QUANTIFYING PRIOR OPINION

by
Persi Diaconis
and
Donald Ylvisaker

TECHNICAL REPORT NO. 207
OCTOBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
QUANTIFYING PRIOR OPINION

by
Persi Diaconis
and
Donald Ylvisaker

TECHNICAL REPORT NO. 207
OCTOBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
QUANTIFYING PRIOR OPINION

Persi Diaconis          Donald Ylvisaker
Department of Statistics Department of Mathematics
Stanford University     U.C.L.A.

ABSTRACT

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.
1. **Introduction.**

definetti 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 $p \in [0,1]$ 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

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

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 \( \beta(a,b;p) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} p^{a-1} (1-p)^{b-1} \). 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
\[
f(p|S_n=k) = \beta(a+k,b+n-k;p).
\]

3. **Subjective Bayesians.** (Ramsey, deFinetti, 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 always 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

\[ \frac{1}{2} \{ \beta(10,20;p) + \beta(20,10;p) \} . \]

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

\[
\begin{align*}
I &= \beta(1,1) \\
II &= .5\beta(10,20) + .5\beta(20,10) \\
III &= .5\beta(10,20) + .2\beta(15,15) + .3\beta(20,10)
\end{align*}
\]

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

\[ f(p|S_n = k) = \sum_{i=1}^{n} w'_i \beta(a_i + k, b_i + n-k; p) . \]

Here the weights \( w'_i \) depend on \( n \) and \( k \) in a simple way.

\[ w'_i = c w_i \int p^k (1-p)^{n-k} \beta(a_i, b_i; p) dp \text{ with } c \text{ chosen so } \sum w'_i = 1. \]

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 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 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)

FIGURE 3. The three posteriors after observing 11 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 density \(\beta(a_i, b_i; p)\) is close to a point mass at \(a_i/(a_i+b_i)\) 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 \(\{w_i, a_i, b_i\}_{i=1}^n\) such that

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

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 \( \sum \binom{n}{i} \frac{x}{i} (1-x)^{n-i} \)
\[
= \frac{1}{n+1} \sum_{i=0}^{n} f \left( -\frac{i}{n} \right) \beta(i+1, n-i+1; p); \text{ this has weights } \frac{f(i/n)}{n+1} = w_i.
\]
The usual non-probabilistic proof of the Weierstrass 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, \( \omega_f(1/\sqrt{n}) = c/\sqrt{n} \) 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 is of order \( \omega_f \left( \frac{1}{n} \right) \). (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, 3 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.
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

$$E(p|S_n = k) = ak + b \quad k = 0, 1, 2, \ldots, n.$$ 

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 \( \mu \). 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 \( \Theta \) is a convex set. It is called the natural parameter space. Throughout we assume that \( \Theta \) is non-empty and open. The exponential family \( \{ P_\theta \} \) of probabilities through \( \mu \) is defined as

\[
dP_\theta = e^{x^T \theta - M(\theta)} \mu(dx) \quad \theta \in \Theta.
\]

As usual, the expectation under \( P \) can be determined by differentiating \( M \):

\[
E_\theta(X) = M'(\theta).
\]

Define a family \( \{ \overline{\Pi}_{n_0, x_0} \} \) of measures on \( \Theta \) by

\[
\overline{\Pi}_{n_0, x_0}(d\theta) = e^{n_0 x_0^T \theta - n_0 M(\theta)} d\theta, \quad n_0 \in \mathbb{R}, \quad x_0 \in \mathbb{R}^d.
\]

If \( \overline{\Pi}_{n_0, x_0} \) can be normalized to a probability \( \Pi_{n_0, x_0} \) on \( \Theta \), 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 $\Theta = \mathbb{R}^d$, $\prod_{n_0} x_0^{(\Theta)} < \infty$ if and only if $n_0 > 0$, $x_0 \in \mathbb{R}$.

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

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 Jewell (1974a,b).

Theorem 3. If $\theta$ has the distribution $\prod_{n_0} x_0$ for $n_0 > 0$ and $x_0 \in \mathbb{R}$ 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 $P_\theta$. Indeed, if $\prod_{n_0} x_0$ is the prior for $\theta$, the posterior density is $\prod_{n_0+n, \frac{n_0 x_0 + \bar{x}}{n_0+n}}$ with $\bar{x}$ the mean of the sample. Theorem 3 yields

$$E(E_\theta(X) \mid X_1 \cdots X_n) = \frac{n_0 x_0 + \bar{x}}{n_0+n}.$$

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 in theorem 3 is integration by parts: consider the one-dimensional case. Then, as usual, $E_\theta(X) = M'(\theta)$ and
\[
\int M'(\theta)e^{n_0x_0\theta-n_0M(\theta)} \, d\theta = \frac{e^{n_0x_0\theta-n_0M(\theta)}}{n_0} \left[ e^{x_0\theta} + x_0 \right] e^{n_0x_0\theta-n_0M(\theta)} \, d\theta.
\]

The boundary terms vanish because \( \Theta \) 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 \( P_\theta \) and suppose the support of \( \mu \) contains an open interval in \( \mathbb{R}^d \). If \( \theta \) has a prior distribution \( \tau \) which is not concentrated at a single point and if

\[
E\{E_\Theta(X)|X\} = aX + b
\]

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

\[
\tau(d\theta) = ce^{-1}b\cdot\theta-a^{-1}(1-a)M(\theta)d\theta.
\]

**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 \( \Theta \) into a prior on any given parameter space to preserve linearity.
Let $\psi: \Theta \to \mathbb{R}^d$ be a diffeomorphism with range $\Theta_\psi$ and inverse $\psi^{-1}$. This transforms Lebesgue measure via multiplication by a Jacobian $\psi'$. The image of the prior $\prod_{n_0,x_0}^{\frac{1}{n_0x_0\psi^{-1}(t)-n_0M(\psi^{-1}(t))}} dt$ for $t \in \Theta_\psi$.

If $t$ is taken as parameter, so the family becomes

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

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

$$E\{E_t(X) | X_1 \cdots X_n\} = \frac{n_0x_0x_n^{n-1}}{n_0+n}.$$

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. Morrie 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 $\Theta$ be the natural parameter space of a $d$-dimensional exponential family. For any probability $\pi$ on $\Theta$, and any $\varepsilon > 0$ there are weights $w_i$, and $(n_i,x_i)$, $n_i > 0$, $x_i \in \mathcal{X}$ such that if
\[
\pi(d\theta) = \sum_{i=1}^{N} w_i c(n_i,x_i) e^{\theta_n x_i - n_i M(\theta)} d\theta,
\]
then
\[d(\pi,\tilde{\pi}) < \varepsilon\]
where $d$ is the Prohorov metric.

Proof. It is well known, and easy to argue directly, 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
\[(2.1) \quad c(n_0,x_0) e^{n_0 x_0 \theta - n_0 M(\theta)} d\theta .\]

To see this, differentiate (2.1), the maximum occurs at the unique value of $\theta$ 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 $\theta$. Finally, for any $\theta_0 \in \Theta$, $M'(\theta_0) = E_{\theta_0}(X) \in \mathcal{X}$, so for any $\theta_0$, there is an $x_0 \in \mathcal{X}$ such that $M'(\theta_0) = x_0$. This completes the argument. $\Box$
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, 1982). 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 6 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 $\theta$ 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

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

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,

$$(3.1) \quad E(\theta|Z) = \frac{n}{n+1} Z .$$
To see this, let $X_1, \ldots, X_n$ be independent of $X$ and each other so that $S_n = X_1 + \cdots + X_n$ has the same distribution as $\Theta$. Then

$$\frac{n+1}{n} E(S_n | S_n + X) = E(S_n | S_n + X) + \frac{1}{n} E(S_n | S_n + X)$$

$$= E(S_n | S_n + X) + E(X | S_n + X) = E(S_n + X | S_n + X) = S_n + X.$$  

For example, if $X$ and $\Theta$ are independent Cauchy variables, $E(\Theta | X+\Theta)$ 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

(3.2) $E(\Theta | Z) = aZ + b$ for real $a$ and $b \in \mathbb{R}^d$.

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 $0 < a < 1$. The following lemma determines what happens at the boundary cases.

**Lemma 1.** Let $X$ and $\Theta$ be independent random variables and suppose $E(|X|) < \infty$. Then $E(\Theta | X+\Theta) = b$ if and only if $\Theta$ is a.s. constant and $E(\Theta | X+\Theta) = (X+\Theta) + b$ if and only if $X$ is a.s. constant.
Proof. The lemma follows from a result in Doob (1953, pg. 314). To bring things into that framework, let \( Y = X + \theta \). Without loss of generality \( b = 0 \). Under the first assumption \( 0 = \mathbb{E}(\theta | X + \theta) = \mathbb{E}(\theta + X | X + \theta) - \mathbb{E}(X | X + \theta) \), so \( \mathbb{E}(X | Y) = Y \). Of course, \( \mathbb{E}(Y | X) = X \). This says that \( X \) and \( Y \) form a 2-term martingale in either order, and this is just what Doob shows is impossible when \( \mathbb{E}|X| < \infty \). The argument under the second assumption is similar. For related theorems see Girshick and Savage (1951, pg. 1653) or Gilat (1971). \( \square \)

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 \( \lambda_{2n} = \int (x_1^{2n} + \ldots + x_d^{2n}) dF \).

**Theorem 6.** Let \( X \) and \( \theta \) be independent \( d \)-dimensional random vectors with neither \( X \) nor \( \theta \) a.s. constant. Assume \( \mathbb{E}|X_i| < \infty \) for \( i = 1, \ldots, d \) and that either

a) the characteristic function of \( X \) has no zeros or

\[
\lim_{n \to \infty} \sum_{n=1}^{\infty} \lambda_{2n}^{-1/2n} = \infty .
\]

If

(3.3) \( \mathbb{E}(\theta | X + \theta) = a(X + \theta) + b \).

Then \( 0 < a < 1 \) and the distribution of \( \theta \) is uniquely determined.

**Proof.** We begin with \( d = 1 \) and show first that \( \mathbb{E}|\theta| < \infty \). 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

(3.4) \[ E(\theta | X+\theta) = \frac{a}{1-a} E(X | X+\theta). \]

Use this in (3.3) to find

(3.5) \[ \theta = \frac{1}{1-a} E(X | X+\theta) - X. \]

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

When \( X \) and \( \theta \) 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 \( \theta \) satisfy

(3.6) \[ (1-a)\phi_\theta'(t)\phi_X(t) = a\phi_X'(t)\phi_\theta(t) \text{ for all } t. \]

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

(3.7) \[ \phi_\theta(t) = \frac{a}{\phi_X(t)^{1-a}}. \]

Observe first that for (3.7) to hold with neither \( X \) nor \( \phi \) constant, it must be that \( \frac{a}{1-a} > 0 \) and so \( 0 < a < 1 \). Now if \( \phi_X \) never vanishes, \( \phi_\theta \) 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 \( \theta \) can be
computed from (3.7) and satisfy the same determinedness condition. The proof for \( d = 1 \) is complete.

For \( d > 1 \), the finiteness of \( E|\theta_1| \) follows readily by the arguments used earlier. Take \( b = 0 \) again without real loss. Now let \( \rho \in \mathbb{R}^d \) and find from (3.3) that

\[
E(\rho \cdot \theta | X + \theta) = a \rho(X + \theta) = E(\rho \cdot \theta | \rho \cdot (X + \theta)) .
\]

If the characteristic function of \( X \) has no zeros then the same is true of the characteristic function of \( \rho \cdot X \). Hence the one-dimensional version of the theorem implies that the distribution of \( \rho \cdot \theta \) is uniquely determined and so therefore is the distribution of \( \theta \). If the distribution of \( X \) satisfies b), the inequality \( |\rho \cdot X|^{2n} \leq m(x_1^{2n} + \ldots + x_d^{2n}) \) with \( m \) depending on \( \rho \) and \( d \) but not on \( n \) or \( X \), implies that the distribution of \( \rho \cdot X \) is determined by its moments. The proof of the theorem is completed by another application of the one-dimensional version. \( \square \)

Remark 1. Here is an example of independent random variables \( X \) and \( \theta \) 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 6 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

\[
\frac{3 \cdot 4^3}{\pi} \left( \frac{\sin \frac{x^4}{4}}{x} \right)^4 < \infty \quad x < \infty .
\]

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 \(\phi_\theta\) which equals \(\phi_X\) on \((-1,1)\) and has period 2 is the characteristic function of an arithmetic distribution. By construction \(\phi_X^* \phi_\theta = \phi_X^* \phi_X\), 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 \(\theta\) 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 \(\phi \ast \phi\). The function \(\psi(t) = \sum_{n=1}^{\infty} \phi(2^nt)\) 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 \(\theta\), we have an example with moments of all orders where \(E[\theta|X+\theta] = \frac{X+\theta}{2}\) but the law of \(\theta\) is not unique.

Remark 2. In theorem 6 we have used Carleman's 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 \(\theta\) is independent of \(X\) and satisfies

\[E(\theta|X+\theta) = \frac{1}{n} (X+\theta)\quad \text{for any fixed integer } n \geq 2,\]

then the distribution of \(\theta\) is uniquely determined.
Proof. From (3.7), $\phi_X(t) = \phi_\theta(t)^{n-1}$ in a neighborhood of 0. This determines the moments of $\theta$. We want to show that $\theta$ is determined by its moments. If not, then, by a fundamental theorem of moment theory, see Theorem 3 of Landau (1980), for any real $t$, there is a probability $\psi$ with all the same moments as $X$, and so the same distribution as $X$, but $\psi^*n$ 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. \[\defend\]

If it were true that being determined by moments was inherited by convolutions, general rational values of $a$ could be handled; for from $\phi_X^n = \phi_\theta^n$ in a neighborhood of zero, and $X$ determined would follow $\psi_X^n$ determined and then $\phi_\theta$ determined by the above argument. Christian Berg (1983) has provided a probability $\mu$, 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 $E|X_1| < \infty$, then the argument used in the introduction to this section shows that $S_n/n$ is a backward martingale. Here $S_n = \sum_{i=1}^n X_i$ and the martingale property is

$$
(3.8) \quad E\left(\frac{S_n}{n} | S_{n+1}, S_{n+2}, \ldots\right) = \frac{S_{n+1}}{n+1}, \quad n = 1, 2, \ldots
$$

using obvious generalizations of the argument in Theorem 5 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 $\phi_i$ is the characteristic function of $X_i$, (3.8) holds if and only if for all $k \geq 2$,

$$\phi_i^k \prod_{j=1}^{k-1} \phi_j(t) = \frac{\phi_k(t)}{k-1} \left( \prod_{j=1}^{k-1} \phi_j(t) \right)' = \phi_{k-1}'(t) \prod_{j=1}^k \phi_j(t) \text{ for all } t.$$ 

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), $\frac{n}{n+1} \in S_F$ 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 $a \in (0,1)$ so $S_F = (0,1)$. Conversely, if $F$ has a mean, $\phi_F(t) \neq 0$ and $S_F = (0,1)$, Theorem 6 implies $F$ is infinitely divisible.

Remark 5. When can any prior be approximated by mixtures of conjugate priors in a location parameter setting? If $X$ is infinitely divisible then the conjugate priors have characteristic functions of the form $e^{-i\theta \phi}$. As $\alpha \to 0$, this approaches a point mass at $\mu$. 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 example, we note the following. If $X$ has a beta density
with parameters $a$ and $b$, and $\theta$ has a beta density with parameters $a+b$ and $c$, then

$$E(\theta|X) = \frac{cX\theta}{b+c} + \frac{b}{b+c}.$$
4. **Some Research Problems.**

The material discussed above suggests many questions, both technical and philosophical. These are discussed here, under the headings: What are we doing: when are two priors close?, Stability, extensions to non-linear regression, matrices and connections with de Finetti's 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 \( \{P_\theta\}_{\theta \in \Theta} \) and loss function \( L \). Suppose that \( \pi_t \) represents a true prior and \( \pi_a \) an approximation. An observation \( x \) yields two posteriors \( \pi_t^x \) and \( \pi_a^x \). Each of these will result in certain decisions ("Bayes rules") \( \delta_t(x) \) and \( \delta_a(x) \). Here \( \delta_t \) minimizes the risk \( R(\pi_t, \delta) = E_x[L(\theta, \delta(x))] \). The difference in risk, if \( \delta_a \) is used instead of \( \delta_t \), 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 $\theta$ 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 $\rho(\cdot, \cdot)$ be a metric between probabilities. Consider

$$\rho^*(\pi_t, \pi_a) = E_t \{ \rho(\pi^X_t, \pi^X_a) \}$$

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 $\rho$ is taken as the Hellinger distance:

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

where $p_i$ is the density of $P_i$ with respect to the dominating measure $\lambda$, then

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

holds for some constant $c$, depending on $L$ and perhaps $\pi_t$, but not on $\pi_a$. Thus, approximation in $\rho^*$ 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 \( \rho \) 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 opens a philosophical
can of worms. Recent work by Shafer (1981), Jeffrey (1982) and Diaconis
and Zabell (1982) emphasizes the constructive nature of forming a prior —
one of us has 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 election.

One way around these difficulties is the conceptually difficult task of
thinking "will it matter if I refine my prior further". Often, in well tra-
velled 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 update 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 \( \{P_{\theta}\}_{\theta \in \Theta} \) and a parameter \( \psi(\theta) \). When does

\[
E(\psi(\theta) | X)
\]

determine a prior distribution on \( \Theta \)? In section 2 we considered 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 \( P_{\theta} \) is a normal location problem and \( \psi(\theta) \) is a polynomial in \( \theta \), 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 \( \theta \) and \( X \) with
(4.1) \[ E(\theta | \theta + X) = A(\theta + X) + b \]

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

\[ \phi_X(s)(I-A)\varphi_\theta(s) = \phi_\theta(s)A\varphi_X(s). \]

The matrix \( A \) may as well be taken non-singular. All possible \( A \)'s occur when \( X \) is normal, see Jewell (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 deFinetti'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 \( i \) occurred among \( X_1, X_2, \ldots, X_n \). This is a 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

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

for all $n$, and all Borel $A \subset \mathbb{R}_+^n$ with $x \in \mathbb{R}_+^n$ satisfying $\sum x_i = 0$ and $A+x \subset \mathbb{R}_+^n$. 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 de Finetti'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 condition 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


