The Annals of Statistics
1979, Vol. 7, No. 3, 558–568

EMPIRICAL BAYES ESTIMATION OF A BINOMIAL PARAMETER VIA MIXTURES OF DIRICHLET PROCESSES

BY DONALD A. BERRY[1] AND RONALD CHRISTENSEN

University of Minnesota

The theory of Dirichlet processes is applied to the empirical Bayes estimation problem in the binomial case. The approach is Bayesian rather than being empirical Bayesian. When the prior is a Dirichlet process the posterior is a mixture of Dirichlet processes. Explicit estimators are given for the case of 2 and 3 parameters and compared with other empirical Bayes estimators by way of examples. Since the number of calculations become enormous when the number of parameters gets larger than 2 or 3 we propose two approximations for estimators of a particular parameter and compare their performance using examples.

1. Introduction. The random variables X_i, $i = 1, \cdots, k$, are binomially distributed with parameters n_i and θ_i and are independent given $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$. The values n_i are known and $\boldsymbol{\theta}$ is a vector of chance observations, each component from a probability distribution G which concentrates its mass on [0, 1]. The measure G and the values of $\theta_1, \ldots, \theta_k$ are unknown. The empirical Bayes problem is to make inferences about G, or a particular θ_i, using the information from $\mathbf{X} = (X_1, \cdots, X_k)$.

There are many applications of empirical Bayes problems. We are particularly interested in clinical trials in which k is the number of medical centers in a multicenter study or the number of strata in the population being considered, or the product of the two. The stratification may be so fine that each individual falls in his own stratum in which case $n_i = 1$ for all i.

Most procedures (see Maritz 1970) for estimating θ_k, say, first use $X_1, \cdots, X_{k-1}$ and $n_1, \cdots, n_{k-1}$ to estimate G. This estimate $\hat{G}$ is taken as the prior distribution of θ_k which is modified by (X_k, n_k) using Bayes' theorem to obtain an estimate of the posterior distribution of θ_k:

$$d\hat{H}(\theta_k | X_k, n_k) \propto \theta_k^{X_k}(1 - \theta_k)^{n_k - X_k}\, d\hat{G}(\theta_k).$$

The approach taken here is in the mainstream of Bayesian statistics, rather than being empirical Bayesian, in that G is regarded as a random probability measure, the distribution of which is modified by $(X_1, n_1; \cdots; X_k, n_k)$ in making inferences about G or θ_k, say, where in the latter case (X_k, n_k) plays a special role. In particular, following the approach developed by Thomas Ferguson, G is assumed to have a Dirichlet process prior distribution.

Received June 1977; revised April 1978.

[1]This author's research supported by NIH Grant No. 5R01-GM 22234-02.

AMS 1970 *subject classifications.* Primary 62C10; secondary 62F15, 62F10.

Key words and phrases. Empirical Bayes estimation, Dirichlet processes, mixtures of Dirichlet processes, approximating mixtures of Dirichlet processes, binomial parameter estimation.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

The reader is referred to Ferguson (1973) and Antoniak (1974) for the fundamentals of Dirichlet processes and mixtures of Dirichlet processes, respectively. These will be essential for following the details of the current development. A theory for the general empirical Bayes problem is presented in the next section. The binomial problem for the special cases $k = 2$ and 3 are treated at length in Section 3. Since, for k large, there are an enormous number of terms to consider in the mixture of Dirichlet processes we propose an approximation of a Dirichlet mixture with a single Dirichlet process in Section 4. In Section 5 we propose two approximations for the problem of estimating a particular θ_i; one uses the approximation developed in Section 4 and the other appeals directly to the results for $k = 2$ obtained in Section 3. These estimators are compared with other empirical Bayes estimators using examples.

2. The empirical Bayes problem. Let the prior distribution of the probability measure G be a Dirichlet process with parameter α, written $G \sim \mathfrak{D}(\alpha)$. Then, as shown by Antoniak (1974), the posterior distribution is a mixture of Dirichlet processes:

$$G|\mathbf{X} \sim \int \mathfrak{D}\big(\alpha + \big(\Sigma_1^k \delta_{\theta_i}\big) \, dF_{\boldsymbol{\theta}|\mathbf{X}} \tag{1}$$

where the measure δ_θ assigns mass 1 to θ and 0 elsewhere, $F_{\boldsymbol{\theta}|\mathbf{X}}$ is the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{X}$.

For example, consider the binomial case with $k = 2$ and take $\alpha = MBe(a, b)$, where M is a positive constant and $Be(a, b)$ is the measure on [0, 1] that has a beta distribution with parameters a and b. Then

$$G|X_1, X_2 \sim \int_{[0,1]^2} \mathfrak{D}\big(MBe(a, b) + \delta_{\theta_1} + \delta_{\theta_2}\big) \, dF_{\boldsymbol{\theta}|\mathbf{X}}(\theta_1, \theta_2)$$

and, applying Proposition 1 of Antoniak (1974),

$$\theta_1 \sim Be(a, b), \quad \theta_2|\theta_1 \sim \frac{1}{M + 1}\big(MBe(a, b) + \delta_{\theta_1}\big).$$

The joint prior distribution of (θ_1, θ_2) is a measure on the unit square that is a weighted sum of the product of two beta measures with parameters a and b and a measure concentrated on the line $\theta_1 = \theta_2$ that has a $Be(a, b)$ distribution. The respective weights are $P(\theta_1 \neq \theta_2) = M/(M + 1)$ and $P(\theta_1 = \theta_2) = 1/(M + 1)$.

In finding the posterior distribution $F_{\boldsymbol{\theta}|\mathbf{X}}$ it will be shown that Bayes' theorem can be applied separately to each part of the measure on (θ_1, θ_2) and then the parts can be combined with their respective posterior weights. Since the general case is no more difficult than the specific, this will be shown in more generality than is needed here. We take X and θ to be univariate with $\theta \in R$, but the proofs hold with only minor modifications when X and θ are multivariate.

Assume $F_{X|\theta=y}$ is a discrete probability distribution function almost surely (dF_θ) defined so that

$$P(X \leqslant x, \theta \in S) = \int_S F_{X|\theta=y}(x) \, dF_\theta(y),$$

for all measurable S, and that it is absolutely continuous with respect to counting measure m. Then the Radon-Nikodym derivative $f_{X|\theta=y}$ exists. It is assumed that $f_{X|\theta=y}(x)$ is a Borel-measurable function of y for all x. The following four propositions are standard and are given without proof.

PROPOSITION 1. *The measure corresponding to $F_X(x)$ is absolutely continuous with respect to counting measure and its Radon-Nikodym derivative $f_X(x)$ is*

$$f_X(x) = \int_\Theta f_{X|\theta=y}(x)\, dF_\theta(y) \qquad \text{a.s.} \quad (m).$$

PROPOSITION 2. *For any measurable S and B,*

$$P(\theta \in S, X \in B) = \int_S \int_B dF_{X|\theta=y}(x)\, dF_\theta(y).$$

The next two propositions are variations of Bayes' theorem.

PROPOSITION 3. *The distribution function of θ given X is*

$$F_{\theta|X=x}(y) = \frac{\int_{(-\infty,y]} f_{X|\theta=t}(x)\, dF_\theta(t)}{\int_\Theta f_{X|\theta=t}(x)\, dF_\theta(t)} \qquad \text{a.s.} \quad (dF_X).$$

For measurable S such that $P(\theta \in S|X = x) > 0$ a.s. (dF_X), define

$$F_{\theta|X=x,\theta\in S}(y) = \frac{P((\theta \leqslant y) \cap (\theta \in S)|X = x)}{P(\theta \in S|X = x)}$$

and

$$F_{\theta|\theta\in S}(y) = \frac{P((\theta \leqslant y) \cap (\theta \in S))}{P(\theta \in S)}.$$

For measurable S such that $0 < P(\theta \in S|X = x) < 1$ a.s. (dF_X),

$$\begin{aligned} (2) \qquad F_{\theta|X=x,\theta\in S}(y)P(\theta \in S|X = x) + F_{\theta|X=x,\theta\notin S}(y)P(\theta \notin S|X = x) \\ = F_{\theta|X=x}(y) \qquad \text{a.s.} \quad (dF_X). \end{aligned}$$

PROPOSITION 4.

$$F_{\theta|X=x,\theta\in S}(y) = \frac{\int_{(-\infty,y]} f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t)}{\int_\Theta f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t)} \qquad \text{a.s.} \quad (dF_X).$$

Equation (2) and Proposition 4 show that we can apply Bayes' theorem to the separate parts of the distribution of (θ_1, θ_2) and recombine them with the appropriate posterior weights. The next proposition demonstrates how to compute the posterior weights.

PROPOSITION 5. *For measurable S such that $P(\theta \in S|X = x) \in (0, 1)$ a.s. (dF_X),* $P(\theta \in S|X = x)$

$$= \frac{P(\theta \in S)\int_S f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t)}{P(\theta \in S)\int_\Theta f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t) + P(\theta \notin S)\int_\Theta f_{X|\theta=t}(x)\, dF_{\theta|\theta\notin S}(t)}.$$

PROOF. Using a relationship similar to Equation (2),

$$P(\theta \in S|X = x) = \int_S dF_{\theta|X=x}(y) = \frac{\int_S f_{X|\theta=t}(x)\, dF_\theta(t)}{\int_\Theta f_{X|\theta=t}(x)\, dF_\theta(t)}$$

$$= \frac{P(\theta \in S)\int_S f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t) + P(\theta \notin S)\int_S f_{X|\theta=t}(x)\, dF_{\theta|\theta\notin S}(t)}{P(\theta \in S)\int_\Theta f_{X|\theta=t}(x)\, dF_{\theta|\theta\in S}(t) + P(\theta \notin S)\int_\Theta f_{X|\theta=t}(x)\, dF_{\theta|\theta\notin S}(t)}.$$

The result follows since the second term in the numerator is zero. □

Proposition 5 will be applied to the binomial problem with $k = 2$ and $k = 3$ in the next section.

3. The binomial problem. The simplest case is $k = 1$:

$$E(\theta_1|X_1) = \frac{a + X_1}{a + b + n_1}, \tag{3}$$

$$E(G|X_1) = \frac{M}{M + 1} Be(a, b) + \frac{1}{M + 1} Be(a + X_1, b + n_1 - X_1).$$

If $k = 2$ then Proposition 5 yields, in terms of odds,

$$\frac{P(\theta_1 = \theta_2|X_1, X_2)}{P(\theta_1 \neq \theta_2|X_1, X_2)}$$

$$= \frac{P(\theta_1 = \theta_2)\dfrac{\Gamma(a + X_1 + X_2)\Gamma(b + n_1 + n_2 - X_1 - X_2)}{\Gamma(n_1 + n_2 + a + b)}\dfrac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}}{P(\theta_1 \neq \theta_2)\dfrac{\Gamma(a + X_1)\Gamma(b + n_1 - X_1 + b)\Gamma(a + X_2)\Gamma(b + n_2 - x_2)\Gamma(a + b)\Gamma(a + b)}{\Gamma(a + b + n_1)\Gamma(a + b + n_2)\Gamma(a)\Gamma(b)\Gamma(a)\Gamma(b)}}$$

$$= \frac{a^{(X_1+X_2)}b^{(n_1+n_2-X_1-X_2)}}{(a + b)^{(n_1+n_2)}} \Big/ \frac{Ma^{(X_1)}a^{(X_2)}b^{(n_1-X_1)}b^{(n_2-X_2)}}{(a + b)^{(n_1)}(a + b)^{(n_2)}},$$

where $c^{(n)} = c(c + 1) \cdots (c + n - 1)$.

Let $p \equiv P(\theta_1 \neq \theta_2|X_1, X_2)$, then

$$E(\theta_2|X_1, X_2) = p\left(\frac{a + X_2}{a + b + n_2}\right) + (1 - p)\left(\frac{a + X_1 + X_2}{a + b + n_1 + n_2}\right) \tag{4}$$

and

$$E(G|X_1, X_2) =$$

$$p\left[\frac{M}{M + 2} Be(a, b) + \frac{1}{M + 2} Be(a + X_1, b + n_1 - X_1) + \frac{1}{M + 2} Be(a + X_2, b + n_2 - X_2)\right]$$

$$+ (1 - p)\left[\frac{M}{M + 2} Be(a, b) + \frac{2}{M + 2} Be(a + X_1 + X_2, b + n_1 + n_2 - X_1 - X_2)\right].$$

The mean of $G|X_1, X_2$ is

$$p\left[\frac{M}{M+2}\left(\frac{a}{a+b}\right)+\frac{1}{M+2}\left(\frac{a+X_1}{a+b+n_1}\right)+\frac{1}{M+2}\left(\frac{a+X_2}{a+b+n_2}\right)\right]$$

$$+(1-p)\left[\frac{M}{M+2}\left(\frac{a}{a+b}\right)+\frac{2}{M+2}\left(\frac{a+X_1+X_2}{a+b+n_1+n_2}\right)\right]$$

$$=\frac{M}{M+2}\left(\frac{a}{a+b}\right)+\frac{p}{M+2}\left[\frac{a+X_1}{a+b+n_1}+\frac{a+X_2}{a+b+n_2}\right]$$

$$+\frac{2(1-p)}{M+2}\frac{a+X_1+X_2}{a+b+n_1+n_2}.$$

The extension for $k = 3$ requires the following weights:

$$P(\theta_1 \neq \theta_2 \neq \theta_3 \neq \theta_1|X_1, X_2, X_3) \equiv p$$

$$P(\theta_1 = \theta_2 \neq \theta_3|X_1, X_2, X_3) \equiv p_{12}$$

$$P(\theta_1 = \theta_3 \neq \theta_2|X_1, X_2, X_3) \equiv p_{13}$$

$$P(\theta_1 \neq \theta_2 = \theta_3|X_1, X_2, X_3) \equiv p_{23}$$

$$P(\theta_1 = \theta_2 = \theta_3|X_1, X_2, X_3) \equiv p_{123}.$$

The computation of these quantities is straightforward but tedious. The desired estimates of θ_3 and G are

$$E(\theta_3|X_1, X_2, X_3) = [p + p_{12}]\left(\frac{a+X_3}{a+b+n_3}\right)+p_{13}\left(\frac{a+X_1+X_3}{a+b+n_1+n_3}\right)$$

$$+p_{23}\left(\frac{a+X_2+X_3}{a+b+n_2+n_3}\right)+p_{123}\left(\frac{a+X_1+X_2+X_3}{a+b+n_1+n_2+n_3}\right)$$

and (where $h \in \{1, 2, 3\} - \{i, j\}$)

$$E(G|X_1, X_2, X_3) = \frac{M}{M+3}Be(a, b)$$

$$+p\left[\frac{1}{M+3}\Sigma_i Be(a+X_i, b+n_i-X_i)\right]$$

$$+\Sigma_{j>i}p_{ij}\left[\frac{2}{M+3}Be(a+X_i+X_j, b+n_i+n_j-X_i-X_j)\right.$$

$$\left.+\frac{1}{M+3}Be(a+X_h, b+n_h-X_h)\right]$$

$$+p_{123}\left[\frac{3}{M+3}Be(a+X_1+X_2+X_3, b+n_1+n_2+n_3-X_1-X_2-X_3)\right].$$

TABLE 1
Empirical Bayes estimates of θ_1 and θ_2

Data (X_1, n_1): Estimator (X_2, n_2):	(4, 5) (9, 10)	(1, 5) (8, 10)	(3, 5) (7, 10)	(1, 5) (9, 10)	(17, 19) (28, 29)	(1, 19) (28, 29)	(10, 19) (21, 29)
Maximum likelihood	.800	.200	.600	.200	.895	.053	.526
	.900	.800	.700	.900	.966	.966	.724
Pooled	.867	.600	.667	.667	.938	.604	.646
Minimax	.707	.293	.569	.293	.821	.136	.521
	.804	.728	.652	.804	.893	.893	.689
Copas' first	.815	.293	.615	.308	.901	.138	.545
	.888	.728	.688	.816	.960	.894	.709
Copas' second	.850	.314	.650	.208	.930	.058	.625
	.850	.730	.650	.855	.930	.962	.643
Griffin-Krutchkoff	.800	.333	.600	.289	.895	.058	.526
	.900	.741	.700	.860	.966	.962	.643
Dirichlet mixtures	.787	.333	.619	.304	.906	.095	.586
with $a = b = M = 1$	.827	.725	.654	.821	.923	.935	.673
$[P(\theta_1 = \theta_2 \mid X_1, X_2)]$	[.663]	[.156]	[.622]	[.064]	[.778]	$[6 \times 10^{-10}]$	[.534]
Dirichlet mixtures	.761	.320	.584	.282	.900	.070	.604
with $a = M = .5, b = 2$	.771	.632	.600	.727	.901	.905	.632
$[P(\theta_1 = \theta_2 \mid X_1, X_2)]$	[.936]	[.349]	[.879]	[.204]	[.987]	$[3 \times 10^{-9}]$	[.852]
Approximation 1	.753	.324	.598	.317	.882	.095	.553
$a = b = M = 1$	.830	.730	.660	.817	.948	.935	.692
Approximation 1	.670	.342	.521	.253	.860	.070	.538
$a = M = .5, b = 2$	.765	.718	.600	.739	.903	.905	.661

Table 1 gives the values of various empirical Bayes estimators of θ_1 and θ_2 for several different combinations of (X_1, n_1), (X_2, n_2); see (Martz and Lian 1974). It also gives the posterior weights for the Dirichlet mixture estimator. The last two rows in Table 1 correspond to an approximation suggested in Section 5. Table 2 gives the values for the same estimators but for θ_1, θ_2, and θ_3 using different combinations of (X_1, n_1), (X_2, n_2), (X_3, n_3). In addition it gives estimates for a second approximation suggested in Section 5.

4. Approximating a mixture of Dirichlet processes. In many applications k is quite large. For example, in clinical trials k may be as large as 1000. It is evident that for k large the bookkeeping difficulties in calculating Dirichlet mixture estimates become enormous. The number of combinations of equality and inequality among $\theta_1, \cdots, \theta_k$ is large indeed, being the number of partitions of k objects, or the Bell exponential number B_k; see (Berge 1971, Section 1.11). The value of B_4 is 15, B_5 is 52, B_6 is 203, and in general $B_k = e^{-1}\Sigma_{j=1}^{\infty} j^k/j!$. The sequence $\{B_k\}$ can be generated by the recursion

$$B_{k+1} = \Sigma_{i=0}^{k}\binom{k}{i}B_i$$

TABLE 2
Empirical Bayes estimates of θ_1, θ_2, and θ_3

Data	(X_1, n_1):	(2, 5)	(1, 5)	(3, 5)	(1, 5)	(9, 10)	(9, 10)	(0, 2)
	(X_2, n_2):	(3, 5)	(3, 10)	(7, 10)	(9, 10)	(10, 12)	(1, 12)	(1, 10)
Estimator	(X_3, n_3):	(2, 5)	(2, 5)	(4, 10)	(8, 10)	(6, 12)	(10, 12)	(20, 20)
		.400	.200	.600	.200	.900	.900	0
Maximum likelihood		.600	.300	.700	.900	.833	.083	.100
		.400	.400	.400	.800	.500	.833	1
Pooled		.467	.300	.560	.720	.735	.588	.656
		.431	.293	.569	.293	.804	.804	.207
Minimax		.569	.348	.652	.804	.759	.177	.196
		.431	.431	.424	.728	.500	.759	.909
		.421	.231	.590	.334	.863	.829	.152
Copas' first		.559	.300	.668	.836	.813	.200	.164
		.421	.369	.440	.760	.555	.782	.884
		.467	.300	.567	.329	.816	.876	.033
Copas' second		.467	.300	.567	.853	.790	.119	.105
		.467	.300	.567	.771	.620	.818	.994
		.400	.200	.600	.355	.810	.876	.054
Griffin-Krutchkoff		.600	.300	.700	.858	.791	.119	.104
		.400	.400	.400	.774	.616	.818	.995
Dirichlet mixtures with		.455	.309	.564	.336	.800	.833	.180
$a = b = M = 1$		.506	.322	.606	.812	.781	.143	.151
		.455	.351	.494	.780	.572	.820	.954
Dirichlet mixtures with		.414	.273	.524	.316	.723	.794	.109
$a = M = .5, b = 2$		.434	.288	.541	.752	.720	.104	.107
		.414	.296	.499	.746	.633	.793	.911
Approximation 1		.440	.299	.569	.346	.815	.829	.232
$a = b = M = 1$		.538	.327	.640	.818	.786	.156	.159
		.440	.386	.453	.762	.552	.801	.950
Approximation 1		.369	.223	.492	.289	.757	.773	.126
$a = M = .5, b = 2$		.392	.283	.579	.762	.741	.133	.112
		.369	.312	.418	.711	.532	.757	.905
Approximation 2		.445	.302	.566	.321	.812	.833	.217
$a = b = M = 1$		.531	.327	.632	.822	.779	.143	.159
		.445	.378	.462	.761	.549	.803	.954
Approximation 2		.387	.251	.510	.301	.731	.777	.120
$a = M = .5, b = 2$		.445	.285	.566	.752	.719	.105	.113
		.387	.306	.447	.703	.583	.758	.910

for $k = 0, 1, \cdots$, where $B_0 = 1$. To avoid extensive calculations and bookkeeping complications it seems desirable to find an approximation of a mixture of Dirichlet processes for which calculations are easy. We propose one form of approximation here, and consider its application to the problems of estimating $G(\theta)$ and, say, θ_k.

The exact distribution of $G|X_1, \cdots, X_k$ is given by (1). Consider the random probability defined by

$$Q(X_1, \cdots, X_k) \sim \mathfrak{D}\left(\alpha + \Sigma_1^k w_i \alpha_{X_i}\right), \tag{5}$$

where α_{X_i} is the posterior probability measure of θ_i given X_i and n_i, assuming the prior is proportional to α; so that

$$d\alpha_{X_i}(\theta_i) = \frac{\theta_i^{X_i}(1-\theta_i)^{n_i - X_i}\, d\alpha(\theta_i)}{\int_{[0,1]} u^{X_i}(1-u)^{n_i - X_i}\, d\alpha(u)}.$$

Assume $\alpha = MBe(a, b)$, then

$$Q(X_1, \cdots, X_k) \sim \mathfrak{D}\big(MBe(a, b) + \Sigma_1^k w_i Be(a + X_i, b + n_i - X_i)\big). \tag{6}$$

Because it has a Dirichlet process distribution, calculations involving $Q(X_1, \cdots, X_k)$ are straightforward. The expected value of Q may be considered an estimate of $G|X_1, \cdots, X_k$:

$$EQ(X_1, \cdots, X_k) = \frac{1}{M + \Sigma w_i}\big[MBe(a, b) + \Sigma_1^k w_i Be(a + X_i, b + n_i - X_i)\big]. \tag{7}$$

For estimating $G(\theta)$ the weight w_i should be a measure of the information in (X_i, n_i). One such measure is the proportion of reduction in variance in observing (X_i, n_i):

$$w_i = w(X_i, n_i) = \frac{\mathrm{Var}_\alpha(\cdot) - \mathrm{Var}_\alpha(\cdot|X_i, n_i)}{\mathrm{Var}_\alpha(\cdot)}.$$

So defined, $0 \leqslant w_i \leqslant 1$, $w_i = 0$ if $n_i = 0$, and $w_i \to 1$ as $n_i \to \infty$. Assuming $\alpha = MBe(a, b)$ these weights are

$$w_i = 1 - \frac{\dfrac{a + X_i}{a + b + n_i}\left[\dfrac{a + X_i + 1}{a + b + n_i + 1} - \dfrac{a + X_i}{a + b + n_i}\right]}{\dfrac{a}{a + b}\left[\dfrac{a + 1}{a + b + 1} - \dfrac{a}{a + b}\right]}. \tag{8}$$

In the same sense that the distribution of G is approximated by Q, a subsequent observation from G is approximated by:

$$E_Q\theta = \frac{1}{M + \Sigma w_i}\left[\frac{Ma}{a + b} + \Sigma_1^k w_i \frac{a + X_i}{a + b + n_i}\right].$$

If, instead of obtaining incomplete information about θ_i from (X_i, n_i), the values of θ_i were themselves observed, then the posterior of G would be a Dirichlet process with parameter $\alpha + \Sigma_1^k \delta(\theta_i)$. The distribution of Q defined in (5) can be viewed as a smoothed version of this distribution. While the distribution of Q cannot be made to converge to that of G as min $n_i \to \infty$ for any definition of w_i (compare, for example, the distribution of the partition $\{[0, \theta_1], (\theta_1, 1]\}$ under Q and under G for n_1 large), the corresponding estimates do converge for w_i defined by (8).

THEOREM 1. *If the parameter α is of the form $MBe(a, b)$ and if $w_i \to 1$ as $n_i \to \infty$ for each i, then EQ given by (7) is a consistent estimator of G in the sense that*

$$\lim_{\min n_i \to \infty} EQ(X_1, \cdots, X_k) = \lim_{\min n_i \to \infty} EG|X_1, \cdots, X_k. \tag{9}$$

PROOF. The right-hand side of (9) is

$$\frac{1}{M+k}\left(\alpha + \Sigma_1^k \delta(\theta_i)\right).$$

Since $w_i \to 1$ it suffices to show that the probability assigned by $Be(a + X_i, b + n_i - X_i)$ to every open interval containing θ_i approaches 1 as $n_i \to \infty$. But the expected value of $Be(a + X_i, b + n_i - X_i)$ is $(a + X_i)/(a + b + n_i)$ which converges to θ_i a.s. as $n_i \to \infty$. □

The weights given in (6) are somewhat arbitrary. Other measures of information that increase to 1 as n_i increases will suffice. If the n_i are moderately large then w_i can be taken to be 1 and the resulting algebra will be simplified.

5. Estimating θ_k. For estimating a particular θ_i, say θ_k, assume G has the same distribution as $Q(X_1, \cdots, X_{k-1})$ defined in (6). Then,

$$\theta_k | X_k \sim \left(M + \Sigma_1^{k-1} w_i\right)^{-1}\left[MBe(a + X_k, b + n_k - X_k) + \Sigma_1^{k-1} w_i Be(a + X_i + X_k, b + n_i - X_i + n_k - X_k)\right],$$

so that

$$(10) \quad \hat{\theta}_k \equiv E_Q \theta_k | X_k = \left(M + \Sigma_1^{k-1} w_i\right)^{-1}\left[M\frac{a + X_k}{a + b + n_k} + \Sigma_1^{k-1} w_i \frac{a + X_i + X_k}{a + b + n_i + n_k}\right].$$

If $n_1, \cdots, n_{k-1}$ are fixed then $\hat{\theta}_k$ is a consistent estimate of θ_k. However, if both n_k and n_i with $i \neq k$ are allowed to increase together, then $\hat{\theta}_k$ is no longer consistent. Using the exact distribution of G given by (1) to estimate θ_k, the contribution of (X_i, n_i) is very small when both n_i and n_k are large, unless X_i/n_i is close to X_k/n_k. This suggests that the weight given to (X_i, n_i) in estimating θ_k should depend on (X_k, n_k) as well as on (X_i, n_i). One way of accomplishing this is to let the weight of α_{X_i} in (5) depend on α_{X_k} as well as on α_{X_i} and α. A definition with some good characteristics that are evident is

$$(11) \quad w_i' = \tfrac{1}{2}\left(w_i y_{ik}^{n_i} + w_k y_{ik}^{n_k}\right),$$

where w_i is defined in (8) and

$$y_{ik} = 1 - \frac{a + X_i}{a + b + n_i} - \frac{a + X_k}{a + b + n_k}.$$

The examples in Tables 1 and 2, for $k = 2$ and 3, respectively, were reconsidered using Equation (10) to estimate θ_k with w_i' defined in (11) in place of w_i. This estimator is called Approximation 1 in those tables. Since w_i' depends on (X_k, n_k) this estimator is no longer the expected value of $\theta_k | X_k$ with respect to Q.

A more appealing approximation for estimating θ_k can be obtained as follows. In each of $k - 1$ separate problems θ_k is estimated from (X_i, n_i) and (X_k, n_k) ignoring the remainder of the data by applying the results of Section 3 with $k = 2$. For each

$i < k$, θ_k is estimated to be $\hat{\theta}_k^i$ using (4) with i and k in place of 1 and 2:

$$\hat{\theta}_k^i = p_i\left(\frac{a + X_k}{a + b + n_k}\right) + (1 - p_i)\left(\frac{a + X_i + X_k}{a + b + n_i + n_k}\right),$$

where

$$\frac{1 - p_i}{p_i} = \frac{a^{(X_1+X_2)}b^{(n_1+n_2-X_1-X_2)}(a + b)^{(n_1)}(a + b)^{(n_2)}}{M(a + b)^{(n_1+n_2)}(a + b)^{(n_1)}(a + b)^{(n_2)}}.$$

The estimate $\hat{\theta}_k^i$ will be close to $(a + X_k)/(a + b + n_k)$, the posterior mean of α_{X_k}, if (X_i, n_i) and (X_k, n_k) suggest that $\theta_i \neq \theta_k$; otherwise, it will be weighted towards $(a + X_i + X_k)/(a + b + n_i + n_k)$. A simple average of these separate estimates then leads to an overall estimate:

$$\theta_k^* = \frac{1}{k - 1}\Sigma_{i=1}^{k-1}\hat{\theta}_k^i. \tag{12}$$

Perhaps a more reasonable definition is the weighted average

$$\Sigma_{i=1}^{k-1}w_i\hat{\theta}_k^i/\Sigma_{i=1}^{k-1}w_i,$$

where w_i is defined in (8). When the n_i are moderately large there is little difference between the two definitions; (12) will be used because it is simpler.

Estimates of θ_1, θ_2, and θ_3 were calculated for the examples in Table 2 using (12), called Approximation 2 in the table. Naturally, Approximation 2 is exact for $k = 2$ so no comparisons are made in Table 1.

An additional example with $k = 6$ is provided by Martz and Lian (1974): "The Portsmouth Naval Shipyard, Portsmouth, N.H., routinely must assess the quality of submitted lots of vendor produced material. The following data consist of the number of defects x_i of a specified type in samples of size $n = 5$ from past lots of welding material. The past data are (0, 1, 0, 0, 5) and in the current, i.e., sixth, lot, $x = 0$."

First consider Approximation 1 for estimating θ_6. Taking $a = b = M = 1$ means $w_1 = w_3 = w_4 = w_5 = w_6 = w_1' = w_3' = w_4' = .8163$, $w_2 = .6939$, $w_2' = .5349$, and $w_5' = .0496$. Evaluating (11) using weights w_i' yields $\hat{\theta}_6 = .1143$. Similarly, $\hat{\theta}_1 = \hat{\theta}_3 = \hat{\theta}_4 = .1143$. For estimating θ_2 the weights become $w_1' = w_3' = w_4' = w_6' = .5349$, $w_5' = .1135$ and $\hat{\theta}_2 = .2178$. For estimating θ_5 the weights become $w_1' = w_3' = w_4' = w_6' = .0496$, $w_2' = .1135$, and $\hat{\theta}_5 = .7795$. These estimates are compared with those of various other estimators in Table 3.

6. Summary. The theory of Dirichlet processes is applied to the empirical Bayes estimation problem in the binomial case. The approach is Bayesian rather than being empirical Bayesian. When the prior is a Dirichlet process the posterior is a mixture of Dirichlet processes. Explicit estimators are given for the case of 2 and 3 parameters and compared with other empirical Bayes estimators by way of examples. Since the number of calculations become enormous when the number of parameters gets larger than 2 or 3 we propose two approximations for estimators of a particular parameter and compare their performance using examples.

TABLE 3
Empirical Bayes estimates for Portsmouth Naval Shipyard data: 0, 1, 0, 0, 5, 0; $n_i = 5$.

Estimator	θ_2	θ_5	θ_6
Maximum likelihood	.200	1	0
Pooled	.200	.200	.200
Minimax	.293	.845	.155
Copas' first	.200	.753	.062
Copas' second	.200	.960	.010
Griffin-Krutchkoff	.200	.960	.010
Approximation 1 $a = b = M = 1$	.218	.780	.114
Approximation 1 $a = b = .5, M = 1$	.185	.817	.072
Approximation 2 $a = b = M = 1$	.231	.850	.119
Approximation 2 $a = b = .5, M = 1$	.189	.906	.073
Approximation 2 $a = M = .5, b = 2$	.162	.717	.064

REFERENCES

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

Berge, C. (1971). *Principles of Combinatorics*. Academic Press, New York and London.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

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

Martz, H. F. and Lian, M. G. (1974). Empirical Bayes estimation of the binomial parameter. *Biometrika* **61** 517–523.

Department of Theoretical Statistics
University of Minnesota
270 Vincent Hall
Minneapolis, Minnesota 55455