Pattern Recognition, Vol. 23, No. 5, pp. 529–537, 1990

EFFICIENCY OF DISCRIMINANT ANALYSIS WHEN INITIAL SAMPLES ARE CLASSIFIED STOCHASTICALLY

T. KRISHNAN and S. C. NANDY
Indian Statistical Institute, 203, Barrackpore Trunk Road, Calcutta 700035, India

(*Received* 26 *November* 1986; *accepted in revised form* 12 *July* 1989)

Abstract—We consider the problem of discriminant analysis of two multivariate normal populations having a common dispersion matrix, where the initial samples are classified stochastically. We assume a beta model for this classification variable and assume it to be independent of the feature vector X, given the group. We study the Efron efficiency of this procedure compared to the situation where the initial classification is done deterministically and correctly. We present tables and charts of this efficiency and conclude that stochastic supervision contains a great deal of information on the discriminant function.

Discriminant analysis Stochastically classified initial samples Asymptotic relative efficiency

1. INTRODUCTION

Discriminant analysis is traditionally performed assuming that the classification of initial samples is done deterministically and correctly. Recently, some applications in remote sensing and in medical diagnosis have led to interest in considering discriminant analysis where the initial classification is prone to error.[1] Aitchison and Begg[2] identify the need for statistical diagnostic techniques based on data sets containing cases which have not been allocated to a single diagnostic type with certainty but for which only an assessment of the probabilities of the types is available. They give an example from medical diagnosis of Conn's syndrome. They discuss some methods of discriminant analysis based on the logistic transform.

In this article, we consider initial samples of this type, for the case of a feature vector X having p-dimensional normal distributions $\mathcal{N}_p(\mu_0, \Sigma)$ and $\mathcal{N}_p(\mu_1, \Sigma)$ in two groups, occurring in proportions π_0 and π_1 respectively; we denote by Δ, the Mahalanobis distance between the two groups. We denote by $Z(0 \leq Z \leq 1)$, the variable indicating the supervisor's assessment of the chance of an unit coming from Group 1 (and $(1 - Z)$ from Group 0). We denote by y the actual group.

In a series of articles in this journal.[3–5] we have investigated the problem of imperfect initial samples. In Katre and Krishnan,[3] we considered the problem where the initial samples are classified deterministically and are subject to a constant and unknown probability of misclassification; this misclassification was assumed to occur independently of the feature vector X; we derived here the maximum likelihood estimators of parameters and gave various procedures for computing them. In Krishnan,[4] we studied the efficiency of this error-prone supervision scheme compared to a perfectly supervised scheme; this efficiency *à la* Efron[5] called the *Asymptotic Relative Efficiency* (Eff) is a measure of the amount of information contained in the error-prone initial samples relative to perfectly supervised initial samples; this efficiency can also be interpreted in terms of the relative sample sizes required in the two schemes to achieve the same expected error rates of the classification scheme using the estimates of the discriminant function derived from these parameter estimates. We[4] presented tables of this Eff for various values of the parameters and interpreted them; our calculations gave an idea of the worth of error-prone initial samples for various parameter values. In Krishnan and Nandy[5] we turned to stochastically supervised initial samples and used the model described here; we derived the EM algorithm of Dempster, Laird and Rubin[6] for maximum likelihood estimation of parameters. In the present article, we work out the Eff of the stochastic supervision scheme compared to a deterministically and correctly supervised scheme to answer questions on the relative information contained in stochastically supervised initial samples and the relative sample size required under stochastic supervision.

Stochastic supervision model

We consider a model for stochastic supervision in which Z is distributed as the beta distribution with parameters m and n (denoted $\mathcal{B}(m, n)$) and independent of X when $y = 0$ and as $\mathcal{B}(n, m)$ and independent of X when $y = 1$. Various choices of m, n give a whole range of cases from the completely unsupervised case (when $m = n$) to the (perfectly) supervised case ($|m - n| \to \infty$) as seen from the cumulative probability curves of Fig. 1. For $m = n$, the distribution is the same for $y = 0$ and $y = 1$ and hence it is the unsupervised case; it does not matter what the common value of m, n is. We show that our efficiency formula when $m = n$ is the same as that obtained in the unsupervised case. For $m \neq n$, the supervisor assessment is probabil-

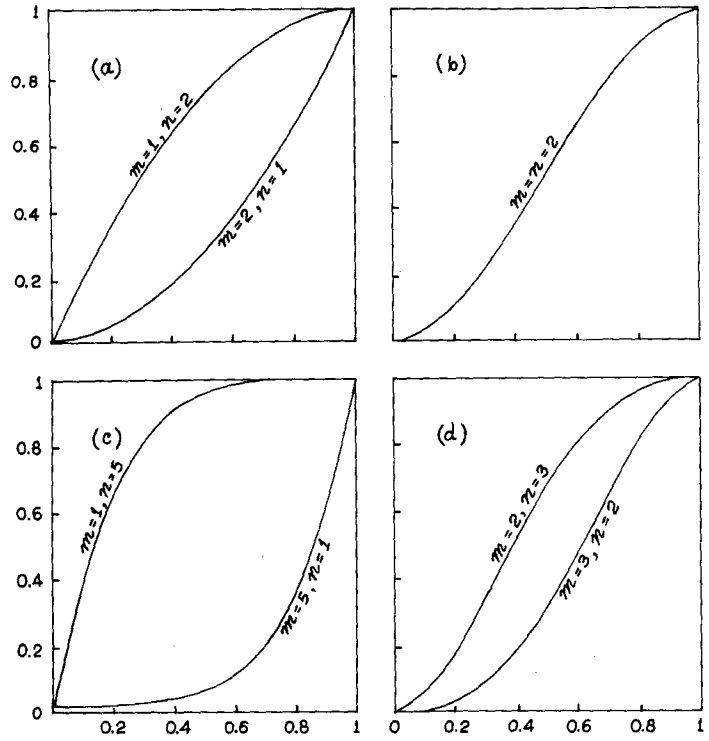

Fig. 1. Cumulative distribution curves of beta distribution for various values of m and n.

istically more on the correct side; the larger $|m-n|$ is, the more correct it is, approaching perfect supervision as $|m-n| \to \infty$. From the nature of the curves it appears that the wider apart the two curves are the better is the supervision. The correctness of the supervision depends not merely on $|m-n|$; for the same value of $|m-n|$, lower values of m, n seem to indicate better supervision. Thus some kind of normalised values of $|m-n|$ may be a suitable indicator of the level of supervision. We discuss this further in Section 4. Thus the beta model may be a reasonable way to describe stochastic supervision. The assumption of $\mathscr{B}(m,n)$ and $\mathscr{B}(n,m)$ models for Z makes the stochastic supervision have a symmetric structure with respect to the groups 0 and 1; it simplifies the mathematics considerably. Although the assumptions of symmetry or that of Z and X being independent given y, may not be completely realistic, it is a useful first model to begin investigation of this phenomenon. A reasonable formulation in which X and Z are dependent is to make the probability of misclassification larger when X is close to the means of both the groups. From the evidence available from the studies of Chhikara and McKeon[1] and Lachenbruch,[7] the efficiency of such a scheme is higher than when X and Z are independent. Thus it is seen that such independent (random) misallocation is the least favourable situation for efficiency and hence is worth studying. Recently Titterington[8] has proposed an alternative to our supervision model using the logistic-normal distribution and worked out the EM algorithm for estimation of parameters under his model. We propose to study efficiency under this model.

The Eff is the ratio of the *Asymptotic Error Rates* (AER) of the correctly supervised and stochastic supervised schemes. The AER of a scheme is a function of the elements of the variance-covariance matrix of the estimators of the linear discriminant coefficients. This AER of a perfectly supervised scheme depends on π_1 and Δ and the AER of a stochastically supervised scheme and its Eff depends on π_1, Δ and the parameters of the stochastic supervision model. Thus in our model, Eff depends on π_1, Δ, m and n. Efron[9] has derived the AER of a perfectly supervised scheme. Thus it only remains for us to derive the AER of the stochastically supervised scheme. Thus following Efron[9] we make a linear transformation on X to reparametrise the model in terms of π_1, Δ, m, n and other parameters of the model. The variance-covariance matrix of the discriminant coefficients is what is required for computing the AER. This AER is a function of the elements of the variance-covariance matrix of estimators of the linear discriminant coefficients. This variance-covariance matrix can be obtained from the likelihood function as follows: obtain the information matrix as the expected value of the negative of the matrix of second mixed derivatives of the loglikelihood with respect to the parameters: invert this matrix to get the required vari-

ance-covariance matrix. Denoting loglikelihood by L, we have

$$L(x,z,y) = L(x,z) + L(y/x,z). \tag{1.1}$$

These loglikelihoods (Ls) respectively correspond to the perfectly supervised scheme, the stochastically supervised scheme and the logistic regression based on X and Z. The information matrices being minus of expected values of second derivatives of these Ls satisfy a similar additive condition. Thus the information matrix of the stochastic supervision scheme can be worked out from those of the perfectly supervised and logistic regression schemes. We proceed to derive this here. Note that we are using logistic regression estimators only as a technique for computing the required information matrix, and logistic regression as such is not our concern in this article. An application of this technique was made by O'Neill[10] to study the efficiency of an unsupervised initial sample *vis-à-vis* a perfectly supervised initial sample.

2. ASYMPTOTIC RELATIVE EFFICIENCY

For the case of p-dimensional normal populations

$$\mathcal{N}_p(\mu_0,\Sigma) \quad \text{and} \quad \mathcal{N}_p(\mu_1,\Sigma)$$

in two groups, occurring in proportions π_0 and π_1 respectively, the Bayes rule uses

$$\beta_0 + \beta' x \tag{2.1}$$

where

$$\beta_0 = \log(\pi_1/\pi_0) - \frac{1}{2}(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0);$$
$$\beta = \Sigma^{-1}(\mu_1 - \mu_0) \tag{2.2}$$

as the discriminant function. The Bayes rule is the one with the least error rate. The Asymptotic Error Rate (AER) of a procedure based on estimates $(a_0,a')_N$ of vector (β_0,β') from a sample of size N is defined to be the limiting value (as $N \to \infty$) of the additional error of (a_0,a') over the Bayes error. This AER will, naturally, depend on the nature of the learning procedure and the values of the parameters. For perfectly supervised, unsupervised and stochastically supervised procedures it will be different and for the same parameter values, the unsupervised procedure will have a larger AER than the stochastically supervised procedure, and the stochastically supervised procedure will have a larger AER than the supervised procedure. When several procedures less efficient than the supervised one are considered, the supervised procedure may be used as the basis of the comparison. This leads to Efron's Asymptotic Relative Efficiency (Eff).

Since error rates of discriminant rules based on β_0,β or their estimates are invariant under linear transformations on the feature vector X, we assume a canonical form for (μ_0,Σ) and (μ_1,Σ) to be $\left(-\frac{\Delta}{2}e_1, I_p\right)$ and $\left(\frac{\Delta}{2}e_1, I_p\right)$ where Δ is the Mahalanobis distance between the two groups, e_1 is the vector $(1, 0, \ldots, 0)$ and I_p is the $p \times p$ identity matrix; this canonical form can be obtained by a linear transformation on X. Let $(a_0,a)_N$ denote the estimate of (β_0,β) based on a sample of N by a certain procedure and let $\mathrm{ER}(a_0,a)_N$ denote the error rate on using $(a_0,a)_N$ for (β_0,β) in (2.1). Let $\lambda = \log(\pi_1/\pi_0)$. Then $\beta_0 = \lambda$, $\beta' = \Delta e_1$.

Efron[9] shows that if

$$\sqrt{N}[(a_0,a)_N - (\beta_0,\beta)] \xrightarrow{L} \mathcal{N}_{p+1}(0,M) \tag{2.3}$$

then

$$N[\mathrm{ER}(a_0,a)_N - \mathrm{ER}(\beta_0,\beta)] \xrightarrow{L} \frac{\pi_1}{2\Delta}\phi\left(\frac{\Delta}{2} - \frac{\lambda}{\Delta}\right)$$
$$\left[r_0^2 - \left(\frac{2\lambda}{\Delta}\right)r_0r_1 + \left(\frac{\lambda}{\Delta}\right)^2 r_1^2 + r_2^2 + \cdots + r_p^2\right] \tag{2.4}$$

where $\xrightarrow{L}$ means convergence in law (distribution), $r = (r_0,r_1,r_2,\ldots,r_p) \sim \mathcal{N}_{p+1}(0,S)$, ϕ is standard normal density function, and 0 the $(p+1)$-null vector. The AER of a procedure with estimates $(a_0,a)_N$ is then defined to be the expectation of the limit above, which is equal to

$$\frac{\pi_1}{2\Delta}\phi\left(\frac{\Delta}{2} - \frac{\lambda}{\Delta}\right)$$
$$\left[s_{00} - \left(\frac{2\lambda}{\Delta}\right)s_{01} + \left(\frac{\lambda}{\Delta}\right)^2 s_{11} + s_{22} + \cdots + s_{pp}\right] \tag{2.5}$$

where $((s_{ij})) = S$. This is denoted for convenience by $AER(a_0,a)$. Then the Asymptotic Relative Efficiency Eff of a procedure with $(c_0,c)_N$ with respect to a procedure yielding estimate $(b_0,b)_N$ is

$$\mathrm{Eff}_p = \mathrm{AER}(b_0,b)/\mathrm{AER}(c_0,c). \tag{2.6}$$

In order to compute this efficiency for stochastically classified initial samples relative to a perfectly supervised sample we need the matrices S for these cases for the maximum likelihood estimates. This is done by computing the information matrix of β_0,β and inverting it; Efron has already computed this for the supervised case as

$$I_c = \pi_0\pi_1\begin{bmatrix} H & 0 \\ 0 & (1+\Delta^2\pi_0\pi_1)^{-1}I_{p-1} \end{bmatrix} \tag{2.7}$$

where

$$H^{-1} = \begin{bmatrix} 1+\Delta^2/4 & -(\pi_0-\pi_1)\Delta/2 \\ -(\pi_0-\pi_1)\Delta/2 & 1+2\pi_0\pi_1\Delta^2 \end{bmatrix}.$$

3. EFFICIENCY OF STOCHASTIC SUPERVISION SCHEME

We have observations (x_j, z_j), $j = 1, 2, \ldots, N$. Let $f_i(x)$ represent the density of $\mathcal{N}_p(\mu_i, \Sigma)$, $i = 0, 1$. Then

$$f_0(x,z) = f_0(x)\frac{1}{\text{Beta}(m,n)}z^{m-1}(1-z)^{n-1}$$

$$f_1(x,z) = f_1(x)\frac{1}{\text{Beta}(n,m)}z^{n-1}(1-z)^{m-1} \quad (3.1)$$

where Beta stands for the complete beta integral, give the density of observation (x, z) in the two groups. For what follows, we need full, various marginal and conditional likelihoods. We use L to denote loglikelihood, whose arguments and the conditioning symbol '/' indicate which loglikelihood is being considered.

$$L(x,z,y) = \log\{[\pi_1 f_1(x,z)]^y[\pi_0 f_0(x,z)]^{1-y}\} \quad (3.2)$$

$$L(x,y) = \log\{[\pi_1 f_1(x)]^y[\pi_0 f_0(x)]^{1-y}\} \quad (3.3)$$

$$L(x,z) = \log\{[\pi_1 f_1(x,z) + \pi_0 f_0(x,z)\} \quad (3.4)$$

$$L(x) = \log\{\pi_1 f_1(x) + \pi_0 f_0(x)\} \quad (3.5)$$

$$L(y/x,z) = \log\{[\pi_1(x,z)]^y[\pi_0(x,z)]^{1-y}\}, \quad (3.6)$$

where

$$\begin{aligned}\pi_0(x,z) &= 1 - \pi_1(x,z)\\ &= \frac{\pi_0 f_0(x,z)}{\pi_0 f_0(x,z) + \pi_1 f_1(x,z)}\\ &= \frac{1}{1 + e^{\beta_0 + \beta' x + (n-m)w}},\end{aligned} \quad (3.7)$$

where $w = \log[z/(1-z)]$ and

$$L(y/x) = \log\{[\pi_1(x)]^y[\pi_0(x)]^{1-y}\}, \quad (3.8)$$

where

$$\begin{aligned}\pi_0(x) &= 1 - \pi_1(x)\\ &= \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}\\ &= \frac{1}{1 + e^{\beta_0 + \beta' x}}.\end{aligned} \quad (3.9)$$

In what follows, we need to deal with information matrices arising out of three types of situations—when the actual group y is known (called conditional), when y is not known (called unconditional) and of the logistic regression type; the observations are correspondingly of the type (x, y), (x) and (y/x). We denote the information matrices based on a single observation for the parameters β_0, β of these three types by I with subscripts C, UC and LR respectively. The relation between I_{UC}, I_C and I_{LR} is obtained by using

$$L(x,y) = L(x) + L(y/x). \quad (3.10)$$

In what follows, we shall also need to consider the case where the stochastic supervision observation z is also available; then, we have again three types of observation (x, z, y), (x, z) and $(y/x, z)$. The information matrices in this case are denoted with an asterisk; thus we have I^*_{UC}, I^*_C and I^*_{LR}. Note that:

(1) we are interested in the parameters β_0, β and the information matrices for them;

(2) the estimates of β_0, β remain the same whether $L(x, y)$ or $L(x, z, y)$ is maximised, since evidently, given y, information of z is redundant as per our model;

(3) I^*_{UC} corresponds to the case of stochastically classified initial samples.

Let us now reparametrise $\mu_0, \mu_1, \Sigma, m, n, \pi_1$ as

$$Q = \pi_1\mu_1 + \pi_0\mu_0$$

$$R = \Sigma + \pi_0\pi_1(\mu_1 - \mu_0)(\mu_1 - \mu_0)'$$

β_0, β, $u = n + m$, and $v = n - m$.

This reparametrisation of m, n into u, v is chosen because the logistic regression involves only v.

Let A, B, C denote the information matrices of $(Q, R, \beta_0, \beta, u, v)$ based on (3.2), (3.4), (3.6) respectively; I^*_C, I^*_{UC} and I^*_{LR} are parts of these matrices respectively corresponding to the parameters β_0, β only.

Let us partition A corresponding to (Q, R), (β_0, β), (u, v) as

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \quad (3.11)$$

and similarly B and C also.

It can be easily checked that because of the independenc of X and Z given y, assumed in our model $A_{13}, A_{23}, A_{31}, A_{32}$ are all 0 matrices. Thus

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix}. \quad (3.12)$$

Further, since $L(y/x, z)$ does not involve (Q, R), the partitioned matrix C is as follows:

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & C_{22} & C_{23} \\ 0 & C_{32} & C_{33} \end{bmatrix}. \quad (3.13)$$

Further, because of (1.1), $A = B + C$. Hence

$$B = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} - C_{22} & -C_{23} \\ 0 & -C_{32} & A_{33} - C_{33} \end{bmatrix}. \quad (3.14)$$

Now, $L(x, z, y)$ breaks up into two factors, one involving m, n and z only and another not involving m, n and z. Thus on the basis of the reparametrisation we can break up $L(x, z, y)$ into two factors, one involving u, v and z only and the other not involving these. We can obtain the information matrix A_{33} using observation z only because of the structure of A. Then

$$A_{33} = \frac{1}{4}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} M \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \tag{3.15}$$

where the information matrix M of m, n is easily seen to be

$$M = \begin{bmatrix} \frac{a_{20}}{a_{00}} - \frac{a_{10}^2}{a_{00}^2} & \frac{a_{11}}{a_{00}} - \frac{a_{01}a_{10}}{a_{00}^2} \\ \frac{a_{11}}{a_{00}} - \frac{a_{01}a_{10}}{a_{00}^2} & \frac{a_{02}}{a_{00}} - \frac{a_{01}^2}{a_{00}^2} \end{bmatrix} \tag{3.16}$$

where

$$a_{ij} = \int_0^1 (\log z)^i (\log(1-z))^j z^{m-1}(1-z)^{n-1} dz$$

$$\forall i,j = 0, 1, 2. \tag{3.17}$$

Let A^{-1} be partitioned similar to A with superscripts, that is, with components A^{11}, etc. Then

$$\begin{aligned} I_C^* = I_C &= (A^{22})^{-1} \\ &= A_{22} - A_{21}A_{11}^{-1}A_{12} \\ &= B_{22} + C_{22} - B_{21}B_{11}^{-1}B_{12}. \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} I_{UC}^* &= (B^{22})^{-1} \\ &= B_{22} - (B_{21} \quad B_{23}) \begin{bmatrix} B_{11} & B_{13} \\ B_{31} & B_{33} \end{bmatrix}^{-1} \begin{pmatrix} B_{12} \\ B_{32} \end{pmatrix} \\ &= B_{22} - (B_{21} \quad C_{23}) \begin{bmatrix} B_{11}^{-1} & 0 \\ 0 & B_{33}^{-1} \end{bmatrix} \begin{pmatrix} B_{12} \\ C_{32} \end{pmatrix} \\ &= B_{22} - B_{21}B_{11}^{-1}B_{12} - C_{23}B_{33}^{-1}C_{32} \\ &= B_{22} - B_{21}B_{11}^{-1}B_{12} - C_{23}[A_{33} - C_{33}]^{-1}C_{32}. \end{aligned} \tag{3.19}$$

Hence

$$I_C^* = I_{UC}^* + C_{22} + C_{23}[A_{33} - C_{33}]^{-1}C_{32}. \tag{3.20}$$

Note that under the linear transformation we have used, $\beta_0 = \lambda$ and $\beta' = \Delta e_1$. The matrix $I_C = I_C^*$ was computed by Efron as (2.7).

Now it remains to compute C. For this we follow the technique of Efron for his Lemma 3, which makes essential use of the exponential family form of (3.6) and (3.7). The conditional loglikelihood given $\{(x_j, z_j)\}$ is

$$\begin{aligned} &\sum_j [y_j \log \pi_1(x_j, z_j) + (1 - y_j)\log \pi_0(x_j, z_j)] \\ &= \sum_j y_j \log \frac{\pi_1(x_j, z_j)}{\pi_0(x_j, z_j)} + \sum_j \log \pi_0(x_j, z_j) \\ &= \sum_j y_j \log \exp[\beta_0 + \beta' x_j + (n-m)w_j] \\ &+ \sum_j \log \frac{1}{1 + \exp[\beta_0 + \beta' x_j + (n-m)w_j]} \\ &= \sum_j y_i(\beta_0 + \beta' x_j + (n-m)w_j) \\ &- \sum_j \log[1 + \exp(\beta_0 + \beta' x_j + (n-m)w_j)] \\ &= (\beta_0, \beta', v)T - \psi(\beta_0, \beta', v) \end{aligned} \tag{3.21}$$

where

$$T = \sum_{j=1}^n \begin{pmatrix} 1 \\ x_j \\ w_j \end{pmatrix} y_j$$

and

$$\psi(\beta_0, \beta', v) = \sum_j \log[1 + \exp(\beta_0 + \beta' x_j + (n-m)w_j)].$$

Now

$$\begin{aligned} C &= \lim_{N\to\infty} \frac{1}{N} Cov_{\beta_0, \beta', v} \\ &= \int_0^1 \int_{R_p} \begin{pmatrix} 1 \\ x \\ w \end{pmatrix} (1, x, w)\pi_1(x, z)\pi_0(x, z) dF(x, z) \end{aligned} \tag{3.22}$$

where

$$dF(x,z) = \left(\pi_0 f_0(x) \frac{z^{m-1}(1-z)^{n-1}}{\text{Beta}(m,n)} + \pi_1 f_1(x) \frac{z^{n-1}(1-z)^{m-1}}{\text{Beta}(n,m)} \right) dx dz$$

since $f(x, z)$ is a mixture of $f_0(x, z)$ and $f_1(x, z)$ in proportions of π_0 and π_1.

As in Efron,

$$\begin{aligned} &\int_{R^n} \pi_1(x)\pi_0(x)(\pi_0 f_0(x) + \pi_1 f_1(x)) dx \\ &= \pi_0\pi_1 \int_{-\infty}^{\infty} \frac{\exp[-\Delta^2/8]\phi(x)}{\pi_1 \exp[\Delta x/2] + \pi_0 \exp[-\Delta x/2]} dx. \end{aligned} \tag{3.23}$$

Thus

$$\begin{aligned} E(x^i, w^j) &= \frac{\pi_0\pi_1 \exp[-\Delta^2/8]}{\sqrt{2\pi}\text{Beta}(m,n)} \int_0^1 w^j z^{m+n-2}(1-z)^{m+n-2} \\ &\quad \int_{-\infty}^{\infty} \frac{x^i \exp[-x^2/2]}{\pi_1 z^{n-1}(1-z)^{m-1}\exp[\Delta x/2] + \pi_0 z^{m-1}(1-z)^{n-1}\exp[-\Delta x/2]} dx dz \\ &= \frac{\pi_0\pi_1 \exp[-\Delta^2/8]}{\sqrt{2\pi}\text{Beta}(m,n)} \int_0^1 \left[\log\left(\frac{z}{1-z}\right) \right]^j z^{m+n-2}(1-z)^{m+n-2} \\ &\quad \int_{-\infty}^{\infty} \frac{x^i \exp[-x^2/2]}{\pi_1 z^{n-1}(1-z)^{m-1}\exp[\Delta x/2] + \pi_0 z^{m-1}(1-z)^{n-1}\exp[-\Delta x/2]} dx dz \\ &= \pi_0\pi_1 D_{ij} \end{aligned} \tag{3.24}$$

thus defining notation D_{ij}.

Now

$$\begin{bmatrix} C_{22} & C_{23} \\ C_{32} & C_{33} \end{bmatrix} = \pi_0\pi_1$$

$$\left[\begin{array}{cccc|cc} D_{00} & D_{10} & 0\cdots & 0 & 0 & D_{01} \\ D_{10} & D_{20} & 0\cdots & 0 & 0 & D_{11} \\ 0 & 0 & D_{00}\cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0\cdots & D_{00} & 0 & 0 \\ - & - & - & - & - & - \\ 0 & 0 & 0\cdots & 0 & 0 & 0 \\ D_{01} & D_{11} & 0\cdots & 0 & 0 & D_{02} \end{array}\right]. \quad (3.25)$$

Now the matrix $C_{22} + C_{23}[A_{33} - C_{33}]^{-1}C_{32}$ required in formula (3.20) becomes

$$F = C_{22} + C_{23}[A_{33} - C_{33}]^{-1}C_{32} \quad (3.26)$$

where

$$[A_{33} - C_{33}]^{-1} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} - D_{02} \end{bmatrix}^{-1}$$

$$= \frac{1}{M_{11}(M_{22} - D_{02}) - M_{12}^2}\begin{bmatrix} M_{22} - D_{02} & -M_{21} \\ -M_{12} & M_{11} \end{bmatrix}$$

where the partitioned matrix M is defined in (3.16). Let us take

$$d = M_{11}/[(M_{22} - D_{02})M_{11} - M_{12}^2]. \quad (3.27)$$

Then

$$F = \begin{bmatrix} D_{00} + dD_{01}^2 & D_{10} + dD_{01}D_{11} \\ D_{10} + dD_{01}D_{11} & D_{20} + dD_{11}^2 \end{bmatrix}. \quad (3.28)$$

Hence we obtain the following result giving a formula for the asymptotic efficiency of stochastic supervision relative to perfect supervision, where the stochastic supervisor's classification follows a beta model and is independent of the feature vector.

$$\mathrm{Eff}_p(\pi_1, \Delta, m, n) = \frac{q(\pi_1, \Delta, m, n,)\mathrm{Eff}_1(\pi_1, \Delta, m, n) + (p-1)\mathrm{Eff}_\infty(\pi_1, \Delta, m, n)}{q(\pi_1, \Delta, m, n) + (p-1)} \quad (3.29)$$

where

$$q(\pi_1, \Delta, m, n) = \frac{\left(1, -\frac{\lambda}{\Delta}\right)[H - F]^{-1}\left(1, -\frac{\lambda}{\Delta}\right)'\mathrm{Eff}_\infty}{1 + \pi_1\pi_0\Delta^2}. \quad (3.30)$$

$\mathrm{Eff}_1(\pi_1, \Delta, m, n)$ and $\mathrm{Eff}_\infty(\pi_1, \Delta, m, n)$ are asymptotic relative efficiencies for estimating the intercept and angle respectively, of the discriminant function and are given by

$$\mathrm{Eff}_1(\pi_1, \Delta, m, n) = \frac{\left(1, -\frac{\lambda}{\Delta}\right)H^{-1}\left(1, -\frac{\lambda}{\Delta}\right)'}{\left(1, -\frac{\lambda}{\Delta}\right)[H - F]^{-1}\left(1, -\frac{\lambda}{\Delta}\right)'}; \quad (3.31)$$

$$\mathrm{Eff}_\infty(\pi_1, \Delta, m, n) = 1 - D_{00}(1 + \pi_0\pi_1\Delta^2). \quad (3.32)$$

Eff_1 and Eff_∞ can also be interpreted as Eff when the dimensions of the feature vectors are 1 and ∞ respectively. The Eff_p is a convex combination of these two quantities.

We observe easily that when $m = n$, the matrix F in (3.30) and (3.31) reduces to (in O'Neill's[10] notation)

$$\begin{pmatrix} a_0 & a_1 \\ a_1 & a_2 \end{pmatrix}$$

where $a_i = D_{i0}$, giving Eff_p as the same formula as for the unsupervised case derived by O'Neill.[10] Further, if $\pi_1 = \pi_0 = \frac{1}{2}$, then $\lambda = 0$, and $\mathrm{Eff}_1 = \mathrm{Eff}_p = \mathrm{Eff}_\infty$.

4. COMPUTATION OF EFFICIENCY AND INTERPRETATION

We have computed Eff_1 and Eff_∞ as given by the formulae derived above for various values of π_1, Δ, m and n. The formulae derived above involve single and double integrals. For these integrals we have used subroutines of the NAG (Numerical Analysis Group) package. Efficiency values for some selected values of the parameters are presented in Table 1. A summary of the efficiencies is presented in Fig. 2.

As noted earlier, our beta distribution is a model for the stochastic nature of the supervision and $m = n$ is a case of lack of supervision and as $|m - n|$ increases the supervision gets better, reaching perfect supervision in the limit as $|m - n| \to \infty$. The area between the cumulative distribution function curves of $\mathscr{B}(m, n)$ and $\mathscr{B}(n, m)$, which is equal to the difference between the means of the two distributions $\mathscr{B}(m, n)$ and $\mathscr{B}(n, m)$ is $\frac{|m-n|}{m+n}$. This quantity increases with $|m - n|$ from 0 for $m = n$ to 1 as $|m - n| \to \infty$. However, we find that the efficiency does not depend only on $|m - n|$; for the same $|m - n|$, it is larger for smaller m and n; so a normalising factor should be a quantity less than $m + n$. At the suggestion of the referee, we tried $\sqrt{m+n}$ and $\frac{|m-n|}{\sqrt{m+n}}$ seems to bear an increasing relationship to efficiency in the range 1 to 5 of m, n that we have considered. This relationship is given in Fig. 2.

From Table 1 and Fig. 2, we notice that:

(1) Eff increases with supervision;

(2) Eff increases with Δ, the distance between the groups;

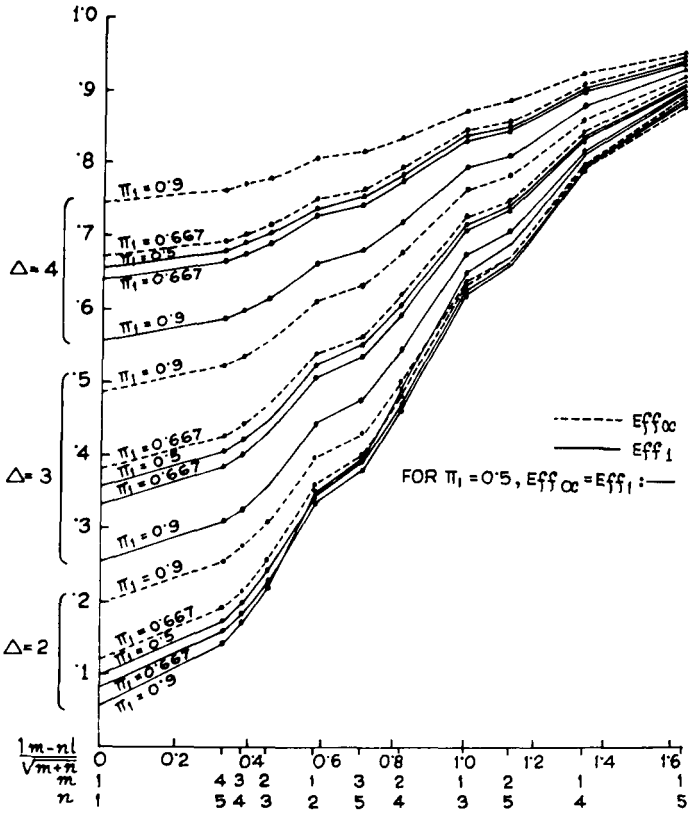

Fig. 2. Eff for various values of parameters and supervision index.

(3) For $\pi = 0.5$, the dimension of the feature vector is immaterial for Eff; this is evident from the formulae also, whereby $\text{Eff}_1 = \text{Eff}_p = \text{Eff}_x$ as we noted earlier;

(4) Eff_1 decreases with the value of π_1 away from $\frac{1}{2}$;

(5) Eff_x increases with the value of π_1 away from $\frac{1}{2}$.

(6) For $m = n$, Eff_1 and Eff_x coincide with the corresponding Eff of unsupervised learning as per O'Neill's[10] tables; this is also evident from the fact that our formula coincides with O'Neill's in the unsupervised case, as pointed out at the end of Section 3.

From (4) and (5) above, it follows that unsymmetric groups need a larger number of features for the same distance between groups.

In an earlier article,[4] we had computed these efficiencies, for a deterministic but error-prone supervision scheme, with a constant probability α of mis-supervision. From a comparison of these two situations, it turns out that:

$\alpha = 0.01$ and $m = 2$, $n = 7$ with the index $\frac{|m-n|}{\sqrt{m+n}} = 1.67$ are similar with Eff in the range of 90–95%;

$\alpha = 0.05$ and $m = 2$, $n = 6$ with the index $\frac{|m-n|}{\sqrt{m+n}} = 1.42$ are similar with Eff in the range of 74–90%;

Table 1. Asymptotic relative efficiency of normal discrimination with stochastic (Beta) supervision

			Δ = 2		Δ = 3		Δ = 4	
π_1	m	n	Eff_1	Eff_∞	Eff_1	Eff_∞	Eff_1	Eff_∞
0.5	1	5	0.8980	0.8980	0.9194	0.9194	0.9536	0.9536
	2	4	0.4762	0.4762	0.6094	0.6094	0.7852	0.7852
	3	3	0.1016	0.1016	0.3590	0.3590	0.6570	0.6570
0.667	1	5	0.8972	0.8988	0.9171	0.9217	0.9518	0.9554
	2	4	0.4694	0.4844	0.5972	0.6223	0.7762	0.7943
	3	3	0.0847	0.1217	0.3375	0.3820	0.6422	0.6719
0.9	1	5	0.9086	0.8958	0.9097	0.9312	0.9400	0.9647
	2	4	0.4927	0.5026	0.5475	0.6801	0.7222	0.8403
	3	3	0.0595	0.1996	0.2537	0.4892	0.5580	0.7483

$\alpha = 0.20$ and $m = 1$, $n = 2$ with the index $\frac{|m-n|}{\sqrt{m+n}} = 0.4$ are similar with Eff in the range of 35–80%;

$\alpha = 0.35$ and $m = 4$, $n = 5$ with the index $\frac{|m-n|}{\sqrt{m+n}} = 0.3$ are similar with Eff in the range of 18–70%;

$\alpha = 0.5$ and $m = n$ with the index $\frac{|m-n|}{\sqrt{m+n}} = 0$ are similar with Eff in the range of 13–75%.

Thus, our formula and the computations thereof give an idea of the worth of stochastic supervision. Stochastic supervision is useful if it is sufficiently far away from an unsupervised scheme. In situations

wherein the design of supervision systems is under consideration and a choice of supervision systems is available at various costs, the formulae above may help one to choose a system on the basis of cost-efficiency analysis. For instance, when $m = 2$, $n = 4$, $p = 1$, $\Delta = 4$, $\pi_1 = 0.667$, Eff $= 0.78$; this means that for these parameter values, 78 stochastic supervision samples are equivalent to 100 perfectly supervised samples. Of course, such an analysis depends upon a knowledge of the above parameters; these parameters can be estimated from a pilot sample of stochastically or perfectly supervised samples. In an earlier article,[5] we have given methods of estimating these parameters under stochastic supervision.

5. SUMMARY

Motivated by some situations in medical diagnosis and remote sensing, we consider the problem of discriminant analysis when the supervisor's classification is stochastic and deal with the problem of efficiency of this supervision relative to perfect supervision. For this, we formulate a model for stochastic supervision in terms of the beta distribution; this distribution enables us to include a variety of situations from perfect supervision to complete lack of supervision and also to quantify the amount of stochastic supervision. Our model which assumes the supervisor classification to be independent of the feature vector is not totally realistic; however, this independent situation is the most unfavourable case from the point of view of efficiency and hence it is worth studying. Under this model of supervision and for the case of two p-dimensional normal populations with a common dispersion matrix, we derive formulae for Efron efficiency of stochastic supervision, which is an index of the amount of statistical information contained in the stochastic supervision *vis-à-vis* perfect supervision; another way of looking at Efron efficiency is the relative sample size required under stochastic supervision compared to perfect supervision to achieve the same estimation efficiency of the discriminant function coefficients. We present tables and charts of this efficiency for various values of parameters of the two p-dimensional normal populations (the relevant parameters are the distance between the two populations and the mixing proportions) and various supervision situations in terms of our beta model. We find that stochastic supervision is quite useful unless the two beta parameters are nearly the same; and if the cost of stochastic supervision is much less than perfect supervision, it is quite worthwhile to use it.

Acknowledgements—The authors are grateful to an anonymous referee for his/her perceptive remarks, which considerably improved the presentation, especially section 4. The idea of finding an "index of supervision", the suggestion of $\frac{|m-n|}{\sqrt{m+n}}$ as a suitable index, the idea of presenting a chart of efficiency in terms of this index rather than extensive tables and interpretation of the efficiency results, are all due to the referee.

REFERENCES

1. R. S. Chhikara and J. McKeon, Linear discriminant analysis with misallocation in training samples, *J. Am. Statist. Ass.* **79**, 899–906 (1984).
2. J. Aitchison and C. B. Begg, Statistical diagnosis when basic cases are not classified with certainty, *Biometrika* **63**, 1–12 (1976).
3. U. A. Katre and T. Krishnan, Pattern recognition with imperfect supervision, *Pattern Recognition* **22**, 423–431 (1989).
4. T. Krishnan, Efficiency of learning with imperfect supervision, *Pattern Recognition* **21**, 183–188 (1988).
5. T. Krishnan and S. C. Nandy, Discriminant analysis with a stochastic supervisor, *Pattern Recognition* **20**, 379–384 (1987).
6. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Stat. Soc.* **39B**, 1–38 (1977).
7. P. A. Lachenbruch, Discriminant analysis when the initial samples are misclassified. II. Nonrandom misclassification models, *Technometrics* **16**, 419–424 (1974).
8. D. M. Titterington, An alternative stochastic supervisor in discriminant analysis, *Pattern Recognition* **22**, 91–95 (1989).
9. B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, *J. Am. Stat. Ass.* **70**, 892–898 (1975).
10. T. J. O'Neill, Normal discrimination with unclassified observations, *J. Am. Stat. Ass.* **73**, 821–826 (1978).

APPENDIX—NOMENCLATURE

$\mathcal{B}(m, n)$: beta distribution with parameters m, n.
Beta(.,.): complete beta integral.
ER: Error Rate; AER: Asymptotic Error Rate; Eff: Efron efficiency.
L: loglikelihood.
X: p-dimensional feature-vector and x: values taken thereof.
Z: classification of stochastic supervisor, indicating the probability of unit belonging to group 1; $0 \leq Z \leq 1$, z: values taken thereof.
y: actual group
$\beta_0 + \beta' x$: Bayes discriminant function.
$\pi_1(\pi_0)$: proportion of group 1(0); $\pi_0 + \pi_1 = 1$.
μ_0, μ_1: mean vectors of X in groups 0 and 1 respectively.
Σ: common dispersion matrix of X in the two groups.
N: sample size from the mixture of the two groups.
$(a_0, a)_N$, $(b_0, b)_N$, $(c_0, c)_N$: estimates of (β_0, β) based on a sample of size N by various schemes.
Δ: Mahalanobis distance between the two groups.
e_1: $(1, 0, \ldots, 0)$.
$\lambda = \log\left(\frac{\pi_1}{\pi_0}\right)$.
$M = ((m_{ij}))$: variance-covariance matrix of estimates $(a_0, a)_{N'}$ of (β_0, β).
I: information matrix (with various subscripts and with or without an asterisk) of (β_0, β).
$f_i(x)$: density of X in group i, $i = 0, 1$.
$f_i(x, z)$: density of X, Z in group i, $i = 0, 1$.
$\pi_i(x)$: posterior probability of group i given x, $i = 0, 1$.
$\pi_i(x, z)$: posterior probability of group i given x, z, $i = 0, 1$.
$u = n + m$, $v = n - m$.
A, B, C: information matrices of all the parameters.

About the Author—T. KRISHNAN received his Master's degree in Mathematics from Madras University in 1958 and Master's degree in Statistics and the Ph.D. degree from the Indian Statistical Institute in 1965 and 1968, respectively. He has been with the Indian Statistical Institute since then and is now a Professor of Applied Statistics. He has held visiting positions in the University of Southampton and the University of Western Australia. His research interests are in Pattern Recognition, Biostatistics and Psychometry.

About the Author—SUBHAS CHANDRA NANDY received his Master's degree in Statistics from Calcutta University in 1982 and the Master of Technology degree in Computer Science from the Indian Statistical Institute in 1985. He has been with the Indian Statistical Institute since then and is now a Programmer. His research interests are in Pattern Recognition and related areas.