Journal of the Korean
Statistical Society
Vol. 23, No. 1, 1994

On EM Algorithm For Discrete Classification With Bahadur Model: Unknown Prior Case

Hea-Jung Kim[1] and Hun-Jo Jung[2]

ABSTRACT

For discrimination with binary variables, reformulated full and first order Bahadur model with incomplete observations are presented. This allows prior probabilities associated with multiple populations to be estimated for the sample-based classification rule. The EM algorithm is adopted to provide the maximum likelihood estimates of the parameters of interest. Some experiences with the models are evaluated and discussed.

KEYWORDS: Multivariate binary populations, Reformulated Bahadur procedure, Classification rule, Monte Carlo study.

1. INTRODUCTION

It is sometimes of interest to an investigator to assess in some way the relative odds or probability that a multivariate observation $\mathbf{x}$ belongs to one of K multivariate binary populations Π_k, $k = 1, \ldots, K$. When the prior probability q_k

[1] Department of Statistics, Dongguk University, Seoul, 100-715, Korea
[2] Department of Computer Science and Statistics, Hanseo University, Chungnam, 352-820, Korea

that $\mathbf{x}$ belongs to Π_k is known where $\sum_{k=1}^{K} q_k = 1$, but the distribution of Π_k is unknown, several models for Π_k have been suggested for the discrimination of $\mathbf{x}$(cf. Dillon and Goldstein , 1978a). Among them Bahadur model(1961) and its lower order models by Moore(1973) gained popularity in the following senses: (i) These models, through the introduction of correlations and higher order terms, can be used to characterize any population distribution. (ii) The multiple classification rule, based upon the models, has been well developed. Solomon(1961), Moore(1973), and Dillon and Goldstein(1978b) evaluated their performances using the empirical cases and the Monte Carlo studies. However, when the q_k's are a priori unknown, their classification rule becomes intuitive(cf. Dillon and Goldstein, 1978a) in a sense that it incorporates an intuitive estimates of q_k's(the k-th training sample proportions).

There have been a couple of methods to estimate q_k's. Using the decision theory, Anderson(1984) suggested a trial-and-error method to find the estimates that minimize expected cost of misclassification. Its difficulty, especially for the multiple discrimination, makes an investigator use the intuitive estimates. The other is the predictive method by Geisser(1964) that cannot be applicable to the estimative approach.

Our concern in this paper is to suggest a way out of the problem consonant with the classification by the full and first order Bahadur models. In doing this, the Bahadur models are reformulated by introducing a set of unobservable indicator variables $\{T_k(\mathbf{x})\}$ which indicate a p-dimensional binary observation(or response pattern) $\mathbf{x}$ belongs to k-th population. Then, using distribution of the reformulated models, a method for estimating q_k's and a classification rule via EM algorithm are suggested.

2. BAHADUR PROCEDURE

The Bahadur(1961) representation for multinomial probabilities allows simple approximation to likelihood used in defining classification rule. Suppose that $\mathbf{X} = (X_1, \ldots, X_p)$, where X_j is a Bernoulli(or binary) random variable, $j = 1, \ldots, p$, such that $\mathbf{X}$ is a multinomial with the sample space consists of 2^p states or response patterns, $\{(x_1, \ldots, x_p)\}$, composed of a series of 0's and 1's. To distinguish the distribution under Π_k from that under $\Pi_{k'}$, $k = 1, \ldots, K$, further let

$$\theta_{kj} = P(X_j = 1 \mid \Pi_k), \quad Z_{kj} = (X_j - \theta_{kj})\{\theta_{kj}(1 - \theta_{kj})\}^{-1/2},$$

and the corresponding correlation terms

$$\gamma_k(j\ell) = E[Z_{kj}Z_{k\ell}], \ \ldots, \ \gamma_k(1,2,\ldots,p) = E[Z_{k1}Z_{k2}\cdots Z_{kp}].$$

Then for any response pattern $\mathbf{x} = (x_1,\ldots,x_p)$ generated from the k-th population, Bahadur(1961) has shown that the multinomial distribution of $\mathbf{X}$ can be reparameterized as

$$P_{[0]}(\mathbf{X}=\mathbf{x} \mid \Pi_k) = \prod_{j=1}^{p} \theta_{kj}^{x_j}(1-\theta_{kj})^{1-x_j}\{1 + \sum_{j<\ell}\gamma_k(j\ell)z_{kj}z_{k\ell} + \sum_{j<\ell<g}\gamma_k(j\ell g)z_{kj}z_{k\ell}z_{kg} + \ldots + \gamma_k(1,2,\ldots,p)z_{k1}z_{k2}\cdots z_{kp}\}, \quad (2.1)$$

where z_{kj} is the observed value of Z_{kj} corresponding to x_j. The appeal of this representation lies in its ability to describe the multinomial probabilities in terms of the means θ_{kj} and the correlation terms. The expression (2.1) is called the full Bahadur model.

In many applications it makes good sense to assume that higher-order correlations are zero, and hence reducing the parameters required for estimation. Moore considered the following lower order models. A first order approximation(First Order Bahadur Model) denoted by $P_{[1]}(\mathbf{X}=\mathbf{x} \mid \Pi_k)$ is obtained by omitting all correlation terms. This is equivalent to assuming independence among the binary variables. A second order approximation(Second Order Bahadur Model) denoted by $P_{[2]}(\mathbf{X}=\mathbf{x} \mid \Pi_k)$, results when the θ_{kj} and $\gamma_k(j\ell)$ terms are retained and all higher order terms are omitted.

In all three separate models utilized from the representation in (2.1), the classification procedures are formed by constructing the likelihoods and classifying $\mathbf{X} = \mathbf{x}$ into Π_{k^*} if

$$q_{k^*}P_{[u]}(\mathbf{X}=\mathbf{x} \mid \Pi_{k^*}) = \text{Max } \{q_kP_{[u]}(\mathbf{X}=\mathbf{x} \mid \Pi_k)\}, \quad (2.2)$$

where $k = 1,\ldots,K, \ u = 0,1,2.$

In most practical situations $P_{[u]}(\mathbf{X}=\mathbf{x} \mid \Pi_k)$ and q_k are unknown and hence the classification rule in (2.2) needs to be estimated from the training samples. Formal sample based estimate of $\hat{P}_{[u]}(\mathbf{X}=\mathbf{x} \mid \Pi_k)$ under the full or reduced Bahadur model is well developed(cf. Moore, 1973). However, that of q_k have not been seen yet. In practice we use an intuitive estimate of q_k to estimate

the classification rule. The intuitive estimate is the training sample proportion $N_k / \sum_{k'=1}^{K} N_{k'}$, where N_k is value of the k-th training sample size.

3. REFORMULATED BAHADUR PROCEDURE

On the basis of the simulation studies of the two-group case, Moore(1973) and Dillon and Goldstein(1978b) concluded that use of the second order Bahadur model for the binary variables classification is not recommended. Besides this, maximum likelihood estimators of the model are not a closed form and require increasingly burdensome computations to evaluate them as the number of the binary variables increases. Thus we will drop the second order Bahadur model in our study.

3.1. Reformulated Full and First Order Bahadur Representation

Let $\{T_k(\mathbf{x}) : k = 1, \ldots, K\}$ be a set of indicator variables for a binary response pattern $\mathbf{x} = (x_1, \ldots, x_p)$ generated from full or first order Bahadur model such that

$$T_k(\mathbf{x}) = \begin{cases} 1 & \text{if true population of } \mathbf{x} \text{ is } \Pi_k, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

When if r were the true population of $\mathbf{x}$, *i.e.* $T_r(\mathbf{x}) = 1$, the values of response of each binary variables actually obtained would be distributed according to

$$P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_r) \equiv P_{[u]}(\mathbf{X} = \mathbf{x} \mid T_r(\mathbf{x}) = 1), \quad u = 0, 1. \tag{3.2}$$

Thus unconditional joint distribution of $\mathbf{x}$ and $T_k(\mathbf{x})$'s would be

$$P_{[u]}(\mathbf{X} = \mathbf{x}, T_1(\mathbf{x}), \ldots, T_r(\mathbf{x}), \ldots, T_K(\mathbf{x})) = \prod_{k=1}^{K} q_k \{P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_k)\}^{T_k(\mathbf{x})}, \tag{3.3}$$

where $q_k = P(T_k(\mathbf{x}) = 1)$ is the prior probability of Π_k, $k = 1, \ldots, K$. It consists of a product of K terms, K-1 of which equal to 1(as $T_k(\mathbf{x}) = 0$, $k \neq r$) and one term of the form $q_r P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_r)$. We call (3.3) as the reformulated Bahadur representation.

At this point, let us consider the classification rule for the case when q_k's are known(unknown case will be discussed in the next part) but the true population of $\mathbf{x}$ is unknown such that $T_k(\mathbf{x})$'s in (3.3) are unobservable. Then Bayes' Theorem may be used to obtain the probability of $T_k(\mathbf{x}) = 1$. Under the representation (3.3), if $\mathbf{x}$ is given, by Bayes' Theorem

$$P_{[u]}(T_k(\mathbf{x}) = 1 \mid \mathbf{X} = \mathbf{x}) \propto q_k \, P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_k), \tag{3.4}$$

where all terms not involving k are absorbed into the proportionality sign. That is,

$$P_{[u]}(T_k(\mathbf{x}) = 1 \mid \mathbf{X} = \mathbf{x}) = q_k \, P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_k) \{\sum_{k'=1}^{K} q_{k'} \, P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_{k'})\}^{-1}. \tag{3.5}$$

Thus we obtain that, under (3.3), the classification rule is to classify $\mathbf{X} = \mathbf{x}$ into Π_{k^*} if

$$P_{[u]}(T_{k^*}(\mathbf{x}) = 1 \mid \mathbf{X} = \mathbf{x}) = \text{Max } P_{[u]}(T_k(\mathbf{x}) = 1 \mid \mathbf{X} = \mathbf{x}), \quad k = 1, \ldots, K. \tag{3.6}$$

It can be noticed that, from equations (3.4) and (3.5), when q_k's are priori known the classification rules defined in (2.2) and (3.6) are essentially the same.

3.2. Maximum Likelihood Estimation Using EM Algorithm

An advantage of the reformulated Bahadur representation (3.3) over the Bahadur representation (2.1) is that it contains q_k's as parameters so that, when q_k's are priori unknown, they can be estimated from the training samples.

Suppose that independent p-variate binary observations $\mathbf{x}_1(k), \ldots, \mathbf{x}_{N_k}(k)$ from k-th population Π_k (k-th training sample) with distribution $P_{[u]}(\mathbf{X} = \mathbf{x} \mid \Pi_k)$, $k = 1, \ldots, K$, $u = 0, 1$, are available. For notational convenience, further suppose that data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $N = \sum_{k=1}^{K} N_k$, to denote total observations of the K training samples so that, for example, $\mathbf{x}_1$ and $\mathbf{x}_N$ denote $\mathbf{x}_1(1)$ and $\mathbf{x}_{N_K}(K)$, respectively. Let $\{T_k(\mathbf{x}_i) : k = 1, \ldots, K\}$ be a set of unobservable indicator values for the i-th observation $\mathbf{x}_i$, $i = 1, \ldots, N$. Such

that if Π_r is the true population for the observation then its response pattern $\mathbf{x}_i$ would have

$$T_k(\mathbf{x}_i) = \begin{cases} 1 & \text{for } k = r, \\ 0 & \text{for } k \neq r. \end{cases}$$

As the data from all populations are assumed to be independent, the likelihood for the complete data is obtained from (3.3):

$$\prod_{k=1}^{K} \prod_{i=1}^{N} \{q_k P_{[u]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k)\}^{T_k(\mathbf{x}_i)}, \quad u = 0, 1, \tag{3.7}$$

where q_k denotes the prior probability that an observation $\mathbf{x}_i$ drawn from Π_k. The maximum likelihood estimates can be calculated analytically and we obtain the estimates as follows.

Proposition 1. Under the full Bahadur model, the likelihood (3.7) yields the maximum likelihood estimates

$$\begin{aligned} \hat{q}_k &= \sum_{i=1}^{N} T_k(\mathbf{x}_i) / \sum_{k=1}^{K} \sum_{i=1}^{N} T_k(\mathbf{x}_i), \quad k = 1, \ldots, K, \\ \hat{P}(\mathbf{X} = \mathbf{x}_s \mid \Pi_k) &= \sum_{i=1}^{N} T_k(\mathbf{x}_i) \cdot I_{\mathbf{x}_s}(\mathbf{x}_i) / \sum_{i=1}^{N} T_k(\mathbf{x}_i), \quad s = 1, \ldots, 2^p, \end{aligned} \tag{3.8}$$

where $\mathbf{x}_s$ denotes the s-th pattern among the 2^p p-variate binary response patterns, and

$$I_{\mathbf{x}_s}(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ has the same response pattern as } \mathbf{x}_s, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The full Bahadur model assumes that the sample space consists of 2^p different response patterns generated by the p-variate binary random vector $\mathbf{X} = (X_1, \ldots, X_p)$ having a multinomial distribution(cf. Dillon and Goldstein, 1978a). Therefore $P_{[0]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k)$ can be expressed in terms of response pattern probabilities:

$$P_{[0]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k) = \prod_{s=1}^{2^p} P(\mathbf{X} = \mathbf{x}_s \mid \Pi_k)^{I_{\mathbf{x}_s}(\mathbf{x}_i)},$$

where $\sum_{s=1}^{2^p} P(\mathbf{X} = \mathbf{x}_s \mid \Pi_k) = 1$.

Thus, under the full Bahadur model, the likelihood (3.7) becomes

$$\prod_{k=1}^{K}\prod_{i=1}^{N}\{q_k \prod_{s=1}^{2^p} P(\mathbf{X} = \mathbf{x}_s \mid \Pi_k)^{I_{\mathbf{x}_s}(\mathbf{x}_i)}\}^{T_k(\mathbf{x}_i)}. \tag{3.9}$$

Upon maximizing the likelihood under the conditions, $\sum_{k=1}^{K} q_k = 1$ and $\sum_{s=1}^{2^p} P(\mathbf{X} = \mathbf{x}_s \mid \Pi_k) = 1$, we get the results.

Proposition 2. Under the first Bahadur model, the likelihood (3.7) gives the maximum likelihood estimates

$$\begin{aligned} \hat{q}_k &= \sum_{i=1}^{N} T_k(\mathbf{x}_i) / \sum_{k=1}^{K}\sum_{i=1}^{N} T_k(\mathbf{x}_i), \quad k = 1, \ldots, K, \\ \hat{\theta}_{kj} &= \sum_{i=1}^{N} x_{ij} T_k(\mathbf{x}_i) / \sum_{i=1}^{N} T_k(\mathbf{x}_i), \quad j = 1, \ldots, p, \end{aligned} \tag{3.10}$$

where x_{ij} denotes the j-th component of the p-variate observation vector $\mathbf{x}_i$.

Proof. When we set $P_{[u]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k) = P_{[1]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k)$, the likelihood in (3.7) becomes

$$\prod_{k=1}^{K}\prod_{i=1}^{N}\{q_k \prod_{j=1}^{p} \theta_{kj}^{x_{ij}}(1-\theta_{kj})^{1-x_{ij}}\}^{T_k(\mathbf{x}_i)}. \tag{3.11}$$

Maximizing the likelihood with the condition, $\sum_{k=1}^{K} q_k = 1$, yields the result.

Under the situation of discrete discriminant analysis the true class of an individual i from the k-th training sample (i.e., the value of $T_k(\mathbf{x}_i)$) is unknown. This leads to a major purpose of the analysis that finds the true class of the individual with certain response pattern $\mathbf{x}_i$ by estimating the classification rule in (3.6). In our problem, the unobservable data $T_k(\mathbf{x}_i)$ preclude the straightforward maximum likelihood estimation of the parameters of interest and hence the classification rule in (3.6). However, if these parameters are known the unobservable data can be estimated by (3.5). Dempster *et al.*(1977) describes a numerical method of maximum likelihood estimation which is ideally suited

to this particular problem of estimating the classification rule. The method is known as the EM algorithm.

If, in the problem at hand, the unobservable indicator variables $\{T_k(\mathbf{x}_i) : i = 1,\ldots,N,\ k = 1,\ldots,K\}$ are treated as missing data then conditions of the EM algorithm are satisfied(see Proposition 3), and the algorithm for our estimation proceeds as follows:

(i) Take initial estimates of the missing data $T_k(\mathbf{x}_i)$'s.

(ii) Calculate the maximum likelihood estimates for the quantities of interest in (3.8) or (3.10) as if the missing data had been found.

(iii) Use the equation (3.5) and the estimates obtained from step (ii) to calculate new estimates of the missing data.

(iv) Repeat steps (ii) and (iii) until the results converge.

For E-step in (iii), the expectations of the missing data $T_k(\mathbf{x}_i)$'s can be obtained from (3.5). These would be expressed as

$$E[T_k(\mathbf{x}_i) \mid \mathbf{X} = \mathbf{x}_i] \ = \ P_{[u]}(T_k(\mathbf{x}_i) = 1 \mid \mathbf{X} = \mathbf{x}_i), \ \ u = 0, 1. \tag{3.12}$$

Thus the estimates of $T_k(\mathbf{x}_i)$'s are expressed as the probability that the true population for the observation with response pattern $\mathbf{x}_i$ is Π_k. Such probabilities may also be used as initial input in step (i). It is also noted, from (3.6) and (3.12), that final estimates of $T_k(\mathbf{x}_i)$'s in E-step directly estimate the classification rule (3.6). Hence we would classify $\mathbf{x}_i$ into Π_{k^*} if the final estimates gives the relation $T_{k^*}(\mathbf{x}_i) \ = \ Max\{T_k(\mathbf{x}_i)\}$.

Proposition 3. The above EM algorithm converges to a stationary point in the parameter space defined in (3.7).

Proof. We need to show that the algorithm satisfies two conditions regarding the convergence(cf. Dempster *et al.*(1977) and Wu(1983)) such that it is sufficient to show that the probability density of the form (3.7) comes from a regular exponential family and that the logarithm of (3.7), i.e. loglikelihood, is bounded. The latter is immediate from the definition of (3.7). The former can

be seen in the following way, following Sundberg(1974). As seen in (3.9) and (3.11), each $P_{[u]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k)$, $u = 0, 1$, is of the exponential type. Therefore, in accordance with Lehmann(1957), it can be expressed as

$$P_{[u]}(\mathbf{X} = \mathbf{x}_i \mid \Pi_k) \;=\; C_k(\delta_k)^{-1} exp\{\delta_k V_k(\mathbf{x}_i)\},$$

where δ_k denote a row parameter vector, column vector $V_k(\mathbf{x}_i)$ indicates a sufficient statistics of same dimension. $C_k(\delta_k)$ is a forming constant. Thus the density of complete data in (3.7) has the form

$$\begin{aligned} &\prod_{k=1}^{K}\prod_{i=1}^{N}\{q_k C_k(\delta_k)^{-1} exp\{\delta_k V_k(\mathbf{x}_i)\}\}^{T_k(\mathbf{x}_i)} \\ &= exp\{\sum_{k=1}^{K}\sum_{i=1}^{N}\{\delta_k T_k(\mathbf{x}_i) V_k(\mathbf{x}_i) \;+\; T_k(\mathbf{x}_i) log\frac{q_k}{C_k(\delta_k)}\}\} \quad (3.13) \end{aligned}$$

and as a consequence it is of regular exponential type for any $u = 0, 1$.

It remains to suggest initial values of the unobservable data $\{T_k(\mathbf{x}_i) : i = 1, \ldots, N_k,\ k = 1, \ldots, K\}$. One possibility is to assign $T_k(\mathbf{x}_i) = 1/K$. A second possibility is to use the data to calculate initial estimates. For example, one could use

$$\hat{T}_k(\mathbf{x}_i) \;=\; \sum_{i'=1}^{N_k} I_{\mathbf{x}_i}(\mathbf{x}_{i'}(k)) \;/\; \sum_{i'=1}^{N} I_{\mathbf{x}_i}(\mathbf{x}_{i'}) \qquad (3.14)$$

as starting values, where

$$I_{\mathbf{x}_i}(\mathbf{x}_{i'}) \;=\; \begin{cases} 1 & \text{if } \mathbf{x}_{i'} \text{ has the same response pattern as } \mathbf{x}_i, \\ 0 & \text{otherwise.} \end{cases}$$

and the same definition applies for the indicator variable $I_{\mathbf{x}_i}(\mathbf{x}_{i'}(k))$. Here $\{\mathbf{x}_{i'}(k) : i' = 1, \ldots, N_k\}$ denote the k-th training sample. Thus (3.14) means the ratio of number of observations with response pattern $\mathbf{x}_i$ in the k-th training sample over those in all K sets of the training samples.

In practice, it is advisable to repeat the algorithm for several different sets of starting values. The EM algorithm is only guaranteed to converge to a local maximum and in situations where relatively few data are used one must usually be content with choosing from a set of estimates corresponding to

different local maxima on the basis of the magnitude of the likelihood. In our experience the starting values defined by (3.14) have been particularly useful in locating the best local maximum, if not the global maximum of interest.

4. NUMERICAL EXAMPLES

In this section we give an empirical data example and simulation results to show the performance of the EM algorithm under the reformulated Bahadur model which is named as reformulated Bahadur procedure. As shown before, when q_k's are known, classification rule constructed by the reformulated model is equivalent to that under the Bahadur model. Therefore, our prime interest in the following examples is the behavior of the reformulated Bahadur procedure when q_k's are unknown. This will be done by comparing the actual error and the apparent error between the reformulated procedure and the Bahadur procedure which uses the intuitive estimates of q_k's for estimating the classification rule (2.2). For two group(or population) classification case, the actual error and the apparent error are respectively defined(cf. Dillon and Goldstein (1978a) and Moore(1973)) as

$$q_1 \sum_{s=1}^{2^p} B(\mathbf{x}_s;[u])P_u(\mathbf{X}=\mathbf{x}_s \mid \Pi_1) + q_2 \sum_{s=1}^{2^p}[1 - B(\mathbf{x}_s;[u])]P_u(\mathbf{X}=\mathbf{x}_s \mid \Pi_2), \tag{4.1}$$

$$\hat{q}_1 \sum_{s=1}^{2^p} B(\mathbf{x}_s;[u])\hat{P}_u(\mathbf{X}=\mathbf{x}_s \mid \Pi_1) + \hat{q}_2 \sum_{s=1}^{2^p}[1 - B(\mathbf{x}_s;[u])]\hat{P}_u(\mathbf{X}=\mathbf{x}_s \mid \Pi_2), \quad u=0,1, \tag{4.2}$$

where $B(\mathbf{x}_s;[u]) = 1$ if a response pattern $\mathbf{x}_s$ is classified into population 2 under each procedure constructed by the Bahadur model u. Otherwise, $B(\mathbf{x}_s;[u]) = 0$.

4.1. An Illustrated Example

The following data are "Attitude Toward Science" reported by Solomon(1961). A total of 2,982 response patterns on four binary variables were collected, of which 1,491 were identified as being from a low I.Q. group (Π_1), while the remaining 1,491 were from high I.Q. group(Π_2). The observed frequency distribution for the four binary variables with 16 response patterns(states) is

given in Table 1. Final estimates of $T_k(\mathbf{x}), k = 1, 2$, of the reformulated first order Bahadur procedure are also given in the table for illustrative purpose. Using the data, we consider two group discrete discriminant analysis with unknown prior probabilities q_1 and q_2. The question now is how well the reformulated Bahadur procedure performs. When we use the full and the first order Bahadur procedures which use the intuitive estimates (the sample proportion of each population) $\hat{q}_1 = \hat{q}_2 = 1/2$, the procedures yield the apparent error 0.4413 and 0.4423, respectively. While, applying EM algorithm to the data of Table 1 gives the following apparent error (a) 0.4411 for the reformulated full Bahadur procedure (b) 0.4390 for the reformulated first order Bahadur procedure. Although in this case relatively poor performance results no matter which procedure is selected, we do see that use of either of the reformulated procedures yields smaller apparent error than those obtained by the Bahadur procedures.

Table 1. Observed Frequency Distribution For Solomon Data And Final Estimates Of Indicator Variables

State	LOW IQ(Π_1)		HIGH IQ(Π_2)		Final Estimates	
$(x_1x_2x_3x_4)$	Observed	Relative	Observed	Relative	$T_1(\mathbf{x})$	$T_2(\mathbf{x})$
1 1 1 1	62	0.042	122	0.082	.2925	.7075
1 1 1 0	70	0.047	68	0.046	.3960	.6040
1 1 0 1	31	0.021	33	0.022	.3961	.6039
1 1 0 0	41	0.027	25	0.017	.5098	.4902
1 0 1 1	283	0.190	329	0.221	.4400	.5600
1 0 1 0	253	0.170	247	0.166	.5547	.4453
1 0 0 1	200	0.134	172	0.115	.5548	.4452
1 0 0 0	305	0.205	217	0.146	.6640	.3360
0 1 1 1	14	0.009	20	0.013	.2096	.7904
0 1 1 0	11	0.007	10	0.007	.2960	.7040
0 1 0 1	11	0.007	11	0.007	.2961	.7039
0 1 0 0	14	0.009	9	0.006	.4001	.5999
0 0 1 1	31	0.021	56	0.037	.3351	.6649
0 0 1 0	46	0.031	55	0.037	.4441	.5559
0 0 0 1	37	0.025	64	0.043	.4443	.5557
0 0 0 0	82	0.055	53	0.035	.5590	.4410
TOTAL	1491	1.000	1491	1.000		

4.2. A Simulation Study

The aim of this study is to show the reformulated Bahadur procedure gives good classification result when q_k's are unknown. Generating a Monte Carlo sample from the Full Bahadur model with 2^p state(or response pattern) probabilities gives rise to not only the problem of zero cell(state cell) observation for the small sample case but also massive computation for estimating classification results under the model. For these reasons the first order Bahadur populations with six binary variables were selected for Monte Carlo sampling. The population distributions of Π_1 and Π_2 for the 2^6 states of $\mathbf{x}$ were then determined from $P_{[1]}(\mathbf{X} = \mathbf{x} \mid \Pi_k)$ which have parameters $\theta_{k1}, \ldots, \theta_{k6}$, $k = 1, 2$, defined in Section 2. These population pairs are described in Table 2. The parameter values tabulated are chosen to reflect various types of difference between the two population distributions.

Table 2. Sample Population Pairs

Pair	Population $1(\Pi_1)$ $\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}$	Population $2(\Pi_2)$ $\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25}$
1	.2, .2, .2, .2, .2, .2	.4, .4, .4, .4, .4, .4
2	.2, .2, .2, .2, .2, .2	.6, .6, .6, .6, .6, .6
3	.2, .25, .3, .35, .4, .45	.4, .45, .5, .55, .6, .65
4	.2, .25, .3, .35, .4, .45	.6, .6, .6, .6, .6, .6

Once a set of population pair in Table 2 has been determined, the method for generating Monte Carlo samples and evaluating the classification results can be outlined as follows:

(i) Samples of N_1 and N_2 are taken from the set of probability distributions $\{P_{[1]}(\mathbf{X} = \mathbf{x} \mid \Pi_1),\ P_{[1]}(\mathbf{X} = \mathbf{x} \mid \Pi_2)\}$ by using SAS random number generator. These samples are used to determine the estimated probabilities $\hat{P}_{[1]}(\mathbf{X} = \mathbf{x}_s \mid \Pi_k), k = 1, 2; s = 1, \ldots, 2^6$ and the classification rules (2.2) and (3.6) for the first order Bahadur procedure and the reformulated first order Bahadur procedure, respectively.

(ii) As one criterion for evaluating the two procedures, the actual errors using the prior probability $q_1 = 1/2,\ 1/3$ are determined from (4.1) for each

procedure. For the second criterion, the correlation between the true log likelihood ratios $P_{[1]}(\mathbf{X} = \mathbf{x}_s \mid \Pi_2)/P_{[1]}(\mathbf{X} = \mathbf{x}_s \mid \Pi_1), s = 1, \ldots 2^6$, and the estimated values of them under the reformulated first order Bahadur model is used for evaluation since it provides a measure of overall performance of the EM algorithm.

(iii) Steps 1 and 2 are repeated 50 times for each set of $\{N_1, N_2\}$. Then the mean and the standard deviation of resulting 50 actual errors and the mean correlation for the 50 trials are tabulated in Table 3.

Table 3. Mean Actual Error and Correlation Based On 50 Monte Carlo Samples

Pair	q_1	Optimum Error	Sample Size N_1	Sample Size N_2	Actual Error Bahadur	Actual Error Reformulated	Correlation
1	1/2	.28896	50	50	.30958(.01350)	.30910(.01545)	.90220
			100	100	.29980(.00629)	.30203*(.09228)	.95385
			1000	1000	.28896(.00000)	.28896(.00000)	.99610
	1/3	.27040	50	100	.29046(.01593)	.28895(.01635)	.83315
			100	200	.27784(.01455)	.27533(.00608)	.90268
			500	1000	.27326(.00312)	.27124(.00518)	.98697
2	1/2	.13904	50	50	.14987(.00797)	.14859(.11211)	.96645
			100	100	.14458(.00534)	.14358(.00515)	.98290
			1000	1000	.13904(.00000)	.13904(.00000)	.99817
	1/3	.14219	50	100	.14725(.00894)	.14906*(.01317)	.92397
			100	200	.14629(.00146)	.14509(.00317)	.96982
			500	1000	.14443(.00131)	.14381(.00206)	.99572
3	1/2	.29782	50	50	.32147(.01288)	.32137(.13732)	.93662
			100	100	.31085(.00713)	.30825(.01472)	.96610
			1000	1000	.29782(.00000)	.29782(.00000)	.99737
	1/3	.26804	50	100	.29287(.01370)	.29313*(.01623)	.78972
			100	200	.27892(.00773)	.27812(.00716)	.91877
			500	1000	.26945(.00223)	.26832(.00145)	.99077
4	1/2	.22633	50	50	.24053(.00931)	.24336*(.01026)	.90409
			100	100	.23494(.00734)	.23378(.00930)	.94928
			1000	1000	.22633(.00000)	.22633(.00000)	.99426
	1/3	.20732	50	100	.23401(.02237)	.23328(.02314)	.88548
			100	200	.21365(.00528)	.21331(.00571)	.95761
			500	1000	.20813(.00072)	.20811(.00114)	.99584

Note 1: We use "Bahadur" to denote the Bahadur first order procedure and "Reformulated" to denote the reformulated Bahadur first order procedure. Note 2: Value in the parenthesis is the standard deviation of 50 actual errors. Note 3: The optimum error can be obtained by substituting true value of $\mathbf{B}(\mathbf{x}_s; [u])$ based upon the true likelihood in the expression (3.15).

A couple of points are indicated from Table 3. Smaller values of the mean actual errors for the reformulated procedure(except for the values with "*" notation) show that it performs better than the Bahadur procedure. This shows that the former is more favorable to the binary variables discrimination with unknown prior probabilities. Moreover, relatively high mean correlations indicate that the suggested EM algorithm works well for estimating true parameters in the likelihood (3.7).

5. CONCLUDING REMARKS

In this paper a sample based discrete discrimination procedure which takes care of unknown prior probabilities, q_k's, has been suggested. It is shown that the suggested procedure is different from the usual Bahadur procedure in that it enables us to get maximum likelihood estimates of q_k's via EM algorithm while the latter uses intuitive estimates. This was done by introducing the reformulated likelihood (3.7). Limited examples of Section 4 advocate the suggested procedure and indicate favorable performance of the EM algorithm. A sampling experiment pertaining to the multiple classification with K groups has not been done here. However, as seen in (2.2) and (3.6), the K-group classification rule can be viewed as a multiple comparison of ${}_K\mathrm{C}_2$ two-group classification results. Thus the examples in Section 4 would suffice the purpose of studying the performance of the suggested procedure.

As noted, the suggested procedure needs more calculation than the Bahadur procedure. However, we may give more credit to it in that it gives a way of overcoming the problem of estimating q_k's. Applications of the suggested procedure to the discrete discrimination models are appreciable. A model to which the procedure may be immediately applicable is the polytomous variables model.

ACKNOWLEDGEMENT

The detailed and constructive comments of referees were very helpful in making the paper more readable and correcting some errors. Their efforts are greatly appreciated.

REFERENCES

(1) Anderson, T.W.(1984). *An Introduction to Multivariate Analysis,* John Wiley & Sons, New York.

(2) Bahadur, R.R.(1961). A representation of the joint distribution of response to n dichotomous items, in *Studies in Item Analysis and Prediction,* H. Solomon Ed., Palo Alto, Calif.: Standford Univ. Press, 158-168.

(3) Dempster, A.P., Laird, N.M., and Rubin, D.B.(1977). Maximum likelihood estimation from incomplete data via the EM algorithm(with discussion), *Journal of Royal Statistical Society,* B39, 1-38.

(4) Dillon, W.R. and Goldstein, M.(1978a). *Discrete Discriminant Analysis,* John Wiley & Sons, New York.

(5) Dillon, W.R. and Goldstein, M.(1978b). On the performance of some multinomial classification rules, *Journal of the American Statistical Association*, Vol. 73, 305-313.

(6) Geisser, S.(1964). Posterior odds for multivariate normal classifications, *Journal of Royal Statistical Society,* B, Vol. 26, 69- 76.

(7) Lehmann, E.L. (1959). *Testing Statistical Hypothesis,* John Wiley & Sons, New York.

(8) Moore, D.H., II(1973). Evaluation of five discriminant procedures for binary variables, *Journal of the American Statistical Association,* Vol. 68, 399-404.

(9) Solomon, H.(Ed.)(1961). *Studies in Item Analysis and Prediction,,* Palo Alto, Calif.: Standford Univ. Press.

(10) Sundberg, R. (1974). Maximum likelihood theory of incomplete data from an exponential family, *Scandinavian Journal of Statistics*, Vol.1, 49-58.

(11) Wu, C.F.J. (1983). On convergence properties of the EM algorithm, *Annals of Statistics*, Vol. 11, 95-103.