Ann. Inst. Statist. Math.
Vol. 41, No. 1, 81-99 (1989)

EMPIRICAL BAYES APPROACH TO MULTIPARAMETER ESTIMATION: WITH SPECIAL REFERENCE TO MULTINOMIAL DISTRIBUTION

T. LWIN[1] AND J. S. MARITZ[2]

[1]*Division of Mathematics and Statistics, CSIRO, Private Bag 10, Clayton, Victoria, 3168, Australia*
[2]*Department of Statistics, LaTrobe University, Bundoora, Victoria, 3083, Australia*

(Received February 10, 1987; revised May 2, 1988)

Abstract. Empirical Bayes approach to estimation of many parameters is considered. Special features of the techniques discussed are: (i) the handling of unequal sample sizes at various stages of an Empirical Bayes sampling scheme and (ii) a general iterative procedure for estimating the parameters of a parametric prior distribution based on the likelihood approach. Linear empirical Bayes estimation is also considered. Application of the general techniques is demonstrated with special reference to a multinomial data distribution.

Key words and phrases: Empirical Bayes estimation, multiparameter, multinomial.

1. Introduction

Let X be a p-vector random variable (r.v.) with a multinomial probability distribution (p.d.) with parameters (m, θ) where θ is a p-vector and m is a scalar parameter taken to be known. We have

$$\Pr(X = x|m,\theta) = p(x|\theta,m) = \left\{ m! \Big/ \prod_{j=1}^{p} x_j! \right\} \theta_1^{x_1},\ldots,\theta_p^{x_p}, \tag{1.1}$$

where, for every i, $0 \le x_i \le m$, $0 < \theta_i < 1$ and $\sum_{i=1}^{p} x_i = m$, $\sum_{i=1}^{p} \theta_i = 1$. The above probability model is fundamental in the analysis of categorical data and contingency table analyses.

In the above the quantity m often represents the sample size of a sample of objects classified into p classes. Suppose that in a current experiment, m observations are made on a p-category Bernoulli r.v. Y with probability θ_i of belonging to category i. Let X_i be the number of Y's in i-th category. Then $X = (X_1,\ldots, X_p)$ has a distribution given by (1.1).

In this paper we are concerned with Empirical Bayes (EB) methods of estimating the unknown parameter vector θ. To adopt the EB approach, a special sampling scheme called an EB scheme is necessary. Let x be the current observation of X corresponding to the parameter θ. Let θ be a realization of an r.v. Θ. At the time when the current observation is made, it is assumed that there are available past observation vectors $x_1,\ldots,x_n$ corresponding to independent realizations θ_i $(i = 1,\ldots,n)$ of Θ. Such EB schemes in multinomial set up are often encountered in practice.

As a practical example, consider the following. Each of a number of subjects is given m questions or propositions to each of which 3 mutually exclusive responses are possible. For example, the responses might be "strong agreement", "strong disagreement" and "neutral". Each subject can be regarded as having probabilities $\theta_1, \theta_2, \theta_3$ of registering a response in the three categories, respectively, and the observed numbers of "strong agreement" etc. responses are multinomial observations. If $\theta = (\theta_1, \theta_2, \theta_3)$ varies randomly from subject to subject, we have, for n subjects a sequence θ_i $(i = 1,\ldots,n)$ of parameters and observations $x_1,\ldots,x_n$ in a typical EB sampling scheme.

As a second example, consider a two-way contingency table with n rows and p-columns. If the row effects can be regarded as corresponding to realizations of a random effect, the results of each row can be taken as a multinomial observation, with randomly varying θ and the problem could well be the estimation of row effects.

2. The problem of empirical Bayes estimation

We consider the squared error loss function given by

$$L(\delta, \theta) = (\delta - \theta)^T C(\delta - \theta)\,, \tag{2.1}$$

where C is a known positive definite matrix and δ is an estimator for θ and is a function of X. The average risk function of δ is defined as

$$\rho(\delta) = E_G E_p L(\delta(X), \Theta)\,, \tag{2.2}$$

where $E_G(\cdot)$ is the expectation with respect to (w.r.t.) the prior distribution function (d.f.) G of Θ and E_p is the expectation w.r.t. the p.d. $p(x|\theta, m)$ of (1.1).

Let $\hat{\theta}^*$ be the mean of the posterior d.f. of θ, i.e.,

$$\hat{\theta}^* = E(\Theta | X = x)\,,$$

where the expectation is w.r.t. the posterior d.f.

$$dB(\theta|x,m) = \{P_{G,m}(x)\}^{-1} p(x|\theta,m)dG(\theta)\,, \tag{2.3}$$

where

$$P_{G,m}(x) = \int p(x|\theta,m)dG(\theta)\,. \tag{2.4}$$

Now letting E stand for the repeated operation $E_G E_p$, we can rewrite (2.2) as

$$\rho(\delta) = E[\{(\delta - \hat{\theta}^*)^T C(\delta - \hat{\theta}^*)\} + \{(\Theta - \hat{\theta}^*)^T C(\Theta - \hat{\theta}^*)\}]\,,$$

which is minimized when δ is chosen to be $\hat{\theta}^*$. Thus the Bayes estimator of θ is the posterior mean of Θ. Further when $\delta = \hat{\theta}^*$, the above expression for $\rho(\delta)$ can be reduced to

$$\rho(\hat{\theta}^*) = E_{p_{G,m}} \operatorname{tr}\{C \operatorname{Cov}(\Theta|X)\}\,, \tag{2.5}$$

or equivalently to

$$\rho(\hat{\theta}^*) = E_G(\Theta^T C\Theta) - E_{p_{G,m}}\{\hat{\theta}^* C\hat{\theta}^*\}\,, \tag{2.6}$$

where $\operatorname{Cov}(\Theta|X)$ is the covariance of Θ w.r.t. the posterior d.f. of θ, $E_{p_{G,m}}(\cdot)$ is the expectation w.r.t. the marginal p.d. of (2.4).

The expression (2.5) is already given by DeGroot (1970), that of (2.6) generalizes a result of Johns (1957) to multi-dimensional cases.

Suppose now that the d.f. $G(\theta)$ is not known. Suppose also that an EB sampling scheme is available such that at the i-th stage a vector x_i is observed corresponding to a realization θ_i of Θ, where

$$x_i = (x_{1i},\ldots,x_{pi})^T\,,$$

and

$$\sum_{j=1}^{p} x_{ji} = m_i\,.$$

Our problem is to obtain EB estimators of θ based on the data $\{x, x_1,\ldots,x_n\}$ under various assumptions on G.

We consider four subcases depending on the nature of specifications on G. The first case deals with the situation where the parametric form of G is known and only a set of parameters α are to be determined to specify G; here we consider (a) the Dirichlet prior distribution and (b) the logistic normal prior distribution. The second case deals with the situation where G is specified only up to second order moments; an investigation into this

case was made possible by restricting the class of estimators of θ to the one linear in the current data x. The third case deals with an approximation of G by a finite step function and thus reduces the problem of estimating a continuous G to that of estimating a finite set of parameters. Finally, in the fourth case, we deal with the situation where no specification of G was made except the fact that it was assumed to belong to a family of continuous distributions whose second order moments exist.

3. Estimation with parametric priors

3.1 *A general procedure*

Consider the case when $G(\theta)$ is specified up to a set of unknown parameters $\alpha = (\alpha_1, \ldots, \alpha_q)$. Let $g(\theta|\alpha)$ be the probability density function (p.d.f.) of θ where the form of $g(\cdot|\alpha)$ is completely known. The Bayes estimator is

$$\hat{\theta}^* = \{P_{G,m}(x)\}^{-1} \int \theta p(x|\theta, m) g(\theta|\alpha) d\theta .$$

The unknown element α in the above expression can be estimated from the previous data, $x_1, \ldots, x_n$.

The likelihood function of these observations can be written as

$$L(\alpha) = \prod_{i=1}^{n} \int p(x_i|\theta, m_i) g(\theta|\alpha) d\theta \tag{3.1}$$

$$= \prod_{i=1}^{n} h(x_i|\alpha, m_i) , \tag{3.2}$$

where

$$h(x|\alpha, m) = \int p(x|\theta, m) g(\theta|\alpha) d\theta .$$

Now the likelihood equations for α_j can be written as

$$0 = \sum_{i=1}^{n} E\left(\frac{\partial \ln g(\Theta|\alpha)}{\partial \alpha_j} \,\middle|\, X = x_i \right) \quad j = 1, \ldots, q .$$

One can expand the function

$$E\left(\frac{\partial \ln g(\Theta|\alpha)}{\partial \alpha_j} \,\middle|\, X = x_i \right)$$

in Taylor series and apply the Newton-Rapson technique to the above set of q equations. This provides us an iterative solution of the likelihood equations as

$$\hat{\alpha}^{(i+1)} = \hat{\alpha}^{(i)} + J^{-1}(\hat{\alpha}^{(i)})U(\hat{\alpha}^{(i)}), \tag{3.3}$$

where $J(\alpha)$ is a $q \times q$ matrix whose (i, j)-th element is

$$J_{ij}(\alpha) = n^{-1} \sum_{t=1}^{n} E\left(\frac{\partial^2 \ln g(\Theta|\alpha)}{\partial\alpha_i \partial\alpha_j} \,\middle|\, X = x_t, m_t \right), \tag{3.4}$$

$U(\alpha)$ is a $q \times 1$ vector whose i-th element is

$$U_i(\alpha) = n^{-1} \sum_{t=1}^{n} E\left(\frac{\partial \ln g(\Theta|\alpha)}{\partial\alpha_i} \,\middle|\, X = x_t, m_t \right), \tag{3.5}$$

and $\hat{\alpha}^{(i)}$ is the i-th step estimate of α. Often in special cases moment estimators of α are readily available, so that these can be used as the initial estimate $\hat{\alpha}^{(0)}$ to start the iterative process. We note here that the iterative sequence will need $J(\alpha)$ to be positive definite at every step. This will be the case for a given g whose information matrix is positive definite; conditions for this are same as classical regularity conditions.

The above general procedure will be illustrated for the multinomial distribution using (a) the Dirichlet prior distribution and (b) the logistic normal prior distribution.

For the special case of the binomial data distribution the Dirichlet prior d.f. becomes a type I beta distribution with parameters (α_1, α_2) and the estimation of α reduces to the well-known problem of estimating the parameters of a negative hypergeometric distribution. Estimation of the parameters by the method of moments is straightforward (see e.g., Maritz (1970), p. 53). ML estimation of (α_1, α_2) leads to the same iterative equations as given above with $p = 2$. The Dirichlet prior has also been applied in an EB approach using parametric prior processes for the unknown d.f. G. Essentially this latter approach is based on assigning a two stage prior d.f. on the usual EB scheme and had been applied to the binomial case (Berry and Christensen (1979)). However, the latter approach did not discuss the estimation of the parameters of the Dirichlet process assumed and hence is not strictly an EB approach in the sense of the present paper.

3.2 *Dirichlet prior and the multinomial distribution*

A Dirichlet prior density for $\theta_1, \ldots, \theta_p$ is given by

$$g(\theta|\alpha) = [\Gamma(\alpha_0)/\{\Gamma(\alpha_1)\cdots\Gamma(\alpha_p)\}]\theta_1^{\alpha_1 - 1}\cdots\theta_p^{\alpha_p - 1},$$

where $0 < \theta_i < 1$; $\Sigma\theta_i = 1$ and $0 < \alpha_i$; $\alpha_0 = \alpha_1 + \cdots + \alpha_p$. This is a conjugate prior p.d.f. for the multinomial distribution. The parameter vector α is unknown. Straightforward calculation gives the posterior distribution of θ

given x as a Dirichlet distribution with parameters $\alpha^* = (\alpha_1^*, \ldots, \alpha_p^*)$ where $\alpha_i^* = x_i + \alpha_i$. The Bayes estimator of Θ is the posterior mean

$$\hat{\theta}^*(x, \alpha) = E(\Theta | x) = (x + \alpha)/(\alpha_0 + m) .$$

The Bayes risk of $\hat{\theta}_r^*(x, \alpha)$ is $\alpha_r(\alpha_0 - \alpha_r)/\{\alpha_0(\alpha_0 + m)^2\}$. If $C = I$ in (2.1), the overall average risk of $\hat{\theta}^*$ is

$$E(\hat{\theta}^* - \Theta)^T(\hat{\theta}^* - \Theta) = \left(\alpha_0^2 - \sum_{i=1}^{p} \alpha_i^2 \right) \Big/ \{\alpha_0(\alpha_0 + m)^2\} .$$

Now consider the case when α is unknown and an EB scheme is available. The method of moments estimators $\check{\alpha}$, for α can easily be obtained by using the marginal moments of $X_i = (X_{1i}, \ldots, X_{pi})^T$ as given by

$$E(X_{ji}) = m_i \alpha_j / \alpha_0 ,$$

$$E\{X_{ji}(X_{ji} - 1)\} = m_i(m_i - 1)\alpha_j(\alpha_j + 1)/\alpha_0(\alpha_0 + 1) .$$

An estimate of α_j is then obtained as

$$\check{\alpha}_j = (\check{\alpha}_0 / n) \left(\sum_{i=1}^{n} X_{ji}/m_i \right) ,$$

where $\check{\alpha}_0$, a moment estimator of α_0, is given by the equation

$$\check{\alpha}_0 = \frac{\Sigma m_i^2 - \Sigma\Sigma X_{ji}^2}{\Sigma\Sigma X_{ji}^2 - \Sigma m_i - \Sigma m_i^2 \, \Sigma \bar{X}_j^2 + \Sigma m_i \, \Sigma \bar{X}_j^2} .$$

An alternative method for estimating α_0 is to equate the sum of the determinants of the sample covariance matrix to the theoretical value (Mosimann (1962)):

$$\sum_{i=1}^{n} \{(m_i + \alpha_0)/(1 + \alpha_0)\}^{p-1} D_i ,$$

where D_i is a determinant of the form

$$D_i = \| d_{jk}^{(i)} \|$$

with

$$d_{jk}^{(i)} = \begin{cases} m_i(\alpha_j/\alpha_0)(1 - \alpha_j/\alpha_0) & j = k , \\ -m_i(\alpha_j/\alpha_0)(\alpha_k/\alpha_0) & j \neq k . \end{cases}$$

The quantity α_j/α_0 is estimated by $\bar{X}_j/m_i$ where

$$\bar{X}_j = \sum_{i=1}^{n} X_{ji}/n .$$

A more efficient way of estimating α_j's and α_0 is to use the maximum likelihood technique of Subsection 3.1. We have

$$\frac{\partial \ln g(\theta|\alpha)}{\partial \alpha_i} = [\psi(\alpha_0) - \psi(\alpha_i)] - \ln \theta_i \quad (i = 1,\ldots,p) ,$$

$$\frac{\partial^2 \ln g(\theta|\alpha)}{\partial \alpha_i \partial \alpha_j} = \begin{cases} \psi'(\alpha_0) - \psi'(\alpha_i) & \text{if} \quad i = j , \\ \psi'(\alpha_0) & \text{if} \quad i \neq j , \end{cases}$$

where $\psi(x) = d \ln \Gamma(x)/dx$ is the digamma function. Thus

$$U_i(\alpha) = \psi(\alpha_0) - \psi(\alpha_i) - n^{-1} \sum_{t=1}^{n} E(\ln \Theta_i | x_t, m_t)$$

and

$$J_{ij}(\alpha) = \begin{cases} \psi'(\alpha_0) - \psi'(\alpha_i) & \text{if} \quad i = j , \\ \psi'(\alpha_0) & \text{otherwise} . \end{cases}$$

From the Dirichlet property of the posterior distribution of Θ given (x, m), we have

$$E(\ln \Theta_i | x, m) = \frac{\alpha_i + x_i}{\alpha_0 + m} [\psi(\alpha_i + 1 + x_i) - \psi(\alpha_0 + m + 1)] .$$

Hence the (i, j)-th element of $J^{-1}(\alpha)$ is given by

$$J^{(i,j)}(\alpha) = \begin{cases} -\dfrac{1}{\psi'(\alpha_i)} - k_0 \left\{ \dfrac{1}{\psi'(\alpha_i)} \right\}^2 & i = j , \\ -\dfrac{k_0}{\psi'(\alpha_i)\psi'(\alpha_j)} & i \neq j , \end{cases}$$

where

$$k_0 = \frac{\psi'(\alpha_0)}{1 + \psi'(\alpha_0) \sum_{i=1}^{p} \left\{ \dfrac{1}{\psi'(\alpha_i)} \right\}} ,$$

and the i-th element of $U(\alpha)$ is given by

$$U_i(\alpha) = n^{-1} \sum_{t=1}^{n} \frac{\alpha_i + x_{it}}{\alpha_0 + m_t} [\psi(\alpha_i + 1 + x_{it}) - \psi(\alpha_0 + m_t + 1)] .$$

Hence the iterative procedure of equation (3.3) can be readily carried out.

3.3 *The logistic normal prior and the multinomial distribution*

A prior distribution that has been widely used in Bayesian analysis of categorical data is the logistic normal prior distribution (see Leonard (1972), Aitchison and Shen (1980)). We now assume a logistic normal prior d.f. for the random parameter Θ. Let

$$\Lambda_i = \ln (\Theta_i / \Theta_p) \qquad i = 1,\ldots,p-1 . \tag{3.6}$$

The joint d.f. of $\Lambda = (\Lambda_1,\ldots,\Lambda_{p-1})^T$ is a multivariate normal distribution with mean vector $\xi = (\xi_1,\ldots,\xi_{p-1})^T$ and a $(p-1)\times(p-1)$ covariance matrix Γ. The prior d.f. of Θ can then be expressed in terms of parameters ξ and Γ. Thus, if $G(\theta|\alpha)$ is the prior d.f. of Θ, then α is composed of elements of ξ and Γ and $g(\theta|\alpha)$, the p.d.f. of Θ can be written as in Aitchison and Shen (1980).

Consider now the Bayes estimator of θ_i. For any d.f. $G(\theta|\alpha)$, the Bayes estimator of θ_i is given by

$$\hat{\theta}_i^*(x,\alpha) = \frac{(x_i+1)}{(m+1)} \frac{h(x_1,\ldots,x_i+1,\ldots,x_p|\alpha,m+1)}{h(x|\alpha,m)} , \tag{3.7}$$

where

$$h(x|\alpha,m) = P_{G,m}(x) = \int p(x|\theta,m)\,dG(\theta|\alpha) ,$$

and $h(x_1,\ldots,x_i+1,\ldots,x_p|\alpha,m+1)$ is also defined as $h(x|\alpha,m)$ with x_i+1 and $m+1$, respectively, in place of x_i and m.

For the logistic normal d.f. $G(\theta|\alpha)$, $h(x|\alpha,m)$ can be obtained in principle as an integral depending on ξ and Γ.

Suppose we are interested in the Bayes estimator of the quantities in (3.6). The posterior distribution of Λ can be written as

$$dB(\lambda|y,\xi,\Gamma) = \{C(y,\xi,\Gamma)\}^{-1} \cdot \exp\left\{ y^T\lambda - mD(\lambda) - \frac{1}{2}\ln|\Gamma| - \frac{1}{2}(\lambda-\xi)^T\Gamma^{-1}(\lambda-\xi) \right\}, \tag{3.8}$$

where $y = (x_1,\ldots, x_{p-1})^T$, $D(\lambda) = \ln\left\{1 + \sum_{i=1}^{p-1} \exp(\lambda_i)\right\}$ and

$$C(y,\xi,\Gamma) = \int \exp\left\{ y^T\lambda - mD(\lambda) - \frac{1}{2}\ln|\Gamma| - \frac{1}{2}(\lambda-\xi)^T\Gamma^{-1}(\lambda-\xi)\right\} d\lambda . \tag{3.9}$$

The moment generating function of the p.d.f. (3.8) is

$$E(e^{t^T\Lambda}|y,\xi,\Gamma) = \chi(t,y,\xi,\Gamma) ,$$

where $t = (t_1,\ldots,t_{p-1})^T$ is a dummy vector. We have

$$\chi(t,y,\xi,\Gamma) = C(y+t,\xi,\Gamma)/C(y,\xi,\Gamma) . \tag{3.10}$$

Hence the posterior expectation of λ_i is

$$\hat{\lambda}_i^*(y,\xi,\Gamma) = \frac{\partial\chi(t,y,\xi,\Gamma)}{\partial t_i} k \Big|_{t=0} = \frac{\partial}{\partial t_i} C(y+t,\xi,\Gamma)\Big|_{t=0} \Big/ C(y,\xi,\Gamma) . \tag{3.11}$$

Hence for given ξ, Γ, y, the problem of obtaining the Bayes estimator of θ_i is to compute integrals of the form $h(x|\alpha,m)$ and that of obtaining the Bayes estimator of λ_i is to evaluate the integral of the form $C(y,\xi,\Gamma)$ and its derivatives.

When ξ and Γ are unknown, it is required to estimate them using an EB scheme. For this purpose we work with the reparametrization (3.6). Using (3.9), the likelihood function of (ξ,Γ) can be expressed as

$$L^*(\xi,\Gamma) \propto \prod_{i=1}^{n} C(y_i, m_i, \xi, \Gamma) , \tag{3.12}$$

where the constant of proportionality is independent of ξ and Γ. Now the likelihood equations for ξ and Γ are obtained as follows.

Let $\partial \ln L^*/\partial\xi$ be a $(p-1)\times 1$ vector with i-th element

$$\partial \ln L^*/\partial\xi_i \qquad i = 1,\ldots,p-1 .$$

Let γ_{ij} be the (i,j)-th element of Γ. Let $\partial \ln L^*/\partial\Gamma$ be a $(p-1)\times(p-1)$ matrix whose (i,j)-th element is

$$\partial \ln L^*/\partial \gamma_{ij} .$$

Then we have, using the matrix differentiation results (Graybill (1969), pp. 262–267),

$$\partial \ln L^*/\partial \xi = -\sum_{i=1}^{n} \Gamma^{-1}\{E(\Lambda \mid y_i, m_i, \xi, \Gamma) - \xi\} ,$$

$$\partial \ln L^*/\partial \Gamma = -\frac{1}{2}\sum_{i=1}^{n} E[\{\Gamma^{-1}(\Lambda - \xi)(\Lambda - \xi)^T\Gamma^{-1} - \Gamma^{-1}\} \mid y_i, m_i, \xi, \Gamma] ,$$

where the expectations in the above expressions are with respect to the posterior p.d.f. (3.8).

The likelihood equations are

$$\hat{\xi} = n^{-1}\sum_{i=1}^{n} E(\Lambda \mid y_i, m_i, \hat{\xi}, \hat{\Gamma}) \tag{3.13}$$

and

$$\hat{\Gamma} = n^{-1}\sum_{i=1}^{n} E[(\Lambda - \xi)(\Lambda - \xi)^T \mid y_i, m_i, \hat{\xi}, \hat{\Gamma}] . \tag{3.14}$$

The likelihood equations themselves present an iterative procedure for determining $\hat{\xi}$ and $\hat{\Gamma}$ if we use i-th step estimates $\hat{\xi}^{(i)}$ and $\hat{\Gamma}^{(i)}$ on the right hand sides of (3.13) and (3.14) to produce $(i+1)$-th step estimates. There remains the technical problem of evaluating expressions for the posterior expectations involved.

In (3.11) it was shown that the expectations of Λ can be obtained in terms of $C(y, \xi, \Gamma)$ and its derivatives. Following similar lines of (3.11), it can be shown that the second order posterior moments of Λ can be obtained by using the second order derivatives of $\chi(t, y, \xi, \Gamma)$ evaluated at $t = 0$. Thus the problem now reduces to that of evaluating the integral (3.9) and its first order derivatives. These quantities can be evaluated by numerical integration using Gaussian quadrature.

4. Linear EB estimation

Consider the class of estimators linear in the current observation x, i.e.,

$$\hat{\theta} = AX + b , \tag{4.1}$$

where X is the r.v. whose realization is x, b is a $p \times 1$ vector and A is a $p \times p$ matrix of quantities not depending on X.

First we want to construct a linear Bayes estimator of θ in the above class. We assume that the prior d.f. of G belongs to the class

$$\mathscr{G}: \{G: E_G(\Theta) = \xi, \operatorname{Cov}_G(\Theta) = \Gamma\} .$$

No specific parametric form of G is assumed. Since the Bayes estimator is sought in the class (4.1), this means that the quantities A and b must be determined such that

$$E(\Theta | X = x) = Ax + b . \tag{4.2}$$

We now consider the data distribution of X to be such that

$$E_p(X) = \theta, \quad \operatorname{Cov}_p(X) = \Sigma(\theta) .$$

This covers the multinomial p.d. (1.1). Using (4.2)

$$\begin{aligned} &E_G(\Theta) = E_H\{E(\Theta|X)\} = E_H(AX + b) = AE_G(\Theta) + b = A\xi + b \quad \text{i.e.,} \\ &b = (I - A)\xi , \end{aligned} \tag{4.3}$$

where I is a $p \times p$ identity matrix.

Next we consider two different evaluations of $E(X\Theta^T)$. First,

$$\begin{aligned} E(X\Theta) &= E_G\{E_p(X)\Theta^T\} \\ &= E_G(\Theta\Theta^T) = \Gamma + \xi\xi^T . \end{aligned} \tag{4.4}$$

On the other hand using (4.2),

$$\begin{aligned} E(X\Theta^T) &= E_H\{XE(\Theta|X)\} \\ &= E_H\{X(X^TA^T + b^T)\} \\ &= (E_HXX^T)A^T + \xi b^T \\ &= \{\operatorname{Cov}_H(X) + \xi\xi^T\}A^T + \xi b^T . \end{aligned} \tag{4.5}$$

We can use the standard result:

$$\operatorname{Cov}_H(X) = E_G \operatorname{Cov}(X|\Theta) + \operatorname{Cov}_G\{E(X|\Theta)\} .$$

This gives

$$E(X\Theta^T) = [E_G\{\Sigma(\Theta)\} + \Gamma + \xi\xi^T]A^T + \xi b^T . \tag{4.6}$$

Equating the two expressions (4.4) and (4.5), we get

$$A = \{E_G\Sigma(\Theta) + \Gamma\}^{-1}\Gamma ,$$

and hence from (4.3), we have

$$b = [E_G\{\Sigma(\Theta)\} + \Gamma]^{-1}E_G\{\Sigma(\Theta)\}\xi .$$

The above general result is readily specialized to the multinomial case by noting that $\Sigma(\Theta)$ has an (i, j)-th element

$$\sigma_{ij} = \begin{cases} m\theta_i(1 - \theta_i) & i = j , \\ -m\theta_i\theta_j & i \neq j . \end{cases}$$

We then have

$$E_G\Sigma(\Theta) = m(D - \xi\xi^T - \Gamma) = mB(\xi, \Gamma) ,$$

where $D(\xi)$ is a $p \times p$ diagonal matrix whose i-th diagonal element is ξ_i. This leads to the linear Bayes estimator for the multinomial parameter θ as

$$\hat{\theta}^* = \{B(\xi, \Gamma) + \Gamma/m\}^{-1}\{\Gamma X/m + B(\xi, \Gamma)\xi\} . \tag{4.7}$$

We now have the question of obtaining estimates of ξ and Γ when an EB scheme is available. This gives us an EB analogue of the linear Bayes estimator $\hat{\theta}^*$. Consider the quantities

$$\bar{Z} = n^{-1}\sum_{i=1}^{n}(X_i/m_i) = n^{-1}\sum_{i=1}^{n}Z_i$$

and

$$S = \sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^T .$$

It can be readily established that

$$E\bar{Z} = \xi$$

and

$$E(S) = (n - 1)\left[\Gamma + n^{-1}\sum_{i=1}^{n}B(\xi, \Gamma)/m_i\right] .$$

Thus simple moment estimators of ξ and Γ are given by:

$$\hat{\xi} = \bar{Z}, \tag{4.8}$$

$$\hat{\Gamma} = \left[\frac{S}{n-1} - \frac{D(\hat{\xi}) - \hat{\xi}\hat{\xi}^T}{n} \sum_{i=1}^{n} m_i^{-1} \right] \Big/ \left(1 - \frac{1}{n} \sum_{i=1}^{n} m_i^{-1} \right). \tag{4.9}$$

Linear EB estimation was introduced in the context of the exponential family by Jackson *et al.* (1970) and more generally by Griffin and Krutchkoff (1971). Both these developments concentrate on a scalar parameter θ and the latter gave an application of linear EB estimator of θ defined by (4.7), (4.8) and (4.9) which then reduced to the Griffin-Krutchkoff estimator. The success of the linear EB technique depends on the extent of the linear approximation to the exact Bayes estimator corresponding to the true prior d.f. G. The Bayes estimator is linear for the conjugate prior d.f. Thus for cases where the true prior d.f. is close to the conjugate prior, the linear EB techniques can be effective. In general, the Bayes estimator may be a nonlinear function of the current observation x and the nature of the prior d.f. is unknown. To overcome this problem, a completely nonparametric family of d.f. G have been considered as is shown in Section 6. By a completely nonparametric d.f. G we mean a continuous d.f. G, whose parametric form is assumed to be unknown. An approximate nonparametric approach that had been found to perform well in the scalar parameter case is to use a finite step function approximation to the unknown G (see Maritz (1970)). This approach has been extended to a vector parameter case in Section 5 below.

5. Finite approximation to G

An intermediate step between adopting a fully parametric prior d.f. $G(\theta|\alpha)$ and adopting a completely nonparametric family of d.f.'s G is to approximate the specified G by a finite step function of the form

$$G_k(\theta) = \sum_{j=1}^{i} \alpha_j \quad \text{at} \quad \theta = \lambda_i \quad (i = 1,\ldots,k), \tag{5.1}$$

where G_k has jumps of size $\alpha_1,\ldots,\alpha_k$ at the points $\lambda_1,\ldots,\lambda_k$ of the parameter space of θ. Here the jump sizes satisfy the constraints ($\alpha_1 + \cdots + \alpha_k = 1$, $0 < \alpha_i < 1$). A likelihood approach can still be developed to estimate $\alpha = (\alpha_1,\ldots,\alpha_k)$ and/or $\lambda = (\lambda_1,\ldots,\lambda_k)$. We can approximate the likelihood function (3.1) by

$$L_k(\alpha, \lambda) = \prod_{i=1}^{n} h_k(x_i|\alpha, \lambda, m_i), \tag{5.2}$$

where

$$h_k(x_i|\alpha,\lambda,m_i)=\sum_{j=1}^{k}p(x_i|\lambda_j,m_i)\alpha_j\,. \tag{5.3}$$

We may treat both α and λ as unknown parameters. But this could lead to over-parametrization. We consider only the two alternative cases:

1. *α is assumed to be known; λ unknown*

We may assume that $\alpha_i=1/k$ and consider estimating $\lambda_1,\ldots,\lambda_k$ by maximum likelihood. For identifiability one may need to impose some restrictions on λ_i's such as an ordering sequence. The estimating sequence would be as in Section 3 with $L_k(\alpha,\lambda)$ in place of $L(\alpha,\lambda)$.

2. *λ is assumed to be known; α unknown*

Assume that $\lambda_1,\ldots,\lambda_k$ are known selected values and $\alpha_1,\ldots,\alpha_k$ are to be estimated. In this case, too, a likelihood procedure based on $L_k(\alpha,\lambda)$ can be developed and an iterative sequence analogous to that of Section 3 can be obtained. Indeed this is a typical problem of estimating mixing proportions of a finite mixture of known components. Thus, it can again be treated by EM algorithm approach. In this case a set of iterative equations for α_j's will result as follows:

$$\hat{\alpha}_j^{(i+1)}=\sum_{u=1}^{n}\hat{\alpha}_j^{(i)}(x_u)/n\,, \tag{5.4}$$

where

$$\hat{\alpha}_j^{(i)}(x_u)=\hat{\alpha}_j^{(i)}p(x_u|\lambda_j,m_u)\Big/\sum_{j=1}^{k}\hat{\alpha}_j^{(i)}p(x_u|\lambda_j,m_u)\,.$$

In (5.4) $\hat{\alpha}_j^{(i)}$ is an estimate of α_j at the i-th stage of the iteration. At the initial stage one can take $\hat{\alpha}_j^{(0)}=1/k$.

6. Simple empirical Bayes estimation

6.1 *A general method*

This section considers the multinomial extension of simple empirical Bayes estimation techniques which do not require an explicit estimation of G. A Bayes identity useful for this purpose in the case of a uniparameter distribution was given in general terms by Maritz and Lwin (1975). Analogue of this identity for multiparameter data distributions can be made along similar lines; the following lemma provides this.

LEMMA 6.1. *Let $X=(X_1,\ldots,X_p)^T$ be a vector random variable with p.d. (p.d.f.) $p(x|\theta)$. Let θ be a realization of a vector r.v. $\Theta=(\Theta_1,\ldots,\Theta_p)^p$ with d.f. $G(\theta)$. Let $M_u=(M_{u(1)},\ldots,M_{u(p)})^T$ be a vector of linear operators*

such that for any real-valued function $\alpha(u)$ *of* $u = (u_1,\ldots,u_p)^T$, *the vector*

$$M_u\alpha(u) = [M_{u(1)}\alpha(u),\ldots, M_{u(p)}\alpha(u)]^T$$

is well defined. Further, for any two operators, M_{u_i} *and* M_{u_j}, *the repeated operator*

$$M_{u_i}M_{u_j}\alpha(u)$$

is well defined. The marginal p.d. (or p.d.f.) of X *is*

$$P(x; G) = \int p(x|\theta)dG(\theta) .$$

Then the following relationships hold:

$$\frac{M_{x_i}P(x; G)}{P(x; G)} = E\left[\frac{M_{x_i}p(x|\Theta)}{p(x|\Theta)} \,\middle|\, x\right],$$

$$\frac{M_{x_i}M_{x_j}P(x; G)}{P(x; G)} = E\left[\frac{M_{x_i}M_{x_j}p(x|\Theta)}{p(x|\Theta)} \,\middle|\, x\right].$$

PROOF. Along the same lines as Lemma 3.1 of Maritz and Lwin (1975).

6.2 *A simple EB method for the multinomial distribution*

We now apply the general relationships of Subsection 6.1 to the multinomial p.d. of (1.1). Let the operator $M_{x_i} = M_i$ be defined as

$$M_ip(x|\theta, m) = \Pr\{X_1 = x_1,\ldots, X_i = x_i + 1,\ldots, X_p = x_p - 1|\theta, m\} \qquad (i \neq p) .$$

Also let the repeated operator $M_{x_i}M_{x_j} = M_iM_j$ be defined as

$$M_iM_jp(x|\theta, m) = \Pr\{X_1 = x_1,\ldots, X_i = x_i + 1,\ldots, X_j = x_j + 1,\ldots, X_p = x_p - 2|\theta, m\} .$$

Applying the above lemma and using $P_{G,m}(x)$ of (2.4) for $P(x; G)$, we have

$$(6.1) \qquad a_i = E\{\Omega_i|x, m\} = \frac{x_i + 1}{x_p}\{M_iP_{G,m}(x)/P_{G,m}(x)\} ,$$

$$(6.2) \qquad a_{ij} = E(\Omega_i\Omega_j|x, m) = \frac{(x_i + 1)(x_j + 1)}{x_p(x_p - 1)}\{M_iM_jP_{G,m}(x)/P_{G,m}(x)\} ,$$

where

$$\Omega_i = \Theta_i / \Theta_p \qquad i = 1,\ldots,p-1\,,$$
$$\Omega_p = 1\,. \tag{6.3}$$

We can express Θ_i's in terms of Ω_i's as

$$\Theta_p = 1 \Big/ \sum_{i=1}^{p} \Omega_i = 1/\Omega_T\,,$$
$$\Theta_i = \Omega_i / \Omega_T \qquad i = 1,\ldots,p-1\,. \tag{6.4}$$

The Bayes estimator of Θ_i can be expressed approximately in terms of a_i and a_{ij}. Consider the Taylor expansion of (6.4) to the second order terms:

$$\begin{aligned}\Theta_i = a_i/a_T &+ (\Omega_i - a_i)\left\{\frac{\Omega_T - \Omega_i}{\Omega_T^3}\right\}_{\Omega=a} \\ &+ \sum_{j\neq i}(\Omega_j - a_j)\left\{\frac{-\Omega_i}{\Omega_T^2}\right\}_{\Omega=a} \\ &+ \frac{1}{2}(\Omega_i - a_i)^2\left\{-2\left(\frac{\Omega_T - \Omega_i}{\Omega_T^3}\right)\right\}_{\Omega=a} \\ &+ 2\sum_{j\neq i}(\Omega_i - a_i)(\Omega_j - a_j)\left\{\frac{2\Omega_i - \Omega_T}{\Omega_T^3}\right\}_{\Omega=a} \\ &+ \sum_{j\neq i}\sum_{k\neq i}(\Omega_j - a_j)(\Omega_k - a_k)\left\{\frac{2\Omega_i}{\Omega_T^3}\right\}_{\Omega=a}\,,\end{aligned}$$

where $\Omega = (\Omega_1,\ldots,\Omega_p)^T$ and $a = (a_1,\ldots,a_p)^T$. Then for $x_p \neq 0, 1$, we have

$$\begin{aligned}E(\Theta_i|x,m) = a_i/a_T &- (a_{ii} - a_i^2)\left(\frac{a_T - a_i}{a_T^3}\right) \\ &+ \sum_{j\neq i}(a_{ij} - a_i a_j)\frac{2a_i - a_T}{a_T^3} + \sum_{j\neq i}\sum_{k\neq i}(a_{jk} - a_j a_k)\frac{a_i}{a_T^3} \\ &\qquad (i = 1,\ldots,p-1)\,,\end{aligned} \tag{6.5}$$

where $a_T = \sum_{j=1}^{p} a_j$.

Suppose now that an EB scheme is available. Also assume that the sample sizes at different stages of the EB scheme are all identical and equal to m. Then the functionals of G in the expressions for a_i and a_{ij} can be estimated from previous data. Let $N_{n,m}(x)$ be the number of previous

observations $x_1,\ldots,x_n$ that are equal to x. Then we have an estimate of $P_{G,m}(x)$ as

$$\hat{P}_{G,m}(x) = N_{n,m}(x)/n\,. \tag{6.6}$$

Similarly, $M_iP_{G,m}(x)$ and $M_iM_jP_{G,m}(x)$ can be estimated by $M_iN_{n,m}(x)/n$ and $M_iM_jN_{n,m}(x)/n$. Thus EB estimators of a_i and a_{ij} can be constructed as

$$\hat{a}_{i,n} = \frac{x_i+1}{x_p}\{M_iN_{n,m}(x)/N_{n,m}(x)\}\,, \tag{6.7}$$

$$\hat{a}_{ij,n} = \frac{(x_i+1)(x_j+1)}{x_p(x_p-1)}\{M_iM_jN_{n,m}(x)/N_{n,m}(x)\}\,. \tag{6.8}$$

An approximate EB estimator can now be constructed for θ_i by substituting EB estimators of a_i and a_{ij} in (6.5).

6.3 *An alternative simple EB estimation for the multinomial distribution*

An alternative simple EB estimator θ can be constructed by considering an extension of an EBE for a binomial parameter originally proposed by Robbins (1956).

Let M_i^* be an operator defined by the relationship

$$M_i^*h(z_1,\ldots,z_p) = h(z_1,\ldots,z_i+1,\ldots,z_p)\,, \tag{6.9}$$

where h is a real valued function. Then the Bayes estimator of θ_i can be expressed as

$$\hat{\theta}_i^*(x,G,m) = \frac{x_i+1}{m+1}M_i^*\{P_{G,m+1}(x)\}/P_{G,m}(x)\,, \tag{6.10}$$

where $P_{G,m}(x)$ is as defined in (1.2) and $P_{G,m+1}(x)$ is also defined as $P_{G,m}$ with $(m+1)$ in place of m.

When an EB scheme with $m_i = m$ $(i = 1,\ldots,n)$ is available, the quantity $P_{G,m}(x)$ can be estimated as in Subsection 6.2. However, $P_{G,m+1}(x)$ cannot be estimated from the data with m trials, thus the quantity $E_i\{P_{G,m+1}(x)\}$ cannot be estimated from the EB scheme.

Following Robbins (1956), we consider the Bayes estimator of θ_i based on $(m-1)$ multinomial trials:

$$\hat{\theta}_i^*(x,G,m-1) = \frac{x_i+1}{m}M_i^*\{P_{G,m}(x)\}/P_{G,m-1}(x)\,. \tag{6.11}$$

The unknown components of the above estimator can now be estimated by using the following.

Let $N_{n,m-1}(x)$ be the number of previous x_i's which are equal to x where the calculation is based on a selected subset of $(m-1)$ observations at each stage of the EB scheme. Let $N_{n,m}(x)$ be defined as before. Then a simple EBE for θ_i can be constructed as

$$\hat{\theta}_i(x, n, m-1) = \frac{x_i + 1}{m} [M_i^*\{N_{n,m}(x)\}/N_{n,m-1}(x)] . \tag{6.12}$$

The estimator (6.12) does not make use of all of the available information, namely all trial results at each stage of the EB scheme. To utilize all the available data, we may regard (6.12) as a randomized estimator based on one of the possible choices of m-results since $N_{n,m-1}(x)$ can be based on any subset of $(m-1)$ trials in the current experiment. By computing (6.12) for every permutation of the current results and averaging them, we obtain an estimator

$$\hat{\theta}_i^{\dagger}(x, n, m) = \sum_{i=1}^{m} x_i M_i^{\dagger}\{\hat{\theta}_i(x, n, m-1)\}/m ,$$

where the operator $M_i^{\dagger}$ is defined as

$$M_i^{\dagger}\{h(z_i, \ldots, z_p)\} = h(z_1, \ldots, z_i - 1, \ldots, z_p) .$$

The estimator $\hat{\theta}_i^{\dagger}(x, n, m)$ is also a simple EB estimator since it does not require an explicit estimation of G.

7. Discussion and conclusion

The binomial case has been widely considered in the literature. A number of non-EB and EB type estimators have been compared in terms of their average risks by Martz and Lian (1974). A similar study for the multinomial case is lacking. The present paper concentrates on the construction of EB estimators for a vector parameter case and specializes in the multinomial distribution. The EB estimators proposed in this paper are asymptotically optimal, in the sense of Robbins (1956), under fairly general conditions. Small sample (n) properties of these estimators need further study.

REFERENCES

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses, *Biometrika*, **67**, 261–272.

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

DeGroot, M. H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.

Graybill, F. A. (1969). *Introduction to Matrices with Applications in Statistics*, Wadsworth, U.S.A.

Griffin, B. S. and Krutchkoff, R. G. (1971). Optimal linear estimators: An empirical Bayes version with application to the binomial distribution, *Biometrika*, **58**, 195–201.

Jackson, D. A., O'Donovan, T. M., Zimmer, W. J. and Deely, J. J. (1970). G_2-minimax estimators in the exponential family, *Biometrika*, **57**, 439–443.

Johns, M. V. (1957). Nonparametric empirical Bayes procedures, *Ann. Math. Statist.*, **28**, 649–669.

Leonard, T. (1972). A Bayesian method for histograms, *Biometrika*, **60**, 297–308.

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

Maritz, J. S. and Lwin, T. (1975). Construction of simple empirical Bayes estimators, *J. Roy. Statist. Soc. Ser. B*, **37**, 421–425.

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

Mosimann, J. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlation among proportions, *Biometrika*, **49**, 65–82.

Robbins, H. (1956). An empirical Bayes approach to statistics, *Proc. 3rd Berkeley Symp. Math. Statist. Probability*, Vol. I, 157–164, University of California Press, Berkeley, California.