Ann. Inst. Statist. Math.
Vol. 52, No. 3, 544–556 (2000)

DISCRIMINANT ANALYSIS WHEN A BLOCK OF OBSERVATIONS IS MISSING

HIE-CHOON CHUNG[1] AND CHIEN-PAI HAN[2]

[1] *Department of Industrial and Information Engineering, Kwangju University, Kwangju 503-703, South Korea*
[2] *Department of Mathematics, University of Texas at Arlington, Arlington, TX76019, U.S.A.*

(Received January 13, 1998; revised March 2, 1999)

Abstract. We consider the problem of classifying a $p \times 1$ observation into one of two multivariate normal populations when the training samples contain a block of missing observations. A new classification procedure is proposed which is a linear combination of two discriminant functions, one based on the complete samples and the other on the incomplete samples. The new discriminant function is easy to use. We consider the estimation of error rate of the linear combination classification procedure by using the leave-one-out estimation and bootstrap estimation. A Monte Carlo study is conducted to evaluate the error rate and the estimation of it. A numerical example is given to illustrate its use.

Key words and phrases: Block of missing observations, linear combination of two discriminant functions, linear combination classification, leave-one-out estimate, bootstrap estimate, Monte Carlo study.

1. Introduction

Let X be a $p \times 1$ vector of observations from one of two normal populations $\pi_i : N(\mu^{(i)}, \Sigma)$, $i = 1, 2$. The observation is to be classified by a rule based on the linear discriminant function. When the population parameters are unknown, Anderson (1951) suggested the statistic

$$W = \left[X - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})\right]' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}), \tag{1.1}$$

where $\bar{X}^{(i)}$ and S are the usual unbiased estimators of $\mu^{(i)}$, $i = 1, 2$, and Σ respectively. The statistic W in (1.1) is called Anderson's classification statistic. The error rate corresponding to this classification rule is called the unconditional error rate, which is

$$\gamma = \frac{1}{2}[\Pr(W < 0 \mid X \in \pi_1) + \Pr(W \geq 0 \mid X \in \pi_2)]. \tag{1.2}$$

The distribution of Anderson's classification statistic is quite complicated. Anderson (1951) and Wald (1944) considered the distribution of W, and Okamoto (1963) obtained an asymptotic expansion for the expectation of the conditional error rate. Since the exact expression for the unconditional error rate is very complicated, the conditional error rate is considered by assuming $\bar{X}^{(1)}$, $\bar{X}^{(2)}$, and S fixed. The conditional probability of misclassifying an observation X from π_1 into π_2 by W is

$$P_1 = \Pr(W < 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S; X \in \pi_1)$$
$$= \Phi\left\{\frac{\frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \mu^{(1)'} S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} \Sigma S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}}\right\},$$

where Φ denotes the cumulative distribution function of the univariate standard normal distribution.

Similarly the conditional probability of misclassifying an observation X from π_2 into π_1 by W is

$$P_2 = \Pr(W \geq 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S; X \in \pi_2) = \Phi\left\{ \frac{\mu^{(2)\prime} S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} \Sigma S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}} \right\}.$$

Hence the conditional error rate is

$$\gamma^* = \frac{1}{2}(P_1 + P_2). \tag{1.3}$$

In this paper we consider the situation when the training sample includes missing values. Chan and Dunn (1972, 1974) investigated the problem of handling incomplete observation vectors. They presented several methods of ignoring and estimating the values of these vectors, and used the resulting vectors in the discriminant function. There appears to be no uniformly best method. They suggested some guidelines in choosing the method for different situations.

Bohannon and Smith (1975) applied Hocking and Smith (1968) estimation procedure to estimate the parameters and compared this procedure to the standard procedure of ignoring the missing values in the construction of the classification rule and the estimation of the error rate.

Twedt and Gill (1992) examined the impact of different methods for replacing missing data in discriminant analysis. They concluded that the methods of replacing missing data were better than the one of ignoring the observation vectors with missing data.

The EM algorithm (Dempster *et al.* (1977)) may be used to estimate the parameters in the classification statistic. This algorithm consists of an iterative calculation involving two steps: i.e., the prediction and the estimation steps.

Anderson (1957) considered the maximum likelihood estimates of parameters of multivariate normal distributions when special patterns of missing observations are obtained in the training samples. The estimators are then used for substituting the unknown parameters in the classification rule.

2. Linear Combination Classification Procedure

We consider a special pattern which contains a block of missing observations. Instead of estimating the parameters, we construct two different discriminant functions from the complete data and incomplete data, respectively, and then a linear combination of these two linear discriminant functions is used to obtain the classification rule.

Let us partition the $p \times 1$ observation X as follows.

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix},$$

where Y is a $k \times 1$ vector and Z is a $(p-k) \times 1$ vector $(1 \leq k < p)$. Suppose random samples of sizes m_i, containing no missing values,

$$X_j^{(i)} = \begin{bmatrix} Y_{j(k\times 1)}^{(i)} \\ Z_{j(p-k)\times 1}^{(i)} \end{bmatrix}, \quad i = 1, 2;\ j = 1, 2, \ldots, m_i,$$

are available from

$$N_p(\mu^{(i)}, \Sigma) = N_p\left(\begin{bmatrix} \mu^{(i)}_{y(k\times 1)} \\ \mu^{(i)}_{z(p-k)\times 1} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy_{(k\times k)}} & \Sigma_{yz_{k\times(p-k)}} \\ \Sigma_{zy_{(p-k)\times k}} & \Sigma_{zz_{(p-k)\times(p-k)}} \end{bmatrix}\right),$$

and random samples of sizes $n_i - m_i$, which contain only the first k-components $Y^{(i)}_{j(k\times 1)}$, $i = 1, 2$; $j = m_i + 1, \ldots, n_i$, are available from $N_k(\mu^{(i)}_y, \Sigma_{yy})$. We denote by $X^{(i)}_j$, $i = 1, 2$; $j = 1, \ldots, m_i$, the complete observations, and by $Y^{(i)}_j$, $i = 1, 2$; $j = 1, \ldots, n_i$, the incomplete observations. Hence the data have the special pattern of missing values where a block of variables is missing on $n_i - m_i$ observations, and the remaining observations are all complete. For π_i the data are given as

$$\begin{array}{cccccc} Y^{(i)}_{11} & Y^{(i)}_{12} & \cdots & Y^{(i)}_{1m_i} & Y^{(i)}_{1(m_i+1)} \cdots & Y^{(i)}_{1n_i} \\ Y^{(i)}_{21} & Y^{(i)}_{22} & \cdots & Y^{(i)}_{2m_i} & Y^{(i)}_{2(m_i+1)} \cdots & Y^{(i)}_{2n_i} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ Y^{(i)}_{k1} & Y^{(i)}_{k2} & \cdots & Y^{(i)}_{km_i} & Y^{(i)}_{k(m_i+1)} \cdots & Y^{(i)}_{kn_i} \\ Z^{(i)}_{11} & Z^{(i)}_{12} & \cdots & Z^{(i)}_{1m_i} & & \\ Z^{(i)}_{21} & Z^{(i)}_{22} & \cdots & Z^{(i)}_{2m_i} & & \\ \vdots & \vdots & & \vdots & & \\ Z^{(i)}_{(p-k)1} & Z^{(i)}_{(p-k)2} & \cdots & Z^{(i)}_{(p-k)m_i} & & \end{array} \qquad i = 1, 2. \tag{2.1}$$

Then the sample means are given by

$$\begin{aligned} \bar{Y}^{(i)}_1 &= \frac{1}{m_i}\sum_{j=1}^{m_i} Y^{(i)}_j, \qquad i = 1, 2, \\ \bar{Y}^{(i)}_2 &= \frac{1}{n_i - m_i}\sum_{j=m_i+1}^{n_i} Y^{(i)}_j, \qquad i = 1, 2 \\ \bar{Z}^{(i)} &= \frac{1}{m_i}\sum_{j=1}^{m_i} Z^{(i)}_j, \qquad i = 1, 2. \end{aligned} \tag{2.2}$$

Let

$$\bar{Y}^{(i)} = \frac{1}{n_i}[m_i\bar{Y}^{(i)}_1 + (n_i - m_i)\bar{Y}^{(i)}_2], \qquad i = 1, 2. \tag{2.3}$$

We can construct two linear discriminant functions. The first linear discriminant function is based on the complete observations, $X^{(i)}_{j(p\times 1)}$, $i = 1, 2$; $j = 1, 2, \ldots, m_i$. We have

$$W_x = (\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1}_{xx}\left[X - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})\right],$$

where

$$\bar{X}^{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} X^{(i)}_j = \begin{bmatrix} \bar{Y}^{(i)}_1 \\ \bar{Z}^{(i)} \end{bmatrix}, \qquad i = 1, 2,$$

$$S_{xx} = \sum_{i=1}^{2}\sum_{j=1}^{m_i}(X^{(i)}_j - \bar{X}^{(i)})(X^{(i)}_j - \bar{X}^{(i)})'/\nu_x, \qquad \nu_x = m_1 + m_2 - 2.$$

The second linear discriminant function is based on the incomplete observations, $\bar{Y}^{(i)}_{j(k\times 1)}$, $i = 1, 2;\ j = 1, 2, \ldots, n_i$. We have

$$W_y = (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} \left[Y - \frac{1}{2}(\bar{Y}^{(1)} + \bar{Y}^{(2)}) \right],$$

where $\bar{Y}^{(i)}$ is given in (2.3), and

$$S_{yy} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_j^{(i)} - \bar{Y}^{(i)})(Y_j^{(i)} - \bar{Y}^{(i)})' / \nu_y, \qquad \nu_y = n_1 + n_2 - 2.$$

Now we combine the two linear discriminant functions and construct the classification rule which is a linear combination of W_x and W_y, namely

$$W_c = cW_x + (1 - c)W_y, \qquad 0 \le c \le 1. \tag{2.4}$$

We call W_c the linear combination classification statistic. An advantage of W_c is that it is easy to use. The observation X is classified into π_1 if

$$W_c = cW_x + (1 - c)W_y \ge 0;$$

otherwise it is classified into π_2. This classification procedure is called the linear combination classification procedure. This classification procedure depends on the value of c. The choice of c will be discussed later.

The probability of misclassifying an observation from π_1 into π_2 is given by

$$\beta_1 = \Pr\{W_c < 0 \mid X \in \pi_1\}.$$

Similarly the probability of misclassifying an observation from π_2 into π_1 is given by

$$\beta_2 = \Pr\{W_c \ge 0 \mid X \in \pi_2\}.$$

The unconditional error rate, with equal prior probability, is defined as

$$\beta = \frac{1}{2}(\beta_1 + \beta_2).$$

In order to find the error rate β, we need to know the distribution of W_c. However, this distribution is extremely complicated. Hence we consider the conditional error rate. The conditional distribution of W_c given $\bar{X}^{(1)}$, $\bar{X}^{(2)}$, S_{xx}, $\bar{Y}^{(1)}$, $\bar{Y}^{(2)}$, S_{yy} is obtained as follows. Let

$$\begin{aligned} W_x &= (\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1} X - \frac{1}{2}(\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1} (\bar{X}^{(1)} + \bar{X}^{(2)}) \\ &= \boldsymbol{a}' X + b = \boldsymbol{a}_1' Y + \boldsymbol{a}_2' Z + b, \end{aligned}$$

where

$$\boldsymbol{a}'_{(1\times p)} = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1}, \qquad \boldsymbol{a}_{(p\times 1)} = \begin{bmatrix} \boldsymbol{a}_{1(k\times 1)} \\ \boldsymbol{a}_{2(p-k)\times 1} \end{bmatrix},$$

$$b = -\frac{1}{2}(\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1} (\bar{X}^{(1)} + \bar{X}^{(2)}).$$

Also let

$$\begin{aligned} W_y &= (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} \left[Y - \frac{1}{2}(\bar{Y}^{(1)} + \bar{Y}^{(2)}) \right] \\ &= (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} Y - \frac{1}{2}(\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} (\bar{Y}^{(1)} + \bar{Y}^{(2)}) \\ &= \boldsymbol{d}' Y + e, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{d}' &= (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1}, \\ e &= -\frac{1}{2}(\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} (\bar{Y}^{(1)} + \bar{Y}^{(2)}). \end{aligned}$$

Then

$$\begin{aligned} W_c &= cW_x + (1-c)W_y \\ &= c(\boldsymbol{a}_1' Y + \boldsymbol{a}_2' Z + b) + (1-c)(\boldsymbol{d}' Y + e) \\ &= [c\boldsymbol{a}_1 + (1-c)\boldsymbol{d}]' Y + c\boldsymbol{a}_2' Z + cb + (1-c)e \\ &= A'Y + B'Z + F = H'X + F, \end{aligned}$$

where

$$\begin{aligned} A &= c\boldsymbol{a}_1 + (1-c)\boldsymbol{d}, \\ B &= c\boldsymbol{a}_2, \\ F &= cb + (1-c)e, \\ H &= \begin{bmatrix} A_{(k\times 1)} \\ B_{(p-k)\times 1} \end{bmatrix}. \end{aligned}$$

Since $W_c = H'X + F$ is a linear combination of the random variable X given $\bar{X}^{(1)}$, $\bar{X}^{(2)}$, S_{xx}, $\bar{Y}^{(1)}$, $\bar{Y}^{(2)}$, S_{yy}, and X is distributed as $N_p(\mu^{(i)}, \Sigma)$, hence W_c is distributed as $N(H'\mu^{(i)} + F, H'\Sigma H)$, $i = 1, 2$. Then the conditional probability of misclassifying an observation X from π_1 into π_2 by W_c is given by

$$\begin{aligned} \beta_1^* &= \Pr(W_c < 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S_{xx}, \bar{Y}^{(1)}, \bar{Y}^{(2)}, S_{yy}; X, Y \in \pi_1) \\ &= \Phi\left(\frac{-H'\mu^{(1)} - F}{\sqrt{H'\Sigma H}} \right). \end{aligned} \tag{2.5}$$

Similarly,

$$\begin{aligned} \beta_2^* &= \Pr(W_c \geq 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S_{xx}, \bar{Y}^{(1)}, \bar{Y}^{(2)}, S_{yy}; X, Y \in \pi_2) \\ &= 1 - \Phi\left(\frac{-H'\mu^{(2)} - F}{\sqrt{H'\Sigma H}} \right) = \Phi\left(\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}} \right). \end{aligned} \tag{2.6}$$

Hence the conditional error rate, with equal prior probability, is defined as

$$\beta^* = \frac{1}{2}(\beta_1^* + \beta_2^*). \tag{2.7}$$

Using the linear combination classification statistic in (2.4), X is classified to π_1 if $W_c > 0$; otherwise it is classified to π_2. Given the training samples, the conditional error rate β^* depends on the value of c. The best value of c may be determined so that the conditional error rate is minimized. However, the minimization process is very tedious and intractable. Hence we propose to use the following value of c.

Let $\bar{X}^{(i)}$ and $S^{(i)}$ be the sample mean and sample covariance matrix of the complete observation vectors of sizes m_i, and $\bar{Y}^{(i)}$ and $S_y^{(i)}$ be the sample mean and sample covariance matrix of the incomplete observation vectors of sizes n_i for each population π_i, $i = 1, 2$ (see data in (2.1)). Since it is assumed that the two populations have the same covariance matrix Σ, the sample covariance matrices $S^{(1)}$ and $S^{(2)}$ are pooled to obtain an unbiased estimate of Σ,

$$S = \frac{(m_1 - 1)S^{(1)} + (m_2 - 1)S^{(2)}}{(m_1 + m_2 - 2)}.$$

Similarly, an unbiased estimate of Σ_{11} is

$$S_y = \frac{(n_1 - 1)S_y^{(1)} + (n_2 - 1)S_y^{(2)}}{(n_1 + n_2 - 2)}.$$

From these sample quantities, we propose to use the operational c^* which is given by

$$c^* = \frac{\left(\frac{1}{m_1} + \frac{1}{m_2}\right)^{-1} D^2}{\left(\frac{1}{m_1} + \frac{1}{m_2}\right)^{-1} D^2 + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D_y^2}, \tag{2.8}$$

where

$$\begin{aligned} D^2 &= (\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}), \\ D_y^2 &= (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_y^{-1} (\bar{Y}^{(1)} - \bar{Y}^{(2)}). \end{aligned} \tag{2.9}$$

The rationale of using this value c^* is given as follows. It is known that the error rates will depend on the Mahalanobis distance and the information from the samples. Usually the error rate is small when the Mahalanobis distance is large or the sample size is large. The operational c^* in (2.8) can be justified in the sense of the training sample sizes of n_i and m_i, $i = 1, 2$, and the squared distances of D^2 and D_y^2 in (2.9). The values of m_i and D^2, $i = 1, 2$, for the complete data characterize the performance of W_x in (2.4); while the values of n_i and D_y^2, $i = 1, 2$, for the incomplete data characterize the performance of W_y in (2.4). When D^2 is much larger than D_y^2, it shows that the component Z of the variable X has large discriminant power. We should use W_x and c^* is made to be large and close to one. Similarly when m_1 and m_2 are large and near the values of n_1 and n_2 respectively, this indicates that the numbers of observations with missing values are small in the two samples, so W_y is not as efficient as W_x. Hence c^* is made to be large again. On the contrary, when D^2 is close to D_y^2 (indicating Z does not provide additional discriminant power) and when m_1 and m_2 are small, c^* becomes small, and W_y has a larger weight. For the special case of $n_1 = n_2$, $m_1 = m_2$, c^* in (2.8) reduces to

$$c_s^* = \frac{m_1 D^2}{m_1 D^2 + n_1 D_y^2}.$$

Table 1. Values of parameters in the Monte Carlo study.

p	k	$n = 20$	$n = 50$	$n = 100$
2	1	$m = 6, 10, 14, 18$	$m = 6, 10, 14, 18, 30, 46$	$m = 6, 10, 14, 18, 30, 46, 70, 90$
		$\Delta_x^2 = .64, 1, 4, 9, 16$	$\Delta_x^2 = .64, 1, 4, 9, 16$	$\Delta_x^2 = .64, 1, 4, 9, 16$
5	1	$m = 10, 14, 18$	$m = 10, 14, 18, 30, 46$	$m = 10, 14, 18, 30, 46, 70, 90$
	3	$\Delta_x^2 = .64, 1, 4, 9, 16$	$\Delta_x^2 = .64, 1, 4, 9, 16$	$\Delta_x^2 = .64, 1, 4, 9, 16$
10	1	$m = 10, 14, 18$	$m = 10, 14, 18, 30, 46$	$m = 10, 14, 18, 30, 46, 70, 90$
		$\Delta_x^2 = 1, 4$	$\Delta_x^2 = 1, 4$	$\Delta_x^2 = 1, 4$

There may be other choices of the value of c, for example, selecting c to minimize the leave-one-out estimate of β^*. This would involve an intensive search in the interval (0,1). Another way to determine c is to minimize the asymptotically unbiased estimator given in McLahlan (1974). This would require the asymptotic expansion of the distribution of the error rate for W_c which is very complicated. In this paper we will use c^* in (2.8) because it compares favorable with Anderson's procedure and it is easy to use. In the numerical example in Section 5, we will also consider the procedure of selecting c to minimize the bootstrap estimate of β^*.

3. Comparison of the Error Rates

In order to compare the different classification procedures we need to evaluate the error rates. We evaluate the performance of the linear combination classification procedure in (2.4) and compare its conditional error rate β^* in (2.7) with the conditional error rate obtained by substituting the parameter estimates into the usual linear discriminant function. The parameter estimates are found by Anderson (1957), Hocking and Smith (1968), and the EM algorithm (Dempster *et al.* (1977)). Since the distributions of the discriminant functions for the different procedures are intractable, we use a Monte Carlo study to simulate the error rates. It can be shown that the linear combination classification statistic is invariant under nonsingular linear transformations when the data contain missing observation. In view of the invariance property, we may let, without loss of generality, $\mu^{(1)} = 0$, $\mu^{(2)} = [\Delta_y, 0, \ldots, \Delta_z, \ldots, 0]'$, and $\Sigma = I$. Using the canonical form, we have the Mahalanobis distance $\Delta_x^2 = (\mu^{(1)} - \mu^{(2)})'(\mu^{(1)} - \mu^{(2)}) = \mu^{(2)'}\mu^{(2)} = \Delta_y^2 + \Delta_z^2$. So $\Delta_z = \sqrt{\Delta_x^2 - \Delta_y^2}$. Let $R = \Delta_y^2/\Delta_x^2$, where $0 \leq R \leq 1$. Thus when we fix Δ_x^2, the parameter R changes as Δ_y^2 varies. For fixed Δ_x^2, the error rates of the linear combination classification procedure W_c in (2.4), Anderson's procedure, the EM algorithm, and Hocking and Smith procedure will be simulated as R changes from 0 to 1. Table 1 gives the combinations of the choices of k, m and Δ_x^2 in the simulation experiments.

The comparisons of the error rates are given in Table 2 and Table 3 for some combinations of p, k, n, m, Δ_x^2, and R. The number of simulations is 1000. We can see that the three error rates obtained by Anderson's procedure, the EM algorithm and Hocking and Smith (H-S) procedure are almost the same for any combination of p, k, n, m, Δ_x^2, and R. Let us now define the difference of error rates between Anderson's procedure and the linear combination classification procedure as

DER = [average of conditional error rate γ^* in (1.3) obtained by Anderson's procedure]
− [average of conditional error rate β^* in (2.7)].

Table 2. Comparison of error rates.

$p = 2$, $k = 1$, $n = 50$, $m = 46$					
Δ_x^2	R	W_c (S. D.)	Anderson and EM* (S. D.)	H-S (S. D.)	DER
1.0	0.0	0.3152 (0.0100)	0.3139 (0.0070)	0.3139 (0.0070)	−0.0013
	0.2	0.3174 (0.0118)	0.3139 (0.0070)	0.3139 (0.0071)	−0.0035
	0.4	0.3184 (0.0109)	0.3140 (0.0070)	0.3140 (0.0070)	−0.0044
	0.6	0.3177 (0.0085)	0.3141 (0.0070)	0.3141 (0.0070)	−0.0036
	0.8	0.3151 (0.0057)	0.3143 (0.0070)	0.3143 (0.0070)	−0.0008
	1.0	0.3109 (0.0031)	0.3144 (0.0071)	0.3144 (0.0071)	0.0035
4.0	0.0	0.1629 (0.0044)	0.1627 (0.0041)	0.1627 (0.0042)	−0.0002
	0.2	0.1649 (0.0072)	0.1626 (0.0041)	0.1626 (0.0041)	−0.0024
	0.4	0.1672 (0.0087)	0.1624 (0.0040)	0.1624 (0.0039)	−0.0048
	0.6	0.1679 (0.0077)	0.1623 (0.0039)	0.1623 (0.0039)	−0.0056
	0.8	0.1657 (0.0051)	0.1624 (0.0040)	0.1624 (0.0040)	−0.0033
	1.0	0.1606 (0.0022)	0.1627 (0.0047)	0.1627 (0.0047)	0.0021

* For each combination, the error rates and the standard deviations of Anderson and EM algorithm are the same, respectively.

Table 3. Comparison of error rates.

$p = 5$, $k = 1$, $n = 20$, $m = 10$					
Δ_x^2	R	W_c (S. D.)	Anderson and EM* (S. D.)	H-S (S. D.)	DER
1.0	0.0	0.3827 (0.0463)	0.3795 (0.0459)	0.3797 (0.0459)	−0.0032
	0.2	0.3804 (0.0429)	0.3803 (0.0444)	0.3799 (0.0442)	−0.0001
	0.4	0.3757 (0.0413)	0.3804 (0.0428)	0.3794 (0.0424)	−0.0047
	0.6	0.3692 (0.0404)	0.3801 (0.0408)	0.3786 (0.0403)	0.0109
	0.8	0.3615 (0.0399)	0.3799 (0.0387)	0.3779 (0.0382)	0.0184
	1.0	0.3526 (0.0405)	0.3795 (0.0374)	0.3770 (0.0370)	0.0269
4.0	0.0	0.2181 (0.0411)	0.2166 (0.0399)	0.2168 (0.0399)	−0.0015
	0.2	0.2168 (0.0398)	0.2169 (0.0402)	0.2163 (0.0398)	0.0001
	0.4	0.2108 (0.0348)	0.2168 (0.0399)	0.2155 (0.0392)	0.0060
	0.6	0.2028 (0.0306)	0.2168 (0.0394)	0.2150 (0.0386)	0.0140
	0.8	0.1935 (0.0284)	0.2172 (0.0388)	0.2148 (0.0375)	0.0237
	1.0	0.1839 (0.0275)	0.2188 (0.0391)	0.2160 (0.0382)	0.0349

* For each combination, the error rates and the standard deviations of Anderson and EM algorithm are the same, respectively.

From Table 2 and Table 3, we can see that there is a point where the sign of DER changes when R goes from 0 to 1. Let us call this point cut-off point R^*. Then R^* divides the parameter space $(0 \leq R \leq 1)$ into two regions with $0 \leq R \leq R^*$ and $R^* < R \leq 1$. The linear combination classification procedure is better than Anderson's procedure, the EM algorithm and Hocking and Smith procedure if R is greater than R^*. We found that R^* depends on the combination of p, k, n, m, and Δ_x^2. For example, R^* appears to be very small for $p = 5$, $k = 1$, $n = 20$, $m = 10$, $\Delta_x^2 = 4$ in Table 3, but very large for

Table 4. Cut-off point R^*.

$p=5,\ k=1$		m						
Δ_x^2	n	10	14	18	30	46	70	90
0.64	20	0.28	0.29	0.29	-	-	-	-
	50	0.16	0.26	0.29	0.37	0.38	-	-
	100	0.12	0.18	0.25	0.37	0.45	0.52	0.53
1.0	20	0.22	0.28	0.31	-	-	-	-
	50	0.17	0.24	0.31	0.42	0.48	-	-
	100	0.14	0.18	0.30	0.45	0.54	0.63	0.63
4.0	20	0.21	0.29	0.40	-	-	-	-
	50	0.22	0.39	0.48	0.65	0.70	-	-
	100	0.31	0.50	0.59	0.75	0.81	0.85	0.87
9.0	20	0.22	0.31	0.46	-	-	-	-
	50	0.23	0.49	0.61	0.69	0.79	-	-
	100	0.43	0.64	0.72	0.82	0.85	0.89	0.89
16.0	20	0.17	0.38	0.46	-	-	-	-
	50	0.37	0.55	0.64	0.77	0.82	-	-
	100	0.48	0.66	0.73	0.83	0.88	0.91	0.91

Table 5. Bootstrap and leave-one-out estimates for β^*.

p	k	Δ_x^2	n	m	R	$\bar{\beta}^*$	Boot (S. D.)	Leave (S. D.)
2	1	1	20	10	0.2	0.3443	0.3317 (0.1188)	0.3331 (0.1255)
					0.9	0.3238	0.3313 (0.1092)	0.3230 (0.1081)
2	1	1	20	18	0.2	0.3295	0.3246 (0.0844)	0.3204 (0.0839)
					0.9	0.3200	0.3215 (0.0792)	0.3148 (0.0806)
2	1	1	50	10	0.2	0.3483	0.3370 (0.1231)	0.3407 (0.1253)
					0.9	0.3180	0.3234 (0.1068)	0.3161 (0.1017)
5	1	1	20	10	0.2	0.3765	0.3353 (0.1162)	0.3844 (0.1309)
					0.9	0.3537	0.3378 (0.1188)	0.3618 (0.1221)
5	3	1	20	10	0.2	0.3810	0.3331 (0.1189)	0.3704 (0.1285)
					0.9	0.3591	0.3315 (0.1190)	0.3424 (0.1137)
5	3	1	20	18	0.2	0.3600	0.3396 (0.0849)	0.3450 (0.0888)
					0.9	0.3458	0.3370 (0.0681)	0.3265 (0.0840)
2	1	4	20	18	0.2	0.1726	0.1673 (0.0646)	0.1661 (0.0640)
					0.9	0.1670	0.1662 (0.0623)	0.1620 (0.0617)
2	1	4	50	46	0.2	0.1649	0.1660 (0.0389)	0.1651 (0.0385)
					0.9	0.1638	0.1630 (0.0387)	0.1614 (0.0389)
5	1	4	20	18	0.2	0.1937	0.1809 (0.0683)	0.1894 (0.0695)
					0.9	0.1773	0.1746 (0.0665)	0.1733 (0.0644)
5	3	4	20	18	0.2	0.1964	0.1821 (0.0665)	0.1879 (0.0665)
					0.9	0.1843	0.1781 (0.0679)	0.1710 (0.0656)
5	3	4	20	10	0.2	0.2211	0.1859 (0.0981)	0.2163 (0.1027)
					0.9	0.1954	0.1806 (0.0925)	0.1805 (0.0872)

$p=2$, $k=1$, $n=50$, $m=46$, $\Delta_x^2=1$ in Table 2. Table 4 gives the cut-off points of R^*. From the simulations, we obtain the following properties of the linear combination classification procedure.

(a) For fixed p, k, n, and Δ_x^2, the value of R^* increases as m increases.

(b) For fixed p, n, m, and Δ_x^2, the value of R^* increases as k increases.

From the properties (a) and (b), we conclude that the linear combination classification is better than Anderson's procedure, the EM algorithm, and Hocking and Smith procedure for given p, n, and Δ_x^2 as the proportion of missing observation gets larger.

4. Estimation of Error Rates

The performance of a classification procedure is measured by its error rate. Since error rate depends on unknown parameters, we must estimate it by samples. We will consider the estimates of the conditional error rate β^* in (2.7) for the linear combination classification procedure. The algorithm of McLachlan (1980) can be extended to obtain the bootstrap estimate of the bias correction when the training samples contain missing value. Also the leave-one-out estimate of the error rate will be obtained. A Monte Carlo study is conducted to obtain the bootstrap and the leave-one-out estimate of β^* for some combinations of $n = 20(m = 10, 18)$, $50(m = 10, 46)$, $p = 2$, $5(k = 1, 3)$, $\Delta_x^2 = 1, 4$ in Table 1 with $R = 0.2, 0.9$. The number of simulations is 1000, and 300 bootstrap samples are generated for each simulation.

Table 5 shows the properties of the bootstrap and leave-one-out estimates for β^* in (2.7). We summarize our findings from the Monte Carlo study as follows:

1) When n and m are moderately larger than p, i.e., $p = 2$, $n = 20$, $m = 10$ and 18, both estimates appear to be nearly unbiased.

2) When n and m are sufficiently larger than p, i.e., $p = 2$, $n = 50$, $m = 46$, both estimates are improved compared to the case in 1).

3) When n and m are not moderately larger than p, i.e., $p = 5$, $k = 1$, 3, $n = 20$, $m = 10$, the estimates for the leave-one-out method generally appear to be nearly unbiased but not for the bootstrap, specially for $R = 0.2$. This happens since information for the discrimination depends on the variables in which data contain missing values.

4) The standard deviations for both estimates are almost the same for each combination.

The conditional error rate can be estimated by substituting the estimates $\hat{\Sigma}$, $\hat{\mu}^{(i)}$ for Σ, $\mu^{(i)}$ in (2.5) and (2.6). Le $\hat{\mu}^{(i)} = [\bar{Y}^{(i)}, \bar{Z}^{(i)}]'$ in (2.2) and (2.3) be the estimate of $\mu^{(i)}$. For the covariance matrices, let $\hat{\Sigma}_{xc}^{(i)} = \begin{bmatrix} \hat{\Sigma}_{yyc}^{(i)} & \hat{\Sigma}_{yzc}^{(i)} \\ \hat{\Sigma}_{zyc}^{(i)} & \hat{\Sigma}_{zzc}^{(i)} \end{bmatrix}$ be the estimate from the complete observations of sizes m_i. Also let $\hat{\Sigma}_{yyi}^{(i)}$ be the estimate from the incomplete observations of sizes $n_i - m_i$ using only the Y observations in (2.1), $i = 1, 2$. Then we suggest the combined estimates,

$$\hat{\Sigma}^{(i)} = \begin{bmatrix} \frac{m_i}{n_i}\hat{\Sigma}_{yyc}^{(i)} + \frac{n_i - m_i}{n_i}\hat{\Sigma}_{yyi}^{(i)} & \hat{\Sigma}_{yzc}^{(i)} \\ \hat{\Sigma}_{zyc}^{(i)} & \hat{\Sigma}_{zzc}^{(i)} \end{bmatrix} \quad \text{for} \quad \Sigma^{(i)}, l = 1, 2.$$

Now the pooled estimate of the covariance matrices is given by

$$\hat{\Sigma} = \frac{n_1}{n_1 + n_2}\hat{\Sigma}^{(1)} + \frac{n_2}{n_1 + n_2}\hat{\Sigma}^{(2)}.$$

Table 6. Population 1: Success.

x_1	x_2	x_3	x_4	x_5	The variables are:
2.97	420	800	600	497	x_1 = Undergraduate GPA
3.80	330	710	380	563	x_2 = GRE Verbal
2.50	270	700	340	510	x_3 = GRE Quantitative
2.50	400	710	600	563	x_4 = GRE Analytic
3.30	280	800	450	543	x_5 = TOEFL Score
2.60	310	660	425	507	
2.70	360	620	590	537	
3.10	220	530	340	543	
2.60	350	770	560	580	
3.20	360	750	440	577	
3.65	440	700	630		
3.56	640	520	610		
3.00	480	550	560		
3.18	550	630	630		
3.84	450	660	630		
3.18	410	410	340		
3.43	460	610	560		
3.52	580	580	610		
3.09	450	540	570		
3.70	420	630	660		

5. Numerical Example

Application of the bootstrap method to estimate the error rate, β^* in (2.7) is illustrated by using unpublished real data sets. They are given by the Admissions Office at the University of Texas at Arlington. The data sets contain two populations. One population is the Success Group that the students receive their master's degree. The other population is the Failure Group that they do not complete their master's degree. For each population, there are 10 foreign students and 10 United States students. Each foreign student has 5 variables which are x_1 = undergraduate GPA, x_2 = GRE verbal, x_3 = GRE quantitative, x_4 = GRE analytic, and x_5 = TOEFL score. For each United States student, one variable, x_5 = TOEFL score is missing. The data sets are shown in Table 6 and Table 7.

Using this data set, we obtain the discriminant function

$$W_c = cW_x + (1-c)W_y,$$

where

$$W_x = \boldsymbol{a}'X + b,$$
$$\boldsymbol{a}' = [-1.9957 \ -0.0170 \ -0.0004 \ 0.0034 \ 0.0242], \quad b = -2.5252,$$
$$W_y = \boldsymbol{d}'X + e,$$
$$\boldsymbol{d}' = [0.5302 \ -0.0042 \ -0.0023 \ 0.2406], \quad e = 0.2846,$$
$$c = 0.7532.$$

For this example, we generate 300 bootstrap samples to estimate β^*. The result of using $c^* = 0.7532$ is that the bootstrap estimate of β^* is 0.3435. We also consider the

Table 7. Population 2: Failure.

x_1	x_2	x_3	x_4	x_5	The variables are:
3.75	250	730	460	513	x_1 = Undergraduate GPA
3.11	320	760	610	560	x_2 = GRE Verbal
3.00	360	720	525	540	x_3 = GRE Quantitative
2.60	370	780	500	500	x_4 = GRE Analytic
3.50	300	630	380	507	x_5 = TOEFL Score
3.50	390	580	370	587	
3.10	380	770	500	520	
2.30	370	640	200	520	
2.85	340	800	540	517	
3.50	460	750	560	597	
3.15	630	540	600		
2.93	350	690	620		
3.20	480	610	480		
2.76	630	410	530		
3.00	550	450	500		
3.28	510	690	730		
3.11	640	720	520		
3.42	440	580	620		
3.00	350	430	480		
2.67	480	700	670		

procedure of selecting c to minimize the bootstrap estimate of β^*. We search c in the interval $[0.05(0.05)0.95]$. The best value occurs at $c = 0.60$ with a bootstrap estimate of β^* 0.2933. The error rate is less than that of c^* but not by much.

6. Conclusion

Discriminant analysis is a multivariate technique concerned with classifying a $p \times 1$ observation X to one of several distinct populations. If the training samples do not contain missing values, the Anderson's classification statistic is used to classify the observation. In this paper, we consider situation that the training samples contain incomplete observation vectors which have a special pattern of missing data; i.e., all missing values occur on the same variables. There are several methods to deal with missing value in discriminant analysis. One method is to estimate the unknown parameters first, and then the estimates of them are substituted into the usual discriminant functions for classification. We call these methods substitution methods for the incomplete data. A new classification procedure in this situation is proposed. The proposed discriminant function is a linear combination of two well defined Fisher's linear discriminant functions. It does not require the estimation of the missing values. The performance of this classification rule is compared to the substitution methods. We found that the linear combination classification procedure is better than the substitution methods as the proportion of missing observations gets larger. A numerical example is given and it is shown that the linear combination classification procedure is easy to use.

REFERENCES

Anderson, T. W. (1951). Classification by multivariate analysis, *Psychometrika*, **16**, 31–50.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *J. Amer. Statist. Assoc*, **52**, 200–203.

Bohannon, Tom R. and Smith, W. B. (1975). Classification based on incomplete data records *ASA Proc. Social Statistics Section*, 214–218.

Chan, L. S. and Dunn, O. J. (1972). The treatment of missing values in discriminant analysis-1, The sampling experiment, *J. Amer. Statist. Assoc.*, **67**, 473–477.

Chan, L. S. and Dunn, O. J. (1974). A note on the asymptotical aspect of the treatment of missing values in discriminant analysis, *J. Amer. Statist. Assoc.*, **69**, 672–673.

Dempster, A. P., Laird, N. M. and Rubin, R. J. A. (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Hocking, R. R. and Smith, W. B. (1968). Estimation of parameters in the mutivariate normal distribution with missing observation, *J. Amer. Statist. Assoc.*, **63**, 159–173.

McLachlan, G. J. (1974). An asymptotic unbiased technique for estimating the error rates in discriminant analysis, *Biometrics*, **30**, 239–249.

McLachlan, G. J. (1980). The efficiency of Efron's bootstrap approach applied to error rate estimation in discriminant analysis, *J. Statist. Comput. Simulation*, **11**, 273–279.

Okamoto, Masashi (1963), An asymptotic expansion for the distribution of the linear discriminant function, *Ann. Math. Statist.*, **34**, 1286–1301.

Twedt, Daniel J. and Gill, D. S. (1992), Comparison of algorithm for replacing missing data in discriminant analysis, *Comm. Statist. Theory Methods*, **21**, 1567–1578.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, *Ann. Math. Statist.*, **15**, 145–162.