EMPIRICAL BAYES ESTIMATION FOR THE MULTINOMIAL DISTRIBUTION

by

L. Billard¹
Florida State University

and

M. R. Meshkani²
College of Statistics and Informatics, Tehran, Iran

FSU Statistics Report M475
July, 1978
Department of Statistics
The Florida State University
Tallahassee, Florida 32306

¹Research supported by National Science Foundation Grant No. MCS76-10453.
²Work performed while at Florida State University.
Empirical Bayes estimation for the multinomial distribution

L. Billard\textsuperscript{1} and M. R. Meshkani\textsuperscript{2}

\textbf{Abstract}

Empirical Bayes estimators for the parameter of a multinomial distribution are constructed. A Monte Carlo study for the particular case of one parameter (i.e., the binomial) shows that the simple empirical Bayes estimators cannot be considered better than the maximum likelihood estimator especially when the variance of the binomial parameter is large. However, the Bayes estimator and the smoothed empirical Bayes estimator obtained by the method of moments, though biased, from a mean square error point of view perform better than the unbiased maximum likelihood estimator.

Keywords: Estimation, Dirichlet prior, Monte Carlo simulation.


\textsuperscript{1}Florida State University. Research supported by National Science Foundation Grant No. M376-10453.

\textsuperscript{2}College of Statistics and Informatics, Tehran. Work performed at Florida State University.
1. INTRODUCTION

In this work, we are concerned with empirical Bayes procedures for estimating the parameter $\lambda = (\lambda_1, \ldots, \lambda_s)$ of a multinomial distribution $M_2(N, \lambda)$. Not only is this of interest in its own right, such a study serves as a first step toward the estimation of the transition probability matrix for a finite state, stationary Markov chain since each row of such a matrix can be compared with the cell probabilities of the multinomial distribution. Although traditional methods of estimation have been used, there is virtually no record of empirical Bayes procedures for Markov chains. Yet, there are many situations which are ideally suited to such techniques, for example, the brand switching application reported by Draper and Nolin (1964).

The concept of empirical Bayes estimation began with Robbins (1955) when he derived asymptotically optimal estimates for parameters of certain common univariate distributions. Neyman (1962) called it a breakthrough in the theory of statistical decision making. The scope was broadened in Robbins (1963, 1964), Samuel (1963) and Hudimoto (1968). Recently, Deely and Zimmer (1976) returned to Robbins' basic idea of asymptotic optimality. Many of these results concern large sample properties. The distributions of empirical Bayes estimators for small sizes of past data are extremely hard to derive. Therefore, their small sample properties have generally only been studied by simulation, Maritz (1970).

In earlier works, the empirical Bayes estimates were constructed utilizing sample frequencies from past data, Robbins (1955), giving rise to so-called simple empirical Bayes estimators. Various simulations
for example, Maritz (1970) showed these estimators were not very efficient. Consequently, in a series of works culminating in a monograph, Maritz (1970) set forth the idea of estimating the distribution function from the past data, producing smoothed empirical Bayes estimators. To achieve these estimators, we may either assume an unspecified prior for the parameter or we may assume a parametric prior whose parameters are themselves estimated from the past data. The former has been considered by many workers, for example, Lin (1975). The latter approach, sometimes with additional assumptions on the prior distribution, has attracted a good deal of attention, for example, Rutherford and Krutchkoff (1967, 1969), Griffin and Krutchkoff (1971), and Nichols and Tsokos (1972).

In a few instances, the empirical Bayes procedure has been specialized to some specific distributions. One of these is the binomial distribution, $\text{B}(N, \theta)$. The earliest work in this spirit is Skelam (1948). In the area of mental test theory where mostly the binomial distribution is used, Meredith and Kearns (1973) and Kearns and Meredith (1975) have worked on a simple empirical Bayes estimate of $\theta$. They found that simple empirical Bayes estimators perform poorly compared to smoothed empirical Bayes estimators. From the Monte Carlo study of the present work (see Section 5), the simple empirical Bayes estimators cannot be considered better than the maximum likelihood estimator, especially when $\text{Var}(\theta)$ is large. Maritz (1970) has suggested some methods of smoothing. Martin and Lian (1974) have carried out a fairly large simulation to compare some empirical Bayes estimators of $\theta$. 
When we consider the multinomial distribution, we find the problem of empirical Bayes estimation for the vector of cell probabilities is practically untouched. A work in this spirit, but for a completely different purpose, is Mosimann (1962). He used a Dirichlet prior for the vector of cell probabilities and estimated the parameters of the prior. Recently, Alam and Ahlers (1977) considered the problem from a multiple decision point of view.

Our aim then is to help breach this gap in the current knowledge for the multinomial distribution. Thus, after giving the basic probability model in Section 2, the Bayes estimates corresponding to a general prior and to the natural conjugate prior, viz., the Dirichlet prior, are obtained in Section 3. Empirical Bayes estimates are constructed in Section 4. Finally, in Section 5, a Monte Carlo study is reported.
2.1. Notations and assumptions.

Let \( Y \) be a discrete random vector with \( d \geq 2 \) components. Let \( S = \{1, \ldots, d\} \). Let \( \Omega = \{\Lambda : \Lambda_j \geq 0, \sum_{j \in S} \Lambda_j = 1\} \). For a given \( \Lambda = (\Lambda_1, \ldots, \Lambda_d) \in \Omega \) and a fixed sample size \( N \), \( Y \) has a multinomial distribution \( M_\Lambda(N, \Lambda) \). That is,

\[
(2.1) \quad P(Y = y | \Lambda) = \frac{N!}{\prod_{j \in S} y_j!} \prod_{j \in S} \Lambda_j^{y_j}, \quad y \in \mathcal{V},
\]

where

\[
\mathcal{V} = \{y : y_j = 0, \ldots, N, j \in S, \sum_{j \in S} y_j = N\}.
\]

Let \( \Lambda \) be a realization of a random stochastic vector whose distribution function is \( G \). For a general prior distribution \( G \), the unconditional probability mass function of \( Y \) is found to be

\[
(2.2) \quad P_G(Y = y) = \frac{N!}{\prod_{j \in S} y_j!} \int \left( \prod_{j \in S} \Lambda_j^{y_j} \right) dG(\Lambda), \quad y \in \mathcal{V}.
\]

It is assumed (2.2) gives positive probabilities for all possible \( y \in \mathcal{V} \).

The natural conjugate prior, the Dirichlet prior \( D(\alpha) \), is of particular interest. Here,

\[
(2.3) \quad q(\Lambda) = q(\Lambda | \alpha) = g(\alpha) \prod_{j \in S} \Lambda_j^{\alpha_j - 1}, \quad \Lambda \in \Omega,
\]

where

\[
g(\alpha) = \Gamma(\alpha_+) / \prod_{j \in S} \Gamma(\alpha_j)
\]
with $\alpha_1 = \sum_{j \in S} \alpha_j$. For this prior, the unconditional probability mass function of $Y$ becomes

$$(2.4) \quad P_D(Y = y) = \left( \frac{N!}{\prod_{j \in S} y_j!} \right) g(\alpha + y), \; y \in V.$$ 

Let $Z$ be a random vector, of the same dimension as $Y$, indicating the occurrence of $\Lambda$ mutually exclusive and collectively exhaustive events with probability vector $\Lambda$. The random vector $Z$ is independent of $Y$, given $\Lambda$. It can be viewed as the indicator of the $(N + 1)$th trial in an experiment where the first $N$ trials give rise to $Y$.

The empirical Bayes assumption is that we assume there is a sequence of independent and identically distributed triplets $(\Lambda, Y, Z)$

$$\{(\Lambda_1, Y_1, Z_1) \; i \in N_1\},$$

where $N_1 = \{1, \ldots, n + 1\}$; and where $\Lambda_1, \; i \in N_1$ remain unobservable. The data $\{(Y_i, Z_i) \; i \in N_1\}$ are available and the objective is to estimate $\Lambda \equiv \Lambda_{n+1}$. When working with the Dirichlet prior, we do not need to know the outcome of the $Z_1$, separately. Hence, the data $\{Y_i, \; i \in N_1\}$ is sufficient to estimate $\Lambda$.

It is well known that the random vectors $X_i$ where

$$(2.5) \quad X_i = Y_i + Z_i, \; i \in N,$$

are independently and identically distributed with probability mass function
\[ P_G(x = x) = \frac{(N + 1)!}{(\prod x_j!)} \left( \prod_{j \in \Omega} \left( \int_{x_j}^\Lambda \right) dG(\Lambda) \right), \ x \in X, \]

where \[ X = \{x: x_j = 0, \ldots, N + 1, j \in S, \prod_{j \in S} x_j = N + 1\}; \]
or

\[ P_D(x = x) = (N + 1)! g(x) / (g(x) + \delta_j), \ x \in X, \]

for prior distributions \( G \), or \( D(a) \), respectively.

2.2. Moments of \( X \) and \( Y \).

We shall derive the moments of \( X \) and \( Y \) in this section. These are required for the estimation of \( \Lambda \) later.

**Theorem 2.1.**

Let \( G \) be an unspecified prior distribution function for \( \Lambda \).

Assume \( G \) has finite second moments. Then, \( G(\Lambda) \)

\[ E(Y) = \psi_G(X) \]

and

\[ E(Y - E(Y)) (Y - E(Y))' = \int_G \]

where \( \psi_G \) has elements \( \psi_G; j \) given by

\[ \psi_G; j = \int_{x_j}^\Lambda dG(\Lambda), j \in S, \]

and \( \int_G \) has elements \( \sigma_{G; jk} \) given by

\[ \sigma_{G; jk} = N(N - 1)\psi_{G; jk} + N\psi_{G; j} (\delta_{jk} - N\psi_{G; k}), j, k \in S, \]

\[ \psi_{G; j} = \int_{x_j}^\Lambda dG(\Lambda), j \in S. \]
where $\delta_{jk}$ is Kronecker's delta and

$$\psi_G; jk = \int_{\Omega} \Lambda^j \Lambda^k \, dG(\Lambda).$$

(2.12)

**Proof.**

This is straightforward using conditional expectation. $\Box$

**Corollary 2.2.**

Let $G$ be an unspecified prior distribution for $\Lambda$ as in Theorem 2.1. Then,

$$E(X) = (N + 1)\psi_G,$$

(2.11)

and

$$E[X - E(X)](X - E(X))' = \Sigma_{Gl}$$

(2.12)

where $\psi_G$ is defined in (2.10) and $\Sigma_{Gl}$ has elements $\sigma_{Gl;jk}$ given by

$$\sigma_{Gl;jk} = N(N + 1)\psi_G;jk + (N + 1)\psi_G;j [\delta_{jk} - (N + 1)\psi_G;k].$$

(2.13)

**Proof.**

We recall $X = \overline{Y} + Z$ has a $N_\delta(N + 1, \Lambda)$ distribution for each $\Lambda \in \Omega$ and use Theorem 2.1. $\Box$

**Theorem 2.3.**

Suppose the prior distribution $G$ in Theorem 2.1 is specialized to a Dirichlet prior $D(\alpha)$. Then,

$$E(X) = (N/\alpha) \cdot \alpha$$

(2.14)
and

\[ E[Y - E(Y)][Y - E(Y)]' = \sum_D \sigma_{D;jk} \delta_{jk} \sigma_{D;jk} \]

where \( \sum_D \) has elements \( \sigma_{D;jk} \) given by

\[ \sigma_{D;jk} = N[(N + \alpha_+)/(1 + \alpha_+)](\alpha_j/\alpha_+)(\delta_{jk} - \alpha_k/\alpha_+), \quad j, k \in S. \]

**Proof.**

The proof is similar to Theorem 2.1. However, in this case the moments are explicitly derived as

\[ \psi_{D;j} = \psi(\alpha) \int_{\Omega} \left( \prod_{k \in S} \alpha_k^{-1} \right) d(\Lambda) = \alpha_j/\alpha_+, \quad j \in S, \]

and

\[ \psi_{D;jk} = \psi(\alpha) \int_{\Omega} \left( \prod_{k \in S} \alpha_k^{-1} \right) d(\Lambda) = \alpha_j \alpha_k/(\alpha_+ + 1 + \alpha_+), \quad j, k \in S. \]

Finally, let us define a matrix \( \sum_D \) with elements \( \sigma_{jk} \) given by

\[ \sigma_{jk} = E(Y_j)[\delta_{jk} - N^{-1} E(Y_k)], \quad j, k \in S. \]

Then,

\[ \sum_D = \sigma \]

where

\[ \sigma = (N + \alpha_+)/(1 + \alpha_+). \]
3. BAYES ESTIMATE OF $\Lambda$

3.1. The loss function.

The Bayes risk relative to a prior $G$ is defined as

$$W(d, G) = \int \int L[d(x), \Lambda] dF(x|\Lambda) dG(\Lambda)$$

where $L[d(x), \Lambda]$ is the loss incurred by taking action $d(x)$ upon observing $x \in X$. Following DeGroot (1970), we assume the squared error loss function

$$L(d, \Lambda) = \sum_{j \in S} L_j(d_j, \Lambda_j) = \sum_{j \in S} (d_j - \Lambda_j)^2$$

where $d = (d_1, \ldots, d_\delta)$. Consequently,

$$W(d, G) = \sum_{j \in S} W_j(d_j, G)$$

where

$$W_j(d_j, G) = \int \int (d_j - \Lambda_j)^2 dF(x|\Lambda) dG(\Lambda).$$

We immediately have

Lemma 3.1.

For $L(d, \Lambda)$ given in (3.1), the minimum Bayes risk for $d$ is achieved by having minimum Bayes risk for all $d_j, j \in S$.

Proof.

It is clear from (3.2) that if $d_j$ has the minimum Bayes risk $W_j(d_j, G)$ for all $j \in S$, then $W(d, G)$ achieves its minimum. \(\square\)
The implication of Lemma 3.1 to our estimation problem is that to obtain the Bayes estimate $d_G$ relative to $G$, we need to obtain the Bayes estimate of $d_{Gj}$ relative to $G$ for each $j \in S$.

3.2. Posterior distribution of $\Lambda$.

For the squared error loss function, it is well known that the Bayes estimate is the posterior mean of the parameter [see, for example, Maritz (1970)].

**Lemma 3.2.**

Let $Y$ be a random vector with multinomial distribution $M_\lambda(N, \Lambda)$ for a given $\Lambda$. Let $\Lambda$ have a prior distribution $G$. Then, the posterior distribution of $\Lambda$ given $Y = \hat{y}$, is

$$dQ_\Lambda^*(\Lambda) = \left( \prod_j \Lambda_j^{y_j} \right) dG(\Lambda) / \left( \prod_j \Lambda_j^{y_j} \right) dG(\Lambda).$$

**Proof.**

By Bayes Theorem, the result is immediate.

**Lemma 3.3.**

Let $Y$ and $\Lambda$ be the same as in Lemma 3.2. Let $\Lambda$ have a Dirichlet prior distribution $D(\alpha)$. Then, the posterior distribution of $\Lambda$ given $Y = \hat{y}$ is a Dirichlet distribution $D(\alpha + \hat{y})$.

**Proof.**

Again, by Bayes Theorem and from (2.4),

$$q_\Lambda^*(\lambda) = g(\alpha + \hat{y}) \prod_j \Lambda_j^{y_j + \alpha_j - 1}, \Lambda \in \Omega.$$
This is obviously $D(\alpha + \gamma)$. □

By the aid of these lemmas, we can obtain the Bayes estimate of

$$\mathbf{\Lambda}, \mathbf{\Lambda}_B = (\Lambda_{B;1}, \ldots, \Lambda_{B;\delta}).$$

**Lemma 3.4.**

Let $\mathbf{Y}$ have a multinomial distribution, $\mathcal{H}_\delta(N, \Lambda)$, $\Lambda \in \Omega$. Then, the marginal distribution of each component $Y_i$, is $\text{Bin}(N, \Lambda_i)$. Moreover, the conditional distribution of the remaining components is $\mathcal{H}_{\delta-1}(N^*, \Lambda^*)$ given $Y_j = y_j$, where $N^* = N - y_j$ and $\Lambda^* = (1 - \Lambda_j)^{-1}[(\Lambda_1, \ldots, \Lambda_j-1, \Lambda_{j+1}, \ldots, \Lambda_{\delta})].$

**Proof.**

From (2.1), we have

$$P(Y = \mathbf{y} | \Lambda) = N! \prod_{k \in S} \frac{y_k}{\Lambda_k} \Lambda_k^{y_k}$$

$$= N! \left(1 - \Lambda_j\right)^{N - y_j} \left(\Lambda_j / y_j\right)! \left((N - y_j)! \prod_{k \neq j} \left[\Lambda_k / (1 - \Lambda_j) \right] \right)$$

$$= P(Y_j = y_j | \Lambda_j) P(\mathbf{y}^* = \mathbf{y}^* | \Lambda^*, y_j)$$

where

$$\mathbf{y}^* = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_\delta).$$

We shall write the result of Lemma 3.4 as

$$\mathcal{H}_\delta(N, \Lambda) = \text{Bin}(N, \Lambda_j) \cdot \mathcal{H}_{\delta-1}(N^*, \Lambda^*), j \in S.$$
Lemma 3.5.

Let \( \underline{Y} \) be a random vector as in Lemma 3.4. Let \( \underline{\Lambda} \) have a prior d.f. \( G \) such that \( P_{G} (Y_{j} = y_{j}) > 0, \; y_{j} \in Y \). Then, for each \( j \in S \), the posterior distribution of \( \Lambda_{j} \) only depends on \( y_{j} \).

Proof.

By definition,

\[
dQ_{j}^{*}(\Lambda_{j}) = \int_{\Omega} q_{G}(\underline{\Lambda}) q_{j}^{*}(\Lambda_{j}) \prod_{k \neq j} d\Lambda_{k}.
\]

We decompose \( G \) as

\[
G(\underline{\Lambda}) = G_{j}(\Lambda_{j}) G^{*}(\Lambda_{j}; j),
\]

where \( G^{*}(\Lambda_{j}; j) \) denotes the conditional joint distribution of \( \Lambda_{k}, \; k \in S, \; k \neq j \), given \( \Lambda_{j} \). Applying (3.4) and 3.5, we have

\[
dQ_{j}^{*}(\Lambda_{j}) = \int_{\Omega} \frac{P(\underline{Y} = \underline{y}|\underline{\Lambda}) dG(\underline{\Lambda})}{\Omega} \prod_{k \neq j} d\Lambda_{k}.
\]

\[
= \int_{\Omega} \frac{P(Y_{j} = y_{j}|\Lambda_{j}) dG_{j}(\Lambda_{j}) P(\underline{y}_{\ast}|\Lambda_{j})}{\Omega} \prod_{k \neq j} d\Lambda_{k}.
\]

\[
= P(Y_{j} = y_{j}|\Lambda_{j}) dG_{j}(\Lambda_{j}) / \Omega P(Y_{j} = y_{j}) dG_{j}(\Lambda_{j}).
\]

We now obtain the Bayes estimate of \( \Lambda \), denoted by \( \Lambda_{BG} \) or \( \Lambda_{BD} \) according to the prior used.
Theorem 3.6.

Let $Y$ and $\Lambda$ be the same as in Lemma 3.5. Let $Y$ be an outcome of $Y$. Then for the squared error loss function, the Bayes estimate of $\Lambda$ relative to $G$ is

$$\Lambda_{BG} = \left( \Lambda_{BG}^* \right)^*$$

where

$$\Lambda_{BG}^* = \left( \Lambda_{BG}^* \right)^* = (y_j + 1) p_G(x_j = y_j + 1)/(N + 1) p_G(y_j = y_j), \ j \in S.$$ 

Proof.

By Lemma 3.1, we need to find $E(\Lambda_j | Y)$, $j \in S$. From Lemma 3.5, we have

$$\Lambda_{BG}^* = \left( \Lambda_{BG}^* \right)^* = \frac{\int \Lambda_j p(Y_j = y_j) dG_j(\Lambda_j)}{\int p(Y_j = y_j) dG_j(\Lambda_j)}, \ j \in S.$$ 

Since $Y_j$ is Bin$(N, \Lambda_j)$, by (2.2) and (2.6),

$$\Lambda_{BG}^* = \left[ (y_j + 1)/(N + 1) \right] p_G(x_j = y_j + 1)/p(Y_j = y_j).$$

Theorem 3.7.

Let the Dirichlet prior $D(\alpha)$ be used in place of $G$ in Theorem 3.6. Then, the Bayes estimate of $\Lambda$ relative to $D(\alpha)$ is

$$\Lambda_{BD} = \left( \Lambda_{BD}^* \right)^*$$

where

$$\Lambda_{BD}^* = \left( \Lambda_{BD}^* \right)^* = (y_j + \alpha_j)/(N + \alpha_0), \ j \in S.$$ 

Proof.

As in Theorem 3.6, $\Lambda_{BD}^* = E(\Lambda_j | Y)$, $j \in S$. From Lemma 3.3 and equation (2.3), this is
\[ \hat{\Lambda}_{BD;j} = \frac{\alpha_j}{\alpha_j + \gamma_j} \sum_{k=1}^{a_j} \int_{\Omega_j} \hat{\Lambda}_{j,k} d\hat{\Lambda}_{k} = \frac{\gamma_j + a_j}{\gamma_j + n + a_j} \hat{\Lambda}_j. \]

The Bayes estimate (3.7) can be expressed as a convex combination of the maximum likelihood estimate of \( \hat{\Lambda}_j \), \( \hat{\Lambda}_j \), and of \( E(\hat{\Lambda}_j) \).

That is,

\[ \hat{\Lambda}_{BD;j} = \left[ \frac{n}{n + \alpha_j} \right] \hat{\Lambda}_j + \left[ \frac{\alpha_j}{n + \alpha_j} \right] E(\hat{\Lambda}_j). \]

Then, as Bishop et. al. (1975, p. 408) point out there are points along the line connecting \( \hat{\Lambda}_j \) and \( E(\hat{\Lambda}_j) \) that are closer to \( \hat{\Lambda}_j \) than is \( \hat{\Lambda}_j \).

This fact leads one to consider estimates of \( \hat{\Lambda}_j \) that are formed by shrinking \( \hat{\Lambda}_j \) toward the origin \( E(\hat{\Lambda}_j) \).

4. EMPIRICAL BAYES ESTIMATE OF \( \Lambda \)

4.1. Preliminaries.

As indicated in Section 2, we assume that \( \Lambda \) has a prior distribution.

However, even if we choose a family of distributions to which the prior of \( \Lambda \) belongs, we do not have enough information to completely specify the prior. Thus, we encounter a problem in using the Bayes estimate of \( \Lambda \).

One way round this problem is to use the information provided by the "past data" in situations similar to the "current" one. This is accomplished by considering the sequence \( \{(\mathbf{X}_i, \mathbf{Z}_i), i \in N\} \), where \( N = \{1, \ldots, N\} \), the outcomes of independent and identically distributed pairs \( (\mathbf{X}_i, \mathbf{Z}_i) \), given in (2.5). We note that the unconditional distribution of \( \mathbf{X} \) as given in (2.4) is completely determined by the parameter \( \alpha \).
On the other hand, we have a set of observations on $Y$ which can be used to estimate $\alpha$. Now, we utilize the Bayes estimate given in (3.7) and substitute for $\alpha$. The resulting value will be called the empirical Bayes estimate of $\alpha$.

It is clear that the degree of usefulness of this estimate depends on the efficiency of the estimation of $\alpha$. Various methods of estimation can be employed such as method of moments, maximum likelihood, etc. They would give rise to different empirical Bayes estimators which can be compared by Monte Carlo methods for small samples (see Section 5) or by asymptotic properties for large samples.

4.2. Simple empirical Bayes estimate of $\alpha$.

When nothing is assumed about the prior distribution except that it exists and has a finite second moment, the empirical Bayes estimate is called the simple empirical Bayes estimate. In so doing, we arrive at the Bayes estimate given in (3.6).

It is recalled that the data

\[(4.1)\quad \{(y_{i1}, x_{ij}), i \in N\},\]

are available, [see (2.5)]. Thus, we can estimate the quantities

\[
P_0(x_{ij} = y_j + 1) \quad \text{and} \quad P(y_j = y_j), \quad j \in S, \quad \text{needed in (3.6). Let us define}
\]

\[(4.2)\quad S_{nj} = \text{(no. of terms equal to } y_j \text{ in the sequence } \{y_{ij}\}, i \in N)\]

\[(4.3)\quad T_{nj} = \text{(no. of terms equal to } (y_j + 1) \text{ in the sequence } \{x_{ij}\}, i \in N)\]
where $y_{ij}$ and $x_{ij}$ are the $j^\text{th}$ elements of $y_i$ and $x_i$, respectively.

We estimate $P_G(Y_j = y_j)$ and $P_G(X_j = y_j + 1)$ by sample frequencies and denote them respectively by

$$P_{1j} = \frac{s_{nj}}{n + 1}, \quad j \in S,$$

and

$$P_{2j} = \frac{f_{nj}}{n}, \quad j \in S.$$

Hence, the simple empirical Bayes estimate of $\Lambda$, denoted by $\hat{\Lambda}_F = (\hat{\Lambda}_{F;j}),$

is defined as

$$(4.4) \quad \hat{\Lambda}_{F;j} = \frac{[n + 1](Y_j + 1)T_{nj}}{n(N + 1)(s_{nj} + 1)}, \quad j \in S.$$  

Since estimation by frequencies does not extract the whole information in the data, we do not expect a very good performance from our simple empirical Bayes estimators, unless the past data is very large. Hence, we consider the other methods while assuming a parametric function for the prior distribution of $\Lambda$.

4.3. Smooth empirical Bayes estimate of $\Lambda$.  

The smooth empirical Bayes estimate of $\Lambda$ is the one in which the estimated parameters of the prior are used. Thus, there will be different versions of the empirical Bayes estimate of $\Lambda$ depending on the method of estimation used.

First, we consider the method of moments. Now, (2.14) provides $s$ equations in $s$ unknowns $\alpha_j$, $j \in S$. However, since $\sum_{j \in S} Y_j = N$, we have
Thus, we need one additional equation to be able to solve for \( \alpha_j, j \in S \).

This can be supplied by (2.18). It is known that \( \Sigma_D \) and \( \Sigma \) are singular
due to \( \sum_{j \in S} Y = N \). Let superscript * on a matrix denote that its last row
and column are deleted. Then, as Mosimann (1962) has suggested, (2.18)
is solved for \( \beta \) by taking determinants of both sides

\[
\beta = \left( |\Sigma_D^*| / |\Sigma| \right)^{1/(\delta-1)}.
\]

Let us define the sample mean and covariances

\[
\bar{Y} = \frac{1}{n} \sum_{i \in N} Y_i = \langle Y_j \rangle, j \in S,
\]

\[
\hat{\sigma}_{D;jk} = \frac{1}{n} \sum_{i \in N} (y_{ij} - \bar{Y}_j)(y_{ik} - \bar{Y}_k), j, k \in S,
\]

and

\[
\hat{\sigma}_{jk} = \bar{Y}_j (\delta_k - n^{-1} \bar{Y}_k), j, k \in S.
\]

Let \( \Sigma_D^* \) and \( \Sigma \) have as their \((j, k)\)th element the terms defined in (4.7) and (4.8), respectively.

Now, an estimate of \( \beta \), denoted by \( b \), is

\[
b = \hat{b} = \left( |\Sigma_D^*| / |\Sigma| \right)^{1/(\delta-1)}.
\]

From (2.19) for \( \delta = 1 \), we have
Using \( \sum_j a_j (2.14) \) in place of \( E(Y) \), we obtain

\begin{equation}
\hat{\alpha}_j = \sum_j a_j = (N - b)/(b - 1), \quad j \in S
\end{equation}

and

\begin{equation}
0, \quad j \in S_1 = S - \{j\}
\end{equation}

where we have used \( a_k \) as an estimate of \( \alpha_k \). We solve (4.10) and (4.11) simultaneously for \( a_k, k \in S \). The solution is

\begin{equation}
\hat{a}_k = \left[ \frac{(N - b)/N(b - 1)} \right] \bar{y}_k, \quad k \in S.
\end{equation}

Having obtained the estimates of \( \alpha_k, k \in S \), we construct the empirical Bayes estimate of \( \Lambda \), denoted by \( \Lambda_1(D) = \{\Lambda_1; j(D)\}, j \in S \). Thus, we have

\begin{equation}
\Lambda_1(D) = \{\Lambda_1; j(D)\}, \quad j \in S.
\end{equation}

**Definition 4.1.**

The empirical Bayes estimate of \( \Lambda \) obtained by the method of moments is the vector \( \Lambda_1(D) \) whose elements \( \Lambda_1; j(D), j \in S \), are given by

\begin{equation}
\Lambda_1; j(D) = (a_j + y_j)/(\sigma^2 + a_j), \quad j \in S.
\end{equation}

An alternative way of estimating \( \alpha \) is the maximum likelihood method. From (2.4), the likelihood of \( \{y_i, i \in N\} \) is

\begin{equation}
L(\alpha|y_i, i \in N) = \prod_{i \in N} g(\alpha + y_i).
\end{equation}

Levin and Reeds (1977) have shown that this is unimodal. Although this equation does not lend itself to a tractable analytical method of maxi-
mization in $a_j$, $j \in S$, there are iterative methods such as Newton-Raphson which can be used to find the maximizing value of $a$. Let the final solution be $\hat{a}$. Then, the empirical Bayes estimator of $\Lambda$, denoted by $\Lambda_2(D) = [\Lambda_{2;j}(D)]$, $j \in S$, is given in

\textbf{Definition 4.2.}

The empirical Bayes estimate of $\Lambda$ obtained by the method of maximum likelihood is the vector $\Lambda_2(D)$ whose elements $\Lambda_{2;j}(D)$, $j \in S$, are given by

\begin{equation}
\Lambda_{2;j}(D) = (\hat{a}_j + y_j)/(N + \hat{a}_j).
\end{equation}

\textbf{5. Preliminaries.}

A Monte Carlo study was designed to compare the average performance of the empirical Bayes estimators with Bayes and maximum likelihood estimators. The objectives of this study were

(i) To ascertain the effect of the shape of the prior distribution;

(ii) To ascertain the effect of the size of the past data; and

(iii) To ascertain the effect of the sample size on the performance of the empirical Bayes estimators.

Since the $\text{Var} (\Lambda)$ determines the shape of the beta distribution it will be used as the indicator of the shape of the prior distribution.

We have considered the binomial distribution, i.e., $\delta = 2$. For a fixed sample size $N$, from (2.1) and (2.5) $Y \sim \text{Bin}(N, \Lambda)$ and $X \sim \text{Bin}(N + 1, \Lambda)$ where $\Lambda$ has a beta distribution, denoted by $\text{Beta}(\alpha, \beta)$. The expected value and variance of $\Lambda$ are given by

\begin{align*}
\text{E}(\Lambda) &= \frac{N + \alpha}{N + \alpha + \beta + 1}, \\
\text{Var}(\Lambda) &= \frac{\alpha \beta}{(N + \alpha + \beta + 1)^2 (N + \alpha + \beta + 2)}.
\end{align*}
In this study, values of \((\rho_1, \rho_2)\) were chosen so that \(E(\Lambda)\) and \(\text{Var}(\Lambda)\) present the situations that we can expect in practice, viz.,

\[
E(\Lambda) \approx \begin{array}{llllllllll}
0.50 & 0.40 & 0.36 & 0.32 & 0.28 & 0.24 & 0.20 & 0.16 & 0.12 & 0.08 \\

decreases & decreases & decreases & decreases & decreases & decreases & decreases & decreases & decreases & decreases
\end{array}
\]

\[
\text{Var}(\Lambda) \approx \begin{array}{llllllllll}
0.50 & 0.15 & 0.10 & 0.08 & 0.05 & 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\
\end{array}
\]

\((\rho_1, \rho_2)\): \((3, 3)\) \((4, 6)\) \((9, 13)\) \((14, 25)\) \((50, 22)\) \((53, 5)\) \((55, 8)\).

Sample sizes \(N = 5, 10, 15, 20\) and the sizes of the past data \(n = 10, 20,\) and 50 have been considered.

5.2. Computations.

For an observed \(y\), the maximum likelihood estimate of \(\Lambda\) is

\[
\hat{\Lambda}_1 = y/N;
\]

while the Bayes estimate is

\[
\hat{\Lambda}_2 = (y + \rho_1)/(N + \rho_1 + \rho_2);
\]

Suppose we have the past data in the form of a sequence

\[
\{y_i, x_i, 1 \leq i \leq N\}
\]

Then, the simple empirical Bayes estimate of \(\Lambda\) is

\[
\hat{\Lambda}_3 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i + 1}{n + 1} \right) \left( \frac{x_i + 1}{n + 1} \right);
\]

where \(S_n\) and \(T_n\) are defined in (4.2) and (4.3). If we have the past data only as

\[
\{y_i, 1 \leq i \leq N\},
\]

then the smooth empirical Bayes estimate of \(\Lambda\) is

\[
\hat{\Lambda}_4 = \frac{y + r_1}{N + r_1 + r_2}
\]

where \(r_1\) and \(r_2\) are obtained in a manner similar to (4.12).
The Bayes risk for \( \Lambda_1 \) and \( \Lambda_2 \) are, respectively,

\[
W_1 = E[E[(Y/N - \Lambda)^2] | \Lambda] = \rho_1 \rho_2 / (N \rho_1 + \rho_2)(\rho_1 + \rho_2 + 1)
\]

and

\[
W_2 = \rho_1 \rho_2 / (N \rho_1 + \rho_2)(\rho_1 + \rho_2)(\rho_1 \rho_2 + 1) = N W_1 / (N \rho_1 + \rho_2).
\]

For each case, \( W_1 \) and \( W_2 \) are given so that comparisons with the risks of other estimators can be made.

For each pair \((\rho_1, \rho_2)\), a sequence of past data like (5.1) with size \((n+1)\) was generated. In each case \( \Lambda \equiv \Lambda_{n+1} \) was to be estimated. This process was repeated 500 times and then averaged. In Meshkami (1978), tabulated values for the estimates of \( \Lambda \) for most (but not all) combinations of these \((\rho_1, \rho_2)\), \(N\) and \(n\) are provided. Also given are tables in which a comparison of the estimates with respect to variance is made, as well as plots of the mean square error against \( \text{Var}(\Lambda) \). In the next subsection, we present just a representative sample of these results. The conclusions that can be drawn from these particular cases apply equally to the other cases.

In Tables 5.1 and 5.2, the estimates so obtained for \( \Lambda_i \), \(i = 1, \ldots, 4\), are given for \( N = 10 \) and \((\rho_1, \rho_2) = (5, 22)\), and for \( N = 20 \) and \((\rho_1, \rho_2) = (3, 3)\), respectively. Also given are values of the respective mean square errors (MSE) given by

\[
\text{MSE} = (499)^{-1} \sum_{j=1}^{500} (\Lambda_j - \Lambda)^2.
\]
5.3. Conclusions.

From these results, we can see that $\Lambda_1$ is unbiased but $\Lambda_2$, $\Lambda_3$, and $\Lambda_4$ are biased. The M.S.E. for $\Lambda_1$ is generally larger than the M.S.E. for $\Lambda_2$ and $\Lambda_4$. However, the M.S.E. for $\Lambda_1$ is smaller than the M.S.E. for $\Lambda_3$ in most cases. Thus, while $\Lambda_2$ and $\Lambda_4$ are biased estimators, from an M.S.E. point of view they perform better than $\Lambda_1$ in various ranges of situations.

We recall that in computing $\Lambda_1$ and $\Lambda_2$, the past data were not used. Hence, we expect to see the effect of an increase in the past data only on the M.S.E. of $\Lambda_3$ and $\Lambda_4$. This effect is indeed observed in all cases.

It is noted that $\text{MSE}(\Lambda_4)$ on average drops more than $\text{MSE}(\Lambda_3)$, indicating effective use of information in the past data by $\Lambda_4$. The overall examination of these tables indicates that in various situations, the estimator $\Lambda_2$ is the best, $\Lambda_4$ is better than $\Lambda_3$ and $\Lambda_1$, and $\Lambda_1$ is better than $\Lambda_3$.

To observe this superiority, we make other comparisons as follows.

Tables 5.3 and 5.4 show the variation of the M.S.E. for $(N, n) = (10, 10)$ and $(N, n) = (15, 20)$, respectively. It is noted that as $\text{Var}(\Lambda)$ increases, the $\text{MSE}(\Lambda_j)$, $j = 1, 2, 3, 4$, generally increases.

Figures 5.1, 5.2, and 5.3 show the effect of sample size on the M.S.E. for a fixed $n$. It is seen that the sample size $N$ has no drastic effect on the M.S.E. Going from $N = 10$ to $N = 20$, we can see a little change in the values of the M.S.E. This conclusion is also true for the
other values of $n$ and $N$. For each pair $(N, n)$, the general pattern indicates that

$$\text{MSE}(\Lambda_2) < \text{MSE}(\Lambda_4) < \text{MSE}(\Lambda_1) < \text{MSE}(\Lambda_3)$$

except for a few points. The difference between $\text{MSE}(\Lambda_3)$ and the M.S.E. for other estimators is considerable, indicating inefficiency of $\Lambda_3$. We note that $\text{MSE}(\Lambda_4)$ almost always is less than $\text{MSE}(\Lambda_1)$. In some cases, like $(N, n) = (10, 20)$ and $(N, n) = (10, 50)$, the difference between $\text{MSE}(\Lambda_1)$ and $\text{MSE}(\Lambda_4)$ is also considerable, showing the superiority of $\Lambda_4$ to $\Lambda_1$.

Comparing Figures 5.1 and 5.4, we can see the effect of $n$ on the M.S.E. for $N = 10$. It is observed that as $n$ increases, $\text{MSE}(\Lambda_j)$, $j = 3, 4$ decreases and $\text{MSE}(\Lambda_4)$ approaches $\text{MSE}(\Lambda_2)$, while $\text{MSE}(\Lambda_3)$ remains far away. This conclusion also applied to the other cases.

With regard to all aspects discussed above, we can conclude that the simple empirical Bayes estimate $\Lambda_3$ performs relatively poorly and cannot be recommended as a better alternative to the maximum likelihood estimator $\Lambda_1$. The smooth empirical Bayes estimator $\Lambda_4$ is in almost every situation better than $\Lambda_1$ and it is proposed as a better alternative to $\Lambda_1$. 
REFERENCES


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

Neyman, J. (1962). Two breakthroughs in the theory of statistical

Nichols, W. G. and Tsokos, C. P. (1972). Empirical Bayes point estima-
tion in a family of probability distributions. In. Statist. Rev. 40,
147-151.


Robbins, H. (1964). The empirical Bayes approach to statistical deci-

approach: estimating the prior distribution. Biometrika 54,
326-328.

techniques in point estimation. Biometrika 56, 133.

Samuel, E. (1963). An empirical Bayes approach to the testing of cer-

Skelam, J. G. (1948). Probability distribution derived from the bi-
### TABLE 5.1

Estimates for $\Lambda$

<table>
<thead>
<tr>
<th>$N = 10$</th>
<th>$\left(\rho_1, \rho_2\right) = (50, 22)$, $\left(\mu_1, \mu_2\right) = (0.02093, 0.00255)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 10$</td>
<td>$\Lambda$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot70$</td>
</tr>
<tr>
<td>$n = 20$</td>
<td>$\cdot67$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot02454$</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>$\cdot71$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot02112$</td>
</tr>
</tbody>
</table>

### TABLE 5.2

Estimates for $\Lambda$

<table>
<thead>
<tr>
<th>$N = 20$</th>
<th>$\left(\rho_1, \rho_2\right) = (3, 3)$, $\left(\mu_1, \mu_2\right) = (0.01071, 0.00824)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 10$</td>
<td>$\Lambda$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot38$</td>
</tr>
<tr>
<td>$n = 20$</td>
<td>$\cdot63$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot01230$</td>
</tr>
<tr>
<td>$n = 50$</td>
<td>$\cdot57$</td>
</tr>
<tr>
<td>MSE</td>
<td>$\cdot01327$</td>
</tr>
</tbody>
</table>
### TABLE 5.3
Comparison of estimators with respect to variance

<table>
<thead>
<tr>
<th>( \text{Var}(\Lambda) )</th>
<th>.04</th>
<th>.05</th>
<th>.08</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\rho_1, \rho_2))</td>
<td>(53, 5)</td>
<td>(50, 22)</td>
<td>(14, 25)</td>
<td>(9, 13)</td>
<td>(4, 6)</td>
<td>(3, 3)</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_1))</td>
<td>.00556</td>
<td>.02454</td>
<td>.02632</td>
<td>.02319</td>
<td>.02030</td>
<td>.02257</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_2))</td>
<td>.00056</td>
<td>.00037</td>
<td>.00158</td>
<td>.00274</td>
<td>.00834</td>
<td>.01179</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_3))</td>
<td>.07131</td>
<td>.07197</td>
<td>.05681</td>
<td>.04969</td>
<td>.04632</td>
<td>.06140</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_4))</td>
<td>.00267</td>
<td>.01305</td>
<td>.01067</td>
<td>.00856</td>
<td>.01260</td>
<td>.02926</td>
</tr>
</tbody>
</table>

### TABLE 5.4
Comparison of estimators with respect to variance

<table>
<thead>
<tr>
<th>( \text{Var}(\Lambda) )</th>
<th>.05</th>
<th>.08</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\rho_1, \rho_2))</td>
<td>(50, 22)</td>
<td>(14, 25)</td>
<td>(9, 13)</td>
<td>(4, 6)</td>
<td>(3, 3)</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_1))</td>
<td>.01516</td>
<td>.01446</td>
<td>.01573</td>
<td>.01721</td>
<td>.01114</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_2))</td>
<td>.00101</td>
<td>.00192</td>
<td>.00378</td>
<td>.00727</td>
<td>.01256</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_3))</td>
<td>.06743</td>
<td>.04276</td>
<td>.05080</td>
<td>.05620</td>
<td>.03887</td>
</tr>
<tr>
<td>(\text{MSE}(\Lambda_4))</td>
<td>.00657</td>
<td>.00560</td>
<td>.00845</td>
<td>.00734</td>
<td>.01516</td>
</tr>
</tbody>
</table>
FIGURE 5.1
Plot of MSE against Var (\( \Lambda \)): \((N, n) = (10, 10)\).

FIGURE 5.2
Plot of MSE against Var (\( \Lambda \)): \((N, n) = (15, 10)\).
FIGURE 5.3
Plot of MSE against Var (λ): (N, n) = (20, 10).

FIGURE 5.4
Plot of MSE against Var (λ): (N, n) = (10, 20).