BAYESIAN STATISTICS 4, pp. 35–60
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.)

On the Development of Reference Priors*

JAMES O. BERGER and JOSÉ M. BERNARDO
Purdue University, USA and *Generalitat Valenciana, Spain*

SUMMARY

The paper begins with a general, though idiosyncratic, discussion of noninformative priors. This provides the background for motivating the recent and ongoing elaborations of the reference prior method for developing noninformative priors, a method initiated in Bernardo (1979). Included in this description of the reference prior method is a new condition that has not previously appeared. Motivation for this new condition is found, in part, in the Fraser, Monette, Ng (1985) example. Extensive discussion of the motivation for reference priors and the specific steps in the algorithm are given, with reference to new examples where appropriate. Also, technical issues in implementing the algorithm are discussed.

Keywords: NONINFORMATIVE PRIORS; MULTIPARAMETER PROBLEMS; STEPWISE COMPUTATION; KULLBACK–LIEBLER DIVERGENCE.

1. INTRODUCTION

1.1. *Perspective on Noninformative Priors*

In some sense, Bayesian analysis is a distinct field only because of noninformative priors. This can certainly be argued from a historical perspective, noting that for virtually 200 years — from Bayes (1763) and Laplace (1774, 1812) through Jeffreys (1946, 1961) — Bayesian statistics was essentially based on noninformative priors. Even today, the overwhelming majority of applied Bayesian analyses use noninformative priors, at least in part. Indeed the only proper priors that are commonly used in practice are those in the early stages of hierarchical models, and these can virtually be thought of as part of the model. (Of course, thinking of such hierarchical distributions as priors rather than, say, random effects models is more natural and is inferentially superior.)

On a philosophical level, things are a bit murkier, but one can still argue for the centrality of noninformative priors. Basically, Bayesian analysis with proper priors is not clearly distinct from probability theory. Indeed, there have been a multitude of Bayesian analyses done throughout history that were viewed as simply being probability analyses. Bayesian analysis with noninformative priors typically does not fit within the usual probability calculus, however. Some Bayesians use foundational arguments to attempt to exclude noninformative priors from consideration, but this also is murky. While axiomatic perspectives typically do suggest that priors should be proper, sensible axiomatics do not rule out proper finitely additive distributions, which operationally can be equivalent to noninformative priors: cf.,

* Research supported by the National Science Foundation, Grants DMS8702620 and DMS8923071, and the Ministerio de Educación y Ciencia, Grant PB870607C0201. Professor José M. Bernardo in on leave of absence from the *Departamento de Estadística e I. O., Universidad de Valencia*, Spain.

Cifarelli and Regazzini (1987), Consonni and Veronese (1989), and Heath and Sudderth (1978).

Finally, even from a pragmatic viewpoint, it might pay to strongly associate Bayesian analysis with use of noninformative priors. How often do we hear "I'm not a Bayesian because statistical inference must be objective" or "I use Bayesian analysis if I actually have usable subjective information, but that is very rare." Statements such as these are, of course, contestable, but the rejoinders "Objectivity is a useless pursuit," and "It may be hard, but you always must try to quantify subjective information," are much less effective arguments than "If your statement were true, the best method of inference would nevertheless be Bayesian analysis with noninformative priors."

It is important, of course, to keep a balanced perspective. Thus today it is obviously to the advantage of Bayesians to claim as their own all true probability inference and to promote the use of subjective priors (especially for problems such as testing of precise hypotheses in which there are no remotely sensible objective answers). And it is important for noninformative prior Bayesians to acknowledge that they are treading on "improper" ground, upon which they do not have the automatic coherency protection provided by proper priors. The noninformative prior Bayesian can run afoul of the likelihood principle (see Berger and Wolpert, 1988, for discussion, but see Wasserman, 1991, for a contrary view), marginalization paradoxes (Dawid, Stone, and Zidek, 1973; but see Jaynes, 1980), strong inconsistency or incoherency (cf. Stone, 1971), and can even encounter the disaster of an improper posterior (see Ye and Berger, 1991, for an example.)

In recognition of these dangers, there are two types of safeguards that are typically pursued by noninformative prior Bayesians. The first, which is the subject of this paper, is the development of a method of generating noninformative priors that seems to avoid the potential problems. The second safeguard is to investigate robustness with respect to the prior, possibly by Bayesian sensitivity studies but more commonly by frequentist evaluation of the performance of the noninformative prior in repeated use. This last type of safeguard is obviously controversial and must be used and interpreted with caution, but it has historically been the most effective approach to discriminating among possible noninformative priors. (Note that the perspective of this second safeguard is that of studying a particular, or several, noninformative priors for a given model, and evaluating their sensibility or performance.)

1.2. *Perspective on Reference Priors*

Bernardo (1979) initiated the reference prior approach to development of noninformative priors, following in the tradition of Laplace and Jeffreys. This tradition is the pragmatic tradition that results are most important; the method should work. If examples are found in which the method fails, it should be modified or adjusted to correct the problem. Thus Laplace (1774, 1812) found that, for the problems he encountered, it worked exceptionally well to simply always choose the prior for θ to be the constant $\pi(\theta) = 1$ on the parameter space Θ. For very small sample sizes, however, it was observed that this led to a significant inconsistency, in that the answer could change markedly depending on the choice of parameterization. (A constant prior for one parameter will not typically transform into a constant prior for another).

This led Jeffreys (1946, 1961) to propose the now famous Jeffreys prior,

$$\pi(\theta) = \sqrt{\det(I(\theta))},$$

where $I(\theta)$ is the Fisher information (see (1.3.1)) and "det" stands for determinant. This method is invariant in the sense of yielding properly transformed priors under reparameterization, and has proved to be remarkably successful in one-dimensional problems. Jeffreys

himself, however, noticed difficulties with the method when θ is multi–dimensional, and would then provide *ad hoc* modifications to the prior.

Bernardo (1979) sought to remove the need for *ad hoc* modifications by systematically dividing multi–dimensional $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ into the "parameters of interest" and the "nuisance parameters," developing the noninformative prior in corresponding stages. As with Jeffreys, this approach was based on information concepts, and indeed the approach yielded the Jeffreys prior in usual one–dimensional problems.

Over the subsequent years and scores of applications the reference prior method has been progressively defined and refined. The papers recording the evolutions in the method that are summarized here include Berger and Bernardo (1989, 1992a, 1992b), Berger, Bernardo, and Mendoza (1989), and Ye (1990). It is noteworthy that the primary impetus for refinement has come from examples, especially the continually–being–invented "counterexamples" to noninformative priors. This explains some of the apparent arbitrariness in the details of the current reference prior method; where different choices were possible, it was through extensive study of examples of application that the ambiguity was resolved. This ongoing process is reviewed in this paper, with several previously unpublished conditions and examples being highlighted.

The above should not be construed as an admission that the reference prior method is solely *ad hoc*. Far from it, the method is grounded in a very appealing heuristic which even today is the source of new insight. For instance, the condition (2.2.5) in Section 2.2 has only recently been added to our description of the reference prior method. This condition arose out of study of the delightful Fraser, Monette, Ng (1985) counterexample (discussed in Section 3.2), the resolution of which required us to return to the fundamental heuristic.

1.3. *Perspective on Methods for Deriving Noninformative Priors*

First, it is important to clarify that we are concerned here with *methods* of developing noninformative priors, not noninformative priors themselves. A method takes as input the statistical model (possibly including the design and / or stopping rule) and possibly the actual data, and produces as output a prior distribution. (Ultimately, of course, it is the posterior distribution which is desired; in some situations it might even be possible to directly develop a reference posterior.) Thus the Jeffreys "method" takes the sample density $f(x|\theta)$ for the data $X \in \mathcal{X}$, computes the Fisher information $I(\theta)$, i.e. the $(k \times k)$ matrix with elements

$$I_{ij}(\theta) = -E_\theta \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta) \right), \tag{1.3.1}$$

with E_θ denoting expectation over X with θ given, and finally produces the prior density

$$\pi(\theta) = \sqrt{\det(I(\theta))}. \tag{1.3.2}$$

In comparing methods of producing noninformative priors, a variety of criteria are involved. The three most important criteria are simplicity, generality, and trustworthiness.

By far the simplest method is to follow Laplace and always choose $\pi(\theta) = 1$. In practice this is, indeed, often quite reasonable, since (as partly argued by Laplace) parameterizations are often chosen to reflect a vague notion of prior uniformity. This simple choice fails on enough problems of interest, however, that a more reliable general method is needed.

On the simplicity scale, the reference prior approach is at the opposite extreme. Indeed, computation of a reference prior is so complex that it typically requires the involvement of a research statistician. Of course, for each statistical model computation of the associated

reference priors need be done only once, with the resulting reference priors (or perhaps posteriors) being made available in the literature.

In terms of generality, Laplace's method and the reference prior method are virtually universally applicable. The Jeffreys method is quite universal, but does require existence of $I(\theta)$ and, typically, additional regularity conditions such as asymptotic normality of the model. Other methods vary widely in terms of generality, some applying only in univariate problems, some requiring special group invariant or transformation structures, etc. Our goal has been the development of a universal method.

Trustworthiness of the method is a rather nebulous concept, essentially referring to how often the method yields a noninformative prior with undesirable properties. Undesirable properties include impropriety of the posterior (clearly the worst possibility), inconsistency or incoherency of resulting statistical procedures, lack of invariance to reparameterization, marginalization paradoxes, lack of reasonable coverage probabilities for resulting Bayesian credible sets, and unremovable singularities in the posterior. The best way to gauge the trustworthiness of a method is to try it on the large set of challenging "counterexamples" to noninformative priors that have been developed over the years. In this sense the reference prior method is very trustworthy; it does not yield a bad answer in any of the counterexamples.

Conspicuously absent in this discussion of methods for developing noninformative priors has been the notion of how to define "noninformative." Most methods begin with some attempt at measuring the amount of information in a prior or the amount of influence that the prior has on the answer. One could debate the sensibility or value of each such measure (and, of course, we are supporters of the measure underlying reference priors) but, on the whole, we feel that this is a somewhat tangential issue. No sensible absolute way to define "noninformative" is likely to ever be found, and often the most natural ways give the silliest answers (cf. Berger, Bernardo, and Mendoza, 1989).

Another aspect of this is the debate over the name "noninformative" versus, say, "reference." Many object to the former, feeling that it carries a false promise. Reference priors are sensibly named (see Bernardo, 1979) and less objectionable in this regard. Other names such as the "standard" or "default" prior have been proposed, the idea being that the profession should ultimately agree on a standard default prior for use with each particular model. Trying to change historical nomenclature is, however, generally a waste of time, so we have chosen to continue using "noninformative" to refer to the general area, and "reference" to refer specifically to reference priors.

No attempt is made here to survey the wide variety of methods for deriving a noninformative prior and to evaluate them by the above criteria. Those most similar to the reference prior approach, in the sense of explicitly considering parameters of interest and nuisance parameters separately, include Stein (1985), Tibshirani (1989), and Ghosh and Mukerjee (1992).

2. THE REFERENCE PRIOR METHOD

The reference prior method is presented here, in full detail. Unfortunately, it is notationally quite complex. The reader interested only in the ideas can skip to Section 3. For a description of the algorithm in the much simpler two-parameter case, see Berger and Bernardo (1989).

2.1. *Introduction and Notation*

In Section 2.2, the general reference prior method will be described. This method is typically very hard to implement. For the regular case, in which asymptotic normality of the model holds, a considerable simplification of the algorithm occurs. This simplification is given in Section 2.3, which is a review of Berger and Bernardo (1992a and 1992b).

We assume that the θ_i are separated into m groups of sizes $n_1, n_2, \ldots, n_m$, and that these groups are given by

$$\theta_{(1)} = (\theta_1, \ldots, \theta_{n_1}),\ \theta_{(2)} = (\theta_{n_1+1}, \ldots, \theta_{n_1+n_2}),$$
$$\ldots \theta_{(i)} = (\theta_{N_{i-1}+1}, \ldots, \theta_{N_i}), \ldots, \theta_{(m)} = (\theta_{N_{m-1}+1}, \ldots, \theta_k),$$

where $N_j = \sum_{i=1}^{j} n_i$. Also, define

$$\theta_{[j]} = (\theta_{(1)}, \ldots, \theta_{(j)}) = (\theta_1, \ldots, \theta_{N_j}),$$
$$\theta_{[\sim j]} = (\theta_{(j+1)}, \ldots, \theta_{(m)}) = (\theta_{N_j+1}, \ldots, \theta_k),$$

with the conventions that $\theta_{[\sim 0]} = \theta$ and $\theta_{[0]}$ is vacuous.

We will denote the Kullback–Liebler divergence between two densities g and h on Θ by

$$D(g, h) = \int_\Theta g(\theta) \log[g(\theta)/h(\theta)] d\theta. \tag{2.1.1}$$

Finally, let $Z_t = \{X_1, \ldots, X_t\}$ be the random variable that would arise from t conditionally independent replications of the original experiment, so that Z_t has density

$$p(z_t|\theta) = \prod_{i=1}^{t} f(x_i|\theta). \tag{2.1.2}$$

2.2. *The General Case*

The general reference prior method can be described in four steps. Justification and motivation will be given in Section 3.

Step 1. Choose a nested sequence $\{\Theta^\ell\}$ of compact subsets of Θ such that $\bigcup_{\ell=1}^{\infty} \Theta^\ell = \Theta$. (This step is unnecessary if the reference priors turn out to be proper.)

Step 2. Order the coordinates $(\theta_1, \ldots, \theta_k)$ and divide them into the m groups $\theta_{(1)}, \ldots, \theta_{(m)}$. Usually it is best to have $m = k$, and the order should typically be according to inferential importance; in particular, the first parameters should be the parameters of interest.

Step 3. For $j = m, m-1, \ldots, 1$, iteratively compute densities $\pi_j^\ell(\theta_{[\sim(j-1)]}|\theta_{[j-1]})$, using

$$\pi_j^\ell(\theta_{[\sim(j-1)]}|\theta_{[j-1]}) \propto \pi_{j+1}^\ell(\theta_{[\sim j]}|\theta_{[j]}) h_j^\ell(\theta_{(j)}|\theta_{[j-1]}), \tag{2.2.1}$$

where $\pi_{m+1}^\ell \equiv 1$ and h_j^ℓ is computed by the following two steps.

Step 3a: Define $p_t(\theta_{(j)}|\theta_{[j-1]})$ by

$$p_t(\theta_{(j)}|\theta_{[j-1]}) \propto \exp\left\{ \int p(z_t|\theta_{[j]}) \log p(\theta_{(j)}|z_t, \theta_{[j-1]}) dz_t \right\}, \tag{2.2.2}$$

where (using $p(\cdot)$ generically to represent the conditional density of the given variables)

$$\begin{aligned} p(z_t|\theta_{[j]}) &= \int p(z_t|\theta) \pi_{j+1}^\ell(\theta_{[\sim j]}|\theta_{[j]}) d\theta_{[\sim j]}, \\ p(\theta_{(j)}|z_t, \theta_{[j-1]}) &\propto p(z_t|\theta_{[j]}) p_t(\theta_{(j)}|\theta_{[j-1]}). \end{aligned} \tag{2.2.3}$$

Step 3b: Assuming the limit exists, define

$$h_j^{\ell}(\theta_{(j)}|\theta_{[j-1]}) = \lim_{t\to\infty} p_t(\theta_{(j)}|\theta_{[j-1]}). \tag{2.2.4}$$

Comment: In (2.2.2), p_t is only defined implicitly, since $p(\theta_{(j)}|z_t, \theta_{[j-1]})$ on the right hand side also depends on p_t (see (2.2.3)). In practice, it is thus usually very difficult to compute the p_t and find their limit. In the regular case discussed in the next section, however, this difficulty can be circumvented.

Step 4. Define a reference prior, $\pi(\theta)$, as any prior for which

$$E_{\ell}^{X} D(\pi_1^{\ell}(\theta|X), \pi(\theta|X)) \longrightarrow 0 \text{ as } \ell \to \infty, \tag{2.2.5}$$

where D is defined in 2.1.1 and E_{ℓ}^{X} is expectation with respect to

$$p^{\ell}(x) = \int_{\Theta} f(x|\theta)\pi_1^{\ell}(\theta)d\theta$$

(writing $\pi_1^{\ell}(\theta)$ for $\pi_1^{\ell}(\theta_{[\sim 0]}|\theta_{[0]})$). Typically one determines $\pi(\theta)$ by the simple relation

$$\pi(\theta) = \lim_{\ell\to\infty} \frac{\pi_1^{\ell}(\theta)}{\pi_1^{\ell}(\theta^*)}, \tag{2.2.6}$$

where θ^* is any fixed point in Θ with positive density for all π_1^{ℓ}, and then verifies that (2.2.5) is satisfied.

Comment: Note that (2.2.5) really defines a *reference posterior*; we convert to a reference prior mainly for pedagogical reasons.

2.3. *The Regular Case*

If the model is regular, in the sense that the replicated $p(z_t|\theta)$ is asymptotically normal, then Step 3 in Section 2.2 can be done in an explicit fashion. The following notation is needed, where $I(\theta)$ is the Fisher information matrix with elements given by (1.3.1) and $S(\theta) = (I(\theta))^{-1}$. Often, we will write just I and S for these matrices. Write S as

$$S = \begin{pmatrix} A_{11} & A_{21}^t & \cdots & A_{m1}^t \\ A_{21} & A_{22} & \cdots & A_{m2}^t \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix}$$

so that A_{ij} is $(n_i \times n_j)$, and define

$$S_j \equiv \text{ upper left } (N_j \times N_j) \text{ corner of } S, \text{ with } S_m \equiv S, \text{ and } H_j \equiv S_j^{-1}.$$

Then the matrices

$$h_j \equiv \text{ lower right } (n_j \times n_j) \text{ corner of } H_j, j = 1, \ldots, m$$

will be of central importance. Note that $h_1 \equiv H_1 \equiv A_{11}^{-1}$ and, if S is a block diagonal matrix (i.e., $A_{ij} \equiv 0$ for all $i \neq j$), then $h_j \equiv A_{jj}^{-1}, j = 1, \ldots, m$. Finally, if $\Theta^* \subset \Theta$, we will define

$$\Theta^*(\theta_{[j]}) = \{\theta_{(j+1)}: (\theta_{[j]}, \theta_{(j+1)}, \theta_{[\sim(j+1)]}) \in \Theta^* \text{ for some } \theta_{[\sim(j+1)]}\}. \tag{2.3.1}$$

We will use the common symbols $|A|$ = determinant of A, and $1_\Omega(y)$ equals 1 if $y \in \Omega$, 0 otherwise.

Step 3 from Section 2.2 can, in the regular case, by replaced by the following, which is essentially taken from Berger and Bernardo (1992b).

Step 3': To start, define

$$\pi_m^l(\theta_{[\sim(m-1)]}|\theta_{[m-1]}) = \pi_m^l(\theta_{(m)}|\theta_{[m-1]}) = \frac{|h_m(\theta)|^{1/2} 1_{\Theta^l(\theta_{[m-1]})}(\theta_{(m)})}{\int_{\Theta^l(\theta_{[m-1]})} |h_m(\theta)|^{1/2} d\theta_{(m)}} \tag{2.3.2}$$

For $j = m-1, m-2, \ldots, 1$, define

$$\pi_j^l(\theta_{[\sim(j-1)]}|\theta_{[j-1]}) = \frac{\pi_{j+1}^l(\theta_{[\sim j]}|\theta_{[j]}) \exp\{\frac{1}{2} E_j^l[(\log|h_j(\theta)|)|\theta_{[j]}\} 1_{\Theta^l(\theta_{[j-1]})}(\theta_{(j)})}{\int_{\Theta^l(\theta_{[j-1]})} \exp\{\frac{1}{2} E_j^l[(\log|h_j(\theta)|)|\theta_{[j]}\} d\theta_{(j)}} \tag{2.3.3}$$

where

$$E_j^l[g(\theta)|\theta_{[j]}] = \int_{\{\theta_{[\sim j]}:(\theta_{[j]},\theta_{[\sim j]})\in\Theta^l\}} g(\theta)\pi_{j+1}^l(\theta_{[\sim j]}|\theta_{[j]}) d\theta_{[\sim j]}. \tag{2.3.4}$$

The calculation of the m–group reference prior is greatly simplified under the condition

$$|h_j(\theta)| \text{ depends only on } \theta_{[j]}, \text{ for } j = 1, \ldots, m. \tag{2.3.5}$$

Then (see Lemma 2.2.1 in Berger and Bernardo, 1992b)

$$\pi^l(\theta) = \left(\prod_{i=1}^m \frac{|h_i(\theta)|^{1/2}}{\int_{\Theta^l(\theta_{[i-1]})} |h_i(\theta)|^{1/2} d\theta_{(i)}} \right) 1_{\Theta^l}(\theta). \tag{2.3.6}$$

3. MOTIVATION FOR THE REFERENCE PRIOR METHOD

3.1. *Information and Replication*

For simplicity, suppose there is a single parameter θ with a compact Θ (or that we are operating on the compact $\Theta^l \subset \Theta$). Suppose that it is desired to define a noninformative prior, $\pi(\theta)$, as that prior which "maximizes the amount of information about θ provided by the data, x." The most natural measure of the expected information about θ provided by X, when π is the prior distribution, is (Shannon, 1948; Lindley, 1956) $I^\theta = E^X D(\pi(\theta|X), \pi(\theta))$, where D is the Kullback–Liebler divergence defined in (2.1.1) and E^X stands for expectation with respect to the marginal density of X, $p(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$.

Unfortunately, basing the analysis on I^θ is not very satisfactory, as is discussed in Berger, Bernardo, and Mendoza (1989). Indeed, it is shown therein that the $\pi(\theta)$ which maximizes I^θ (possibly with θ restricted to the compact Θ^ℓ) is typically a *discrete* distribution, even when Θ is, say, a connected subset of Euclidean space. Clearly such a $\pi(\theta)$ would be a very unappealing noninformative prior.

Bernardo (1979) considered a variant of this approach, defining

$$I_t^\theta = E^{Z_t} D(\pi(\theta|Z_t), \pi(\theta)), \tag{3.1.2}$$

where Z_t consists of t replicates of X as discussed in Section 2.1. The underlying idea is that, as $t \to \infty$, Z_t will typically provide perfect information about θ, in which case $I^\theta_\infty = \lim_{t\to\infty} I^\theta_t$ can be thought of as the missing information about θ when π describes the initial state of knowledge. Thus the π maximizing I^θ_∞ could reasonably be called "least informative." Unfortunately, it is typically the case that I^θ_∞ is infinite for almost all π, so that this approach also does not work. However, it suggests finding, for each t, the prior π_t which maximizes I^θ_t, and then passing to a limit in t. Using a variational argument it can be shown, under certain conditions, that π_t satisfies

$$\pi_t(\theta) \propto \exp\left\{\int p(z_t|\theta) \log \pi(\theta|z_t) dz_t\right\}. \tag{3.1.3}$$

This equation, reproduced in (2.2.2) for the multiparameter case, is the heart of the reference prior algorithm, and (2.2.4) defines the limit in t.

As observed in Section 2.2, (3.1.3) only defines π_t implicitly. However, as $t \to \infty$, both $p(z_t|\theta)$ and $\pi(\theta|z_t)$ will typically converge to their asymptotic distributions, and (3.1.3) will become an explicit equation. For instance, in the regular case of asymptotic normality of the posterior, it can be shown (cf. Bernardo, 1979, Berger and Bernardo, 1992a) that, for large t, $\pi_t(\theta)$ is approximately proportional to $\sqrt{I(\theta)}$, which is thus the reference prior.

For the case of two parameters, $\theta = (\theta_1, \theta_2)$, with $m = 2$ stages to be used in Section 2.3, the argument proceeds by first determining $\pi_2(\theta_2|\theta_1)$, the conditional reference prior for θ_2 assuming that θ_1 is given. This is done exactly as in the previous univariate argument, and results in the analogue of (2.3.2). The idea is then to use $\pi_2(\theta_2|\theta_1)$ to integrate θ_2 out of the model, leaving a marginal model $p^*(z_t|\theta_1)$, for which a reference prior $\pi(\theta_1)$ can (as $t \to \infty$) be found. The overall reference prior on Θ is then $\pi_1(\theta) \propto \pi_2(\theta_2|\theta_1)\pi(\theta_1)$, which is the analogue of (2.3.3); the expression for $\pi(\theta_1)$ in (2.3.3) follows from another asymptotic argument (cf., Bernardo, 1979, Berger and Bernardo, 1992a).

Extension to more than two groupings and multi–dimensional groupings is straightforward. The result is the algorithm described in Section 2.3.

3.2. *Compact Θ^ℓ and Condition (2.2.5)*

In Berger, Bernardo, and Mendoza (1989) it was shown that, for noncompact Θ, there typically exist priors for which I^θ_t in (3.1.2) is infinite, making useless any attempt to define "least informative prior" directly through I^θ_t. The most direct way to circumvent the problem is to operate on compact Θ^ℓ, passing to the limit as $\Theta^\ell \to \Theta$. The issue, then, is how to choose the Θ^ℓ. Usually the choice does not matter, but sometimes it does (cf., Berger and Bernardo, 1989 and 1992a). And even when the choice does matter, it seems to require quite pathological choices of Θ^ℓ to achieve different results.

Choosing the Θ^ℓ to be natural sets in the original parameterization has always worked well in our experience. Indeed, the way we think of the Θ^ℓ is that there is some large compact set on which we are really noninformative, but we are unable to specify the size of this set. We might, however, be able to specify a shape, Ω, for this set, and would then choose $\Theta^\ell = \ell\Omega \cap \Theta$, where $\ell\Omega$ consists of all points in Ω multiplied by ℓ.

Condition (2.2.5) is a new qualification that we have added to the reference prior method. The motivation for this condition is that the pointwise convergence in (2.2.6), that we had previously used in defining the method, does not necessarily imply convergence in an information sense, which is the basis of the reference prior method. Note that (2.2.5) is precisely convergence in the information measure defined by (3.1.1).

Because this is a new condition in the reference prior method, we present two examples, one in which the condition is satisfied and one in which it is not.

Example 1. Suppose $\mathcal{X} = \Theta = (-\infty, \infty)$ and X given θ is $\mathcal{N}(\theta, 1)$. Define $\Theta^\ell = [-\ell, \ell]$. It is easy to apply the reference prior method here, obtaining

$$\pi_1^\ell(\theta) = \frac{1}{2\ell} \text{ on } \Theta^\ell, \quad \pi_1^\ell(\theta|x) = \frac{f(x|\theta)}{[\Phi(\ell - x) - \Phi(-\ell - x)]} \text{ on } \Theta^\ell,$$
$$\pi(\theta) = 1, \quad \pi(\theta|x) = f(x|\theta),$$

and $p^\ell(x) = \int f(x|\theta)\pi_1^\ell(\theta)d\theta = [\Phi(\ell - x) - \Phi(-\ell - x)]/(2\ell)$, where Φ denotes the standard normal c.d.f. Thus

$$D(\pi_1^\ell(\theta|x),\ \pi(\theta|x)) = \int \pi_1^\ell(\theta|x) \log \frac{\pi_1^\ell(\theta|x)}{\pi(\theta|x)} d\theta = -\log([\Phi(\ell - x) - \Phi(-\ell - x)]),$$

and

$$E_\ell^X D(\pi_1^\ell(\theta|X),\ \pi(\theta|X)) = \int p^\ell(x) D(\pi_1^\ell(\theta|x),\ \pi(\theta|x))dx$$
$$= -\frac{1}{2\ell} \int_{-\infty}^{\infty} [\Phi(\ell - x) - \Phi(-\ell - x)] \log([\Phi(\ell - x) - \Phi(-\ell - x)])dx$$
$$= -\int_1^{\infty} [\Phi(y\ell) - \Phi((y-2)\ell)] \log([\Phi(y\ell) - \Phi((y-2)\ell)])dy,$$

the last step using symmetry and making the transformation $y = (\ell - x)/\ell$. Break this integral into $\int_1^3 + \int_3^\infty$. Since $-v \log v \le e^{-1}$ for $0 \le v \le 1$, the dominated convergence theorem can be applied to the first integral to show that it converges to 0 as $\ell \to \infty$. For the second integral, the inequality

$$1 - \frac{0.5}{v} e^{-\frac{1}{2}v^2} \le \Phi(v) \le 1 - \frac{0.3}{v} e^{-\frac{1}{2}v^2}$$

for large v can be used to prove convergence to 0 as $\ell \to \infty$. Hence Condition 2.2.5 is satisfied.

◁

Example 2. Fraser, Monette, and Ng (1985) considered a discrete problem with $\mathcal{X} = \Theta = \{1, 2, 3, \ldots\}$ and

$$f(x|\theta) = \tfrac{1}{3} \text{ for } x \in \{[\tfrac{\theta}{2}],\ 2\theta,\ 2\theta + 1\},$$

with $[v]$ denoting the integer part of v (and $[\frac{1}{2}]$ separately defined as 1). Note that, when x is observed, θ must lie in $\{[\frac{x}{2}],\ 2x,\ 2x + 1\}$, and that the likelihood function is constant over this set. It is immediate that, if one used the noninformative prior $\pi(\theta) = 1$, then

$$\pi(\theta|x) = \tfrac{1}{3} \text{ for } \theta \in \{[\tfrac{x}{2}],\ 2x,\ 2x + 1\}. \tag{3.2.1}$$

This is a very unsatisfactory answer, as discussed in Fraser, Monette, and Ng (1985). As a simple example of this inadequacy, consider the credible set $C(x) = \{2x,\ 2x + 1\}$, which according to (3.2.1) would have probability 2/3 of containing θ for each x. But it is easy to check that the frequentist coverage probability of $C(x)$, considered as a confidence set, is $P_\theta(C(X) \text{ contains } \theta) = \frac{1}{3}$ for all θ. This is an example of "strong inconsistency" (see Stone (1971) for other examples) and indicates a serious problem with the noninformative prior. For later discussion, it is interesting to note that the noninformative prior $\pi(\theta) = \theta^{-1}$ performs

perfectly satisfactorily here, resulting in posterior probabilities and coverage probabilities that are in essential agreement. The prior of Rissanen (1983) is also fine here.

Now, to apply the reference prior method to this problem one must first choose compact subsets Θ^ℓ. Clearly any such sets will here be finite sets, and it can easily be shown that the $\pi_1^\ell(\theta)$ must be constant on finite sets. If now one attempted to pass to the limit in (2.2.6), the result would be the unsatisfactory $\pi(\theta) = 1$.

This turns out, however, to be a situation in which the limit from (2.2.6) violates (2.2.5). To see this take, for instance, the Θ^ℓ to be $\Theta^\ell = \{1, 2, \ldots, 2\ell\}$. As previously mentioned, $\pi_1^\ell(\theta)$ then becomes uniform on Θ^ℓ, so that (3.2.1) is modified to be

$$\pi_1^\ell(\theta|x) = \begin{cases} \frac{1}{3} \text{ for } \theta \in \{[\frac{x}{2}], 2x, 2x+1\} & \text{if } x < \ell \\ \frac{1}{2} \text{ for } \theta \in \{[\frac{x}{2}], 2x\} & \text{if } x = \ell \\ 1 \text{ for } \theta = [\frac{x}{2}] & \text{if } \ell < x \leq 4\ell + 1 \\ \text{nonexistent} & \text{if } 4\ell + 1 < x. \end{cases}$$

Also, it is easy to see that

$$p^\ell(x) = \sum_{\theta=1}^{\infty} f(x|\theta)\pi_1^\ell(\theta) = \begin{cases} 1/\ell & \text{if } x < \ell \\ 2/(3\ell) & \text{if } x = \ell \\ 1/(3\ell) & \text{if } \ell < x \leq 4\ell + 1 \\ 0 & \text{if } 4\ell + 1 < x. \end{cases}$$

An easy calculation then yields

$$\begin{aligned} E_\ell^X D(\pi_1^\ell(\theta|X),\ \pi(\theta|X)) &= \sum_{x=1}^{\infty} p^\ell(x) \sum_{\theta=1}^{\infty} \pi_1^\ell(\theta|x) \log[\pi_1^\ell(\theta|x)/\pi(\theta|x)] \\ &= \frac{2}{3\ell} \log\left(\frac{3}{2}\right) + \frac{(3\ell+1)}{3\ell} \log(3) \longrightarrow \log(3) \text{ as } \ell \longrightarrow \infty, \end{aligned} \tag{3.2.2}$$

so that (2.2.5) is violated.

At this point, all that can be concluded is that a reference prior, as we have defined it, does not exist. There is a fascinating hint, however, that our approach of approximating by compact sets and passing to a limit in "information divergence" may be too crude in this situation. The hint arises from consideration of priors $\pi(\theta) \propto \theta^{-\alpha}$. Repeating the computation done earlier for $\alpha = 0$ yields the interesting fact that the analogue of (3.2.2) does not converge to 0 for $\alpha < 1$ but does converge to 0 for $\alpha = 1$. This suggests that a more clever truncation or way of looking at the truncated problems would yield $\pi(\theta) \propto \theta^{-1}$ as the reference prior (which, as mentioned earlier, is perfectly satisfactory), but we have been unable to devise such a formulation. ◁

We have concentrated on condition (2.2.5) here because this is the first discussion of it in print. Our feeling, however, is that it would be highly unusual for $\pi(\theta)$, defined by (2.2.6), to lead to a violation of (2.2.5). Hence we hesitate to recommend routine verification of the condition, unless there is reason to suspect some pathology.

As one final comment, the need to use (2.2.5) rather than (2.2.6) to define a limit in ℓ suggests that an analogous condition might be needed to replace the pointwise limit in t in (2.2.4). As we have no examples of the necessity of such, however, we have stayed with the simple (2.2.4).

3.3. *Parameters of Interest and Stepwise Computation*

As mentioned in Section 1.2, the separation of θ into parameters of interest and nuisance parameters has been a cornerstone of the reference prior method. In the notation of Sections 2.1 and 2.2, θ would be divided into $m = 2$ groups, with $\theta_{(1)}$ being the parameters of interest and $\theta_{(2)}$ being the nuisance parameters. We begin the discussion of this with a historical example, that will subsequently be put to a new use.

Example 3. Neyman and Scott (1948) introduced an example that has since become a standard test for all new methods of inference. The model consists of $2n$ independent observations,

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2),\ i = 1, \ldots, n,\ j = 1, 2.$$

Reduction to sufficient statistics $X = (\overline{X}_1, \ldots, \overline{X}_n, S^2)$, where $\overline{X}_i = (X_{i1} + X_{i2})/2$ and $S^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij} - \overline{X}_i)^2$, and use of the prior $\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-\alpha}$ yields

$$\pi(\mu_1, \ldots, \mu_n, \sigma | x) \propto \frac{1}{\sigma^{(2n+\alpha)}} \exp\left\{ -\frac{1}{2\sigma^2} \left[s^2 + 2 \sum_{i=1}^{n} (\overline{x}_i - \mu_i)^2 \right] \right\}, \tag{3.3.1}$$

for which the posterior mean of σ^2 is $E[\sigma^2 | x] = s^2/(n + \alpha - 3)$.

The original interest in this example, from a noninformative prior perspective, is that the unmodified Jeffreys prior is $\pi(\mu_1, \ldots, \mu_n, \sigma) = \sqrt{\det I} \propto \sigma^{-(n+1)}$, leading to a posterior mean for σ^2 of $E[\sigma^2 | x] = s^2/(2n - 2)$. This would be inconsistent as $n \to \infty$, since it can be shown that $S^2/n \to \sigma^2$ with probability one (frequentist) so that $S^2/(2n - 2) \to \sigma^2/2$.

Bernardo (1979) and Jeffreys (for related problems) overcame this difficulty by separately dealing with $\theta_{(1)} = \sigma$ and $\theta_{(2)} = (\mu_1, \ldots, \mu_n)$. To apply the reference prior algorithm to these two groups, compute $I(\theta)$ and write it as

$$I(\theta) = \begin{pmatrix} I_{11}(\theta) & 0 \\ 0 & I^*(\theta) \end{pmatrix}, \tag{3.3.2}$$

where $I_{11}(\theta) = 8n/\sigma^2$ and $I^*(\theta) = (2/\sigma^2) I_{(n-1)\times(n-1)}$. Computation yields $|h_1(\theta)| = 8n/\sigma^2$ and $|h_2(\theta)| = 2^n/\sigma^{2n}$, so that condition (2.3.5) is satisfied. Choosing $\Theta^\ell = (\ell^{-1}, \ell) \times (-\ell, \ell) \times \cdots \times (-\ell, \ell)$ (virtually any choice would give the same answer here), (2.3.6) can thus be used to yield, on Θ^ℓ,

$$\pi^\ell(\theta) = \frac{\sqrt{8n/\sigma^2}}{\int_{\ell^{-1}}^{\ell} \sqrt{8n/\sigma^2}\, d\sigma} \cdot \frac{\sqrt{2n/\sigma^{2n}}}{\int_{-\ell}^{\ell} \cdots \int_{-\ell}^{\ell} \sqrt{2n/\sigma^{2n}}\, d\mu_1 \cdots d\mu_n} = k_\ell/\sigma, \tag{3.3.3}$$

where k_ℓ is a constant. Finally, applying (2.2.6) (verification of (2.2.5) is rather tedious here), yields $\pi(\theta) = 1/\sigma$.

This reference prior is perfectly satisfactory, yielding a posterior for which the posterior mean is the very sensible $s^2/(n - 2)$. Thus if σ^2 (or σ) is the parameter of interest with $(\mu_1, \ldots, \mu_n)$ being nuisance parameters, all is well with the reference prior algorithm.

Unfortunately, this simple method of grouping does not always work. Suppose, for instance, that $\theta_{(1)} = \mu_1$ and $\theta_{(2)} = (\mu_2, \ldots, \mu_n, \sigma)$, i.e., that μ_1 is the parameter of interest with the rest being nuisance parameters. Now, $I(\theta)$ is as in (3.3.2) but with $I_{11}(\theta) = 2/\sigma^2$ and $I^*(\theta) = \text{diag}\{\frac{2}{\sigma^2}, \ldots, \frac{2}{\sigma^2}, \frac{8n}{\sigma^2}\}$. Thus $h_1(\theta) = 2/\sigma^2$ and $h_2(\theta) = n2^{(n+2)}/\sigma^{2n}$. Define

$\Theta^\ell = (-\ell, \ell) \times \Theta^*$, where $\Theta^* = (-\ell, \ell) \times \ldots \times (-\ell, \ell) \times (\ell^{-1}, \ell)$. The start of the iteration for the reference prior yields (see (2.3.2))

$$\pi_2^\ell(\theta_{(2)}|\theta_{(1)}) = \frac{\sqrt{n2^{(n+2)}/\sigma^{2n}}}{\int_{\ell^{-1}}^{\ell}\int_{-\ell}^{\ell}\cdots\int_{-\ell}^{\ell}\sqrt{n2^{(n+2)}/\sigma^{2n}}d\mu_2\ldots d\mu_n d\sigma} 1_{\Theta^*}(\theta_{(2)}) = \frac{k_\ell}{\sigma^n}1_{\Theta^*}(\theta_{(2)}).$$

Since $h_1(\theta)$ does not depend on $\theta_{(1)} = \mu_1$, it is easy to see that (2.3.3) becomes $\pi_1^\ell(\theta) = k_\ell\sigma^{-n}1_{\Theta^\ell}(\theta)$. Passing to the limit in ℓ results in the reference prior $\pi(\theta) = 1/\sigma^n$.

For this prior, a standard Bayesian computation yields that the marginal posterior for μ_1 given x is a t–distribution with $(2n-1)$ degrees of freedom, median $\overline{x}_1$, and scale parameter $s^2/[2(2n-1)]$. Thus, for instance, a 95% HPD credible set for μ_1 is

$$C(\overline{x}_1, s) = \left(\overline{x}_1 - t_{(2n-1)}(.975)\frac{s}{\sqrt{2(2n-1)}},\ \overline{x}_1 + t_{(2n-1)}(.975)\frac{s}{\sqrt{2(2n-1)}}\right),$$

where $t_{(2n-1)}(.975)$ is the .975 quantile of a standard t with $(2n-1)$ degrees of freedom.

Now, from a frequentist perspective, it is easy to see that $(\overline{X}_1 - \mu_1)/(S/\sqrt{2n})$ has a standard t–distribution with n degrees of freedom. It follows that $C(\overline{X}_1, S)$ has frequentist coverage probability

$$P_\theta(C(\overline{X}_1, S) \text{ contains } \mu_1) = 2F_n\left(\sqrt{\frac{n}{(2n-1)}}t_{(2n-1)}(.975)\right) - 1,$$

where F_n is the standard t c.d.f. For large n, F_n is approximately the standard normal c.d.f. Φ, and $t_{(2n-1)}(.975) \cong 1.96$, so that $P_\theta(C(\overline{X}_1, S) \text{ contains } \mu_1) \cong 2\Phi\left((1.96)/\sqrt{2}\right) - 1 = 0.83$. This, again, is a strong inconsistency, indicating that the noninformative prior is highly inadequate. It is of interest to note that $\pi(\theta) = 1/\sigma$ would here result in perfect agreement between posterior probability and frequentist coverage. ◁

The above example clearly demonstrates that it is not sufficient to merely divide θ into parameters of interest and nuisance parameters. Once separation of θ into more groups is considered, the natural suggestion is to completely separate the coordinates into k groups of one element each.

Example 3 (continued). If one sets $m = k$, letting each coordinate of θ be a grouping for the reference prior algorithm, it can be checked that $\pi(\theta) = 1/\sigma$ is the resulting reference prior regardless of the ordering of the coordinates of θ. This one–at–a–time reference prior is thus excellent for this problem.

Example 4. In Ye (1990), the development of reference priors for problems in sequential analysis is considered. If N is the stopping time in a sequential problem with independent observations, the Fisher information matrix is $I(\theta) = (E_\theta N)I_1(\theta)$, where $I_1(\theta)$ is the Fisher information for a sample of size one. Then the Jeffreys prior becomes $\pi(\theta) = (E_\theta N)^{k/2}\sqrt{\det(I_1(\theta))}$, which can easily be terrible if k is large because of the presence of $(E_\theta N)^{k/2}$. Grouping and iterating the reference prior method will typically reduce the power of $k/2$, but does not necessarily cure the problem (see Ye, 1990, for examples). But if one uses the one–at–a–time reference prior, then under reasonable conditions (see Ye, 1990) the result is $\pi(\theta) = \sqrt{E_\theta N}\pi^*(\theta)$, where $\pi^*(\theta)$ is the one–at–a–time reference prior for the fixed sample size problem. This is a very reasonable prior. (Of course, use of this method of determining a prior violates the Stopping Rule Principle, but this appears to be one of the unavoidable penalties in use of noninformative priors.) ◁

Other arguments for use of the one–at–a–time reference prior can be found in Berger and Bernardo (1992a and 1992b). Bayarri (1981) gives an example where at least 3 groupings are necessary (and the one–at–a–time reference prior is fine). The bottom line is that we have not yet encountered an example in which the one–at–a–time reference prior is unappealing, and so our pragmatic recommendation is to use this reference prior unless there is a specific reason for using a certain grouping (see Berger and Bernardo, 1992b, for a possible example).

There remains the problem of how to order the parameters before applying the one–at–a–time reference prior algorithm. Currently, we recommend ordering the parameters according to "inferential importance," but beyond putting the "parameters of interest" first, this is too vague to be of much use. Using an average of the reference priors arising from the various acceptable orderings has some appeal, but seems a bit too *ad hoc*. In practice, we have typically computed all one–at–a–time reference priors (and, indeed, all possible reference priors). We have not yet encountered an example in which this could not be done. Having a variety of possible noninformative priors is actually rather useful, since it allows a sensitivity study to choice of the noninformative prior. For additional discussion of this issue see Ghosh and Mukerjee (1992) and the ensuing comments.

3.4. *Other Issues*

3.4.1. Technical Considerations

In computation of the reference prior in the regular case, the two most difficult steps would appear to be evaluation of the expectation E_j^ℓ in (2.3.3) and passing to the limit in (2.2.6). Fortuitously, the latter typically makes the former relatively easy. This is because the expectation in (2.3.3) is with respect to π_{j+1}, which typically is tending towards an improper prior as $\ell \to \infty$. When this happens, it will usually be the case that $E_j^\ell[(\log|h_j(\theta)|)|\theta_{[j]}]$ can be expanded in a Taylors series as

$$K_\ell + C_\ell \psi(\theta) + D_\ell(\theta),$$

where $K_\ell \to \infty$, $C_\ell \to C$, and $D_\ell(\theta) \to 0$ as $\ell \to \infty$. When inserted into (2.3.3), the K_ℓ term typically cancels in the numerator and denominator, and the $D_\ell(\theta)$ term is typically irrelevant (both because of the exponentiation of the E_j^ℓ term). Thus the contribution of the E_j^ℓ term to the final answer will be $\exp\{\frac{1}{2}C\psi(\theta)\}$. Many variants on this theme are possible. What is important is the recognition that (i) exact computation of the E_j^ℓ is typically not needed — computing the first few terms of a Taylors expansion (in ℓ) usually suffices; and (ii) since the expansion is then being exponentiated, all terms except those going to zero (in ℓ) are important.

3.4.2. Prediction and Hierarchical Models

Two classes of problems that are not covered by the reference prior methods so far discussed are hierarchical models and prediction problems. The difficulty with these problems is that there are unknowns (that are indeed even usually the unknowns of interest) that have specified distributions. For instance, if one wants to predict Y based on X when (Y, X) has density $f(y, x|\theta)$, the unknown of interest is Y, but its distribution is conditionally specified. One needs a noninformative prior for θ, not Y. Likewise, in a hierarchical model with, say, $\mu_1, \mu_2, \ldots, \mu_p$ being i.i.d. $\mathcal{N}(\xi, \tau^2)$, the $\{\mu_i\}$ may be the parameters of interest but a noninformative prior is needed only for the hyperparameters ξ and τ^2.

The obvious way to approach such problems is to integrate out the variables with conditionally known distributions (Y in the predictive problem and the $\{\mu_i\}$ in the hierarchical

model), and find the reference prior for the remaining parameters based on this marginal model. The difficulty that arises is how to then identify parameters of interest and nuisance parameters to construct the ordering necessary for applying the reference prior method; the real parameters of interest were integrated out!

We currently deal with this difficulty by defining the parameter of interest in the reduced model to be the conditional mean of the original parameter of interest. Thus, in the prediction problem, $E[Y|\theta]$ (which will be either θ or some transformation thereof) will be the parameter of interest, and in the hierarchical model $E[\mu_i|\xi, \tau^2] = \xi$ will be defined to be the parameter of interest. This technique has worked well in the examples to which it has been applied, but further study is clearly needed.

3.4.3. Invariance

When $\pi(\theta) = \sqrt{\det I(\theta)}$ is the reference prior (typically recommended only for one–dimensional problems), one automatically has invariance with respect to one–to–one transformations of θ, in the sense that the reference prior for a different parameterization would be the correct transform of $\pi(\theta)$. For the iterative reference prior of Section 2.3, certain types of invariance also exist. For instance, in the case of two groupings, $\theta_{(1)}$ and $\theta_{(2)}$, the reference prior is invariant (in the above sense) with respect to choice of the "nuisance parameter" $\theta_{(2)}$, and is also invariant with respect to one–to–one transformations of $\theta_{(1)}$. The reference prior can depend dramatically, however, on which parameters are chosen to be $\theta_{(1)}$. Some results on invariance for more than two groupings are known, but the general issue is still under study.

REFERENCES

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18–31.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53**, 370–418.

Berger, J. (1984). The robust Bayesian viewpoint. *Robustness of Bayesian Analysis* (J. Kadane, ed.), Amsterdam: North-Holland, 63–124.

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

Berger, J. and Bernardo, J. M. (1992a) Ordered group reference priors with application to a multinomial problem. *Biometrika* **79**, (to appear).

Berger, J. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Proceedings of the Indo–USA Workshop on Bayesian Analysis in Statistics and Econometrics* (P. Goel, ed.), New York: Springer, 323–340

Berger, J., Bernardo, J. M. and Mendoza, M. (1989). On priors that maximize expected information. *Recent Developments in Statistics and Their Applications* (J. P. Klein and J. C. Lee, eds.), Seoul: Freedom Academy Publishing, 1–20.

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

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

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

Consonni, G. and Veronese, P. (1989). A note on coherent invariant distributions as non–informative priors for exponential and location–scale families. *Comm. Statist. Theory and Methods* **18**, 2883–2907.

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

Fraser, D. A. S., Monette, G., and Ng, K. W. (1985). Marginalization, likelihood, and structural models. *Multivariate Analysis VI* (P. R. Krishnaiah, ed.), Amsterdam: North-Holland, 209–217.

Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 195–210, (with discussion).

Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333–345.

Jaynes, E. T. (1980). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics* (A. Zellner, ed.), Amsterdam: North-Holland, 43–78.

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

Jeffreys, H. (1961) *Theory of Probability*. Oxford: University Press.

Laplace, P. (1774). Mémoire sur la probabilité des causes par les évenements. *Mem. Acad. R. Sci. Presentés par Divers Savans* **6**, 621–656. (Translated in *Statist. Sci.* **1**, 359–378).

Laplace, P. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Neyman, J. and Scott, B. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

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

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

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. *Sequential Methods in Statistics*, Banach Centre Publications **16**, 485–514.

Stone, M. (1971). Strong inconsistency from uniform priors, with comments. *J. Amer. Statist. Assoc.* **58**, 480–486.

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

Wasserman, L. (1991). An inferential interpretation of default priors. *Tech. Rep.* **516**, Carnegie-Mellon University.

Ye, K. Y. (1990). *Noninformative Priors in Bayesian Analysis*. Ph.D. Thesis, Department of Statistics, Purdue University.

Ye, K. Y. and Berger, J. O. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika* **78**, 645–656.

DISCUSSION

R. McCULLOCH (*University of Chicago, USA*)

My discussion will touch on four points: i) the notion of a parameter of interest; ii) what I think the key idea is; iii) how a subjectivist might use the results; (iv) the role of compact sets.

The notion of a parameter of interest. From the outset the reference prior method has had the notion of a "parameter of interest" and the complementary notion "nuisance parameter" as central concepts. I did my graduate work at Minnesota under Seymour Geisser. If I ever uttered the phrase "paramer of interest" he would beat me up. Geisser smiles a lot and he makes jokes but he's a mean dude. If a parameter has no effect on predictions then by the principle of parsimony it would be eliminated from the model. If it does have an effect on predictions then it can hardly be called a nuisance parameter. Thus, the predictive viewpoint indicates that the need to identify parameters which are a "nuisance" severely restricts the applicability of the method.

I think that it is evident that much of statistical practice has been severely hurt by the nuisance parameter concept. Consider the simple $N(\mu, \sigma^2)$ model. In elementary statistics courses we teach methods for testing hypotheses about the parameter of interest μ without

having to specify σ. But surely the conclusion that μ is close to 0 has quite different implications depending on whether σ is small or large.

What I think the key idea is. The reference prior method has become quite complicated. To me it ends up being something of a black box. The authors explicitly state that with all the asymptotic approximations and limits that the method should be viewed as heuristic and invite us to judge it by how well it works in various examples. The examples are interesting. By that I mean that they excite, in the reader, the psychological and emotional state we label with the word interest. In the product of means example I'm not sure I don't like the uniform prior best. In the Monette, Fraser, Ng example it works only in the sense that it warns us (rightly) that the resultant prior may not be quite right. I have never been happy about the Neyman Scott example because the number of parameters goes to infinity just as fast as the sample size.

And yet the method *does* generate interesting priors. If the method is applied to the entire parameter vector without breaking it up and iterating, the result is Jeffreys' prior. Jeffreys himself found fault with his prior is cases where there was more than one parameter and modified the result in *ad hoc* ways. By breaking up the parameter vector and more or less applying Jeffreys method iteratively the method produces, by a formal mechanism, the kind of results that Jeffreys himself seemed to prefer. This is the key idea in the method. Perhaps it is in one of the earlier papers that I didn't read, but I would like to know if the authors have any intuition for why this seems to work.

How a subjectivist might use the results. The first sentence of the paper is, "Bayesian analysis is a distinct field only because of noninformative priors". My reaction to this sentence was something of a personal epiphany. It revealed something of myself to me. I had thought that I was the kind of guy who was pretty much willing to try anything (an attitude that gets me into trouble in other walks of life). But upon reading the opening line my soul shuddered and then emitted a dark, twisted howl, of which, the only discernible syllable was "no".

Are these results of interest to a subjectivist? I think the answer is yes. Suppose I don't want to bother choosing my prior. Well, I will go ahead and see what I get from a "noninformative" prior. For example I might use the Laplace uniform prior, hopefully after giving some thought to the parametrization. Then, broadly, three things can happen to me. The simplest case is where my posterior distribution is quite tight. In which case I am tempted to conclude that the likelihood dominates the prior and if I bothered to elicit a subjective prior I would just get the same posterior anyway. The second case is where the posteriors is very diffuse. In this case I am tempted to conclude that there isn't much information in the data so that whatever prior I put in will be highly influential in that the posterior will be much like the prior so I might be better off getting more data than eliciting my prior. The third case is anything in between the other two. In this case we would like to check to see how influential the choice of prior is. Ideally we would like to know if the posterior is really any different from that which would be obtained from a carefully elicited prior. To gauge this (without eliciting a prior) we would compare the posterior based on the Laplace prior with that obtained from the reference prior. If the difference between these two posteriors is substantively important you probably can't get away without thinking about your prior.

Compact sets. Limits of compact sets play an important role in the reference prior method. It seems to me that this is one limit that, in our modern computational environment we could avoid taking. In the old days people wanted analytical results and given the set of mathematical tools it was actually more convenient to let the parameter space be infinite.

Now most of our work is done numerically so that, in effect, we are using a compact set. Also, it may be that the choice of a compact subset of the parameter space is something that could be done fairly easily based upon prior information even in high dimensional problems.

Well, all of the above seems like a lot of complaining and whining. We all use "noninformative" priors and this work is probably the most important current work in the area. I found the papers very um...er...ah...interesting. If the authors obtain impossible solutions it is because they are working on an impossible problem. It is comforting to see that Professor Bernardo is keeping the spirit of Don Quixote alive in Spain and I should not be surprised to see the aged Knight, some dark and stormy night, pursuing his quest yet, in the town of West Lafayette.

B. CLARKE (*Purdue University, USA*)

Introduction. Implicit in the work of Berger and Bernardo is a physical interpretation which merits direct examination. They note that in certain examples, the information-theoretic merging of two posteriors may depend on the sequence of compact sets supporting the prior which defines one them. This motivates the definition of reference priors given by expression (2.2.5). Although they have written that such dependence indicates the necessity for subjective input, it can also be given a physical interpretation, in terms of universal noiseless source coding.

In addition, the stepwise prior which appears in expression (2.2.5) can be given a physical interpretation in terms of the capacity of a certain information-theoretic channel. While expression (2.2.5) itself can also be interpreted physically in the context of channel coding, this seems somewhat artificial. Since channel coding and source coding are quite distinct, we raise the question of how to physically interpret the reference prior method.

As this may sound like a criticism, we also argue that the unsatisfactory results obtained in the Fraser-Monette-Ng example, and in the Neyman-Scott example are not a failure of the method, but instead reflect unreasonable expectations.

Some asymptotics will be used and we follow the notation of the paper. For instance, we use $Z_t = (X_1, \ldots, X_t)$, a vector of iid outcomes from $f(\cdot|\theta)$. Henceforth, we only note our occasional necessary departures.

Channel coding and source coding. First consider the function

$$K(t,l) = E_l^{Z_t} D\left(\pi_1^l(\cdot|Z_t), \pi(\cdot|Z_t)\right).$$

The criterion in (2.2.5) is that $\pi(\theta)$ satisfy

$$\lim_{l\to\infty} K(t,l) = 0,$$

so that π is a limit point of the sequence $< \pi_l >|_{l=1}^{\infty}$. The definition of $K(t,l)$ gives

$$K(t,l) = D(\pi_1^l, \pi) + \int \pi_1^l(\theta)\left[D\left(f(Z_t|\theta), p_l(Z_t)\right) - D(f(Z_t|\theta), p(Z_t))\right] d\theta. \qquad (1)$$

When Z_t is discrete, the integrand is essentially the change in redundancy due to using the Shannon code based on p_l, rather than the Shannon code based on p, when the true source is $f(\cdot|\theta)$. Integration over θ gives the Bayes redundancy, and the Shannon code based on a mixture of distributions with respect to a given prior is essentially the code achieving minimal Bayes redundancy, as defined by that prior. Consequently, if the integration over the second term of the integrand were with respect to π, we would say that the sequence

of Bayes codes given by the sequence of mixtures p_l tends to the Bayes code for the entire family of $f(\cdot|\theta)$'s. However, both terms are integrated with respect to π_l which is intended to approximate the limit π.

The first term on the right in (1) represents the redundancy of coding with respect to π when the true prior is π_1^l. However, in the statistical context it is not clear what this means. Perhaps it is sensible to replace (2.2.5) with $D(\pi_1^l, \pi) \to 0$: If π is proper, regularity conditions already imply that for fixed $l, k(t,l) \to 0$ as t increases. The result might be finding rates at which l may be let to increase as a function of t.

Next we turn to a channel coding interpretation for the stepwise reference prior. A conditional density effectively defines a channel. The Shannon mutual information gives, typically, an achievable rate of transmission across the channel. The supremal value of that rate is called the channel capacity. For compact parameter spaces the reference prior is usually the source distribution which gives the channel capacity. In the two step case, there is a formula in Ghosh and Mukerjee (1992) which implies that π_1^l is the source distribution for the channel defined by $m(Z_t|\theta_1) = \int f(Z_t|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2$ which achieves the maximal rate of transmission, asymptotically in t.

This channel has the following interpretation. The message sent is θ_1. There are t receivers, and they pool their data to decode the message. The l defines the range of messages we are able to transmit. The effect of the mixing in $m(Z_t|\theta_1)$ amounts to saying that unbeknownst to the sender, once θ_1 is sent, an auxilliary message θ_2 is sent, with probability $\pi(d\theta_2|\theta_1)$. The decoding is affected in that the constant term in the expansion for the mutual information changes.

Contrasting the two interpretations, we note that adding and subtracting the integral $\int \pi(\theta) D(f(Z_t|\theta), p(Z_t))\, d\theta$ in (1) gives a difference of mutual informations which makes sense in terms of channel coding. However, the other terms are problematic. Also, the l defines an increasing sequence of parametric families in source coding, but a range of messages in channel coding. It is not clear what this means in a statistical context.

Comments on the examples. Finally, we pick a few knits. In the Neyman-Scott example, the two step reference prior approach is sensible when σ is the parameter of interest, but breaks down when μ_1 is the parameter of interest. This is not really surprising since the number of parameters is growing as a linear function of the data, so there is no hope to estimate all of them well. On a technical note, to control error terms in certain proofs, it is essential that the number of parameters grow slowly, if at all.

Regarding the Fraser-Monette-Ng example, it is important to note that the parameter space is discrete. The asymptotics in the discrete case are quite distinct from those in the continuous case: There is no dependence on t or $I(\theta)$. In the absence of nuisance parameters, the mutual information converges to the entropy of the prior. As a result, the reference prior is, asymptotically, the maximum entropy distribution. So, it is not surprising that anomalous results are obtained. Some constraint on the class over which the maximization occurs may be necessary.

In any event, the authors have made a valuable contribution, for which they are to be complimented.

M. GHOSH (*University of Florida, USA*)

The present article is yet another masterly contribution from Berger and Bernardo on the development of reference priors. These authors, over the last few years, have made several important extensions of the original work of Bernardo (1979), where the reference priors were first introduced. One of the major accomplishments of this ongoing research is a

systematic development of reference priors in the presence of nuisance parameters, and the present article is yet another important step in that direction.

I will confine my discussion to the Neyman-Scott example, one of the major examples in this paper, a problem that has fascinated statisticians for more than four decades. I am particularly impressed by the simple reference prior $\pi(\mu_1, \ldots, \mu_n, \sigma) \propto \sigma^{-1}$ which leads to the consistent Bayes estimator $S^2/(n-2)$ of σ^2, consistency being achieved in a frequentist sense.

I now show that an alternative consistent estimator of σ^2 can be derived using a hierarchical Bayes approach, though the proof of consistency requires certain mild conditions on the μ_i's. The derivation proceeds as follows:

First note that $(\overline{X}_1, \ldots, \overline{X}_n, S^2)$ is minimal sufficient, where $\overline{X}_i = (X_{i1} + X_{i2})/2$, $i = 1, \ldots, n$, and $S^2 = \sum_{i=1}^n \sum_{j=1}^2 (X_{ij} - \overline{X}_i)^2$. Consider now the following hierarchical model:

I. Conditional on $\mu = (\mu_1, \ldots, \mu_n), \sigma^2, M = m$, and $\Lambda = \lambda, \overline{X}_1, \ldots, \overline{X}_n$ and S^2 are mutually independent with $\overline{X}_i \sim N(\mu_i, \frac{1}{2}\sigma^2), i = 1, \ldots, n$, and $S^2 \sim \sigma^2 \chi^2_{n'}$

II. Conditional on $M = m, \sigma^2$, and $\Lambda = \lambda, \mu_i$'s are iid $N(m, \lambda^{-1}\sigma^2)$.

III. Marginally M, σ^{-2} and $\Lambda\sigma^{-2}$ are mutually independent with $M \sim$ uniform $(-\infty, \infty)$, $\sigma^{-2} \sim$ Gamma $(0, \frac{1}{2}g_0)$, where $g_0 (\leq 0)$ is some specified number, and $\Lambda\sigma^{-2}$ is Gamma $(0, -1)$, where we use the notation Gamma(α, β) for a (possibly improper) distribution with pdf $f(y) \propto \exp(-\alpha y) y^{\beta - 1}$.

Based on the above hierarchical model, one obtains the following results:

(i) Conditional on $\overline{X}_i = \overline{x}_i (i = 1, \ldots, n), S^2 = s^2$, and $\Lambda = \lambda$,

$$\sigma^{-2} \sim \text{Gamma}\left(\frac{1}{2}\left(s^2 + \frac{2\lambda}{2+\lambda}\sum_{i=1}^n (\overline{x}_i - \overline{\overline{x}})^2\right), \frac{1}{2}(2n - 3 + g_0)\right),$$

where $\overline{\overline{x}} = n^{-1}\sum_{i=1}^n \overline{x}_i$;

(ii) conditional on $\overline{X}_i = \overline{x}_i (i = 1, \ldots, n)$, and $S^2 = s^2$, Λ has conditional pdf

$$f(\lambda | \overline{x}_1, \ldots, \overline{x}_n, s^2) \propto (\lambda/(2+\lambda))^{\frac{1}{2}(n-1)} \lambda^{-2} (s^2 + 2\lambda(2+\lambda)^{-1} \sum_{i=1}^n (\overline{x}_i - \overline{\overline{x}})^2)^{-\frac{1}{2}(2n-3+g_0)}.$$

It is convenient to reparametrize Λ into $U = \Lambda/(2 + \Lambda)$ so that posterior pdf of U is

$$f(u | \overline{x}, \ldots, \overline{x}_n, s^2) \propto u^{\frac{1}{2}(n-5)} (1 + uF)^{-\frac{1}{2}(2n-3+g_0)} I_{[0<u<1]}. \tag{1}$$

Based on (i) and (ii), we obtain

$$E\left(\sigma^{-2} | \overline{x}_1, \ldots, \overline{x}_n, s^2\right) = s^2 (2n - 5 + g_0)^{-1} \left[1 + E(UF | \overline{x}_1, \ldots, \overline{x}_n, s^2)\right] \tag{2}$$

where $F = 2\sum_{i=1}^n (\overline{x}_i - \overline{\overline{x}})^2 / s^2$, a multiple of the usual F statistic. Using (1), it follows after some simplifications that

$$E(UF | \overline{x}_1, \ldots, \overline{x}_n, s^2) = \frac{\int_0^{\frac{F}{(1+F)}} v^{\frac{1}{2}(n-3)} (1-v)^{\frac{1}{2}(n-4+g_0)} dv}{\int_0^{\frac{F}{(1+F)}} v^{\frac{1}{2}(n-5)} (1-v)^{\frac{1}{2}(n-2+g_0)} dv} \tag{3}$$

Integrating by parts, it follows from (3) that

$$E(UF|\overline{x}_1,\ldots,\overline{x}_n,s^2) = (n-3)(n+g_0-2)^{-1} - \frac{2F^{\frac{1}{2}(n-3)}}{(n+g_0-2)(1+F)^{\frac{1}{2}(2n-5+g_0)}\int_0^{F/(1+F)} v^{\frac{1}{2}(n-5)}(1-v)^{\frac{1}{2}(n-2+g_0)}dv} \tag{4}$$

Combining (2) and (4), one gets

$$E\left(\sigma^{-2}|\overline{x}_1,\ldots,\overline{x}_n,s^2\right) = s^2/(n+g_0-2) - \frac{s^2}{2n-5+g_0} \cdot \frac{2F^{\frac{1}{2}(n-3)}}{(n+g_0-2)(1+F)^{\frac{1}{2}(2n-5+g_0)}\int_0^{F/(1+F)} v^{\frac{1}{2}(n-5)}(1-v)^{\frac{1}{2}(n-2+g_0)}dv} \tag{5}$$

As $n \to \infty$, the first term in the right hand side of (5) converges to σ^2 in probability for every fixed g_0. Using some heavy and tedious algebra, it can also be shown that the second term in the right hand side of (5) converges to zero in probability as $n \to \infty$ if $n^{-1}\sum_{i=1}^{n}(\mu_i - \overline{\mu})^2 \to A$ (some fixed positive number) as $n \to \infty$. Thus, the consistency of the Bayes estimate given in (5) holds under some mild conditions. It may be interesting to note that the first term in the right hand side of (5) equals the Berger-Bernardo reference prior estimate when $g_0 = 0$.

Barnard (1970) suggested that the Neyman-Scott problem could be resolved using an empirical Bayes approach. Barnard never did spell out how the empirical Bayes approach should be used, but it seems quite plausible that an empirical Bayes approach will also meet with success if one estimates m and λ rather than use a hyperprior on these parameters.

I wish to thank Professor J. K. Ghosh for posing a question which led to the development of (5).

M. GOLDSTEIN (*University of Durham, UK*)

The subject of reference priors seems to divide Bayesians between those who like and use them, and those who find them rather puzzling. As one of the latter group, I would like to pick up on the link asserted between reference priors and "objectivity", and the related claim that, in some general sense, reference priors "work".

Consider the Fraser-Monette-Ng example dealt with in the paper. The authors seem to imply that a reference-type prior distribution of form $\pi(\theta) \propto \theta^{-1}$ would be "perfectly satisfactory". Now, I can agree that such a prior distribution would not be inherently contradictory. However, what puzzles me is how such a prior could have some claim to objectivity. What is "objective" in placing four times higher probability on the smallest of the three allowable θ values, when we have seen x, than that placed on the other two values? Either this corresponds to a genuine prior judgement that small values are, a priori, more likely than large values, or it looks like an arbitrary fix. The idea that a scientist could "objectively" demonstrate that smaller values of θ were more likely than larger values, without making any subjective inputs, is rather weird, so maybe the authors might like to comment on what they view as the criteria for judging a successful reference prior for this problem.

R. E. KASS (*Carnegie Mellon University, USA*)

I have several comments on this interesting paper.

1. The authors' work is very much in the spirit of Jeffreys, who judged rules for determining prior distributions according to how they worked in specific examples. As a

matter of historical record, however, there is not much support for the claim that "Jeffreys himself ... noticed difficulties with the method, i.e., his general rule, taking the prior to be proportional to $\det[I(\theta)]^{1/2}$ when θ is multidimensional and would then provide ad hoc modifications to the prior". What Jeffreys did was (i) suggest that location parameters should be treated specially, (ii) note that simple alternative solutions exist in many problems (such as taking a uniform prior on the Binomial proportion), which do not agree with those produced by his general rule, and (iii) encourage further investigation of "invariance theory" in determining priors. The defects he noted in his general rule were present in one-parameter problems; application to problems of higher dimension *per se* did not seem to bother him.

2. As far as nomenclature is concerned, I think "noninformative prior" is sufficiently problematic that introducing an alternative is desirable. Since "reference prior" is often understood to refer to Bernardo's method, perhaps a better choice would be "conventional prior". This would clearly be true to Jeffreys's intent in his suggestion that such priors be determined "by international agreement ... as ... in the choice of units of measurement and many other standards of reference" (*Brit. J. Phil. Sci.*, 1955, p. 277).

3. Another matter of nomenclature involves the term "parameters of interest". Apart from location problems, it is by no means clear that parameters may be ordered according to "interest". Why not simply refer to ordered parameters, and let the arbitrariness in the choice remain obvious? (In Jeffreys's scenario, the specifics could be determined by the international committee that will table the results.)

4. In the Neyman-Scott example, it should be remarked that for any fixed n, Jeffreys's method was to take the prior proportional to $1/\sigma$. I do not see how his method is any more "ad hoc" than the authors'. Also, it would seem that a hierarchical prior would be of interest in this example; the authors treated the conjugate case in full detail in their previous work.

5. Hierarchical models present very important cases for any conventional prior methodology and I would hope to see further work in this direction. Computation, however, is likely to be a very serious difficulty. The information matrix is already hard to compute and the brief remarks made in Section 3.4.2 are not specific enough to offer much comfort: the matrix would have to be computed at a large number of values of θ and it is not clear how we would combine that computation with some Gaussian quadrature or simulation method for computing posterior quantities in a reasonably efficient scheme. (By the way, the problem when using asymptotic approximations is greatly reduced because the matrix need only be computed at a few points.) The authors note that their full iterative algorithm may not be feasible in analytically-intractable cases. What, then, are we to do in these commonly-encountered situations?

6. In answering such ultimately practical questions, it is perhaps inappropriate to separate prior selection from sensitivity analysis. In any real problem we will want to perform some kind of sensitivity analysis and we are then led to ask what the role will be for a conventional prior in such an alaysis. This seems to me to be an important outstanding problem. My own experience is dragging me toward subjective sensitivity analysis, but if one were to go to the trouble of performing a sensitivity analysis based on subjectively-determined priors, why would one need a conventional prior? It would seem that the best answer is that it would assist in scientific reporting. On the other hand, it may be possible to construct a method for assessing sensitivity that would be both useful and more convenient than one that requires detailed prior elicitation; perhaps conventional priors could play a role in this process.

G. KOOP (*Boston University, USA*) and
M. F. J. STEEL (*Tilburg University, The Netherlands*)

We congratulate Professors Berger and Bernardo for developing an elegant general approach to reference priors for independent experiments. Our comments are not so much a criticism of their approach as they are a query concerning an extension which we judge to be important. That is, among econometricians there has been a great deal of discussion lately on what constitute reasonable noninformative priors for non-independent experiments. One subject of controversy is the elicitation of noninformative priors for variants of the simple AR(1) model, $y_t = \rho y_{t-1} + \varepsilon_t$ (where the ε_t's are i.i.d. $N(0, \sigma^2)$ and $t = 1, \ldots, T$). Phillips (1991) develops Jeffreys prior for this model, which is often used to test for a unit root ($\rho = 1$). It is our opinion that the development of noninformative priors for dynamic models such as the AR(1) model has great relevance indeed for practitioners of applied econometrics. In the following, we discuss the use of the reference prior in dynamic models and describe the problems that arise in this context.

In the simple AR(1) model, conditional on $y_0 = 0$, there are two parameters, ρ and σ^2. By treating either σ^2 or ρ as the nuisance parameter, the model satisfies (2.3.5) in Berger and Bernardo and hence (2.3.6) holds. The reference prior calculated using Berger and Bernardo's method is the same as the Jeffreys' prior, which possesses tails of $0(\rho^{T-2})$. Although it has all the advantages ascribed to it by Berger and Bernardo, the prior also has several disadvantages. First, the Jeffreys' prior depends on sample size, T, (i.e., is data based) and violates the Likelihood Principle. Second, this dependence on sample size occurs in such a way that the prior influences the posterior even as sample size gets large. That is, the likelihood does not dominate the posterior, even for large samples, which precludes "calibration" (i.e., two Bayesians can continue to disagree as information accrues). Third, econometricians are frequently interested in testing whether y_t is stationary against the hypothesis that it is non-stationary ($|\rho| < 1$ versus $|\rho| \geq 1$). The prior odds for the stationarity hypothesis against a hypothesis containing any finite interval of comparable length in the explosive region are virtually zero. Relative to what econometricians think is reasonable, the Jeffreys' prior places far too much weight on explosive alternatives. This is because the Jeffreys' procedure takes expectations over the sample space. In the AR(1) model, the sampling properties of explosive models dominate those from stationary models. Fourth, the Jeffreys' prior for the AR(1) model depends on the order in which data are observed. Sequential updating is precluded. For all of these reasons, many econometricians consider the Jeffreys' prior to be unreasonable and strongly criticize its indiscriminate use in dynamic models (see the discussion to Phillips (1991)).

On the basis of these objections, we contend that the reference prior approach described in Berger and Bernardo does not extend inmediately to non-independent experiments. On reading the unit root literature in econometrics, we find that a great demand for noninformative priors appears to exist. Hence the development of reference priors that circumvent the above objections might just convince classical econometricians–a very challenging audience indeed–of the merits of Bayesian methods. Perhaps the authors could propose a convincing procedure for such models. Or should researchers just stick with the simple Laplace rule?

D. J. POIRIER (*University of Toronto, Canada*)

The authors readily acknowledge that pursuit of "reference" priors leads to priors that often have a variety of discomforting properties, two of which can include incoherency and violation of the Likelihood Principle. In addition the authors admit that reference priors are often not easy to derive, and are unlikely to achieve broad agreement because they depend on issues such as the parameters of interests. Given these latter pragmatic problems, coupled

with the former distasteful theoretical violations, I think the reader deserves more elaboration on why this pursuit is worthwhile (other than the obvious reply that so many previous authors have gone before). If one is to violate basic principles, then at least the violator should outline the cases in which such violations may be palatable, and if pragmatic expediency is not the reason, what then is the reason?

L. WASSERMAN (*Carnegie Mellon University, USA*)

This is a very interesting paper. I have two comments; both are aimed at promoting more widespread use of the techniques in this paper. First, I think it may be possible to provide a rationale for the sample space dependence of reference priors. The argument goes like this. Let E be the experiment selected by the experimenter. Let I_E be the event that the experimenter preferred E to all other experiments and let π be the experimenter's prior. Suppose I try to guess π. Let J be my prior on the set $\mathcal{P}$ of all priors. It can be shown that, under suitable assumptions, my best guess at π conditional on I_E, is the Jeffreys' prior. In other words, $E_J(\pi|I_E) = \pi_E^*$ where π_E^* is Jeffreys' prior for experiment E. The details are in Wasserman (1991). Thinking of Jeffreys' prior as a guess at π conditional on I_E obviates the criticism that there is a violation of the likelihood principle. It might be possible to justify the stepwise prior in a similar way, by conditioning on the information that the experimenter has chosen a "parameter of interest".

My second comment is a minor point about terminology. As mentioned by the authors, the alernative "default prior" has been suggested in place of "non-informative prior" to refer to priors chosen by scientific convention. Kass (1989) uses this term too. I suggest we abandon the term "non-informative prior" and use "default prior" instead. The former is emotionally charged and, besides, we all agree that there is no such thing as a noninformative prior. Also, the term "reference prior" is ambiguous. Does it refer to (a) priors chosen by scientific convention, (b) priors chosen by the missing information argument or (c) priors chosen by a stepwise argument? To add to the confusion, Box and Tiao (1973) also use "reference prior". I suggest "default prior" for (a), "missing information prior" (MIP) for (b) and "stepwise prior" for (c).

REPLY TO DISCUSSION

We thank the discussants for their interesting comments and questions. Because several of the discussants raised certain common questions, we will respond by topic.

The Name. Kass suggests we replace the name "noninformative" prior with "conventional" prior, while Wasserman prefers "default" prior. Assuming "reference" is to be the name associated with the particular method we advocate for derivation of a noninformative prior, we would slightly prefer the name "conventional" to "default" for the general concept, simply because "default" sounds somewhat unscientific. Basically, however, it is so difficult to change a historical name that we do not advocate such a change. Perhaps when we finally have a true statistical convention to select our official noninformative priors, we should meet in Geneva and then we can call them the "Geneva Convention" priors.

Parameters of Interest. Kass and McCulloch express various concerns about the definition and meaning of "parameters of interest" and "nuisance parameters". We refer to these partly because the historical development of reference priors was heavily influenced by these notions, and partly because the concepts do still seem to provide some guidance in choosing the parameter ordering or the parameters (see Section 3.4.2) to which the reference prior algorithm should be applied. But we must admit that we are drifting away from these concepts; in particular we no longer recommend dividing the parameters exclusively into

these two classes. And we are close to just recommending trying all parameters orderings, regardless of which parameters are of interest.

Dependent Data. The example discussed by Koop and Steel is fascinating, pointing out a serious potential difficulty in information-based methods of deriving noninformative priors. What seems to be happening is that the data can be made more and more informative by having the prior concentrate on larger and larger values of the parameter ρ. Therefore, as the sample size increases, the prior will shift to larger ρ to "increase the information provided by the data". We agree with Koop and Steel that the net effect of this does not seem to be good. Is there a solution within the reference prior theory? We will certainly think about it, but the answer might well be — No!

Why Do All This? This is a very good question, raised to different degrees by Kass, McCulloch, and Poirier. There are actually two distinct questions here. The first is: Has the reference prior method become so involved that we have lost the original motivation for noninformative priors — simplicity? The key to the answer is recognizing that noninformative priors are typically used in a "look-up" scenario, with the practitioner choosing a model and then searching the literature for the "correct" noninformative prior. It will be the job of the reference prior researchers to determine the reference priors for common models, and provide tables of such. The highly sophisticated practitioner who operates by inventing and studying many completely new models will probably find the reference prior algorithm too difficult to employ for each new model, but might well choose to derive it for the model ultimately selected.

The second important question here is: What is the alternative? Let us consider two possibilities:

(i) McCulloch considers use of a constant noninformative prior, lists three possible things that can happen, and suggests that the reference prior is at best useful only for checking if the answer is sensitive to the choice of a constant prior. There are at least two other possibilities that need to be considered, however. The first is that the posterior need not be proper for a constant prior, and impropriety of the posterior may not be easy to recognize in this age of analysis by computer. The exponential regression model referred to in the paper is an example. The other troubling possibility is that the posterior for a constant prior could be quite concentrated, but concentrated in the "wrong" place. The famous Stein example of estimating the squared norm of a multivariate normal mean is one such example; one of us recently even encountered a variant of this problem in a major consulting project. Reference priors are not guaranteed to avoid these difficulties, but their track record is certainly better.

(ii) Subjective Bayesian analysis, perhaps with sensitivity analysis as mentioned by Kass, is the obvious possible alternative. And as the noninformative prior theory grows in complexity, the difficulties in subjective Bayesian analysis start to seem less foreboding. We are not sure, however, if subjective Bayesian analysis is the cure to the difficulties in multi-dimensional problems. The point is simply that subjective elicitation is so difficult in even moderate dimensional problems (hierarchical prior and other structured scenarios excepted) that there is no guarantee that the subjective approach will even be superior to the noninformative prior approach. Only a small number of features of a multivariate prior are ever specified, and the simplifying assumptions that one typically must make (e.g., independence of the parameters) can be extremely influential without one being aware of it. It is nice to say that one will conduct a sensitivity study, but what is the chance, in a high dimensional space, of happening to encounter the truly influential features of the prior? Current research on Bayesian robustness may provide solutions

to this concern, but the jury is still out. The surprising success of the reference prior method on the various known "difficult" multiparameter problems could be viewed as an indication that it somehow seeks and neutralizes potentially serious high-dimensional "confounding" of parameters, but at the moment this is sheer speculation. At the very least, it would be sound practice in a subjective Bayesian sensitivity study to include the reference prior (and probably the constant prior).

Violation of the Likelihood Principle and other Nasties. Several of the discussants express concern over the various foundational inconsistencies that can be encountered with common methods of developing noninformative priors. We have learned to live with these as one of the prices that must be paid. In this regard, we were extremely interested in the statement by Wasserman that, if one adopts a broad enough perspective, the information-based noninformative priors may not be in violation of these principles. His argument sounds plausible, and we await its fleshing out with considerable anticipation.

Koop and Steel discuss a number of unappealing properties of the Jeffreys (and probably reference) prior for the AR(1) model. Two of these properties fall in the category of general problems with noninformative priors, and are thus worth highlighting. First, the fact that the usual noninformative priors do not work well for testing is not too surprising, since noninformative priors rarely work well for testing. Indeed, they only work when there is symmetry between hypotheses or for eliminating nuisance parameters common to the hypotheses. Likewise, it is not uncommon for noninformative priors to be inconsistent with sequential updating, since they often will depend on the amount and even the nature of the data to be obtained. Of course, if the sequential sampling plan is completely known in advance, then one can obtain the noninformative prior for that sequential experiment —see Example 4. We are not saying that these are pleasing properties, but they do seem to be unavoidable.

The Neyman-Scott Example. Kass observes that Jeffreys method (not prior) for location-scale problems was to use $1/\sigma$ as the noninformative prior, and asks — why is this method (which gives the "right" answer for the Neyman-Scott example) more *ad hoc* than the reference prior method? In a sense, Jeffreys method here could be considered to be the stepwise reference prior method, since it is based on somehow attempting to separate the location and the scale parameters. Indeed, our recommended reference prior method could just be viewed as a general way to accomplish separation of parameters.

Kass also observes that a hierarchical model might be natural in this example. We agree, but it is important to note that choosing a hierarchical prior is a major subjective judgement, and is far from being noninformative.

Ghosh provides an interesting analysis showing that use of a hierarchical model here works well, in the sense of providing consistent estimators under weak assumptions. The key feature of his analysis to note is that he does not prove consistency only under the condition that the μ_i arise from the indicated hierarchical prior, but under much weaker assumptions. On a technical point, we suspect that his consistency condition can be weakened to

$$\limsup_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}(\mu_i-\overline{\mu})^2\right]\leq k$$

which is of some interest because it states that the hierarchical prior works even if the μ_i are not arising as i.i.d. observations from a distribution with a finite variance.

Coding. The efforts by Clarke to explain, through notions of coding, the various motivations for the reference prior steps are very interesting. It is perhaps unfortunate that not everything seems to be completely explainable in this regard.

The suggestion that one might consider, instead of (2.2.5), the condition $D(\pi_1^\ell, \pi) \to 0$ is reasonable for proper priors, but for improper π this quantity typically converges to infinity. The suggestion of choosing ℓ to depend on the asymptotic repetition number is an interesting possibility, but in some sense there are already too many options in developing a reference prior; we would recommend adding more options only if the current structure proves to be inadequate.

Why Does the Stepwise Method Work?

McCulloch asks this question. We do not really know the answer. Examination of examples such as the Neyman-Scott example reveals the problem with considering all parameters jointly, but our insight is not much deeper than that. More generally, one of us has never been exactly sure why the entire reference prior method works, and has continually been very pleasantly surprised at its success. At the very least, one must agree with McCulloch's statement "and yet the method **does** generate interesting priors".

Miscellaneous Comments.

(i) The comments of Kass concerning the attitude of Jeffreys towards higher dimensional problems are interesting. To us, the key point is that Jeffreys was at least willing to modify his rule in higher dimensions.

(ii) The computational difficulties mentioned by Kass, especially in regards to determining reference priors for hierarchical models, are very real. Undoubtedly there are problems for which reference priors are not effectively computable, even numerically.

(iii) McCulloch suggests subjectively choosing a large compact set on which to operate, thereby avoiding the need to perform the limiting operation over compact sets. Actually, however, the limiting operation over compact sets is typically a simplifying operation as discussed in section 3.4.1; trying to do the exact computation of a reference prior over a fixed compact set would typically be much more difficult.

ADDITIONAL REFERENCES IN THE DISCUSSION

Barnard, G. A. (1970). Discussion of the paper by Kalbfleisch and Sprott. *J. Roy. Statist. Soc. B* **32**, 194–195.

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

Kass, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188–234, (with discussion).

Phillips, P. C. B. (1991). To criticize the critics: an objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, (with discussion), (to appear).