Communications of the Korean Statistical Society
2009, Vol. 16, No. 4, 675–686

Bootstrap Confidence Intervals of Classification Error Rate for a Block of Missing Observations

Hie-Choon Chung[1,a]

[a]Department of e-Business, Gwangju University

Abstract

In this paper, it will be assumed that there are two distinct populations which are multivariate normal with equal covariance matrix. We also assume that the two populations are equally likely and the costs of misclassification are equal. The classification rule depends on the situation when the training samples include missing values or not. We consider the bootstrap confidence intervals for classification error rate when a block of observation is missing.

Keywords: Bootstrap Confidence interval, error rate, block of missing observations, linear combination classification statistic, Jackknife method, Monte Carlo study.

1. Introduction

In discriminant analysis the problem is to classify a $p \times 1$ observation X of unknown origin to one of several distinct populations using an appropriate classification rule. In this paper it will be assumed that there are two distinct populations which are multivariate normal with equal covariance matrix. We also assume that the two populations are equally likely and the costs of misclassification are equal. The classification rule depends on the situation when the training samples include missing values or not.

1.1. Discriminant analysis with complete data

If the population π_i has density $f_i(X)$, $i = 1, 2$, the Bayes procedure (see Section 6.2, Anderson, 1984) 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,

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

This research is supported by Gwangju University, 2009.

[1] Associate Professor, Department of e-Business, Gwangju University, Jinwol-Dong 592, Nam-Gu, Gwangju 503-703, Korea. E-mail: hcc@gwangju.ac.kr.

Then the random variable $U = [X - 1/2(\mu^{(1)} + \mu^{(2)})]'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ 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 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

$$P(1|2) = \Pr(U \geq 0|X \in \pi_2) = \Phi\left(-\frac{\Delta}{2}\right)$$

Then the optimal error rate(see Equation 11–27, Johnson and Wichern, 2002) is defined as

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

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

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

is called Anderson's classification statistic. The error rate corresponding to this classification rule is called the unconditional error rate, which is $\gamma = 1/2[\Pr(W < 0|X \in \pi_1) + \Pr(W \geq 0|X \in \pi_2)]$.

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

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

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

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

Hence the conditional error rate (see Section 6.3, Anderson, 1984) is

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

It is clear that the expectation of the conditional error rate is simply the unconditional error rate. Three error rates are used to judge the performance of a classification rule. Since the three error rates, *i.e.*, the optimal error rate, the unconditional error rate and the conditional error rate, are all functions of unknown parameters, they need to be estimated. Estimation of error rates has received considerable attention since the 1930s. There are several methods for estimating the error rate given

in the literature. The plug-in (Fisher, 1936) estimator is obtained by substituting unbiased estimates, $\bar{X}^{(1)}$, $\bar{X}^{(2)}$ and S for $\mu^{(1)}$, $\mu^{(2)}$ and Σ into (1.4). Then the estimator for γ^* in (1.4) is given by

$$\hat{\gamma}^* = \Phi\left(-\frac{D}{2}\right), \tag{1.5}$$

where $D^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})' S^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)})$ is the sample analog of Mahalanobis squared distance Δ^2. We can obtain the same expression,

$$\hat{\alpha} = \Phi\left(-\frac{D}{2}\right) \tag{1.6}$$

by substituting the estimate D for Δ directly into the optimal error rate α in (1.3). Hence this plug-in estimator can be used to estimate both the optimal error rate and conditional error rate.

1.2. Discriminant analysis with incomplete data

When the training samples contain incomplete observation vectors, there are several methods of dealing with missing values in discriminant analysis. One is to ignore these incomplete observation vectors in the construction of a classification rule. But this method is usually ineffective since information has been lost. Other methods (Chan and Dunn, 1972, 1974; Bohannon and Smith, 1975; Twedt and Gill, 1992; Anderson, 1957) incorporate these incomplete observation vectors in the construction of the classification rule and the estimation of the error rate.

In this paper 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,

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

are available from

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

and random samples of sizes $n_i - m_i$, which contain only the first k-components $Y_{j\,(k\times 1)}^{(i)}$, $i = 1,2$; $j = m_i + 1,\ldots,n_i$, 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. We can construct two linear discriminant functions.

The first linear discriminant function is based on the observations, $X_j^{(i)}$, $i = 1,2$; $j = 1,\ldots,m_i$. We have

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

where $\bar{X}^{(i)} = 1/m_i \sum_{j=1}^{m_i} X_j^{(i)} = \begin{bmatrix} Y_1^{(i)} \\ Z^{(i)} \end{bmatrix}$,

$$\bar{Y}_1^{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_j^{(i)}, \quad i = 1,2, \qquad \bar{Z}^{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} Z_j^{(i)}, \quad i = 1,2, \tag{1.7}$$

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

The second linear discriminant function is based on the incomplete observations, $\bar{Y}_{j\,(k\times 1)}^{(i)}$, $i = 1,2$; $j = 1,2,\ldots,n_i$. We have $W_y = (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1}[Y - 1/2(\bar{Y}^{(1)} + \bar{Y}^{(2)})]$, where

$$\bar{Y}^{(i)} = \frac{1}{n_i}\left[m_i \bar{Y}_1^{(i)} + (n_i - m_i)\bar{Y}_2^{(i)}\right], \tag{1.8}$$

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

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

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

$$W_c = cW_x + (1-c)W_y, \quad 0 \le c \le 1. \tag{1.9}$$

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.

Let $W_x = \boldsymbol{a}'X + b$, where $\boldsymbol{a}'_{(1\times p)} = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1}$, $b = -1/2(\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1}(\bar{X}^{(1)} + \bar{X}^{(2)})$. Also let $W_y = \boldsymbol{d}'Y + e$, where $\boldsymbol{d}' = (\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1}$, $e = -1/2(\bar{Y}^{(1)} - \bar{Y}^{(2)})' S_{yy}^{-1}(\bar{Y}^{(1)} + \bar{Y}^{(2)})$. Then $W_c = c(\boldsymbol{a}_1'Y + \boldsymbol{a}_2'Z + b) + (1-c)(\boldsymbol{d}'Y + e) = A'Y + B'Z + F = H'X + F$, where $A = c\boldsymbol{a}_1 + (1-c)\boldsymbol{d}$, $B = c\boldsymbol{a}_2$, $F = cb + (1-c)e$, $H = \begin{bmatrix} A_{(k\times 1)} \\ B_{(p-k)\times 1} \end{bmatrix}$. 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 $N(H'\mu^{(i)} + F, H'\Sigma H)$, $i = 1,2$. Then the conditional probability of misclassifying an observation X from π_1 by π_2 is W_c given by

$$\beta_1^* = \Phi\left(-\frac{H'\mu^{(1)} + F}{\sqrt{H'\Sigma H}}\right). \tag{1.10}$$

similarly,

$$\beta_2^* = \Phi\left(\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}}\right). \tag{1.11}$$

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

$$\beta^* = \frac{1}{2}(\beta_1^* + \beta_2^*). \tag{1.12}$$

Using the linear combination classification statistic in (1.9), 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 operational c^* which is given by

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

where $D^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$, $D_y^2 = (\bar{Y}^{(1)} - \bar{Y}^{(2)})'S_y^{-1}(\bar{Y}^{(1)} - \bar{Y}^{(2)})$.

From the simulations, Chung *et al.* (2000) showed that the linear combination classification is better than Anderson's procedure (Anderson, 1957), the EM algorithm (Dempster *et al.*, 1977) and Hocking and Smith procedure (Hocking and Smith, 1968) as the proportion of missing observation gets larger.

In this paper, we propose to construct confidence intervals of the error rates using a bootstrap method. Bootstrap confidence intervals of those are compared to the jackknife confidence interval derived by Dorvlo (1992). Then the real data sets illustrate the application of the bootstrap method.

2. Bootstrap Confidence Interval for the Error Rate When Training Samples Do Not Contain Missing Values

The usual confidence intervals are based on an asymptotic approximation that can be quite inaccurate in practice (see Buckland, 1983; Diciccio and Efron, 1996). However, the bootstrap confidence intervals can be applied to more realistic situations.

In this section, we consider the bootstrap confidence interval for the optimal error rate in (1.3), when the data contain no missing values. Then the bootstrap confidence interval of the error rate will be extended to the case that the data consist of missing observation in Section 3. Another method to construct a confidence interval for α is jackknife method which is described as follows.

2.1. Jackknife confidence interval

Dorvlo (1992) considered an interval estimator based on the jackknife method of estimating the optimal error rate for W, when the training samples have no missing values. He proposed the jackknife estimator $\hat{\alpha}_1$, defined as

$$\hat{\alpha}_1 = nf(\hat{\beta}) - \frac{n-1}{n}\sum_{j=1}^{n} f(\hat{\beta}_j), \tag{2.1}$$

where $n = n_1 + n_2$, n_1 and n_2 are the sample sizes taken from populations $\pi_1 : N(\mu^{(1)}, \Sigma)$ and $\pi_2 : N(\mu^{(2)}, \Sigma)$ respectively, and

$$\begin{aligned}
\hat{\beta} &= \bar{X}_1 - \bar{X}_2, \\
\hat{\beta}_j &= \bar{X}_{1j} - \bar{X}_2, \quad j = 1, \ldots, n_1, \\
&= \bar{X}_1 - \bar{X}_{2j}, \quad j = n_1 + 1, \ldots, n, \\
\bar{X}_{1j} &= \frac{n_1 \bar{X}_1 - X_j}{n_1 - 1}, \quad j = 1, \ldots, n_1, \\
\bar{X}_{2j} &= \frac{n_2 \bar{X}_2 - X_j}{n_2 - 1}, \quad j = n_1 + 1, \ldots, n, \\
f(\hat{\beta}) &= \Phi\left[-\frac{1}{2}\left(\hat{\beta}' \Sigma^{-1} \hat{\beta}\right)^{\frac{1}{2}}\right], \\
f(\hat{\beta}_j) &= \Phi\left[-\frac{1}{2}\left(\hat{\beta}_j' \Sigma^{-1} \hat{\beta}_j\right)^{\frac{1}{2}}\right], \quad j = 1, \ldots, n,
\end{aligned}$$

where Φ denotes the cumulative standard normal distribution. Here $\bar{X}_{1j}$ $(j = 1, \ldots, n_1)$ and $\bar{X}_{2j} (j = n_1 + 1, \ldots, n)$ denote the sample means obtained by deleting the j^{th} observation $(j = 1, \ldots, n)$. Let S_{1j} and S_{2j} denote the corresponding covariance matrices based on $(n - 3)$ degrees of freedom, and S denotes the covariance matrix based on $(n - 2)$ degrees of freedom. Also let

$$\hat{\alpha}_{1j} = n f(\hat{\beta}) - (n-1) f(\hat{\beta}_j), \quad j = 1, \ldots, n.$$

Then we can replace Σ^{-1} in the expression of $f(\hat{\beta})$ and $f(\hat{\beta}_j)$ by S^{-1} and S_{ij}^{-1} $(i = 1, j = 1, \ldots, n_1; i = 2, j = n_1 + 1, \ldots, n)$ respectively, since those tend to Σ^{-1} in the limit. Dorvlo (1992) concluded that the interval estimate of α could be written as

$$\left\{\hat{\alpha}_1 - t_{\frac{\eta}{2}} \sqrt{\frac{\sum_{j=1}^{n} \left(\hat{\alpha}_{1j} - \hat{\alpha}_1\right)^2}{n(n-1)}},\ \hat{\alpha}_1 + t_{\frac{\eta}{2}} \sqrt{\frac{\sum_{j=1}^{n} \left(\hat{\alpha}_{1j} - \hat{\alpha}_1\right)^2}{n(n-1)}}\right\},$$

where $t_{\eta/2}$ denotes the $100(1-\eta/2)$ percentage point of the t-distribution with $n-1$ degrees of freedom.

2.2. Bootstrap confidence interval

Now we consider the bootstrap confidence interval for the optimal error rate α in (1.3), when the training samples contain no missing values. The bootstrap method is a resampling technique using Monte Carlo simulation (Efron 1982). In our situation, independent random samples of sizes n_1 and n_2 with replacement are taken from the two training samples respectively. An estimator $\hat{\alpha}^*$ of α based on the bootstrap sample is obtained by using (1.6). This process is repeated independently a large number B of times. Then bootstrap confidence interval for α can be obtained from the B values of $\hat{\alpha}^*$. Let $\hat{\alpha}^*_{(i)}$ denote the i^{th} ordered value, so that

$$\hat{\alpha}^*_{(1)} \leq \hat{\alpha}^*_{(2)} \leq \cdots \leq \hat{\alpha}^*_{(B)}.$$

There are several methods to construct the bootstrap confidence interval. We will consider the percentile method, bias-corrected percentile method, accelerated bias-corrected percentile method

to construct the confidence interval (Efron, 1982, 1987; Buckland, 1983, 1984, 1985; Hall, 1986a, 1986b; Hinkley, 1988; DiCiccio and Romano, 1988; among others). These three types of $100(1-2\eta)\%$ confidence interval are presented as follows:

Percentile method. The confidence interval is given by $(\hat{\alpha}^*_{(r)}, \hat{\alpha}^*_{(s)})$, where $r = (B+1)\eta$ and $s = (B+1)(1-\eta)$, both rounded to nearest integer, subject to $r + s = B + 1$.

Bias-corrected percentile method. Suppose $\hat{\alpha}^*_{(q)} < \hat{\alpha} < \hat{\alpha}^*_{(q+1)}$, where $\hat{\alpha}$ is calculated from the original samples. That is, q of the B bootstrap estimates for α are smaller than $\hat{\alpha}$. Define

$$z_o = \Phi^{-1}\left(\frac{q}{B}\right), \quad \eta_{BL} = \Phi(2z_o - z_\eta) \text{ and } \eta_{BR} = \Phi(2z_o + z_\eta),$$

where $\Phi(z_\eta) = 1 - \eta$ and Φ denotes the cumulative standard normal distribution. Then the confidence interval is given by $(\hat{\alpha}^*_{(j)}, \hat{\alpha}^*_{(k)})$, where $j = (B+1)\eta_{BL}$ and $k = (B+1)\eta_{BR}$.

Accelerated bias-corrected percentile method. Define

$$\eta_{AL} = \Phi\left(z_o + \frac{z_o - z_\eta}{1 - a(z_o - z_\eta)}\right) \text{ and } \eta_{AR} = \Phi\left(z_o + \frac{z_o + z_\eta}{1 - a(z_o + z_\eta)}\right),$$

where $a = 1/6[\sum_{i=1}^{B}(\hat{\alpha}^*_i - \bar{\hat{\alpha}}^*)^3/[\sum_{i=1}^{B}(\hat{\alpha}^*_i - \bar{\hat{\alpha}}^*)^2]^{3/2}]$, which is called the acceleration constant, and $\bar{\hat{\alpha}}^*$ is the mean of the B bootstrap estimates for $\hat{\alpha}^*_j$, $i = 1, \ldots, B$.

Then the confidence interval is given by $(\hat{\alpha}^*_{(u)}, \hat{\alpha}^*_{(v)})$, where $u = (B+1)\eta_{AL}$ and $v = (B+1)\eta_{AR}$. Note that η_{AR} and η_{AL} become η_{BR} and η_{BL} if a equals 0. If z_0 is zero, then η_{BR} and η_{BL} become η.

In order to evaluate the properties of the confidence interval for α, a Monte Carlo study is proposed. In this study, bivariate normal random deviates are generated from $\pi_1 : N(0, I)$ and $\pi_2 : N([\Delta_x, 0]', I)$ by using subroutine in the International Mathematical and Statistical Library(IMSL), where Δ_x^2 is the Mahalanobis distance. For each Monte Carlo study, 500 iterations will be obtained. In each iteration, $B = 5000$ bootstrap samples are generated. Then the bootstrap confidence intervals for α are obtained from the B values of $\hat{\alpha}^*$ which is an estimator of α based on the bootstrap sample by D method which is suggested by Fisher (1936). In order to construct the bootstrap confidence intervals for α, we apply Algorithm AS214 given in Buckland (1985). Then the coverage probability and average length of the confidence intervals are computed. The average length is computed by subtracting the average lower limit for the confidence interval of conditional error rate from the average upper limit for it, whose average limit are obtained by taking average of the 500 lower limits and the 500 upper limits respectively. The coverage probability is also considered from the 500 training samples, in which the conditional error rate is checked whether it is between the lower limit and the upper limit for each training sample. The coverage probability is obtained by dividing the number covered by both limits by 500. The bootstrap confidence intervals are compared with the jackknife confidence interval given in Dorvlo (1992) based on the average length and coverage probability.

From the Table 1, we recommend both the bias-corrected percentile method and the jackknife method to obtain the confidence interval for α in (1.3).

3. Bootstrap Confidence Interval When Training Samples Contain Missing Values

We will extend the bootstrap confidence interval for α to the case that the training samples contain missing values. We will not consider the jackknife confidence interval in this case since the jackknife method does not improve the bootstrap method when training samples do not contain missing values.

Table 1: 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	1.0	0.3085	P	0.1685	0.3872	0.2186	89.2
				B	0.2028	0.4156	0.2128	88.8
				A	0.2114	0.4156	0.2042	88.4
				J	0.1871	0.4350	0.2479	91.8
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
2	50	4.0	0.1587	P	0.0968	0.2089	0.1121	92.8
				B	0.1063	0.2202	0.1139	95.0
				A	0.1062	0.2202	0.1140	95.8
				J	0.0992	0.2174	0.1182	94.0
5	30	1.0	0.3085	P	0.1561	0.3337	0.1775	69.8
				B	0.2168	0.3911	0.1743	90.4
				A	0.2239	0.3911	0.1672	90.4
				J	0.2043	0.4086	0.2043	94.4
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
5	50	4.0	0.1587	P	0.0840	0.1929	0.1089	85.6
				B	0.1060	0.2203	0.1143	92.6
				A	0.1097	0.2205	0.1108	91.2
				J	0.0987	0.2179	0.1192	93.8

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

We will consider the bootstrap confidence interval for the conditional error rate β^* in (1.12) using W_c. The conditional error rate can be estimated by substituting the estimates $\hat{\Sigma}$, $\hat{\mu}^{(i)}$ for Σ, $\mu^{(i)}$ in (1.10) and (1.11) respectively. Let $\hat{\mu}^{(i)} = [\bar{Y}^{(i)}, \bar{Z}^{(i)}]'$ be the estimate of $\mu^{(i)}$ from (1.7) and (1.8). 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, $i = 1, 2$. Then we suggest the combined estimates,

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

for $\Sigma^{(i)}$, $i = 1, 2$. Now the pooled estimate of the covariance matrices is given by

Table 2: Comparison of 95% Confidence Interval for β^*

$p=2, k=1$						Average	Average	Average	Coverage
n	m	R	Δ_x^2	$\bar{\beta}^*$	Method*	Lower Limit	Upper Limit	Length	Prob.
					P	0.1009	0.3426	0.2417	64.6
20	10	0.8	1	0.3268	B	0.1538	0.3926	0.2388	84.4
					A	0.1626	0.3922	0.2296	65.8
					P	0.0447	0.2239	0.1792	79.0
20	10	0.8	4	0.1748	B	0.0827	0.2766	0.1939	92.6
					A	0.0914	0.2775	0.1861	76.4
					P	0.1479	0.3687	0.2208	79.0
20	18	0.8	1	0.3229	B	0.1902	0.4082	0.2180	93.0
					A	0.1937	0.4081	0.2144	78.0
					P	0.0599	0.2209	0.1650	80.4
20	18	0.8	4	0.1698	B	0.0877	0.2734	0.1857	92.0
					A	0.0934	0.2746	0.1812	78.8
					P	0.0717	0.2347	0.1630	88.6
50	20	0.3	4	0.1776	B	0.0883	0.2529	0.1646	93.6
					A	0.0978	0.2531	0.1553	87.8
					P	0.0882	0.2152	0.1270	87.6
50	20	0.8	4	0.1712	B	0.1091	0.2365	0.1274	94.0
					A	0.1108	0.2365	0.1257	87.2
					P	0.2232	0.3672	0.1440	89.8
50	46	0.3	1	0.3152	B	0.2404	0.3843	0.1439	94.0
					A	0.2406	0.3843	0.1437	89.0
					P	0.2215	0.3694	0.1479	89.6
50	46	0.8	1	0.3183	B	0.2364	0.3842	0.1478	93.8
					A	0.2367	0.3842	0.1476	87.8

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

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

We will use these estimates in the construction of the bootstrap confidence intervals for the conditional error rate β^* in (1.12) when the training samples contain missing observations. Basically the same procedure described for α is applied in this situation for getting the three types of the bootstrap confidence intervals for β^*, *i.e.*, the percentile method, the bias-corrected percentile method, and the accelerated bias-corrected method. In order to evaluate the properties of the confidence intervals for β^*, we conduct a similar Monte Carlo study described for the optimal error rate, α in (1.3).

Basically the same procedure described for α is applied in this situation for getting the three types of the bootstrap confidence intervals for β^*. We generated bivariate normal random deviates from $\pi_1 : N(0, I)$ and $\pi_2 : N([\Delta_y, \Delta_z]', I)$ by using IMSL subroutines, where Δ_y^2 and Δ_z^2 are Mahalanobis distance based on the variable Y and the variable Z respectively. Note that

$$\Delta_x^2 = \Delta_y^2 + \Delta_z^2 \text{ for } X = [X, Z]', \quad R = \Delta_y^2/\Delta_x^2 \text{ where } 0 \le R \le 1.$$

For each Monte Carlo study, 500 iterations will be obtained. In each iteration, $B = 1000$ bootstrap samples are generated.

From the Table 2, the bias-corrected percentile method appears to be resonable compared to those of the other two methods.

Table 3: Population 1: Success

X_1	X_2	X_3	X_4	X_5
2.97	420	800	600	497
3.80	330	710	380	563
2.50	270	700	340	510
2.50	400	710	600	563
3.30	280	800	450	543
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	

X_1 = Undergraduate GPA; X_2= GRE Verbal; X_3= GRE Quantitative; X_4= GRE Analytic; X_5 = TOEFL Score

Table 4: Population 2: Failure

X_1	X_2	X_3	X_4	X_5
3.75	250	730	460	513
3.11	320	760	610	560
3.00	360	720	525	540
2.60	370	780	500	500
3.50	300	630	380	507
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	

X_1 = Undergraduate GPA; X_2= GRE Verbal; X_3= GRE Quantitative; X_4= GRE Analytic; X_5 = TOEFL Score

4. Numerical Example

Application of the bootstrap method to estimate the error rate, β^* in (1.12) is illustrated by using real data sets. They are given by the Admissions Office at the University of Texas at Arlington. The data sets contain two populations, which are shown in Table 3 and Table 4. One population is the Success Group that the students receive their masters'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.

Using this data set, we obtain the discriminant function

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

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

For this example, we generate 300 bootstrap samples to estimate β^*, and 1,000 bootstrap samples to construct the bootstrap confidence interval for β^*. The result of using $c^* = 0.7532$ is that the bootstrap estimate of β^* is 0.3435. The 95% confidence interval of β^* is (0.2721, 0.4609) which is obtained by the bias-corrected percentile method.

5. Conclusion

Discriminant analysis is a multivariate technique concerned with classifying a $p \times 1$ observation X to one of several distinct populations using an appropriate classification rule. The classification rule depends on the situation when the training samples include missing values or not. In this paper, we consider the situation that the training samples contain incomplete observation vectors which have a pattern of missing data; *i.e.*, all missing values occur on the same variables. In this situation, we use the discriminant function which is a linear combination of two well defined Fisher's linear discriminant functions. The performance of a classification procedure is evaluated by its error rate which depends on unknown parameters. For the situation, we consider the bootstrap confidence interval for the conditional error rate β^* in (1.12) using W_c. We recommend the bias-corrected method. A numerical example is given and it is shown that the linear combination classification procedure is easy to use for the incomplete case.

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, *Journal of the American Statistical Association*, **52**, 200–203.

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

Bohannon, T. R. and Smith, W. B. (1975). ASA Proceedings of Social Statistics Section, 214–218.

Buckland, S. T. (1983). Monte Carlo methods for confidence interval estimation using the bootstrap technique, *Bias*, **10**,194–212.

Buckland, S. T. (1984). Monte Carlo confidence intervals, *Biometrics*, **40**, 811–817.

Buckland, S. T. (1985). Calculation of Monte Carlo confidence intervals, *Royal Statistical Society*, Algorithm AS214, 297–301.

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.

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.

Chung, H. C. and Han, C. P. (2000). Discriminant analysis when a block of observations is missing, *Annals of the Institute of Statistical Mathematics*, **52**, 544–556.

Dempster, A. P., Laird, N. M. and Rubin, R. J. A. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 302–306.

Diciccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals, *Statistical Science*, **11**, 189–228.

DiCiccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals, *Journal of the Royal Statistical Society, Series B*, **50**, 338–354.

Dorvlo, A. S. S. (1992). An interval estimation of the probability of misclassification, *Journal of Mathematical Analysis and Application*, **171**, 389–394.

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.

Efron, B. (1987). Better bootstrap confidence intervals, *Journal of the American Statistical Association*, **82**, 171–200.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188.

Hall, p. (1986a). On the bootstrap and confidence intervals, *Annals of Statistics*, **14**, 1431–1452.

Hall, P. (1986b). On the number of bootstrap simulations required to construct a confidence interval, *Annals of Statistics*, **14**, 1453–1462.

Hinkley, D. V. (1988). Bootstrap methods, *Journal of the Royal Statistical Society, Series B*, **50**, 321–337.

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.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall

Twedt, D. 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.

Received March 2009; Accepted May 2009