The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 49-63

Discriminant Analysis with Incomplete Pattern Vectors

Hie-Choon Chung[1)]

Abstract

We consider the problem of classifying a p x 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.

1. Introduction

We consider the problem of classifying a p x 1 observation X of unknown origin to one of two distinct populations using an appropriate classification rule. If the population π_i has density $f_i(X)$, i = 1, 2, the Bayes procedure classifies X into π_1 if

$$\frac{f_1(X)}{f_2(X)} \geq c, \tag{1.1}$$

where c is a constant which depends on the prior probabilities and costs of misclassification; otherwise X is classified into π_2. In the particular case of two populations being equally likely and the costs of misclassification being equal, c = 1.

If the populations are multivariate normal with equal covariance matrix, that is π_i : $N(\mu^{(i)}, \Sigma)$, (1.1) becomes, after taking logarithm,

$$[X - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})]' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \geq 0. \tag{1.2}$$

Then the random variable,

$$U = [X - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})]' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}),$$

is distributed as $N(\Delta^2/2, \Delta^2)$ if X comes from π_1 and as $N(-\Delta^2/2, \Delta^2)$ if X comes

1) Full-time Lecturer, Department of Industrial Information Engineering, Kwangju University, Kwangju, 502-703, Korea

from π_2, where $\Delta^2 = (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ is the Mahalanobis squared distance between the two populations. When X comes from π_1 , the probability of misclassification is

$$P(2|1) = \Pr(U < 0 | X \in \pi_1) = \Phi(-\Delta/2).$$

Similarly, the probability of misclassifying X from π_2 to π_1 is

$$P(1|2) = \Pr(U \geq 0 | X \in \pi_2) = \Phi(-\Delta/2).$$

Then the optimal error rate is defined as

$$\alpha = \frac{1}{2}[P(2|1) + P(1|2)] = \Phi(-\Delta/2). \tag{1.3}$$

In practice the population parameters are usually unknown. Then independent random samples $\{ X_1^{(i)}, X_2^{(i)}, \ldots, X_{n_i}^{(i)} \}$ of sizes n_i, $i = 1, 2$, are taken from the two populations. When the training samples do not contain missing values, Anderson (1951) suggested the method of simple substitution of $\bar{X}^{(i)}$ for $\mu^{(i)}$ and S for Σ in (1.2), where $\bar{X}^{(i)}$ and S are the usual unbiased estimators of $\mu^{(i)}$, $i = 1, 2$, and Σ respectively. The statistic

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

is called Anderson's classification statistic. The error rate correponding to this classification rule is called the unconditional error rate, which is

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

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

$$P_1 = \Pr(W < 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S ; X \in \pi_1)$$

$$= \Phi\left\{ \frac{\frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \mu^{(1)'} S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} \Sigma S^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)})}} \right\}.$$

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

$$P_2 = \Pr(W \geq 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S ; X \in \pi_2)$$

$$= \Phi\left\{ \frac{\mu^{(2)'} S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} \Sigma S^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)})}} \right\}.$$

Hence the conditional error rate is

$$\gamma^* = \frac{1}{2}(P_1 + P_2). \tag{1.4}$$

In this paper we consider the situation when the training sample includes incomplete observation vectors. Chan and Dunn (1972, 1974) presented several methods of ignoring and estimating the values of these vectors, and used the resulting vectors in the discriminant function.

Bohannon and Smith (1975) applied Hocking-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 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 (1.2).

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 x 1 observation X as follows.

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

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

$$X_j^{(i)} = \begin{bmatrix} Y_j^{(i)} \\ Z_j^{(i)} \end{bmatrix}, \; i = 1, 2 \; ; \; j = 1, 2, \ldots, m_i,$$

are available from

$$N_p(\mu^{(i)}, \Sigma) = N_p\left(\begin{bmatrix} \mu_y^{(i)} \\ \mu_z^{(i)} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{zz} \end{bmatrix} \right),$$

and random samples of sizes $n_i - m_i$, which contain only the first k-components $Y_j^{(i)}$, $i = 1, 2$; $j = m_i + 1, \ldots, n_i$, are available from $N_k(\mu_y^{(i)}, \Sigma_{yy})$. We denote by $X_j^{(i)}$, $i = 1, 2$; $j = 1, \ldots, m_i$, the complete observations, and by $Y_j^{(i)}$, $i = 1, 2$; $j = 1, \ldots, n_i$, 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. Then the sample means are given by

$$\bar{Y}_1^{(i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_j^{(i)}, \quad i = 1, 2, \tag{2.1}$$

$$\bar{Y}_2^{(i)} = \frac{1}{n_i - m_i} \sum_{j=m_i+1}^{n_i} Y_j^{(i)}, \quad i = 1, 2, \tag{2.2}$$

$$\bar{Z}^{(i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} Z_j^{(i)}, \quad i=1, 2. \tag{2.3}$$

Let

$$\bar{Y}^{(i)} = \frac{1}{n_i} [m_i \bar{Y}_1^{(i)} + (n_i - m_i) \bar{Y}_2^{(i)}], \quad i=1, 2. \tag{2.4}$$

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

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

where

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

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

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

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

where $\bar{Y}^{(i)}$ is given in (2.4), 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

$$W_c = c\,W_x + (1-c)\,W_y, \quad 0 \le c \le 1. \tag{2.5}$$

We call W_c the linear combination classification statistic. An advantage of W_c is that it is easy to use. 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

$$\beta = \frac{1}{2}(\beta_1 + \beta_2). \tag{2.6}$$

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

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

$$= \boldsymbol{a}' X + b = \boldsymbol{a}_1' Y + \boldsymbol{a}_2' Z + b ,$$

where

$$\boldsymbol{a}' = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1} ,$$

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix},$$

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

Also let

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

$$= \boldsymbol{d}' Y + e ,$$

where

$$\boldsymbol{d}' = (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} ,$$

$$e = -\frac{1}{2}(\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1} (\bar{Y}^{(1)} + \bar{Y}^{(2)}) .$$

Then

$$W_c = c\,W_x + (1-c)\,W_y$$

$$= c(\boldsymbol{a}_1'Y + \boldsymbol{a}_2'Z + b) + (1-c)(\boldsymbol{d}'Y + e)$$
$$= A'Y + B'Z + F = H'X + F,$$

where

$$A = c\,\boldsymbol{a}_1 + (1-c)\,\boldsymbol{d},$$
$$B = c\,\boldsymbol{a}_2,$$
$$F = c\,b + (1-c)\,e,$$
$$H = \begin{bmatrix} A \\ B \end{bmatrix}.$$

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 $N(H'\mu^{(i)} + F, H'\Sigma H)$. Then the conditional probability of misclassifying an observation X from π_1 into π_2 by W_c is given by

$$\beta_1^* = \Pr(W_c < 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S_{xx}, \bar{Y}^{(1)}, \bar{Y}^{(2)}, S_{yy}; X, Y \in \pi_1)$$
$$= \Phi\left(-\frac{H'\mu^{(1)} + F}{\sqrt{H'\Sigma H}}\right). \tag{2.7}$$

Similarly,

$$\beta_2^* = \Pr(W_c \geq 0 \mid \bar{X}^{(1)}, \bar{X}^{(2)}, S_{xx}, \bar{Y}^{(1)}, \bar{Y}^{(2)}, S_{yy}; X, Y \in \pi_2)$$
$$= 1 - \Phi\left(-\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}}\right) = \Phi\left(\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}}\right). \tag{2.8}$$

Hence the conditional error rate for β in (2.6), with equal prior probability, is defined as

$$\beta^* = \frac{1}{2}(\beta_1^* + \beta_2^*). \tag{2.9}$$

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 $S_x^{(i)}$ 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. Since it is assumed that the two populations have the same covariance matrix Σ, the sample covariance matrices $S_x^{(1)}$ and $S_x^{(2)}$ are pooled to obtain an unbiased estimate of Σ,

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

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

$$c^* = \frac{(\frac{1}{m_1}+\frac{1}{m_2})^{-1} D_x^2}{(\frac{1}{m_1}+\frac{1}{m_2})^{-1} D_x^2 + (\frac{1}{n_1}+\frac{1}{n_2})^{-1} D_y^2}, \tag{2.10}$$

where

$$D_x^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_x^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}), \tag{2.11}$$

$$D_y^2 = (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_y^{-1} (\bar{Y}^{(1)} - \bar{Y}^{(2)}). \tag{2.12}$$

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 (2.10) can be justified in the sense of the training sample sizes of n_i and m_i, and the squared distances of D_x^2 in (2.11) and D_y^2 in (2.12). The values of m_i and D_x^2, for the complete data characterize the performance of W_x in (2.5); while the values of n_i and D_y^2, for the incomplete data characterize the performance of W_y in (2.5). When D_x^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_x^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 (2.10) reduces to

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

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 (2.5) and compare its conditional error rate β^* in (2.9) with the conditional error rate obtained by substituting the parameter estimates into the usual linear discriminant function. Since the distributions of the discriminant functions for the different procedures are intractable, we use a Monte Carlo study to simulate the error rates. We found 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 $\Delta_x^2 = (\mu^{(1)} - \mu^{(2)})'(\mu^{(1)} - \mu^{(2)}) = \mu^{(1)'}\mu^{(2)} = \Delta_y^2 + \Delta_z^2$. So $\Delta_z = \sqrt{\Delta_x^2 - \Delta_y^2}$. Let $R = \Delta_y^2/\Delta_x^2$, where $0 \le R \le 1$. 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 (2.5), Anderson's procedure, the EM algorithm, and Hocking-Smith (AEH) procedure will be simulated as R changes from 0 to 1. < Table 3.1> 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 3.2> and <Table 3.3> for some combinations of p, k, n, m, Δ_x^2, and R. The number of repetitions is 1000. We can see that the three error rates obtained by AEH 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 (1.4) obtained by Anderson's procedure] - [average of conditional error rate β^* in (2.9)].

From <Table 3.2>, 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 ($0 \le R \le 1$) into two regions with $0 \le R \le R^*$ and $R^* < R \le 1$. The linear combination classification procedure is better than AEH 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, Δ_x^2=4 in <Table 3.2>. <Table 3.3> 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 AEH 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 (2.9) 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), Δ_x^2= 1, 4 in <Table 3.1> with R = 0.2, 0.9. The number of repetitions is 1000, and 300 bootstrap samples are generated for each repetition.

<Table 4.1> shows the properties of the bootstrap and leave-one-out estimates for β^* in (2.9). 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 appears 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.

Now we consider the bootstrap confidence interval for the optimal error rate α in (1.3), when the data contain no missing values. The percentile method, bias-corrected percentile method, and accelerated bias-corrected percentile method will be considered. In order to evaluate the properties of the confidence interval for α, a Monte Carlo study is carried out. In this study, bivariate normal random deviates are generated from π_1 : N(0, I) and π_2 : N([Δ_x, 0] ′, I)

by using subroutines in the IMSL, where Δ_x^2 is the Mahalanobis distance. The Monte Carlo study is conducted for the combinations of Δ_x^2= 1, 4, p=2, 5; and equal training sample sizes, $n_1 = n_2 = 20$ and 50 for p=2, and $n_1 = n_2 = 30$ and 50 for p=5. For each combination of sample size, parameter Δ_x and variable p, 500 iterations will be obtained. In each iteration, 5000 bootstrap samples are generated, except for p=5 and n=50, in which case, 1000 bootstrap samples are generated. In order to construct the bootstrap confidence intervals for α, we apply Algorithm AS214 given Buckland (1985). Then the coverage probability and average length of the confidence intervals are computed from the 500 training samples. The bootstrap confidence intervals are compared with the jackknife confidence intervals given in Dorvlo (1992) based on the average length and the coverage probability. <Table 4.2> shows 95% confidence intervals, average lengths and coverage probabilities of the confidence intervals in the case that training samples do not contain missing values.

Now we will extend the bootstrap confidence interval for α to the case that the training samples contain missing values. The jackknife confidence interval is not applicable in this case because of missing values. We will consider the bootstrap confidence interval for the conditional error rate β^* in (2.9) using W_c. The conditional error rate can be estimated by substituting the estimates $\hat{\Sigma}$, $\hat{\mu}^{(i)}$ for Σ, $\mu^{(i)}$ in (2.7) and (2.8). Let $\hat{\mu}^{(i)} = [\ \bar{Y}^{(i)},\ \bar{Z}^{(i)}\ ]'$ in (2.3) and (2.4) be the estimate of $\mu^{(i)}$. For the covariance matrices, let

$$\hat{\Sigma}_{xc}^{(i)} = \begin{bmatrix} \hat{\Sigma}_{yyc}^{(i)} & \hat{\Sigma}_{yyc}^{(i)} \\ \hat{\Sigma}_{yyc}^{(i)} & \hat{\Sigma}_{yyc}^{(i)} \end{bmatrix}$$ 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. Then for $\Sigma^{(i)}$, we suggest the combined estimates,

$$\hat{\Sigma}^{(i)} = \begin{bmatrix} \frac{m_i}{n_i}\hat{\Sigma}_{yyc}^{(i)} + \frac{n_i - m_i}{n_i}\hat{\Sigma}_{yyi}^{(i)} & \hat{\Sigma}_{yzc}^{(i)} \\ \hat{\Sigma}_{xyc}^{(i)} & \hat{\Sigma}_{zzc}^{(i)} \end{bmatrix}.$$

Now the pooled estimate of the covariance matrices is given by

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

We will use these estimates in the construction of the bootstrap confidence intervals for the conditional error rate β^* in (2.9) when the training samples contain missing observations. We generate bivariate normal random deviates from π_1 : N (0, I) and π_2 : N ([$[\Delta_y, \Delta_z]$' , I) by using IMSL subroutines. For each combination of p, k, Δ_x^2, R, n, and m, 500 iterations will be performed, and in each iteration, 1,000 bootstrap samples are generated.

<Table 4.3> shows average lengths and coverage probabilities of the 95 % confidence intervals in the case that training samples contain missing values. The bias-corrected percentile method appears to be reasonable if we consider the coverage probabilities and the average lengths of the confidence intervals compared to those of the other two methods.

5. Concluding Remarks

Discriminant analysis is a multivariate technique concerned with classifying a p x 1 observation X to one of several distinct populations. In this paper, it is assumed that there are two distinct populations which are multivariate normal with equal covariance matrix; that is, π_i : N ($\mu^{(i)}$, Σ). 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, which can be obtained by using AEH procedure. Then the estimates 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 is better than the substitution methods as the proportion of missing observations gets larger. Bootstrap method is a statistical methodology using extensive Monte Carlo simulation (Efron 1982). We use bootstrap method to construct a confidence interval for the error rate in discriminant analysis.

<Table 3.1>. Values of Parameters in the Monte Carlo Study

p	k	n = 20	n = 50	n = 100
2	1	m = 6, 10, 14, 18 Δ_x^2 = .64, 1, 4, 9, 16	m = 6, 10, 14, 18, 30, 46 Δ_x^2 = .64, 1, 4, 9, 16	m = 6, 10, 14, 18, 30, 46, 70, 90 Δ_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	Δ_x^2 = .64, 1, 4, 9, 16	Δ_x^2 = .64, 1, 4, 9, 16	Δ_x^2 = .64, 1, 4, 9, 16
10	1	m = 10, 14, 18 Δ_x^2 = 1, 4	m = 10, 14, 18, 30, 46 Δ_x^2 = 1, 4	m = 10, 14, 18, 30, 46, 70, 90 Δ_x^2 = 1, 4

<Table 3.2>. Comparison of Error Rates

p = 5, k = 1, n = 20, m = 10

Δ_x^2	R	W_c (S. D.)	Anderson* (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.

<Table 3.3>. 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 4.1>. Bootstrap and Leave-one-out Estimates for β^*

p	k	Δ_x^2	n	m	R	$\overline{\beta^*}$	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)

<Table 4.2>. Comparison of 95 % Confidence Interval for α

p	n	Δ_x^2	Optimal Error Rate	Method*	Average Lower Limit	Average Upper Limit	Average Length	Coverage Prob.
2	20	4.0	0.1587	P	0.0619	0.2266	0.1647	86.2
				B	0.0823	0.2528	0.1705	91.2
				A	0.0832	0.2529	0.1697	90.0
				J	0.0643	0.2517	0.1874	91.0
2	50	1.0	0.3085	P	0.2238	0.3670	0.1432	92.8
				B	0.2370	0.3804	0.1434	93.2
				A	0.2399	0.3804	0.1405	92.2
				J	0.2319	0.3816	0.1497	93.4
5	30	4.0	0.1587	P	0.0591	0.1913	0.1322	76.2
				B	0.0925	0.2322	0.1397	92.0
				A	0.1031	0.2323	0.1292	88.8
				J	0.0800	0.2348	0.1548	92.0
5	50	1.0	0.3085	P	0.2001	0.3413	0.1412	83.6
				B	0.2381	0.3825	0.1444	91.8
				A	0.2400	0.3823	0.1423	91.2
				J	0.2313	0.3853	0.1540	94.2

* P = percentile method,
B = bias-corrected percentile method,
A = accelerated bias-corrected percentile method,
J = jackknife method.

<Table 4.3>. Comparison of 95 % Confidence Interval for β^*

p = 2						Average Lower	Average Upper	Average	Coverage
n	m	R	Δ_x^2	$\overline{\beta^*}$	Method*	Limit	Limit	Length	Prob.
20	10	0.8	4	0.1748	P	0.0447	0.2239	0.1792	79.0
					B	0.0827	0.2766	0.1939	92.6
					A	0.0914	0.2775	0.1861	76.4
20	18	0.8	1	0.3229	P	0.1479	0.3687	0.2208	79.0
					B	0.1902	0.4082	0.2180	93.0
					A	0.1937	0.4081	0.2144	78.0
50	20	0.3	4	0.1776	P	0.0717	0.2347	0.1630	88.6
					B	0.0883	0.2529	0.1646	93.6
					A	0.0978	0.2531	0.1553	87.8
50	46	0.3	1	0.3152	P	0.2232	0.3672	0.1440	89.8
					B	0.2404	0.3843	0.1439	94.0
					A	0.2406	0.3843	0.1437	89.0

* P = percentile method,
B = bias-corrected percentile method,
A = accelerated bias-corrected percentile method.

References

[1] Anderson, T. W. (1951). Classification by multivariate analysis, *Psychometrika*, 16, 31-50.

[2] Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association*, 52, 200-203.

[3] Bohannon, Tom R. and Smith, W. B. (1975). *ASA Proceedings of Social Statistics Section*, 214-218.

[4] Buckland, Stephen T. (1985). Calculation of Monte Carlo confidence intervals, *Royal Statistical Society*, Algorithm AS214, 297-301.

[5] Chan, L. S. and Dunn, O. J. (1972). The treatment of missing values in discriminant analysis-1, The sampling experiment, *Journal of the American Statistical Association*, 67, 473-477.

[6] Chan, L. S. and Dunn, O. J. (1974). A note on the asymptotical aspect of the treatment of missing values in discriminant analysis, *Journal of the American Statistical Association*, 69, 672-673.

[7] Dorvlo, Atsu S.S. (1992). An interval estimation of the probability of misclassification, *Journal of Mathematical Analysis and Application*, 171, 389-394.

[8] Efron, B. (1982), The jackknife, the bootstrap, and other resampling plans, *CBMS-NSF Regional Conference Series in Applied Mathematics, 38. Society for Industrial and Applied Mathematics (SIAM)*, Philadelphia.

[9] Hocking, R. R. and Smith, W. B.(1968). Estimation of parameters in the mutivariate normal distribution with missing observation, *Journal of the American Statistical Association*, 63, 159-173.

[10] McLachlan, G. J.(1980). The efficiency of Efron's bootstrap approach applied to error rate estimation in discriminant analysis, *Journal of Statistical Computation and Simulation*, 11, 273-279.

[11] Twedt, Daniel J. and Gill, D. S. (1992). Comparison of algorithm for replacing missing data in discriminant analysis, *Communications in Statistics-Theory and Methods*, 21, 1567-1578.