BAYESIAN STATISTICS 2, pp. 133-156
J.M. Bernardo, M.H. DeGroot, D. V. Lindley, A.F.M. Smith (Eds.)
© Elsevier Science Publishers B. V. (North-Holland), 1985

Quantifying Prior Opinion

PERSI DIACONIS
Stanford University

DONALD YLVISAKER
U.C.L.A.

SUMMARY

We investigate the approximation of a general prior by the more tractable mixture of conjugate priors. We suggest a new definition of conjugate prior for exponential families and offer a definition of conjugate prior for location families A practical example, involving Bernoulli variables, is treated in detail.

Keywords: CONJUGATE PRIORS; EXPONENTIAL FAMILIES; MIXTURES

1. INTRODUCTION

Bruno de Finneti has often emphasized the difference between the Bayesian standpoint and Bayesian techniques:

> "Bayesian Techniques, if considered as merely formal devices are no more trustworthy than any other tool (or *ad hoc* method) of the plentiful arsenal of 'objectivist statistics'."

In other words, there is more to Bayesian statistics than slapping down a convenient prior and computing Bayes rules. Let us illustrate these concerns through a simple example. Consider taking a specific penny and spinning it on its edge 50 times on a table. After observing the first 50 spins we are to predict the proportion of spins in a new series of spins and give an indication of how sure we are of our answer.

Any coherent Bayesian treatment of this problem can be interpreted as follows: Let S_n be the number of heads in the first n tosses. A parameter $p \epsilon [0,1]$ can be introduced so that the law of S_n given p is binomial; further, there is a prior distribution (here taken to have a density) $f(p)$ on $[0,1]$, such that

$$P(S_n = k) = \binom{n}{k} \int_0^1 p^k(1-p)^{n-k} f(p)dp .$$

After observing $S_n = k$, the predictions will be based on the posterior

$$f(p \mid S_n = k) = \frac{p^k(1-p)^{n-k} f(p)}{\int p^k(1-p)^{n-k} f(p)dp} .$$

At issue here is the prior $f(p)$. Let us distinguish three categories of Bayesians (certainly a crude distinction in light of Good's (1971) 46,656 lower bound on the possible types of Bayesians).

1. *Classical Bayesians* (Like Bayes, Laplace and Gauss) took $f(p) \equiv 1$. A so called flat prior.

2. *Modern Parametric Bayesians* (Raiffa, Lindley, Mosteller) took $f(p)$ as a beta density $\beta(a,b;p) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$. They note that this family contains a wide variety of distributional shapes, including the uniform prior ($a=b=1$). With a beta prior, the posterior becomes especially simple $f(p \mid S_n=k) = \beta(a+k,b+n-k;p)$.

3. *Subjective Bayesians* (Ramsey, de Finetti, Savage). Take the prior as a quantification of what is known about the coin and spinning process.

As an example of this third approach, consider the way that Diaconis quantified his prior for the coin spinning experiment: "To begin with, there is a big difference between spinning a coin on a table and tossing it in the air. While tossing often leads to about an even proportion of heads and tails (indeed one can sort of prove this from the physics involved) spinning often leads to proportions like 1/3 or 2/3. Some basis for this opinion can be reported: I remember reading a story in the New York Times about a high-school teacher who had his class spin a penny 5000 times. The result was 80 % tails. When I was a graduate student, Arthur Dempster spun a coin on edge 50 times with a similar, skew result. It is a well known proposition around certain pool rooms that some coins have very strong regular biases when spun on edge (1964D pennies favor tails). The reasons for the bias are not hard to infer. The shape of the edge will be a strong determining factor - indeed, magicians have coins that are slightly shaved; the eye cannot detect the shaving, but the spun coin *always* comes up heads".

With this experience as a base, a bimodal prior seemed appropriate-spun coins tend to be biased, but not alway to heads. No beta prior is bimodal of course. A simple class of bimodal priors is given by mixtures of symmetric beta densities. Figure 1 shows the density

$$\{\beta(10,20;p)+\beta(20,10;p)\}/2.$$

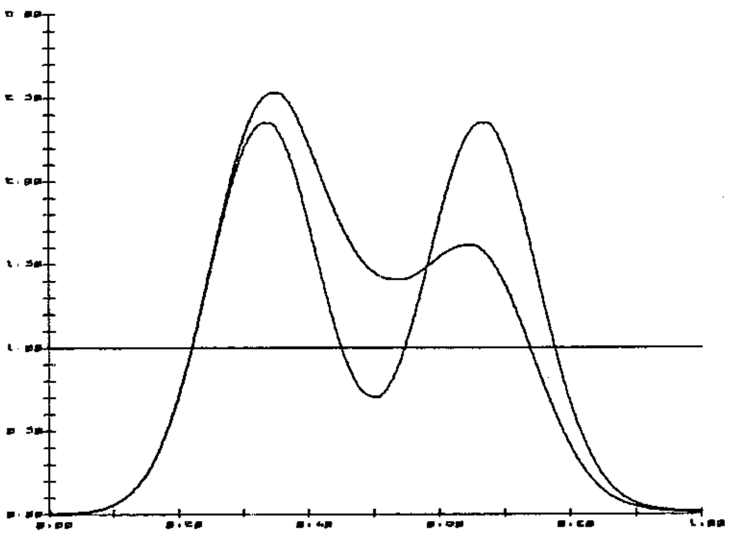

FIGURE 1. *Three prior distributions on [0,1]*

I = $\beta(1,1)$

II = $.5\beta(10,20) + .5\beta(20,10)$

III = $.5\beta(10,20) + .2\beta(15,15) + .3\beta(20,10)$

On reflection, it was decided that tails had come up more often than heads in the past; further some coins seemed likely to be symmetric. A final approximation to the prior was taken as

$$5\beta(10,20;p) + .2\beta(15,15;p) + .3\beta(20,10;p).$$

All of these priors are of the form

$$f(p) = \sum_{i=1}^{n} w_i\beta(a_i,b_i;p)$$

for weights w_i and parameters a_i,b_i. Notice that the posterior of such a mixture of beta densities is again a mixture of beta densities

$$f(p \mid S_n=k) = \sum_{i=1}^{n} w_i'\beta(a_i+k_1\,b_i+n-k;p).$$

Here the weights w_i' depend on n and k in a simple way

$$w_i' = c\,w_i \int p^k(1-p)^{n-k}\beta(a_i,b_i;p)dp, \qquad \text{with } c \text{ chosen so } \Sigma\, w_i' = 1.$$

The mixture prior can be thought of as a weighted combination of "beta populations", the w_i measuring the prior degree of belief that the actual coin was chosen from the i^{th} population. The posterior weights w_i' are proportional to the product of w_i and the relative likelihood of observing k successes in n trials in the i^{th} population.

The penny was actually spun. After 10 spins there were 3 heads and 7 tails. Figure 2 shows the posterior distributions corresponding to the 3 priors of Figure 1. Note that the 3

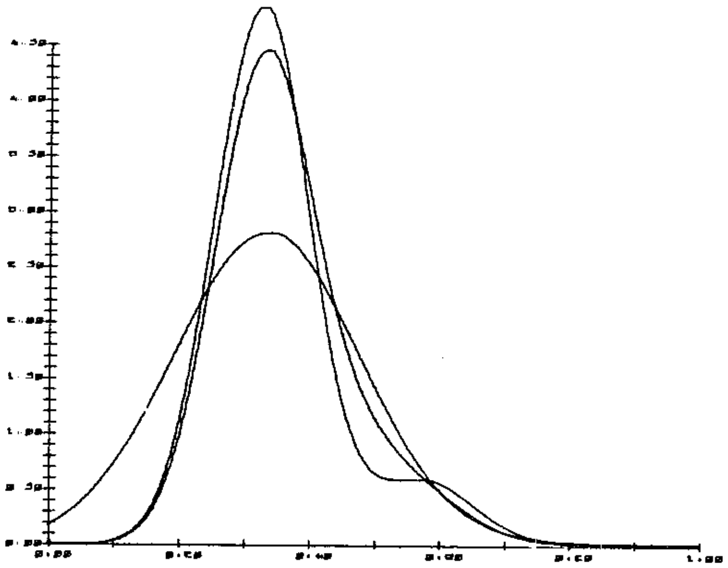

FIGURE 2. *The three posteriors after 3 heads in 10 trials*

I = $\beta(4,8)$

II = $.84\beta(13,27) + .11\beta(23,17)$

III = $.77\beta(13,27) + 16\beta(18,22) + .07\beta(23,17)$

modes agree (and point prediction from the 3 priors would be close) but the spreads are different, so that the variability assigned to predictions depend on the prior. After 50 spins there were 14 heads and 36 tails. The 3 priors are shown in Figure 3. They seem fairly close for any practical purpose.

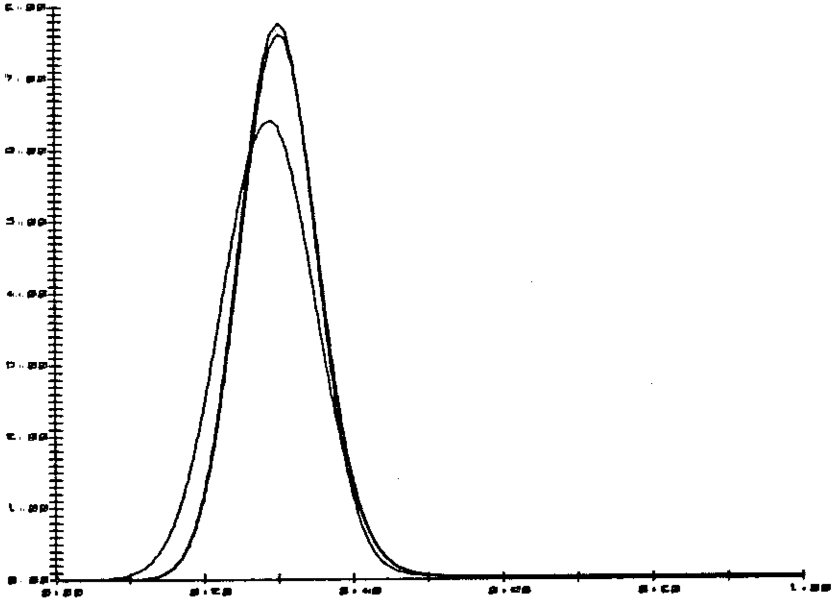

FIGURE 3 *The three posteriors after observing 14 heads in 50 trials*

I = $\beta(15,37)$

II = $.997\beta(24,56) + .003\beta(34,46)$

III = $.95\beta(24,56) + .047\beta(29,51) + .003\beta(34,46)$

The point of the example is that it is pretty easy to be an honest Bayesian using mixtures of conjugate priors. The computations for updating are straightforward. Of course, for "large samples" such careful quantification will not be important (at least in low-dimensional problems). However, for small or moderate samples, the prior matters.

It is natural to consider if we have gone far enough in considering mixtures of beta densities. Can any density be well approximated now, or are some opinions still ruled out. It is easy to see that any prior (density or not) can be well approximated by a mixture of beta densities. The reason is simple: by choosing a_i and b_i large, the densities $\beta(a_i,b_i;p)$ is close to a point mass at $a_i/(a_i+b)$ and mixtures of point masses are clearly dense in the weak star topology.

A more quantitative argument follows from

Theorem 1 Let $f(p)$ be a continuous density on [0,1]. Then, there exist $\{w_i,a_i,b_i\}_{i=0}^{n}$ such that

$$\max_{p} |f(p) - \sum_{i=0}^{n} w_i\beta(a_i,b_i;p)| \leq \tfrac{7}{4}\,\omega_f(1/\sqrt{n})$$

where the modulus of continuity is defined as

$$\omega_f(t) = \max_{|x-y|<t} |f(x)-f(y)|.$$

Proof. Consider the modified Bernstein polynomial

$$\sum_{i=0}^{n} w_i\,\beta(i+1,n-i+1;p), \quad w_i = \int_{i/n+1}^{i+1/n+1} f(x)dx.$$

(Note: The more usual Bernstein polynomial is

$$\Sigma f(\tfrac{i}{n})\binom{n}{i}x^i(1\text{-}x)^{n\text{-}i} = \frac{1}{n+1}\sum_{i=0}^{n} f(\tfrac{i}{n})\beta(i+1,n\text{-}i+1,n\text{-}i+1;p)$$

this has weights $\dfrac{f(i/n)}{n+1} = w_i.$)

The usual non-probabilistic proof of the Wierstrass approximation theorem, using Bernstein polynomials, as given in chapter 2 of Lorentz (1966), goes through with the weights as given to yield the stated result. □

Remarks. For differentiable functions, $\omega_f(1/\sqrt{n}) \doteq c/\sqrt{n}$ for a constant c, so the approximation is of order $1/\sqrt{n}$. This is known to be the best possible rate of approximations by Bernstein polynomials as consideration of the density $f(x) = 4|x-\frac{1}{2}|$ shows. It is known that the best degree n polynomial approximation to a continuous function if of order $\omega_f(\frac{1}{n})$ (Jackson theorem). Indeed, it is possible to characterize the functions which can be well approximated by Bernstein polynomials, this work can be found by looking in the book by Lorentz or recent years of the Journal of Approximation Theory under the heading of saturation classes of Bernstein polynomials.

The purpose of the proof is *not* to suggest direct use of the Bernstein polynomial approximation, rather just to show that approximation is possible. In the example, 2 terms were chosen. In other examples, a small number of terms may be chosen so that the moments or a few quantities match exactly.

The purpose of the present paper is to indicate that the techniques used in the example apply fairly generally. A version of it holds for mixtures of conjugate priors in multivariate exponential families, and more generally yet. The structure of this paper is as follows: in section 2 we review the work of Diaconis and Ylvisaker (1979) on conjugate priors for exponential families. The main result here is that the standard families of priors can be characterized by a simple property of posterior linearity. In section 3 we indicate how the proposed definition of conjugate priors carries over to non-exponential families. This overlaps with work of Goldstein (1975). The final section discusses some definitions of approximation suitable for Bayesian inference. Here the results are less complete and there are many open research problems. Some of these problems are solved by Dalal and Hall (1983) who also work with mixtures.

2. CONJUGATE PRIORS FOR EXPONENTIAL FAMILIES

Conjugate priors are widely used for the usual exponential families of parametric statistics (normal, binomial, Poisson, gamma, etc.). We will suggest mixtures of such priors to approximate any prior. We begin by pointing out that the usual definitions are essentially vacuous.

Most often, a conjugate family is defined either as a family of priors that is closed under sampling or as a family of priors which is proportional to the likelihood. Consider the beta priors for coin tossing and observe that for any continuous non-negative function h the family

$$c\, p^a(1-p)^b h(p)dp$$

is closed under sampling and has a density (with respect to the carrier $h(p)dp$) which is proportional to the likelihood. Since h is an essentially arbitrary function, it would seem that *any* prior on $[0,1]$ is a conjugate prior. The usual family has $h(p) = 1$. In Diaconis and Ylvisaker (1979) we asked what additional properties of a prior give the families usually called conjugate priors. It turns out that such priors can be characterized by a condition of posterior linearity For the binomial distribution this becomes

$$E\{p \mid S_n = k\} = ak+b \qquad k=0,1,2, \ldots, n.$$

A result like this holds for any of the standard families and actually characterizes the prior. Three of the main results will now be stated. We begin by stating the results for exponential families in their natural parametrizations. Following this we describe how the results transform to the more usual parametrizations. The results are equivalent. Any exponential family can be written in terms of the natural parametrization, and this allows a unified treatment.

Start with a fixed σ-finite measure μ on the Borel sets of $\mathbb{R}^d$--*the carrier measure*. Let $\mathcal{X}$ be the interior of the convex hull of the support of μ. Assume $\mathcal{X}$ is non-empty. Define $M(\theta) = \log \int e^{\theta \cdot x}\mu(dx)$ and let $\Theta = \{\theta: M(\theta) < \infty\}$. As usual, Holder's inequality shows that Θ is a convex set. It is called the *natural parameter space*. Throughout we assume that Θ is non-empty and open. The *exponential family* $\{P_\theta\}$ *of probabilities through* μ is defined as

$$dP_\theta = e^{x \cdot \theta - M(\theta)}\,\mu(dx) \quad \theta \in \Theta.$$

As usual, the expectations under P can be determined by differentiating M:

$$E_\theta(X) = M'(\theta).$$

Define a family $\{\tilde{\Pi}_{n_0,x_0}\}$ of measures on Θ by

$$\tilde{\Pi}_{n_0,x_0}(d\theta) = e^{n_0 x_0 \cdot \theta - n_0 M(\theta)}\, d\theta, \quad n_0 \in \mathbb{R}, \quad x_0 \in \mathbb{R}^d.$$

If $\tilde{\Pi}_{n_0,x_0}$ can be normalized to a probability Π_{n_0,x_0} on Θ, it will be termed a *distribution conjugate to the exponential family* $\{P_\theta\}$ *of 2.1.* The next theorem determines for which (n_0,x_0) normalization is possible:

Theorem 2 a) If $\Theta = \mathbb{R}^d$, $\tilde{\Pi}_{n_0,x'_0 s}(\Theta) < \infty$ if and only if $n_0 > 0$, $x_0 \in \mathcal{X}$.

b) If $\Theta \neq \mathbb{R}^d$ and $n_0 > 0$, $\tilde{\Pi}_{n_0,x_0}(\Theta) < \infty$ if and only if $x_0 \in \mathcal{X}$.

The next theorem unifies many standard Bayesian calculations. It shows that x_0 is the prior mean of the parameter $E_\theta(X)$. It has been part of the folklore for years. A rigorous proof of the 1-dimensional case appears in Jewel (1974a,b).

Theorem 3. If θ has the distribution Π_{n_0,x_0} for $n_0 > 0$ and $x_0 \in \mathcal{X}$ then

$$E(E_\theta(X)) = x_0.$$

Remark 1. The result gives posterior linearity for a sample $X_1, X_2, \ldots, X_n$ of size n from P_θ. Indeed, if Π_{n_0,x_0} is the prior for θ, the posterior density is $\Pi_{n_0+n,\, \frac{n_0x_0+n\bar{x}}{n_0+n}}$ with $\bar{x}$ the mean of the sample. Theorem 3 yields

$$E\{E_\theta(X) \mid X_1 \ldots X_n\} = \frac{n_0x_0+n\bar{x}}{n_0+n}.$$

Thus the posterior expectation of the mean parameter is a convex combination of the prior expectation of the mean parameter and $\bar{x}$. The weights are proportional to n_0 and the sample size n--in this sense n_0 may be thought of as a prior sample size. Novick and Hall (1965) have considered negative values of n_0 which yield improper "ignorance priors".

Remark 2. The argument for theorem 3 is integration by parts: consider the one-dimensional case. Then, as usual, $E_\theta(X) = M'(\theta)$ and

$$\int M'(\theta)e^{n_0x_0\theta - n_0M(\theta)}\, d\theta = \frac{e^{n_0x_0\theta - n_0M(\theta)}}{n_0}\Bigg|_{\underline{\theta}}^{\overline{\theta}} + x_0 \int e^{n_0x_0\theta - n_0M(\theta)}\, d\theta.$$

The boundary terms vanish because Θ is open and the right side is x_0 times the correct norming constant. In higher dimensions, the argument is more complicated.

The next theorem gives a converse to Theorem 3 which characterizes conjugate priors.

Theorem 4 Let X be a sample of size one from P_θ and suppose the support of μ contains an open interval in $\mathbb{R}^d$. If θ has a prior distribution τ which is not concentrated at a single point and if

$$E\{E_\theta(X) \mid X\} = aX+b$$

for some constant a and vector b, then $a \neq 0$, τ is absolutely continuous with respect to Lebesgue measure, and

$$\tau(d\theta) = ce^{a^{-1}b\,\theta - a^{-1}(1-a)M(\theta)}d\theta.$$

Remarks. Versions of Theorem 4 appropriate for discrete data are given in Diaconis and Ylvisaker (1979). The known results handle all the usual families, but they are still annoyingly incomplete.

Thus far we have assumed that the exponential family was given in its natural parametrization. Often, standard families are parametrized in other terms such as the parametrization involving p for the binomial. We now show how to transform the prior on Θ into a prior on any given parameter space to preserve linearity.

Let $\psi:\Theta \to \mathbb{R}^d$ be a diffeomorphism with range Θ_ψ and inverse ψ^{-1}. This transforms Lebesgue measure via multiplication by a Jacobian ψ'. The image of the prior Π_{n_0,x_0} becomes

$$e^{n_0x_0\psi^{-1}(t) - n_0M(\psi^{-1}(t))}\ \frac{1}{\psi'(\psi^{-1}(t))}\, dt \quad \text{for } t \in \Theta_\psi.$$

If t is taken as parameter, so the family becomes

$$e^{x\,\psi^{-1}(t) - M(\psi^{-1}(t))}\,\mu(dx) \quad t \in \Theta_\psi.$$

Then by standard properties of conditional expectation, we have as before

$$E\{E_t(X) \mid X_1 \ldots X_n\} = \frac{n_0x_0+n\bar{x}}{n_0+n}.$$

The conjugate family in terms of t is still closed under sampling and has posterior proportional to likelihood. The Jacobian factor simply specifies a choice of carrier measures which gives the additional property of posterior linearity. It is instructive to carry out the calculations for the standard families and see the standard conjugates, in their usual forms emerge at the end. Morris DeGroot has observed that the Jacobian factors always seem to merge in a nice way with the norming constant $M(\psi^{-1}(t))$. We do not know a theorem that makes this precise.

The final result of this section is the analog of Theorem 1 for a d-dimensional exponential family. The result shows that any prior (with or without a density) can be well approximated by a finite mixture of conjugate priors. The notation and assumptions are the same as in theorems 2,3, and 4.

Theorem 5. Let Θ be the natural parameter space of a d-dimensional exponential family. For any probability π on Θ, and any $\epsilon > 0$ there are weights w_i, and (n_i, x_i), $n_i > 0$, $x_i \epsilon \mathcal{X}$ such that if

$$\tilde{\pi}(d\theta) = \sum_{i=1}^{N} w_i\, c(n_i, x_i)\, e^{n_i x_i \theta - n_i M(\theta)}\, d\theta .$$

Then

$$d(\pi, \tilde{\pi}) < \epsilon$$

where d is the Prohorav metric.

Proof. It is well known, and easy to argue direetly, that finite mixtures of point masses are weak star dense. A proof in a general setting is in Theorem 12.11 of Choquet (1969). Hence we must only show that any point mass can be weak star approximated by a prior of the form

$$c(n_0, x_0) e^{n_0 x_0 \theta - n_0 M(\theta)}\, d\theta . \qquad (2.1)$$

To see this, differentiate (2.1), the maximum occurs at the unique value of θ satisfying

$$\nabla M(\theta) = x_0$$

It is straightforward to show that as n_0 tends to infinity, the prior (2.1) concentrates at this θ. Finally, for any $\theta_0 \epsilon \Theta$, $M'(\theta_0) = E_{\theta_0}(X) \epsilon \mathcal{X}$, so for any θ_0, there is an $x_0 \epsilon \mathcal{X}$ such that $M'(\theta_0) = x_0$. This completes the argument. □

Remarks. One can also emulate the proof using Bernstein polynomials. This has been carried out to yield an approximation with error term in unpublished thesis work of Mark Jacobson at Stanford University. Related results are in Lorentz (1953), Dubins (1983), Dalal (1978), and Dalal and Hall (1980, 1983). Again, the theorem above is just meant to indicate that mixtures are capable of approximating any prior.

3. LOCATION PARAMETERS

In Section 2 we suggest using linear posterior expectation as a definition of conjugate priors. This offers the possibility of moving away from the exponential family setting. The present section carries out this program for location parameter problems. The main characterization result is theorem 5 which extends a theorem proved by Goldstein (1975). Conjugate priors are suggested only as a convenient building block: We can show, in certain circumstances, that any prior can be well approximated by a mixture of conjugate priors.

We begin by giving a class of priors with linear posterior expectation. Let X be a d-dimensional random vector with distribution function F, and let θ be a random vector independent of X with prior distribution F^{*n} (n-fold convolution). For the location problem, $F(x-\theta)$, the observed variable is $Z = X + \theta$. In what follows, we do not want to assume that the prior has a mean We thus define

$$E(\theta \mid Z) = g(Z) \text{ if and only if } E(\theta_i^+ \mid Z) - E(\theta_i^- \mid Z) = g_i(Z)\ a.s.\ 1 \leq i \leq d$$

where the subscript denotes the i^{th} coordinate Conditional expectation in this sense is still linear and of course agrees with the usual notion when means are finite c.f. Strauch (1965) With this definition we have, provided the expectations exist,

$$E(\theta \mid Z) = \frac{n}{n+1} Z . \qquad (3.1)$$

To see this, let $X_1, \ldots, X_n$ be independent of X and each other so that $S_n = X_1 + \ldots + X_n$ has the same distribution as θ. Then

$$\frac{n+1}{n} E\{S_n \mid S_n + X\} = E\{S_n \mid S_n + X\} + \frac{1}{n} E\{S_n \mid S_n + X\}$$

$$= E\{S_n \mid S_n + X\} + E\{X \mid S_n + X\} = E\{S_n + X \mid S_n + X\} = S_n + X .$$

For example, if X and θ are independent Cauchy variables, $E(\theta \mid X + \theta)$ exists and equals $\frac{X+\theta}{2}$. The class of priors can be widened by allowing inclusion of a known location parameter for the prior. With this in mind, we define a conjugate prior for the location parameter problem through F as any prior on $\mathbb{R}^d$ such that

$$E(\theta \mid Z) = aZ + b \quad \text{for real } a \text{ and } b \in \mathbb{R}^d \qquad (3.2)$$

For location problems, posterior linearity only holds for samples of size 1. For larger samples, $\bar{X}$ is not a sufficient statistic unless the distribution of X is normal, see Ferguson (1954) or DeGroot and Goel (1981) for more on this point.

It will surface in the proof of theorem 6 that (3.2) can hold only if $0 \leq a \leq 1$. The following lemma determines what happens at the boundary cases.

Lemma 1. Let X and θ be independent random variables and suppose $E(|X|) < \infty$. Then $E(\theta \mid X + \theta) = b$ if and only if θ is a.s. constant and $E(\theta \mid X + \theta) = (X + \theta) + b$ if and only if X is a.s. constant.

Proof. The lemma follows from a result in Doob (1953, p 314) To bring things into that framework, let $Y = X + \theta$. Without loss of generality $b = 0$. Under the first assumption $0 = E(\theta \mid X + \theta) = E(\theta + X \mid X + 0) - E(X \mid X + \theta)$, so $E(X \mid Y) = Y$. Of course, $E(Y \mid X) = X$. This says that X and Y form a 2-term martingale in either order, and this is just what Doob shows is impossible when $E|X| < \infty$. The argument under the second assumption is similar. For related theorems see Girshick and Savage (1951, p. 1653) or Gilat (1971). □

Assume now that (3.2) holds for a location parameter problem and consider the question of uniqueness of the underlying prior distribution To state theorem 6, let $X = (X_1, \ldots, X_d)'$ have distribution functions F not concentrated at a point and write $\lambda_{2n} = \int \{x_1^{2n} + \ldots x_n^{2n}\} dF$

Theorem 6. Let X and θ be independent d-dimensional random vectors with neither X nor θ a.s. constant. Assume $E|X_i| < \infty$ for $i = 1, \ldots, d$ and that either

a) the characteristic function of X has no zeros or

b) $\sum_{n=1}^{\infty} \lambda_{2n}^{-1/2n} = \infty$.

If

$$E(\theta|X+\theta) = a(X+\theta)+b \qquad (3.3)$$

Then $0 < a < 1$ and the distribution of θ is uniquely determined.

Proof: We begin with $d = 1$ and show first that $E|\theta| < \infty$. Now $a^2 \neq a$ from lemma 1 and, by translating X if necessary, one can take $b = 0$. From (3.3) and the linearity of expectation,

$$E(\theta|X+\theta) = a(X+\theta) = aE(X+\theta|X+\theta) = aE(X|X+\theta)+aE(\theta|X+\theta).$$

Hence

$$E(\theta|X+\theta) = \frac{a}{1-a}E(X|X+\theta) \qquad (3.4)$$

Use this in (3.3) to find

$$\theta = \frac{1}{1-a}E(X|X+\theta)-X \qquad (3.5)$$

By assumption, the right side of (3.5) is absolutely integrable so $E|\theta| < \infty$

When X and θ have finite expectations, it follows from lemma 1.1.1 of Kagan, Linnik and Rao (1973) that (3.3) holds if and only if the characteristic functions of X and θ satisfy

$$(1-a)\phi_\theta'(t)\phi_X(t) = a\phi_X'(t)\phi_\theta(t) \text{ for all } t \qquad (3.6)$$

Since $\phi_X(t)$ does not vanish in some interval I about 0, (3.6) gives for $t \in I$,

$$\phi_\theta(t) = \phi_X(t)^{\frac{a}{1-a}} \qquad (3.7)$$

Observe first that for (3.7) to hold with neither X nor ϕ constant, it must be that $a/(1-a) > 0$ and so $0 < a < 1$. Now if ϕ_X never vanishes, ϕ_θ is determined by (3.7). On the other hand, if the distribution of X satisfies *b* and so is determined by its moments, the corresponding moments of θ can be computed from (3.7) and satisfy the same determinedness condition. The proof of part *a* is complete.

Now suppose that X satisfies (b). We first argue that θ has moments of all order. For the rest of the argument, suppose we are given a pair of random variables θ_1, X_1 satisfying $E(\theta_1|\theta_1+X_1) = a(\theta_1+X_1)$. For fixed $0 < a < 1$; we show that the existence of moments for X_1, $\lambda_{2m}^{(1)} = EX_1^{2m}$, implies the existence of moments for θ_1, $\mu_{2m}^{(1)} = E\theta_1^{2m}$ To see this let (θ_2,X_2) be an independent copy of (θ_1,X_1) and set $\theta = \theta_2-\theta_1$, $X = X_2-X_1$. It follows that

$$E(\theta|\theta+X) = a(\theta+X) \ ,$$

and X are symmetric with

i) $E(\theta_1-E\theta_1)^{2m} \leq E(\theta_2-\theta_1)^{2m} = \mu_{2m}$

ii) $\lambda_{2m} = EX^{2m} = \sum\limits_{j=0}^{2m} \binom{2m}{j}(-1)^j\lambda_j^{(1)}\lambda_{2m-j}^{(1)} \leq 2^{2m}\lambda_{2m}^{(1)}$.

We shall show below that there is a C for which $\mu_{2m} \leq C^{2m}\lambda_{2m}$ for all m and then from i) and ii) we will have

$$E(\theta_1-E\theta_1)^{2m} \leq \mu_{2m} \leq C^{2m}\lambda_{2m} \leq (2C)^{2m}\lambda_{2m}^{(1)}.$$

From a knowledge of X_1 and its moments, we have $E(\theta_1) = \frac{a}{1-a}E(X_1)$ so it will generally be seen that if X_1 satisfies (b), so does $\bar{\theta} = (\theta_1-E\theta_1)$ and then, so does θ_1. To show that there is a C with $\mu_{2m} \leq C^{2m}\lambda_{2m}$ first observe that $E(\theta(\theta+X)^{2m-1}) = aE(\theta+X)^{2m}$. Thus

$$\sum_{j=1\ \ j\text{ odd}}^{2m-1}\binom{2m-1}{j}\mu_{j+1}\lambda_{2m-1-j} = a\sum_{j=0\ \ j\text{ even}}^{2m}\binom{2m}{j}\mu_j\lambda_{2m-j}$$

so that

$$0 = \sum_{j=0\ \ j\text{ even}}^{2m}\left[\frac{a\binom{2m}{j}-\binom{2m-1}{j-1}}{\binom{2m}{j}}\right]\binom{2m}{j}\mu_j\lambda_{2m-j}.$$

We write this in the form

$$(*) \qquad \sum_{0\leq r\leq ma}\left(a-\frac{r}{m}\right)\binom{2m}{2r}\mu_{2r}\lambda_{2m-2r} = \sum_{ma\leq r\leq m}\left(\frac{r}{m}-a\right)\binom{2m}{2r}\mu_{2r}\lambda_{2m-2r}.$$

Determine the smallest r_0 so that $\frac{r_0}{r_0+1} > a$ and a C at least as large as $(\frac{2}{1-a})^{\{1/(1-a)\}}$ so that

$$(**) \qquad \mu_{2r} \leq C^{2r}\lambda_{2r} \quad \text{for} \quad r = 1,2,\ldots,r_0.$$

For $m = r_0+1$, note that $ma = (r_0+1)a < r_o$ so that in the left hand side of (*) all $r \leq r_0$. Then the left hand side is

$$\leq \sum_{0\leq r\leq ma}\binom{2m}{2r}C^{2r}\lambda_{2r}\lambda_{2m-2r} \leq C^{2ma}2^{2m}\lambda_{2m}$$

while the right hand side of (*) exceeds $(1-a)\mu_{2m}$. Hence

$\mu_{2m} \leq \frac{1}{1-a}C^{2ma}2^{2m}\lambda_{2m} \leq C^{2m}\lambda_{2m}$ provided $C^{2m} \geq \frac{1}{1-a}C^{2ma}2^{2m}$. This requires that $C^{2m(1-a)}$ $\geq \frac{1}{1-a}2^{2m}$ or $C \geq \frac{2^{\{1/(1-a)\}}}{(1-a)^{\{/(2m(1-a))\}}}$. But in fact $C \geq (\frac{2}{1-a})^{\{1/(1-a)\}} \geq \frac{2^{\{1/(1-a)\}}}{(1-a)^{\{1/(2m(1-a))\}}}$.

Therefore (**) is extended to r_0+1 with the same C as before. Performing the same induction step again requires only that $ma = (r_1+1)a$ *be* $< r_1$ and this is guaranteed inasmuch as $\frac{r_1}{r_1+1} > \frac{r_0}{r_o+1} > a$

For $d>1$, the fimiteness of $E|\theta_i|$ follows readily by the arguments used earlier. Take $b = 0$ again without real loss. Now let $\varrho \in \mathbb{R}^d$ and find from (3.3) that

$$E(\varrho\cdot\theta|X+\theta) = a\varrho\cdot(X+\theta) = E(\varrho\cdot\theta|\varrho\cdot(X+\theta)).$$

If the characteristic function of X has no zeros then the same is true of the characteristic function of $\varrho\cdot X$. Hence the one-dimensional version of the theorem implies that the distribution of $\varrho\cdot\theta$ is uniquely determined and so therefore is the distribution of θ. If the distribution of X satisfies b), the inequality $|\varrho\cdot X|^{2n} \leq m(X_1^{2n}+\ldots+X_d^{2n})$ with m depending on ϱ and d but not on n or X, implies that the distribution of $\varrho\cdot X$ is determined by its moments. The proof of the theorem is completed by another application of the one-dimensional version. □

Remark 1. Here is an example of independent random variables X and θ having finite means, different distributions and satisfying (3.3) with $a = \frac{1}{2}$, $b = 0$. The example makes it clear that some hypothesis on the distribution of X are required in Theorem 5 in order to guarantee uniqueness. Now (3.3) holds with $a = \frac{1}{2}$ and $b = 0$ if and only if $\phi_X'\phi_\theta = \phi_\theta'\phi_X$ as at (3.6). Let X have density

$$\frac{3 \cdot 4^3}{\pi} \left(\frac{\sin \frac{x}{4}}{x} \right)^4 \qquad -\infty < x < \infty .$$

Such an X has a finite mean and a real characteristic function which is continuously differentiable and vanishes outside $(-1,1)$. Using Theorem 4.32 of Lucas (1970) we see that the function ϕ_θ which equals ϕ_X on $(-1,1)$ and has period 2 is the characteristic function of an arithmetic distribution. By construction $\phi_X'\phi_\theta = \phi_\theta'\phi_X$, since the two sides are equal in $(-1,1)$ and both sides vanish outside this interval

The construction can be varied to give X and θ having different distributions but moments of all orders: Let $\phi(t)$ be the well known "tent" characteristic function supported on $(-1,1)$. Convolving $\phi(t)$ with itself leads to smoother and smoother functions with compact support--the example above is based on $\phi * \phi$ The function

$$\psi(t) = \mathop{*}_{n=1}^{\infty} \phi(2^n t)$$

may be shown to be an infinitely differentiable characteristic function with support on $[-1,1]$. If it is used to define X and its periodic continuation is used to define θ, we have an example with moments or all orders where $E\{\theta \mid X+\theta\} = \{X+\theta\}/2$ but the law of θ is not unique.

Remark 2 In theorem 5 we have used Carlemans sufficient condition for a distribution to be determined by its moments. We wondered if this could be replaced by the weaker condition that X was determined by its moments. Here is one result in that direction.

Proposition 1 If X is determined by its moments and θ is independent of X and satisfies

$$E(\theta \mid X+\theta) = \frac{1}{n}(X+\theta) \text{ for any fixed integer } n \geq 2 ,$$

then the distribution of θ is uniquely determined

Proof. From (3.7), $\phi_X(t) = \phi_\theta(t)^{n-1}$ in a neighborhood of 0. This implies the moments of θ. We want to show that θ is determined by its moments. If not, then, by a fundamental theorem of moment theory, see Theorem B of Landau (1980), for any real t, there is a probability ψ with all the same moments as X, and so the same distribution as X, but ψ^{*n} has an atom at the point nt. Since t is arbitrary, there is a probability with the same moments as X but having an atom at any specified place. This contradiction proves the proposition. □

If it were true that being determined by moments was inherited by convolutions, general rational values of a could be handled; for from $\phi_X^n = \phi_\theta^m$in a neigborhood of zero, and X determined would follow ϕ_X^n determined and then ϕ_θ determined by the above argument. Cristian Berg (1983) has provided a probability μ, which is determined by its moments but such that $\mu * \mu$ is not determined! For more on these matters, see Devinatz (1959).

Remark 3. We note the connection of the present section to a result in Martingale theory. If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables with $E|X_i| < \infty$, then the argument used in the introduction to this section shows that S_n/n is backward martingale. Here $S_n = \Sigma_{i=1}^n X_i$ and the martingale property is

$$E\{ \frac{S_n}{n} \mid S_{n+1}, S_{n+2}, \ldots \} = \frac{S_{n+1}}{n+1}, \qquad n = 1,2, \ldots \tag{3.8}$$

using obvious generalizations of the argument in Theorem 6 we can show that if $\{X_i\}$ are independent random variables such that (3.8) holds and $E|X_1| < \infty$, then all the X_i have finite first moments. If ϕ_i is the characteristic function of X_i, (3.8) holds if and only if for all $k \geq 2$,

$$\phi_k' \prod_{j=1}^{k-1} \phi_j(t) = \frac{\phi_k(t)}{k-1} \Big(\prod_{j=1}^{k-1} \phi_j(t) \Big)' = \phi_{k-1}'(t) \prod_{j=1}^{k} \phi_j(t) \qquad \text{for all } t$$

If $\phi_1(t) \neq 0$ then all the $\phi_i = \phi_1$ Using variants of the construction in remark 2 we can construct an infinite sequence of independent random variables, each with a different distribution, which satisfy (3.8)

Remark 4. For a distribution function F, let S_F be the set of real numbers a which can occur in (3.2). From (3.1), $n/(n+1) \epsilon S_F$ for all $n = 1,2, \ldots$. If F is uniform on $[0,1]$, these are the only numbers in S_F. If F is infinitely divisible with finite mean then ${\phi_X}^{a/1-a}$ is a characteristic function for every $a \epsilon (0,1)$ so $S_F = (0,1)$. Conversely, if Φ has a mean, $\phi_F(t) \neq 0$ and $S_F = (0,1)$, Theorem 6 imphes F is infinitely divisible.

Remark 5. When can any prior be approximated by mixtures of conjugate priors in a location parameters setting? If X is infinitely divisible then the conjugate priors have characteristic functions of the form $e^{-it\mu}\phi_X^\alpha$. As $\alpha \rightarrow 0$, this approaches a point mass at μ. We conclude that approximation is possible when X is infinitely divisible. It may be that the converse of this holds. If X is uniform on $[0,1]$ then the only conjugate priors are translates of convolutions of uniforms; it is easy to show that finite mixtures of these are not weak star dense in all probabilities on the line.

Remark 6. It is possible to develop some theory of the above sort for scale parameters. As an exâmple, we note the following. If X has a beta density with parameters a and b, and θ has a beta density with parameters $a+b$ and c, then

$$E\{\theta \mid X\theta\} = \frac{cX\theta}{b+c} + \frac{b}{b+c} .$$

4. SOME RESEARCH PROBLEMS

The material discussed above suggests many questions, both technical and philosophical. These are discussed here under the following headings: What are we doing: when are two priors close?; Stability; extensions to non-linear regression; matrices; and connections with de Finettis' Theorem.

4.1. *What are we doing: When are two priors close?*

The approximation theorems (1 and 6) involve a topology on the space of all priors. It is both practically relevant and philosophically natural to link the topology to the actual problem in hand. Thus to say when two priors are close entails specifying a use for the priors. Here are some specific suggestions drawn from work of Stein (1965).

Consider a decision problem specified by a family of probabilities $\{P_\theta\}_{\theta\epsilon\Theta}$ and loss function L. Suppose that π_t represents a true prior and π_a an approximation. An observation x yields two posteriors π_t^x and π_a^x. Each of these will result in certain decision ("Bayes rules") $\delta_t(x)$ and $\delta_a(x)$. Here δ_t minimizes the risk $R(\pi_t,\delta) = E_t\{L(\theta,\delta(x))\}$. The difference in risk, if δ_a is used instead of δ_t can serve as a measure of separation between the two priors:

$$S(\pi_t,\pi_a) = R(\pi_t,\delta_a) - R(\pi_t,\delta_t) .$$

Remark. Of course the separation S is not a metric. Stein (1965) shows that it is not even symmetric. Some further discussion and interpretation is in Diaconis and Stein (1983). We have verified that when the statistical problem is estimation of a binomial parameter θ based on a sample of size n, with squared error as loss, then any prior can be approximated, in the sense of the separation S, by a mixture of beta priors. This is an easy case. Other loss functions, and unbounded parameter spaces certainly merit careful study. Jim Berger has shown us arguments that suggest that for estimating a normal mean with squared error as loss a Cauchy prior for the mean cannot be approximated in the sense indicated.

Stein has suggested an intermediate notion of separation which does not depend on the loss function: Let $\varrho(\cdot,\cdot)$ be a metric between probabilities. Consider

$$\varrho^*(\pi_t,\pi_a) = E_t\{\varrho(\pi_t^x,\pi_a^x)\}$$

where the expectation is taken with x given its true marginal distribution. Stein has sketched an argument to show that if $L(\cdot,\cdot)$ is smooth in its second argument, and ϱ is taken as the Hellinger distance:

$$\varrho(P_1,P_0) = \int(p_1^{1/2}-p_0^{1/2})\lambda$$

where p_i is the density of P_i with respect to the dominating measure λ, then

$$S(\pi_t,\pi_a) \leq c\, \varrho^*(\pi_t,\pi_a)$$

holds for some constant c, depending on L and perhaps π_t, but not on π_a. Thus, approximation in ϱ^* entails approximation in separation for a wide variety of problems. Again, it seems worthwhile to have examples and some "honest" theorems. It seems likely that if ϱ is taken as any metric metrizing the weak star topology, any prior can be approximated by mixtures of conjugate priors. Related issues are discussed by Kadane and Chuang (1978).

The language of "approximate" and "true" priors open a philosophical can of worms. Recent works by Shafer (1981), Jeffrey (1982) and Diaconis and Zabell (1982) emphasized the *constructive* nature of forming a prior - none of us have a true prior, sitting inside, waiting to be "elicited". Clearly, as we think about things different possibilities and refinements will leap to mind. The very act of thinking provides valuable "data" so that the true prior is always unknown, as is the "true position" of an electron.

One way around these difficulties is the conceptually difficult task of thinking "will it matter if I refine my prior further". Often, in well travelled problems, the fine details of a prior will not matter much. This is likely to be the case in low dimensional problems with reasonable sized samples.

A rather different argument against very careful specifications of a prior follows from work of Jeffrey (1982) and Diaconis and Zabell (1982). They argue that we often up date a prior to a posterior by methods different from Bayes theorem. This will tend to be true in "exploratory data analysis" situations where it is practically impossible to quantify a prior over a sufficiently high dimensional space, and our reaction to data is likely to be "I forgot all about that possibility".

4.2. *Stability*

It would be useful to have some more quantitative measures of the effect of small changes in prior specification on final decisions. One approach is to use the separation $S(\pi_t,\cdot)$ as a basis of influence function calculations. Ramsey and Novick (1980) have suggested similar things be done simultaneously for prior, likelihood and loss function.

4.3 *Extensions to Non-Linear Regression*

Consider now a general family of probabilities $\{P_\theta\}_{\theta\epsilon\theta}$ and a parameter $\psi(\theta)$. When does

$$E\{\psi(\theta)\,|\,X\}$$

characterie a prior distribution on Θ? In section 2 we consider the special case of exponential families and linear regression. For exponential families in their natural parametrization and $\psi(\theta)=\theta$, a characterization result can be shown to hold. Thus, for a normal location problem, with scale 1, if the prior is proportional to $\theta^2\, e^{-\theta^2/2}$, $E\{\theta\,|\,X\} = \{X^3+6X\}/\{2X^2+4\}$, and the prior is uniquely characterized by this relation.

However, we cannot settle the uniqueness problem in any generality: For example, if P_θ is a normal location problem $\psi(\theta)$ is a polynomial in θ, then for a normal prior,

$$E\{\psi(\theta)\,|\,X\}$$

is a polynomial in X. We do not know if this characterizes normal priors

4.4. *Matrices*

In exponential families and location problems, we can ask about priors with posterior linearity in terms of matrices. We here discuss the location problem in d-dimensions. When can one find independent θ and X with

$$E(\theta\,|\,\theta+X) = A(\theta+X)+b \tag{4.1}$$

for some $d\times d$ A and $b \in \mathbb{R}^d$? Assume $E|X| < \infty$ with $EX = E\theta = b = 0$. From page 11 of Kagan, Linnick and Rao (1973), (4.1) holds if and only if

$$\phi_X(s)(I-A)\nabla\phi_\theta(s) = \phi_\theta(s)A\nabla\phi_X(s)$$

The matrix A may as well be taken non-singular. All possible A's occur when X is normal, see Jewel (1982) and the references cited there. If A is non singular and $A \neq aI$, then perhaps normality is the only case.

4.5. *Connections with de Finetti's Theorem*

It is possible to give parameter-free versions of some of the characterization results of section two. An elegant classical version is W. E. Johnson's theorem as discussed by Zabell (1982). Imagine a process $\{X_i\}$ taking $k \geq 3$ values. Suppose that $P\{X_{n+1}=i\,|\,X_1,\ldots,X_n\}$ only depends on the number of times i ocurred among $X_1,X_2,\ldots,X_n$. This is necessary and sufficient condition for the law of X_i to be a Dirichlet mixture of multinomials.

This can be generalized to characterize the exchangeable sequences which are conjugate prior mixtures of specified exponential families. It would take us too far afield to try to develop the modern theory of partial exchangeability here. A survey, in the language of Bayesian statistics, can be found in Diaconis and Freedman (1983). In general terms, researchers in this field have found additional notions of "symmetry" which imply that a sequence is a mixture of standard parametric families. We propose that a further condition can be given in terms of how one would predict X_{n+1} given $X_1,X_2,\ldots,X_n$, that will result in characterizations of the mixing measure We will content ourselves with a single example.

Theorem 7 Let X_i $1 \leq i < \infty$ take values in $\mathbb{R}_+$ and satisfy,

$$P\{(X_1, \ldots ,X_n) \epsilon A\} = P\{(X_1, \ldots ,X_n) \epsilon A + x\}$$

for all n and all Borel $A \subset \mathbb{R}^n_+$ with $x \epsilon \mathbb{R}^n$ satisfying $\Sigma x_i = 0$ and $A + x \epsilon \mathbb{R}^n_+$. Then X_i are a scale mixture of exponentials. If in addition,

$$E\{X_2 | X_1\} = aX_1 + b$$

then the mixing measure is a gamma distribution.

Sketch of Proof. It is easily verified that the symmetry condition implies X_i is exchangeable. As usual, the condition still applies when the law of X is conditioned on the tail field T. By deFinetti's theorem, the process conditioned on the tail field is i.i.d.. But this entails for all positive a_1 and a_2

$$P\{X_1 > a_1 + a_2 | T\} = P\{X_1 > a_1, X_2 > a_2 | T\} = P\{X_1 > a_1 | T\} P\{X_1 > a_2 | T\}.$$

So X_i are exponential.

Because of exchangeability, the second conditions gives

$$E\{X_n | X_1\} = aX_1 + b \quad \text{for any} \quad n \geq 2.$$

By the strong law for exchangeable variables the average of $X_2, X_3, \ldots, X_n$ converge to the mean $E_\theta(X_1)$, so we have

$$E\{E_\theta(X_1) | X_1\} = aX_1 + b$$

and this was shown to characterize gamma priors in section two.

REFERENCES

BERG, C (1983). On the Presevation of Determinacy Under Convolution. *Proc. Amer. Math. Soc* (to appear).

CHOQUET, G. (1969). *Lectures on Analysis, Vol. I.* Benjamin, Reading, MA.

DALAL, S (1978). A Note on the Adequacy of Mixtures of Dirichlet Processes. *Sakhyā, A,* **40**, 185-191.

DALAL, S and HALL, J H (1980). On Approximating Parametric Bayes Models by Non Parametric Bayes Models. *Ann. Statist.*, **8**, 664-672.

DALAL, S. and HALL, W. J. (1983). Approximating Priors by Mixtures of Natural Conjugate Priors. *J. Roy. Statist. Soc. B,* **45**, 278-286.

DEVINATZ, A. (1959). On a Theorem of Levy-Raikov. *Ann. Math. Statist.*, **30**, 583-586.

DIACONIS, P and FREEDMAN, D (1983). Partial Exchangeability and Sufficiency. *Sakhyā.* (to appear).

DIACONIS, P and STEIN, C (1983). *Lectures in Decision Theory.* To appear.

DIACONIS, P and YLVISAKER, D (1979) Conjugate Priors for Exponential Families. *Ann. Statist* **7**, 269-281.

DIACONIS, P. and ZABELL, S (1982). Updating Subjective Probability. *J. Amer. Statist. Assoc.*, 822-830

DOOB, J. (1953). *Stochastic Processes.* New York: Wiley.

DUBINS, L. (1983). Bernstein-Like Polynomial Approximation in Higher Dimensions. *Zeit. Wahr* (to appear)

FERGUSON, T. (1955). On the Existence of Linear Regression in Linear Structural Relations. *Univ Calif. Publications in Statistics,* **2**, 143-166.

GILAT, D. (1971). Some Conditions Under Which Two Random Variables are Equal Almost Surely and a Simple Proof of a Theorem of Chung and Fuchs. *Ann. Math. Statist.*, **42**, 1647-1665.

GIRSHICK, M.A. and SAVAGE, L.J. (1951). Bayes and Minimax Estimates for quadratic loss functions. *Proceeding of the Second Berkeley Symposium*, 53-74

GOLDSTEIN, M. (1975). Uniqueness Relations for Linear Posterior Expectations. *J. Roy. Statist. Soc. B,* **37**, 402-405.

GOOD, I. J. (1971) 46,656 Varieties of Bayesians. Letter in *Amer. Statist.*, **25**, 62-63.

JEFFREY, R (1982) *The Logic of Decision.* Chicago University of Chicago Press.

JEWEL, W.S. (1974). Credible Means are Exact Bayesian for Exponential Families. *Astin Bull.* **8**, 77-90

— (1975). Regularity Conditions for Exact Credibility. *Astin Bull.*, **8**, 336-341.

— (1982). Enriched Multinormal Priors Revisited. *Tech. Report*, University of California, Berkeley.

KADANE, J. and CHUANG, D (1978). Stable Decision Problems. *Ann. Statist.*, **6**, 1095-1110

KAGAN, A.M., LINNICK, Yu.V., and RAO, C.R (1973). *Characterization Problems in Mathematical Statistics.* New York: Wiley

LANDAU, H. (1980). The Classical Moment Problem: Hilbertian Proofs. *J. Funct. Anal.*, **38**, 255-272.

LORENTZ, G.G (1966). *Bernstein Polynomials.* University of Toronto Press, Toronto.

LORENTZ, H.C. (1980) *Approximation of Functions.* New York: Holt, Rinehart and Winston.

MORGAN, R.I. (1970). A Class of Conjugate Prior Distributions. *Tech. Rep.* Dep. of Statistics, University of Missouri, Columbia, MO.

NOVICK, M R and HALL, W J. (1965). A Bayesian Indifference Procedure. *J. Amer. Statist. Assoc* **60**, 1104-1117.

RAMSEY, J.O. and NOVICK, M R. (1980). PLU Robust Bayesian Decision Theory. *J. Amer. Statist. Assoc.*, **75**, 901-907.

SHAFER, G (1981). Constructive Probability. *Synthese,* **48**, 1-60

STRAUCH, R.E. (1965). Conditional Expectations of Random Variables Without Expectations. *Ann. Math. Statist.*, **36**, 1556-1559

STEIN, C. (1965). Approximation of Improper Prior Measures by Prior Probability Measures. *Bernoulli, Bayes, Laplace,* (J. Neyman, L. LeCam eds.), New York: Springer-Verlag.

ZABELL, S (1982). W.E. Johnson's "Sufficientness" Postulate. *Ann. Statist.*, **10**, 1091-1099.

ZELLNER, A (1980). On Bayesian Regression Analysis with *g*-Prior Distributions. *Tech. Rep.*, Graduate School Business, University of Chicago

DISCUSSION

F.J. GIRON (*Universidad de Málaga, Spain*)

This paper deals with two main issues. The first one is an extension of the notion of conjugate families outside the exponential family, following the suggestion of Goel and DeGroot (1980) that "... perhaps one should use linear posterior expectation as the defining property of a conjugate distribution" in view of a previous result of the authors in his 1979 paper.

The other important result is theorem 5, namely, that *any* prior can be approximated, in the sense of the weak star topology, by a finite mixture of conjugate distributions.

As stated in the introduction, the authors carry through the first point for the case of a location parameter, opening the way for further generalizations to scale parameters. They succeed in their attempt, for which I congratulate them, and develop a fine theory for the location parameter problem.

However, I want to make two minor points. Actually, the first one cannot really be considered an objection at all. In fact, in their setting for the location parameter problem, posterior linearity only holds for samples of size 1. Of course, outside the regular exponential family, sufficient statistics are not to be expected, so that posterior linearity can only hold for the case X is normally distributed, as proven in Goel and DeGroot (*loc. cit.*)

On the other hand, their definition of conjugate prior for the location parameter problem through F, in terms of linear posterior expectation, broadens the concept of conjugate distributions.

The second point is of a more fundamental nature. Consider the following examples, for which we follow the author's notation

Example 1 Let X be uniformly distributed in $[0,1]$, θ uniformly distributed in $[-1,2]$, independently of X and let $Z = X + \theta$, so that $Z|\theta$ is uniform in $[\theta,\theta + 1]$,

Then the likelihood function $1(\theta|z)$ is

$$l(\theta|z) = I_{[z-1 \cdot z]}(\theta)\,,$$

and the posterior distribution of θ given $Z=z$ is uniform in the interval [max $(z-1, -1)$, min $(z, 2)$], so that the posterior expectation

$$E[\theta|z] = \tfrac{1}{2}\{\max(z-1, -1) + \min(z, 2)\}\,.$$

Example 2 Suppose X follows an exponential distribution ($\lambda = 1$) and is independently distributed in θ, which is distributed as a right-truncated exponential with density

$$p(\theta) = \begin{cases} \mu e^{-\mu(\theta_o-\theta)} & \text{if } \theta \leq \theta_o \\ 0 & \text{otherwise} \end{cases}$$

If $Z = X + \theta$, then the conditional density of the observable Z given θ

$$f(z|\theta) = \begin{cases} e^{-(z-\theta)} & \text{if } z \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the posterior distribution of θ given $Z=z$ is

$$p(\theta|z) = \begin{cases} (\mu+1)\, e^{-(\mu+1)[min(z,\theta_o)-\theta]} & \text{if } \theta \leq \min(z,\theta_o) \\ 0 & \text{otherwise,} \end{cases}$$

and the posterior expectation is

$$E[\theta/z] = \min(z,\theta_o) - \frac{1}{1+\mu}\,.$$

In the first example, the posterior distribution is always uniform so that, according to the classical definition, the family of uniform priors is a conjugate family. However, the

posterior expectation is *not* linear in the observation but *piece-wise linear*. Also note that in this case the boundary values of a (0 and 1) are admissible, and this does not contradict lemma 1 in section 3 of the paper.

The second example exhibits the same behaviour, so that the right-truncated exponential family is conjugate and the posterior expectation is piece-wise linear and continuous.

As a suggestion, piece-wise posterior linearity might serve as a basis for a definition of a conjugate family when the sample space depends on the parameter.

Turning now to the second important result, namely, theorem 5, I had found myself that for a few special families, such as normal, beta, gamma, etc., the above mentioned result was true using the same technique of approximating the constants, or degenerate, distributions by a suitable member of the conjugate family. Yet, I lacked a general proof of the result, which the authors provide in the paper.

This result is also extended in section 3 to the case of a location parameter problem when the underlying distribution is infinitely divisible.

Incidentally, note that if piece-wise linearity is considered, as in examples 1 and 2, then the family of all finite mixtures of uniform priors and the family of finite mixtures of right-truncated, two parameter, exponentials are both weak star dense in the set of all probability measures on the real line.

This theoretical result is also important from a practical view-point as follows from considerations take from the introductory example and from the following result which complements theorem 5.

With a slight change of notation, theorem 5 can be rewritten in the following form

Theorem 5. Let **b** be a conjugate family for the likelihood $f(x|\theta)$. Then for any prior Q on θ and $\epsilon > 0$, there exists weights w_i and $P_i \epsilon$ **b** such that, if $R = \sum_{i=1}^{N} w_i P_i$, $d_p(Q, R) < \epsilon$.

Furthermore if we denote by Q_x, R_x, P_{ix} the posterior distributions of Q, R, P_i, respectively, then there exist $\delta > 0$ such that

$$d_p(Q_x, R_x) < \delta,$$

where

$$R_x(d\theta) = \sum_{i=1}^{N} w_i' P_{ix}(d\theta) \qquad ; P_{ix} \epsilon \mathbf{b}\,,$$

$$w_i' = \frac{w_i f(x|P_i)}{\sum_{i=1}^{N} w_i f(x|P_i)}\,,$$

and $f(x|P_i) = \int_\theta f(x|\theta)\, P_i(d\theta)$ is the predictive density of x given P_i.

D V LINDLEY (*Somerset, UK*)

I would like to use this interesting paper as an excuse to raise the question of what modern mathematics has to say for us statisticians. Diaconis mentions a result in the folklore. I look at this; it does not seem familiar but easily yields to integration by parts. Later we are told that this is correct "up to rigour". But what is rigour? Is integration by parts not rigorous? At one point in my hand-waving proof a limit has to be shown to be zero. Diaconis has to show a set is open. Is the difference material? I have recently read Morris

Kline's "Loss of Certainty" and find that he argues that mathematicians delude themselves in thinking their methods are totally rigorous, and that mathematics should relate more to the real world. Is it true that modern mathematics is just "papering up the cracks" or has it something original to say to us? I have heard John Hammersley express similar doubts concerning one branch, functional analysis. Perhaps these remarks of mine are simply caused by a generation gap and I am too wedded to the mathematics of 1940's Cambridge to understand 1980's Californian. But it would be good to have examples of original contributions to statistics that owe their ideas to later mathematics than mine. (One example, given to me after the discussion, was Stein's brilliant demonstration of the inadmissibility of the sample mean. This result arises from a desire to provide a rigorous proof of an "obvious" result, subsequently shown to be false, but nowhere appears to use mathematics that was not available to Neyman and Fisher.)

Take the case of the exponential family considered in this paper. The original idea was almost simultaneously produced by Pitman, Darmois and Koopmans. Pitman's paper was beautiful and unsophisticated. Koopmans' was heavy, more rigorous and indigestible. Darmois' I cannot remember, but knowing something of French mathematics of the period, it was probably both elegant and rigorous (by the standards of the day). Conjugate families are due to Wetherill and to Raiffa & Schlaifer. Does modern mathematics add substantially to these ideas? The converse type of result, that linearity implies conjugacy, is important because of the restriction it places on the usefulness of the linear tool. (Goel mentioned the importance of this, but does it affect the highly succesful GLIM programme?)

Techniques are not as important as concepts and originality. I am for a bit of "hand waving": it rarely lets one down and in the hands of brilliant people is almost always reliable. But I ask the question in a neutral spirit: has modern mathematics a substantial contribution to make to statistical science? The question is surely an important one. Forgive me, Persi, for using your paper as an excuse for raising the question. You are perhaps the best person to answer it.

J.M. BERNARDO (*Universidad de Valencia*)

The idea that most prior opinions may be well approximated by mixtures of conjugate priors has been tacitily accepted for a long time; I remember, for instance, that the issue was discussed in one of the London University Friday Statistical Seminars back in 1975. However, it is always dangerous in mathematics to take for granted 'obvious' results, which too often are later shown not to be true; thus, the precise statements given in this paper should be most welcome.

On the other hand, the distance between a theorem of existence and a constructive procedure to find the solution may be very large. We now *know* that opinions *may* be well described by some mixture of conjugate priors; I would like the author to comment on the procedures, as general as possible, which he suggests for *actually* expressing opinions in terms of such mixtures.

J. BERGER (*Purdue University*)

My comments are based on a (possibly faulty) memory of the very stimulating talk and ensuing lively discussion. Concerning the plethora of interesting results presented, I have a question, a picky comment, and a more serious comment.

The question concerns the analysis of the coin spinning experiment, in which Professor Diaconis used a mixture of beta priors for the probability p, that a spun coin would land heads, apparently as an illustration of a situation in which a mixture prior was needed. The motivation for this was not completely clear to me, however. Presumably, the probability of a head would be related to some physical difference, d (in, say, weights), between the two sides of the coin. For a completely unknown coin, the prior distribution of p should then have a similar shape to the distribution (among coins) of d, and I wonder why a mixture model for this would seem necessary? I am not trying to argue that single conjugate priors are sufficient - far from it; because of an inherent distrust of conjugate priors, for robustness reasons, I would like to see this as another example where conjugate priors are clearly inappropriate.

Professor Diaconis also discussed using finite mixtures of conjugate priors to approximate an arbitrary prior, and presented some very nice theorems to the effect that this could be done to any desired degree of accuracy. This possibility is being raised in many Bayesian quarters these days, due to the calculational ease of working with finite mixtures of conjugate priors. There is a very serious issue concerning such an approximation, however, namely the issue of whether this good approximation to the prior ensures that the posterior will also be well approximated. I think the answer, in general, is no

To see this, consider the simple situation where $X \sim N(\theta,1)$, and my "true" prior, π_T, is Cauchy (0,1). Let

$$\pi_A = \sum_{i=1}^{m} \lambda_i \pi_i, \qquad \pi_i \text{ being } N(\mu_i, A_i),$$

be an approximating finite mixture of conjugate normal priors. Now it is easy to show that the posterior corresponding to π_T can be drastically different than that corresponding to π_A for specific observations x. Indeed, as $x \rightarrow \infty$, $\pi_T(\theta \mid x)$ becomes approximately equal to a $N(x,1)$ distribution while $\pi_A(\theta \mid x)$ becomes approximately equal to a $N(\mu(x),\varrho)$ distribution, where

$$\varrho = \frac{A^*}{1+A^*}, \qquad \mu(x) = x - \frac{1}{1+A^*}(x-\mu^*),$$

$$A^* = \max_i \{A_i\}, \qquad \text{and } \mu^* = \max_{i: A_i = A^*} \{\mu_i\}.$$

These distributions clearly differ drastically for large x.

Perhaps even more telling is that the approximation can be bad, from a posterior viewpoint, "on the average". For instance, if one were trying to estimate θ under squared error loss, the posterior mean would be used; call it $\delta_T(x)$ or $\delta_A(x)$ for π_T or π_A, respectively. Also let m_T denote the marginal distribution of X, and note that, as $x \rightarrow \infty$, a simple Taylor's expansion shows that

$$m_T(x) = \frac{1}{\pi(1+x^2)}(1+o(I)).$$

Now the "average" performance of δ_T is

$$\int\int (\theta - \delta_T(x))^2 \, \pi_T(\theta \mid x) \, m_T(x) \, d\theta \, dx < 1$$

(since δ_T must have sinaller Bayes risk than the minimax rule), while the average performance of δ_A (with respect to the true prior) is

$$\int\int (\theta - \delta_A(x))^2 \, \pi_T(\theta \mid x) \, m_T(x) \, d\theta \, dx$$

$$\geq \int_K^{\infty} \int_{-\infty}^{\infty} (\theta - \delta_A(x))^2 \, \pi_T(\theta \mid x) \, m_T(x) \, d\theta \, dx$$

$$\cong \int_K^\infty \int_{-\infty}^\infty (\theta-\mu(x))^2 \frac{1}{\sqrt{2\pi}} e^{-1/2(\theta-x)^2} d\theta \frac{1}{\pi(1+x^2)} dx \quad \text{(for large } K\text{)}$$

$$\geq \int_K^\infty \frac{(x-\mu^*)^2}{(1+A^*)^2} \cdot \frac{1}{\pi(1+x^2)} dx = \infty$$

This is just another illustration that the "tail" of the prior can be crucial to Bayesian robustness, and there is simply no way to appropriately approximate a polynomial tail (such as that of π_T) by a finite mixture of exponential tails (such as in π_A).

The final, nitpicky, comment I had concerned the "definition" of conjugate priors given by Professor Diaconis, essentially that given in Diaconis and Ylvisaker (1979). Although I appreciate the elegant characterization that this definition allows, I object to the fact that, if X has a p-variate normal distribution with unknown mean vector θ and known covariance matrix Σ, then a p-variate normal prior for θ (with mean vector μ and covariance matrix A not commuting with Σ) does not qualify under the definition as a conjugate prior. This is such an important situation and so "conjugate" in all other senses, that the definition of Diaconis and Ylvisaker should probably be called something else

A.F.M SMITH (*University of Nottingham*)

Would Diaconis like to comment further on the alternative approach to approximating arbitrary priors by mixtures of natural conjugate priors that was put forward recently by Dalal and Hall (1983)?

REPLY TO THE DISCUSSION

To Professor Bernardo:

I do not believe there are general rules for expressing prior belief Each real problem has to be thought about on its own merits. Of course we have examples of previous analyses, familiar models, and notions of symmetry. We can try to break our problem into pieces such that each piece is similar to a familiar problem, where risks and benefits of particular assumptions are well understood. Mixtures are useful when there is a partition and a believable prior specification given each piece of the partition Not all problems are amenable to such a decomposition, but I think there are enough examples to justify the present analysis

To Professor Berger:

The point of the coin spinning example is that the coin is not "completely unknown". It was selected from a population displaying strong biases in both directions (presumably because of physical differences in the way the edges of coins are finished) A bimodal prior is forced by experience, reflecting (roughly) the results of dozens of coins spun hundreds of times each, together with a bit of physics.

I am delighted with your examples. For me, they emphasize the need for careful statements and honest proofs They *also* show how "tails can wag dogs" and encourage me to find a metric for the weak star topology which is not so affected by what happens off at infinity. I've put some comments toward this end in the body of the paper

On your comment about normal priors: as indicated in the section on 'matrices' I believe the only time we can have matrix scaling is for normal location problems; here, as elsewhere, the normal is special and I'd prefer to have a unified theory of exponential families and call the priors "augmented conjugate" as Jewel does.

To Professor Giron

Our paper tries to characterize the usual notion of conjugate priors and shows how, sometimes, any prior can be well approximated by mixtures. This works reasonably well for exponential families and less generally for location (and other families). Your examples seem like just the direction to head in: find tractable, understandable classes of priors and show that we can approximate any prior

To Professor Lindley:

Why do I try to prove theorems instead of working in the older tradition of "+ ..."? Three reasons are correctness, communication, and aesthetics Let me elaborate: On correctness - given my choice, I'd rather be "brilliant" than careful. Given my limitations, I simply can't see another route to "getting it right" than trying it out in real examples and trying to do the mathematics carefully. There are certainly great examples of the "+ ..." school - you don't even have to be Bayesian (e.g., Fisher or Cox) but there are so many more examples where the magic insight is missing and all one is left with is useless heuristics.

On communication - one of my delights at the Valencia meetings was meeting many young statisticians who learned modern mathematical statistics. When push come to shove, we would constantly revert back to standard mathematical usage to clarify discussions. This can be contrasted with trying to figure out an article written in "+ ..." language: there we find "theorem" with often no assumptions or clear conclusions; I find it frustrating to try to figure out what someone else has in mind without hints, and feel that someone should do their homework before publishing One of the achievements of modern mathematical statistics is a common universal language, free of the mysteries of "inference".

On aesthetics - I find modern mathematics beautiful The best work in mathematical statistics has the same appeal So, if you like, I prove theorems because it makes me happy.

Professor Lindley asks for contributions of modern mathematics to statistical science. Many constructions that implicitly involve infinite dimensions are nowadays correctly and routinely used after very rocky beginnings. For example, robustness with its notion of influence curve (a function space derivative), the whole cannon of techniques from weak convergence and invariance principles as used in connection with Kaplan-Meier and Cox modelling

Turning to Bayesian statistics we have the complete class theorems which have helped bridge the gap between decision theoretic and Bayesian methods - these are impossible without some topology and functional analysis. The work of Savage and de Finetti (and many of the rest of us) on de Finetti's theorem rests solidly on functional analysis - indeed the Hewitt-Savage paper is often cited by functional analysts as a first "Choquet theorem".

Finally, let us consider what is perhaps the finest synthesis of the Bayesian and decision theoretic schools: Stein's results and the whole sea of related work. It's true, that nowadays there are reasonably simple variants of Stein's result. *But* the result was born, bred, nurtured and raised in the heart of complete class theorems and invariance theory. The many far reaching extensions by Berger, Brown, Efron, Morris and a host of others use tools from every area of mathematics - from potential theory, through difussion through differential equations. One of the profound results of this work is an intimate connection between all of these areas, so tools developed one place can be used in any other. Thanks to this work many other simultaneous estimation problems (e.g., many Poisson variables) can be handled and thought about as they arise, without having to consult a few Guru's.

One virtue of mathematical language is its power to unify. Take Professor Lindley's question about putting the assumption "θ is open" in place of "the bounday terms vanish", in Theorem 3. The openness of θ also serves in Theorems 2 and 4; without it, we have a spate of extra conditions, one for each theorem; with it a modicum of order. There is also a practical reason: I believe openness is easier to check in several dimensions, while understanding the analog if integration by parts, Stoke's theorem, is hard.

As Professor Lindley points out, brilliant work can be done without modern mathematics. Further, the use of mathematical machinery does not guarantee correctness. To take the case suggested by Professor Lindley, consider the Koopman-Pitman-Darmois theorem: roughly, if a statistical problem admits finite dimensional sufficient statistics, then it is an exponential family. To make this precise, one needs a precise notion of "statistic" because there are continuous 1-1 functions from $\mathbb{R}^n$ into $\mathbb{R}$, some smoothness assumptions are needed: but how much smoothness? The earlier writers didn't worry about this in too detailed a way, so that it is difficult to determine if statistics involving absolute values are admissible. Later writers, like Dynkin and Brown did worry about it, and used substantial mathematical tools. Yet major theorems in their papers are simply wrong. The matter seems to be usefully settled, see Hipp (1974) which contains the relevant references.

To Professor Smith:

After our manuscript was finished, the very relevant paper by Dalal and Hall (1983) appeared. This contains a more general framework and some of the examples (Bernstein-like approximations) are developed in great detail. Additionally, they treat some of the problems suggested in section 4. Dalal has used mixtures of priors in other ways over the past 10 years (see their bibliography)

It would be interesting to trace the history of the use of mixtures. At the conference, George Barnard said he had used mixtures of priors in the 1940's, and several other groups seem to have discussed the possibility. Most of the results presented here were developed in 1974-75; they were written up because of the systematic prodding of Jay Kadane.

REFERENCES IN THE DISCUSSION

GOEL, P.K. and DEGROOT, M.H. (1980). Only normal distribution have linear posterior expectations in linear regression. *J. Amer. Statist. Assoc.* **75**, 895-900

HIPP, C. (1974) Sufficient statistics and exponential families, *Ann. Statist.* **2**, 1283-1292.

BAYESIAN STATISTICS 2, pp. 157-182
J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (Eds.)
© Elsevier Science Publishers B.V. (North-Holland), 1985

Direct Subjective-Probability Modelling Using Ellipsoidal Distributions

JAMES M. DICKEY and CHONG-HONG CHEN
State University of New York at Albany

SUMMARY

Interactive elicitation and fitting methods previously proposed for assesing Bayesian prior distributions in normal linear sampling are here extended to assessment of general subjective-probability models involving ellipsoidal distributions and concomitant variables without explicit sampling models. Ellipsoidal (sometimes called "elliptical") distributions are related to their marginal and conditional distributions in simple ways that allow assessment through univariate quantiles. Dependence on concomitant variables is modeled through assessable functional forms in the ellipsoidal location and scale parameters. New families of tractable spherical and ellipsoidal distributions are given based on Carlson's two-way multiple hypergeometric function. These generalize the multivariate normal and multivariate-t distributions. Assessment methods are developed in detail for scale mixtures of multivariate normal distributions.

Keywords: ASSESSMENT; ELICITATION; ELLIPSOIDAL DISTRIBUTIONS; ELLIPTICAL DISTRIBUTIONS; EXPERT OPINION; MULTIPLE HYPERGEOMETRIC FUNCTIONS; SUBJECTIVE PROBABILITY.

1. INTRODUCTION

It was asserted in the first Valencia conference (Dickey 1980) that subjective-probability modelling of expert opinion is more widely applicable than Bayes-theorem inference methods. Bayes-theorem methods require the expert assessment of "prior" probabilities, that is, prior to modeled statistical data. Often, however, important decisions must be based *primarily* on beliefs of experts, for example when proper sample data are not available, or not feasible to obtain. Commonly, the available data refer to something rather different than the context of the decision problem at hand. Or, there are times when it cannot honestly be said that the data are sampled according to a recognized formal statistical model.

In the end, it must be admitted that a *responsible* decision maker (same person here as the expert) will prefer to use his or her actual posterior opinion to some formal Bayes-theorem-generated posterior distribution, *especially* when the final opinion disagrees with the Bayes thing. Of course, the opinion must be formed in light of all available data and all economically obtainable data, and in light of feasible Bayes-theorem calculations and other probability calculations. It is difficult to include in a formal likelihood function all relevant information imparted by a body of data and its reporting process (Dawid and Dickey, 1977). It is difficult, posterior to data, to assess a prior distribution by pretending to consult ones state of mind prior to experiencing the report of the data. And then, what