and 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, pp. 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 (e.g., as reported in Wolfinger and Kass 1996). 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 lead to posteriors with undesirable properties. These include incoherence, inadmissibility of Bayes estimators, marginalization paradoxes, sample space dependence, impropriety, 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 The Use of Diffuse Proper Priors

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) called 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. But it does not have any fundamental ability to avoid the difficulties that arise in using reference priors. To specify the meaning of "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 θ is one-dimensional, taking the normal standard deviation to be 10^{10} times the largest imaginable value of θ). But then the posterior based on $\pi^*(\theta)$ will also approximate the formal posterior that would be obtained from $\pi(\theta)$. Although it is true that mathematically the posterior based on $\pi^*(\theta)$ will be proper, computationally the two posteriors will behave in much the same way, and thus any serious data analytical difficulties present with the original posterior will remain with its modification.

One kind of difficulty that a diffuse proper prior fails to avoid involves possible effects of prior domination, outlined in Section 4.2.2. A second involves improper posteriors. To be more specific about the latter, let us return to the normal hierarchical model mentioned in Section 4.2.5 and consider what might happen if we try to replace the prior $\pi(\mu, \tau) \propto 1/\tau$ (equivalently, a uniform prior on $(\mu, \log \tau)$), which leads to an improper posterior, with a diffuse proper prior. Suppose that we observe a small sample that produces a likelihood mildly peaked at the boundary $\tau = 0$, providing modest information that τ is small. We would like to express our knowledge about τ using the posterior, but a ramification of impropriety is that we will be unable to obtain an inference interval for τ having 95% posterior probability. Suppose that we try to get around this problem by using a diffuse proper prior, say a normal with large variances on $(\mu, \log \tau)$. Although that maneuver does create a proper posterior, it is of no practical use in this situation, because the posterior interval that we construct will be extremely sensitive to our choice of prior variance on $\log \tau$; we will find no range of variance values for which the location of our inference interval remains about the same. This limited-data situation is not unrealistic and rare, but rather quite common; in many hierarchical modeling problems there is not much information about certain second-stage variance parameters, which may be nearly zero. Thus introducing diffuse proper priors on all unknown parameters can be dangerous in many frequently encountered settings.

On the other hand, as we said, it is often possible to choose the spread in a proper prior to be suitably large while still obtaining reasonable results. But, as we indicate in the next section, this occurs when the improper prior itself will provide satisfactory results even if it leads to an improper posterior. Our point is that the introduction of diffuse proper priors does not provide an automatic solution to a serious problem: When difficulties with reference priors arise for a particular model it should serve as a warning *about the likelihood function* that care will be needed with proper priors as well. As we have said, we consider 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 viewpoint articulated by Box and Tiao (1973) appears when we consider what they called "data-dominated" cases, which could also be called large-sample cases; they occur when the posterior is dominated by a peaked likelihood function. Box and Tiao emphasized these situations, as did Jeffreys (in many places in his 1961 book, for example). 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. We return once again to the one-dimensional normal hierarchical model discussed in Sections 4.2.2 and 5.1, for which the prior $\pi(\mu, \tau) = \tau^{-1}$ leads to an improper posterior. If the sample size is reasonably large and the data provide information that τ is positive, then the likelihood function will have a sharp peak away from the boundary $\tau = 0$. In this situation, if one ignores a region for τ near the boundary, then the posterior becomes integrable and well behaved; this amounts to substituting for the improper prior a proper version obtained by truncation to a compact set. (The set becomes compact if we also ignore very large values of τ and both large and small values of μ.) Alternatively, we could use a normal prior on $(\mu, \log \tau)$ that has very large variances. In principle, the choice of compact set, or the choice of normal variances, could be very influential on the results—as it would be in the small-sample scenario discussed in Section 5.1. But in this situation, in which the data provide a lot of information, there would be much leeway in the choice: A 95% posterior probability interval for τ would, for instance, be quite stable for various alternative choices among these replacement priors. Furthermore, numerical procedures for posterior calculations to produce a 95% probability interval, for example, would likely perform well with the original prior, producing seemingly sensible results. Here the impropriety of the posterior becomes a mere technicality that may be ignored. We note that Jeffreys (1961, p. 212) also was not worried when he discussed an improper posterior distribution for a median.

To summarize our viewpoint, we see a dichotomy between large-sample and small-sample problems. The discussion of "default" methods should be confined primarily to problems of the former kind, whereas the latter require much more serious attention, beyond what reference analysis can yield. In practice, it may not be immediately apparent whether a particular posterior is likely to be data dominated. In such intermediate cases, well-chosen reference priors (leading to proper posteriors, for example) may play an additional role by allowing a data analyst to obtain preliminary results that would help determine whether the likelihood is highly peaked, and how much additional effort should be expended on getting inferences from a particular model.

With this large-sample motivation in mind, we note that several of the methods that we discussed rely specifically 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 that 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, then two problems present themselves: computation of it and verification that it leads to a proper posterior. For some models, such as the normal families mentioned in Section 1, it is not difficult to compute the prior of Jeffreys's general rule. But for others, such as in many nonnormal hierarchical models, it may not be clear how the prior may be efficiently computed.

Although we have pointed out that results based on improper posteriors are sometimes quite sensible, they will remain worrisome unless the data analyst has good reason to think that 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, then 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, moreover, may be more informative about the two alternative priors than about the data. A useful partial answer ought to involve asymptotics, because 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 directly assess how much the posterior would change if an appropriate informative prior were to replace the reference prior. The negative results of Wasserman (1995) mentioned in Section 4.4 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 that this aspect of judgment may be improved by statistical research and experience as are the many other data analytic judgments that statistical scientists must make.

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

[Received December 1993. Revised November 1995.]

REFERENCES AND 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. A locally homogeneous parameterization can be found and this leads to (1).

——— (1980), "The Interpretation of Improper Prior Distributions as Limits of Data-Dependent Proper Prior Distributions," *Journal of the Royal Statistical Society*, Ser. B, 42, 46–52.

Suggests that improper priors be regarded as limits of data dependent proper priors. Considers an example of a strong inconsistency (sec. 4.2.1) and an example of a marginalization paradox (sec. 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.

Ash, R. B. (1965), *Information Theory*, New York: Dover Publications.

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. Notes 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," *Philosophical Transactions of the Royal Society*, 53, 370–418. Reprinted in *Biometrika*, 45, 243–315, 1958.

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 was presented by Stigler (1986).

Beale, E. M. L. (1960), "Confidence Regions in Nonlinear Estimation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 22, 41–88.

Berger, J. O. (1992), Comment on "Non-Informative Priors," by J. K. Ghosh and R. Mukerjee, in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 205–206.

Berger, J. O. (1984), "The Robust Bayesian Viewpoint" (with discussion), in *Robustness in Bayesian Statistics*, ed. J. Kadane, Amsterdam: North-Holland.

——— (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

——— (1990), "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25, 303–328.

——— (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.

——— (1994), "An Overview of Robust Bayesian Analysis," *TEST*, 3, 5–124.

Berger, J., and Bernardo, J. (1989), "Estimating a Product of Means: Bayesian Analysis With Reference Priors," *Journal of the American Statistical Association*, 84, 200–207.

Applies the method of reference priors 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.

——— (1991), "Reference Priors in a Variance Components Problem," in *Bayesian Inference in Statistics and Econometrics*, eds. P. Goel and N. S. Iyengar, New York: Springer-Verlag.

Applies the Berger–Bernardo method to balanced variance components problems. Derives various priors depending on how the parameters are grouped.

——— (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.

——— (1992b), "On the Development of the Reference Prior Method," in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 35–60.

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 to deal with impropriety. Also, there is some discussion of nonregular 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*, eds. J. Klein and J. Lee, Seoul: Freedom Academy.

Deals with some technical matters related to the Berger–Bernardo approach, including 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, J. O., and Strawderman, W. (1993), "Choice of Hierarchical Priors: Admissibility in Estimation of Normal Means," Technical Report 93-34C, Purdue University, Dept. of Statistics.

Berger, J. O., and Wolpert, R. L. (1988), *The Likelihood Principle*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Vol. 6, Hayward, CA: Institute of Mathematical Statistics.

Berger, J. O., and Yang, R. (1994a), "Noninformative Priors and Bayesian Testing for the AR(1) Model," *Econometric Theory*, 10, 461–482.

See Section 3.5.1.

——— (1994b), "Estimation of a Covariance Matrix Using the Reference Prior," *The Annals of Statistics*, 22, 1195–1211.

The problem is to estimate the covariance matrix Σ in a $N(0, \Sigma)$ model. The authors argue that Jeffreys's prior does not "appropriately shrink the eigenvalues." They decompose Σ as $\Sigma = \mathbf{O}'D\mathbf{O}$, where $\mathbf{O}$ is an orthogonal matrix and D is diagonal with decreasing elements. Then they apply the method of Berger and Bernardo (1992b), treating the parameters as being ordered in importance, with the elements of D being the most important.

Bernardo, J. M. (1979a), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 113–147.

See Section 3.5.

——— (1979b), "Expected Information as Expected Utility," *The Annals of Statistics*, 7, 686–690.

Views the task of reporting a posterior distribution as a decision problem. Shows that if $u(p^*, \theta)$ is the utility of reporting a distribution p^* when θ is the true value of the parameter and if u satisfies certain conditions, then $u(p^*, \theta) = A \log p^* + B(\theta)$ for some constant A and some function B.

——— (1980), "A Bayesian Analysis of Classical Hypothesis Testing," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia, University Press, pp. 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 using Jeffreys's rule (possibly on a truncated space) for the parameter under the alternative is suggested.

Berti, P., Regazzini, P., and Rigo, P. (1991), "Coherent Statistical Inference and Bayes Theorem," *The Annals of Statistics*, 19, 366–381.

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

Bhattacharyya, A. (1943), "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions," *Bulletin of the Calcutta Mathematical Society*, 35, 99–109.

Bondar, J. V. (1977), "A Conditional Confidence Principle," *The Annals of Statistics*, 5, 881–891.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

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

See Section 3.3.

Brunk, H. D. (1991), "Fully Coherent Inference," *The Annals of Statistics*, 19, 830–849.

Investigates coherence in the spirit of Dawid and Stone (1972, 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. Introduces a notion of compatibility between the prior and the posterior to rule out such behavior.

Buehler, R. J. (1959), "Some Validity Criteria for Statistical Inference," *Annals of Mathematical Statistics*, 30, 845–863.

Buehler, R. J., and Feddersen, A. P. (1963), "Note on a Conditional Property of Student's t," *Annals of Mathematical Statistics*, 34, 1098–1100.

Chang, T., and Eaves, D. (1990), "Reference Priors for the Orbit in a Group Model," *The Annals of Statistics*, 18, 1595–1614.

See Section 3.2.

Chang, T., and Villegas, C. (1986), "On a Theorem of Stein Relating Bayesian and Classical Inferences in Group Models," *Canadian Journal of Statistics*, 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.

Chentsov, N. N. (1990), "The Unfathomable Influence of Kolmogorov," *The Annals of Statistics*, 18, 987–998.

Chernoff, H. (1954), "Rational Selection of Decision Functions," *Econometrica*, 22, 422–443.

Derives the principal of insufficient reason for finite spaces based on eight postulates of rational decision making. 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," in *Probability and Bayesian Statistics*, ed. R. Viertl, New York: Plenum Press, pp. 83–95.

See Section 3.11.

Cifarelli, D. M., and Regazzini, E. (1983), "Qualche Osservazione Sull'uso di Distribuzioni Iniziali Coniugate Alla Famiglia Esponenziale," *Statistica*, 43, 415.

See Section 3.11.

Clarke, B., and Barron, A. (1994), "Jeffreys Prior is Asymptotically Least Favorable Under Entropy Risk," *Journal of Statistical Planning and Inference*, 41, 36–60.

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.)

——— (1990), "Information-Theoretic Asymptotics of Bayes Methods," *IEEE Transactions on Information Theory*, 36, 453–471. Clarke, B., and Sun, D. (1992), "Reference Priors Under the Chi-Squared Distance," technical report, Purdue University, Dept. of Statistics.

After noting that Jeffreys's prior can be obtained by maximizing expected Kullback-Leibler distance between prior and posterior, considers 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.

Clarke, B., and Wasserman, L. (1995), "Information Trade-off," *TEST*, 4, 19–38.

See Section 3.5.3.

——— (1993), "Noninformative Priors and Nuisance Parameters," *Journal of the American Statistical Association*, 88, 1427–1432.

See Section 3.5.3.

Consonni, G., and Veronese, P. (1987), "Coherent Distributions and Lindley's Paradox," in *Probability and Bayesian Statistics*, ed. R. Viertl, New York: Plenum, pp. 111–120.

Discuss the Jeffreys–Lindley paradox 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.

——— (1988), "A Note on Coherent Invariant Distributions as Non-Informative Priors for Exponential and Location-Scale Families," Studi Statistici, 19, Universita L. Bocconi, Milano.

Uses Dawid's notion of context invariance (Dawid 1983) to derive noninformative priors for exponential and location-scale families.

Cox, D. R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society*, Ser. B, 49, 1–18.

Datta, G. S., and Ghosh, M. (1994), "On the Invariance of Noninformative Priors," Technical Report 94-20, University of Georgia, Dept. of Statistics.

Explores the invariance (or lack of invariance) of a multitude of priors, including Berger–Bernardo priors and coverage matching priors.

——— (1995a), "On Priors Providing Frequentist Validity for Bayesian Inference," *Biometrika*, 82, 37–45.

Derives a differential equation that characterizes priors π such that

$$P_\theta\left[\frac{\sqrt{n}(t(\theta)-t(\hat{\theta}))}{\sqrt{b}} \le z\right] = P_\pi\left[\frac{\sqrt{n}(t(\theta)-t(\hat{\theta}))}{\sqrt{n}} \le z|X\right] + O_p(n^{-1}),$$

where $\hat{\theta}$ is the posterior mode and b is the asymptotic posterior variance.

——— (1995b), "Some Remarks on Noninformative Priors," *Journal of the American Statistical Association*, 90, 1357–1363.

Compares Berger–Bernardo priors to reverse reference priors (i.e., the Berger–Bernardo prior with the role of nuisance parameter and parameter of interest switched). In particular, gives attention to coverage matching properties. Constructs a general class that matches coverage for each parameter.

——— (1995c), "Hierarchical Bayes Estimators of the Error Variance in One-Way ANOVA Models," *Journal of Statistical Planning and Inference*, 45, 399–411.

Considers a large class of hierarchical Bayes estimators for σ^2 in oneway ANOVA models. Discusses the minimax properties of the estimators. Chooses priors for the second stage using coverage matching arguments. Studies Jeffreys's prior and the Berger–Bernardo prior.

——— (1995d), "Noninformative Priors for Maximal Invariant Parameter in Group Models," *TEST*, 4, 95–114.

Compares several priors, including the Berger–Bernardo prior and the Chang–Eaves prior, in certain models with group structure. Particular attention is given to the marginalization paradox and coverage matching properties.

Dawid, A. P. (1983), "Invariant Prior Distributions," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp. 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. Shows that inferences from Bayesian posteriors are 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.

——— (1973), "Expectation Consistency and Generalized Bayes Inference," *The Annals of Statistics*, 1, 478–485.

Extends work of Dawid and Stone (1972). Drops the assumption that the model gives positive probability to all data points. Characterizes priors that produce a given expectation consistent posterior.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973), "Marginalization Paradoxes in Bayesian and Structural Inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 35, 189–233.

See Section 4.2.4.

de Finetti, B. (1937), "Foresight: Its Logical Laws, Its Subjective Sources," translated and reprinted in *Studies in Subjective Probability*, eds. H. Kyburg and H. Smokler, New York: John Wiley, pp. 93–158.

——— (1972), *Probability, Induction, and Statistics*, New York: John Wiley.

——— (1974, 1975), *Theory of Probability*, Vols. 1 and 2, New York: John Wiley.

DeGroot, M. H. (1973), "Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio," *Journal of the American Statistical Association*, 68, 966–969.

DeRobertis, L., and Hartigan, J. A. (1981), "Bayesian Inference With Intervals of Measures," *The Annals of Statistics*, 9, 235–244.

Dey, D. K., Gelfand, A. E., and Peng, F. (1993), "Overdispersed Generalized Linear Models," Technical Report, University of Connecticut, Dept. of Statistics.

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

DiCiccio, T. J., and Martin, M. M. (1993), "Simple Modifications for Signed Roots of Likelihood Ratio Statistics," *Journal of the Royal Statistical Society*, Ser. 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, T. J., and Stern, S. E. (1994), "Frequentist and Bayesian Bartlett Correction of Test Statistics Based on Adjusted Profile Likelihood," *Journal of the Royal Statistical Society*, Ser. B, 56, 397–408.

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

Dickey, J. M. (1976), "Approximate Posterior Distributions," *Journal of the American Statistical Association*, 71, 680–689.

Eaton, M. (1992), "A Statistical Diptych: Admissible Inferences—Recurrence of Symmetric Markov Chains," *The Annals of Statistics*, 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, M. L., and Sudderth, W. D. (1993a), "The Formal Posterior of a Standard Flat Prior in MANOVA is Incoherent," unpublished manuscript, University of Minnesota, Dept. of Statistics.

Shows 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 by Heath and Sudderth (1978, 1989) and Lane and Sudderth (1983).

——— (1993b), "Prediction in a Multivariate Normal Setting: Coherence and Incoherence," *Sankhya*, 55, 481–493.

Shows that the prior $d\Sigma/|\Sigma|^{(p+1)/2}$ for the covariance matrix of a 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).

——— (1983b), "Minimally Informative Prior Analysis of a Non-Linear Model," *The Statistician*, 32, 117.

Describes work applying the scheme of Bernardo (1979a) to partially nonlinear models (see Eaves 1983a).

——— (1985), "On Maximizing Missing Information About a Hypothesis," *Journal of the Royal Statistical Society*, Ser. B, 47, 263–266.

Discusses the problem of choosing a prior in testing problems from the missing information (Berger–Bernardo) point of view (see also Bernardo 1980 and Pericchi 1984).

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Efron, B. (1970), "Comments on Blyth's Paper," *Annals of Mathematical Statistics*, 41, 1049–1054.

——— (1973), Discussion of "Marginalization Paradoxes in Bayesian and Structural Inference," by A. P. Dawid, M. Stone, and J. V. Zidek (1973) *Journal of the Royal Statistical Society*, Ser. B, 35, 219.

——— (1986), "Why Isn't Everyone a Bayesian?" (with discussion), *The American Statistician*, 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.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transitions of the Royal Society of London*, Ser. A, 222, 309–368.

Fraser, D. A. S. (1968), *The Structure of Inference*, Huntington, NY: Krieger.

Freedman, D., and Purves, R. (1969), "Bayes's Method for Bookies," *The Annals of Mathematical Statistics*, 40, 1177–1186.

Friedman, K., and Shimony, A. (1971), "Jayne's Maximum Entropy Prescription and Probability Theory," *Journal of Statistical Physics*, 3, 381–384.

Gatsonis, C. A. (1984), "Deriving Posterior Distributions for a Location Parameter: A Decision-Theoretic Approach," *The Annals of Statistics*, 12, 958–970.

See Section 3.9.

Geisser, S. (1964), "Posterior Odds for Multivariate Normal Classifications," *Journal of the Royal Statistical Society*, Ser. B, 26, 69–76.

——— (1984), "On Prior Distributions for Binary Trials," *The American Statistician*, 38, 244–251.

Argues that in estimating the success probability θ from a binomial or negative binomial sample, the interval (0, 1) of possible values of θ is a convenient representation of the finitely many values of θ 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 Sec. 3.1). Thus a uniform prior on θ should be used for binomial or negative binomial sampling. Gives predictive distribution calculations as a way of formalizing this argument (see also Stigler 1982).

Geisser, S., and Cornfield, J. (1963), "Posterior Distributions for the Multivariate Normal Distribution," *Journal of the Royal Statistical Society*, Ser. 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. Proposes a class of priors indexed by a parameter ν. 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 ν that gives the usual Student intervals for a single mean and Hotelling's regions for the joint problem. Stone (1964) gave a criticism of this prior in the special case $\nu = 2$—namely, the prior is not a probability limit of proper priors. Geisser (1964) recommended $\nu = p + 1$.

George, E. I., and McCulloch, R. (1993), "On Obtaining Invariant Prior Distributions," *Journal of Statistical Planning and Inference*, 37, 169–179.

Motivated by Jeffreys, defines 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 were also considered by Good (1969), Hartigan (1964), and Kass (1981).

Ghosh, J. K., and Mukerjee, R. (1992a), "Non-Informative Priors" (with comments), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 195–210.

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

——— (1992b), "Bayesian and Frequentist Bartlett Corrections for Likelihood Ratio and Conditional Likelihood Ratio Tests," *Journal of the Royal Statistical Society*, Ser. 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.

——— (1993), "On Priors that Match Posterior and Frequentist Distribution Functions," *Canadian Journal of Statistics*, 21, 89–96.

See Section 3.7.

Ghosh, M. (1994), "On Some Bayesian Solutions of the Neyman–Scott Problem," in *Statistical Decision Theory and Related Topics V*, eds. S. S. Gupta, and J. Berger, New York: Springer-Verlag, pp. 267–276.

Good, I. J. (1960), "Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments," *Journal of the Royal Statistical Society*, Ser. B, 22, 319–331. Corr. 30, 203.

——— (1966), "A Derivation of the Probabilistic Explication of Information," *Journal of the Royal Statistical Society*, Ser. B, 28, 578–581.

——— (1967), "A Bayesian Significance Test for Multinomial Distributions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 29, 399–431.

——— (1969), "What is the Use of a Distribution?" in *Multivariate Analysis*, ed. Krishnaiah, New York: Academic Press, pp. 183–203.

Defines $U(G|F)$ to be "the utility of asserting that a distribution is G when, in fact, it is F." Studies various functional form for U. For a particular form of U, a minimax argument establishes (1) as the least favorable distribution.

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge, U.K.: Cambridge University Press.

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," *Annals of Mathematical Statistics*, 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 θ. An asymptotic version leads to an asymptotically locally invariant (ALI) prior defined in the one-dimensional case by

$$\left(\frac{\partial}{\partial\theta}\right)\log h(\theta) = -E(f_1 f_2)/E(f_2),$$

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 (μ, σ^2) model, we get $\pi(\mu, \sigma) = \sigma^{-5}$.

——— (1965), "The Asymptotically Unbiased Prior Distribution," *Annals of Mathematical Statistics*, 36, 1137–1152.

See Section 3.9.

——— (1966), "Note on the Confidence-Prior of Welch and Peers," *Journal of the Royal Statistical Society*, Ser. B, 28, 55–56.

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 $1 - \alpha + 0(1/\sqrt{n})$.

——— (1971), "Similarity and Probability," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart and Winston, pp. 305–313.

See Section 3.11.

Hartigan, J. A. (1979), Discussion of "Reference posterior distributions for Bayesian inference" by J. M. Bernardo (1979), *Journal of the Royal Statistical Society*, Ser. B, 41, 113–147.

——— (1983), *Bayes Theory*, New York: Springer-Verlag.

Heath, D., and Sudderth, W. (1978), "On Finitely Additive Priors, Coherence, and Extended Admissibility," *The Annals of Statistics*, 6, 333–345.

See Section 4.2.1.

——— (1989), "Coherent Inference From Improper Priors and From Finitely Additive Priors," *The Annals of Statistics*, 17, 907–919.

Gives conditions such that the formal posterior obtained from an improper prior are coherent in the sense of Heath and Sudderth (1978).

Hill, B. M. (1980), "On Some Statistical Paradoxes and Nonconglomerability," in *Bayesian Statistics*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: University Press, pp. 39–65.

Hill, B. M., and Lane, D. (1985), "Conglomerability and Countable Additivity," *Sankhyā*, Ser. A, 47, 366–379.

——— (1986), "Conglomerability and Countable Additivity," in *Bayesian Inference and Decision Techniques*, eds. P. Goel and A. Zellner, Amsterdam: Elsevier, pp. 45–57.

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 nonidentifiable.

Hodges, J. (1992), "Who Knows What Alternative Lurks in the Hearts of Significance Tests?," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 247–266.

Howson, C., and Urbach, P. (1989), *Scientific Reasoning: The Bayesian Approach*, La Salle, IL: Open Court.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley.

Ibrahim, J. G., and Laud, P. W. (1991), "On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior," *Journal of the American Statistical Association*, 86, 981–986.

Gives sufficient conditions for the propriety of the posterior and the existence of moments for generalized linear models. In particular, shows that Jeffreys's prior leads to proper posteriors for many models.

Ibrigamov, I. A., and H'asminsky, R. Z. (1973), "On the Information Contained in a Sample About a Parameter," *2nd International Symposium on Information Theory*, 295–309.

Jaynes, E. T. (1957), "Information Theory and Statistical Mechanics I, II, *Physical Review*, 106, 620–630; 108, 171–190.

——— (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. Makes a connection between maximum entropy and frequency distributions. When a parameter is continuous, a base measure is needed. Recommends using group-invariant measures for this purpose when they are available. A critique of this approach was given by Seidenfeld (1987).

——— (1980), "Marginalization and Prior Probabilities," in *Bayesian Analysis in Econometrics and Statistics*, ed. A. Zellner, Amsterdam: North-Holland, pp. 43–87.

A rebuttal to the Dawid, Stone, and Zidek (1973) paper. 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's thesis is that if we are more careful about notation and write $p(A|x, I_i)$ instead of $p(A|x)$, then 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.

——— (1982), "On the Rationale of Maximum Entropy Methods," *Proceedings of IEEE*, 70, 939–952.

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

——— (1983), *Papers on Probability, Statistics and Statistical Physics*, ed. R. Rosenkrantz, 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," *Proceedings of the Royal Society of London*, Ser. A, 186, 453–461.

Proposes his prior. (Material essentially contained in Jeffreys 1961.)

——— (1955), "The Present Position in Probability Theory," *British Journal for Philosophy of Science*, 5, 275–289.

——— (1957), *Scientific Inference* (2nd ed.) Cambridge, U.K.: Cambridge University Press.

——— (1961), *Theory of Probability* (3rd ed.) London: Oxford University Press.

An extremely influential text that lays the foundation for much of Bayesian theory. Jeffreys's rule is defined and hypothesis testing is studied in great detail. See Section 2.

——— (1963), Review of *The Foundations of Statistical Inference*, by L. J. Savage, M. S. Bartlett, G. A. Barnard, D. R. Cox, E. S. Pearson, and C. A. B. Smith, *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," *Journal of the American Statistical Association*, 75, 845–854.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1986), "Statistical Implications of Finitely Additive Probability," in *Bayesian Inference and Decision Techniques*, eds. P. Goel and A. Zellner, Amsterdam: Elsevier, pp. 59–76.

Discusses various paradoxes that occur with finitely additive probabilities. Some of these problems are discussed in Section 4.2.1.

Kakutani, S. (1948), "On Equivalence of Infinite Product Measures," *The Annals of Mathematics*, 2nd Series, 49, 214–224.

Kashyap, R. L. (1971), "Prior Probability and Uncertainty," *IEEE Transactions on Information Theory*, IT-14, 641–650.

See Section 3.9.

Kass, R. E. (1981), "The Geometry of Asymptotic Inference," Technical Report 215, Carnegie Mellon University, Dept. of Statistics.

——— (1982), Comment on "Is Jeffreys a 'Necessarist'?," by A. Zellner, *The American Statistician*, 36, 390–391.

——— (1989), "The Geometry of Asymptotic Inference," *Statistical Science*, 4, 188–234.

See Section 3.6.

——— (1990), "Data-Translated Likelihood and Jeffreys's Rule," *Biometrika*, 77, 107–114.

See Section 3.3.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors and Model Uncertainty," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Slate, E. H. (1992), "Reparameterization and Diagnostics of Posterior Non-Normality" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 289–306.

——— (1994), "Some Diagnostics of Maximum Likelihood and Posterior Normality," *The Annals of Statistics*, 22, 668–695.

Kass, R. E., and Vaidyanathan, S. (1992), "Approximate Bayes Factors and Orthogonal Parameters, With Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society*, Ser. B, 54, 129–144.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Keynes, J. M. (1921), *A Treatise on Probability*, London: Macmillan.

Lane, D. A., and Sudderth, W. D. (1983), "Coherent and Continuous Inference," *The Annals of Statistics*, 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 prior.

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, chap. 3.

Lindley, D. V. (1956), "On a Measure of the Information Provided by an Experiment," *Annals of Mathematical Statistics*, 27, 986–1005.

——— (1958), "Fiducial Distributions and Bayes's Theorem," *Journal of the Royal Statistical Society*, Ser. 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).

——— (1990), "The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics" (with discussion), *Statistical Science*, 5, 44–49.

Lindley, D. V., Tversky, A., and Brown, R. V. (1979), "On the Reconciliation of Probability Assessments" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 142, 146–180.

McCullagh, P. (1992), "Conditional Inference and Cauchy Models," *Biometrika*, 79, 247–259.

Mitchell, A. F. S. (1967), Comment on "A Bayesian Significance Test for Multinomial Distributions," by I. J. Good, *Journal of the Royal Statistical Association*, Ser. B, 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 ρ yields an improper posterior. Says that the nonlocation Jeffreys prior is unsatisfactory "on commonsense grounds" and proposes an alternative class of priors. (See also Ye and Berger 1991.)

Moulton, B. R. (1993), "Bayesian Analysis of Some Generalized Error Distributions for the Linear Model," unpublished manuscript, Bureau of Labor Statistics, Division of Price and Index Number Research.

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

Mukerjee, R., and Dey, D. K. (1993), "Frequentist Validity of Posterior Quantiles in the Presence of a Nuisance Parameter: Higher-Order Asymptotics," *Biometrika*, 80, 499–505.

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

Mukhopadhyay, S., and DasGupta, A. (1995), "Uniform Approximation of Bayes Solutions and Posteriors: Frequentistly Valid Bayes Inference," *Statistics and Decisions*, to appear.

See Section 4.2.1.

Mukhopadhyay, S., and Ghosh, M. (1995), "On the Uniform Approximation of Laplace's Prior by t-Priors in Location Problems," *The Journal of Multivariate Analysis*, to appear.

See Section 4.2.1.

Nachbin, L. (1965), *The Haar Integral*, New York: van Nostrand.

Natarajan, R., and McCulloch, C. E. (1995), "A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses," *Biometrika*, 82, 639–643.

Neyman, J., and Scott, E. L. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Nicolaou, A. (1993), "Bayesian Intervals With Good Frequentist Behavior in the Presence of Nuisance Parameters," *Journal of the Royal Statistical Society*, Ser. B, 55, 377–390.

Novick, M. R. (1969), "Multiparameter Bayesian Indifference Procedures" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 31, 29–64.

Extends the procedure of 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.

Novick, M. R., and Hall, W. J. (1965), "A Bayesian Indifference Procedure," *Journal of the American Statistical Association*, 60, 1104–1117.

See Section 3.11.

Peers, H. W. (1965), "On Confidence Points and Bayesian Probability Points in the Case of Several Parameters," *Journal of the Royal Statistical Society*, Ser. B, 27, 9–16.

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

——— (1968), "Confidence Properties of Bayesian Interval Estimates," *Journal of the Royal Statistical Society*, Ser. B, 30, 535–544.

Finds priors to make various two-sided intervals—equal-tailed regions, likelihood regions and HPD regions—posterior probability content and frequentist coverage match to order $O(n^{-1})$.

Peisakoff, M. P. (1950), "Transformation Parameters," unpublished doctoral thesis, Princeton University.

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 used by Box and Cox (1964). The resulting priors lead to inferences that mimic the maximum likelihood analysis.

——— (1984), "An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models," *Biometrika*, 71, 575–586.

Argues that in choosing between models $M_1, \ldots, M_k$, the usual posterior tends to favor models having a small expected gain in information and thus offers an explanation for the Jeffreys–Lindley paradox. Suggests avoiding this situation via an unequal prior weighting of the models.

Perks, W. (1947), "Some Observations on Inverse Probability, Including a New Indifference Rule," *Journal of the Institute of 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. 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," *Communications in Statistics*, 4, 455–468.

See Section 4.2.2.

Phillips, P. C. B. (1991), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics*, 6, 333–364.

See section 3.5.1.

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.

——— (1977), "Predictive Distributions and Non-Informative Priors," in *Transactions of the 7th Prague Conference on Information Theory*, Prague: Publishing House of the Czechoslovak Academy of Sciences, pp. 399–407.

See Section 3.11.

Pierce, D. A. (1973), "On Some Difficulties in a Frequency Theory of Inference," *The Annals of Statistics*, 1, 241–250.

Pinkham, R. S. (1966), "On a Fiducial Example of C. Stein," *Journal of the Royal Statistical Society*, Ser. B, 37, 53–54.

See Section 4.2.2.

Poirier, D. (1994), "Jeffreys's Prior for Logit Models," *Journal of Econometrics*, 63, 327–339.

Polson, N. G. (1988), "Bayesian Perspectives on Statistical Modeling," unpublished doctoral dissertation, University of Nottingham, Dept. of Mathematics.

——— (1992a), "On the Expected Amount of Information From a Non-Linear Model," *Journal of the Royal Statistical Society*, Ser. B, 54, 889–895.

——— (1992b), Discussion of "Non-Informative Priors," by J. K. Ghosh and R. Mukerjee, in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, Oxford, U.K.: Clarendon Press, pp. 203–205.

Press, S. J. (1996), "The de Finetti Transform," in *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, Boston: Kluwer Academic Publishers.

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.

Raftery, A. E. (1995), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, to appear.

Regazzini, E. (1987), "De Finetti's Coherence and Statistical Inference," *The Annals of Statistics*, 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," *The Annals of Statistics*, 11, 416–431.

See Section 3.10.

Robinson, G. K. (1978), "On the Necessity of Bayesian Inference and the Construction of Measures of Nearness to Bayesian Form," *Biometrika*, 65, 49–52.

——— (1979a), "Conditional Properties of Statistical Procedures," *The Annals of Statistics*, 7, 742–755.

——— (1979b), "Conditional Properties of Statistical Procedures for Location and Scale Parameters," *The Annals of Statistics*, 7, 756–771.

Rosenkrantz, R. D. (1977), *Inferencė, Method and Decision: Towards a Bayesian Philosophy of Science*, Boston: Reidel.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

Savage, L. J. (1962), "Bayesian Statistics," in *Recent Developments in Information and Decision Theory*, eds. R. F. Machol and P. Gray, New York: Macmillan, reprinted in *The Writings of Leonard Jimmie Savage—A Memorial Selection* (1981), Washington, DC: American Statistical Association and Institute of Mathematical Statistics.

——— (1972), *The Foundations of Statistics* (2nd ed.), New York: Dover.

Savage, L. J., Bartlett, M. S., Barnard, G. A., Cox, D. R., Pearson, E. S., and Smith, C. A. B. (1962), *The Foundations of Statistical Inference*, London: Methuen.

Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (1984), "The Extent of Non-Conglomerability of Finitely Additive Probabilities," *Zeitschrift fur Wahrscheinlictkeitstheorie und Verwandte Gebiete*, 66, 205–226.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Seidenfeld, T. (1979), "Why I am not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz," *Theory and Decision*, 11, 413–440.

Critiques works of Rosenkrantz (1977) and, more generally, objective Bayesian inference. Emphasizes inconsistencies that arise from invariance arguments and from entropy methods based on partial information.

——— (1981), "Paradoxes of Conglomerability and Fiducial Inference," in *Proceedings of the 6th International Congress on Logic Methodology and Philosophy of Science*, eds. J. Los and H. Pfeiffer, Amsterdam: North-Holland.

——— (1987), "Entropy and Uncertainty," in *Foundations of Statistical Inference*, eds. I. B. MacNeill and G. J. Umphrey, Boston: Reidel, pp. 259–287.

See Section 3.4.

Severini, T. A. (1991), "On the Relationship Between Bayesian and Non-Bayesian Interval Estimates," *Journal of the Royal Statistical Society*, Ser. B, 53, 611–618.

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

——— (1993), "Bayesian Interval Estimates Which are Also Confidence Intervals," *Journal of the Royal Statistical Society*, Ser. B, 55, 533–540.

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

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

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell Systems Technical Journal*, 27, 379–423, 623–656.

Shimony, A. (1973), "Comment on the Interpretation of Inductive Probabilities," *Journal of Statistical Physics*, 9, 187–191.

Sinha, S. K., and Zellner, A. (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.

Skala, H. J. (1988), "On σ-Additive Priors, σ-Coherence, and the Existence of Posteriors," in *Risk, Decision and Rationality*, ed. B. R. Munier, Dordrecht: Reidel, pp. 563–574.

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

Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

Spall, J. C., and Hill, S. D. (1990), "Least-Informative Bayesian Prior Distributions for Finite Samples Based on Information Theory," *IEEE Transactions on Automatic Control*, 35, 580–583.

See Section 3.11.

Spiegelhalter, D. J., and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-Linear Models With Vague Prior Information," *Journal of the Royal Statistical Society*, Ser. B, 44, 377–387.

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

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, Berkeley: University of California Press, pp. 197–206.

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 3.

——— (1959), "An Example of Wide Discrepancy Between Fiducial and Confidence Intervals," *Annals of Mathematical Statistics*, 30, 877–880.

See Section 4.2.2.

——— (1965), "Approximation of Improper Prior Measures by Prior Probability Measures," in *Bernoulli-Bayes-Laplace Anniversary Volume: Proceedings of an International Research Seminar Statistical Laboratory*, eds. Jerzy Neyman and Lucien M. Le Cam. New York: Springer-Verlag, pp. 217–240.

——— (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, pp. 485–514.

Examines the argument of Welch and Peers (1963) showing that one-sided α-level posterior Bayesian intervals based on (1) have coverage $\alpha + O(1/n)$. Gives a different proof and then makes an extension for the case where the parameter space is multidimensional and there is one parameter of interest. (This is the basis of Tibshirani 1989.)

Stigler, S. M. (1982), "Thomas Bayes's Bayesian Inference," *Journal of the Royal Statistical Society*, Ser. A, 145, 250–258.

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

——— (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: The Belknap Press of Harvard University Press.

Stone, M. (1963), "The Posterior t Distribution," *Annals of Mathematical Statistics*, 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, were proved by Stone (1965, 1970) and are related to the notion of coherence (Sec. 4.2.1).

——— (1964), "Comments on a Posterior Distribution of Geisser and Cornfield," *Journal of the Royal Statistical Society*, Ser. B, 26, 274–276.

Establishes that one of the priors discussed 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. (Also, see Geisser 1964.)

——— (1965), "Right Haar Measures for Convergence in Probability to Invariant Posterior Distribuiions," *Annals of Mathematical Statistics*, 36, 440–453.

See Section 3.2.

——— (1970), "Necessary and Sufficient Conditions for Convergence in Probability to Invariant Posterior Distributions," *Annals of Mathematical Statistics*, 41, 1349–1353.

See Section 3.2.

——— (1976), "Strong Inconsistency From Uniform Priors" (with discussion), *Journal of the American Statistical Association*, 71, 114–125.

See Section 4.2.1.

——— (1982), "Review and Analysis of Some Inconsistencies Related to Improper Priors and Finite Additivity," in *Logic, Methodology and Philosophy of Science VI, Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science,* Amsterdam: North-Holland, pp. 413–426.

See Section 4.2.1.

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 have been considered by 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 μ than a Bayesian who uses the whole likelihood. (See Box and Tiao 1973, pp. 303–304, for a comment on this paper.)

Sudderth, W. D. (1980), "Finitely Additive Priors, Coherence and the Marginalization Paradox," *Journal of the Royal Statistical Society*, Ser. 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, D. (1995), "A Note on Noninformative Priors for Weibull Distributions," unpublished manuscript.

Sun, D., and Ye, K. (1995), "Reference Prior Bayesian Analysis for Normal Mean Products," *Journal of the American Statistical Association*, 90, 589–597.

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

——— (1994a), "Inference on a Product of Normal Means With Unknown Variances," Technical Report 94-13, Virginia Polytechnic Institute and State University, Dept. of Statistics.

——— (1996b), "Frequentist Validity of the Posterior Quantiles for a Two-Parameter Exponential Family," *Biometrika*, to appear.

Sweeting, T. J. (1984), "On the Choice of Prior Distribution for the Box–Cox Transformed Linear Model," *Biometrika*, 71, 127–134.

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

——— (1985), "Consistent Prior Distributions for Transformed Models," in *Bayesian Statistics 2*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: Elsevier Science Publishers, pp. 755–762.

Constructs priors for models that are transformations of standard parametric models. This generalizes the work of 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," *Journal of the Royal Statistical Society*, Ser. 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.

See Section 3.7.

Tukey, J. W. (1960), "A Survey of Sampling From Contaminated Distributions," in *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford, CA: Stanford University Press, pp. 448–485.

Villegas, C. (1971), "On Haar Priors," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart & Winston, pp. 409–414.

Argues for the right Haar measure when the parameter space is the group of nonsingular linear transformations. Then derives the marginal distribution for the covariance matrix. Also shows that the marginal distribution for the subgroup of upper triangular matrices is right invariant. See Section 3.2.

——— (1972), "Bayes Inference in Linear Relations," *Annals of Mathematical Statistics*, 43, 1767–91.

——— (1977a), "Inner Statistical Inference," *Journal of the American Statistical Association*, 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 μ 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.

——— (1977b), "On the Representation of Ignorance," *Journal of the American Statistical Association*, 72, 651–654.

Uses a scale-invariance argument to justify the prior $\pi(\lambda) \propto 1/\lambda$ for a Poisson model. In a multinomial model, the prior $\pi(p_1, \ldots, p_k) \propto \Pi_i p_i^{-1}$ is justified by requiring permutation invariance and consistency with respect to the collapsing of categories.

——— (1981), "Inner Statistical Inference II," *The Annals of Statistics*, 9, 768–776.

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

von Kries, J. (1886), *Die Principien der Wahrscheinlichkeitsrechnung*, Freiburg: Eine Logische Untersuchung.

Wallace, D. L. (1959), "Conditional Confidence Level Properties," *Annals of Mathematical Statistics*, 30, 864–876.

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

Wasserman, L. (1992), "Recent Methodological Advances in Robust Bayesian Inference" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 583–602.

——— (1995), "The Conflict Between Improper Priors and Robustness," *Journal of Statistical Planning and Inference*, to appear.

See Section 4.4.

Welch, B. L. (1965), "On Comparisons Between Confidence Point Procedures in the Case of a Single Parameter," *Journal of the Royal Statistical Society*, Ser. B, 27, 1–8.

Compares Bayesian intervals based on (1) to some other asymptotically accurate confidence intervals (see also Welch and Peers 1963 and Sec. 3.7).

Welch, B. L., and Peers, H. W. (1963), "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 25, 318–329.

See Section 3.7.

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

Wolfinger, R. D., and Kass, R. E. (1996), Bayesian Analysis of Variance Components Models Via Rejection Sampling, Technical Report, Department of Statistics, Carnegie Mellon University.

Yang, R., and Chen, M.-H. (1995), Bayesian Analysis for Random Coefficient Regression Models Using Noninformative Priors, *Journal of Multivariate Analysis*, 55.

Ye, K. (1993), "Reference Priors When the Stopping Rule Depends on the Parameter of Interest," *Journal of the American Statistical Association*, 88, 360–363.

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.

——— (1994), "Bayesian Reference Prior Analysis on the Ratio of Variances for the Balanced One-Way Random Effects Model," *Journal of Statistical Planning and Inference*, 41, 267–280.

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

Ye, K., and Berger, J. (1991), "Noninformative Priors for Inferences in Exponential Regression Models," *Biometrika*, 78, 645–656.

For the exponential regression model $Y_{\imath\jmath} \sim N(\alpha + \beta\rho^{x+x_i a}, \sigma^2)$, the prior $\pi(\alpha, \beta, \sigma, \rho) \propto \sigma^{-1}$ yields an improper posterior. Posits that Jeffreys's prior has undesirable features, citing Mitchell (1967). Considers the Berger–Bernardo prior for this problem and studies the frequentist coverage properties of the resulting intervals.

Zabell, S. L. (1992), "R. A. Fisher and the Fiducial Argument," *Statistical Science*, 7, 369–387.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley.

——— (1977), "Maximal Data Information Prior Distributions," in *New Developments in the Applications of Bayesian Methods*, eds. A. Aykac and C. Brumat, Amsterdam: North-Holland, pp. 201–215.

See Section 3.8.

——— (1982), "Is Jeffreys a 'Necessarist'?," *The American Statistician*, 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.

——— (1991), "Bayesian Methods and Entropy in Economics and Econometrics," in *Maximum Entropy and Bayesian Methods*, eds. W. T. Grandy, Jr. and L. H. Schick, Boston: Kluwer, pp. 17–31.

——— (1995), "Models, Prior Information and Bayesian Analysis," *The Journal of Econometrics*, to appear.

Considers using entropy methods, not just for finding priors but also for constructing models. Addresses such problems as common parameters in different data densities, iid and non-iid observations, exchangeable sequences, hyperparameters for hierarchical models, multinomial models, prior odds for alternative models and the derivation of statistical models by maximizing entropy subject to particular side conditions.

Zellner, A. (1996), "Past and Recent Results on Maximal Data Information Priors," Technical Report, Graduate School of Business, University of Chicago.

Zellner, A., and Min, C. (1993), "Bayesian Analysis, Model Selection and Prediction," in *Physics and Probability: Essays in Honor of Edwin T. Jaynes*, eds. W. T. Grandy, Jr. and P. W. Milonni, Cambridge, U.K.: Cambridge University Press, pp. 195–206.

Considers several problems, including a discussion of maximal data information priors (see sec. 3.8 with applications to some time series models) and a discussion on model selection and prediction.

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: University of Valencia Press, pp. 585–647.

Extends Jeffreys's approach to hypothesis testing for normal mean to deal with the normal linear multiple regression model.