Biometrika (1992), **79**, 1, *pp.* 25-37
Printed in Great Britain

Ordered group reference priors with application to the multinomial problem

BY JAMES O. BERGER
Department of Statistics, Purdue University, West Lafayette, Indiana 47907, U.S.A.
AND JOSE M. BERNARDO
Departamento de Estadística, Presidencia de la Generalidad and Universidad de Valencia, Caballeros 2, 46001 Valencia, Spain

SUMMARY

Noninformative priors are developed, using the reference prior approach, for multiparameter problems in which there may be parameters of interest and nuisance parameters. For a given grouping of parameters and ordering of the groups, intuitively, according to inferential importance, an algorithm for determining the associated reference prior is presented. The algorithm is illustrated on the multinomial problem, with discussion of the variety and success of various groupings and ordering strategies.

Some key words: Bayesian inference; Multiparameter problem; Noninformative prior.

1. INTRODUCTION

In the development of noninformative prior distributions, Bernardo (1979) explicitly recognized the importance of identifying the parameters of interest and the nuisance parameters, and tailoring the noninformative prior to this choice; a global noninformative prior distribution, e.g. that of Jeffreys (1961), will not always be adequate for inferences about different parameters within a model. Many of the 'counterexamples' to noninformative priors, e.g. those of Stein (1959) or Dawid, Stone & Zidek (1973), provide dramatic illustrations of this fact.

The reference prior approach of Bernardo (1979) addresses this problem by suggesting a two-step reference prior. First, find the conditional reference prior for the nuisance parameters given the parameters of interest; then find the reference prior for the parameters of interest in the marginal model formed by integrating out the nuisance parameters. This procedure worked well in the examples considered by Bernardo (1979) and in subsequent work such as Bayarri (1981, 1985), Bernardo (1980, 1981, 1982, 1985), Bernardo & Girón (1988), Eaves (1983, 1985), Ferrandiz (1982), Lindley (1988), Mendoza (1987, 1988) or Sendra (1982).

Recently, two limitations of the method have been observed. The first is somewhat technical, but often crucial. The conditional reference prior found in the first step is often improper, and yet is subsequently used to form the marginal model for the parameter of interest. Attempts to justify this step rigorously revealed a rather surprising necessity: one must 'normalize' even improper conditional reference priors. The normalization, and indeed the entire calculation, is done by a limiting operation on proper versions of the problem; Berger & Bernardo (1989, 1992) illustrated this in estimating a product of normal means, and for balanced variance components, respectively.

The second recent observation is that merely grouping the parameters of a model into parameters of interest and nuisance parameters may not go far enough. Allowing multiple groups 'ordered' in terms of importance may be needed, with the reference prior being determined through a succession of analyses for the implied conditional problems. In fact, experience leads us to recommend providing a complete ordering of all parameters of a model, so that the reference prior is determined through a series of one-dimensional conditional steps.

In § 2, we introduce the general m-group reference prior algorithm. Section 3 applies the algorithm to the multinomial problem. Section 4 presents conclusions and discussion.

As background for the developments in the paper, we mention our overall philosophy concerning noninformative priors. This begins with the observation that noninformative priors seem to be popular in applied Bayesian work; even the most avowed subjectivists seem to use heavily noninformative priors, perhaps with profuse apologies, when analyzing data. The second cornerstone of our philosophy is that no one has succeeded, or is ever likely to succeed, in defining unambiguously 'noninformative' priors in an absolute sense. Our goal is the more modest one of developing an algorithm for generation of priors that have a minimal impact on the Bayesian analysis when compared with the impact provided by the data. The concern is that, in higher dimensions, noninformative priors, such as the Jeffreys prior, can have hidden features that have a dramatic, and unrecognized, effect on the answer.

The reference prior approach, especially the new approach discussed herein of development through a series of one-dimensional conditional steps, seems to be remarkably successful in obtaining noninfluential priors in higher dimensions. Evidence of this can be found in the papers mentioned above; Ye & Berger (1991) successfully deal with the notorious exponential regression model, and there are numerous, as yet unpublished, examples. Indeed we have, as yet, found no example in which the recommended reference prior algorithm has led to a 'bad' prior, bad in the sense that the resulting inferences seem undesirable. We certainly do not claim that this reference prior approach is guaranteed to produce results with no undesirable characteristics, but its successes and lack of counterexamples are impressive.

We do not seek in this paper to explore or justify the above comments. Rather, the goal is to present carefully the modified reference prior algorithm, so as to allow study and application by others. Because of this, the example chosen is probably the most perplexing for the theory, namely the multinomial example. For this, the issues of 'parameter of interest' and parameter 'ordering' are particularly relevant, and lead to a recognition that there are numerous possible definitions of 'noninformative'. Anyone seeking to apply the reference prior algorithm should be aware of this ambiguity.

The results for the multinomial problem are of independent interest. For instance, one of the reference priors developed has the appealing property of being consistent with respect to marginalization over 'nuisance' cells, a property not shared by, say, the Jeffreys prior.

2. NOTATION AND THE ALGORITHM

2·1. *Notation*

We consider a parametric statistical problem in which the random observation X has density $p(x \mid \theta)$, where $\theta \in \Theta \subset R^k$ is the unknown parameter. We assume that the Fisher

information matrix

$$H(\theta) = -E_{x|\theta}\left[\left\{\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log p(x\,|\,\theta)\right\}\right]$$

exists and has rank k, so that

$$S(\theta) = H^{-1}(\theta)$$

also exists. Often, we will just write H and S.

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

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

where $N_j = n_1 + \ldots + n_j$ for $j = 1, \ldots, m$. Also we define, for $j = 1, \ldots, m$,

$$\theta_{[j]} = (\theta_{(1)}, \ldots, \theta_{(j)}), \quad \theta_{[\sim j]} = (\theta_{(j+1)}, \ldots, \theta_{(m)}).$$

Finally, we write S as

$$S = \begin{bmatrix} A_{11} & A_{21}^{\mathrm{T}} & \ldots & A_{m1}^{\mathrm{T}} \\ A_{21} & A_{22} & \ldots & A_{m2}^{\mathrm{T}} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mm} \end{bmatrix}$$

so that A_{ij} is $n_i \times n_j$, and define S_j to be the upper left $N_j \times N_j$ corner of S, with $S_m \equiv S$, and $H_j \equiv S_j^{-1}$; then the matrices h_j, defined to be the lower right $n_j \times n_j$ corner of H_j, for $j = 1, \ldots, m$, will be of central importance. Note that $h_1 \equiv H_1 \equiv A_{11}^{-1}$ and, if S is a block diagonal matrix, that is $A_{ij} \equiv 0$ for all $i \neq j$, then $h_j \equiv A_{jj}^{-1}$, for $j = 1, \ldots, m$. Defining $B_j = (A_{j1} \ldots A_{j(j-1)})$, for $j = 2, \ldots, m$, iterative expressions for computing these quantities, in general, are

$$h_j = (A_{jj} - B_j H_{j-1} B_j^{\mathrm{T}})^{-1},$$
$$H_j = \begin{pmatrix} H_{j-1} + H_{j-1}B_j^{\mathrm{T}} h_j B_j H_{j-1} & -H_{j-1}B_j^{\mathrm{T}} h_j \\ -h_j B_j H_{j-1} & h_j \end{pmatrix}, \tag{2·1}$$

where any entry containing a factor of H_0 is to be omitted. In the important special case where each $n_j = 1$, no matrix inversions are needed above. An even greater simplification occurs if, in addition, $B_{i+1} = (c_i B_i, A_{i+1\,i})$ for some constant c_i. Then,

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

Finally, if $\Theta^* \subset \Theta$, we will define

$$\Theta^*(\theta_{[j]}) = \{\theta_{(j+1)} : (\theta_{[j]}, \theta_{(j+1)}, \theta_{[\sim(j+1)]}) \in \Theta^* \text{ for some } \theta_{[\sim(j+1)]}\}; \tag{2·2}$$

we will use the common symbols $|A|$ = determinant of A,

$$1_\Omega(y) = \begin{cases} 1 & (y \in \Omega), \\ 0 & \text{otherwise}, \end{cases}$$

and will throughout the paper adopt the conventions that

$$\sum_{i=l}^{l-1}(.) = 0, \quad \prod_{i=l}^{l-1}(.) = 1.$$

2·2. *The m-group reference prior*

We suppose the θ_i have been ordered and divided into the m groups $\theta_{(1)}, \ldots, \theta_{(m)}$. Note that the ordering within the groups does not matter; see § 2·3 for discussion of the grouping and ordering.

When the reference priors that are developed turn out to be proper, see, e.g. § 3, matters are straightforward. Often, however, they are improper, and care must be taken in their definition. In the improper case we proceed by specifying, see § 2·3 for discussion, a nested sequence $\Theta^1 \subset \Theta^2 \subset \ldots$ of compact subsets of Θ such that

$$\bigcup_{l=1}^{\infty} \Theta^l = \Theta.$$

A reference prior is determined on each compact Θ^l, for which the result is typically a proper prior, followed by performing a limiting operation. Specifically, one follows the following algorithm.

Start. Define

$$\pi_m^l(\theta_{[\sim(m-1)]} \mid \theta_{[m-1]}) = \pi_m^l(\theta_{(m)} \mid \theta_{[m-1]}) = \frac{|h_m(\theta)|^{\frac{1}{2}} 1_{\Theta^l(\theta_{[m-1]})}(\theta_{(m)})}{\int |h_m(\theta)|^{\frac{1}{2}} \, d\theta_{(m)}}, \tag{2·3}$$

where the integral is over the range $\Theta^l(\theta_{[m-1]})$.

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

$$\pi_j^l(\theta_{[\sim(j-1)]} \mid \theta_{[j-1]}) = \frac{\pi_{j+1}^l(\theta_{[\sim j]} \mid \theta_{[j]}) \exp\{\frac{1}{2} E_j^l[(\log|h_j(\theta)|) \mid \theta_{[j]}]\} 1_{\Theta^l(\theta_{[j-1]})}(\theta_{(j)})}{\int \exp\{\frac{1}{2} E_j^l[(\log|h_j(\theta)|) \mid \theta_{[j]}]\} \, d\theta_{(j)}}, \tag{2·4}$$

where the integral is over the range $\Theta^l(\theta_{[j-1]})$ and where

$$E_j^l[g(\theta) \mid \theta_{[j]}) = \int g(\theta) \pi_{j+1}^l(\theta_{[\sim j]} \mid \theta_{[j]}) \, d\theta_{[\sim j]}, \tag{2·5}$$

where the integral is over the range $\{\theta_{[\sim j]} : (\theta_{[j]}, \theta_{[\sim j]} \in \Theta^l\}$. 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. For $j = 1$, interpret $\theta_{[\sim 0]}$ as θ and $\theta_{[0]}$ as vacuous, and write

$$\pi^l(\theta) = \pi_1^l(\theta_{[\sim 0]} \mid \theta_{[0]}). \tag{2·6}$$

Finish. Define the m-group reference prior, assuming it yields a proper posterior, by

$$\pi(\theta) = \lim_{l \to \infty} \frac{\pi^l(\theta)}{\pi^l(\theta^*)}, \tag{2·7}$$

where θ^* is some point in Θ^1.

Note that, if the integrals and expectation in (2·3) and (2·4) are finite when the 'l' is removed, i.e. when Θ^l is replaced by Θ everywhere, then the reference prior is defined simply by π_1, so that (2·7) is not needed.

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

$$|h_j(\theta)| \text{ depends only on } \theta_{[j]} \quad (j = 1, \ldots, m). \tag{2·8}$$

LEMMA 2·1. *If* (2·8) *holds, then*

$$\pi^l(\theta) = \left(\prod_{i=1}^{m} \frac{|h_i(\theta)|^{\frac{1}{2}}}{\int |h_i(\theta)|^{\frac{1}{2}} \, d\theta_{(i)}} \right) 1_{\Theta^l}(\theta), \tag{2·9}$$

where the integral is over the range $\Theta^l(\theta_{[i-1]})$.

Proof. Using (2·8) it is clear that

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

The result is immediate from (2·4). □

2·3. *Motivation and explanation*

Ordering and Grouping. What ordering should be chosen for the θ_i? In nonhierarchical models, as considered here, we suggest the ordering be in terms of the inferential importance of the θ_i. In particular, the parameter, or parameters, of interest should be first. Note that, as argued by Bernardo (1979), a cornerstone of the reference prior approach is that the reference prior may change as one focuses on different parameters, even within the same study.

On the issue of grouping of coordinates, our advice is: do not group without a very good reason. Thus the k-group reference prior, each stage having $n_i = 1$, is generally recommended. At one time (Berger & Bernardo, 1989) we advocated creating two groups, with $\theta_{(1)}$ being the 'parameters of interest' and $\theta_{(2)}$ being the 'nuisance parameters'. Examples of unsuitable performance, to be discussed elsewhere, led us to consider additional groups, eventually leading to the present recommendation. An example in which one might choose to group is discussed in § 3·4. Incidentally, within groups the ordering of the θ_i is immaterial.

Choice of the Θ^l. To reiterate, when the reference priors are proper there is no need to consider compact Θ^l. And even when improper, the reference prior is often unaffected by the particular sequence $\{\Theta^l\}$ chosen.

When needed, our typical choice of the $\{\Theta^l\}$ is simply a collection of nested rectangles in Θ, or other appropriate shape if Θ is not an 'infinite' rectangle. This is based on the heuristic idea that the Θ^l should reflect the type of set on which we would state 'noninformativeness' if we had to choose a compact set, though by choosing a nested infinite sequence we do not commit ourselves to any particular compact, and it is often the case that parameterizations are chosen so that one is 'noninformative' about natural regions, e.g. rectangles, in that parameterization. This is admittedly quite vague and, to be honest, we are unhappy when the choice of $\{\Theta^l\}$ matters. Note that consideration of limits of compacts is also necessary in certain other approaches to development of noninformative priors; Cifarelli & Regazzini (1987) and Consonni & Veronese (1989) are two recent such references.

The motivation for the k-group reference prior algorithm. Bernardo (1979) discussed the motivation for the reference prior approach. The idea is basically to choose the prior which, in a certain asymptotic sense, maximizes the information in the posterior that is provided by the data. We will not repeat the discussion here.

Berger & Bernardo (1989) presented a treatment of the case $k = 2$. The idea is to first find the reference prior for θ_2, at each given value of θ_1, calling this the 'conditional reference prior' $\pi^l(\theta_2|\,\theta_1)$. Assuming asymptotic normality for the model, the argument of Bernardo (1979) leads (Berger & Bernardo, 1989) to

$$\pi^l(\theta_2|\,\theta_1) \propto |h_2(\theta_1, \theta_2)|^{\frac{1}{2}} 1_{\Theta^l(\theta_1)}(\theta_2). \tag{2·10}$$

Since this would subsequently be combined with a 'marginal' reference prior for θ_1, it was realized that normalization would be important. Hence the Θ^l were introduced, in case $|h_2(\theta_1, \theta_2)|^{\frac{1}{2}}$ was not integrable, and $\pi^l(\theta_2|\,\theta_1)$ was actually defined as the normalized

version of (2·10). This is directly analogous to the 'start' of the reference prior algorithm in § 2·2, which gives $\pi^l(\theta_{(m)}|\theta_{(1)}, \ldots, \theta_{(m-1)})$.

Reverting to the two parameter case for simplicity, the natural next step is to form the marginal model for θ_1, by integrating out θ_2 with respect to $\pi^l(\theta_2|\theta_1)$, and then to find the reference prior for θ_1 in this marginal model. This approach unfortunately requires the determination of $H(\theta_1)$ for the convolution of $p(x|\theta_1, \theta_2)$ and $\pi^l(\theta_2|\theta_1)$. Such is frequently not available in closed form, limiting the usefulness of the approach. Thus, we consider $Z = (X_1, \ldots, X_t)$, where the X_i are independently and identically distributed $p(x|\theta_1, \theta_2)$, and derive the marginal model

$$p^l(z|\theta_1) = \int \prod_{i=1}^{T} p(x_i|\theta_1, \theta_2)\pi^l(\theta_2|\theta_1)\, d\theta_2.$$

Applying the reference prior algorithm of Bernardo (1979) results in the 'marginal' reference prior for θ_1

$$\exp\left\{\tfrac{1}{2}\int \pi^l(\theta_2|\theta_1) \log|h_1(\theta_1, \theta_2)|\, d\theta_2\right\}.$$

Multiplying this by the 'conditional' reference prior $\pi^l(\theta_2|\theta_1)$ gives the overall reference prior

$$\pi^l(\theta_1, \theta_2) \propto \pi^l(\theta_2|\theta_1) \exp\left\{\tfrac{1}{2}\int \pi^l(\theta_2|\theta_1) \log|h_1(\theta_1, \theta_2)|\, d\theta_2\right\}.$$

But this is just the numerator in (2·4), when $m = k = 2$ and for $j = 1$. The denominator in (2·4) is just the appropriate normalizing constant. Details of this argument are given by Berger & Bernardo (1992).

Further stages, when $m > 2$, are handled in exactly the same manner yielding (2·4) as the stage-to-stage updating formula. The net result is $\pi^l(\theta)$, the m-stage reference prior on the compact Θ^l. Under reasonable conditions, the reference prior can now be obtained by passing to the limit via

$$\pi(\theta) = \lim_{l\to\infty} \frac{\pi^l(\theta)}{\pi^l(\theta^*)}, \tag{2·11}$$

where $\theta^* \in \Theta^1$ is any fixed point. Note that the main condition is that the posterior obtained from this $\pi(\theta)$ be proper. The above argument is meant to be only heuristic, and does not provide our definition of a grouped reference prior; the definition is given via the algorithm in § 2·2. For a partial indication of the difficulties of making the heuristic argument precise see Berger, Bernardo & Mendoza (1989).

3. THE MULTINOMIAL DISTRIBUTION

3·1. *Preliminaries*

Calculation of m-group reference priors for the multinomial distribution is comparatively simple because all distributions involved turn out to be proper, and the integrations in (2·3) and (2·4) can be done in closed form. First, some preliminary formulae are given; § 3·2 develops the m-group reference prior; § 3·3 investigates properties of the reference prior; § 3·4 is discussion.

We write the multinomial density for $k+1$ cells as

$$p(r_1, \ldots, r_k \mid \theta_1, \ldots, \theta_k) = \left[n! \Big/ \left\{ \left(\prod_{i=1}^{k} r_i! \right)(n-r)! \right\} \right] \left(\prod_{i=1}^{k} \theta_i^{r_i} \right)(1-\delta_k)^{n-r}, \quad (3{\cdot}1)$$

where r_i is the observed frequency in cell i, θ_i is the probability of cell i, n is the total number of observations,

$$r = \sum_{i=1}^{k} r_i, \quad \delta_j = \sum_{i=1}^{j} \theta_i.$$

Note that, in our notation, we will suppress the cell count and probability for the $(k+1)$st cell.

We assume that the θ_i have been ordered and grouped as discussed in § 2·3, see also §§ 3·4 and 4, and we freely use the associated § 2 notation. Calculation yields

$$H(\theta_1, \ldots, \theta_k) = n \operatorname{diag}\{\theta_1^{-1}, \ldots, \theta_k^{-1}\} + n(1-\delta_k)^{-1} 1_k,$$

where diag $\{.\}$ stands for the diagonal matrix with given entries, and 1_k stands for the $k \times k$ matrix of all ones. Further calculation yields

$$S(\theta_1, \ldots, \theta_k) = \frac{1}{n} \operatorname{diag}\{\theta_1, \ldots, \theta_k\} - \frac{1}{n} \theta^{\mathrm{T}} \theta.$$

From this, it is clear that

$$S_j = \frac{1}{n} \operatorname{diag}\{\theta_1, \ldots, \theta_{N_j}\} - \frac{1}{n} \theta_{[j]}^{\mathrm{T}} \theta_{[j]}$$

and, since S_j has the same structure as S, it must be the case that

$$H_j \equiv S_j^{-1} = n \operatorname{diag}\{\theta_1^{-1}, \ldots, \theta_{N_j}^{-1}\} + n(1-\delta_{N_j})^{-1} 1_{N_j}.$$

Furthermore, an easy calculation yields the determinant of the lower right $n_j \times n_j$ corner of H_j,

$$|h_j| = n^{n_j} \left(\prod_{i=N_{j-1}+1}^{N_j} \theta_i^{-1} \right)(1-\delta_{N_{j-1}})(1-\delta_{N_j})^{-1}. \quad (3{\cdot}2)$$

Thus we have available, in closed form, all the quantities needed to apply the reference prior algorithm. For use in the following, define the constants

$$C_{2l-1} = \frac{\pi^l}{(l-1)!}, \quad C_{2l} = \frac{(2\pi)^l}{\{(2l-1)(2l-3)\ldots(1)\}} \quad (3{\cdot}3)$$

for all positive integers l.

3·2. *The multinomial m-group reference prior and posterior*

All distributions that will be encountered have finite mass, so that there is no need to consider a compact sequence $\{\Theta^l\}$. Hence all formulae in § 2·2 will be applied with the 'l' superscripts removed. Note also that, here, (2·2) becomes

$$\Theta(\theta_{[j]}) = \left\{ \theta_{(j+1)}\text{: all elements of } \theta_{(j+1)} \text{ are positive with sum less than } (1-\delta_{N_j}) \right\}.$$

The following lemma provides the crucial calculational development for the reference prior. It can be proved by iteratively integrating over $\theta_{N_j}, \ldots, \theta_{N_{j-1}+1}$.

LEMMA 3·1. *For* $j = 1, \ldots, m$,

$$|h_j(\theta)|^{\frac{1}{2}} \Big/ \left\{ \int_{\Theta(\theta_{[j-1]})} |h_j(\theta)|^{\frac{1}{2}} \, d\theta_{(j)} \right\} = C_{n_j}^{-1} \left(\prod_{i=N_{j-1}+1}^{N_j} \theta_i^{-\frac{1}{2}} \right) (1-\delta_{N_{j-1}})^{(1-n_j)/2} (1-\delta_{N_j})^{-\frac{1}{2}}. \quad (3\cdot4)$$

THEOREM 3·2. *The m-group reference prior is given by*

$$\pi(\theta) = \left(\prod_{i=1}^{m} C_{n_i}^{-1} \right) \left(\prod_{i=1}^{k} \theta_i^{-\frac{1}{2}} \right) \left\{ \prod_{i=1}^{m-1} (1-\delta_{N_i})^{-\frac{1}{2}n_{i+1}} \right\} (1-\delta_{N_m})^{-\frac{1}{2}}. \quad (3\cdot5)$$

The m-group reference posterior is

$$\pi(\theta \mid r_1, \ldots, r_k) \propto \left(\prod_{i=1}^{k} \theta_i^{r_i - \frac{1}{2}} \right) \left\{ \prod_{i=1}^{m-1} (1-\delta_{N_i})^{-\frac{1}{2}n_{i+1}} \right\} (1-\delta_{N_m})^{n-r-\frac{1}{2}}. \quad (3\cdot6)$$

Proof. Since the $|h_j(\theta)|$ satisfy (2·8), Lemma 2·1 yields

$$\pi_j(\theta_{[\sim(j-1)]} \mid \theta_{[j-1]}) = \pi_{j+1}(\theta_{[\sim j]} \mid \theta_{[j]}) |h_j(\theta)|^{\frac{1}{2}} 1_{\Theta(\theta_{[j-1]})}(\theta_{(j)}) \Big/ \left\{ \int_{\Theta(\theta_{[j-1]})} |h_j(\theta)|^{\frac{1}{2}} \, d\theta_{(j)} \right\}.$$

From (2·6), (2·4), and Lemma 3·1 it follows that

$$\pi(\theta) = \left(\prod_{j=1}^{k} \theta_i^{-\frac{1}{2}} \right) \prod_{j=1}^{m} \{ C_{n_j}^{-1} (1-\delta_{N_{j-1}})^{(1-n_j)/2} (1-\delta_{N_j})^{-\frac{1}{2}} \}.$$

Telescoping the product yields (3·5), and (3·6) is immediate from (3·1). □

Two interesting special cases are the 1-group and the k-group reference priors.

Case 1: *The one-group reference prior.* If $m = 1$, (3·5) yields

$$\pi(\theta) = C_k^{-1} \left(\prod_{i=1}^{k} \theta_i^{-\frac{1}{2}} \right) (1-\delta_k)^{-\frac{1}{2}}, \quad (3\cdot7)$$

which is, of course, Jeffreys's noninformative prior.

Case 2. *The k-group reference prior.* If $m = k$, that is all group sizes are $n_i = 1$, then (3·5) yields

$$\pi(\theta) = (\pi^{-k}) \prod_{i=1}^{k} \{ \theta_i^{-\frac{1}{2}} (1-\delta_i)^{-\frac{1}{2}} \}. \quad (3\cdot8)$$

This is actually the reference prior that we will recommend for typical use.

3·3. *Properties of the reference prior and posterior*

Marginal distributions. The marginal probability distribution of $(r_1, \ldots, r_l)$ is also multinomial, with cell probabilities $\theta_1, \ldots, \theta_l$ and sample size n, so that all other observations are lumped together into the new $(l+1)$st cell, which is the union of the $(l+1)$st through $(k+1)$st cells in the original multinomial. It is of considerable interest to see whether or not the m-group reference prior 'marginalizes' consistently, in that the reference prior for the 'collapsed' $(l+1)$-cell multinomial be the same as that obtained by finding the marginal distribution of $\theta_1, \ldots, \theta_l$ from the original m-group reference prior. The following lemma provides the answer. For simplicity we assume, with the

exception of (3·11), that the marginalization is done over groups. The proof of this lemma is tedious but standard.

LEMMA 3·3. *For the prior* $\pi(\theta)$ *in* (3·5), *the marginal reference prior for* $\theta_{(1)}, \ldots, \theta_{(j)}$ *is*

$$\pi_{[j]}(\theta_{(1)}, \ldots, \theta_{(j)}) = \left(\prod_{i=1}^{j} C_{n_i}^{-1}\right)\left(\prod_{i=1}^{N_j} \theta_i^{-\frac{1}{2}}\right)\left\{\prod_{i=1}^{j-1} (1-\delta_{N_i})^{-\frac{1}{2}n_{i+1}}\right\}(1-\delta_{N_j})^{-\frac{1}{2}}. \tag{3.9}$$

The marginal reference prior for $\theta_{(1)}$ *is*

$$\pi_{[1]}(\theta_{(1)}) = C_{n_1}^{-1}\left(\prod_{i=1}^{n_1} \theta_i^{-\frac{1}{2}}\right)(1-\delta_{n_1})^{-\frac{1}{2}}, \tag{3·10}$$

while that for $\theta_1, \ldots, \theta_l$, *when* $l < n_1$, *is*

$$\pi_{[1,l]}(\theta_1, \ldots, \theta_l) = \frac{C_{n_1-l}}{C_{n_1}}\left(\prod_{i=1}^{l} \theta_i^{-\frac{1}{2}}\right)(1-\delta_l)^{(n_1-l-1)/2}. \tag{3·11}$$

Typically, of course, one will be interested in the marginal posteriors, rather than the marginal priors. These are immediate from the marginal priors, however; simply multiply by the likelihood from the corresponding marginal multinomial distribution. For instance, the marginal posterior for $\theta_{(1)}$ is

$$\pi_{[1]}(\theta_{(1)} | r_1, \ldots, r_{n_1}) \propto \left(\prod_{i=1}^{n_1} \theta_i^{r_i-\frac{1}{2}}\right)(1-\delta_{n_1})^{n-\sum r_i-\frac{1}{2}},$$

where the sum is over the range $i = 1, \ldots, n_1$. This could, of course, also have been obtained by calculating the marginal density of $\theta_{(1)}$ from (3·6).

To return to the question posed at the beginning of this section, we see that (3·9) and (3·10) are of exactly the same form as (3·5). Hence, if we reduce consideration to the first j groups of parameters $\theta_{(1)}, \ldots, \theta_{(j)}$, the answers obtained by marginalizing from the original m-group reference prior are identical to the answers obtained by treating the j groups as a 'new' multinomial problem. This property may be viewed to be valuable, because of the following well-known example.

Example. Suppose we have the multinomial model with $\theta_1, \ldots, \theta_{n_1}$. Consider the 1-group reference prior, which is Jeffreys's prior, here

$$\pi(\theta_1, \ldots, \theta_{n_1}) \propto \left(\prod_{i=1}^{n_1} \theta_i^{-\frac{1}{2}}\right)(1-\delta_{n_1})^{-\frac{1}{2}}.$$

The posterior means of the θ_i are then

$$E(\theta_i | r_1, \ldots, r_{n_1}) = \frac{r_i + \frac{1}{2}}{n + \frac{1}{2}(n_1+1)}.$$

Now suppose one notices that the $(n - r_1 - \ldots - r_{n_1})$ observations in the $(n+1)$st cell could have been further subdivided into n_2+1 new categories; say one discovers a new classification scheme into n_2 categories for elements in this cell. Then one has the apparent option of adding $r_{n_1+1}, \ldots, r_{n_1+n_2}$ and $\theta_{n_1+1}, \ldots, \theta_{n_1+n_2}$ to the multinomial model. If one did so and used Jeffreys's prior, which would then be

$$\pi(\theta_1, \ldots, \theta_{n_1+n_2}) \propto \left(\prod_{i=1}^{n_1+n_2} \theta_i^{-\frac{1}{2}}\right)(1-\delta_{n_1+n_2})^{-\frac{1}{2}},$$

a calculation shows that the posterior means of the θ_i would now be

$$E(\theta_i | r_1, \ldots, r_{n_1+n_2}) = \frac{r_i + \frac{1}{2}}{n + \frac{1}{2}(n_1 + n_2 + 1)}.$$

The creation of new cells can thus have a pronounced effect on posterior beliefs about existing cells.

The m-group reference prior is essentially immune to this difficulty, since the marginal prior, and posterior, for, say, $\theta_{(1)}$ is the same no matter how many additional groups, or cells, are added. This needs two qualifications, however. The first is that the marginalization property does not hold for all groups; it holds only for an initial sequence $(\theta_{(1)}, \ldots, \theta_{(j)})$. Of course, by construction it is $\theta_{(1)}$ that is supposed to be of interest, so that this should not be an objection.

The second limitation of the marginalization property is that it does not hold within, say, $\theta_{(1)}$. This can be seen from (3·11), where the marginal reference prior is not of the form (3·10), and is hence different from that which would have been obtained had the problem been originally confined to $(\theta_1, \ldots, \theta_l)$. We defer further discussion of this issue to § 3·4.

Moments of the reference priors. For comparing and understanding the group reference priors, it is useful to have expressions for their moments. Proof of the following is straightforward.

LEMMA 3·4. *For $N_{j-1}+1 \leq l \leq N_j$ and $\pi(\theta)$ defined by* (3·5),

$$E^{\pi}(\theta_l^s) = \prod_{i=1}^{j} \prod_{p=1}^{s} \left\{1 + \frac{n_i}{(2p-1)}\right\}^{-1}.$$

If all $n_i = 1$,

$$E^{\pi}(\theta_j^s) = \left\{\frac{(2s)!}{(2^s s!)^2}\right\}^j.$$

If $s = 1$, but n_i is arbitrary,

$$E^{\pi}(\theta_l) = \prod_{i=1}^{j} (1+n_i)^{-1},$$

Finally, the mean for the $(k+1)$st cell is

$$E^{\pi}(1 - \delta_{N_m}) = \prod_{i=1}^{m} (1+n_i)^{-1}.$$

3·4. *Discussion of the multinomial problem*

The multinomial scenario dramatically demonstrates how the m-group reference prior can 'decouple' groups of coordinates. Thus the inferences obtained for $\theta_{(1)}$ will depend only on $(r_1, \ldots, r_{n_1})$ and n, and not on what happens in other cells, or how many other cells there are. This is a natural property when, indeed, $\theta_{(1)}$ is of interest and the other parameters are nuisance parameters. Note that standard noninformative priors, such as Jeffreys's prior, do not have this property.

The desirability of this property can, however, be questioned. It requires an asymmetric treatment of the θ_i, and in problems where there is a small number of fixed 'indistinguishable' cells, such asymmetry may be unappealing.

To dramatize the difference, consider the two extremes of the 1-group and the k-group reference priors in (3·7) and (3·8). From Lemma 3·4, one sees, for instance, that the prior means of the θ_i are $(1+k)^{-1}$ for the 1-group Jeffreys's reference prior, but are 2^{-i} for the k-group reference prior. Thus the 1-group reference prior treats the θ_i equally, while the k-group reference prior gives exponentially decreasing mass to the θ_i as i increases.

This situation clearly demonstrates the impossibility of unambiguously defining 'noninformative'. Initially it seems reasonable to insist that a noninformative prior for a multinomial problem be exchangeable, and to require that it have the marginalization property; but these requirements are completely incompatible. Through consideration of a variety of examples we have convinced ourselves that the marginalization property is typically more important, and hence that the k-group reference prior is typically more attractive, but some flexibility is clearly required. In particular, if one has a small number of cells of interest, between which exchangeability seems very natural, it would clearly be tempting to use a 2-group reference prior, guaranteeing the marginalization property for the group of parameters of interest, while preserving exchangeability within the group. Thus, one might well want to be 'subjectively noninformative'.

If it is only $\theta_{(1)}$ that is of interest, note that there is no reason to even formally consider use of an m-group reference prior. The result will simply be that obtained by collapsing the original multinomial to the (n_1+1)-cell multinomial, with cell probabilities determined by $\theta_{(1)}$, and then using Jeffreys's prior for $\theta_{(1)}$. Thus, in practice, one needs to formally use the m-group reference prior only if more than the first group is of interest. Of course, this will typically be the case if the recommended full k-group reference prior in (3·8) is utilized and several of the θ_i are of interest.

4. Conclusions

We have considered m-group reference priors as a possible solution to the clear-cut need in multiparameter problems for developing noninformative priors with limited dependencies between groups of parameters, especially parameters of interest and nuisance parameters. There are different possible views on the success of the solution.

The least commital view is that the m-group reference prior method succeeds in generating a variety of interesting possible noninformative priors. For instance, (3·8) is very interesting for its marginalization property and is, to our knowledge, new. As candidates for in-depth study or for Bayesian sensitivity studies, these can be very useful noninformative priors, especially because of their ability to 'decouple' parameters. In this regard, the 1-group Jeffreys and k-group reference priors are likely to exhibit the greatest differences and, if a Bayesian analysis yields essentially the same answer for either prior there is reason to be confident in the answer.

The more optimistic view about m-group reference priors, in particular about k-group reference priors, so that each group has only one parameter, is that they provide the best available 'automatic' priors for general use. Our preference for the k-group reference prior is, for the most part, empirically based. In all examples we have considered, including many of the 'counterexamples' to Jeffreys's or other noninformative priors, the k-group reference priors have yielded very sensible results.

Our enthusiasm for k-group reference priors is slightly tempered by two issues we have touched on. First, they can be technically difficult or ambiguous to derive, especially when limits over $\{\Theta^l\}$ are needed. This can obviously reduce their pragmatic appeal,

although derivation of the k-group reference prior could be considered to be the theoretician's job, in which case the user is not affected.

The second difficulty with k-group reference priors is that they can depend on the ordering of the parameters and it can be difficult to decide on a complete ordering, especially for the nuisance parameters. One possibility is to order the parameters of interest, but group all nuisance parameters together, or maybe have several groups, when natural. Ordering within groups does not matter, and grouping nuisance parameters is rarely harmful, so this is often a sensible resolution of the ordering problem. A second possible solution is to try several different orderings, and see if it matters. Note , indeed, that the ordering or groupings of nuisance parameters frequently is immaterial, as in the multinomial problem. A third possible solution is to use the average of all the k-group reference priors from feasible orderings.

Many other issues could be raised. One of the most important is that of inference about functions $\varphi(\theta_1, \ldots, \theta_k)$ of the parameters. Reference prior theory, see, e.g. Bernardo (1979) or Berger & Bernardo (1989), requires a reparameterization, with $\varphi(\theta_1, \ldots, \theta_k)$ defined as the 'new' $\theta_{(1)}$, before the reference prior can be determined. Luckily short-cuts appear to be available, so that it is not necessary to completely redo the reference prior development for every function that is of interest. This work will be reported elsewhere.

ACKNOWLEDGEMENTS

This work was supported by the U.S.-Spain Joint Committee for Scientific and Technological Cooperation, and by Ministerio de Educación y Ciencia and National Science Foundation grants.

REFERENCES

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. *Statist. Decisions*, Suppl. **2**, 17-22.

BERGER, J. O. & BERNARDO, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Am. Statist. Assoc.* **84**, 200-7.

BERGER, J. O. & BERNARDO, J. M. (1992). Reference priors in a variance components problem. In *Proceedings of the Indo-USA Workshop on Bayesian Analysis in Statistics and Econometrics*, Ed. P. Goel, pp. 323-40. New York: Springer-Verlag.

BERGER, J. O., BERNARDO, J. M. & MENDOZA, M. (1989). On priors that maximize expected information. In *Recent Developments in Statistics and their Applications*, Ed. J. Klein and J. C. Lee, pp. 1-20. Seoul: Freedom Academy.

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. R. Statist. Soc.* B **41**, 113-47.

BERNARDO, J. M. (1980). A Bayesian analysis of classical hypothesis testing (with discussion). In *Bayesian Statistics* **1**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 605-47. Valencia University Press.

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. & GIRÓN, F. J. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics* **3**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 67-78. Oxford University Press.

BOX, G. E. P. & TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass: Addison-Wesley.

CIFARELLI, D. M. & REGAZZINI, E. (1987). Priors for exponential families which maximize the association between past and future observations. In *Probability and Bayesian Statistics*, Ed. R. Viertl, pp. 83–95. New York: Plenum.

CONSONNI, G. & VERONESE, P. (1989). A note on coherent invariant distributions as non-informative priors for exponential and location-scale families. *Comm. Statist.* A **18**, 2883–7.

DAWID, A. P., STONE, M. & ZIDEK, J. V. (1973). Marginalization paradox in Bayesian and structural inference (with discussion). *J. R. Statist. Soc.* B **35**, 189–233.

EAVES, D. M. (1983). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373–9.

EAVES, D. M. (1985). On maximizing the missing information about a hypothesis. *J. R. Statist. Soc.* B **47**, 263–5.

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

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

HILL, B. M. (1965). Inference about variance components in the one-way model. *J. Am. Statist. Assoc.* **60**, 806–25.

LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 896–1005.

LINDLEY, D. V. (1988). Statistical inference concerning Hardy–Weinberg equilibrium. In *Bayesian Statistics* **3**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 307–26. Oxford University Press.

MENDOZA, M. (1987). A Bayesian analysis of a generalized slope ratio bioassay. In *Probability and Bayesian Statistics*, Ed. R. Viertl, pp. 357–64. London: Plenum.

MENDOZA, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression model. In *Bayesian Statistics* **3**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 705–11. Oxford University Press.

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

STEIN, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–80.

YE, K. & BERGER, J. O. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika* **78**, 645–56.

[*Received November* 1989. *Revised January* 1991]