The Annals of Statistics
1991, Vol. 19, No. 3, 1191–1224

BAYES EMPIRICAL BAYES ESTIMATION FOR NATURAL EXPONENTIAL FAMILIES WITH QUADRATIC VARIANCE FUNCTIONS

BY G. G. WALTER AND G. G. HAMEDANI

University of Wisconsin-Milwaukee and Marquette University

Certain orthogonal polynomials are employed to estimate the prior distribution of the parameter of natural exponential families with quadratic variance functions in an approach which combines Bayesian and nonparametric empirical Bayesian methods. These estimates are based on samples from the marginal distribution rather than the conditional distribution.

1. Introduction. The univariate natural exponential families (NEF) with quadratic variance functions (QVF) include many of the most widely used distributions (normal, Poisson, gamma, binomial, and negative binomial; indeed these are five of the six basic NEF–QVF distributions). These were studied by Morris (1982, 1983), who presented many of their properties in a unified way. Among other things, Morris constructed an associated family of orthogonal polynomials which in each particular case reduced to a family of standard classical orthogonal polynomials. These polynomials were then used to find estimators for arbitrary analytic functions.

The conjugate families needed for the prior distribution in Bayesian analysis were also studied by Morris (1983). These are not themselves NEF–QVF distributions, but belong to a Pearson family and have a simple form which can be exploited to obtain formulas for Bayes and parametric empirical Bayes estimation.

In previous papers, Walters and Hamedani (1987, 1989) have exploited certain classical orthogonal polynomials to obtain estimates for a prior distribution in an approach which combined Bayesian and nonparametric empirical Bayesian methods. These estimates are based on samples from the marginal distribution.

In this work we shall show that Bayes empirical Bayes procedure works in this general setting of NEF–QVF. However, the orthogonal polynomials must be related to the prior distribution rather than the conditional distribution, and therefore must be defined differently than those of Morris (1982).

We shall suppose that an initial prior distribution, based on subjective knowledge, has been selected from a member of the conjugate family. This is the best one can do [Morris (1983), Theorem 5.5] if only the first two moments of the prior distribution are known. We then use our sample to improve the

Received September 1988; revised July 1990.

AMS 1980 *subject classifications*. Primary 60E05, 62F10, 62F15.

Key words and phrases. Exponential families, natural exponential families, quadratic variance function, normal distribution, Poisson distribution, gamma distribution, binomial distribution, negative binomial distribution, hyperbolic secant distribution, orthogonal polynomials, moments.

estimate, that is, by getting a better approximation to the true prior distribution as the partial sum of a series of these orthogonal polynomials.

The orthogonal polynomials defined here are exactly the "classical" orthogonal polynomials considered by Tricomi (1955). In each of the six NEF–QVF distributions, the polynomials are identified as particular types of classical orthogonal polynomials. In some cases, however, only a finite number of them can be used, since the conjugate prior distributions may not have moments of all orders.

In Section 2 we shall review, for subsequent use, some of the properties of NEF–QVF distributions given in Morris (1982). In Section 3 we define our family of orthogonal polynomials and show their relation to those defined by Morris (1982). Some basic properties are also discussed, including the differential equation and recurrence formulas satisfied by the polynomials. More detailed properties are relegated to Appendix A. Section 4 introduces a biorthogonal system related to the polynomials which is used to recover the prior distribution from the marginal distribution. This is applied to the empirical Bayes estimation problems in Section 5. Appendix B contains the results for each of the six basic NEF–QVF distributions. These results are summarized in Table 1.

In the standard Bayesian approach it is assumed that the parameter, say θ, is fixed but not precisely known. The prior probability law $g(\theta)$ has a different character than the probability law $f(x|\theta)$ of the random variable X. It is assumed to be a subjective measure of the investigator's prior knowledge of θ. The observations are of the function $f(x|\theta)$, and a sample $\mathbf{X} = (X_1, X_2, \dots, X_N)$ therefore has the probability law

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{N} f(x_i|\theta),$$

and the marginal distribution of $\mathbf{X}$ is determined by

$$f(\mathbf{x}) = \int \prod_{i=1}^{N} f(x_i|\theta) g(\theta)\, d\theta.$$

The nonparametric empirical Bayes procedure referred to earlier is due principally to Robbins (1956). It assumes that the parameter θ is a bona fide random variable. A sample consists of independent pairs

$$(X_1, \Theta_1), (X_2, \Theta_2), \dots, (X_N, \Theta_N)$$

with the joint probability law

$$\prod_{i=1}^{N} f(x_i|\theta_i) g(\theta_i).$$

The $X_1, X_2, \dots, X_N$ are observable, but the $\Theta_1, \Theta_2, \dots, \Theta_N$ are not. The marginal distribution of $\mathbf{X} = (X_1, X_2, \dots, X_N)$ is

$$f(\mathbf{x}) = \int\int \cdots \int \prod_{a=1}^{N} f(x_i|\theta_i)g(\theta_i)\, d\theta_1\, d\theta_2 \cdots d\theta_N$$
$$= \prod_{i=1}^{N} \int f(x_i|\theta)g(\theta)\, d\theta.$$

The assumption here is that $(X_1, \Theta_1), (X_2, \Theta_2), \dots, (X_N, \Theta_N)$ is an independent sample from the distribution with density function $f(x|\theta)g(\theta)$. The conditional probability law of X given θ, namely, $f(x| \cdot)$, is assumed known; $g(\cdot)$ is assumed unknown but a smooth density.

Most approaches to the problem of estimating $g(\theta)$ have been indirect in that estimators are obtained not for $g(\theta)$ itself, but for the moments of $g(\theta)$; see Maritz (1970). These approaches, while simple, often suffer from excess "jumpiness" [as was observed by Berger (1985)] and should be smoothed. The direct methods in which $g(\theta)$ itself is estimated have usually been based on step functions [see, e.g., Deely and Kruse (1968)] or Dirichlet processes [see, e.g., Berry and Christensen (1979)] or maximum likelihood [see, e.g., Laird (1978) or Leonard (1984)]. Laird pointed out that her method is equivalent to the simultaneous estimation of several exchangeable parameters and leads to an estimator with finite support. Since we shall assume that g is a smooth density, such estimators suffer from the same difficulty as the empirical distribution, viz. they are not mean-squared consistent.

At the other extreme lie the parametric methods in which g depends only on a finite number of parameters. The simplest method depends on the assumption that g belongs to a class of conjugate densities, for example, the assumption that g is a beta density in the binomial case.

The method presented here is intermediate between the two, and may involve a finite or infinite number of parameters. It is similar to that in Walter and Hamedani (1987, 1989) and is based on orthogonal polynomials. It involves a preliminary choice of a conjugate prior and of two parameters, the prior mean and variance of which may be subjective (Bayesian) or estimated from the data (parametric empirical Bayesian), followed by an improved estimate of g based on the sample from the mixture. Then Bayes and empirical Bayes methods are combined, but in a fashion somewhat different than that of Deely and Lindley (1981).

2. A review of certain properties of NEF–QVF.

A natural exponential family is one with a cumulative distribution function (CDF) F_θ given by

$$dF_\theta(x) = \exp(\theta x - \psi(\theta))\, dF_0(x), \qquad x \in \mathbb{R}, \tag{2.1}$$

where $\theta \in \Theta \subseteq \mathbb{R}$, F_0 is a univariate CDF possessing a moment-generating function in a neighborhood of zero and

$$\psi(\theta) \equiv \log \int \exp(\theta x)\, dF_0(x), \qquad \theta \in \Theta.$$

The mean and variance of F_θ are given by

$$\mu = \psi'(\theta) \tag{2.2}$$

and

$$V(\mu) = \psi''(\theta) = \left(\frac{d\theta}{d\mu}\right)^{-1}. \tag{2.3}$$

A NEF has quadratic variance function if V has the form

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2. \tag{2.4}$$

The orthogonal polynomials defined by Morris (1982) are given by the Rodrigues formulas

$$P_n(x,\mu) = V^n(\mu)\left\{\frac{d^n}{d\mu^n}f(x,\theta)\right\}\Big/f(x,\theta), \qquad n = 0,1,2,\dots, \tag{2.5}$$

where $f(x,\theta) = \exp(\theta x - \psi(\theta))$ is a multiple of a NEF–QVF probability law. These are polynomials of exact degree n in both x and μ which are orthogonal with respect to (2.1) as a function of x. Their normalizing factor is

$$E_\theta\big(P_n^2(X,\mu)\big) = \alpha_n V^n(\mu), \tag{2.6}$$

where $\alpha_n = n!\Pi_{k=0}^{n-1}(1 + kv_2)$. If the parameter μ is changed to μ_0, the expected value with respect to $\theta(\mu)$ is

$$E_\theta(P_n(X,\mu_0)) = \frac{\alpha_n}{n!}(\mu - \mu_0)^n. \tag{2.7}$$

This is used to obtain the unique minimum variance unbiased estimator of an analytic function [Morris (1983), Theorem 3.1]

$$g_1(\mu) = \sum_{n=0}^{\infty} \frac{c_n(\mu - \mu_0)^n}{n!} \tag{2.8}$$

by

$$\hat{g}_1(X) = \sum_{n=0}^{\infty} \frac{c_n}{\alpha_n} P_n(X,\mu_0), \tag{2.9}$$

where $c_n = g_1^{(n)}(\mu_0)$, the nth derivative at μ_0.

This unbiased estimator (2.9) leads to the standard moment estimators when $g_1(\mu) = (\mu - \mu_0)^n$, for example, if $n = 1$, $\hat{g}_1(\overline{X}) = \overline{X} - \mu_0$. This estimator is of course sufficient for θ as well by the Rao–Blackwell theorem. This can also be shown directly by the factorization theorem.

A conjugate prior distribution has a density with respect to $d\mu$ of the form

$$g(\mu) = K\{\exp(m\mu_0\theta - m\psi(\theta))\}/V(\mu), \tag{2.10}$$

where θ is now treated as a function of μ and $m > 0$ is a convolution parameter. This now depends upon two prior parameters μ_0 and m, and is a two-parameter exponential family; in fact, it is a Pearson family, but is not in general a NEF.

3. A system of orthogonal polynomials. Since the system of polynomials given by Morris (2.5) is orthogonal with respect to x and not μ, we must define a different system for use with prior distributions. We define

$$r_n(\mu) = r_n(\mu, m, \mu_0) = (-1)^n \frac{1}{g(\mu)} \frac{d^n}{d\mu^n}(g(\mu)V^n(\mu)), \qquad n = 0, 1, 2, \dots, \tag{3.1}$$

where $g(\mu)$ is the prior distribution given by (2.10). Then

$$\begin{aligned} r_0(\mu) &= 1, \\ r_1(\mu) &= m(\mu - \mu_0), \\ r_2(\mu) &= m^2(\mu - \mu_0)^2 - m(\mu - \mu_0)V'(\mu) + (2v_2 - m)V(\mu), \\ &\vdots \end{aligned}$$

and $r_n(\mu)$ is a polynomial of exact degree n except for exceptional values of the parameters. For example, if $m = v_2$ or $m = 2v_2$, then the leading coefficient of $r_2(\mu)$ is zero; otherwise it is not.

We next observe that the polynomials of Morris may be obtained from the prior distribution by means of formula (2.5) since $g(\mu) = Kf_1(m\mu_0, \theta)/V(\mu)$, where $f_1(x, \theta)$ is the modification to $f(x, \theta)$ which includes the convolution parameter m [Morris (1983), (4.1)],

$$f_1(x, \theta) = \exp(\theta x - m\psi(\theta)).$$

This is equivalent to multiplying μ and $V(\mu)$ by m. Indeed we have

$$P_n(m\mu_0, m\mu) = V^{n-1}(\mu)\left(\frac{d^n}{d\mu^n}(V(\mu)g(\mu))\right)\Big/ g(\mu), \qquad n = 1, 2, \dots, \tag{3.2}$$

and hence find that r_n is given by

$$\begin{aligned} r_n &= (-1)^n (VgV^{n-1})^{(n)}/g \\ &= (-1)^n \sum_{k=0}^{n} \binom{n}{k} (V^{n-1})^{(k)} (Vg)^{(n-k)}/g \\ &= (-1)^n \left(\sum_{k=0}^{n} \binom{n}{k} V^{n-k-1}(Vg)^{(n-k)}/g \right) \left\{ (V^{n-1})^{(k)} V^{1+k-n} \right\} \\ &= (-1)^n \sum_{k=0}^{n} \binom{n}{k} P_{n-k} \left\{ (V^{n-1})^{(k)} V^{1+k-n} \right\}. \end{aligned} \tag{3.3}$$

Since the expression in the braces is a polynomial of degree at most k, it follows that r_n is a polynomial of degree at most n. It is of exact degree n except possibly in a discrete set of values of the parameters.

We now consider the orthogonality. We assume that $g(\mu)$ has moments up to the $2n$th order. Then by Theorem 5.2 of Morris (1983) and the fact that the integrated terms in integration by parts vanish for NEF–QVF distributions,

$$
\begin{aligned}
\int_a^b \mu^k r_n(\mu) g(\mu)\, d\mu &= \int_a^b \mu^k (-1)^n (V^n(\mu) g(\mu))^{(n)}\, d\mu \\
&= \int_a^b \frac{d^n}{d\mu^n} \mu^k V^n(\mu) g(\mu)\, d\mu \\
&= \begin{cases} 0, & k < n, \\ n! \int_a^b V^n(\mu) g(\mu) d\mu, & k = n, \end{cases}
\end{aligned}
\tag{3.4}
$$

and hence r_n and r_k are orthogonal. Let r_n have leading coefficient k_n. Then the normalizing factor is

$$
\begin{aligned}
\gamma_n &= \int_a^b r_n^2(\mu) g(\mu)\, d\mu = k_n \int_a^b \mu^n r_n(\mu) g(\mu)\, d\mu \\
&= k_n n! \int_a^b V^n(\mu) g(\mu)\, d\mu = k_n n! \nu_n.
\end{aligned}
\tag{3.5}
$$

This sort of Rodrigues formula is satisfied by the classical orthogonal polynomials of Jacobi, Laguerre and Hermite. In fact polynomials defined by (3.1) have a long history and many of their properties appear in Tricomi (1955). He showed not only the orthogonality mentioned in (3.4), but also derived the form of $g(\mu)$ in the case of finite, semiinfinite and infinite intervals. These forms correspond to the three classical cases. However, he assumed special forms of $V(\mu)$ which do not always hold for us. In particular, $V(\mu)$ may have a nonzero leading coefficient in some cases of infinite intervals.

All orthogonal polynomials satisfy a three-term recurrence formula of the form [Szegö (1967), page 42]

$$
\mu r_n(\mu) = A_n r_{n+1}(\mu) + B_n r_n(\mu) + C_n r_{n-1}(\mu). \tag{3.6}
$$

The coefficients may be evaluated by observing that

$$
\begin{gathered}
\int_a^b \mu r_n^2(\mu) g(\mu)\, d\mu = B_n \int_a^b r_n^2(\mu) g(\mu)\, d\mu = B_n \gamma_n, \\
\int_a^b \mu r_n(\mu) r_{n+1}(\mu) g(\mu)\, d\mu = A_n \gamma_{n+1}, \\
\int_a^b \mu r_n(\mu) r_{n+1}(\mu) g(\mu)\, d\mu = \int_a^b k_n \mu^{n+1} r_{n+1}(\mu) g(\mu)\, d\mu = \frac{k_n}{k_{n+1}} \gamma_{n+1}, \\
\int_a^b \mu r_n(\mu) r_{n-1}(\mu) g(\mu)\, d\mu = C_n \gamma_{n-1} = \frac{k_{n-1}}{k_n} \gamma_n.
\end{gathered}
\tag{3.7}
$$

In addition, the $r_n(\mu)$ are related to the $P_n(x,\mu)$ of Morris by the relation

$$\begin{aligned}\int_a^b f(x,\theta)P_n(x,\mu)g(\mu)\,d\mu &= \int_a^b \left(\frac{d^n}{d\mu^n}f(x,\theta)\right)V^n(\mu)g(\mu)\,d\mu\\ &= \int_a^b (-1)^n f(x,\theta)\frac{d^n}{d\mu^n}(V^n(\mu)g(\mu))\,d\mu \qquad (3.8)\\ &= \int_a^b f(x,\theta)r_n(\mu)g(\mu)\,d\mu.\end{aligned}$$

Then $r_n(\mu)$ correspond to particular classical polynomials in some of the special cases (see Table 1). These classical polynomials usually are defined as eigenfunctions of a second-order differential operator. Therefore, it is not surprising that the following proposition holds.

TABLE 1

Natural exponential families with quadratic variance functions, their conjugate prior distributions and associated orthogonal polynomials

Name	Normal	Poisson
Density	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\lambda)^2/2\sigma^2}$	$\frac{\lambda^x e^{-\lambda}}{x!}$
θ	$\frac{\lambda}{\sigma^2}$	$\log\lambda$
$\theta(\mu)$	$\frac{\mu}{\sigma^2}$	$\log\mu$
$\psi(\theta)$	$\frac{\sigma^2\theta^2}{2}$	e^θ
$V(\mu)$	σ^2	μ
(a,b)	$(-\infty,\infty)$	$(0,\infty)$
Zero of $V(\mu)$	—	0
Std. $g_0(\mu)^\dagger$	$e^{-\mu^2}$	$\mu^\alpha e^{-\mu}$
for $m =$	$2\sigma^2$	1
$\mu_0 =$	0	$\alpha+1$
$\nu_n = \int v^n g_0{}^\dagger$	σ^{2n}	$\frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}$
Standard polynomial	Hermite	Laguerre
Usual symbol	$H_n(x)$	$L_n^{(\alpha)}(x)$
$r_n(\mu)$	$\left(\frac{m}{2}\right)^{n/2}\sigma^n H_n\left(\sqrt{\frac{m}{2}}\left(\frac{\mu-\mu_0}{\sigma}\right)\right)$	$(-1)^n n!\,L_n^{(\alpha)}(m\mu)$
$r_n(a)$	—	$r_n(0)=(-1)^n\frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}$
k_n	m^n	m^n
$n_0 > \max n$	∞	∞

†Up to a multiplicative constant and possible change of scale and/or location.

TABLE 1
Continued

Name	Gamma	Binomial
Density	$\left(\frac{x}{\lambda}\right)^{r-1}\frac{e^{-x/\lambda}}{\lambda\Gamma(r)}$	$\binom{r}{x}p^x(1-p)^{r-x}$
θ	$\frac{-1}{\lambda}$	$\log\left(\frac{p}{(1-p)}\right)$
$\theta(\mu)$	$\frac{-r}{\mu}$	$\log\left(\frac{\mu}{r-\mu}\right)$
$\psi(\theta)$	$-r\log(-\theta)$	$r\log(1+e^{\theta})$
$V(\mu)$	$\frac{\mu^2}{r}$	$\mu-\frac{\mu^2}{r}$
(a,b)	$(0,\infty)$	$(0,r)$
Zero of $V(\mu)$	0	$0, r$
Std. $g_0(\mu)^\dagger$	$\mu^{\alpha}e^{-1/\mu}$	$\mu^{\beta}(r-\mu)^{\alpha}$
for $m=$	$\frac{-(2+\alpha)}{r}$	$\frac{\alpha+\beta+2}{r}$
$\mu_0=$	$-(2+\alpha)^{-1}$	$\frac{\beta+1}{m}$
$\nu_n=\int v^n g_0{}^\dagger$	$\frac{\Gamma(-1-\alpha-2n)}{\Gamma(-\alpha-1)r^n}$	$r^nB(n+\alpha+1, n+\beta+1)$
Standard polynomial	Bessel on $(0,\infty)$	Jacobi on $(-1,1)$
Usual symbol	$Y_n^{(\alpha)}(x)$	$P_n^{(\alpha,\beta)}(x)$
$r_n(\mu)$	$(2+\alpha)^n\mu_0^nY_n^{(\alpha)}\left(\frac{\mu}{\mu_0}\left(\frac{-\alpha}{2+\alpha}\right)\right)$	$n!r^nP_n^{(\alpha,\beta)}\left(\frac{2\mu}{r}-1\right)$
$r_n(a)$	$r_n(0)=(2+\alpha)^n\mu_0^n$	$(-1)^nr^n\frac{\Gamma(n+\beta+1)}{\Gamma(\beta+1)}=r_n(0)$
k_n	$(-1)^n\frac{\Gamma(2n+\alpha+1)}{\Gamma(n+\alpha+1)}$	$\frac{\Gamma(2n+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}$
$n_0>\max n$	$-1-\alpha$	∞

†Up to a multiplicative constant and possible change of scale and/or location.

PROPOSITION 3.1. *Let $\nu_n=\int_a^b V^n(\mu)g(\mu)\,d\mu<\infty$ for $n<n_0$. Then $r_n(\mu)$, given by* (3.1)*, satisfy the differential equation*

$$\frac{d}{d\mu}(g(\mu)V(\mu)r_n'(\mu))=\xi_nr_n(\mu)g(\mu), \tag{3.9}$$

where $\xi_n=n((n-1)v_2-m)$.

PROOF. Since r_n' is a polynomial of degree $n-1$, the left-hand side of (3.9) may be expressed as

$$(gV)'r_n'+gVr_n''=g(Vr_n''-r_1r_n')=g\rho,$$

where ρ is a polynomial of degree less than or equal to n. Let $k<n$. We shall

TABLE 1
Continued

Name	Negative binomial	Hyperbolic Secant
Density	$\binom{x+r-1}{x} p^x (1-p)^r$	$(1+\lambda^2)^{-r/2} e^{x \tan^{-1}\lambda} f_{r,0}(x)$
θ	$\log p$	$\tan^{-1}\lambda$
$\theta(\mu)$	$\log\left(\frac{\mu}{r+\mu}\right)$	$\tan^{-1}\left(\frac{\mu}{r}\right)$
$\psi(\theta)$	$-r\log(1-e^\theta)$	$-r\log(\cos\theta)$
$V(\mu)$	$\mu + \frac{\mu^2}{r}$	$r + \frac{\mu^2}{r}$
(a,b)	$(0,\infty)$	$(-\infty,\infty)$
Zero of $V(\mu)$	$(0,-r)$	$\pm ir$
Std. $g_0(\mu)^\dagger$	$\mu^\beta(r+\mu)^\alpha$	$(r+i\mu)^\alpha(r-i\mu)^\beta$
for $m =$	$\frac{\alpha+\beta+2}{r}$	$\frac{-(\alpha+\beta+2)}{r}$
$\mu_0 =$	$\frac{\beta+1}{m}$	$\frac{\beta-\alpha}{mi}$
$\nu_n = \int v^n g_0{}^\dagger$	$r^n B(n+\beta+1, -2n-\alpha-\beta-1)$	$(4r)^n \frac{\Gamma(-2n-\alpha-\beta-1)}{\Gamma(-n-\alpha)\Gamma(-n-\beta)}$
Standard polynomial	Jacobi on $(1,\infty)$	Jacobi on $(-i\infty, i\infty)$
Usual symbol	$P_n^{(\alpha,\beta)}(x)$	$P_n^{(\alpha,\beta)}(x)$
$r_n(\mu)$	$n!r^n P_n^{(\alpha,\beta)}\left(\frac{2\mu}{r}+1\right)$	$(-i)^n n! 2^n P_n^{(\alpha,\beta)}\left(\frac{-i\mu}{r}\right)$
$r_n(a)$	$r_n(0) = r^n \frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}$	$r_n(ir) = (-i)^n 2^n \frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}$
k_n	$(-1)^n \frac{\Gamma(2n+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}$	$\frac{(-1)^n\Gamma(2n+\alpha+\beta+1)}{r^n\Gamma(n+\alpha+\beta+1)}$
$n_0 > \max n$	$-1-\alpha-\beta$	$-1-\alpha-\beta$

†Up to a multiplicative constant and possible change of scale and/or location.

show that

$$\int_a^b g(\mu)\rho(\mu)\mu^k \, d\mu = 0$$

and hence that $\rho(\mu)$ is orthogonal to all polynomials of degree less than n. Thus $\rho(\mu)$ is a multiple of $r_n(\mu)$, say $\xi_n r_n(\mu)$. To see this we integrate by parts twice,

$$\int_a^b (g(\mu)V(\mu)r_n'(\mu))'\mu^k \, d\mu = -\int_a^b g(\mu)V(\mu)r_n'(\mu)k\mu^{k-1} \, d\mu$$

$$= k\int_a^b r_n(\mu)(g(\mu)V(\mu)\mu^{k-1})' \, d\mu = 0, \tag{3.10}$$

since $(g(\mu)V(\mu)\mu^{k-1})'/g(\mu)$ is a polynomial of degree less than or equal to k.

The leading coefficient of $\rho(\mu)$ is the same as that of $Vr_n'' - r_1 r_n'$, namely, $-mk_n n + v_2 n(n-1)k_n = nk_n((n-1)v_2 - m)$. This in turn must be $\xi_n k_n$, whence the conclusion. □

A more detailed version similar to this proof may be found in Tricomi (1955).

COROLLARY 3.2. *The polynomials* $\{r_n\}$ *satisfy the differential equation*

$$Vy'' - m(\mu - \mu_0)y' - n((n-1)v_2 - m) = 0.$$

We may also obtain an expression for the derivatives of r_n. Indeed $r_n'V$ is a polynomial of degree less than or equal to $n+1$; it satisfies the recurrence formula

$$r_n'V = \alpha_n r_{n+1} + \beta_n r_n + \delta_n r_{n-1}, \tag{3.11}$$

where

$$\alpha_n = nv_2 k_n/k_{n+1},$$

$$\beta_n = \frac{1}{2}\int_a^b r_n^2(\mu) r_1(\mu) g(\mu)\, d\mu/\gamma_n = \frac{1}{2}m(B_n - \mu_0),$$

$$\begin{aligned}\delta_n &= \int_a^b r_n'(\mu) r_{n-1}(\mu) V(\mu) g(\mu)\, d\mu/\gamma_{n-1}\\ &= \int_a^b r_1(\mu) r_{n-1}(\mu) r_n(\mu) g(\mu)\, d\mu - \int_a^b V(\mu) r_{n+1}'(\mu) r_n(\mu) g(\mu)\, d\mu\\ &= (k_{n-1}/k_n)\gamma_n(m - (n-1)v_2).\end{aligned}$$

All the coefficients of these two recurrence formulas may be given in terms of k_n, ν_n and B_n. This last coefficient may be found in terms of the others as well if the polynomials are known at one point [usually a zero of $V(\mu)$]. Then

$$ar_n(a) = A_n r_{n+1}(a) + B_n r_n(a) + C_n r_{n-1}(a)$$

or

$$B_n = -A_n \frac{r_{n+1}(a)}{r_n(a)} - \frac{C_n r_{n-1}(a)}{r_n(a)} + a. \tag{3.12}$$

In Table 1, we give the values of k_n, ν_n and $r_n(a)$ for each of the six families. More detailed general calculations are found in Appendix A and for particular cases in Appendix B.

4. A related biorthogonal sequence. In Bayesian simultaneous estimation methods [see Leonard (1984) for references] it is assumed that the density g belongs to a parametrized family, and then assumptions about the parameters of g are introduced. As pointed out by Leonard (1984), the choice of g very often involves a unimodal density with thin tails (e.g., normal or gamma). While these choices of prior will be adequate in numerous situations, they will be less appropriate in many others. Dawid (1973) investigated prior densities with thicker tails than normal and showed that it is unreasonable to expect the same results from analysis based upon a normal prior. Alternatively, g might possess more than one mode, in which case fairly complex analysis might be involved. In view of these observations due to Leonard (1984), he studied the empirical estimation of the general prior density g, that is, under no prior information about g. He pointed out that if some partial information about g were available, then it could be used for smoothing densities.

We are therefore interested in prior distributions which are not necessarily conjugate distributions but are more general. In this section we shall denote the conjugate density by $g_0(\mu)$ and shall allow $g(\mu)$ to be any density in the (topological) span of $\{r_n\}$ in $L^2(g_0)$. This is not a restriction if the $\{r_n\}$ are complete as they are for finite intervals.

In this case, if

$$g(\mu) = \sum_{n=0}^{\infty} a_n r_n(\mu) g_0(\mu), \tag{4.1}$$

then the marginal distribution is formally given by

$$\begin{aligned} f(x) &= \int_a^b f(x,\theta) \sum_n a_n r_n(\mu) g_0(\mu)\, d\mu \\ &= \sum_n a_n \int_a^b f(x,\theta) r_n(\mu) g_0(\mu)\, d\mu, \end{aligned} \tag{4.2}$$

times some fixed measure $dF_0(x)$. We denote by l_n the functions

$$l_n(x) = \int_a^b f(x,\theta) r_n(\mu) g_0(\mu)\, d\mu. \tag{4.3}$$

These may also be expressed in terms of Morris's polynomials as

$$l_n(x) = \int_a^b f(x,\theta) P_n(x,\mu) g_0(\mu)\, d\mu, \tag{4.4}$$

by (3.8).

We shall be interested in turning the problem around and going from $f(x)$ in (4.2) to $g(\mu)$, that is, in finding coefficients a_n such that

$$f(x) = \sum a_n l_n(x), \tag{4.5}$$

which we may then use to recover $g(\mu)$ by using (4.1). To do so we find a sequence $\{\lambda_n(x)\}$ of polynomials biorthogonal to $\{l_n(x)\}$ by using (2.8) for $r_n(\mu)$. The μ_0 used there is taken to be the parameter in $g_0(\mu)$ given by (2.10). Thus by (2.9) we have the unique minimum variance unbiased estimator of $r_n(\mu)$ in the form

$$r_n(\mu) = \sum_{k=0}^{n} c_{nk} \frac{(\mu - \mu_0)^k}{k!} \tag{4.6}$$

given by

$$\lambda_n(X) = \sum_{k=0}^{n} \frac{c_{nk}}{\alpha_k} P_k(X, \mu_0). \tag{4.7}$$

Then we have

$$\begin{aligned} E_0(\lambda_k(X) l_n(X)) &= \int_a^b \lambda_k(x) \left\{ \int_a^b f(x,\theta) r_n(\mu) g_0(\mu)\, d\mu \right\} dF_0(x) \\ &= \int_a^b E_\theta(\lambda_k(X)) r_n(\mu) g_0(\mu)\, d\mu \\ &= \int_a^b r_k(\mu) r_n(\mu) g_0(\mu)\, d\mu = \delta_{nk}\gamma_n, \end{aligned} \tag{4.8}$$

by Morris (1983), page 517.

Thus if $f(x)\, dF_0(x)$ denotes the marginal distribution, we may express $f(x)$ in the form (4.5) by taking

$$a_n = E_0(\lambda_n(X) f(X))/\gamma_n \tag{4.9}$$

and subsequently use (4.1) to find $g(\mu)$.

We have been rather cavalier with questions of convergence in this section. A number of problems may arise which we shall allude to in the next sections.

1. $g_0(\mu)$ may not have moments of all orders so that $r_n(\mu) g_0(\mu)$ may not be integrable for large n.
2. $g(\mu)$ may not be identifiable. This may happen if the l_n are not linearly independent.
3. The topological span of $\{r_n\}$ in $L^2(g_0)$ may not include all the prior distributions of interest.

The variance of $\lambda_n(X)$ may also be calculated from the general formula in Morris (1983). It is

$$\begin{aligned} \operatorname{Var}_\theta\{\lambda_n(X)\} &= \{r_n'(\mu)\}^2 V(\mu) + \{r_n''(\mu)\}^2 V^2(\mu)/2(1+v_2) \\ &\quad + \sum_{k=3}^{n} \{r_n^{(k)}(\mu)\}^2 V^k(\mu)/\alpha_k, \end{aligned} \tag{4.10}$$

where $\alpha_k = k! \prod_{j=0}^{k-1}(1 + jv_2)$.

However, we shall be interested in the variance of $\lambda_k(X)$ when X is the random variable with distribution

$$\int_a^b f(x,\theta) g_0(\mu)\, d\mu\, dF_0(x),$$

which we shall denote by Var_{g_0}. We shall use repeated integration by parts to evaluate integrals of the form

$$I_{nk} = \int_a^b \{r^{(k)}(\mu)\}^2 V^k(\mu) g_0(\mu)\, d\mu. \tag{4.11}$$

The appropriate formula is

$$\begin{aligned}\int_a^b p(\mu)q'(\mu)V(\mu)g_0(\mu)\, d\mu &= -\int_a^b p'(\mu)q(\mu)V(\mu)g_0(\mu)\, d\mu \\ &\quad + \int_a^b r_1(\mu)p(\mu)q(\mu)g_0(\mu)\, d\mu,\end{aligned} \tag{4.12}$$

where p and q are polynomials such that

$$\int_a^b |p(\mu)q(\mu)V(\mu)| g_0(\mu)\, d\mu < \infty.$$

For $n = k$, the integral (4.11) is easy to evaluate. It is

$$I_{nn} = (n!k_n)^2 \int_a^b V^n(\mu) g_0(\mu)\, d\mu = (n!k_n)^2 \frac{\gamma_n}{k_n n!} = n!k_n\gamma_n. \tag{4.13}$$

For $k = 0$, $I_{n_0} = \gamma_n$. In the other cases we apply (4.12) repeatedly to obtain:

LEMMA 4.1. *Let $g_0(\mu)$ have a finite $2n$th moment. Then*

$$I_{nk} = \frac{n!}{(n-k)!} \prod_{j=0}^{k-1} (m - (n+j-1)v_2)\gamma_n, \qquad k = 1, 2, \dots, n. \tag{4.14}$$

For $k = 1$ we use (4.12) with $p = r_n'$ and $q = r_n$. Then we find that

$$\begin{aligned}&\int_a^b r_n'(\mu) r_n'(\mu) V(\mu) g_0(\mu)\, d\mu \\ &\quad = -\int_a^b r_n''(\mu) r_n(\mu) V(\mu) g_0(\mu)\, d\mu + \int_a^b r_1(\mu) r_n'(\mu) r_n(\mu) g_0(\mu)\, d\mu \\ &\quad = -\xi_n\gamma_n.\end{aligned}$$

For general k we take $q = r_n^{(k-1)}$ and $p = r_n^{(k)} V^{k-1}$ to find

$$\begin{aligned}&\int_a^b r_n^{(k)}(\mu)\big(r_n^{(k)}(\mu) V^{k-1}(\mu)\big) V(\mu) g_0(\mu)\, d\mu \\ &\quad = -\int_a^b r_n^{(k-1)}(\mu)\big(r_n^{(k)}(\mu) V^{k-1}(\mu)\big)' V(\mu) g_0(\mu)\, d\mu \\ &\qquad + \int_a^b r_1(\mu) r_n^{(k-1)}(\mu) r_n^{(k)}(\mu) V^{k-1}(\mu) g_0(\mu)\, d\mu \\ &\quad = -\int_a^b r_n^{(k-1)}(\mu) \\ &\qquad \times \{r_n^{(k+1)}(\mu) V(\mu) + r_n^{(k)}(\mu)(k-1)V'(\mu) - r_1(\mu) r_n^{(k)}(\mu)\} \\ &\qquad \times V^{k-1}(\mu) g_0(\mu)\, d\mu,\end{aligned} \tag{4.15}$$

which is simplified by using the differential equation for $r_n^{(k-1)}$,

$$\begin{aligned} &Vy'' - (r_1 - (k-1)V')y' \\ &\quad = \{(n(n-1) - (k-1)(k-2))v_2 + m(k-1) - mn\}y \\ &\quad = (n-k+1)((n+k-2)v_2 - m)y = \eta_{n,k-1}y. \end{aligned} \tag{4.16}$$

By substituting (4.16) into the right-hand side of (4.15), we find that

$$\begin{aligned} &\int_a^b \left(r_n^{(k)}(\mu)\right)^2 V^k(\mu) g_0(\mu)\, d\mu \\ &\quad = -\eta_{n,k-1}\int_a^b \left\{r_n^{(k-1)}(\mu)\right\}^2 V^{k-1}(\mu) g_0(\mu)\, d\mu \end{aligned} \tag{4.17}$$

and the conclusion follows by induction.

COROLLARY 4.2. *The variance of $\lambda_n(X)$ is given by*

$$\operatorname{Var}_{g_0}(\lambda_n(X)) = \sum_{k=1}^{n} \binom{n}{k} \prod_{j=0}^{k-1} \frac{(m - (n-1+j)v_2)}{(1+jv_2)} \Big/ \gamma_n. \tag{4.18}$$

Since the I_{nk} of (4.11) must be nonnegative, it follows by (4.14) that $\eta_{n,k} \le 0$ and in particular that

$$\eta_{n,n} = n((2n-1)v_2 - m) \le 0$$

or

$$m \ge (2n-1)v_2. \tag{4.19}$$

This is not a contradiction, since in those cases in which $v_2 > 0$, the conjugate prior distribution has only a finite number of moments. If $v_2 < 0$, the binomial case only, then we must have $1 + jv_2 > 0$ for $j = 0, 1, \dots, k-1$, that is, $r = k$, where $V(\mu) = \mu - \mu^2/r$.

5. Estimation. In this section we suppose that we have an i.i.d. sample $X_1, X_2, \dots, X_N$ of the mixture with the probability law given by

$$dF(x) = \int_a^b f(x, \theta(\mu)) g(\mu)\, d\mu\, dF_0(x) = f(x)\, dF_0(x). \tag{5.1}$$

We shall first estimate $f(x)$ by using density estimators similar to those used with orthogonal functions. Then we estimate $g(\mu)$ by employing the procedure mentioned in the last section. Finally we obtain Bayes empirical Bayes estimates of the moments.

If $g(\mu)$ is a conjugate prior density, then the (Bayesian) posterior estimate of the mean is a weighted average of μ_0 and the sample mean $\bar{X}$ as is well known. However, this assumption is excessively restrictive, since such conjugate priors are usually univalent. This excludes the common assumption that mixtures consist of a linear combination of the $f(x|\theta_i)$. This in turn corresponds to prior distributions of the form $\sum p_i \delta(\theta - \theta_i)$. A "smeared" smooth version of this would be $\Sigma p_i \delta_n(\theta - \theta_i)$, where $\{\delta_n\}$ is smooth delta family [Walter and Blum (1979)]. Prior distributions of this form arise from MLE [Laird (1978)] and are the form considered in Leonard (1984). If $g(\mu)$ is not the conjugate prior density, this is no longer necessarily true and the posterior mean is

$$\hat{\mu} = \int_a^b \mu \bar{f}(\bar{x}, \theta) g(\mu)\, d\mu \Big/ \int_a^b \bar{f}(\bar{x}, \theta) g(\mu)\, d\mu, \tag{5.2}$$

where $\bar{f}$ is the probability law of the sample mean which is also a NEF–QVF. This can either be estimated directly or by first estimating $g(\mu)$ from the sample. We shall adopt the latter approach, which has the advantage of giving estimates of other moments as well.

We shall assume that $g_0(\mu)$, an initial Bayesian conjugate prior distribution, has been found and has moments up to $2n_0$ which may be infinity. If our Bayesian is reluctant to specify μ_0 and m based on his subjective knowledge, other procedures may be used. One such is to assume a noninformative prior distribution as the initial guess for g_0. This only works if the interval (a, b) is bounded. Another procedure is to estimate μ_0 and m from a portion of the data by using MLE or other methods and then using the conjugate prior distribution

$$\hat{g}_0(\mu) = K \exp\{\hat{m}\hat{\mu}_0 \theta(\mu) - \hat{m}\psi(\theta(\mu))\} / V(\mu)$$

as the estimate. If the true prior density $g(\mu)$ is of the form

$$g(\mu) = h(\mu) g_0(\mu), \qquad \mu \in \Omega, \tag{5.3}$$

where $h(\mu)$ is a polynomial of degree less than or equal to $n_1 = \min\{n_0, \mathrm{card}(\mathrm{supp}\, F(x))\}$, if $n_1 < \infty$, and an element of $L^2(\Omega; g_0)$, if $n_1 = \infty$, but is unknown, we assume that $f(x)$ is given by (5.1) with that $g(\mu)$.

We use the sample to estimate the coefficients in the expression (4.1) of g and (4.2) of f. They are

$$\hat{a}_k = \frac{1}{N} \sum_{i=1}^{N} \lambda_k(X_i) \gamma_k^{-1}. \tag{5.4}$$

The estimators of $g(\mu)$ and of $f(x)$ are, respectively,

$$\hat{g}_p(\mu) = \hat{h}_p(\mu) \hat{g}_0(\mu) = \sum_{k=0}^{p} \hat{a}_k r_k(\mu) \hat{g}_0(\mu), \qquad p = 0, 1, \ldots, n_1, \tag{5.5}$$

and

$$\hat{f}_p(x) = \sum_{k=0}^{p} \hat{a}_k l_k(x), \qquad p = 0, 1, \ldots, n_1. \tag{5.6}$$

Here, as in the orthogonal function estimation, p is a smoothness parameter with decreasing value corresponding to increasing smoothness. If one is interested primarily in smoothness, p should be chosen as small as possible consistent with the maximum number of anticipated modes in $g(\mu)$. Since

$\hat{h}_p(\mu)$ is a polynomial of degree less than or equal to p, it can pick up $p-1$ modes.

In general $p = p(N)$ will increase with the sample size and may approach infinity. This happens only if $n_1 = \infty$, in which case we obtain mean-squared consistency for general prior densities (below). However, it is also possible to restrict p to some value less than n_1. In this case the problem becomes parametric with parameters $\mu_0, m, a_0, a_1, \dots, a_p$. The choice of p again may be subjective, chosen on the basis of smoothness, or it may be data-based. For the latter, the choice of p based on the penalized MLE method of Schwarz (1978) seems the most promising, but has not as yet been explored.

PROPOSITION 5.1. *If $h(\mu)$ is a polynomial given by*

$$h(\mu) = \sum_{k=0}^{n_2} a_k r_k(\mu), \qquad n_2 \le n_1,$$

then

(i) $\hat{a}_k$ *is an unbiased estimator of* a_k, $k = 0, 1, \dots, n_2$;
(ii) $\hat{h}_p(\mu)$ *is an unbiased estimator of* $h(\mu)$, $n_2 \le p \le n_1$;
(iii) $\hat{f}_p(x)$ *is an unbiased estimator of* $f(x)$, $n_2 \le p \le n_1$.

PROOF. By (5.4) we have

$$\begin{aligned} E_g \hat{a}_k &= \gamma_k^{-1} E_g \lambda_k(X) \\ &= \gamma_k^{-1} \int_{\mathbb{R}} \int_{\Omega} \lambda_k(x) f(x,\theta) h(\mu) g_0(\mu)\, d\mu\, dF_0(x) \qquad (5.7) \\ &= \gamma_k^{-1} \int_{\mathbb{R}} r_k(\mu) h(\mu) g_0(\mu)\, d\mu = a_k, \end{aligned}$$

where E_g denotes the expectation with respect to the distribution given in (5.1). The proofs of (ii) and (iii) follow from (i). □

The variance of $\hat{a}_k$ may be obtained from that of $\lambda_k(X)$, which in turn may be based on (4.10). Indeed we have

$$\operatorname{Var}_g \hat{a}_k = \operatorname{Var}_g \lambda_k(X) \Big/ N\gamma_k^2 = \int_a^b \rho_{2k}(\mu) h(\mu) g_0(\mu)\, d\mu \Big/ N\gamma_k^2,$$

where $\rho_{2k}(\mu)$ is the polynomial on the right-hand side of (4.10).

From this we obtain:

COROLLARY 5.2. *Let $h(\mu)$ be a polynomial of degree $p \le n$. Then the integrated mean-squared errors (IMSE) of $\hat{h}_p$ and $\hat{f}_p$ satisfy*

$$\int_a^b E\left[\hat{h}_p(\mu) - h(\mu)\right]^2 g(\mu)\, d\mu = O\left(\frac{1}{N}\right),$$

$$\int_a^b E\left[\hat{f}_p(x) - f(x)\right]^2 dF_0(x) = O\left(\frac{1}{N}\right).$$

However, if h is not a polynomial of degree less than or equal to p, then the estimators (ii) and (iii) will be biased and the mean-squared error will not converge at the same rates. In fact we shall now assume that h is not a polynomial but is a bounded continuous function. This requires that $g_0(\mu)$ have moments of every order if h is to be approximated by polynomials.

THEOREM 5.3. *Let $g_0(\mu)$ have moments of every order and let $h(\mu)$ be a bounded continuous function such that $T^q h \in L^2(g_0)$, for some positive integer q, where T is the differential operator given by*

$$T\phi = V\phi'' - r_1\phi'.$$

Let $\hat{h}_p$ and $\hat{f}_p$ be given by (5.5) *and* (5.6), *respectively. Then for some constants C_1 and C_2 independent of N and p and each $\varepsilon > 0$,*

$$\int_a^b E\big[\hat{h}_p(\mu) - h(\mu)\big]^2 g_0(\mu)\,d\mu \le \frac{C_1}{N}(m + p|v_2|)^{p+1} + C_2(p+1)^{1+\varepsilon-2q},$$

$$\int_a^b E\big[\hat{f}_p(x) - f(x)\big]^2\,dF_0(x) \le \frac{C_1}{N}(m + p|v_2|)^{p+1} + C_2(p+1)^{1+\varepsilon-2q}.$$

PROOF. The mean-squared error of $\hat{h}_p$ is given by

$$E\big[\hat{h}_p - h\big]^2 = E\left[\sum_{k=0}^{p}(\hat{a}_k - a_k)r_k\right]^2 + \left[\sum_{k=p+1}^{\infty} a_k r_k\right]^2,$$

where now h is given by the convergent [in $L^2(g_0)$] series

$$h(\mu) = \sum_{k=0}^{\infty} a_k r_k(\mu).$$

The coefficients are given by

$$\begin{aligned} a_k &= \int_a^b h(\mu) r_k(\mu) g_0(\mu)\,d\mu/\gamma_k \\ &= \int_a^b T^q h(\mu) r_k(\mu) g_0(\mu)\,d\mu/\gamma_k \xi_k^q \end{aligned} \tag{5.8}$$

and

$$|a_k|^2 \le \int_a^b [T^q h(\mu)]^2 g_0(\mu)\,d\mu/\xi_k^{2q}\gamma_k, \tag{5.9}$$

by Schwarz's inequality. Thus the integrated bias term satisfies

$$\int_a^b\left[\sum_{k=p+1}^{\infty} a_k r_k(\mu)\right]^2 g_0(\mu)\,d\mu = \sum_{k=p+1}^{\infty} a_k^2\gamma_k = O\left(\sum_{k=p+1}^{\infty}\xi_k^{-2q}\right). \tag{5.10}$$

Since by (3.9), $\xi_k = O(k)$ if $v_2 = 0$, $m \ne 0$ and $\xi_k = O(k^2)$ if $v_2 \ne 0$, it follows that (5.10) is dominated by $(p+1)^{1+\varepsilon-2q}$ for each $\varepsilon > 0$.

The integrated variance (IV) term is given by

$$\mathrm{IV} = \int_a^b E\left[\sum_{k=1}^{p} (\hat{a}_k - a_k) r_k(\mu)\right]^2 g_0(\mu)\, d\mu$$

$$(5.11) \qquad = \sum_{k=1}^{p} \left\{E(\hat{a}_k - a_k)^2/\gamma_k\right\}$$

$$= \sum_{k=1}^{p} \left\{\int_a^b \sum_{j=1}^{k} \left(\left[r_k^{(j)}(\mu)\right]^2 V^j(\mu)/\alpha_k\right) h(\mu) g_0(\mu)\, d\mu\right\} \Big/ N\gamma_k.$$

In order to evaluate this expression, we use (4.18) to find that for $1 + iv_2 > 0$, $i = 1, 2, \dots, p$,

$$\mathrm{IV} \le \frac{1}{N} \sum_{k=1}^{p} \gamma_k^{-1} \sum_{j=1}^{k} \binom{k}{j} \prod_{i=0}^{j-1} \frac{(m - (k - 1 + i)v_2)}{(1 + iv_2)} \gamma_k \|h\|_\infty$$

$$(5.12) \qquad \le \frac{1}{N} \sum_{k=1}^{p} \sum_{j=1}^{k} \binom{k}{j} (m - (k-1)v_2)^j \|h\|_\infty$$

$$= \frac{1}{N} \sum_{k=1}^{p} (1 + m - (k-1)v_2)^k \|h\|_\infty$$

$$\le \text{const.} \frac{1}{N} (m + (p-1)|v_2|)^{p+1}.$$

Hence by combining (5.12) and (5.10) we reach the first conclusion. The second follows from the first by Schwarz's inequality. □

This IMSE can be made to converge to 0 as $N \to \infty$ provided $p = p(N)$ converges to infinity at such a rate that $p^{p+1}/N \to 0$ as $N \to \infty$ for $v_2 \neq 0$ and at a rate of $m^{p+1}/N \to 0$ as $N \to \infty$ for $v_2 = 0$. In the first case the rate of convergence will be extremely slow. However, in the case of $v_2 > 0$ there are only a finite number of possible terms in the expansion of $h(\mu)$, while for $v_2 < 0$ there are only a finite number of possible values of X. Hence in neither case is it possible to allow $p \to \infty$.

If $v_2 = 0$, the convergence is more rapid, though still quite slow. Indeed, if $p + 1 = O(\log N / 2 \log m)$, we have:

COROLLARY 5.3. *If $v_2 = 0$, then the IMSE satisfies*

$$IMSE = O\left([\log N]^{1-2q}\right)$$

for both $\hat{h}_p$ and $\hat{f}_p$, where $p + 1 = O(\log N / 2 \log m)$.

The estimates of the posterior mean and variance arising from the estimate of $g(\mu) = h(\mu)g_0(\mu)$ are

$$\hat{\theta}_p = \left[\frac{1}{\hat{f}_p(x)}\int_a^b \theta f(x,\theta)\hat{h}_p(\mu)g_0(\mu)\,d\mu\right]_{(a,b)}$$

and

$$\hat{V}_p = \left[\frac{1}{\hat{f}_p(x)}\int_a^b \left(\theta - \hat{\theta}_p\right)^2 f(x,\theta)\hat{h}_p(\mu)g_0(\mu)\,d\mu\right]_{(0,\infty)}.$$

Both may be shown to be asymptotically optimal [see Walter and Hamedani (1989)] with rate $O(N^{-1/2+\varepsilon})$ by Corollary 5.2 if $h(\mu)$ is a polynomial in $L^2((a,b); g_0)$. If $v_2 = 0$ and $h(\mu)$ is a bounded continuous function, then both are again asymptotically optimal but with a slow rate by Corollary 5.3.

This is also true of the posterior mean estimate. We can use the properties of the orthogonal polynomials to find an expression for the posterior mean [$h(\mu)$ a polynomial]

$$\begin{aligned}\mu_1 &= \int_a^b \mu f(x,\theta)\sum_{k=0}^{n} a_k r_k(\mu)g_0(\mu)\,d\mu \Big/ \sum_{i=0}^{n} a_k l_k(x) \\ &= \sum_{k=0}^{n} a_k\left(A_k l_{k+1}(x) + B_k l_k(x) + C_k l_{k-1}(x)\right)\Big/ \sum_{k=0}^{n} a_k l_k(x).\end{aligned} \tag{5.13}$$

The estimate can then be obtained by replacing a_k by $\hat{a}_k$ and truncating the sum to p.

An alternative point of view is found by observing that

$$l_k(x) = E(r_k(\mu)|X = x)$$

and using the moment calculations obtained from Theorem 5.2 of Morris (1983),

$$E((\mu - x_0)r_k(\mu)|X = x) = \frac{1}{m+1}E(r_k'(\mu)V(\mu)|X = x), \tag{5.14}$$

where $x_0 = (x + m\mu_0)/(m+1)$. This is useful when p is small, for example, $p = 1$. Then we have

$$l_0(x) = 1, \qquad l_1(x) = r_1(x_0), \qquad E(\mu|X = x) = x_0$$

and

$$\begin{aligned}E(\mu r_1(\mu)|X = x) &= E((\mu - x_0)r_1(\mu) + x_0 r_1(\mu)|X = x) \\ &= \frac{m}{m+1}E(V(\mu)|X = x) + x_0 r_1(x_0) \\ &= V(x_0)\frac{m}{(m+1)(m+1-v_2)} + x_0 r_1(x_0).\end{aligned}$$

Hence for $p = 1$, the posterior mean is

$$\mu_1 = \frac{a_0 x_0 + a_1(x_0 r_1(x_0) + V(x_0)m/(m+1)(m+1-v_2))}{a_0 + a_1 r_1(x_0)} \tag{5.15}$$
$$= x_0 + \frac{a_1 r(x_0) m}{(m+1)(m+1-v_2)(a_0 + a_1 r_1(x_0))}.$$

We have proposed a general method which encompasses six particular cases. Our results have been primarily theoretical and of an asymptotic nature. However, the small sample behavior of our method is potentially more useful. A simulation study is being undertaken, but is not as yet complete. We present here an example from our previous work and results of two simulations.

EXAMPLE 1. In the binomial case, since the interval (a, b) is bounded, the use of a noninformative prior is possible. If the interval is normalized to $(0, 1)$ by using $p = \mu/r$ (see Table 1) as the parameter, the setting is exactly the same as in Walter and Hamedani (1987). The resulting polynomials are the Legendre polynomials. These were used to estimate p from the past data $(5, 4, 5, 5, 0)$ and current value 5 from a binomial mixture with $r = 5$. The results were

$$\hat{p}_1 = 0.865, \qquad \hat{p}_2 = 0.930, \qquad \hat{p}_3 = 0.939, \qquad \hat{p}_4 = 0.945,$$

where the subscript denotes the number of terms in the estimator. This may be compared to the estimates ranging from 0.886 to 0.936 from the same data arising from an estimator based on a Dirichlet process [Berry and Christensen (1979)].

EXAMPLE 2. A simulation in the binomial case $r = 5$ in which the prior was bimodal was also done (J. Letelier, personal communication). The prior was assumed to be the density

$$g(p) = \frac{3\pi}{14}(2\sin(\pi p) + \sin(3\pi p)),$$

and samples of size 15 were taken from the resulting marginal distribution. The results were

$$\hat{p}_1 = 0.509, \qquad \hat{p}_2 = 0.605, \qquad \hat{p}_3 = 0.769, \qquad \hat{p}_4 = 0.528,$$

in which the subscript denotes the number of terms in the estimate of $g(p)$. In this example a different sample was used in each of these cases as well as a different current value. These were also generated randomly and were, respectively, $x = 3, 5, 5, 3$ for the four cases. The true value of the $E(p)$ was of course 0.5.

The expression for $\hat{g}(p)$ was compared to that of $g(p)$. For approximation by a fourth degree polynomial (five terms in the series), the correct shape was observed even when samples as small as 5 were taken.

EXAMPLE 3. We consider a simulation in which the prior distribution has a point mass at 0 and at 2, that is, $g(\mu) = \frac{1}{2}\delta(\mu) + \frac{1}{2}\delta(\mu - 2)$, and the NEF–QVF is $N(\mu, 1)$, that is,

$$f(x, \mu) = \exp\left\{x\mu - \frac{\mu^2}{2}\right\}.$$

Since the conjugate prior is also normal in this case, it has the form

$$g_0(\mu) = K \exp\left\{m\mu_0\theta - \frac{\theta^2}{2}\right\},$$

where $\theta = \mu$. In this case we cannot take the trial prior to be noninformative since the interval (a, b) is infinite. Accordingly, we take it to be as simple as possible with $\mu_0 = m = 1$. The polynomials $r_n(\mu)$ are by (B.1.3)

$$r_n(\mu) = \left(\frac{1}{2}\right)^{n/2} H_n\left(\frac{\mu - 1}{\sqrt{2}}\right)$$

for

$$g_0(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-(\mu - 1)^2}{2}\right\}.$$

The polynomials $P_n(x, \mu)$ similarly may be found to be

$$P_n(x, \mu) = \left(\frac{1}{2}\right)^{n/2} H_n\left(\frac{x - \mu}{\sqrt{2}}\right).$$

From this a simple calculation gives us

$$\begin{aligned} l_n(x) &= \int_{-\infty}^{+\infty}\left(\exp\left\{x\mu - \frac{\mu^2}{2}\right\}\right)\left(\frac{1}{2}\right)^{n/2} H_n\left(\frac{\mu - 1}{\sqrt{2}}\right) g_0(\mu)\, d\mu \\ &= \left(\frac{1}{\sqrt{2}} \exp\left\{\frac{x^2}{2}\right\}\right)\left(\frac{1}{2}\right)^n H_n\left(\frac{x - 1}{2}\right) \exp\left\{-\left(\frac{x - 1}{2}\right)^2\right\}. \end{aligned}$$

The biorthogonal system $\lambda_n(x)$ which satisfies

$$\int_{-\infty}^{+\infty} \lambda_k(x) l_n(x)\, dF_0(x) = \delta_{kn}\gamma_n$$

must satisfy

$$\int_{-\infty}^{+\infty} \lambda_k(x) l_n(x) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = \delta_{kn} 2^n n! \sqrt{\pi}.$$

Hence

$$\lambda_k(x) = H_k\left(\frac{x - 1}{2}\right) 2^k \sqrt{\pi}.$$

The estimator of the prior, given a sample $X_1, \dots, X_N$, in this case is

$$\hat{g}(\mu) = \sum_{n=0}^{p} \hat{a}_n 2^{-n/2} H_n\left(\frac{\mu - 1}{\sqrt{2}}\right)(2\pi)^{-1/2} \exp\left\{-\frac{(\mu-1)^2}{2}\right\},$$

where

$$\hat{a}_n = \frac{1}{N}\sum_{i=1}^{N} H_n\left(\frac{X_i - 1}{2}\right) \Big/ n!\sqrt{\pi}.$$

A number of samples of the mixture were generated and then used to try to recover $g(\mu)$. The first was (0.639, 0.0049, 1.456, −1.083, 1.376). The resulting estimates of a_n were

$$\hat{a}_0 = 1, \quad \hat{a}_1 = -0.521, \quad \hat{a}_2 = -0.419, \quad \hat{a}_3 = 0.191, \quad \hat{a}_4 = 0.085,$$

which gave an estimator of

$$\begin{aligned}\hat{g}(\mu) = \frac{1}{\sqrt{2\pi}}&\left(\exp\left\{-\frac{(\mu-1)^2}{2}\right\}\right)\\ &\times\left[1 - \frac{0.521}{\sqrt{2}}H_1\left(\frac{\mu-1}{\sqrt{2}}\right) - \frac{0.419}{2}H_2\left(\frac{\mu-1}{\sqrt{2}}\right)\right.\\ &\qquad\left. + \frac{0.191}{2\sqrt{2}}H_3\left(\frac{\mu-1}{\sqrt{2}}\right) + \frac{0.085}{4}H_4\left(\frac{\mu-1}{\sqrt{2}}\right)\right].\end{aligned}$$

This estimate is very crude given the small sample size and the small number of terms used. The mean is just 1 plus the coefficient of H_1, in this case, $\hat{\mu} = 0.64$. (This is not the posterior mean, but rather an estimate of the prior mean.)

For a sample of size 10 with the same seed we have

$$\hat{a}_0 = 1, \quad \hat{a}_1 = -0.736, \quad \hat{a}_2 = -0.356,$$
$$\hat{a}_3 = 0.402, \quad \hat{a}_4 = -0.0052,$$

while for a sample of size 20 we have

$$\hat{a}_0 = 1, \quad \hat{a}_1 = -0.015, \quad \hat{a}_2 = -0.140,$$
$$\hat{a}_3 = 0.034, \quad \hat{a}_4 = -0.135.$$

In this last case, the estimate for μ is $\hat{\mu} = 0.989$, while the correct value is of course 1.

REMARK. Leonard (1984) has proposed an estimator of the form

$$\hat{g}(\mu(\theta)) = \frac{1}{m}\sum_{i=1}^{m} \delta(\theta - \theta_i)$$

for the prior density. The θ_i are chosen based on a sample $x_1, x_2, \dots, x_m$ by maximizing the likelihood function

$$L = \prod_{j=1}^{m} \sum_{i=1}^{m} \exp\left(x_j\theta_i - \psi(\theta_i)\right)$$

with respect to each θ_i. Since this estimator shares the shortcomings (and advantages) of the empirical distribution, it cannot be mean-squared consistent. However, a smoothed version should be. This can easily be obtained in terms of our orthogonal polynomials by approximating $\delta(\mu - \mu_i)$ by the partial sums of its polynomial expansion,

$$\delta(\mu - \mu_i) = \sum_{n=0}^{\infty} r_n(\mu) r_n(\mu_i) g_0(\mu),$$

that is, since $\delta(\theta - \theta_i) = \delta(\mu - \mu_i)/V(\mu)$,

$$\hat{g}_p^*(\mu(\theta)) = \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{p} r_n(\mu(\theta)) r_n(\mu(\theta_i)) \right) g_0(\mu(\theta)). \tag{5.16}$$

This approach has not yet been explored but shows promise.

APPENDIX A

Estimates of parameters associated with the orthogonal polynomials $\{r_n\}$. The basic relations are the Rodrigues formula

$$r_n = (-1)^n (V^n g_0)^{(n)}/g_0, \tag{3.1}$$

the differential equation

$$(V g_0 r_n')' = \xi_n r_n g_0, \qquad \xi_n = n((n-1)v_2 - m), \tag{3.9}$$

the recurrence formula

$$\mu r_n(\mu) = A_{n+1} r_{n+1}(\mu) + B_n r_n(\mu) + C_n r_{n-1}(\mu) \tag{3.6}$$

and the derivative expression

$$V r_n' = \alpha_n r_{n+1} + \beta_n r_n + \delta_n r_{n-1}. \tag{3.11}$$

We have also taken the leading coefficient to be k_n and the normalizing factor to be γ_n. A central quantity that occurs repeatedly is the expression

$$\int_a^b V^n(\mu) g_0(\mu)\, d\mu = \nu_n, \tag{A.1}$$

which may be calculated explicitly if the QVF, $V(\mu)$ and $g_0(\mu)$ are known (see Table 1).

The normalizing factor may be calculated easily: It is given by (3.5),

$$\gamma_n = k_n n! \nu_n \tag{A.2}$$

and may be found explicitly if k_n is calculated as well.

A recurrence formula for the leading coefficients may be found by means of (3.1) [Tricomi (1955), page 136],

$$\frac{k_{n+1}}{k_n} = \frac{(2nv_2 - m)((2n-1)v_2 - m)}{(n-1)v_2 - m}. \tag{A.3}$$

This gives us expressions for A_n and C_n in (3.1),

$$A_n = \frac{k_n}{k_{n+1}}, \qquad C_n = \frac{k_{n-1}}{\gamma_{n-1}} \frac{\gamma_n}{k_n} = \frac{n\nu_n}{\nu_{n-1}}. \tag{A.4}$$

However B_n involves both the leading coefficient and the coefficient k'_n of μ^{n-1} in $r_n(\mu)$. It is [Tricomi (1955), page 126]

$$B_n = \frac{k'_n}{k_n} - \frac{k'_{n+1}}{k_{n+1}}. \tag{A.5}$$

By again using the differential equation, we can find k'_n to be [Tricomi (1955), page 137]

$$\frac{k'_n}{k_n} = n\left\{\frac{m\mu_0 + (n-1)v_1}{-m + 2(n-1)v_2}\right\}. \tag{A.6}$$

The three coefficients in the other recurrence formula (3.11) may be found in terms of k_n and ν_n by using these expressions. Alternately one can use (3.12) which gives B_n in terms of $r_n(a)$ if it is known.

Since the differential equation has as its highest order coefficient $V(\mu)$, a quadratic function, it is possible to convert it into a standard form by a linear change of variable. This form depends on whether or not $v_2 = 0$ and, if it does, whether $v_1 = 0$ as well.

In case $v_2 \neq 0$, we may divide (3.9) by $-v_2$ to obtain the differential equation satisfied by $y = r_n$:

$$\left(-\mu^2 - \frac{v_1}{v_2}\mu - \frac{v_0}{v_2}\right)\frac{d^2y}{d\mu^2} + \left(\frac{m}{v_2}\mu - \frac{m\mu_0}{v_2}\right)\frac{dy}{d\mu} + \frac{\xi_n}{v_2}y = 0. \tag{A.7}$$

We then change scale and location by letting $r = a + bt$ where $a = -(v_1 + d)/(2v_2)$ [i.e., $V(a) = 0$] and $b = d/v_2$ where d^2 is the discriminant of $V(\mu)$, $d = \sqrt{v_1^2 - 4v_0v_2}$. Then (A.7) becomes

$$\begin{gathered} t(1-t)\frac{d^2y}{dt^2} - \left(\frac{m}{2v_2}\left(\frac{v_1 + d + 2v_2\mu_0}{d}\right) - \frac{m}{v_2} - t\right)\frac{dy}{dt} + \frac{\xi_n}{v_2}y = 0, \\ \frac{\xi_n}{v_2} = n\left(n - 1 - \frac{m}{v_2}\right) \end{gathered} \tag{A.8}$$

the hypergeometric equation of Gauss. Hence our polynomial r_n is expressed as the hypergeometric function

$$\text{(A.9)} \quad r_n(\mu) = c_n F\left(-n, n-1-\frac{m}{v_2}; \frac{-m}{2dv_2}(d+v_1+2v_2\mu_0); \frac{\mu-a}{b}\right),$$

where c_n is a constant. Since $F(a,b;c;0) = 1$, we find that

$$c_n = r_n(a).$$

In the case $v_2 = 0$, but $v_1 \neq 0$ the equation in (3.9) may be expressed as

$$\text{(A.10)} \quad \left(\mu + \frac{v_0}{v_1}\right)\frac{d^2y}{d\mu^2} - \frac{m}{v_1}(\mu-\mu_0)\frac{dy}{d\mu} + n\frac{m}{v_1}y = 0.$$

By letting $\mu = a + bt$, $a = -v_0/v_1$, $b = v_1/m$, we obtain the confluent hypergeometric equation

$$\text{(A.11)} \quad t\frac{d^2y}{dt^2} + \left(\frac{mv_0}{v_1^2} - \frac{m\mu_0}{v_1} - t\right)\frac{dy}{dt} - (-n)y = 0,$$

whence it follows that

$$r_n(\mu) = {}_1F_1\left(-n; \frac{mv_0}{v_1^2} - \frac{m\mu_0}{v_1}; \frac{\mu-a}{b}\right),$$

where ${}_1F_1$ is the confluent hypergeometric function.

In case both v_1 and $v_2 = 0$, then $V(\mu)$ is just a constant v_0, and (3.9) may, by the transformation $\mu = a + bt$, $a = \mu_0$, $b^2 = 2v_0/m$, be converted to the Hermite equation

$$\frac{d^2y}{dt^2} - 2t\frac{dy}{dt} + 2ny = 0.$$

These formulas can be used to obtain an expression for the derivative of r_n. Indeed in the case $v_2 \neq 0$, we use the fact that

$$\text{(A.12)} \quad \begin{aligned} \frac{d}{dx}F(a,b;c,x) &= \sum_{\nu=1}^{\infty} \frac{(a)_\nu(b)_\nu}{\nu!(c)_\nu}\nu x^{\nu-1} \\ &= \sum_{\nu=0}^{\infty} \frac{(a)_{\nu+1}(b)_{\nu+1}}{\nu!(c)_{\nu+1}}x^\nu = \frac{ab}{c}\sum_{\nu=0}^{\infty}\frac{(a+1)_\nu(b+1)_\nu}{\nu!(c+1)_\nu}x^\nu \\ &= \frac{ab}{c}F(a+1, b+1; c+1; x). \end{aligned}$$

Hence for our polynomials we have

$$\text{(A.13)} \quad \begin{aligned} \frac{d}{d\mu}r_n(\mu,\mu_0) = r_n(a,\mu_0)e_n F\bigg(&-n+1, n-\frac{m}{v_2}; \\ &1 - \frac{m}{2dv_2}(d+v_1+2v_2\mu_0); \frac{\mu-a}{b}\bigg), \end{aligned}$$

where

$$e_n = \frac{2n(n-1-m/v_2)v_2^2}{m(d+v_1+2v_2\mu_0)}.$$

But

$$\begin{aligned} r_{n-1}&\left(\mu, \mu_0 - \frac{d}{m}\right) \\ &= r_{n-1}\left(a, \mu_0 - \frac{d}{m}\right) \\ &\quad \times F\left(-n+1, n-\frac{m}{v_2}; \frac{-m}{2dv_2}\left(d+v_1+2v_2\left(\mu_0 - \frac{d}{m}\right)\right); \frac{\mu-a}{b}\right), \end{aligned}$$

since a and b depend only on $V(\mu)$ and not the prior parameters m and μ_0. Thus we have

$$\frac{d}{d\mu} r_n(\mu, \mu_0) = \frac{r_n(a, \mu_0)e_n}{r_{n-1}(a, \mu_0 - d/m)} r_{n-1}\left(\mu, \mu_0 - \frac{d}{m}\right). \tag{A.14}$$

The constant coefficients in (A.14) may be evaluated by using the fact that the leading coefficients of both sides must be equal. Also, it should be observed that this leading coefficient does not involve the parameter μ_0. Indeed by (3.3) we see that

$$\begin{aligned} r_n(\mu, \mu_0) = (-1)^n \sum_{k=0}^{n} \binom{n}{k} P_{n-k}(m\mu_0, m\mu) \\ \times \left\{ \left(V^{n-1}(\mu)\right)^{(k)} V^{1+k-n}(\mu) \right\}, \end{aligned} \tag{3.3}$$

where $P_{n-k}(m\mu_0, m\mu)$ is a polynomial of exact degree $n-k$ in both $m\mu_0$ and μ. Its leading coefficient in μ does not depend upon μ_0 [Morris (1982), (8.2)]. The expression $(V^{n-1}(\mu))^{(k)}V^{1+k-n}(\mu)$ is a polynomial of degree k whenever $v_2 \neq 0$. Hence each term on the right-hand side of (3.3) is a polynomial of degree n whose leading coefficient is independent of μ_0 and so is $r_n(\mu, \mu_0)$. Thus by equating the leading coefficients in (A.14) we find

$$\frac{d}{d\mu} r_n(\mu, \mu_0) = \frac{nk_n}{k_{n-1}} r_{n-1}\left(\mu, \mu_0 - \frac{\mu_0}{m}\right). \tag{A.15}$$

Other formulas involving the derivative may be found in [Tricomi (1955), page 136]. As before, the coefficients may be expressed in terms of ν_n and the parameters of $V(\mu)$, $g_0(\mu)$.

For the case $v_2 = 0$, the polynomials are either Laguerre or Hermite and the derivative expressions are well known. Another expression for the derivative of $r_n g_0$ is easily derived from the definition

$$(r_n g_0)' = \rho_n (r_{n-1} g_0)' + n\rho_n' (r_{n-1} g_0), \tag{A.16}$$

where $\rho_n = r_1 - (n-1)V'$.

APPENDIX B

The six individual cases. In this section we present detailed calculations for each of the six basic NEF–QVF distributions by Morris (1982).

B.1. Normal. In the normal case we have $\mu = \theta\sigma^2$, $\psi(\theta) = \theta^2\sigma^2/2$ and $V(\mu) = \sigma^2$. The conjugate prior distribution is a constant times

$$\begin{aligned} g_0(\mu) &= \frac{1}{\sigma^2}\exp\big(m\big(\mu_0\theta - \sigma^2\theta^2/2\big)\big) \\ &= \frac{1}{\sigma^2}\exp\big((m\mu/2\sigma^2)(2\mu_0 - \mu)\big). \end{aligned} \tag{B.1.1}$$

The polynomials satisfy the Rodrigues formula

$$\begin{aligned} r_n(\mu) &= (-1)^n \exp\big((m\mu/2\sigma^2)(\mu - 2\mu_0)\big) \\ &\quad\times \frac{d^n}{d\mu^n}\big(\sigma^{2n}\exp\big(-(m\mu/2\sigma^2)(\mu - 2\mu_0)\big)\big). \end{aligned} \tag{B.1.2}$$

For $m = 2\sigma^2$ and $\mu_0 = 0$, $r_n(\mu) = \sigma^{2n}H_n(\mu)$, where H_n are the Hermite polynomials. The polynomial r_n may be obtained from H_n by a change of scale and location:

$$r_n(\mu) = \big(m/2\sigma^2\big)^{n/2}\sigma^{2n}H_n\Big(\big(m/2\sigma^2\big)^{1/2}(\mu - \mu_0)\Big). \tag{B.1.3}$$

The leading coefficient of H_n is 2^n and hence that of $r_n(\mu)$ is

$$k_n = m^n\sigma^{2n}. \tag{B.1.4}$$

Since $V(\mu) = \sigma^2$ is a constant, $\nu_n = \sigma^{2n}$.

The formulas for H_n are well known; the recurrence formulas are

$$xH_n(x) = \tfrac{1}{2}H_{n+1}(x) + nH_{n-1}(x) \tag{B.1.5}$$

and

$$H_n'(x) = 2nH_{n-1}(x), \tag{B.1.6}$$

while the normalizing factor is

$$\gamma_n = \int_{-\infty}^{\infty} e^{-x^2}H_n(x)\,dx = 2^n n!\pi^{1/2}. \tag{B.1.7}$$

B.2. Poisson. In the Poisson case the parameters θ and μ are related by $\mu = e^\theta = \psi(\theta) = V(\mu)$, where μ may take values in $\Omega = (0,\infty)$. There is an immense literature in this case, most of which deals with estimation of the parameter θ [Hudson (1978)]. That their mean-squared error is often better than ours is not surprising given the generality of our method. The conjugate prior is a multiple of

$$g_0(\mu) = \frac{1}{\mu}\exp(m\mu_0\log\mu - m\mu) = \mu^{m\mu_0-1}\exp(-m\mu). \tag{B.2.1}$$

The Rodrigues formula is

$$r_n(\mu) = (-1)^n \mu^{-m\mu_0+1} e^{m\mu} \frac{d^n}{d\mu^n}\left(\mu^{n+m\mu_0-1} e^{-m\mu}\right), \tag{B.2.2}$$

from which it follows that r_n with a change of scale and normalization is related to the standard Laguerre polynomial $L_n^{(\alpha)}$ with $\alpha = m\mu_0 - 1$,

$$r_n(\mu) = (-1)^n n! L_n^{(\alpha)}(m\mu), \tag{B.2.3}$$

and hence $r_n(0) = (-1)^n \Gamma(n+\alpha+1)/\Gamma(\alpha+1)$, $k_n = m^n$.

The recurrence formulas are

$$\begin{aligned} xL_n^{(\alpha)}(x) &= -(n+1)L_{n+1}^{(\alpha)}(x) + (2n+\alpha+1)L_n^{(\alpha)}(x) \\ &\quad -(n+\alpha)L_{n-1}^{(\alpha)}(x) \end{aligned} \tag{B.2.4}$$

and

$$x\left(L_n^{(\alpha)}(x)\right)' = nL_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x). \tag{B.2.5}$$

The ν_n again are straightforward:

$$\nu_n = \int_0^\infty \mu^n \mu^\alpha e^{-\mu} d\mu \Big/ \Gamma(\alpha+1) = \Gamma(n+\alpha+1)/\Gamma(\alpha+1). \tag{B.2.6}$$

B.3. Gamma. In the case of the gamma distribution, the parameter θ is given by $\theta = -r/\mu$, $\psi(\theta) = -r\log(-\theta)$ and $V(\mu) = \mu^2/r$. Hence the conjugate prior is a constant times

$$\begin{aligned} g_0(\mu) &= (r/\mu^2)\exp(-m\mu_0 r/\mu + mr\log(r/\mu)) \\ &= r^{mr+1}\mu^{-mr-2}\exp(-m\mu_0 r/\mu), \qquad \mu \in (0,\infty). \end{aligned} \tag{B.3.1}$$

The Rodrigues formula is

$$r_n(\mu) = (-1)^n \mu^{2+mr} e^{m\mu_0 r/\mu} \frac{d^n}{d\mu^n}\left(\mu^{2n-mr-2} e^{-m\mu_0 r/\mu}\right). \tag{B.3.2}$$

If we take $m\mu_0 r/\mu = 2/x$ and let $mr + 2 = -\alpha$, we obtain the Rodrigues formula (with a different constant) for the generalized Bessell polynomials $Y_n^{(\alpha)}$ [Chihara (1978), page 183],

$$Y_n^{(\alpha)}(x) = 2^{-n} x^{-\alpha} e^{2/x} \frac{d^n}{dx^n}\left(x^{2n+\alpha} e^{-2/x}\right). \tag{B.3.3}$$

Hence we have, since $mr = -(\alpha+2)$,

$$\begin{aligned} r_n(\mu) &= \left(-(2+\alpha)\mu_0/2\right)^n (-1)^n 2^n Y_n^{(\alpha)}(2\mu/-(2+\alpha)\mu_0) \\ &= \left((2+\alpha)\mu_0\right)^n Y_n^{(\alpha)}((-2/(2+\alpha))\mu/\mu_0). \end{aligned} \tag{B.3.4}$$

Thus $r_n(0) = ((2+\alpha)\mu_0)^n Y_n^{(\alpha)}(0) = (2+\alpha)^n \mu_0^n$ and the leading coefficient k_n may be calculated from that of $Y_n^{(\alpha)}$ or directly from (B.3.2). It is

$$k_n = (-1)^n \Gamma(2n+\alpha+1)/\Gamma(n+\alpha+1).$$

In the normalized case $\mu_0 = 1/mr = -1/(2+\alpha)$, the ν_n are found to be

$$\begin{aligned}\nu_n &= \int_0^\infty (\mu^{2n+\alpha}/r^n)e^{-1/\mu}\,d\mu/\Gamma(-\alpha-1)\\ &= \Gamma(-2n-\alpha-1)/r^n\Gamma(-\alpha-1).\end{aligned} \tag{B.3.5}$$

The recurrence formulas for $Y_n^{(\alpha)}$ are given in Chihara (1978), page 183.

The $r_n(\mu)$ may also be expressed in terms of Laguerre polynomials which satisfy the Rodrigues-type formulas [Szegö (1967), page 388]

$$e^{-y}y^{\beta}L_n^{(\beta)}(y) = \frac{(-1)^n}{n!}\mu^{n+1}\frac{d^n}{d\mu^n}\left(\mu^{-\beta-1}e^{-1/\mu}\right), \tag{B.3.6}$$

where $y = 1/\mu$. Hence for $mr + 2 = \beta + 1$ and $\mu_0 = 1/mr$, we have

$$r_n(\mu) = \mu^n n! \sum_{k=0}^{n} \binom{2n}{n+k} L_k^{(\beta)}(1/\mu)(-1)^{n-k}. \tag{B.3.7}$$

This last expression can also be given by [Tricomi (1955), page 218]

$$r_n(\mu) = \mu^n n! L_n^{(\beta-2n)}(1/\mu). \tag{B.3.8}$$

In this case only a finite number of the $r_n(\mu)$ belong to $L^2((0,\infty); g_0(\mu))$. Indeed

$$\begin{aligned}\int_0^\infty V^n(\mu)g_0(\mu)\,d\mu &= \int_0^\infty (\mu^2/r)^n r^{mr+1}\mu^{\alpha}e^{-m\mu_0 r/\mu}\,d\mu\\ &= \int_0^\infty r^{mr+1-n}x^{2n-\alpha-2}e^{-m\mu_0 rx}\,dx\end{aligned}$$

converges if $-2n-\alpha-2 > -1$ and diverges otherwise. Thus only those polynomials with n satisfying this inequality have finite norms.

B.4. Binomial. In the binomial case, the appropriate interval is finite $\Omega = (0, r)$. The mean is $\mu = r/(1+e^{-\theta})$, $\psi(\theta) = r\log(1+e^{\theta})$ and $V(\mu) = \mu - \mu^2/r$. Here r is the total number of trials and $1/(1+e^{-\theta}) = p$, the probability of success. The conjugate prior distribution will be

$$\begin{aligned}g_0(\mu) &= \exp(m(\mu_0\theta - r\log(1+e^{\theta})))/(\mu-\mu^2/r)\\ &= \exp(m\mu_0\log(\mu/(r-\mu)) - mr\log(r/(r-\mu)))/(\mu-\mu^2/r)\\ &= \mu^{m\mu_0-1}(r-\mu)^{mr-m\mu_0-1}r^{1-mr}\\ &= \mu^{\beta}(r-\mu)^{\alpha}r^{-(\alpha+\beta+1)}.\end{aligned} \tag{B.4.1}$$

With a change of scale (i.e., $r = 1$), this leads to the usual Rodrigues formula for the Jacobi polynomials on $(0,1)$,

$$\begin{aligned}\frac{1}{n!}r_n(\mu) &= \frac{1}{n!}(-1)^n\mu^{-\beta}(1-\mu)^{-\alpha}\frac{d^n}{d\mu^n}\left(\mu^{n+\beta}(1-\mu)^{n+\alpha}\right)\\ &= p_n^{(\alpha,\beta)}(\mu)\end{aligned} \tag{B.4.2}$$

The Bayes empirical Bayes problem has already been treated for this case in Walter and Hamedani (1987). They also considered the case of a noninformative initial prior which led to the Legendre polynomials. The more general problem in which the indices n_i of the binomial distribution are allowed to vary was not considered, but may be attacked by the method of Leonard (1976). The recurrence formulas for $r = 1$ are well known [see Szegö (1967), pages 71–72] as is the differential equation. We observe merely that $r_n(0) = \Gamma(n + \alpha + 1)/\Gamma(\alpha + 1)$, that the leading coefficient is

$$k_n = \frac{\Gamma(2n + \alpha + \beta + 1)}{\Gamma(n + \alpha + \beta + 1)}$$

and

$$\gamma_n = \int_0^1 r_n^2(\mu) g_0(\mu)\, d\mu = \frac{\Gamma(n + \alpha + 1)\Gamma(n + \beta - 1)}{n!(2n + \alpha + \beta + 1)\Gamma(n + \alpha + \beta - 1)}.$$

These Jacobi polynomials are related to the standard ones on $(-1, 1)$ by a change of scale and location

$$p_n^{(\alpha,\beta)}(\mu) = P_n^{(\alpha,\beta)}(2\mu - 1).$$

The $\{r_n(\mu)\}$ are complete in $L^2((0, r), \mu^\beta(r - \mu)^\alpha)$ but the corresponding $\{l_n(x)\}$ given by (4.3) are not linearly independent since x has only $r + 1$ distinct values. Hence to avoid problems with identifiability of $g(\mu) = h(\mu)g_0(\mu)$, we must restrict $h(\mu)$ to the span of $\{r_0, r_1, \dots, r_r\}$.

B.5. Negative binomial. In the case of the negative binomial the mean is given by $\mu = r/(e^{-\theta} - 1)$, $\psi(\theta) = -r\log(1 - e^{\theta})$ and $V(\mu) = \mu + \mu^2/r$. The conjugate prior distribution may be expressed as a constant times

$$\begin{aligned} g_0(\mu) &= \exp(m\mu_0 \log(\mu/(r + \mu)) - mr\log(r/(r + \mu)))/(\mu + \mu^2/r) \\ &= \mu^{m\mu_0 - 1}(r + \mu)^{mr - m\mu_0 - 1}r^{1 - mr} \qquad \text{(B.5.1)} \\ &= \mu^\beta (r + \mu)^\alpha r^{-(\alpha + \beta + 1)}, \end{aligned}$$

which is similar to (B.4.1) but the interval is $\Omega = (0, \infty)$. Since $g_0 \in L^1(0, \infty)$, it is restricted to $\beta > -1$ which must hold since both μ_0 and m are positive. However, $\alpha + \beta < -1$ as well and therefore $\alpha < 0$. We can change the scale again which is equivalent to setting $r = 1$ and find the Rodrigues formula to be

$$r_n(\mu) = (-1)^n \mu^{-\beta}(1 + \mu)^{-\alpha} \frac{d^n}{d\mu^n}\left(\mu^{n+\beta}(1 + \mu)^{n+\alpha}\right). \tag{B.5.2}$$

This can be converted to the Rodrigues formula for Jacobi polynomials on $(-1, 1)$ by the change of variable $x = 1 + 2\mu$, to obtain

$$
\begin{aligned}
r_n\left(\frac{x-1}{2}\right) &= (-2)^n \left(\frac{x-1}{2}\right)^{-\beta} \left(\frac{x+1}{2}\right)^{-\alpha} \\
&\quad \times \frac{d^n}{dx^n}\left(\left(\frac{x-1}{2}\right)^{n+\beta}\left(\frac{x+1}{2}\right)^{n+\alpha}\right) \\
&= \left(\frac{(-1)^n}{2^n}\right)(x-1)^{-\beta}(x+1)^{-\alpha}\frac{d^n}{dx^n}\left((x-1)^{n+\beta}(x+1)^{n+\alpha}\right) \\
&= (-1)^n n! P_n^{(\beta,\alpha)}(x) = n! P_n^{(\alpha,\beta)}(x).
\end{aligned}
\tag{B.5.3}
$$

However, since the interval in x is $(0, \infty)$, many of the standard calculations do not hold. The moments ν_n are given by

$$
\nu_n = \int_0^\infty \mu^n (1 + \mu/r)^n \mu^\beta (r + \mu)^\alpha \, d\mu \Big/ \int_0^\infty \mu^\beta (r+\mu)^\alpha \, d\mu, \tag{B.5.4}
$$

which by the change of variable $x = \mu/(r + \mu)$ are seen to be

$$
\nu_n = 4^n B(n + \beta + 1, -2n - \beta - \alpha - 1)/B(\beta + 1, -\beta - \alpha - 1). \tag{B.5.5}
$$

These moments clearly exist only if $\beta > -1$ and $2n + \beta + \alpha < -1$ and hence we do not have a complete set of polynomials.

Other parameters may be calculated in terms of $P_n^{(\alpha,\beta)}$. We find $r_n(0)$ by (B.5.3) to be

$$
\begin{aligned}
r_n(0) &= (-1)^n n! P_n^{(\beta,\alpha)}(1) = (-1)^n n! (-1)^n \binom{n+\alpha}{n} \\
&= \Gamma(n + \alpha + 1)/\Gamma(\alpha + 1)
\end{aligned}
\tag{B.5.6}
$$

and the leading coefficient to be

$$
k_n = (-1)^n \Gamma(2n + \alpha + \beta + 1)/\Gamma(n + \alpha + \beta + 1).
$$

B.6. Generalized hyperbolic secant. The generalized hyperbolic secant distributions introduced by Morris (1982) have as their interval of mean values $\Omega = (-\infty, \infty)$ with $\mu = r \tan \theta$, $\psi(\theta) = -r \log(\cos \theta)$ and $V(\mu) = \mu + \mu^2/r$. The conjugate prior therefore is

$$
\begin{aligned}
g_0(\mu) &= \left(\mu/(r^2 + \mu^2)\right) \exp\left(m\mu_0 \tan^{-1}(\mu/r)\right. \\
&\qquad\qquad \left. + mr \log\left(\cos\left(\tan^{-1}(\mu/r)\right)\right)\right) \\
&= \left(r/(r^2 + \mu^2)\right)(\cos\theta)^{mr} e^{m\mu_0\theta}, \qquad \theta = \tan^{-1}(\mu/r) \\
&= r^{mr+1}(r^2 + \mu^2)^{-1-mr/2}\left((r - i\mu)/(r + i\mu)\right)^{m\mu_0 i/2}.
\end{aligned}
\tag{B.6.1}
$$

The Rodrigues formula is

$$r_n(\mu) = (-1/r)^n (r^2+\mu^2)^{1+mr/2} e^{-m\mu\theta} \frac{d^n}{d\mu^n}\left((r^2+\mu^2)^{n-1-mr/2} e^{m\mu\theta}\right)$$

$$\text{(B.6.2)} \qquad = (-1/r)^n r^{mr+1}(1/g_0(\mu)) \times \frac{d^n}{d\mu^n}\left((r-i\mu)^{(m\mu_0 i/2)-(mr/2)-1+n}(r+i\mu)^{-(m\mu_0 i/2)-(mr/2)-1+n}\right).$$

However, the differential equation is easier to interpret in this case. It is

$$\text{(B.6.3)} \qquad (r^2+\mu^2)y'' - mr(\mu-\mu_0)y' = n(n-1-mr)y.$$

This may be converted into the equation

$$\text{(B.6.4)} \quad (1-x^2)\frac{d^2y}{dx^2} + (m\mu_0 i + mrx)\frac{dy}{dx} + n(n-1-mr)y = 0$$

by the change of variable $\mu = ixr$. But this is the equation of the Jacobi polynomials on $(-1,1)$, with the solution

$$y = P_n^{(\alpha,\beta)}(x),$$

where $\alpha = -(m/2)(r+\mu_0 i) - 1$, $\beta = -(m/2)(r-\mu_0 i) - 1$. Under the same change of variable (B.6.2) becomes

$$\begin{aligned} r_n(xri) &= \left(\frac{-1}{r}\right)^n \left(\frac{1}{ri}\right)^n (r-xr)^{-\alpha}(r+xr)^{-\beta} \\ &\quad \times \frac{d^n}{dx^n}\left((r-xr)^{\alpha+n}(r+xr)^{\beta+n}\right) \qquad \text{(B.6.5)} \\ &= (i)^n(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left((1-x)^{\alpha+n}(1+x)^{\beta+n}\right) \\ &= (-i)^n n! 2^n P_n^{(\alpha,\beta)}(x). \end{aligned}$$

Since $P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$, it follows that

$$\text{(B.6.6)} \qquad r_n(ri) = (-i)^n 2^n \Gamma(n+\alpha+1)/\Gamma(\alpha+1).$$

The leading coefficient of $P_n^{(\alpha,\beta)}(x)$ is $2^{-n}\binom{2n+\alpha+\beta}{n}$ [Szegö (1967), page 63]. Hence that of $r_n(\mu)$ is

$$\text{(B.6.7)} \quad k_n = ((-1)^n/r^n)\Gamma(2n+\alpha+\beta+1)/\Gamma(n+\alpha+\beta+1).$$

The recurrence formulas can be found from those of the Jacobi polynomials.

The moments ν_n are given by

$$\nu_n = \frac{\int_{-\infty}^{\infty}\left((r^2+\mu^2)^n/r^n\right)(r+i\mu)^\alpha (r-i\mu)^\beta \, d\mu}{\int_{-\infty}^{\infty}(r+i\mu)^\alpha (r-i\mu)^\beta \, d\mu,} \tag{B.6.8}$$

and may be calculated by means of the formula [Erdélyi (1954), page 22]

$$\begin{aligned} &\frac{2}{\pi}\int_0^{\pi/2}(\cos\theta)^{\nu-1}\cos(y\theta)\,d\theta \\ &= \frac{2^{1-\nu}\Gamma(\nu)}{\Gamma((\nu+y+1)/2)\Gamma((\nu-y+1)/2)}, \qquad \operatorname{Re}(\nu) > 0. \end{aligned} \tag{B.6.9}$$

By the change of variable $\mu/r = \tan\theta$, we find the integral to be

$$\begin{aligned} &\frac{1}{\pi}\int_{-\pi/2}^{\pi/2}(\cos\theta)^{\nu+1}e^{iy\theta}r(\cos\theta)^{-2}\,d\theta \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} r^{\nu+2}(r^2+\mu^2)^{-(\nu+1)/2}\left(\frac{r-i\mu}{r+i\mu}\right)^{-y/2} d\mu \end{aligned}$$

and hence by taking $y = \alpha - \beta$ and $-(\nu+1)/2 = n + (\alpha+\beta)/2$, we find (B.6.8) to be

$$\nu_n = \frac{(4r)^n\Gamma(-2n-\alpha-\beta-1)\Gamma(-\beta)\Gamma(-\alpha)}{\Gamma(-n-\alpha)\Gamma(-n-\beta)\Gamma(-\alpha-\beta-1).} \tag{B.6.10}$$

Again the moments exist only when $2n + \alpha + \beta < -1$.

Acknowledgments. The authors wish to thank Richard Askey for pointing out certain references and for saving them the pain (and pleasure) of rediscovering properties of classical orthogonal polynomials. We are also grateful to a referee for calling our attention to several important references and for his suggestions which improved our presentation.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERRY, D. A. and CHRISTENSEN, R. (1979). Empirical Bayes estimation of binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7** 558–568.

CHIHARA, T. S. (1978). *An Introduction to Orthogonal Polynomials*. Gordon and Breach, New York.

DAWID, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60** 664–667.

DEELY, J. J. and KRUSE, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.* **39** 268–288.

DEELY, J. J. and LINDLEY, D. V. (1981). Bayes Empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841.

ERDÉLYI, A. (ed.) (1954). *Tables of Integral Transforms* **1**. McGraw-Hill, New York.

HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.

LAIRD, N. M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LEONARD, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika* **63** 69–75.

LEONARD, T. (1984). Some data-analytic modifications to Bayes–Stein estimation. *Ann. Inst. Statist. Math. Part A* **86** 11–21.

MARITZ, J. S. (1970). *Empirical Bayes Methods*. Methuen, London.

MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80.

MORRIS, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Ann. Statist.* **11** 515–529.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 157–164. Univ. California Press, Berkeley.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SZEGÖ, G. (1967). *Orthogonal Polynomials*. Amer. Math. Soc., Providence, R.I.

TRICOMI, F. G. (1955). *Vorlesungen über Orthogonalreiken* [translation of *Serie Ortogonali di Funzloni*, Torino (1948)]. Springer, Berlin.

WALTER, G. G. and HAMEDANI, G. G. (1987). Empiric Bayes estimation of binomial probability. *Commun. Statist. Theory Methods* **16** 559–577.

WALTER, G. G. and HAMEDANI, G. G. (1989). Bayes empirical Bayes estimation for discrete exponential families. *Ann. Inst. Statist. Math.* **41** 101–119.

WALTER, G. G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF WISCONSIN–MILWAUKEE
MILWAUKEE, WISCONSIN 53201

DEPARTMENT OF MATHEMATICS, STATISTICS,
AND COMPUTER SCIENCE
MARQUETTE UNIVERSITY
MILWAUKEE, WISCONSIN 53233