PRIORS WITH LINEAR POSTERIOR EXPECTATION

BY

PERSI DIACONIS and DONALD YLVISAKER

TECHNICAL REPORT NO. 102
OCTOBER 14, 1977

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Priors With Linear Posterior Expectation

by

Persi Diaconis and Donald Ylvisaker

Stanford University
and
The University of California, Los Angeles

TECHNICAL REPORT NO. 102

OCTOBER 14, 1977

1 AMS 1970 Subject Classification: 62 E10, 62 A15

Key Words and Phrases: Conjugate priors, linearity of regression, Bayesian analysis, characterization theorems, exponential families, credibility theory.

2 This author's research was supported in part by NSF Grant MPS 74-21416 and by the Energy Research and Development Administration under contract EX-76-C-03-0515.

3 This author's research was supported in part by NSF Grants MPS 72-4591, MPS 75-7120 and MCS 77-02121.
Abstract.

Let $X$ be a random vector with distribution depending on a parameter $\theta \in \Theta$. We investigate the characterization of prior measures on $\Theta$ which have linear posterior expectation of the mean parameter of $X$: $\mathbb{E}(\mathbb{E}(X|\theta)|X=x) = ax+b$. Our results lend to characterizing properties of all the usual conjugate priors for multivariate exponential families. We also describe which hyperparameters permit such conjugate priors to be proper. Similar results are derived for location parameter problems.
1. **Introduction**

Consider an observable random vector \( X \) with distribution depending on a parameter \( \theta \in \Theta \). In many statistical contexts, estimation of the mean parameter is of interest. If a prior distribution on \( \Theta \) is available and squared error is used as loss, the Bayes estimate, given \( X = x \), is the posterior expectation of the mean parameter. The estimate is particularly simple when the posterior expectation of the mean parameter is a linear function of \( x \):

\[
(1.1) \quad \mathbb{E}(E(X|\theta)|X=x) = ax + b.
\]

This paper deals with a characterization of prior distributions such that (1.1) holds for some scalar \( a \). The characterization problems discussed are of interest in the following two contexts.

**Conjugate Priors in Bayesian Inference.** Modern Bayesian statistics is dominated by the notion of conjugate priors, and yet this notion is not crisply defined. The usual definition is that a prior is conjugate if the posterior density is proportional to the prior density (for example, Lindley [1972], pg. 22-23). A family of priors is often called conjugate if it is closed under sampling (Raiffa and Schlaifer [1961], pg. 43-57, say). An example will show a weakness in these definitions. Let \( S_n \) be the number of heads in \( n \) independent tosses of a coin with unknown parameter \( p \). The accepted family of conjugate priors for \( p \) is the beta family...
with densities \( f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \alpha > 0, \beta > 0 \). Let \( h \) be any positive bounded measurable function on the unit interval and observe that a prior density proportional to \( h(p)f(p; \alpha, \beta) \) leads to a posterior density of \( p \), given \( S_n = x \), proportional to \( h(p)f(p; \alpha+x, \beta+n-x) \). Thus the family \( \{h(\cdot)f(\cdot, \alpha, \beta)\} | \alpha > 0, \beta > 0, \) \( h \) positive bounded measurable\) with each member normalized to be a prior density, is both closed under sampling and has the property that the posterior density is proportional to the prior density. Evidently the defining properties cited are imprecise. Beta priors satisfy (1.1). That is, there are numbers \( a_n, b_n \) such that

\[
E[p | S_n = k] = \frac{\int_0^1 p^{k+1}(1-p)^{n-k}f(p, \alpha, \beta)dp}{\int_0^1 p^k(1-p)^{n-k}f(p, \alpha, \beta)dp} = a_n^{k+b_n}
\]

holds for \( k = 0, 1, 2, \ldots, n \). A principle result of this paper is that, subject to regularity conditions, the conjugate priors typically used for an exponential family satisfy and are characterized by (1.1). The regularity conditions allow such standard examples as the normal prior for normal location, the gamma prior for gamma scale, Wishart prior for normal covariance, and the beta prior for the negative binomial. The Dirichlet prior for the multinomial is also suitably characterized, but this fact requires a separate treatment. While the assumption of (1.1) offers computational convenience and simplicity, we present no philosophical argument in its behalf. It is advanced solely as a property which singles out the customary prior distributions.
Credibility Theory and Linear Bayesian Analysis. Linear Bayes prediction has been used since 1920 by the actuarial professional under the heading of credibility theory (Kahn [1975]). Our Theorem 2, that (1.1) holds for exponential families with their customary conjugate priors, is a rigorous treatment of some recent results of Jewell [1974b] on what is there termed exact credibility. In work unconnected with credibility theory per se, Ericson [1969], [1970] noted that when (1.1) holds, \(a\) and \(b\) can be given expression in terms of the means and variances of the underlying distributions. Independently, Hartigan [1969] made essentially the same observation and went on to use the \(a\) and \(b\) so determined to form a linear Bayes predictor. Efron and Morris [1973] have an extension of the empirical Bayes approach they developed for normal location problems to situations where the Bayes estimate is linear Bayes. Now in fact, when Theorems 3, 4, 5 and 6 below are in force, they imply that the assumption of (1.1) for fixed \(a\) and \(b\) is exactly the assumption of a specific prior distribution.

The characterization theorems here have been given previously, in special cases. Johnson [1956], [1967] characterized the gamma prior for Poisson mean, while Goldstein [1975] extended Ericson's result to obtain characterizations in the location parameter setting through the use of moments. Moreover our theorems are closely related to results found in Chapter 5 and 6 of Kagan, Linnik and Rao [1973]. The results given here are considerably more general and, in some cases, more precise than those previously found.
Sections 2, 3 and 4 of this paper are devoted to exponential families. Section 2 studies the usual notion of conjugate prior and establishes (1.1) under mild regularity conditions. Moreover, Theorem 1 gives precise conditions on the "hyperparameters" of the conjugate prior to guarantee integrability (see Novick and Hall [1965] in this connection). Section 3 is devoted to a proof that (1.1) characterizes conjugate priors when the observation space is sufficiently rich. Section 3 also comments on the possibility of (1.1) holding when \( \alpha \) is a matrix. In Section 4 we consider problems particular to the case of a discrete observation space. Attention is shifted in Section 5 to another rich family of distributions of \( X|\theta \) — those which arise in the location parameter problem. The main result there states that if (1.1) holds and if \( X|\theta = 0 \) has a non-vanishing Fourier transform or satisfies a moment condition, then the prior distribution of the location parameter \( \theta \) is uniquely determined. In particular, the unique prior in the Cauchy location problem is a Cauchy distribution.

2. Conjugate Priors in Exponential Families.

This section contains requisite notation and terminology associated with a \( d \)-parameter exponential family of distributions. Depending on the setting, Theorem 1 gives sufficient or necessary and sufficient conditions on the "hyperparameters" of a conjugate prior distribution for it to be proper. Theorem 2 then establishes linear posterior expectation under regularity conditions. Theorem 2 has been in the folklore of the subject and a full proof for the 1-dimensional case has recently appeared in Jewell [1974a], [1975]. The section closes with a brief Bayesian interpretation of Theorems 1 and 2.
Start with a fixed $\sigma$-finite measure $\mu$ on the Borel sets of $\mathbb{R}^d$. Consider the convex hull of the support set of the measure $\mu$, and then let $\Theta$ be the interior of this convex set. It will always be assumed that $\Theta$ is a non-empty open set in $\mathbb{R}^d$, so that the observation set is genuinely $d$-dimensional. For $\theta \in \mathbb{R}^d$, define $M(\theta) = \ln \int e^{\theta^T x} \mu(x)$ and let $\Theta = \{ \theta | M(\theta) < \infty \}$. The standard use of Hölder's inequality in this context shows that $\Theta$ is a convex set—it is called the natural parameter space. It is further assumed that $\Theta$ is a non-empty open set in $\mathbb{R}^d$—in the terminology of Barndorff-Nielson [1970], we restrict attention to regular exponential families. The openness of $\Theta$ is indeed a regularity condition on the measure $\mu$—one which is employed crucially in Theorem 2.

The exponential family $\{P_\theta\}$ of probability measures through $\mu$ is determined by

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

Expectation under $P_\theta$ will be denoted by $E_\theta$ or $E(\cdot | \theta)$. Now suppose $X$ is a random vector with distribution $P_\theta$. Then if one differentiates the identity $\int_{\Theta} dP_\theta(x) = 1$ in $\theta$ and makes admissible interchanges of differentiation and integration, one finds

$$1) \quad E(X|\theta) = E_\theta(X) = \nabla M(\theta) = \left( \frac{\partial M(\theta)}{\partial \theta_1}, \ldots, \frac{\partial M(\theta)}{\partial \theta_d} \right)^T = (M_1(\theta), \ldots, M_d(\theta))^T,$$

(2.2)

$$2) \quad E_\theta(X) - \nabla M(\theta))(X - \nabla M(\theta))^T = M''(\theta) = \left( \frac{\partial^2 M(\theta)}{\partial \theta_i \partial \theta_j} \right)_{i,j=1}^{d} = (M_{ij}(\theta))^d_{i,j=1}.$$
Because $\mathcal{F}$ is assumed open in $\mathbb{R}^d$, the Hessian $M''(\theta)$ must be positive definite at each $\theta$ - for otherwise there is a $\theta_0$ and a vector $c \neq 0$ so that $c'(X - \nabla M(\theta_0)) = 0$ a.s. $P_{\theta_0}$ and then a.e. $\mu$. Furthermore from (2.2.i), $\nabla M(\theta)$ must be in the convex hull of the support of $\mu$. It is then easy to see that $\nabla M(\theta)$ cannot be a boundary point of this convex set, so $\nabla M(\theta) \notin \mathcal{F}$ for each $\theta \in \Theta$.

Recall that $\Theta$ is to be a non-empty convex open set in $\mathbb{R}^d$ and let $d\theta$ denote the Lebesgue measure on the Borel sets of $\Theta$. Define a family $\{\tilde{\pi}_{n_0,x_0}\}$ of measures on the same space according to

$$d\tilde{\pi}_{n_0,x_0}(\theta) = e^{n_0x_0 \cdot \theta - n_0 M(\theta)} d\theta, \quad n_0 \in \mathbb{R}, \quad x_0 \in \mathbb{R}^d. \quad (2.3)$$

If $\tilde{\pi}_{n_0,x_0}$ can be normalized to a probability measure $\pi_{n_0,x_0}$ on $\Theta$, it will be termed a distribution conjugate to the exponential family $\{P_{\theta}\}$ of (2.1). The province of the parameter $(n_0, x_0)$ which allows such a normalization is the subject matter of the first theorem. An interpretation of the theorem and examples related to it appear at the end of the section as remarks 1, 2, and 3.

**Theorem 1.** a) If $\Theta = \mathbb{R}^d$, $\tilde{\pi}_{n_0,x_0}(\Theta) < \infty$ if and only if $n_0 > 0$ and $x_0 \in \mathcal{F}$.

b) If $\Theta \neq \mathbb{R}^d$ and $n_0 > 0$, $\tilde{\pi}_{n_0,x_0}(\Theta) < \infty$ if and only if $x_0 \in \mathcal{F}$.
Proof. Note first that if \( A \) is a compact subset of \( \mathbb{R}^d \), \( \int_A e^{x \cdot \theta} d\mu(x) \leq \int e^{x \cdot \theta} d\mu(x) < \infty \) for \( \theta \in \Theta \), and hence \( \mu(A) < \infty \). For any such \( A \) with \( \mu(A) > 0 \), let \( \mu_A(B) = \mu(\Lambda \cap B)/\mu(A) \). Now according to Jensen's inequality,

\[
(2.4) \quad e^{x \cdot \theta \cdot M(\theta)} = e^{x \cdot \theta (\int e^{x \cdot \theta} d\mu(z))^{-1}} \leq e^{x \cdot \theta [\mu(A)]^{-1} (\int e^{x \cdot \theta} d\mu_A(z))^{-1}} \leq [\mu(A)]^{-1} e^{\theta \cdot (x - \int x \cdot \mu_A(z))},
\]

It will first be shown that \( \tilde{n}_{0,x}^x(x_0(x)) < \infty \) provided \( n_0 > 0 \) and \( x_0 \in \mathbb{R}^d \). In particular, write \( x_0 = \sum_{j=1}^{d+1} \lambda_j^x x_j^x \) where each \( \lambda_j^x \) is positive, \( \sum_{j=1}^{d+1} \lambda_j^x = 1 \) and \( x_j^x \in (\text{supp } \mu) \) for \( j = 1, \ldots, d+1 \). In such a representation it may be supposed that the \( x_j^x \) do not lie in a \( d-1 \) dimensional hyperplane. Now choose compact sets \( A_0(\mathcal{X}) \), \( \mu(A_0) > 0 \), so that \( x_0 = \sum_{j=1}^{d+1} \lambda_j x_j \), where each \( \lambda_j \) is positive, \( \sum_{j=1}^{d+1} \lambda_j = 1 \) and \( x_j = \int x d\mu_{A_0} \) for \( j = 1, \ldots, d+1 \). Again, one may assume that the \( x_j \) are not in some \( d-1 \) dimensional hyperplane. Invoke the inequality (2.4) to write

\[
(2.5) \quad \int_\Theta e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta \leq \int_\mathbb{R}^d e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta \\
= \sum_{k=1}^{d+1} \int_{x_k = \max_j x_j} e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta \\
\leq \sum_{k=1}^{d+1} [\mu(A_k)]^{-1} \int_{x_k = \max_j x_j} e^{n_0 \theta \cdot (\sum_{j=1}^{d+1} \lambda_j x_j - x_k)} d\theta.
\]
In the $k^{th}$ integral on the right side of (2.5) make the change of variable $\sigma_j = \theta \cdot (x_j - x_k)$, $j \neq k$, with Jacobian $|J_k|$, say. Necessarily $|J_k| \neq 0$ and then the right side of (2.5) is

$$\sum_{k=1}^{d+1} \frac{[\mu(A_k)]^{-1}}{|J_k|} \int_{\sigma_j \leq 0, j \neq k} e^{\frac{1}{n_0} \sum_{j \neq k} \lambda_j \sigma_j} d\theta < \infty.$$  

To complete the proof of a), suppose $\theta = \mathbb{R}^d$ and

$$\int_{\mathbb{R}^d} e^{\frac{1}{n_0} x_0 \cdot \theta - n_0 M(\theta)} d\theta < \infty.$$  

Observe that the matrix of second partials of $e^{\frac{1}{n_0} x_0 \cdot \theta - n_0 M(\theta)}$ is $n_0^2 (x_0 - \nabla M(\theta))^T (x_0 - \nabla M(\theta)) - n_0 M''(\theta)$, and is clearly non-negative definite if $n_0 \leq 0$. This possibility must be ruled out since a positive convex function cannot be integrable over $\mathbb{R}^d$. Suppose then that $n_0 > 0$ and $x_0 \neq x$. Note first that

$$n_0 x_0 \cdot \theta - n_0 M(\theta) = n_0 (x_0 - \nabla M(\theta)) e^{\frac{1}{n_0} x_0 \cdot \theta - n_0 M(\theta)}.$$  

Now separate $x_0$ from $x$ by a hyperplane so that $\xi \cdot x_0 \geq \xi^*$ and $\xi \cdot x \leq \xi^*$ if $x \in X$. If it is recalled that $\nabla M(\theta) \epsilon \mathbb{R}$, one has

$$\xi \cdot \nabla e^{\frac{1}{n_0} x_0 \cdot \theta - n_0 M(\theta)} \geq 0$$  

for all $\theta$. Thus the integral of $e^{\frac{1}{n_0} x_0 \cdot \theta - n_0 M(\theta)}$ with respect to the variable $\sigma_1 = \xi \cdot \theta$ is the integral of a positive monotone function and must be infinite. The assumption of integrability together with $x_0 \neq x$ thus leads to a contradiction and a) is proved.
For (b), it will be shown that the integrability of
\[ n_0 x_0 \cdot \theta - n_0 M(\theta) \]
e over \( \Theta \) with \( n_0 > 0 \) and \( x_0 \neq \emptyset \) leads to a contradiction. Here \( \emptyset \neq \mathbb{R}^d \), so let a unit vector \( \xi \) be chosen so that
\[ \xi \cdot x_0 \geq \xi^* \] and \( \xi \cdot x \leq \xi^* \) for \( x \in \emptyset \). By translating \( \emptyset \) if necessary, \( \xi^* \) may be taken to be 0. In
\[ \int_{\emptyset} e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} \, d\theta \]
make the change of variable \( \sigma_1 = \xi \cdot \theta, \sigma_2 = \xi_2 \cdot \theta, \ldots, \sigma_d = \xi_d \cdot \theta \)
where \( \xi_1, \xi_2, \ldots, \xi_d \) are orthonormal in \( \mathbb{R}^d \). As below (2.6),
\[ \xi \cdot \nabla e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} \geq 0 \] so integration first on \( \sigma_1 \) is integration of a positive non-decreasing function. A necessary condition for integrability is that \( \xi \cdot \theta \) be bounded on \( \emptyset \) for almost every \( (d\theta) \) choice of \( \sigma_2, \ldots, \sigma_d \). Suppose then that \( \theta_0 \in \emptyset \) so
\[ \int_{\emptyset} e^{x_0 \cdot \theta_0} \, d\mu(x) = \int e^{(x \cdot \xi)(\theta_0 \cdot \xi) + (x \cdot \xi_2)(\theta_0 \cdot \xi_2) + \cdots + (x \cdot \xi_d)(\theta_0 \cdot \xi_d)} \, d\mu(x) < \infty. \]
In as much as \( x \cdot \xi \leq 0 \) on \( \emptyset \), \( \theta_0 + u \xi \) must also be in \( \emptyset \) for any \( u > 0 \),
\[ x \cdot (\theta_0 + u \xi) \leq e^{x \cdot \theta_0} \] since \( e^{u \xi} \leq e \) on \( \emptyset \). But \( \xi \cdot (\theta_0 + u \xi) = \xi \cdot \theta_0 + u \) so \( \xi \cdot \theta \) is not bounded above on \( \emptyset \). This contradiction means \( x_0 \) must be in \( \emptyset \) and the proof of the theorem is complete.

The following result unifies many standard Bayesian calculations.

**Theorem 2.** Suppose \( \Theta \) is open in \( \mathbb{R}^d \). If \( \Theta \) has the distribution \( \pi \), \( n_0 > 0 \) and \( x_0 \in \emptyset \), then \( E(\nabla M(\theta)) = x_0 \).
Proof. The required result translates through (2.6) to

\[(2.7) \quad \int_{\Theta} e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta = 0.\]

Consider the first component of (2.7) and assume for now that Fubini's Theorem applies in order to write

\[(2.8) \quad \int_{\Theta} \frac{\partial}{\partial \theta_k} e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta = \int \cdots \lim_{\theta \rightarrow \theta_1} \int e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta_2 \cdots d\theta_d \]

\[- \int \cdots \lim_{\theta \rightarrow \theta} e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} d\theta_2 \cdots d\theta_d\]

where \(\theta_1 = \theta_1 \left(\theta_2, \ldots, \theta_d\right) < \theta_1 < \theta_1 = \theta_1 \left(\theta_2, \ldots, \theta_d\right)\) for \(\theta \in \Theta\). The last two integrals will be shown to be zero. Consider the first of them when \(\theta_1 = +\infty\). Use (2.4) in conjunction with a set \(A\) so that \(x_1 < (\int z d\mu_A(z))_1\), hold \(\theta_2, \ldots, \theta_d\) fixed and let \(\theta_1 \rightarrow +\infty\) to see that the integrand is zero in such a case. If \(\theta_1 \left(\theta_2, \ldots, \theta_d\right) < \infty\), take \(\theta_1^*\) so that \((\theta_1^*, \theta_2, \ldots, \theta_d)^t \in \Theta\). Then

\[\int_{x_1 \leq 0} e^{\theta_1 x_1 + \cdots + \theta_d x_d} d\mu(x) \leq \int_{x_1 \leq 0} e^{\theta_1^* x_1 + \cdots + \theta_d x_d} d\mu(x) < \infty\]

Now since \(M(\theta_1, \theta_2, \ldots, \theta_d) = +\infty\),

\[\int_{x_1 > 0} e^{\theta_1 x_1 \cdots + \theta_d x_d} d\mu(x) \rightarrow \int_{x_1 > 0} e^{\theta_1^* x_1 + \cdots + \theta_d x_d} d\mu(x) = +\infty.\]
as $\theta_1^* \leq \theta_1 \rightarrow \bar{\theta}_1 = \bar{\theta}_1(\theta_2, \ldots, \theta_d)$, by monotone convergence. Thus $M(\theta_1, \ldots, \theta_d) \rightarrow \infty$ as $\theta_1 \rightarrow \bar{\theta}_1$, and the first integral on the right side of (2.8) is zero. A similar argument applies to the second integral.

It remains to be seen that Fubini's theorem has been correctly applied at (2.8). From (2.2.ii) with $\theta_2, \ldots, \theta_d$ fixed, $\frac{\partial}{\partial \theta_1} M(\theta)$ is a strictly increasing function of $\theta_1$. Thus $\frac{\partial}{\partial \theta_1} e^{n_0 \cdot \theta - n_0 M(\theta)}$ changes sign at most once from positive to negative as $\theta_1$ varies over $(\underline{\theta}_1, \bar{\theta}_1)$, (2.6). In particular, one deduces from the argument of the previous paragraph that there is a unique point $\theta_1^* = \theta_1^*(\theta_2, \ldots, \theta_d)$ at which

$$\frac{\partial}{\partial \theta_1} e^{n_0 \cdot \theta - n_0 M(\theta)} = 0$$

and $e^{n_0 \cdot \theta - n_0 M(\theta)}$ has a maximum. Hence with $\theta^* = (\theta_1^*, \theta_2, \ldots, \theta_d)'$, 

$$\int |\frac{\partial}{\partial \theta_1} e^{n_0 \cdot \theta - n_0 M(\theta)}| d\theta_1 = 2 e^{n_0 \cdot \theta^* - n_0 M(\theta^*)} = 2 \max_{\theta_1} e^{n_0 \cdot \theta - n_0 M(\theta)}.$$

Absolute integrability in the left side of (2.8) evidently requires the integrability of $e^{n_0 \cdot \theta^* - n_0 M(\theta^*)}$ in $\theta_2, \ldots, \theta_d$. To proceed on this, let $x_0 = \sum_{j=1}^{d+1} \lambda_j x_j$ as above (2.5). Moreover, by translating $\mu$ if necessary, choose $x_0 = 0$. Then from (2.4),

11
\[ (2.9) \quad \int \cdots \int \max_{\theta_1} \left\{ e^{n_0 x_0 \cdot \theta - n_0 M(\theta)} \right\} \, d\theta_2 \cdots d\theta_d \]

\[ \leq \int \cdots \int \max_{\theta_1} \min_{\theta_k} \mu(A_n) e^{-n_0 \theta \cdot x_k} \, d\theta_2 \cdots d\theta_d \]

\[ \leq c \int \cdots \int e^{-n_0 \min_{\theta_k} \theta \cdot x_k} \, d\theta_2 \cdots d\theta_d \]

For \( 0 \neq \theta^{(2)} = (\theta_2, \ldots, \theta_d)' = \|\theta^{(2)}\|\eta \) one finds

\[ \min_{\theta_1} \max_{\theta_k} \left[ \theta_1 x_{1k} + \theta_2 x_{2k} + \cdots + \theta_k x_{kd} \right] = \|\theta^{(2)}\|\min_{\theta_1} \max_{\theta_k} \left[ \frac{\theta_1}{\|\theta^{(2)}\|} x_{1k} + \eta \cdot x_k^{(2)} \right] \]

\[ = \|\theta^{(2)}\| \min_{\theta_1} \max_{\theta_k} \left[ \theta_1 x_{1k} + \eta \cdot x_k^{(2)} \right] \cdot \]

If

\[ \inf \min_{\theta_1} \max_{\theta_k} \left[ \theta_1 x_{1k} + \eta \cdot x_k^{(2)} \right] = 5 \]

\[ \|\eta\| = 1, \quad \theta_1 \]

is positive the right side of (2.9) is bounded by

\[ c \int \cdots \int_{R^{d-1}} e^{-n_0 \|\theta^{(2)}\|5} \, d\theta_2 \cdots d\theta_d < \infty. \]

To see that 5 is positive, observe that \( 0 = \theta \cdot \sum_{j=1}^{d+1} \lambda_j x_j = \sum_{j=1}^{d+1} \lambda_j \theta \cdot x_j \), so \( \max_{k} \theta \cdot x_k \geq 0 \) for any \( \theta \). Then \( \min_{\theta_1} \max_{\theta_k} \theta \cdot x_k \geq 0 \) for any \( \theta^{(2)} \). But \( \min_{\theta_1} \max_{\theta_k} \theta \cdot x_k \) is a continuous function in \( \theta^{(2)} \) so

\[ \inf \min_{\theta_1} \max_{\theta_k} \left[ \theta_1 x_{1k} + \eta \cdot x_k^{(2)} \right] = 5 > 0 \] unless there is an \( \eta^* \) with \( \|\eta^*\| = 1 \)

12
so that \( \min_{\theta_1} \max_{k} [\theta_1 x_k^* + n^* x_k^{(2)}] = 0 \). Now if this were the case, there
would be a vector \((\theta_1, n^*)' = \xi\) with \(\max_k \xi \cdot x_k = 0\) and also, since
\(\Sigma \lambda_j \xi \cdot x_k = 0\), \(\min_k \xi \cdot x_k = 0\). This contradicts the fact that the \(x_k\)
are not contained in a \(d-1\) dimensional hyperplane. The proof of Theorem
2 is completed by applying the same arguments to the other coordinates in
(2.7).

**Remark 1.** To apply Theorem 2 to a sample \(X_1, X_2, \ldots, X_n\) of size \(n\)
from \(P_\theta\), note that if \(\pi_{n_0,x_0}\) is the prior density for \(\theta\), the
posterior density is

\[
\pi_{n_0+n, x} \propto \frac{n_0 x_0 + n \bar{x}}{n_0 + n}
\]

with \(\bar{x}\) the mean of the sample. Theorem 2 yields

\[
E(\nabla M(\theta)|X_1, \ldots, X_n) = \frac{n_0 x_0 + n \bar{x}}{n_0 + n}
\]

i.e., the conditional expectation of the mean parameter given the sample
is a linear combination of the prior expectation of the mean parameter
and \(\bar{x}\). The weights in the linear combination are proportional to \(n_0\)
and the sample size - in this sense \(n_0\) may be thought of as a prior
sample size. The restrictions of Theorem 1 to guarantee proper conjugate
priors are consistent with this interpretation.
Remark 2. Novick and Hall (1965) and others have considered negative values of \( n_0 \), which yield non-proper "ignorance" priors. It is of interest to note an example with \( \Theta \not\in \mathbb{R}^d \) in which \( n_0 x_0 \) is normable with \( n_0 < 0 \). For this, take \( d = 1 \) and \( \mu([x]) = 1 \) for \( x = 0, 1, \ldots \). The resulting exponential family corresponds to the geometric distribution. In terms of the natural parameter one finds \( e^{M(\theta)} = \frac{1}{1 - e^{\theta}} \) and thus \( \Theta = (-\infty, 0) \). From Theorem 1, if \( n_0 > 0 \), \( e^{n_0 x_0 \theta} \) is integrable over \( (-\infty, 0) \) if and only if \( x_0 \in X = (0, \infty) \). Now let \( n_0 = -t \) with \( t > 0 \) and consider

\[
(2.11) \quad \int_{-\infty}^{0} e^{-tx_0 \theta + tM(\theta)} d\theta = \int_{-\infty}^{0} \frac{1}{\theta x_0 \theta^t - n_0 M(\theta)} d\theta
\]

\[
= \int_{0}^{1} \frac{1}{u^{t+1}} \frac{1}{(1-u)^t} du.
\]

The last integral of (2.11) is clearly finite if and only if \( t < 1 \) and \( x_0 < 0 \). Thus for the present example one may think of prior sample sizes \( n_0 \) between \(-1\) and \( 0 \), together with prior sample averages \( x_0 \) that are negative.

Remark 3. As an example showing that the openness condition on \( \Theta \) is needed for Theorem 2, take \( d = 1 \) and \( \mu(x) = \frac{dx}{1 + x^2} \) on \([0, \infty)\). Then \( \Theta = (-\infty, 0] \) and an easy calculation will show that the resulting prior mean is not linear in \( x_0 \). Further discussion may be found in Jewell [1975].
3. **Characterization of Conjugate Priors — Continuous Case.**

The section is concerned with the converse to Theorem 2: can one conclude from the linearity at (2.10) that \( \theta \) had a conjugate prior? The answer is yes if it is assumed that the support of \( \mu \) is sufficiently rich. The restriction will be clarified somewhat in Section 4.

In the statement of Theorem 5 and that of Theorems 4 and 6 below, there is an assumed form for a conditional expectation. Each univariate expectation can be interpreted in the following way. (cf. Strauch (1965)):

\[
E(Y|Z) = g(Z) \text{ if and only if } E(Y_+|Z) - E(Y_-|Z) = g(Z) \text{ a.s.}
\]

With such an interpretation in mind, we are able to avoid the explicit assumption that means are finite when we postulate the form of a conditional expectation. Conditional expectation as at (3.1) is still linear, a fact we will use later.

**Theorem 5.** Let \( X \) be a sample of size one from \( P_\theta \) of (2.1), and suppose the support of \( \mu \) contains an open interval \( I_0 \) in \( \mathbb{R}^d \). It is assumed that \( \theta \) is non-empty and open in \( \mathbb{R}^d \). If \( \theta \) has a prior distribution \( \tau \) which is not concentrated on a single point, and if

\[
E(\nabla M(\theta)|X) = aX + b
\]
for some constant $a$ and constant vector $b$, then $a \neq 0$, $\tau$ is absolutely continuous with respect to Lebesgue measure, and

$$d\tau(\theta) = c \cdot e^{-b \cdot \theta - a^{-1}(1-a)M(\theta)}d\theta.$$ 

**Proof.** From (3.1) and (3.2), $E(M_1^+ | X) - E(M_1^- | X)$ is finite a.s. so $E(M_1^+ | X) \cdot E(M_1^- | X) < \infty$ for $i=1, \ldots, d$, a.s. Since $X$ has the positive density $f(x) = \int e^{x\theta - M(\theta)}d\tau(\theta)$ with respect to $\mu$, $E(M_1^+ | X) \cdot E(M_1^- | X) < \infty$ a.e. $\mu$ for each $i$. Observe that

$$(3.3) \quad E(M_1^+ | x) = \int_{M_1^-}^+ e^{x\theta - M(\theta)}d\tau(\theta)/f(x)$$

with probability 1 in $x$ for each $i$, and then that (3.3) holds a.e. $\mu$ for each $i$. Therefore all integrals on the right side of (3.3) are finite a.e. $\mu$ and may be freely manipulated. From (3.2) one finds

$$(3.4) \quad \int \nabla M(\theta)e^{x \cdot \theta - M(\theta)}d\tau(\theta) = (ax + b)f(x) \quad \text{a.e. } \mu.$$ 

If $a = 0$ in (3.4), $\int (\nabla M(\theta) - b)e^{x \cdot \theta - M(\theta)}d\tau(\theta)$ vanishes on an open interval $I_0$ of $\mathbb{R}^d$. But then $\nabla M(\theta) - b$ must vanish of the support of $\tau$ and so is zero on at least two points. Such a conclusion violates the strict convexity of $M$ (from (2.2) and below).

In (3.4) replace $x$ by $z = x + iy$ and observe that

$$Q(z) = \int (\nabla M(\theta) - az - b)e^{z \cdot \theta - M(\theta)}d\tau(\theta).$$
vanishes at least on Re $z \in I$. Then for a choice of $x_0 = \text{Re} z \in I$, $Q(x_0 + iy)$ vanishes for all $y$ and

\begin{equation}
(3.5) \quad \int \left( \frac{\nabla M(\theta) - ax_0 - b}{a} \right) e^{x_0 \cdot \theta + iy \cdot \theta - M(\theta)} \, d\tau(\theta) = iy \int e^{x_0 \cdot \theta + iy \cdot \theta - M(\theta)} \, d\tau(\theta).
\end{equation}

In (3.5), write $m(\theta)' = (m_1(\theta), \ldots, m_d(\theta))'$ and let

\[ dF(\theta) = e^{x_0 \cdot \theta - M(\theta)} \, d\tau(\theta). \]

Then one has

\begin{equation}
(3.6) \quad \int e^{iy \cdot \theta} m(\theta) dF(\theta) = iy \int e^{iy \cdot \theta} dF(\theta).
\end{equation}

The argument now proceeds from (3.6) along the lines of Lemma 6.1.1 of Linnik, Kagan and Rao [1973].

Begin with the first equation from (3.6). Multiply both sides of the equation by the factor

\[ \left( \frac{1}{2\pi} \right)^d \frac{d}{T_k} \frac{1-e^{-i\alpha_k y_k}}{iy_k} \cdot e^{-i\alpha_k y_k} - e^{-i\beta_k y_k}, \]

with $\alpha_k < \beta_k$, $k = 1, \ldots, d$, and then integrate over $-T \leq y_k \leq T$, $k=1, \ldots, d$. On the right hand side one finds

\begin{equation}
(3.7) \quad \int_{-T}^{T} \cdots \int_{-T}^{T} \left( \frac{1}{2\pi} \right)^d \frac{d}{T_k} \frac{1-e^{-i\alpha_k y_k}}{iy_k}
\quad \times \left( e^{-i\alpha_1 y_1} - e^{-i\beta_1 y_1} \right) \frac{d}{T_k} \int_{\alpha_k}^{\beta_k} e^{-i\alpha_k y_k} \, du_k \int e^{iy \cdot \theta} dF(\theta)
\end{equation}
\[
\int_{\alpha_2}^{\beta_2} \cdots \int_{\alpha_d}^{\beta_d} du_1 \cdots du_d \int d\theta \int_{-T}^{T} \frac{1}{2\pi} \prod_{k=2}^{d} \left( 1 - \frac{e^{-i\theta_k y_k}}{-iy_k} + e^{-iy_k (u_k - \theta_k)} \right) \\
\times (e^{-i\alpha_1 y_1} - e^{-i(\alpha_1 + h_1) y_1}) \prod_{l=2}^{d} \left( 1 - \frac{e^{-i\beta_1 y_1}}{-iy_1} \right) \prod_{l=2}^{d} \left( 1 - \frac{e^{-i(x_1 + h_1) y_1}}{-iy_1} \right) dy_1 \cdots dy_d
\]

\[
\int_{\alpha_2}^{\beta_2} \cdots \int_{\alpha_d}^{\beta_d} \left[ F(\alpha_1, u_2, \ldots, u_d), (\alpha_1 + h_1, u_2 + h_2, \ldots, u_d + h_d) \right] du_2 \cdots du_d
\]

as \( T \to \infty \), where \( F(\alpha, \beta) \) denotes the \( F \) measure of the \( d \)-dimensional interval \((\alpha, \beta)\). Proceeding in the same way with the left hand side of (3.6), one has

(3.8) \[
\int_{-T}^{T} \cdots \int_{-T}^{T} \frac{1}{2\pi} \prod_{k=1}^{d} \left( 1 - \frac{e^{-i\theta_k y_k}}{-iy_k} \right) \prod_{h=1}^{d} \int_{\alpha_k}^{\beta_k} e^{-iu_h y_h} du_h \int e^{-i\theta_m(\theta) d\theta} d\theta
\]

\[
\int_{\alpha_1}^{\beta_1} \cdots \int_{\alpha_d}^{\beta_d} F(1)(u_1, \ldots, u_d), (u_1 + h_1, \ldots, u_d + h_d) \right) du_1 \cdots du_d
\]

where \( dF(1)(\theta) = m_1(\theta) d\theta \). In (3.7) and (3.8) let \( h_k \to \infty \), \( k=1, \ldots, d \), to obtain

(3.9) \[
\int_{\alpha_2}^{\beta_2} \cdots \int_{\alpha_d}^{\beta_d} \left[ F(\alpha_1, u_2, \ldots, u_d) - F(\beta_1, u_2, \ldots, u_d) \right] du_2 \cdots du_d
\]

\[
= \int_{\alpha_1}^{\beta_1} \cdots \int_{\alpha_d}^{\beta_d} F(1)(u_1, \ldots, u_d) du_1 \cdots du_d
\]

18
for all $\alpha, \beta$. Now for fixed $\alpha_1, \beta_1$ it follows from (3.9) that

$$
(3.10) \quad F(\alpha_1, u_2, \ldots, u_d) - F(\beta_1, u_2, \ldots, u_d) = \int_{\alpha_1}^{\beta_1} F(1)(u_1, \ldots, u_d) du_1
$$

for almost all $u_2, \ldots, u_d$. Hence (3.10) holds simultaneously for all rational $(\alpha_1, \beta_1)$ except possibly for a set $N$ of $(u_2, \ldots, u_d)$ of Lebesgue measure 0. Then aside from $N$, $F(u_1, \ldots, u_d)$ must be absolutely continuous in $u_1$. It will be argued that this conclusion holds for all $u_2, \ldots, u_d$, i.e., that $N = \emptyset$.

Note first that since $F$ is non-decreasing and continuous in $u_1$ for almost $u_2, \ldots, u_d$, it is a fact a continuous function of all $d$ variables. Hence for each fixed $\alpha_1, \beta_1$, the left side of (3.10) is continuous in $u_2, \ldots, u_d$. But the right side of (3.10) is

$$
\int_{-\infty}^{u_2} \cdots \int_{-\infty}^{u_d} \int_{\alpha_1}^{\beta_1} \int_{-\infty}^{\infty} m_1(\theta)dF(\theta)
$$

and so must also be continuous in $u_2, \ldots, u_d$ since $F$ is continuous. Finally then, (3.10) holds for all $u_2, \ldots, u_d$ and, again from continuity, simultaneously for all $\alpha_1, \beta_1$.

Given that $F(u_1, \ldots, u_d)$ is absolutely continuous in $u_1$ for every $u_2, \ldots, u_d$, Fubini's theorem insures that $F$ is absolutely continuous with respect to $d$-dimensional Lebesgue measure and, from (3.10),

$$
\frac{\partial F}{\partial u_1} = -F(1)(u_1, \ldots, u_d).
$$

From the remaining equations at (3.6) get the full relation:

19
\[(3.11) \quad \frac{\partial F}{\partial u_r} = -F(r)(u_1, \ldots, u_d) = - \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_d} \frac{m_r(\theta)}{d \theta} dF(\theta), \quad r = 1, \ldots, d.\]

Write \(dF(\theta) = f(\theta) d\theta = \frac{\partial}{\partial \theta_i} F(\theta_1, \ldots, \theta_d) d\theta \) and use \((3.11)\) to see that
\[
\frac{\partial F}{\partial u_r} \quad \text{is also absolutely continuous with}
\]
\[
\frac{\partial}{\partial \theta_i} F(\theta) = \frac{\partial F}{\partial u_r} = -m_r f, \quad r = 1, \ldots, d.\]

Now
\[
m = \frac{\nabla M - \alpha x_0 - b}{a}
\]
so
\[(3.12) \quad \nabla f = (ba^{-1} + x_0 - a^{-1} \nabla M)f\]

from which it follows that \(f = ce^{ba^{-1} \cdot \theta + x_0 \cdot \theta - a^{-1} M(\theta)} \). Recall that
\[
dF(\theta) = e^{x_0 \cdot \theta - M(\theta)} d\tau(\theta) \quad \text{and so find} \quad d\tau(\theta) = ce^{-a^{-1} \cdot \theta - \frac{1-a}{a} M(\theta)},
\]
the desired conclusion.

**Remark 4.** It is not generally possible to ask that \((1.1)\) hold with \(a = A\) a matrix, \(A\) not proportional to \(I\). Here is a brief look at the situation for the present exponential family setting. Begin with the statement of Theorem 3 but with a matrix \(A\) and a prior measure \(\tau\) not supported on a \(d-1\)-dimensional hyperplane. Proceed through the proof of Theorem 3 to \((3.4)\). At this point one can argue that \(|A| \neq 0\) since if it were, there would be some vector \(\xi\) with \(\xi \cdot (\nabla M(\theta) - b) = 0\) on the support of \(\tau\), a contradiction. Continue through the proof to the conclusion.
\[(3.13) \quad \nabla f = - A^{-1} (\gamma M - A x_0 - b)f \]

where \( f \) is the density of the measure \( e^{\theta \cdot (\gamma M - A x_0 - b)} d\tau(\theta) \). Now it is not generally true that \((3.13)\) has a solution \( f \), but then \((3.2)\) could not have applied. For an example, take \( A \) to be a diagonal matrix with entries along the diagonal all distinct. It can be easily argued that \((3.13)\) can be solved for such an \( A \) only if \( M(\theta) \) has the special form \( \sum_{i=1}^{d} \mu(\theta_i) \). Other types of matrices lead to similar, though less agreeable, conditions on \( M \). When \((3.13)\) does allow a solution, the prior measures \( \tau \) which result are what Jewell [1974b] (in the content of the multivariate normal distribution) refers to as enriched priors.

4. **Characterization of Conjugate Priors -- Discrete Case.**

The converse of Theorem 2 is less complete when the support of \( \mu \) does not contain an interval. Consider for instance the problem of estimating a binomial parameter \( p \) from a sample of size \( n \). If \( \pi \) is a prior measure of the Borel sets of \([0,1]\), the conditions of posterior linearity become:

\[(4.1) \quad \int_0^1 p^{k+1} (1-p)^{n-k} d\pi(p) = (ak+b) \int_0^1 p^k (1-p)^{n-k} d\pi(p) \text{ for } k=0,1,2,\ldots,n. \]

There are merely restrictions on the first \( n+1 \) moments of the measure \( \pi \) and any \( \pi \) which has the same first \( n+1 \) moments as a beta measure will satisfy \((4.1)\). In this section we give theorems characterizing the conjugate priors of all commonly occurring exponential families on the
non-negative integers. Theorem 4 specializes to a characterization of
the beta distribution as the unique family allowing linear posterior
expectation for negative binomial random variables for example. The case
of Poisson variables is not covered by Theorem 4 but this has already
been treated by Johnson [1957,1967]. Theorem 5 then deals with the
assumptions needed to characterize the binomial distribution.

Suppose \( X \) is a sample of size 1 from \( P_{\theta} \) of (2.1) and let the
support of \( \mu \) be the non-negative integers. For this setting, \( \theta \) is
always an interval which is unbounded to the left. Our regularity
assumption would have \( \theta \) be an open interval, and for Theorem 4 it
will be assumed that \( \theta = (-\infty, \theta_0) \) with \( \theta_0 < \infty \), i.e., that \( \mu \) does
not have a moment generating function on all of \( R \). Under this set-up,
we have

**Theorem 4.** Suppose \( \theta \) has a prior distribution \( \tau \) on \( \theta = (-\infty, \theta_0) \)
with \( \theta_0 < \infty \), and assume \( \tau \) is not concentrated on a single point. If

\[
E(M'(\theta)|X=x) = ax + b \quad \text{for} \quad x = 0, 1, \ldots
\]

(4.2) then \( a > 0 \), \( \tau \) is absolutely continuous with respect to Lebesgue measure,
and \( d\tau(\theta) = ce^{-a-1b\theta-a^{-1}(1-a)M(\theta)}d\theta \).

**Proof.** One may proceed as in the proof of Theorem 3 to the equation (3.4).
If \( a \) were zero, (3.4) gives

\[
\int_{-\infty}^{\theta_0} (M'(\theta) - b)e^{x\theta - M(\theta)}d\tau(\theta) = 0 \quad , \quad x = 0, 1, \ldots \quad .
\]

(4.3)
Make the change of variables \( t = e^\theta \) in (4.3) and so produce a signed measure on \((0, e^0)\) having all moments zero. This signed measure must in fact be the zero measure, since the moment problem is determined on a compact interval. This implies \( M'(\theta) - b \) is zero on the support of \( \tau \), a contradiction. So \( a \neq 0 \) and (4.2) can be written as

\[
(4.4) \int_{-\infty}^{e^0} e^{x\theta} (M'(\theta) - b)e^{-M(\theta)} d\tau(\theta) = ax \int_{-\infty}^{e^0} e^{x\theta} e^{-M(\theta)} d\tau(\theta), \quad x = 0, 1, \ldots.
\]

Transform the left side of (4.4) as follows:

\[
(4.5) \int_{-\infty}^{e^0} \left[ \int_{-\infty}^{\theta} xe^{xy} dy \right] (M'(\theta) - b)e^{-M(\theta)} d\tau(\theta)
\]

\[
= \int_{-\infty}^{e^0} xe^{xy} \left[ \int_{y}^{\infty} (M'(\theta) - b)e^{-M(\theta)} d\tau(\theta) \right] dy
\]

\[
= -\int_{-\infty}^{e^0} xe^{x\theta} \left[ \int_{-\infty}^{\theta} (M'(y) - b)e^{-M(y)} d\tau(y) \right] d\theta.
\]

In (4.5) the interchange of integrations can be easily justified and (4.4) has been invoked with \( x = 0 \) to produce the final equality.

Replacing the left side of (4.4) by the right side of (4.5) one has

\[
(4.6) \int_{-\infty}^{e^0} e^{x\theta} \left[ -\int_{-\infty}^{\theta} (M'(y) - b)e^{-M(y)} d\tau(y) \right] d\theta = a \int_{-\infty}^{e^0} e^{x\theta} e^{-M(\theta)} d\tau(\theta),
\]

for \( x = 1, 2, \ldots \).
Make again the change of variable \( t = e^\theta \) in \((4.6)\) to produce a signed measure on \((0, e^{\theta_0})\) all of whose moments are zero except possibly the zeroth. Such a signed measure must concentrate on the origin and, in the present circumstance, puts no weight there. But then

\[
(4.7) \quad a e^{-M(\theta)} d\tau(\theta) = -\left[ \int_{-\infty}^{\theta} (M'(y) - b) e^{-M(y)} d\tau(y) \right] d\theta.
\]

From \((4.7)\), \( \tau \) is absolutely continuous with a density \( f \) which satisfies the differential equation

\[
(4.8) \quad af'(\theta) - aM'(\theta)f(\theta) = -(M'(\theta) - b)f(\theta).
\]

The conclusion follows easily from \((4.8)\).

**Remark 5.** In the conclusions of Theorems 3 and 4 one may identify the \( n_0 \) and \( x_0 \) from \((2.3)\) as \( n_0 = \frac{1-a}{a} \) and \( x_0 = \frac{b}{1-a} \). From Theorem 1, the corresponding measure \( \tau \) is finite when \( n_0 > 0 \) and \( x_0 \in X \). Thus \((3.2)\) or \((4.2)\) can hold provided \( 0 < a < 1 \) and \( \frac{b}{1-a} \in \mathbb{R} \). It is not always true that \( ae(0,1) \). Recall the geometric example of Remark 2. There one has the further possibility of \((4.2)\) with \( a > 1 \), as long as \( b \) is positive. Values of \( a > 1 \) are counter intuitive, for an iteration of \((1.1)\) with \( d = 1 \) yields \( a^n x + \frac{1-a^n}{1-a} b \) and this tends a.s. to an integrable random variable provided \( E(|x| \log^+|x|) < \infty \), cf. Rota [1962]. But this does not apply to the present instance with \( n_0 = -1/2 \) and \( x_0 = -1 \) for example, since one can compute that \( E|X| = E|X| = +\infty \).
As the discussion at the beginning of the section indicates, a different formulation is required if the beta distribution is to be characterized for binomial observations. This is accomplished in Theorem 5. For simplicity, in the statement and proof of this theorem we use the notation of the mean parameter as opposed to the natural parameter.

**Theorem 5.** Let \( \tau \) be a prior distribution for \( p \) on the Borel sets of \([0,1]\) and assume \( \tau \) does not concentrate on a single point. If for each \( n = 1, 2, \ldots \), there are numbers \( a_n \) and \( b_n \) for which

\[
(4.9) \quad \int_0^1 p^{k+1}(1-p)^{n-k} d\tau(p) = (a_n k + b_n) \int_0^1 p^k (1-p)^{n-k} d\tau(p)
\]

for \( k = 0, 1, \ldots, n \), then

\[
(4.10) \quad a_n = \frac{a}{1 + a(n-1)} \quad ; \quad b_n = \frac{b}{1 + a(n-1)}
\]

and \( \tau \) is a beta distribution.

**Proof.** Ericson's [1969] result, in conjunction with the linearity of (4.9), implies that

\[
E(p|S_n = k) = \frac{k \var(p) + E(p)E(1-p)}{n \var(p) + E(p)E(1-p)}
\]

This yields (4.10) with \( a = \frac{\var(p)}{E(p)E(1-p)} \) and \( b = \frac{E(p)E(1-p)}{E(1-p)} \). Take \( n = 1 \) in (4.9) and sum over \( k = 0, 1 \) to obtain \( \int p d\tau(p) = b + a \int d\tau(p) \), so \( \int p d\tau(p) = \frac{b}{1-a} \). Now take \( k = n \) in (4.9) to write
\begin{equation}
\int_0^1 p^{n+1} d\tau(p) = \frac{an+b}{1+a(n-1)} \int_0^1 p^n d\tau(p) \\
= \cdots = \left\{ \frac{n}{\prod_{j=1}^n \frac{a_j+b}{1+a(j-1)}} \right\} \frac{b}{1-a}.
\end{equation}

In this way it is clear that all moments of \( \tau \) are determined by \( a \) and \( b \), hence \( \tau \) is also. Moreover, (4.9) can be achieved by using the beta density with \( \alpha = \frac{b}{a} \) and \( \beta = \frac{1-(a+b)}{a} \).

The proof of Theorem 5 generalizes in a straightforward way to yield a characterization of the Dirichlet family as the unique family allowing linear posterior expectation for multinomial observations.

5. **Prior Measures in Location Problems**

Looking for solutions of (1.1) as a method of providing tractable prior distributions 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 here, Theorem 6, is an extension of a theorem proved by Goldstein [1975].

We begin by giving a class of priors with posterior linear 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 function \( F^\ast \) (n-fold convolution). For the location problem \( F(x-\theta) \) the observed variable is \( Z = X+\theta \). Assuming the conditional expectation exists in the sense of (3.1), we show:

\begin{equation}
E(\theta|Z) = \frac{n}{n+1} Z.
\end{equation}
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

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

\[
= E(S_n | S_n + X) + E(X | S_n + X) = E(S_n + X | S_n + X) = S_n + X.
\]

Note that it is not necessary to assume that $X$ or $\theta$ in (5.1) have finite expected value. For example, if $X$ and $\theta$ are independent Cauchy variables it is easy to verify that $E(\theta | X+\theta)$ exists and equals $\frac{X+\theta}{2}$. The class of priors can be widened by allowing the inclusion of a known location parameter for the prior. When $F$ is infinitely divisible the class of priors can be further widened to the extent that the $\frac{n}{n+1}$ in (5.1) can be replaced by $a$ with $0 < a < 1$, see Remark 7 about this point.

It is important to note that posterior linearity only obtains for samples of size one. For larger samples, $\bar{X}$ is not a sufficient unless the distribution of $X$ is normal.

It is useful for us to prove a non-existence result at this point. Lemma 1 below asserts that (5.1) cannot hold with $\frac{n}{n+1}$ replaced by 0 or 1 under very mild conditions on $X$ and $\theta$. In fact, under the same conditions, it will surface in the proof of Theorem 6 that (5.1) can hold only if $0 < a < 1$.

**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 if parts of both statements are direct from the definitions. Suppose $E(\theta|X+\theta) = b$, and take $b = 0$ without loss of generality. Then $E(X|X+\theta) = X+\theta$ or $\theta = E(X|X+\theta)-X$. Because $E|X| < \infty$, it follows that $E|\theta| < \infty$ and $E\theta = 0$. If $X$ is a.s. constant, $\theta = 0$ w.p.1 so suppose $P(X < x)P(X \geq x) > 0$ for some $x$. From the condition $E(\theta|X+\theta) = 0$, write

$$0 = \int_{X+\theta < x} \theta \, dP = \int_{X+\theta < x} \theta \, dP + \int_{X < x} \theta \, dP + \int_{X > x} \theta \, dP + \int_{\theta < 0} \theta \, dP + \int_{\theta \geq 0} \theta \, dP + 0 = 0 .$$

This equality implies that

$$\int_{X+\theta < x} \theta \, dP = \int_{X \geq x} \theta \, dP = 0 .$$

Thus for any $x$ so that $P(X < x)P(X \geq x) > 0$, we have $\theta < 0$ and $X \geq x$ implies $X+\theta \geq x$ while $X < x$ and $\theta \geq 0$ implies $X+\theta < x$. Clearly $\theta = 0$ a.s. by the independence of $X$ and $\theta$. If one begins with $E(\theta|X+\theta) = (X+\theta)+b$, then $E(X|X+\theta) = b$ and $E|X| < \infty$. Repeating the argument above completes the proof.

Assume now that (1.1) holds for a location parameter problem (as it can following (5.1)), and consider the question of uniqueness of the underlying prior distribution. Theorem 6 gives conditions guaranteeing such uniqueness. For this, let $X = (X_1, \ldots, X_d)'$ have distribution
function $F$ not concentrated at a point, and write $\lambda_{2n} = \int(x_1^{2n} + \cdots + x_d^{2n})dF$.

**Theorem 6.** Let $X$ and $\theta$ be independent $d$-dimensional random vectors with neither $X$ nor $\theta$ a.s. constant. Assume $E|X_i| < \infty$ 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

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

then $0 < a < 1$ and the distribution of $\theta$ 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 (5.2) and the linearity of conditional expectation as at (3.1),

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

and hence

$$E(\theta|X+\theta) = \frac{a}{1-a} E(X|X+\theta). \quad (5.3)$$

Use (5.3) in (5.2) to find

$$\theta = \frac{1}{1-a} E(X|X+\theta) - X. \quad (5.4)$$

By assumption the right side of (5.4) is absolutely integrable, so $E|\theta| < \infty$. When $X$ and $\theta$ have finite expectation, it follows from Lemma 1.1.1 of Kagan, Linnik and Rao [1973] that (5.2) holds if and only if the characteristic functions of $X$ and $\theta$ satisfy:
(5.5) \[ (1-a)\phi_0(t)\phi_X(t) - a\phi_X(t)\phi_0(t) = 0 \] for all \( t \).

Since \( \phi_X(t) \) does not vanish in some interval \( I \) about zero, (5.5) gives, for \( t \in I \)

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

Observe first that for (5.6) to hold with neither \( X \) nor \( \theta \) a.s. 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 (5.6). 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 (5.6) and satisfy the same determinedness condition. The proof for \( d=1 \) is therefore complete.

For \( d > 1 \) the finiteness of \( E|\theta_1| \) follows readily by the arguments used earlier. Take \( b=0 \) again without loss of generality. Now let \( \xi \in \mathbb{R}^d \) and find from (5.2) that

\[ (5.7) \quad E(\xi \cdot \theta | X + \theta) = a\xi \cdot (X + \theta) = E(\xi \cdot \theta | \xi \cdot (X + \theta)). \]

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

**Remark 6.** Here is an example of independent random variables $X$ and $\theta$ having finite means, different distributions and satisfying (5.1) with $n=1$. The example makes it clear that some hypotheses on the distribution of $X$ are required in Theorem 6 in order to guarantee uniqueness. Now (5.1) holds with $n=1$ if and only if $\phi_X^{\prime} \phi_\theta = \phi_\theta' \phi_X$ as at (5.5). Let $X$ have the density

$$\frac{1/4}{\pi} \left( \frac{\sin x}{x} \right)^{1/4}, \quad -\infty < x < \infty.$$  

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 $\phi_\theta'$ which equals $\phi_X'$ on $(-1,1)$ and has period 2 is the characteristic function of an arithmetic distribution. By construction one has $\phi_X^{\prime} \phi_\theta = \phi_\theta' \phi_X$.

**Remark 7.** For a random variable $X$, let $S_X$ be the set of real numbers $a$ which can occur in (5.2). From (5.1) $\frac{n}{n+1} \in S_X$ for each $n = 1, 2, \ldots$. If $X$ is infinitely divisible with finite mean then

$\phi_X^{\prime \prime} \phi_X^{\prime} = \frac{1-a}{a}$

is a characteristic function which offers a solution to (5.5) and thus to (5.2) for any $a$ in $(0,1)$. Conversely if $X$ has a mean, $\phi_X(t) \neq 0$ and $S_X = (0,1)$, Theorem 6 implies that $X$ is infinitely divisible.

**Remark 8.** We note that the connection of the present section to a result in martingale theory. If $(X_n)_{n=1}^{\infty}$ are i.i.d. random variables with
\( E|X_1| < \infty \), then the argument at (5.1) shows, writing \( S_n = \sum_{i=1}^{n} X_i \), that \( S_n/n \) is a backward martingale:

\begin{equation}
(5.8) \quad E\left[ \frac{S_n}{n} \mid S_{n+1}, S_{n+2}, \ldots \right] = \frac{S_{n+1}}{n+1}, \quad n = 1, 2, \ldots.
\end{equation}

Using obvious generalizations of the argument in Theorem 6 we can show that if \( \{X_i\} \) are independent random variables such that (5.8) holds and \( E|X_1| < \infty \) then all the \( X_i \) have finite first moments, and if \( \phi_i \) is the characteristic function of \( X_i \), (5.8) holds if and only if for all \( k \geq 2 \),

\begin{equation}
(5.9) \quad \phi_k'(t) \prod_{j=1}^{k-1} \phi_j(t) = \phi_{k-1}'(t) \prod_{j=1}^{k-1} \phi_j(t) \quad \text{for all } t.
\end{equation}

If \( \phi_1(t) \neq 0 \) then all the \( \phi_i = \phi_1 \). Using variants of the construction in Remark 6 we can construct an infinite sequence of independent random variables, each with a different distribution, which satisfy (5.8).

**Acknowledgement.** We thank Bradley Efron, Joseph Keller, Carl Morris, and David Siegmund for helpful remarks as this work progressed.
References


