ON THE DEVELOPMENT OF THE REFERENCE PRIOR METHOD*

by

James O. Berger
Purdue University

and

José–Miguel Bernardo
Presidencia de la Generalitat,
Valencia

Technical Report # 91-15C

Department of Statistics
Purdue University

April 1991

*Research supported by National Science Foundation Grants DMS8702620 and DMS8923071.

⋆

ON THE DEVELOPMENT OF THE REFERENCE PRIOR METHOD*

by

James O. Berger
Purdue University

and

José–Miguel Bernardo
Presidencia de la Generalitat,
Valencia

ABSTRACT

The paper begins with a general, though ideosynchratic, 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 (1984) 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.

*Research supported by National Science Foundation, Grants DMS8702620 and DMS8923071.

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 (1937, 1962) — 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., Cifarelli and Regazzini (1987), Consonni and Veronese (1989), Heath and Sudderth (1978), Hill and Lane (1984), Stone (1979), and Veronese and Consonni (1986).

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), marginalization paradoxes (Dawid, Stone, and Zidek, 1973; but see Jaynes, 1980), strong inconsistency or incoherency (cf. Stone, 1971 and 1979), 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 (1937, 1962) 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 adhoc modifications to the prior.

Bernardo (1979) sought to remove the need for adhoc 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, including Bayarri (1981, 1985), Bernardo (1981, 1982, 1985), Bernardo and Girón (1988), Eaves (1985), Ferrandiz (1982), Mendoza (1987, 1988), and Ye and Berger (1991), 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, 1991a, 1991b), 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 adhoc. 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 (1984) 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 noninformative or 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 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. The methods include those in Akaike (1978), Box and Tiao (1973), Chang and Eaves (1990), George and McCulloch (1989), Geisser (1984), Good (1983), Hartigan (1964, 1983), Jaynes (1968, 1983), Novick and Hall (1965), Rissanen (1983), Rosenkrantz (1977), Villegas (1977, 1981), and Zellner (1977).

It is of interest to briefly discuss one other method for deriving noninformative priors that is currently being intensely studied by a number of statisticians, and which incorporates the distinction between parameters of interest and nuisance parameters. This is the "frequentist coverage of credible sets" method. The idea is to consider $100(1-\alpha)\%$ one–sided Bayesian credible sets for the parameter of interest, arising from use of the noninformative prior, and to compute the (asymptotic) frequentist coverage of the sets. A prior for which this coverage is (asymptotically) $1-\alpha$ for all values of θ is considered to be optimally noninformative. The literature on this approach includes Stein (1965, 1985), Tibshirani (1989), and Ghosh (1991). (See also, Eaton, 1982).

The simplicity and generality of this method are not yet fully clarified. In one–dimensional problems it again yields the Jeffreys prior. In two–dimensional problems, with a parameter of interest θ_1 and nuisance parameter θ_2, it does provide a partial prescription for determining the noninformative prior (cf., Tibshirani (1989) and Ghosh (1991)). First, one must reparameterize so that θ_2 is orthogonal to θ_1 (i.e., so that $I(\theta)$ is a diagonal

matrix). Then the method specifies the noninformative prior to be any prior of the form

$$\pi(\theta_1, \theta_2) = g(\theta_2)\sqrt{I_{11}(\theta)}. \tag{1.3.3}$$

Methods for choosing g and extending this to higher dimensions are still under development. Due to these uncertainties and the considerable difficulty (often near impossibility) of orthogonalizing, the practicality of the method is uncertain, but it seems to work very well when it can be applied, and may have something very interesting to say about the reference prior method (see Section 3.3).

2. The Reference Prior Method

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 (1991a and 1991b).

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 distance" 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

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

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. Expressions for these matrices are given in the Appendix. In particular, $h_1 \equiv H_1 \equiv A_{11}^{-1}$ 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

$$|A| = \text{ determinant of } A, \quad l_\Omega(y) = \begin{cases} 1 & \text{if } y \in \Omega \\ 0 & \text{otherwise,} \end{cases}$$

and will adopt the conventions that $\sum_{i=l}^{l-1}(\cdot) = 0$ and $\prod_{i=l}^{l-1}(\cdot) = 1$.

Step 3 from Section 2.2 can, in the regular case, by replaced by the following, which is essentially taken from Berger and Bernardo (1991b).

Step 3′: To start, define

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

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

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

(Note that it is easy to check, by integrating in turn over $\theta_{(m)}, \theta_{(m-1)}, \ldots, \theta_{(j)}$, that π_j^l defines a probability distribution.)

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

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

Lemma 1. *If (2.3.5) holds, then*

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

Proof. Using (2.3.5) it is clear that

$$E_j^l[\log |h_j(\theta)||\theta_{[j]}] = \log |h_j(\theta)|.$$

The result is immediate from (2.3.3). □

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)

$$I^\theta = E^X D(\pi(\theta|X),\ \pi(\theta)), \tag{3.1.1}$$

where D is the "Kullback–Liebler distance" 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 $I_\infty^\theta = \lim_{t\to\infty} I_t^\theta$ 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_t^θ, 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, $\pi(\theta|z_t)$ can be approximated for large t by a $\mathcal{N}(\theta|\hat\theta_t, \frac{1}{t}S(\hat\theta_t))$ distribution (i.e., a normal distribution with mean equal to $\hat\theta_t$, the m.l.e., and variance $\frac{1}{t}S(\hat\theta_t)$, with S being the inverse of the Fisher information at $\hat\theta_t$), so that (3.1.3) becomes (with hopefully understandable abuse of notation)

$$\begin{aligned}
\pi_t(\theta) &\propto \exp\left\{\int p(z_t|\theta)\log\mathcal{N}(\theta|\hat\theta_t, \tfrac{1}{t}S(\hat\theta_t))dz_t\right\} \\
&= \exp\left\{\int p(\hat\theta_t|\theta)\log\mathcal{N}(\theta|\hat\theta_t, \tfrac{1}{t}S(\hat\theta_t))d\hat\theta_t\right\} \\
&= \exp\left\{\int p(\hat\theta_t|\theta)\left[\log\sqrt{\frac{t}{2\pi}} - \frac{1}{2}\log S(\hat\theta_t) - \frac{t}{2S(\hat\theta_t)}(\hat\theta_t-\theta)^2\right]d\hat\theta_t\right\}.
\end{aligned} \tag{3.1.4}$$

But $p(\hat\theta_t|\theta)$ can be approximated for large t by a $\mathcal{N}(\hat\theta_t|\theta, \frac{1}{t}S(\theta))$ distribution, so that (3.1.4)

becomes

$$\begin{aligned}\pi_t(\theta) &\propto \sqrt{\frac{t}{2\pi}} \exp\left\{\int \mathcal{N}(\hat{\theta}_t|\theta, \frac{1}{t}S(\theta))\left[-\frac{1}{2}\log S(\hat{\theta}_t) - \frac{t}{2S(\hat{\theta}_t)}(\hat{\theta}_t - \theta)^2\right]\right\} d\hat{\theta}_t \\ &\cong \sqrt{\frac{t}{2\pi}} \exp\left\{-\tfrac{1}{2}\log S(\theta)\right\} \exp\{-1\},\end{aligned} \tag{3.1.5}$$

the approximation to the integral over the first term following from the fact that $\mathcal{N}(\hat{\theta}_t|\theta, \frac{1}{t}S(\theta))$ is converging to a point mass at θ. Thus we have that, for large t, $\pi_t(\theta)$ is approximately proportional to

$$\exp\{-\tfrac{1}{2}\log S(\theta)\} = S(\theta)^{-1/2} = \sqrt{I(\theta)},$$

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) still needs to be explained, however. This arises from the same type of asymptotic argument, noting that the asymptotic marginal posterior distribution of θ_1, given z_t, is $\mathcal{N}(\theta_1|\hat{\theta}_1, \frac{1}{t}S_1(\hat{\theta}))$, where $S_1(\theta)$ is the upper left element of $S(\theta)$ and $\hat{\theta}$ is the m.l.e. Then, starting with the analogue of (3.1.3),

$$\begin{aligned}\pi_t(\theta_1) &\propto \exp\left\{\int p^*(z_t|\theta_1)\log\pi(\theta_1|z_t)dz_t\right\} \\ &\cong \exp\left\{\int p^*(z_t|\theta_1)\log\mathcal{N}(\theta_1|\hat{\theta}_1, \tfrac{1}{t}S_1(\hat{\theta}))dz_t\right\} \\ &= \exp\left\{\int p(\hat{\theta}|\theta_1)\log\mathcal{N}(\theta_1|\hat{\theta}_1, \tfrac{1}{t}S_1(\hat{\theta}))d\hat{\theta}\right\} \\ &= \exp\left\{\int\left[\int p(\hat{\theta}|\theta)\pi_2(\theta_2|\theta_1)d\theta_2\right]\log\mathcal{N}(\theta_1|\hat{\theta}_1, \tfrac{1}{t}S_1(\hat{\theta}))d\hat{\theta}\right\} \\ &= \exp\left\{\int \pi_2(\theta_2|\theta_1)\left[\int p(\hat{\theta}|\theta)\log\mathcal{N}(\theta_1|\hat{\theta}_1, \tfrac{1}{t}S_1(\hat{\theta}))d\hat{\theta}\right]d\theta_2\right\} \\ &\propto \exp\{-\tfrac{1}{2}\int \pi_2(\theta_2|\theta_1)\log|S_1(\theta)|d\theta_2\},\end{aligned}$$

the last step following from the same type of argument as did (3.1.5). This is essentially the expression for $\pi(\theta_1)$ in (2.3.3).

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_t^θ in (3.1.2) is infinite, making useless any attempt to define "least informative prior" directly through I_t^θ. 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 1991a). 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

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

and

$$p^\ell(x) = \int f(x|\theta)\pi_1^\ell(\theta)d\theta = \frac{\Phi(\ell - x) - \Phi(-\ell - x)}{2\ell},$$

where Φ denotes the standard normal c.d.f. Thus

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

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 $-v \log v \leq e^{-1}$ for $0 \leq v \leq 1$, 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

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

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 (1984) 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 (1984) and Berger and Wolpert (1988). 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

$$P_\theta(C(X) \text{ contains } \theta) = \tfrac{1}{3} \text{ for all } \theta.$$

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 (see Berger and Wolpert, 1988).

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

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

It follows that

$$D(\pi_1^{\ell}(\theta|x),\ \pi(\theta|x)) = \sum_{\theta=1}^{\infty} \pi_1^{\ell}(\theta|x) \log \frac{\pi_1^{\ell}(\theta|x)}{\pi(\theta|x)} = \begin{cases} \log(1) = 0 & \text{if } x < \ell \\ \log(3/2) & \text{if } x = \ell \\ \log(3) & \text{if } \ell < x \leq 4\ell + 1 \\ 0 & \text{if } 4\ell + 1 < x, \end{cases}$$

and

$$\begin{aligned} E_{\ell}^{X} D(\pi_1^{\ell}(\theta|X),\ \pi(\theta|X) &= \sum_{x=1}^{\infty} p^{\ell}(x) D(\pi_1^{\ell}(\theta|x),\ \pi(\theta|x)) \\ &= \frac{2}{3\ell} \log\left(\frac{3}{2}\right) + \sum_{x=\ell+1}^{4\ell+1} \frac{1}{3\ell} \log(3) \\ &= \frac{2}{3\ell} \log\left(\frac{3}{2}\right) + \frac{(3\ell+1)}{3\ell} \log(3) \\ &\longrightarrow \log(3) \text{ as } \ell \longrightarrow \infty, \end{aligned} \tag{3.2.2}$$

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 distance" 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 $S^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij} - \overline{X}_i)^2$, results in the density

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

k being a numerical constant. Note that $ES^2 = n\sigma^2$. Finally, for the prior

$$\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-\alpha}, \tag{3.3.2}$$

it follows immediately that

$$\pi(\mu_1, \ldots, \mu_n, \sigma|x) \propto \frac{1}{\sigma^{(2n+\alpha)}} \exp\left\{-\frac{1}{2\sigma^2}\left[s^2 + 2\sum_{i=1}^{n}(\overline{x}_i - \mu_i)^2\right]\right\} \tag{3.3.3}$$

and that the posterior mean of σ^2 is

$$E[\sigma^2|x] = s^2/(n + \alpha - 3). \tag{3.3.4}$$

The original interest in this example, from a noninformative prior perspective, is that the unmodified Jeffreys prior is

$$\pi(\mu_1, \ldots, \mu_n, \sigma) = \sqrt{\det I} \propto \sigma^{-(n+1)}, \tag{3.3.5}$$

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 $S^2/n \to \sigma^2$ 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

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

where $I_{11}(\theta) = 8n/\sigma^2$ and

$$I^*(\theta) = \begin{pmatrix} 2/\sigma^2 & & 0 \\ & \ddots & \\ 0 & & 2/\sigma^2 \end{pmatrix}. \tag{3.3.6}$$

Computation yields

$$|h_1(\theta)| = 8n/\sigma^2 \text{ and } |h_2(\theta)| = 2^n/\sigma^{2n}, \tag{3.3.7}$$

so that condition (2.3.5) is satisfied. Choosing

$$\Theta^\ell = (\ell^{-1}, \ell) \times (-\ell, \ell) \times \ldots \times (-\ell, \ell) \tag{3.3.8}$$

(virtually any choice would give the same answer here), Lemma 1 can thus be applied to yield, on Θ^ℓ,

$$\pi^\ell(\theta) = \frac{\sqrt{8n/\sigma^2}}{\int_{\ell^{-1}}^{\ell} \sqrt{8n/\sigma^2} d\sigma} \cdot \frac{\sqrt{2n/\sigma^{2n}}}{\int_{-\ell}^{\ell} \ldots \int_{-\ell}^{\ell} \sqrt{2n/\sigma^{2n}} d\mu_1 \ldots d\mu_n} = k_\ell/\sigma, \tag{3.3.9}$$

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, $I(\theta)$ becomes

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

where $I_{11}(\theta) = 2/\sigma^2$ and

$$I^*(\theta) = \begin{pmatrix} \frac{2}{\sigma^2} & & 0 \\ & \ddots & \\ 0 & & \frac{2}{\sigma^2} \\ & & & \frac{8n}{\sigma^2} \end{pmatrix}. \tag{3.3.10}$$

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))

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

Since $h_1(\theta)$ does not depend on $\theta_{(1)} = \mu_1$, it is easy to see that (2.3.3) becomes

$$\pi_1^\ell(\theta) = \frac{k_\ell}{\sigma^n} 1_{\Theta^\ell}(\theta).$$

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

$$P_\theta(C(\overline{X}_1, S) \text{ contains } \mu_1) \cong 2\Phi\left(\frac{1}{\sqrt{2}}(1.96)\right) - 1 = 0.83.$$

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 (1991a and 1991b). 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, 1991b, 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 adhoc. 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.

The recent advances in the asymptotic frequentist approach to determination of a noninformative prior, discussed in Section 1.3, have further muddled this issue of ordering of the parameters, as the following Lemma shows.

Lemma 2. *Suppose* $\theta^* = (\theta_1^*, \theta_2^*)$, *with* $\Theta^* = \Theta_1^* \times \Theta_2^*$ *and* $I(\theta^*)$ *diagonal. Suppose the reference prior algorithm of Section 2.3 can be applied, and that the choices* $m = 2$, $\theta_{(1)} = \theta_1 = \theta_2^*$ *and* $\theta_{(2)} = \theta_2 = \theta_1^*$, *and* $\Theta^\ell = \Theta_1^\ell \times \Theta_2^\ell$, *where* $\Theta_1^\ell \subset \Theta_2^*$ *and* $\Theta_2^\ell \subset \Theta_1^*$, *are made. Then the reference prior, as defined by (2.2.6), satisfies (1.3.3), assuming the limit exists.*

Proof. Since $I(\theta^*)$ is diagonal, it is easy to see that $h_1(\theta) = I_{11}(\theta) = I_{22}(\theta^*)$ and $h_2(\theta) =$

$I_{22}(\theta) = I_{11}(\theta^*)$. First, (2.3.2) yields

$$\pi_2^\ell(\theta_2|\theta_1) = \frac{\sqrt{I_{22}(\theta)}1_{\Theta_2^\ell}(\theta_2)}{\int_{\Theta_2^\ell}\sqrt{I_{22}(\theta)}d\theta_2} \equiv \frac{\sqrt{I_{22}(\theta)}}{g_1^\ell(\theta_1)}1_{\Theta_2^\ell}(\theta_2).$$

Next, observe that

$$E_1^\ell[(\log(I_{11}(\theta))|\theta_1] = \int_{\Theta_2^\ell}\log(I_{11}(\theta))\pi_2^\ell(\theta_2|\theta_1) \equiv g_2^\ell(\theta_1).$$

Thus (2.3.3) yields

$$\begin{aligned}\pi_1^\ell(\theta) &= \frac{\pi_2^\ell(\theta_2|\theta_1)\exp\{\frac{1}{2}g_2^\ell(\theta_1)\}1_{\Theta_2^\ell}(\theta_2)1_{\Theta_1^\ell}(\theta_1)}{\int_{\Theta_1^\ell}\exp\{\frac{1}{2}g_2^\ell(\theta_1)\}d\theta_1} \\ &= \frac{\exp\{\frac{1}{2}g_2^\ell(\theta_1)\}1_{\Theta_1^\ell}(\theta_1)}{g_1^\ell(\theta_1)\int_{\Theta_1^\ell}\exp\{\frac{1}{2}g_2^\ell(\theta_1)\}d\theta_1}\cdot\sqrt{I_{22}(\theta)}\;1_{\Theta_2^\ell}(\theta_2) \\ &\equiv g^\ell(\theta_1)\cdot\sqrt{I_{22}(\theta)}\;1_{\Theta_2^\ell}(\theta_2).\end{aligned}$$

Passing to the limit in ℓ (which is assumed to exist), yields (with $g^\ell(\theta_1) \to g(\theta_1)$)

$$\pi(\theta) = g(\theta_1)\sqrt{I_{22}(\theta)} = g(\theta_2^*)\sqrt{I_{11}(\theta^*)}, \tag{3.3.11}$$

which is of the form (1.3.3). □

This lemma shows that if θ_1^* is the parameter of interest and θ_2^* is an orthogonal nuisance parameter, and if the Θ^ℓ are chosen to be product sets in this parameterization, then the reverse reference prior (with the nuisance parameter being ordered first) is that which is suggested by the asymptotic frequentist coverage argument.

In Section 1.3 we mentioned the difficulties with proceeding in this fashion: it is unclear what to do in higher dimensions and orthogonalizing the parameters is very hard. Nevertheless, the situation is far from clear, and we cannot categorically state which ordering of parameters is best.

As a final comment on this issue, note that the reverse reference prior method, as given in Lemma 2, is of interest for application of the asymptotic frequentist coverage method of determining a noninformative prior, since it specifies how $g(\theta_2)$ (see 1.3.3) and (3.3.11) should be chosen. This is left unspecified in the frequentist method.

3.4 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.

The computation of the $\{h_i\}$ is greatly simplified by the expressions in the Appendix. In a nonregular case, one has to replace the asymptotic argument (outlined in Section 3.1) that leads to the $\{h_i\}$ with asymptotics appropriate to the model. Note that, at a minimum, this requires knowing the asymptotic posterior distribution for a given model. See Bernardo (1979) for an example of determination of a reference prior in a nonregular case. (Example 2 was, of course, a nonregular case, but because Θ^ℓ was a finite set one could avoid asymptotics and directly compute the π_1^ℓ.)

3.5 Other Issues

3.5.1 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.

3.5.2 Simulation

Various difficulties can be encountered in construction of the reference prior. For instance, $I(\theta)$ might not be computable in closed form. One possibility to overcome this problem would be to compute the reference prior by simulation. This is particularly true for problems in which the actual Bayesian analysis requires computation by simulation in any case. Adding the additional integrations needed to compute, say, $I(\theta)$ is typically only a moderate complication in such problems. This relative ease in computation would extend to cases such as those covered by Lemma 1. If the full iterative reference prior algorithm of Section 2.3 needed to be applied, however, and $I(\theta)$ were not available in closed form, the difficulties would probably be too complex.

3.5.3 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.

APPENDIX

With the notation of Section 2.3, define $B_j = (A_{j1}A_{j2}\ldots A_{jj-1})$, $j = 2,\ldots,m$, of sizes $(n_j \times N_{j-1})$. It is straightforward to verify that, for $j = 1,\ldots,m$

$$h_j = (A_{jj} - B_j H_{j-1} B_j^t)^{-1} \qquad (A1.1)$$

and

$$H_j = \begin{pmatrix} H_{j-1} + H_{j-1}B_j^t h_j B_j H_{j-1} & -H_{j-1}B_j^t h_j \\ -h_j B_j H_{j-1} & h_j \end{pmatrix}, \qquad (A1.2)$$

where any entry containing a factor of H_0 is to be omitted. Thus one may calculate the matrices $H_1,\ldots,H_m$, and hence $h_1,\ldots,h_m$, iteratively.

In the important special case where each $n_j = 1$, no matrix inversions are needed above, so that calculation of the h_j is trivial if S is available. An even greater simplification occurs if, in addition,

$$B_{i+1} = (c_i B_i, \quad A_{i+1\ i}) \qquad (A1.3)$$

for some constant c_i. Then, (A1.1), (A1.2), and (A1.3) can be used to show that

$$h_{i+1} = [A_{i+1\ i+1} + c_i^2 A_{ii} - 2c_i A_{i+1\ i} - h_i(c_i A_{ii} - A_{i+1\ i})^2]^{-1}. \qquad (A1.4)$$

This is particularly useful when (A1.3) holds for all i, which often occurs in patterned covariance matrices, since then (A1.4) can be used to iteratively determine all the h_i, starting with $h_1 = A_{11}^{-1}$, and defining $c_1 = 1$.

REFERENCES

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

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

Bayarri, M.J. (1985). Bayesian inference on the parameters of the Beta distribution. *Statistics and Decisions*, Suppl. Issue **2**, 17–22.

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

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. (1991a). Reference priors in a variance components problem. *Proceedings of the Indo–USA Workshop on Bayesian Analysis in Statistics and Econometrics*, P. Goel (ed.). Bangalore, India (in press).

Berger, J. and Bernardo, J.M. (1991b). Ordered group reference priors with application to a multinomial problem. *Biometrika* (to appear).

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

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle* (2nd edition), Institute of Mathematical Statistics, Hayward, California.

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

Bernardo, J.M. (1981). Reference decisions. *Symposia Matematica* **25**, 85–94.

Bernardo, J.M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J.M. (1985). On a famous problem of induction. *Trab. Estadist.* **36**, 24–30.

Bernardo, J.M. and Girón, F.J. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics* **3**, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (eds.). Oxford University Press, Oxford.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis.* Addison–Wesley, Reading, Massachusetts.

Chang, T. and Eaves, D. (1990). Reference priors for the orbit in a group model. *Ann. Statist.* **4**, 1595–1614.

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

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

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

Eaton, M.L. (1982). A method for evaluating improper prior distributions. In *Statistical Decision Theory and Related Topics III*, S. Gupta and J. Berger (eds.). Springer–Verlag, New York.

Eaves, D.M. (1983). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373–379.

Eaves, D.M. (1985). On maximizing the missing information about a hypothesis. *J. Roy. Statist. Soc. B* **47**, 263–265.

Ferrandiz, J.R. (1982). Una solución Bayesiana a la paradoja de Stein. *Trab. Estadist.* **33**, 31–46.

Fraser, D.A.S., Monette, G., and Ng, K.W. (1984). Marginalization, likelihood, and struc-

tural models. In *Multivariate Analysis VI*, P.R. Krishnaiah (ed.). North–Holland, Amsterdam.

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

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

Ghosh, J.K. (1991). Noninformative priors: A review and new results. To appear in this volume.

Good, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis.

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

Hartigan, J.A. (1983). *Bayes Theory.* Springer–Verlag, New York.

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

Hill, B. and Lane, D. (1984). Conglomerability and countable additivity. In *Bayesian Inference and Decision Techniques with Applications*, P.K. Goel and A. Zellner (eds.). North–Holland, Amsterdam.

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

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

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

Jeffreys, H. (1937, 1961). *Theory of Probability.* Oxford University Press, London.

Laplace, P. (1774). Mémoire sur la probabilité des causes par les évenemens. *Mem. Acad. R. Sci. Presentés par Divers Savans* **6**, 621–656 (translated in *Statistical Science* **1**,

359–378).

Laplace, P. (1812). *Theorie Analytique des Probabilites.* Courcier, Paris.

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

Mendoza, M. (1987). A Bayesian analysis of a generalized slope ratio bioassay. *Probability and Bayesian Statistics*, R. Viertl (ed.). Plenum Press, London.

Mendoza, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression model. *Bayesian Statistics* **3**, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (eds.). Oxford University Press, Oxford.

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

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

Piccinato, L. (1977). Predictive distributions and noninformative priors. In *Transactions of the 7th Prague Conference.* Reidel, Dordrecht.

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

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

Sendra, M. (1982). Distribución final de referencia para el problema de Fieller–Creasy. *Trab. Estadist.* **33**, 55–72.

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

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli–Bayes–Laplace Festschrift*, 217–240. Springer–Verlag, New York.

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

Stone, M. (1979). Review and analysis of some inconsistencies related to improper priors and finite additivity. In *Proceedings of Sixth International Congress on Logic, Methodology and Philosophy of Science*, L.J. Cohen, J. Los, H. Pfeiffer, and K. Podewski (eds.) North–Holland, Amsterdam.

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

Veronese, P. and Consonni, G. (1986). Noninformative priors for exponential families. *Studi Statistici n. 14*. Istituto di Metodi Quantitative, Universitá L. Bocconi, Milano.

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

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

Ye, K.Y. (1990). Noninformative priors in Bayesian analysis. Ph.D. Thesis, Department of Statistics, Purdue University, W. Lafayette.

Ye, K.Y. and Berger, J. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika* (to appear).

Zellner, A. (1977). Maximal data information prior distributions. In *New Methods in the Applications of Bayesian Methods*, A. Aykac and C. Brumat (eds.). North–Holland, Amsterdam.