ON PRIORS THAT MAXIMIZE EXPECTED INFORMATION *

by

James O. Berger
Department of Statistics
Purdue University
West Lafayette, IN, USA

José M. Bernardo
Departamento de Estadística e I.O.
Universidad de Valencia
Valencia, Spain

Manuel Mendoza
Departamento de Mathemáticas
UNAM
Mexico DF, Mexico

Technical Report # 88-5

Department of Statistics
Purdue University

February, 1988
Revised August, 1988

* Research supported by the U.S.-Spain Joint Committee for Scientific and Technological Cooperation, Grant CC B8409-025, and by the National Science Foundation, Grant DMS 8702620.

△

ON PRIORS THAT MAXIMIZE EXPECTED INFORMATION

by

James O. Berger
Purdue University,

José M. Bernardo
Universidad de Valencia,

Manuel Mendoza
U N A M

ABSTRACT

There have been numerous attempts to derive noninformative priors for Bayesian inference via maximization of information measures. The *reference prior* approach of Bernardo (1979) utilizes an entropy-based measure, typically applied to an asymptotic limit of i.i.d. repetitions of the experiment. In attempting to provide a careful mathematical foundation for the reference prior approach, a number of interesting mathematical issues arise. These include (i) the possible nonexistence of maximizing priors; (ii) possible discreteness of the maximizing priors; and (iii) questions concerning the limiting process. Features of the mathematical foundation are described in this paper, along with examples of the above phenomena. These examples are also relevant to other information-based approaches to determination of prior distributions.

1. INTRODUCTION

Information-based approaches to the development of noninformative priors attempt to find "least informative" priors or, alternatively, priors which "maximize the amount of information provided by the data." Pursuing the latter idea, Bernardo (1975) considered, as a measure of the expected information about a parameter $\theta \epsilon \Theta$ provided by an experiment $\mathcal{E}$ when π is the prior distribution on Θ, the quantity

$$I^{\theta}\{\mathcal{E}, \pi\} = E[\log\{p(X|\theta)/p(X)\}], \tag{1.1}$$

where $p(x|\theta)$ is the density (for convenience) of the data X from $\mathcal{E}$ and

$$p(x) = E^{\pi}[p(x|\theta)]$$

(we will use p generically); here "log" stands for the natural logarithm, E^{π} stands for expectation over θ (w.r.t. π), and E stands for expectation over the joint distribution of (X, θ). The prior, π_1, which maximizes $I^{\theta}\{\mathcal{E}, \pi\}$ has a reasonable claim to being noninformative, especially since I^{θ} is perhaps the most natural measure of expected information from an experiment (cf. Lindley (1956) and Shannon (1948)).

A related approach, from Bernardo (1979), considers the "experiment" $\mathcal{E}(k)$ that is defined as k independent replications of $\mathcal{E}$. Letting $Z = (X_1, \ldots, X_k)$ denote the data arising from $\mathcal{E}(k)$, one can consider $I^{\theta}\{\mathcal{E}(k), \pi\}$. The idea is that $\mathcal{E}(\infty)$, defined as the "limit" of $\mathcal{E}(k)$ as $k \to \infty$, typically will provide perfect information about θ, and so $I^{\theta}\{\mathcal{E}(\infty), \pi\}$ can be thought of as the missing information about θ when π describes the initial state of knowledge. Thus the π maximizing $I^{\theta}\{\mathcal{E}(\infty), \pi\}$ could reasonably be called "least-informative." Typically, however, $I^{\theta}\{\mathcal{E}(k), \pi\}$ grows without bound for almost all π, making formal definition of $I^{\theta}\{\mathcal{E}(\infty), \pi\}$ impossible. Therefore, Bernardo (1979) suggested what has become known as the *reference prior* approach. First maximize $I^{\theta}\{\mathcal{E}(k), \pi\}$ for finite k, resulting in what we will call the *k-reference prior*, π_k. Then define the *reference prior*, π^*, providing it exists, by

$$\pi^*(\cdot) = \lim_{k \to \infty} \frac{\pi_k(\cdot)}{\pi_k(A_o)}, \tag{1.2}$$

where A_o is some fixed compact set. Note that π^* will often be improper.

The k-reference priors are themselves of interest. Indeed, we indicated at the beginning that π_1 has a very natural motivation, while the π_k for $k > 1$ are clearly compromises between π_1 and π^*. While we will be discussing the π_k fairly extensively, note that experience has found π^* to generally be the superior "noninformative" prior.

A further generalization that is of considerable interest occurs when π is restricted to some class Γ of distributions. Many of our examples will refer to the *quantile class*, Γ_Q, given when θ is one-dimensional by

$$\Gamma_Q = \{\pi : \pi((b_i, b_{i+1}]) = \gamma_i \ \text{ for } i = 1, \ldots, \ell, \\ \text{where } -\infty \leq b_1 < b_2 < \ldots < b_{\ell+1} \leq \infty \text{ and } \sum_{i=1}^{\ell} \gamma_i = 1\}, \tag{1.3}$$

or to the *moment class*, Γ_M, given by

$$\Gamma_M = \{\pi : E^\pi[\theta^i] = \gamma_i,\ i = 0, \ldots, \ell\}. \tag{1.4}$$

It is rather natural to define the "$\mathcal{E}(k)$-least informative" prior in Γ as that which maximizes $I^\theta\{\mathcal{E}(k), \pi\}$ over all $\pi \epsilon \Gamma$. We will refer to such a prior (if it exists) as a *k-reference prior with respect to* Γ. If these priors have a limit as in (1.2), the limit will be called the *reference prior with respect to* Γ.

The first issue that will be addressed here is that of the existence of k-reference (with respect to Γ) priors. It will be demonstrated in Section 2 that these often do not exist because $I^\theta\{\mathcal{E}(k), \pi\}$ can frequently be infinite. The solution that will be suggested is to consider a nested sequence of compact subsets of Θ, find the desired reference prior when θ is restricted to the subset, and pass to a limit as the sequence enlarges to Θ.

In Section 3, the nature of k-reference priors (with respect to Γ) is discussed. The general variational equation which defines the k-reference prior is given. When $p(x|\theta)$ is from an exponential family, it is argued that the k-reference prior (with respect to Γ) is typically a discrete measure. As a specific example, the Bernoulli trial situation is studied.

Section 4 considers the existence of an overall reference prior with respect to Γ. When Γ is the quantile class, Γ_Q, it is shown that a reference prior does not exist. In contrast, reference priors typically do exist for the moment class, Γ_M.

The examples in this paper illustrate the difficulties in basing derivation of noninformative priors on maximization of versions of I^θ. While rigorous use of I^θ will often be possible, these difficulties suggest that I^θ be mainly used as a heuristic guide in the construction of noninformative priors. Further discussion is given in Section 5. Note that many of the issues and examples in the paper also apply to other information-based methods for development of noninformative priors, such as the methods of Jaynes (1968), Good (1969), Kashyap (1971), and Zellner (1977).

The following notation will be used throughout the paper. The experiment $\mathcal{E}(k)$ consists of observing $Z = (X_1, \ldots, X_k)$ having density $p(z|\theta) = \prod_{i=1}^{k} p(x_i|\theta)$ (recall our generic use of p). For prior distribution π on Θ,

$$p_\pi(z) = E^\pi[p(z|\theta)]. \tag{1.5}$$

(When clear from context, we will omit the subscript π.) Thus

$$\begin{aligned} I^\theta\{\mathcal{E}(k), \pi\} &= E[\log\{p(Z|\theta)/p_\pi(Z)\}] \\ &= E[\log p(Z|\theta)] - E[\log p_\pi(Z)] \\ &= E^\pi[\psi(\theta)] - E^{p_\pi}[\log p_\pi(Z)], \end{aligned} \tag{1.6}$$

where

$$\psi(\theta) = E^{p(z|\theta)}[\log p(Z|\theta)]; \tag{1.7}$$

here $E^{p(z|\theta)}$ denotes expectation over Z with respect to $p(z|\theta)$, and E^{p_π} denotes expectation over Z with respect to p_π.

2 EXISTENCE OF k-REFERENCE PRIORS AND TRUNCATION

2.1 Examples of Existence and Non-existence

In discussion of Bernardo (1979), J. Hartigan observes that $I^\theta\{\mathcal{E}(k),\pi\}$ is often infinite for a variety of π, in which case a unique (and interesting) π_k cannot exist. As an illustration, and to see that this can happen even when π is constrained, consider the following example, related to that of Hartigan.

Example 1. Suppose $X \sim \mathcal{N}(\theta, 1)$ is to be observed, and that $(\ell - 1)$ quantiles of θ are specified; thus $\pi \epsilon \Gamma_Q$, the quantile class specified in (1.3) (with $b_1 = -\infty$ and $b_{\ell+1} = \infty$). Assume w.l.o.g. that $b_2 < 0$. For $\mathcal{E}(k)$, calculation yields

$$E[\log p(Z|\theta)] = -\frac{k}{2}\log(2\pi e),$$

so that

$$I^\theta\{\mathcal{E}(k),\pi\} = -\frac{k}{2}\log(2\pi e) - \int p_\pi(z)\log p_\pi(z)dz. \tag{2.1}$$

Now consider those $\pi\epsilon\Gamma_Q$ which have density π w.r.t Lebesgue measure satisfying $\pi(\theta) \le C$ and (letting $K = \gamma_1 \log|b_2|$)

$$\pi(\theta) = \frac{K}{|\theta|(\log|\theta|)^2} \quad \text{for } \theta \le b_2.$$

(Note that $\int_{-\infty}^{b_2} \pi(\theta)d\theta = \gamma_1$.) Then it is straightforward to show that, for some $B < b_2$,

$$\frac{(1-\epsilon)K}{|z|(\log|z|)^2} < p_\pi(z) < \frac{(1+\epsilon)K}{|z|(\log|z|)^2} < 1 \quad \text{for } z < B.$$

Direct calculation then yields

$$\int_{-\infty}^{B} p_\pi(z)[\log p_\pi(z)]dz < \int_{-\infty}^{B} \frac{(1-\epsilon)K}{|z|(\log|z|)^2}\left[\log\frac{(1+\epsilon)K}{|z|(\log|z|)^2}\right]dz = -\infty,$$

while

$$\int_{B}^{\infty} p_\pi(z)[\log p_\pi(z)]dz < \int_B^\infty p_\pi(z)[\log C]dz < \log C.$$

Thus $I^\theta\{\mathcal{E}(k),\pi\} = \infty$ for any such π. $\square$

It is not always the case that $I^\theta\{\mathcal{E}(k), \pi\}$ is infinite for certain π, as the following lemma shows.

Lemma 1. *Suppose X has discrete or continuous (w.r.t. Lebesgue measure) density $p(x|\theta)$, where $p(x|\theta) \le C_1$, for all x, θ. If $p(x|\theta)$ is non-zero (i) at only a finite number of values of x (discrete case) or (ii) in a subset of $|x| \le C_2$ (continuous case), then $I^\theta\{\mathcal{E}(k), \pi\} < \infty$ for all k and π.*

Proof: Since $-e^{-1} \le y \log y \le \max\{0, C \log C\}$ for $0 \le y \le C$, $p(z|\theta)$ and $p_\pi(z)$ both satisfy

$$-e^{-1} \le p \log p \le \max\{0, C_1^k \log C_1^k\}.$$

The lemma follows easily from writing

$$I^\theta = E^\pi \left[\int p(z|\theta) \log p(z|\theta) dz \right] - \int p_\pi(z) \log p_\pi(z) dz.$$

□

2.2 Truncation of Θ

To deal with situations such as Example 1, a natural idea is to truncate the parameter space. Thus suppose $\{\Theta_m\}$ is a nested sequence of compact sets with $\Theta = \bigcup_{m=1}^{\infty} \Theta_m$. For a class, Γ, of priors, let Γ^m denote the set of all $\pi \in \Gamma$ that are supported in Θ_m.

Definition. $\mathcal{E}_m$ is defined to be the experiment consisting of observing $X \sim p(x|\theta)$ where, however, the parameter space is restricted to Θ_m. When it exists, $\pi_{k,m}$ will denote the k-reference prior (with respect to Γ^m) corresponding to the base experiment $\mathcal{E}_m$. The reference prior (with respect to Γ^m) corresponding to $\mathcal{E}_m$ will be defined (when it exists) by

$$\pi_m^*(\cdot) = \lim_{k \to \infty} \frac{\pi_{k,m}(\cdot)}{\pi_{k,m}(A_o)}, \tag{2.2}$$

where A_o is some fixed compact subset of Θ_1. The reference prior (with respect to Γ) corresponding to the original experiment $\mathcal{E}$ will be defined (when it exists) by

$$\pi^*(\cdot) = \lim_{m \to \infty} \frac{\pi_m^*(\cdot)}{\pi_m^*(A_o)}. \tag{2.3}$$

That truncation can alleviate the difficulty encountered in Example 1 is indicated by the following lemma.

Lemma 2. *If $X \sim \mathcal{N}(\theta, 1)$, then $I^\theta\{\mathcal{E}_m(k), \pi\} < \infty$ for all π.*

Proof. Since Θ_m is compact, it is straightforward to verify that there exists $\epsilon > 0$ and $C < \infty$ such that

$$p_\pi(z) = \int_{\Theta_m} p(z|\theta)\pi(d\theta) \le e^{-\epsilon z^2} < 1 \quad \text{for } |z| \ge C.$$

Then, since $p_\pi(z)$ is continuous and bounded,

$$\int p(z)\log p(z)dz = K + \int\limits_{|z|\geq C} p(z)\log p(z)dz,$$

and

$$\Big|\int\limits_{|z|>C} p(z)\log p(z)dz\Big| \leq 2\int\limits_{|z|>C} \sqrt{p(z)}\left[\sqrt{p(z)}|\log\sqrt{p(z)}|\right] dz$$

$$\leq K_1 \int\limits_{|z|>C} \sqrt{p(z)}dz < \infty.$$

When used in (2.1), these bounds show that $I^\theta < \infty$. □

We suspect that $I^\theta\{\mathcal{E}_m(k), \pi\}$ will typically be finite for all π, as in the situation of Lemma 2, but have made no effort to verify this. When I^θ is finite for all π, then $\pi_{k,m}$ will typically exist, as will π^*_m. Unfortunately, π^* as defined in (2.3) need not exist, as will be seen in Section 4.

3. DERIVATION OF TRUNCATED REFERENCE PRIORS

3.1 The Variational Equation

We consider here the determination of the $\pi_{k,m}$ reference prior with respect to Γ, when Γ is given by

$$\Gamma = \{\pi : E^\pi[g_i(\theta)] = \gamma_i,\ i = 0, \ldots, \ell\}, \tag{3.1}$$

where for convenience we define $g_0(\theta) = 1$ and $\gamma_0 = 1$ so that π is fixed to have mass one. Note that the quantile class (1.3) can be written in this form by defining

$$g_i(\theta) = \mathbf{1}_{(b_i, b_{i+1}]}(\theta),\ i = 1, \ldots, \ell, \tag{3.2}$$

where $\mathbf{1}_A(\theta)$ is the indicator function on the set A, and the moment class (1.4) can be written in this form by defining

$$g_i(\theta) = \theta^i,\ i = 1, \ldots, \ell. \tag{3.3}$$

For convenience in the following, define

$$\Theta_{k,m} = \{\text{support of } \pi_{k,m}\}, \text{ and}$$

$$p_{k,m}(z) = p_{\pi_{k,m}}(z) = E^{\pi_{k,m}}[p(z|\theta)]. \tag{3.4}$$

Theorem 1. *If $\pi_{k,m}$ exists, then with probability one on $\Theta_{k,m}$ (w.r.t. $\pi_{k,m}$)*

$$E^{p(z|\theta)}[\log p_{k,m}(Z)] - \psi(\theta) + \sum_{i=0}^{\ell} \lambda_i g_i(\theta) = 0, \tag{3.5}$$

where the λ_i are constants. (Note that (3.5) implicitely defines $\pi_{k,m}$.)

Proof: This is a standard variational argument. Let

$$\pi = \pi_{k,m} + \epsilon\eta, \tag{3.6}$$

where η is any signed measure for which

$$\frac{|\eta(d\theta)|}{\pi_{k,m}(d\theta)} \leq K < \infty \quad \text{and} \tag{3.7}$$

$$\int\limits_{\Theta_m} g_i(\theta)\eta(d\theta) = 0 \quad \text{for } i = 0, \ldots, \ell. \tag{3.8}$$

Note that $\pi\epsilon\Gamma$ for $|\epsilon| < K^{-1}$.

Consider $I^\theta\{\mathcal{E}_m(k), \pi\}$ for π of the form (3.6), and expand I^θ in a Taylors series about $\pi_{k,m}$. Using (3.7) and (3.8), it can be shown that

$$-[I^\theta\{\mathcal{E}_m(k), \pi\} - I^\theta\{\mathcal{E}_m(k), \pi_{k,m}\}] = \epsilon E^\eta\{-\psi(\theta) + E^{p(z|\theta)}[\log p_{k,m}(Z)]\} + O(\epsilon^2).$$

Since this must be positive for $\pi_{k,m}$ to be the maximizer of I^θ, it follows that

$$E^\eta\{-\psi(\theta) + E^{p(z|\theta)}[\log p_{k,m}(Z)]\} = 0$$

for all η satisfying (3.7) and (3.8). (If not, choosing ϵ to be sufficiently small in absolute value and of the opposite sign of $E^\eta\{\ \}$, achieves a contradiction.) A standard Lagrangian argument then yields the conclusion of the theorem (the Lagrange multipliers coming from (3.8) and the "probability one w.r.t. $\pi_{k,m}$" restriction from (3.7)). □

3.2 Discreteness of $\pi_{k,m}$

When $p(z|\theta)$ is from the exponential family, it will typically be the case that $\psi(\theta)$ (see (1.7)) and

$$\phi(\theta) = E^{p(z|\theta)}[\log p_{k,m}(Z)] \tag{3.9}$$

are analytic functions in each coordinate of θ for

$$\theta\epsilon\Theta_m^* = \text{ convex hull of } \Theta_m.$$

For instance, when $p(x|\theta)$ is $\mathcal{N}(\theta, \Sigma)(\Sigma$ known), then $\psi(\theta)$ is constant and $\phi(\theta)$ is a convolution transform of $\log p_{k,m}(z)$; both ψ and ϕ are thus trivially analytic in each coordinate.

Suppose, in addition, that for $\theta\epsilon\tilde{\Theta}$, a compact subset of $\Theta_m^*, g_i(\theta)$ is analytic in each coordinate of θ for $i = 1, \ldots, \ell$. Then the left hand side of (3.5) is an analytic function of each coordinate of $\theta\epsilon\tilde{\Theta}$. Defining

$$\Theta_{k,m}^* = \{\theta\epsilon\Theta_m^* : \text{ (3.5) is satisfied}\}, \text{ and}$$
$$\tilde{\Theta}_{k,m} = \tilde{\Theta} \cap \Theta_{k,m}^*,$$

it follows that either

$$\begin{aligned} &(i) \quad \tilde{\Theta}_{k,m} \text{ is a finite set; or} \\ &(ii) \quad \tilde{\Theta}_{k,m} = \tilde{\Theta}. \end{aligned} \tag{3.10}$$

(If $\tilde{\Theta}_{k,m}$ is an infinite set in $\tilde{\Theta}$, it has an accumulation point, and an analytic function which is zero on a set with an accumulation point must be constant over its domain of analyticity.)

Typically one can show that (ii) in (3.10) is impossible, so that $\tilde{\Theta}_{k,m}$ must be a finite set. But since the support of $\pi_{k,m}$ in $\tilde{\Theta}$ must be a subset of $\tilde{\Theta}_{k,m}$, it would follow that $\pi_{k,m}$ is then a discrete measure on $\tilde{\Theta}$. As specific examples, consider $\pi_{k,m}$ for the moment and quantile classes, Γ_M and Γ_Q.

Example 2. For Γ_M, defined by (3.1) and (3.3), it is clear that $g_i(\theta) = \theta^i$ is analytic over all of Θ^*_m, for $i = 0, \ldots, \ell$. Hence the above discussion proves that, if $\pi_{k,m}$ exists and $\psi(\theta)$ and $\phi(\theta)$ (see (1.7) and (3.9)) are analytic functions on Θ^*_m, then either

(i) $\pi_{k,m}$ is a finite discrete measure on Θ_m; or

(ii) for all $\theta \epsilon \Theta^*_m$,

$$\phi(\theta) - \psi(\theta) + \sum_{i=0}^{\ell} \lambda_i \theta^i = 0. \tag{3.11}$$

Example 3. Let Γ_Q be the quantile class given by (3.1) and (3.2), and assume that $\psi(\theta)$ and $\phi(\theta)$ (see (1.7) and (3.9)) are analytic functions in each coordinate of θ for $\theta \epsilon \Theta^*_m$. *Claim:* If $\pi_{k,m}$ exists, it is either

(i) a finite discrete measure on Θ_m; or

(ii) it is the unconstrained reference prior and satisfies, for some constant λ,

$$\psi(\theta) = \phi(\theta) + \lambda \quad \text{for all } \theta \epsilon \Theta^*_m. \tag{3.12}$$

Proof of Claim: Define $\tilde{\Theta}_j = \Theta^*_m \cap [b_j, b_{j+1}]$. Clearly each g_i is either the constant function 0 or the constant function 1 on $\tilde{\Theta}_j$, so the left hand side of (3.5) is an analytic function on $\tilde{\Theta}_j$. Defining $\tilde{\Theta}_{k,m,j} = \tilde{\Theta}_j \cap \Theta^*_{k,m}$, it follows from (3.10) that either $\tilde{\Theta}_{k,m,j} = \tilde{\Theta}_j$, or $\tilde{\Theta}_{k,m,j}$ is a finite set. But, if $\tilde{\Theta}_{k,m,j} = \tilde{\Theta}_j$, then $\phi(\theta) - \psi(\theta) = \lambda_j$ on $\tilde{\Theta}_j$ and hence (by analyticity) $\phi(\theta) - \psi(\theta) = \lambda_j$ on all of Θ^*_m. Since this must then be true for each $j = 1, \ldots, \ell$, it would follow that all λ_j must be equal; (3.5) is then the equation defining the unconstrained reference prior. □

It will obviously be rather rare for the unconstrained reference prior to be in Γ, so that (ii) in Example 3 will almost never occur. Also, it is typically possible to show that (3.11) and (3.12) cannot happen (even in the unrestricted case), and discreteness of $\pi_{k,m}$ follows. Here is an example of such an argument.

Example 4. (Bernoulli Experiment). Suppose that $p(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x = 0$ or 1 and $0 \le \theta \le 1$. Then $\mathcal{E}(k)$ consists of observing $Z = (X_1, \ldots, X_k)$ with $p(z|\theta) = \theta^s(1-\theta)^{k-s}, s = \sum_{i=1}^{k} x_i$. (There is no need to truncate $\Theta = [0,1]$ here; see Lemma 1.) Calculation yields

$$\psi(\theta) = E^{p(z|\theta)}[\log p(Z|\theta)] = k[\theta \log \theta + (1-\theta)\log(1-\theta)]$$

and

$$\phi(\theta) = E^{p(z|\theta)}[\log p_k(Z)] = \sum_{i=0}^{k} \alpha_i \theta^i,$$

where the α_i are constants that depend on π_k.

We seek the unconstrained k-reference prior, so that (3.5) is simply

$$\begin{aligned} 0 &= \phi(\theta) - \psi(\theta) + \lambda_0 \\ &= \sum_{i=0}^{k} \alpha_i \theta^i - k[\theta \log \theta + (1-\theta)\log(1-\theta)] + \lambda_0. \end{aligned} \tag{3.13}$$

Observe that

$$\frac{d^{k+1}}{d\theta^{k+1}}[\phi(\theta) - \psi(\theta) + \lambda_0] = k(k-1)!\{(-1)^k \theta^{-k} - (1-\theta)^{-k}\}.$$

If k is odd, this is always negative, from which it follows that (3.13) can have at most $(k+1)$ solutions. If k is even, then $(k+1)$ is odd, and the same reasoning implies that (3.13) can have at most $(k+2)$ solutions. From Theorem 1, it follows that the support of π_k can be at most $(k+1)$ or $(k+2)$ points, as k is odd or even.

Solving for π_k explicitely is not easy. When $k = 1$, it is easy to check that $E^\pi[\psi(\theta)]$ and $-E^{p(z)}[p(Z)]$ are both maximized when $\pi(0) = \pi(1) = \frac{1}{2}$, so that this is clearly the 1-reference prior. Numerical solution is needed for larger k, however. □

3.3 Heuristic Derivation of π_m^*

Assuming that $\pi_{k,m}$ is a nonzero density (typically discrete) on $\Theta_{k,m} \subset \Theta_m$, equation (3.5) can be rewritten (letting $\pi_{k,m}(\theta|z)$ denote the posterior density of θ given z)

$$\begin{aligned} 0 &= -E^{p(z|\theta)}[\log\{p(Z|\theta)/p_{k,m}(Z)\}] + \sum_{i=0}^{\ell} \lambda_i g_i(\theta) \\ &= -E^{p(z|\theta)}[\log\{\pi_{k,m}(\theta|Z)/\pi_{k,m}(\theta)\}] + \sum_{i=0}^{\ell} \lambda_i g_i(\theta) \\ &= -E^{p(z|\theta)}[\log \pi_{k,m}(\theta|Z)] + [\log \pi_{k,m}(\theta)] + \sum_{i=0}^{\ell} \lambda_i g_i(\theta), \end{aligned}$$

so that, on $\Theta_{k,m}$,

$$\pi_{k,m}(\theta) = \exp\{E^{p(z|\theta)}[\log \pi_{k,m}(\theta|Z)] - \sum_{i=0}^{\ell} \lambda_i g_i(\theta)\}. \tag{3.14}$$

Note that this is still just an implicit formula for $\pi_{k,m}(\theta)$, since $\pi_{k,m}(\theta|z)$ depends on $\pi_{k,m}(\theta)$; the expression might be useful for iterative calculation of $\pi_{k,m}(\theta)$, however.

To find π^*_m, it is necessary to let $k \to \infty$ and use (2.2). This is generally very difficult to do explicitely. It is quite plausible, however, that the limit can be obtained by letting $k \to \infty$ in (3.14). The key idea (from Bernardo, 1979) is that, as $k \to \infty, \pi_{k,m}(\theta|z)$ will typically converge to some asymptotic distribution, which can then (hopefully) be inserted into (3.14). Often,

$$\pi_{k,m}(\theta|z) \approx \mathcal{N}(\hat{\theta}(z), k^{-1}I(\hat{\theta}(z))^{-1}) \tag{3.15}$$

for large k, where $\hat{\theta}(z)$ is the m.l.e. and $I(\theta)$ is the Fisher information matrix for $\mathcal{E}$. (Recall that z consists of k independent replications of $\mathcal{E}$.) Inserting this approximation into (3.14) yields (for large k and ignoring multiplicative constants)

$$\begin{aligned}\pi_{k,m}(\theta) &\propto \exp\left\{E^{p(z|\theta)}[\log|I(\hat{\theta})|^{\frac{1}{2}} - \frac{k}{2}(\hat{\theta}-\theta)^t I(\hat{\theta})(\hat{\theta}-\theta)] - \sum_{i=0}^{\ell} \lambda_i g_i(\theta)\right\} \\ &\propto |I(\theta)|^{\frac{1}{2}} \exp\left\{-\sum_{i=0}^{\ell} \lambda_i g_i(\theta)\right\}.\end{aligned}$$

This suggests that

$$\pi^*_m(\theta) = |I(\theta)|^{\frac{1}{2}} \exp\left\{-\sum_{i=0}^{\ell} \lambda_i g_i(\theta)\right\} \mathbf{1}_{\Theta_m}(\theta), \tag{3.16}$$

where the λ_i are chosen so that

$$\int_{\Theta_m} g_i(\theta)\pi^*_m(\theta)d\theta = \gamma_i, \ i = 0, \ldots, \ell. \tag{3.17}$$

<u>Example 5</u>. If Γ is the moment class specified by $g_i(\theta) = \theta^i, i = 0, \ldots, \ell$, then (3.16) becomes

$$\pi^*_m(\theta) = \sqrt{I(\theta)} \exp\left\{-\sum_{i=0}^{\ell} \lambda_i \theta^i\right\} \mathbf{1}_{\Theta_m}(\theta),$$

where the λ_i are chosen so that

$$\int_{\Theta_m} \theta^i \sqrt{I(\theta)} \exp\left\{-\sum_{j=0}^{\ell} \lambda_j \theta^j\right\} d\theta = \gamma_i, \ i = 0, \ldots, \ell.$$

<u>Example 6</u>. If Γ is the quantile class specified by $g_i(\theta) = \mathbf{1}_{(b_i, b_{i+1}]}(\theta), i \geq 1$, and if $(b_2, b_\ell) \subset \Theta_m$, then (3.17) becomes (defining $A_i = \Theta_m \cap (b_i, b_{i+1}]$)

$$\int_{A_i} \sqrt{I(\theta)} e^{-\lambda_i} d\theta = \gamma_i,$$

so that

$$\lambda_i = \log \left\{ \gamma_i^{-1} \int_{A_i} \sqrt{I(\theta)} d\theta \right\}.$$

Hence π_m^* is given explicitely by

$$\pi_m^*(\theta) = \sum_{i=1}^{\ell} \frac{\gamma_i \sqrt{I(\theta)}}{\int_{A_i} \sqrt{I(\theta)} d\theta} \mathbf{1}_{A_i}(\theta). \tag{3.18}$$

Note that $\pi_m^*(\theta)$ is typically discontinuous here. □

Making precise the heuristic derivation of (3.16), when (3.15) holds, is mathematically formidable, since the $\Theta_{k,m}$ are typically finite and nonnested (though they do "increase" to "fill" Θ_m). We do, however, feel that (3.16) is generally correct. Note that $|I(\theta)|^{\frac{1}{2}}$ is the noninformative prior proposed by Jeffreys (1939/67). Also, expressions similar to (3.16) have been derived in many papers, including Jaynes (1968), Good (1969), and Kashyap (1971).

4. EXISTENCE OF π^*

In Section 2, the frequent need for truncation of Θ was introduced. Using the techniques in Section 3, we feel that it will generally be possible to obtain π_m^*, the truncated reference prior (with respect to Γ^m). In Section 2 we also suggested defining the reference prior (with respect to Γ) by

$$\pi^*(\cdot) = \lim_{m\to\infty} \frac{\pi_m^*(\cdot)}{\pi_m^*(A_0)}, \tag{4.1}$$

where A_0 is a fixed compact set in Θ_1. Does this work?

When π is not restricted to a class Γ, or when the restrictions in Γ are strong enough to induce propriety of π^* in (4.1), then we believe that (4.1) generally yields successful results. Consider, for instance, Example 5.

Example 5 (continued). Suppose $\ell = 2$ and

$$\int_{\Theta} (1+\theta^2)\sqrt{I(\theta)} e^{-c\theta^2} d\theta < \infty$$

for $c > 0$. Then it is easy to check that π^* defined by (4.1) exists. Indeed it can be normalized to be a proper distribution given by

$$\pi^*(\theta) = \sqrt{I(\theta)} \exp\{-\lambda_0 + \lambda_1\theta + \lambda_2\theta^2\}\mathbf{1}_{\Theta}(\theta),$$

where

$$\int_{\Theta} \theta^i \sqrt{I(\theta)} \exp\{-\lambda_0 + \lambda_1\theta + \lambda_2\theta^2\} d\theta = \gamma_i, \; i = 0, 1, 2.$$

If, for instance, $p(x|\theta)$ is normal with known variance, then $I(\theta)$ is constant and π^* can be easily verified to itself be a normal density with mean γ_1 and variance $(\gamma_2 - \gamma_1^2)$. □

Unfortunately, even (4.1) is not always successful in defining a sensible π^*, as a version of Example 6 demonstrates.

Example 6 (continued). Suppose that $\sqrt{I(\theta)}$ is not integrable as $\theta \to \pm\infty$. Then it is easy to see that (4.1) yields (ignoring $\pi_m^*(A_0)$ which here has no effect)

$$\begin{aligned}\pi^*(\theta) &= \lim_{m\to\infty} \pi_m^*(\theta) \\ &= \sum_{i=2}^{\ell-1} \frac{\gamma_i \sqrt{I(\theta)}}{\int_{b_i}^{b_{i+1}} \sqrt{I(\theta)} d\theta} \mathbf{1}_{(b_i, b_{i+1}]}(\theta).\end{aligned}$$

(This can be renormalized to have total mass one if desired.) The problem is that the tails of π^* in $(-\infty, b_2]$ and (b_ℓ, ∞) have disappeared, so that π^* is no longer in the specified quantile class. □

In the above example, the failure of the reference prior approach to find a "good" noninformative prior w.r.t. Γ_Q is less a failure of the approach than an indication that the question is ill-posed. If one only specifies quantiles, it is natural that a "noninformative" prior will seek to spread out the tails as much as possible, and this cannot be done while preserving the specified quantiles if the parameter space is unbounded. (Choosing a specific very large m is also not a viable solution; the ensuing statistical answer can depend very strongly on m.) This is related to the essential impossibility of choosing a "noninformative" prior to test between models of different dimensions. For some problems, there simply do not exist "good" noninformative priors.

5. CONCLUSIONS

We have outlined a mathematical framework for determining a "reference" noninformative prior, a framework which we feel would generally be successful when success is possible. The examples indicate, however, that great care must be taken in the mathematics, and in recognizing the possible difficulties that can occur (providing further verification of Hartigan's cautions in Bernardo, 1979). Our current view is that, due to the mathematical complexities, formal attempts to implement the above framework are probably an inefficient use of time. Rather, one should use the framework to heuristically suggest noninformative priors, which can then be studied from a variety of viewpoints for suitability (see, e.g., Bernardo (1979) and Berger and Bernardo (1988)).

ACKNOWLEDGEMENT

The authors are grateful to a referee for helpful suggestions.

REFERENCES

Berger, J. and Bernardo, J.M. (1988). Estimating a product of means: Bayesian analysis with reference priors. To appear in *J. Amer. Statist. Assoc.*

Bernardo, J.M. (1975). Non-informative prior distributions: a subjectivist approach. *Bull. Internat. Statist. Inst. 46*, 94-97.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B41*, 113-147.

Good, I.J. (1969). What is the use of a distribution? In *Multivariate Analysis, Vol. 2* (P.R. Krishnaiah, ed.). Academic Press, New York.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics, SCC-4*, 227-291.

Jeffreys, H. (1939/1967). *Theory of Probability* (3rd ed.). Clarendon Press, Oxford.

Kashyap, R.L. (1971). Prior probability and uncertainty. *IEEE Transactions on Information Theory, IT-17*, 641-650.

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

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

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