Ann. Inst. Statist. Math.
Vol. 41, No. 1, 101-119 (1989)

BAYES EMPIRICAL BAYES ESTIMATION FOR DISCRETE EXPONENTIAL FAMILIES

G. G. WALTER[1]* AND G. G. HAMEDANI[2]

[1]*Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, U.S.A.*
[2]*Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI 53233, U.S.A.*

(Received August 21, 1987; revised May 24, 1988)

Abstract. Bayes-empiric Bayes estimation of the parameter of certain one parameter discrete exponential families based on orthogonal polynomials on an interval (a, b) is introduced. The resulting estimator is shown to be asymptotically optimal. The application of this method to three special distributions, the binomial, Poisson and negative binomial, is discussed.

Key words and phrases: One parameter exponential families, Jacobi polynomials, Laguerre polynomials, estimation of the prior density.

1. Introduction

The Bayesian approach to estimation of a parameter θ involves choosing an estimator θ^* which minimizes the Bayes risk. By using Bayes theorem it is shown that this estimator (see Bickel and Doksum (1977)) is given by

$$\theta^* = \frac{\int \theta p(x|\theta)dG(\theta)}{\int p(x|\theta)dG(\theta)}, \tag{1.1}$$

where $G(\theta)$ is the (prior) distribution function of θ and $p(x|\theta)$ is the conditional probability law of the random variable X given θ. Usually $p(x|\theta)$ is assumed to be known and $G(\theta)$ is assumed to belong to a parametric family whose parameters are chosen by the investigator on the basis of the past experience.

*The first author was supported by NSF grant DCR-8504620.

In the empiric Bayes procedure, due principally to Robbins (1963), the prior distribution function $G(\theta)$ is not known but must be estimated from a sample $X_1, X_2, \ldots, X_N$ of X. This estimate is then used, together with (1.1) and the next observation X_{N+1} to obtain an estimate of θ_{N+1}, the value of θ which led to this observation. This last estimate should converge to θ^* in mean square as $N \to \infty$. In this case it is called asymptotically optimal.

Deely and Lindley (1981) observed that empiric Bayes methods are not truly Bayesian and proposed a procedure for combining Bayes and empiric Bayes methods. In this work we shall study an alternative procedure doing the same task. It is an extension of the method introduced in Walter and Hamedani (1987) for estimation of the binomial parameter. We shall consider here, discrete exponential families with a single parameter.

We suppose first that $X_1, X_2, \ldots, X_N, X_{N+1}$ form an i.i.d. sample of the mixture

$$p(x) = \int p(x|\theta)g(\theta)d\theta , \tag{1.2}$$

where θ has the unknown "true" absolutely continuous distribution with density function $g(\theta)$. An initial prior approximation $g_0(\theta)$ to $g(\theta)$ based on subjective knowledge of an element from a conjugate family is made. Then a weight function $w(\theta)$ is formed from $g_0(\theta)$ and any factors in $p(x|\theta)$ which are independent of x. Orthogonal polynomials $p_k(\theta)$ based on $w(\theta)$ are constructed. These $p_k(\theta)$ are then used to estimate $g(\theta)$ by functions of the form

$$\hat{g}_m(\theta) = \sum_{k=0}^{m} \hat{a}_k p_k(\theta) g_0(\theta) , \tag{1.3}$$

where $\hat{a}_k$ depends on the sample $X_1, X_2, \ldots, X_N$. The $\hat{g}_m(\theta)$ are in turn used to obtain estimates $\hat{p}_m(x)$ of $p(x)$ from (1.2) and estimates of θ_{N+1} from (1.1) when $X_{N+1} = x$. For $m = 0$, we have exactly the Bayes estimate, while for larger values of m we have progressively less smooth estimators $\hat{p}_m(x)$.

We shall show that the resulting $\hat{g}_m$ and $\hat{p}_m$ are integrated mean square consistent with respect to an appropriate weight function and the estimate of θ_{N+1} is asymptotically optimal.

These methods differ considerably from the previously proposed methods for estimating the prior density function. The methods of Choi and Bulgren (1968), Deely and Kruse (1968), Blum and Susarla (1977), Tortorella and O'Bryan (1979) were all based on step function estimation, while those of Berry and Christensen (1979) were based on Dirichlet processes. O'Bryan and Walter (1979) used Fourier transform methods.

In the next section we shall consider the Bayes-empiric Bayes estimation in the case of discrete exponential families with a single parameter. In this case there is a natural conjugate family from which we choose our initial prior $g_0(\theta)$. The application of this method to three special distributions, the binomial, Poisson, and negative binomial is discussed in the last section. In two of these cases one encounters a standard type of orthogonal polynomials.

2. General formulation

In this section we shall deal with discrete random variables whose probability functions are members of a one parameter natural exponential family. We shall change the parameterization to enable us to construct an orthogonal system of polynomials in the new parameter. We begin with the natural form for the one parameter exponential family,

$$p(x|\phi) = [\exp\{\phi T(x) + d(\phi) + S(x)\}]I_A(x)\,, \tag{2.1}$$

where A is a discrete set in $\mathbb{R}^N$ with indicator function I_A, ϕ is the parameter which is assumed to be real and $d(\phi)$, $T(x)$, $S(x)$ are real-valued functions. The statistic $T(x)$ is sufficient for ϕ when $p(x|\phi)$ has this form (see Bickel and Doksum (1977)). Most standard discrete distributions including the binomial, Poisson, negative binomial, are of this form.

Such exponential families have natural conjugate families with two parameters whose probability density function is given by

$$\pi'_{(\alpha,\beta)}(\phi) = [\exp\{\phi\alpha + d(\phi)\beta\}]I_\Phi(\phi)/\psi(\alpha,\beta)\,, \tag{2.2}$$

where α and β are the parameters, $\psi(\alpha,\beta)$ is a normalizing factor, and Φ is a subset of $\mathbb{R}$ (see Bickel and Doksum (1977)). By Theorem 2.3.1 of Bickel and Doksum (1977), the probability function of $Y = T(X)$, the natural sufficient statistics, which we assume to take integer values, is given by

$$p(y|\phi) = [\exp\{\phi y + d(\phi) + S^*(y)\}]I_{A^*}(y)\,,$$

where A^* is the image of A under T and $S^*(y) = \ln\{\Sigma_{\{x\in A:\, T(x)=y\}}e^{S(x)}\}$ if $y \in A^*$ and 0 otherwise. We now change the parameterization in $p(y|\phi)$ by setting $\theta = \exp\{\phi\}$; and noting that θ belongs to a finite or infinite interval (a, b). Then we have

$$\begin{aligned} p(y|\theta) &= \theta^y \exp\{d(\ln\theta)\}\exp\{S^*(y)\}I_{A^*}(y) \\ &= \theta^y\xi(\theta)\sigma(y)I_{A^*}(y)\,. \end{aligned} \tag{2.3}$$

The conjugate prior probability density function of θ would have the form

$$\pi_{(\alpha,\beta)}(\theta) = \theta^{\alpha}\xi^{\beta}(\theta)I_{\Theta}(\theta)/\psi^{*}(\alpha,\beta)\,, \tag{2.4}$$

where Θ is the interval (a,b).

The form (2.3) was already considered by Maritz (1970) who used it to obtain an estimate for the Bayes rule in terms of the marginal distribution,

$$\delta(y) = c(y)f(y+1)/f(y)\,.$$

This is a form of a representation suggested by Robbins (1955). However, such representations suffer from "jumpiness" as was observed in Berger (1985), and must be smoothed. Our procedure will not only provide an estimator based on the marginal distribution but will also allow the user to specify the degree of smoothness desired.

A pure Bayesian approach would involve estimation of θ under the hypothesis that it has a "measure of uncertainty" given by the prior probability density (2.4). For each fixed y, the posterior probability density function of θ would be proportional to the product of (2.3) and (2.4) and would again be a member of the conjugate family. We assume that a Bayesian has chosen α and β in some way and let $g_0(\theta) = \pi_{(\alpha,\beta)}(\theta)$. This could be done either by using prior knowledge or by treating their estimation as a parametric empiric Bayes problem and using a portion of the data for this purpose.

After α and β have been chosen, perhaps both zero in the non-informative case, we "improve" the estimation by using additional data. This data is used in conjunction with appropriate orthogonal polynomials on (a,b). The orthogonality is with respect to the weight function $w(\theta)$ where

$$w(\theta) = \theta^{\alpha}\xi^{\beta+1}(\theta)I_{\Theta}(\theta)\,. \tag{2.5}$$

We need also to assume that $\int_a^b \theta^n w(\theta)d\theta$ exists for $n = 0, 1, 2, \dots$.

Now we define the polynomials $\{p_n\}$ in a standard way; we let $p_0(\theta) = 1/h_0$, where $h_0^2 = \int_a^b w(\theta)d\theta$, and

$$p_1(\theta) = \left\{\theta - \int_a^b \theta p_0(\theta)w(\theta)d\theta\right\}\Big/ h_1\,,$$

where h_1 is the normalizing factor needed to make $\int_a^b p_1^2(\theta)w(\theta)d\theta = 1$. This is the familiar Gram-Schmidt orthogonalization procedure and may be continued to all positive integers. In this way we obtain a sequence

$$p_0(\theta), p_1(\theta), \dots, p_n(\theta), \dots\,,$$

such that $p_n(\theta)$ is a polynomial of degree n which is orthogonal to $p_m(\theta)$ for $n \neq m$ and in fact if $q(\theta)$ is any polynomial of degree $m < n$, then

$$\int_a^b q(\theta)p_n(\theta)w(\theta)d\theta = 0 .$$

The assumption that the prior distribution has a density will ensure that the above procedure does not terminate.

The expansion of any function $h(\theta) \in L^2(w; (a, b))$ is the series

$$\sum_{k=0}^{\infty} a_k p_k(\theta) ,$$

where

$$a_k = \int_a^b h(\theta)p_k(\theta)w(\theta)d\theta .$$

This series converges in the sense of $L^2(w; (a, b))$ by Bessel's inequality and in fact converges to $h(\theta)$ provided (a, b) is a finite interval (see Szegö (1967), p. 44). The polynomials $p_n(\theta)$ satisfy a three term recurrence formula

$$p_n(\theta) = (A_n\theta + B_n)p_{n-1}(\theta) - C_n p_{n-2}(\theta), \qquad n = 2, 3, 4, \ldots , \tag{2.6}$$

as do all orthogonal polynomials (see Szegö (1967), p. 42).

2.1 *Auxiliary formulae*

We are interested in estimating the prior distribution, which we assume to have a density $g(\theta)$, from a sample of the mixture $f(y)$

$$\begin{aligned} f(y) &= \int_a^b p(y|\theta)g(\theta)d\theta \\ &= \sigma(y)\int_a^b \theta^y(\xi(\theta)g(\theta)/w(\theta))w(\theta)d\theta \\ &= \sigma(y)\int_a^b \theta^y h(\theta)w(\theta)d\theta . \end{aligned} \tag{2.7}$$

The density $g(\theta)$ may not always be identifiable. Any function $h(\theta)$ which is orthogonal to all polynomials with exponents in A^* will map into the zero function in (2.7). In particular, if A^* is a finite set, then any polynomial all of whose terms have degree greater than the largest integer in A^*, will have this property.

Hence we cannot recover an arbitrary $g(\theta)$ from the knowledge of $f(y)$ but must restrict the choice in some way.

If A^* is finite, then the maximum number of linearly independent $f(y)$ equals the cardinality of A^*, and hence we must restrict $h(\theta)$ to a finite dimensional subspace of $L^2(w; (a, b))$. We shall suppose $A^* = \{0, 1, 2, \ldots, n\}$ for simplicity. The restriction on $h(\theta)$ will then be that it belongs to the space spanned by $p_0(\theta), p_1(\theta), \ldots, p_n(\theta)$, i.e.,

$$h(\theta) = \sum_{k=0}^{n} a_k p_k(\theta) . \tag{2.8}$$

In this case $f(y)$ will become

$$f(y) = \sum_{k=0}^{n} a_k l_k(y) , \tag{2.9}$$

where

$$l_k(y) = \int_a^b \theta^y \sigma(y) p_k(\theta) w(\theta) d\theta, \qquad k = 0, 1, \ldots, n . \tag{2.10}$$

Clearly the $l_k(y)$ are linearly independent and therefore the a_k are uniquely determined by $f(y)$. Thus, given $f(y)$ in this form, we may find the solution $h(\theta)$ by using the coefficients from (2.9) in (2.8).

If A^* is infinite, no finite linear combination of the p_k will give all possible $f(y)$ in (2.7). We shall again assume that A^* is a set of contiguous integers $\{0, 1, 2, \ldots\}$. There now is no problem with identifiability if (a, b) is a bounded interval, since in this case, the p_k are complete in $L^2(w; (a, b))$ (Szegö (1967), p. 40). That is, if $h_1(\theta)$ and $h_2(\theta)$ are both in this space and both map into the same $f(y)$, then all the expansion coefficients of h_1 and h_2 with respect to the p_k are the same. Therefore, $h_1 - h_2$ is orthogonal to all the p_k and hence equals zero almost everywhere.

If the interval (a, b) is unbounded, it may be that the p_k are not complete. In this case we must again restrict $h(\theta)$ to a subspace of $L^2(w; (a, b))$. Now, however, we take limits of finite linear combinations of the p_k; i.e., we take the topological rather than algebraic span of our set of polynomials as the allowable set of $h(\theta)$.

In each of these cases we solve equation (2.7) for a unique $h(\theta)$ satisfying the appropriate restriction given on $f(y)$. Our procedure will involve first estimating the coefficients in (2.9) and showing the two sides are equal (n could be ∞). Then, $h(\theta)$ given by (2.8) is the unique solution.

We shall obtain the a_k by using a biorthogonal sequence $(l_k(y), \lambda_k(y))$ satisfying

$$\sum_{y \in A^*} l_k(y) \lambda_j(y) = \delta_{kj} .$$

We first observe that in both the finite and infinite case with complete set $\{p_k\}$, (2.9) holds for some choice of a_k. However, in the case of an unbounded interval in which $\{p_k\}$ is not necessarily complete, we must add this as one of our hypotheses.

The $\lambda_k(y)$ may be obtained from the orthogonal polynomials by the formula (see below)

$$\lambda_k(y) = p_k^{(y)}(0)/(y!\sigma(y)) . \tag{2.11}$$

Then we find that $\lambda_k(y) = 0$ for $k < y$ and therefore

$$\begin{aligned}\sum_y f(y)\lambda_k(y) &= \sum_{y=0}^{k}\sum_{j=0}^{y} a_j l_j(y)\lambda_k(y)\\ &= \sum_{j=0}^{k} a_j \sum_{y=j}^{k} l_j(y)\lambda_k(y) = a_k .\end{aligned}$$

In many calculations it is easier to use other formulae for l_k and λ_k. The l_k may also be given in terms of the coefficients of $p_k(\theta)$ or in terms of the recurrence formula (2.6). Indeed from (2.6) we have

$$\begin{aligned}&A_k\int_a^b \theta\theta^y\sigma(y)p_{k-1}(\theta)w(\theta)d\theta\\ &= (A_k\sigma(y)/\sigma(y+1))\int_a^b \theta^{y+1}\sigma(y+1)p_{k-1}(\theta)w(\theta)d\theta\\ &= (A_k\sigma(y)/\sigma(y+1))l_{k-1}(y+1)\\ &= \int_a^b \theta^y\sigma(y)\{p_k(\theta) - B_kp_{k-1}(\theta) + C_kp_{k-2}(\theta)\}w(\theta)d\theta\\ &= l_k(y) - B_kl_{k-1}(y) + C_kl_{k-2}(y) .\end{aligned} \tag{2.12}$$

The corresponding recurrence formula for the λ_k is given by using (2.11) and is

$$\begin{aligned}&(A_k\sigma(y-1)/\sigma(y))\lambda_{k-1}(y-1)\\ &\qquad = \lambda_k(y) - B_k\lambda_{k-1}(y) + C_k\lambda_{k-2}(y) .\end{aligned} \tag{2.13}$$

Since $p_k(\theta)$ is a polynomial of degree k, it has the form

$$p_k(\theta) = \sum_{y=0}^{k} \alpha_{ky}\theta^y ,$$

and hence

$$l_k(y) = \sum_{j=0}^{k} \alpha_{kj}\mu_{y+j}\sigma(y) , \tag{2.14}$$

where μ_j is the j-th moment of $w(\theta)$, while

$$\lambda_k(y) = \alpha_{ky}/\sigma(y) . \tag{2.15}$$

To show that the two are biorthogonal we calculate

$$\begin{aligned}\sum_{y=0}^{\infty} \lambda_j(y)l_k(y) &= \sum_{y=0}^{j} \lambda_j(y)l_k(y)\\ &= \sum_{y=0}^{j} \frac{\alpha_{jy}}{\sigma(y)}\int_a^b \theta^y p_k(\theta)w(\theta)d\theta\sigma(y)\\ &= \int_a^b \sum_{y=0}^{j} \alpha_{jy}\theta^y p_k(\theta)w(\theta)d\theta\\ &= \int_a^b p_j(\theta)p_k(\theta)w(\theta)d\theta = \delta_{kj} .\end{aligned}$$

2.2 *Estimation*

In the pure Bayesian case the posterior estimate of θ would be

$$\theta^* = \frac{\int_a^b \theta^{y+1}\sigma(y)\xi(\theta)g(\theta)d\theta}{\int_a^b \theta^y\sigma(y)\xi(\theta)g(\theta)d\theta} = \frac{\sigma(y)}{\sigma(y+1)}\cdot\frac{f(y+1)}{f(y)} . \tag{2.16}$$

In the empiric Bayes case one estimates $g(\theta)$ (or the resulting $f(y)$) from a sample $Y_1, Y_2,\ldots, Y_N$ and then uses this estimate of g to obtain an estimate $\hat{\theta}$ from another observation Y_{N+1}. Thus (2.16) would become

$$\hat{\theta} = \frac{\sigma(y)}{\sigma(y+1)}\cdot\frac{\hat{f}(y+1)}{\hat{f}(y)} , \quad Y_{N+1} = y , \tag{2.17}$$

in the empiric Bayes formulation.

Berger (1985) has objected to this estimator when $\hat{f}$ is f_N, the empiric distribution, because of the "jumpiness" of $\hat{\theta}$. However, we shall use smoother estimates based on polynomial expansion of f_N in terms of the $l_k(y)$. This will also enable us to estimate other Bayesian outputs such as the posterior variance or posterior confidence intervals by the same method and thus meet another of Berger's objections.

We begin by choosing α and β as indicated previously and we then define the weight of (2.5). This results in a preliminary Bayesian estimate of θ given by (2.16) in which

$$g_0(\theta) = Cp_0(\theta)w(\theta)/\xi(\theta)\,, \tag{2.18}$$

where C is a positive constant. This estimate is independent of the sample $Y_1, Y_2, \ldots, Y_N$ and uses only the value of $Y_{N+1} = y$. The sample is then used to estimate the expansion coefficients of $h(\theta) = g(\theta)/\theta^{\alpha}\xi^{\beta}(\theta)$ with respect to the orthogonal polynomials $p_k(\theta)$,

$$a_k = \int_a^b h(\theta)p_k(\theta)w(\theta)d\theta = \int_a^b g(\theta)p_k(\theta)\xi(\theta)d\theta\,. \tag{2.19}$$

These coefficients a_k are estimated by

$$\hat{a}_k = \frac{1}{N}\sum_{i=1}^{N}\lambda_k(Y_i), \qquad k = 0, 1, 2, \ldots\,, \tag{2.20}$$

which gives as an estimator of $g(\theta)$

$$\hat{g}_m(\theta) = \sum_{k=0}^{m}\hat{a}_k p_k(\theta)w(\theta)/\xi(\theta), \qquad m = 0, 1, 2, \ldots\,, \tag{2.21}$$

and of $f(y)$

$$\hat{f}_m(y) = \sum_{k=0}^{m}\hat{a}_k l_k(y), \qquad m = 0, 1, 2, \ldots\,. \tag{2.22}$$

For $m = 0$, (2.21) is the same as (2.18) while as m increases, the estimator becomes progressively less smooth. For $m = n$, the size of A^* in the finite case, the estimator of $f(y)$ is exactly the empiric distribution. If A^* is infinite the estimator approaches the empiric distribution as $N \to \infty$.

2.3 *Asymptotic optimality*

We now assume that the discrete set A^* consists of successive integers $0, 1, 2, \ldots, n$ where n is finite or infinite. We cannot quite do so without loss of generality, that is, without losing the parametric form of our family given by (2.3). We also assume that $a \leq 1 \leq b$ and that $\sigma(y)$ is a probability function, merely by shifting the origin in (2.3) which we can do without loss of generality.

We first observe that $\hat{a}_k$ is an unbiased estimator of a_k, the expansion coefficient of $h(\theta)$

$$E[\hat{a}_k] = \frac{1}{N}\sum_{i=1}^{N}E[\lambda_k(Y_i)] = \sum_{y=0}^{k}\lambda_k(y)f(y) = a_k\,,$$

and that the variance of $\hat{a}_k$ is

$$E[(\hat{a}_k - a_k)^2] = \frac{1}{N}\sum_{y=0}^{k}\lambda_k^2(y)f(y) - \frac{a_k^2}{N} \le \frac{C_k}{N}, \tag{2.23}$$

where $C_k = \max_{0 \le y \le k} \lambda_k^2(y)$. Our estimate of the prior density function (2.21) may be expressed as

$$\hat{g}_m(\theta) = \hat{h}_m(\theta)w(\theta)/\xi(\theta), \tag{2.24}$$

where

$$\hat{h}_m(\theta) = \sum_{k=0}^{m} \hat{a}_k p_k(\theta). \tag{2.25}$$

Hence the mean square error of $\hat{h}_m(\theta)$ is

$$\begin{aligned} E[(\hat{h}_m(\theta) - h(\theta))^2] &= E\left[\sum_{k=0}^{m}(\hat{a}_k - a_k)p_k(\theta)\right]^2 + \left[\sum_{k=m+1}^{\infty} a_k p_k(\theta)\right]^2 \\ &\le \sum_{k=0}^{m} p_k^2(\theta)\sum_{k=0}^{m} E[(\hat{a}_k - a_k)^2] + \left[\sum_{k=m+1}^{\infty} a_k p_k(\theta)\right]^2, \end{aligned} \tag{2.26}$$

and the integrated mean square error with respect to $w(\theta)$ is simply

$$\begin{aligned} &\int_a^b E[(\hat{h}_m(\theta) - h(\theta))^2]w(\theta)d\theta \\ &\le \sum_{k=0}^{m}\int_a^b p_k^2(\theta)w(\theta)d\theta \sum_{k=0}^{m}\frac{C_k}{N} + h_0^2\sum_{k=m+1}^{\infty} a_k^2 \\ &= h_0^2\left(\frac{m}{N}\sum_{k=0}^{m} C_k + \sum_{k=m+1}^{\infty} a_k^2\right). \end{aligned} \tag{2.27}$$

If we assume that $h \in L^2(w; (a, b))$ and (a, b) is finite, then

$$\sum_{k=m+1}^{\infty} a_k^2 \to 0 \quad \text{as} \quad m \to \infty .$$

Hence if m is chosen sufficiently large to make $\sum_{k=m+1}^{\infty} a_k^2 < \varepsilon/2$, and the sample size N then chosen large enough to make the first term in (2.27) less than $\varepsilon/2$, then the integrated mean square error can be made less than ε. We observe that N depends on m and as m tends to infinity so does N to assure the convergence of (2.27).

This procedure works if there are an infinite number of elements in A^* since $\hat{a}_k$ is defined for every integer. However, if A^* has only $n+1$ points, then $\hat{a}_k = 0$ for $k > n$ since $\lambda_k = 0$ in that case. Thus $\hat{g}_m = \hat{g}_n$ for all $m \geq n$ and (2.27) holds only for $m \leq n$. We therefore require in this case that $h(\theta)$ be a polynomial of degree $\leq n$.

A similar assumption will be made in the case of an infinite interval (a, b) since then the system $\{p_n\}$ is not necessarily complete.

We summarize the various assumptions we have made:

ASSUMPTIONS 2.1.

(i) The conditional probability $p(y|\theta)$ belongs to a discrete family of the form

$$p(y|\theta) = \theta^y \xi(\theta)\sigma(y) \,,$$

where $y = 0, 1, \ldots, n$; $n = \text{card } A^* \leq \infty$; $\theta \in (a, b)$; and $\xi(\theta)$ and $\sigma(y)$ are positive measurable functions.

(ii) A conjugate prior density of the form

$$\pi_{(\alpha,\beta)}(\theta) \sim \theta^\alpha \xi^\beta(\theta) \,,$$

has been found.

(iii) The integrals

$$\int_a^b \theta^{\alpha+k} \xi^{\beta+1}(\theta) d\theta \,,$$

exist for all integers $k = 0, 1, \ldots$.

(iv) The true prior density

$$g(\theta) = h(\theta)\theta^\alpha \xi^\beta(\theta) \,,$$

is such that $h(\theta)$ is in the closure in $L^2(w; (a, b))$ of the space spanned by $p_0(\theta), p_1(\theta), \ldots, p_n(\theta)$, $n = \text{card } A^*$.

We can now restate the results of the above calculations as:

THEOREM 2.1. *Let Assumptions* 2.1 *hold. Then*

$$\int_a^b E|\hat{g}_m(\theta) - g(\theta)|^2 \theta^{-\alpha} \xi^{1-\beta}(\theta) d\theta \to 0 \quad \textit{as} \quad m \to \infty \,,$$

where $\hat{g}_m(\theta)$ is the estimate (see (2.21)) *of the true prior density $g(\theta)$ based on a sample of size $N(m)$ with common probability function $f(y)$, $y \in A^*$ (see* (2.7)). *Furthermore, $N(m) \to \infty$ as $m \to \infty$.*

COROLLARY 2.1. *Let $Y_1, Y_2, \ldots, Y_N$ be as in Theorem* 2.1, *and let $\hat{\theta}_m$ be given by*

$$\hat{\theta}_m = \left[\frac{1}{\hat{f}_m(y)} \int_a^b \theta^{y+1} \sigma(y) \xi(\theta) \hat{g}_m(\theta) d\theta \right]_{(a,b)},$$

where the subscript (a, b) denotes the restriction to (a, b) of the preceding expression. Then as $m \to \infty$ (and hence $N \to \infty$),

$$E[(\hat{\theta}_m - \theta^*)^2] \to 0,$$

i.e., $\hat{\theta}_m$ is asymptotically optimal.

PROOF. We observe that

$$\text{(2.28)} \quad E[(\hat{\theta}_m - \theta^*)^2] \leq \left(\frac{\sigma(y)}{\sigma(y+1)} \right)^2 E\left[\left(\frac{\hat{f}_m(y+1)}{\hat{f}_m(y)} - \frac{f(y+1)}{f(y)} \right)^2\right]$$
$$\leq \left(\frac{\sigma(y)}{\sigma(y+1)} \right)^2 \frac{2^\gamma}{f^\gamma(y)} \left\{ E|\hat{f}_m(y+1) - f(y+1)|^\gamma \right.$$
$$\left. + \left(\frac{f^\gamma(y+1)}{f^\gamma(y)} + 1 \right) E|\hat{f}_m(y) - f(y)|^\gamma \right\},$$

where $0 < \gamma < 1$. Since

$$\text{(2.29)} \qquad E|\hat{f}_m(y) - f(y)|^\gamma = E\left| \int_a^b \theta^y \sigma(y)(\hat{g}_m(\theta) - g(\theta))\xi(\theta) d\theta \right|^\gamma$$
$$\leq \left\{ \int_a^b \theta^{2y} \sigma^2(y) w^2(\theta) d\theta \right\}^{\gamma/2}$$
$$\cdot \left\{ E\int_a^b |\hat{h}_m(\theta) - h(\theta)|^2 w(\theta) d\theta \right\}^{\gamma/2},$$

which converges to 0, we have $E[(\hat{\theta}_m - \theta^*)^2] \to 0$ as $m \to \infty$.

The same considerations lead to:

COROLLARY 2.2. *Let $Y_1, Y_2, \ldots, Y_N$ be as in Theorem* 2.1, *and let $V = E_{\pi(\theta|y)}[(\theta - \theta^*)^2]$ be the posterior variance, Let $\hat{V}_m$ be given by*

$$\hat{V}_m = \left[\frac{1}{\hat{f}_m(y)} \int_a^b (\theta - \hat{\theta}_m)^2 \theta^y \sigma(y) \xi(\theta) \hat{g}_m(\theta) d\theta \right]_{(0,\infty)},$$

then if $b < \infty$,

$$E[(\hat{V}_m - V)^2] \to 0 \quad \textit{as} \quad m \to \infty .$$

If $h(\theta)$ is in the algebraic span of the polynomials $p_0, p_1, \ldots, p_m$, the problem reduces to a parametric one with $m + 3$ parameters. In this case we have:

COROLLARY 2.3. *Let $Y_1, Y_2, \ldots, Y_N$ be as in Theorem 2.1, let $h(\theta)$ be a polynomial of fixed degree m, then*

$$\int_a^b E|\hat{g}_m(\theta) - g(\theta)|^2 \theta^{-\alpha} \xi^{1-\beta}(\theta) d\theta = O\left(\frac{1}{N}\right).$$

2.4 *Non-informative prior*

If as frequently happens the Bayesian is unwilling or unable to choose α and β, but is equally happy with all choices, then the prior $g_0(\theta)$ becomes a constant. This of course will be improper if (a, b) is an infinite interval. However, the empiric Bayes part of our analysis is still valid in this case. Indeed, the weight function for our orthogonal polynomials is just

$$w(\theta) = \xi(\theta) = e^{-\psi(\ln\theta)} ,$$

where $\psi(\phi)$ is the cumulant generating function on (a, b) and may be given by (Morris (1982))

$$\psi(\phi) = \ln \sum_y e^{\phi y} \sigma(y) = \ln \Sigma \theta^y \sigma(y) .$$

Hence $w^{-1}(\theta) = \sum_{y=0}^{\infty} \theta^y \sigma(y)$.

Morris (1982), constructed polynomials in y and $\mu = \psi'(\phi)$ which are orthogonal with respect to the weight (exp $\{\phi y - \psi(\phi)\}$)$\sigma(y)$. In some cases these can be modified to correspond to the orthogonal polynomials we have used. However, since $w(\theta)$ is not in general a polynomial function, this does not always work.

It should also be remarked that since $w^{-1}(\theta)$ is given by a power series, our family of distributions is also a power series family and results on identifiability for such families can be used (see Patil *et al.* (1975)).

3. Some examples

In this section we consider particular instances of discrete distributions whose probability functions can be put in the form of an exponential family. This will include the binomial, Poisson, negative binomial distributions. In the case of the binomial we use the traditional parameterization rather than converting to the form of the last section; since the latter form will lead to non-standard orthogonal polynomials which would of course be less desirable.

3.1 *Binomial distribution*

This has been studied in Walter and Hamedani (1987). The conditional probability function is

$$(3.1) \qquad p(x|\theta) = B_n(x,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, \ldots, n\,,$$

with conjugate prior given by the Beta density

$$(3.2) \qquad \pi_{(\alpha,\beta)}(\theta) = \theta^\alpha(1-\theta)^\beta / B(\alpha,\beta), \qquad 0 \le \theta \le 1, \quad \alpha,\beta > -1\,.$$

The associated orthogonal polynomials are the well-known Jacobi polynomials $p_k^{(\beta,\alpha)}(\theta)$ satisfying the Rodrigues formula

$$p_k^{(\beta,\alpha)}(\theta) = \frac{(-1)^k}{k!\theta^\alpha(1-\theta)^\beta}\frac{d^k}{d\theta^k}\{\theta^{\alpha+k}(1-\theta)^{\beta+k}\}, \qquad k = 0, 1, 2, \ldots\,.$$

They are orthogonal with respect to the weight function

$$w(\theta) = \theta^\alpha(1-\theta)^\beta\,.$$

In Walter and Hamedani (1987) it was shown that the estimator

$$\hat{\theta}_{n-1} = \left[\frac{\int_0^1 \theta B_{n-1}(x,\theta)\hat{g}_m(\theta)d\theta}{\hat{f}_{n-1}(x)}\right]_{[0,1]},$$

is asymptotically optimal in the sense that

$$E[(\hat{\theta}_{n-1} - \theta^*_{n-1})^2] \to 0 \qquad \text{as} \qquad N \to \infty\,,$$

where θ^*_{n-1} is the Bayes rule for the conditional probabilities given by $B_{n-1}(x,\theta)$. For large n, the difference is negligible. The necessity of using $n-1$ instead of n can be avoided if the prior distribution is assumed to have a density of the form

$$\pi_{(\alpha,\beta)}(\theta) = q(\theta)\theta^\alpha(1-\theta)^\beta\,,$$

where $q(\theta)$ is a polynomial of degree n or less, as was assumed in Theorem 2.1.

Alternatively one could use the parameterization of the last section which in the binomial case would be

$$\theta' = \frac{\theta}{1-\theta},$$

and

$$B_n(x,\theta) = p(x|\theta') = \binom{n}{x}(\theta')^x(1+\theta')^{-n}.$$

However, as was mentioned before, this would lead to non-standard orthogonal polynomials.

3.2 *Poisson distribution*

Compound Poisson distributions were studied in Tucker (1963) and in Walter (1985), in the latter case from a point of view similar to that considered in this work. The conditional probability function is already in the form of Section 2,

$$p(x|\theta) = \theta^x e^{-\theta}/x!, \qquad x = 0,1,2,\dots, \tag{3.3}$$

and the conjugate prior is given by the Gamma density function

$$\pi_{(\alpha,\beta)}(\theta) = \theta^\alpha e^{-\beta\theta}/\Gamma(\alpha,\beta), \qquad 0 < \theta < \infty. \tag{3.4}$$

Hence the weight function for the orthogonal polynomials is

$$w(\theta) = \theta^\alpha e^{-(\beta+1)\theta} = [(\beta+1)\theta]^\alpha e^{-(\beta+1)\theta}/(\beta+1)^\alpha. \tag{3.5}$$

This is the weight function for a well-known family of orthogonal polynomials, the Laguerre polynomials $L_k^\alpha((\beta+1)\theta)$ where

$$L_k^\alpha(\eta) = \frac{\eta^\alpha e^\eta}{k!} D_\eta^k(e^{-\eta}\eta^{k+\alpha}), \qquad k = 0,1,2,\dots. \tag{3.6}$$

In this case we have

$$l_k^{(\alpha,\beta)}(x) = \int_0^\infty \frac{\theta^x}{x!} L_k^\alpha((\beta+1)\theta)w(\theta)d\theta \tag{3.7}$$

$$= \int_0^\infty \frac{\eta^{x+\alpha} L_k^\alpha(\eta) e^{-\eta}}{x!(\beta+1)^{x+\alpha+1}} d\eta$$

$$= (-1)^k \frac{\Gamma(x+\alpha+1)}{x!(\beta+1)^{x+\alpha+1}} \binom{x}{k},$$

from equation (2.3) in Walter (1985). The orthogonal polynomials are not normalized, but can be made so by using

$$C_k^{(\alpha,\beta)} = \int_0^\infty [L_k^\alpha((\beta+1)\theta)]^2 w(\theta) d\theta \tag{3.8}$$

$$= \frac{\Gamma(\alpha+1)}{(\beta+1)^{\alpha+1}} \binom{k+\alpha}{k}, \qquad k = 0, 1, 2, \ldots ,$$

again from Walter (1985). The biorthogonal functions $\lambda_k^{(\alpha,\beta)}$ are particularly simple in this case. Indeed we have

$$l_k^{(\alpha,\beta)}(x) / C_k^{(\alpha,\beta)} = (-1)^k \binom{x+\alpha}{k+\alpha}, \tag{3.9}$$

and hence

$$\sum_{x=k}^{m} \frac{l_x^{(\alpha,\beta)}(m)}{C_x^{(\alpha,\beta)}} \cdot \frac{l_m^{(\alpha,\beta)}(x)}{C_k^{(\alpha,\beta)}} = \delta_{mk} . \tag{3.10}$$

Thus, if we normalize the polynomial by dividing by $C_k^{(\alpha,\beta)}$, we find that

$$\lambda_m^{(\alpha,\beta)}(k) = l_k^{(\alpha,\beta)}(m) / C_k^{(\alpha,\beta)}, \qquad k \le m . \tag{3.11}$$

Thus, the estimator for the prior density

$$g(\theta) = \sum_{k=0}^{\infty} a_k L_k^\alpha((\beta+1)\theta) \theta^\alpha e^{-\beta\theta} / C_k^{(\alpha,\beta)} ,$$

is of the form

$$\hat{g}_m(\theta) = \sum_{k=0}^{m} \hat{a}_k L_k^\alpha((\beta+1)\theta) \theta^\alpha e^{-\beta\theta} / C_k^{(\alpha,\beta)} , \tag{3.12}$$

where

$$\hat{a}_k = \frac{1}{N} \sum_{i=1}^{N} \lambda_k^{(\alpha,\beta)}(X_i) , \tag{3.13}$$

based on the sample $X_1, X_2, \ldots, X_N$ from a distribution with probability function

$$f(x) = \int_0^\infty \frac{\theta^x}{x!} e^{-\theta} g(\theta) d\theta .$$

Our general theory tells us that $\theta^{-\alpha} e^{\beta\theta} \hat{g}_m(\theta)$ is integrated mean square consistent with weight (3.5) and that

$$\hat{\theta}_m = \max \left\{ \frac{\int_0^\infty \theta^{x+1} e^{-\theta} \hat{g}_m(\theta) d\theta}{\hat{f}_m(x)}, 0 \right\}, \tag{3.14}$$

is asymptotically optimal.

3.3 *Negative binomial distribution*

The conditional probability function in this case has the form

$$p(x|\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r, \qquad x = 0, 1, 2, \ldots , \tag{3.15}$$

in which $(1-\theta)$ corresponds to the probability of success and (3.15) gives the probability that x failures occur before the r-th success. In this interpretation r is a fixed positive integer and θ is the parameter. The conjugate prior family is again a Beta family

$$\pi_{(\alpha,\beta)}(\theta) = \theta^\alpha (1-\theta)^\beta / B(\alpha, \beta), \qquad 0 < \theta < 1, \quad \alpha, \beta > -1 .$$

However, the weight function for the orthogonal polynomials will be

$$w(\theta) = \theta^\alpha (1-\theta)^{\beta+r},$$

which again leads to the Jacobi polynomials, in this case

$$\{p_k^{(\beta+r,\alpha)}(\theta)\} .$$

Hence $l_k^{(\alpha,\beta)}$ is given by

$$\begin{aligned} l_k^{(\alpha,\beta)}(x) &= \int_0^1 \binom{r+x+1}{x} \theta^x p_k^{(\beta+r,\alpha)}(\theta) \theta^\alpha (1-\theta)^{\beta+r} d\theta \\ &= \binom{r+x-1}{x} \sum_{j=0}^k a_{kj} \int_0^1 \theta^{j+x+\alpha} (1-\theta)^{\beta+r} d\theta \end{aligned} \tag{3.16}$$

$$=\binom{r+x-1}{x}\sum_{j=0}^{k}a_{kj}\frac{\Gamma(j+x+\alpha+1)\Gamma(\beta+r+1)}{\Gamma(j+x+\alpha+\beta+r+2)}\,.$$

Since this is exactly the setting of Section 2, we may choose $\lambda_k^{(\alpha,\beta)}$ by

$$\lambda_k^{(\alpha,\beta)}(x)=\begin{cases} c_k a_{kx}/\binom{r+x-1}{x}, & x\le k\,,\\ 0, & x>k\,,\end{cases}$$

and find that

$$(3.17)\qquad \sum_{x=0}^{k}\lambda_m^{(\alpha,\beta)}(x)l_k^{(\alpha,\beta)}(x)$$

$$=\int_0^1 C_m p_m^{(\beta+r,\alpha)}(\theta)p_k^{(\beta+r,\alpha)}(\theta)\theta^{\alpha}(1-\theta)^{\beta+r}d\theta$$

$$=C_k\delta_{mk}d_k^{(\beta+r,\alpha)}\,.$$

The normalizing factor $d_k^{(\beta+r,\alpha)}$ (see Szegö (1967)) is given by

$$(3.18)\quad d_k^{(\beta+r,\alpha)}=\frac{\Gamma(k+\beta+r+1)\Gamma(k+\alpha+1)}{\Gamma(k+1)\Gamma(k+\alpha+\beta+r+1)(2k+\alpha+\beta+r+1)}\,.$$

This may be combined with $\lambda_k^{(\alpha,\beta)}$ by taking

$$C_k=1/d_k^{(\beta+r,\alpha)}\,,$$

to obtain the desired biorthogonality. Thus the estimator for the prior density is

$$(3.19)\qquad \hat{g}_m(\theta)=\sum_{k=0}^{m}\hat{a}_k p_k^{(\beta+r,\alpha)}(\theta)\theta^{\alpha}(1-\theta)^{\beta}\,,$$

where, given a sample $X_1, X_2,\ldots, X_N$ from

$$f(x)=\int_0^1\binom{r+x-1}{x}\theta^x(1-\theta)^r g(\theta)d\theta\,,$$

we take

$$(3.20)\qquad \hat{a}_k=\frac{1}{N}\sum_{i=1}^{N}\frac{a_{kX_i}}{\binom{r+X_i-1}{X_i}d_k^{(\beta+r,\alpha)}}\,.$$

This $\hat{g}_m$ is integrated mean square consistent with weight $\theta^{-\alpha}(1-\theta)^{1-\beta}$ and the empiric Bayes posterior estimate of θ, the restriction to [0, 1] of

$$\hat{\theta}_m = \frac{\int_0^1 \theta^{x+1} \binom{r+x-1}{x} (1-\theta)^r \hat{g}_m(\theta) d\theta}{\hat{f}_m(x)}, \tag{3.21}$$

is asymptotically optimal.

REFERENCES

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

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

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*, Holden Day, San Francisco.

Blum, J. R. and Susarla, V. (1977). Estimation of a mixing distributions function, *Ann. Probab.*, **5**, 200–209.

Choi, K. and Bulgren, W. G. (1968). An estimation procedure for mixture of distributions, *J. Roy. Statist. Soc. Ser. B*, **30**, 444–460.

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.

Maritz, J. S. (1970). *Empirical Bayes Methods*, Methum, London.

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

O'Bryan, T. and Walter, G. (1979). Mean square estimation of the prior distribution, *Sankhyā Ser. A*, **41**, 95–108.

Patil, G. P., Kotz, S. and Ord, J. K. (1975). *Statistical Distributions in Scientific Work*, 3, D. Reidel, Dordrecht.

Robbins, H. (1955). An empiric Bayes approach to statistics, *Proc. Third Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 157–164, Univ. of California Press, Berkeley.

Robbins, H. (1963). The empirical Bayes' approach to testing statistical hypothesis, *Rev. Int. Statist. Inst.*, **31**, 195–208.

Szegö, G. (1967). *Orthogonal Polynomials*, American Math. Soc., Providence, Rhode Island.

Tortorella, M. and O'Bryan, T. (1979). Estimation of the prior distribution by best approximation in uniformly convex function spaces, *Bull. Inst. Math. Acad. Sinica*, **7**, 69–85.

Tucker, H. G. (1963). An estimate of the compounding distribution of a compound Poisson distribution, *Theory Probab. Appl.*, **8**, 195–200.

Walter, G. G. (1985). Orthogonal polynomial estimators of the prior distribution of a Compound Poisson distribution, *Sankhyā Ser. A*, **47**, 222–230.

Walter, G. G. and Hamedani, G. G. (1987). Empiric Bayes estimation of binomial probability, *Comm. Statist. A—Theory Methods*, **16**, 559–577.