Journal of the Korean Data & Information Science Society
2013, **24**(1), 189–200

http://dx.doi.org/10.7465/jkdi.2013.24.1.189
한국데이터정보과학회지

Conditional bootstrap confidence intervals for classification error rate when a block of observations is missing†

Hie-Choon Chung[1] · Chien-Pai Han[2]

[1]Department of Healthcare Management, Gwangju University
[2]Department of Mathematics, University of Texas at Arlington

Received 29 November 2012, revised 8 January 2013, accepted 17 January 2013

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 whether the training samples include missing values or not. We consider the conditional bootstrap confidence intervals for classification error rate when a block of observation is missing.

Keywords: Block of observations missing, conditional bootstrap confidence interval, error rate, jackknife method, linear combination classification statistic, 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.

When there are two normal populations, the Bayes procedure classifies X into π_1 if

$$\left[X - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)\right]' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) \geq 0, \tag{1.1}$$

otherwise X is classified into π_2. Then the optimal error rate is defined as

$$\alpha = \Phi\left(-\triangle/2\right), \tag{1.2}$$

where $\triangle^2 = \left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$ is the Mahalanobis squared distance between the two populations, and Φ denotes the cumulative distribution function of the univariate standard normal distribution. Anderson (1951) suggested the method of simple substitution of

† This study was conducted by research funds from Gwangju University in 2012.
[1] Professor, Department of Healthcare Management, Gwangju University, Gwangju 503-703, South Korea. Email: hcc@gwangju.ac.kr
[2] Professor, Department of Mathematics, University of Texas at Arlington, Texas, U.S.A.

$\overline{X}^{(i)}$ for and $\mu^{(i)}$ and the pooled sample covariance matrix S for Σ in (1.1), where $\overline{X}^{(i)}$ and S are the usual unbiased estimators of $\mu^{(i)}, i = 1, 2$, and Σ, respectively. The statistic $W = \left[X - (1/2)\left(\overline{X}^{(1)} + \overline{X}^{(2)}\right)\right]' S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)$ is called Anderson's classification statistic, which is called W statistic. The error rate corresponding to this classification rule is called the unconditional error rate, which is

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

Since the exact expression for the unconditional error rate is very complicated, the conditional error rate is considered by assuming, $\overline{X}^{(1)},\ \overline{X}^{(2)}$ and S fixed. The conditional probability of misclassifying an observation X from π_1 into π_2 by W is

$$\begin{aligned} P_1 &= Pr\left(W < 0 | \overline{X}^{(1)},\ \overline{X}^{(2)},\ S; X \in \pi_1\right) \\ &= \Phi\left\{\frac{\frac{1}{2}\left(\overline{X}^{(1)} + \overline{X}^{(2)}\right)' S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right) - \mu^{(1)\prime} S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)}{\sqrt{\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)' S^{-1}\Sigma S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)}}\right\}. \end{aligned} \tag{1.3}$$

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

$$\begin{aligned} P_2 &= Pr\left(W \geq 0 | \overline{X}^{(1)},\ \overline{X}^{(2)},\ S; X \in \pi_2\right) \\ &= \Phi\left\{\frac{-\frac{1}{2}\left(\overline{X}^{(1)} + \overline{X}^{(2)}\right)' S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right) + \mu^{(2)\prime} S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)}{\sqrt{\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)' S^{-1}\Sigma S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)}}\right\}. \end{aligned} \tag{1.4}$$

Hence the conditional error rate is

$$\gamma^* = \frac{1}{2}(P_1 + P_2) \tag{1.5}$$

Since these three error rates are all functions of unknown parameters, they need to be estimated. The plug-in estimator (Fisher, 1936) is obtained by substituting unbiased estimates, $\overline{X}^{(1)},\ \overline{X}^{(2)}$ and S for $\mu^{(1)}$, $\mu^{(2)}$ and Σ into (1.3) and (1.4), which is called D method. Then the estimator for γ^*in (1.5) is given by

$$\widehat{\gamma}^* = \Phi(-\frac{\widehat{\triangle}}{2}), \tag{1.6}$$

where $\widehat{\triangle}^2 = \left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)' S^{-1}\left(\overline{X}^{(1)} - \overline{X}^{(2)}\right)$ is the sample analog Mahalanobis squared distance $\triangle^2$. We can obtain the same expression,

$$\widehat{\alpha} = \Phi(-\frac{\widehat{\triangle}}{2}) \tag{1.7}$$

by substituting the estimate $\widehat{\triangle}$ for $\triangle$ directly into the optimal error rate α in (1.2). Hence this plug-in estimator can be used to estimate both the optimal error rate and conditional error rate.

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; Anderson,1957; Twedt and Gill, 1992) 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,

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

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 $Y_{j(k\times 1)}^{(i)}$, $i = 1,2; j = m_i+1,\cdots,n_i$, are available from $N_k(\mu_y^{(i)},\Sigma_{yy})$. We denote by $X_j^{(i)}, i = 1,2; j = 1,\cdots,m_i$, the complete observations, and by $Y_j^{(i)}, i = 1,2; j = 1,\cdots,n_i$, the incomplete observations. Hence the data have the special pattern of missing values where a block of variables is missing on 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, $X_j^{(i)}, i = 1,2; j = 1,\cdots,m_i$. We have

$$W_x = (\overline{X}^{(1)} - \overline{X}^{(2)})' \ S_{xx}^{-1}[X - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)})], \tag{2.1}$$

where

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

$$Y_1^{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_j^{(i)}, \overline{Z}^{(i)} = \frac{1}{m_i}\sum_{j=1}^{m_i} Z_j^{(i)} = \begin{bmatrix} \overline{Y}_1^{(i)} \\ \overline{Z}^{(i)} \end{bmatrix}, \ i = 1,2,$$

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

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

$$W_y=\left(\overline{Y}^{(1)}-\overline{Y}^{(2)}\right)'\ S_{yy}^{-1}[Y-\frac{1}{2}\left(\overline{Y}^{(1)}+\overline{Y}^{(2)}\right)], \tag{2.2}$$

where

$$\overline{Y}^{(i)}=\frac{1}{n_i}[m_i\overline{Y}_1^{(i)}+(n_i-m_i)\ \overline{Y}_2^{(i)}],$$

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

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

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

$$W_c=cW_x+(1-c)W_y, 0\le c\le 1. \tag{2.3}$$

We call this 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 $W_x=a'X+b$, where

$$a'_{(1\times p)}=\left(\overline{X}^{(1)}-\overline{X}^{(2)}\right)'\ S_{xx}^{-1},\ b=\frac{-1}{2}(\overline{X}^{(1)}-\overline{X}^{(2)})'\ S_{xx}^{-1}(\overline{X}^{(1)}+\overline{X}^{(2)}).$$

Also let $W_y=d'Y+e$, where

$$d'=\left(\overline{Y}^{(1)}-\overline{Y}^{(2)}\right)'\ S_{yy}^{-1},\ e=\frac{-1}{2}\left(\overline{Y}^{(1)}-\overline{Y}^{(2)}\right)'\ S_{yy}^{-1}\left(\overline{Y}^{(1)}+\overline{Y}^{(2)}\right).$$

Then

$$W_c=c\left(a_1'Y+a_2'Z+b\right)+(1-c)\left(d'Y+e\right)=A'Y+B'Z+F\ =\ H'X+F,$$

where

$$A=ca_1+(1-c)\,d,\ B=ca_2,\ F=cb+(1-c)e,\ H=\begin{bmatrix}A_{(k\times 1)}\\ B_{(p-k)\times 1}\end{bmatrix}.$$

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

$$\beta_1^*=\Phi\left(-\frac{H^{'}\mu^{(1)}+F}{\sqrt{H'\Sigma H}}\right), \tag{2.4}$$

and similarly we have

$$\beta_2^*=\Phi\left(\frac{H'\mu^{(2)}+F}{\sqrt{H'\Sigma H}}\right) \tag{2.5}$$

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

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

Using the linear combination classification statistic in (2.3), 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

$$c^* = \frac{(\frac{1}{m_1}+\frac{1}{m_2})^{-1}D^2}{(\frac{1}{m_1}+\frac{1}{m_2})^{-1}D^2+(\frac{1}{m_1}+\frac{1}{m_2})^{-1}D_y^2},$$

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

From the simulations, Chung and Han (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 lager.

In this paper, we propose to construct confidence intervals of the error rates, β^*in (2.6) 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.

3. Conditional bootstrap confidence interval

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 conditional bootstrap confidence interval for the conditional error rate γ^* in (1.5), when the data contain no missing values. Then the conditional bootstrap confidence interval of the error rate will be extended to the case that the data consist of missing observation in Section 4. Another method to construct a confidence interval for γ^* is conditional jackknife method which is described as follows.

3.1. Conditional 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 $\widehat{\alpha}_1$ defined as

$$\widehat{\alpha}_1 = nf\left(\widehat{\beta}\right) - \frac{n-1}{n}\sum_{j=1}^{n} f(\widehat{\beta}_j), \tag{3.1}$$

where

$$\begin{aligned}
\widehat{\beta} &= \overline{X}_1 - \overline{X}_2,\\
\widehat{\beta}_j &= \begin{cases} \overline{X}_{1j} - \overline{X}_2, \ j = 1, \cdots, n_1,\\ \overline{X}_1 - \ \overline{X}_{2j}, \ j = n_1{+}1, \cdots, n, \end{cases}\\
\overline{X}_{1j} &= (n_1\overline{X}_1{-}X_j)/(n_1{-}1), \ j = 1, \cdots, n_1,\\
\overline{X}_{2j} &= (n_2\overline{X}_2{-}X_j)/(n_2{-}1), \ j = n_1{+}1, \cdots, n,\\
F(\widehat{\beta}) &= [-\frac{1}{2}(\widehat{\beta}'\Sigma^{-1}\widehat{\beta})^{\frac{1}{2}}],\\
f(\widehat{\beta}_j) &= [-\frac{1}{2}(\ \widehat{\beta}_j'\Sigma^{-1}\widehat{\beta}_j)^{\frac{1}{2}}]], \ j = 1, \cdots, n,
\end{aligned}$$

and Φ denotes the cumulative standard normal distribution. Here $\overline{X}_{1j}$ $(j = 1, \cdots, n_1)$ and $\overline{X}_{2j}$ $(j = n_1{+}1, \cdots, n)$ denote the sample means obtained by deleting the j-th observation $(j = 1, \cdots, n)$. 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

$$\widehat{\alpha}_{1j} = nf(\widehat{\beta}) - (n-1)f(\widehat{\beta}_j), \ j = 1, \cdots, n.$$

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

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

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

3.2. Bootstrap confidence interval

(1) Unconditional bootstrap D-method

Fisher (1936) suggested the D method which is the simplest method for estimating the conditional error rate, γ^* in (1.5). An estimator $\widehat{\gamma}^*$ of the conditional error rate γ^* in (1.5) based on the bootstrap sample is obtained by D method.

Now we review the bootstrap confidence interval. An estimator $\widehat{\alpha}^*$ of α based on the bootstrap sample is obtained, and B estimates, $\widehat{\alpha}_1^*, \ldots, \widehat{\alpha}_B^*$ are calculated from the bootstrap samples. Let the sample variance of the $\widehat{\alpha}_i^*$ be s_α^2. Then bootstrap confidence interval for

α can be obtained from the B values of $\widehat{\alpha}^*$. Let $\widehat{\alpha}^*_{(i)}$ denote the i-th ordered value, so that $\widehat{\alpha}^*_{(i)} \leq \widehat{\alpha}^*_{(2)} \leq \cdots \leq \widehat{\alpha}^*_{(B)}$.

Three types of 100(1-2η)% confidence interval are presented in Efron (1982, 1987), Buckland (1983,1984,1985), Hall (1986a, 1986b), Hinkley (1988), DiCiccio and Romano (1988) among others.

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

Bias-corrected percentile method. Suppose $\widehat{\alpha}^*_{(q)} \leq \widehat{\alpha} \leq \ldots \leq \widehat{\alpha}^*_{(q+1)}$, where $\widehat{\alpha}$ is calculated from the original samples. That is, q of the B bootstrap estimates for α are smaller than $\widehat{\alpha}$. Define $z_0{=}\Phi^{-1}(q/B)$, $\eta_{BL}{=}\Phi\,(2z_0{-}z_\eta)$ and , $\eta_{BR}{=}\Phi\,(2z_0{+}z_\eta)$ where $\Phi\,(z_\eta){=}\,1{-}\eta$ and Φ denotes the cumulative standard normal distribution. Then the confidence interval is given by $(\widehat{\alpha}^*_{(j)},\ \widehat{\alpha}^*_{(k)})$, where $j = (B+1)\eta_{BL}$ and where $k = (B+1)\eta_{BR}$.

Accelerated bias-corrected percentile method.

Define

$$\eta_{AL}{=}\Phi(z_0{+}\frac{z_0{-}z_\eta}{1-a\,(z_0{-}z_\eta)}\ ), \eta_{AR}{=}\Phi(z_0{+}\frac{z_0{+}z_\eta}{1-a\,(z_0{+}z_\eta)}\ ),$$

where

$$a = \frac{1}{6}\left[\sum_{i=1}^{B}(\widehat{\alpha}^*_i{-}\overline{\widehat{\alpha}}^*)^3 \Big/ \left[\sum_{i=1}^{B}(\widehat{\alpha}^*_i{-}\overline{\widehat{\alpha}}^*)^2\right]^{3/2}\right],$$

which is called the acceleration constant, and $\overline{\widehat{\alpha}^*}$ is the mean of the B bootstrap estimates for $\widehat{\alpha}^*_i, i = 1, \ldots, B$. Then the confidence interval is given by $(\widehat{\alpha}^*_{(u)},\ \widehat{\alpha}^*_{(v)})$, 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 γ^*, a Monte Carlo study is proposed. In this study, bivariate normal random deviates are generated from $\pi_1 : N(0,\ I)$, and $\pi_2 : N\left([\triangle_x,\ 0]',\ I\right)$ by using subroutine in the International Mathematical and Statistical Library (IMSL), where $\triangle_x^2$ is the Mahalanobis distance. For each Monte Carlo study, 1,000 iterations will be obtained. In each iteration, B=10,000 bootstrap samples are generated. Then the bootstrap confidence intervals for γ^* are obtained from the B values of $\widehat{\alpha}^*$ which is an estimator of γ^* based on the bootstrap sample by D method which is suggested by Fisher (1936). In order to construct the bootstrap confidence intervals for γ^*, 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 1,000 lower limits and the 1,000 upper limits respectively. The coverage probability is also considered from the 1,000 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 1,000. The bootstrap confidence intervals are compared with the conditional jackknife confidence interval given in Dorvlo (1992) based on the average length and coverage probability.

(2) Conditional bootstrap confidence interval

The conditional distribution of W given X$\in \pi_1$ is normal with mean,

$$\left[\mu^{(1)}-\frac{1}{2}\left(\overline{X}^{(1)}+\overline{X}^{(2)}\right)\right]' S^{-1}\left(\overline{X}^{(1)}-\overline{X}^{(2)}\right),$$

and variance $(\overline{X}^{(1)}-\overline{X}^{(2)})'\ S^{-1}\Sigma \overline{S}^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$. Since the conditional distribution of W involves unknown parameters, we may replace $\mu^{(1)}$ and Σ by $\overline{X}^{(1)}$ and S respectively. So the conditional distribution of W given $\overline{X}^{(1)}$, $\overline{X}^{(2)}$, S and X$\in \pi_1$ is approximated by $N\left(\widehat{\triangle}^2/2,\ \widehat{\triangle}^2\right)$ where $\widehat{\triangle}^2= (\overline{X}^{(1)}-\overline{X}^{(2)})'\ \overline{S}^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$. Similarly, the approximate conditional distribution of W given X$\in \pi_2$ is $N\left(-\widehat{\triangle}^2/2,\ \widehat{\triangle}^2\right)$.

We note that the usual bootstrap samples taken from the training samples are not appropriate because each bootstrap sample gives different mean and covariance matrix that changes the conditioning. Now, we obtain the conditional bootstrap samples of size n_1, denoted by $w_{1j}, j = 1, 2, \cdots, n_1$, from the conditional distribution, $N(\widehat{\triangle}^2/2,\ \widehat{\triangle}^2)$ when X comes from π_1. Let n_1^* be the number of negative w_{1j}. Similarly, we can obtain the conditional bootstrap samples of size n_2, denoted by $w_{2j}, j = 1, 2, \cdots, n_2$, from the distribution, $N(-\widehat{\triangle}^2/2,\ \widehat{\triangle}^2)$ when X comes from π_2. Then, we let n_2^* be the number of positive w_{2j}. We can construct the estimator $\widehat{e}$ of γ^* from the conditional bootstrap samples, which is

$$\widehat{e}=\frac{1}{2}\left(\frac{n_1^*}{n_1}+\frac{n_2^*}{n_2}\right), \tag{3.2}$$

This process is repeated independently a large number of B times. Then the conditional bootstrap confidence interval of γ^* in(1.5) can be obtained from the B values of $\widehat{e}$.
It is seen from Table 3.1 that the coverage probability for the conditional bootstrap confidence interval is close to the stated confidence level, but those of the unconditional bootstrap and jackknife confidence intervals are not. Hence we recommend to use the conditional bootstrap confidence intervals. Among the conditional bootstrap methods we recommend the bias-corrected percentile method to obtain the confidence interval for γ^* in (1.5).

Table 3.1 Comparison of confidence intervals for γ^*

%	p	n_1	$\overline{\gamma^*}$	Method†		Average Lower Limit	Averagee eUpper Limit	Average Length	Coverage Prob.
7	2	15	0.1721	UNC	P	0.0583	0.1863	0.1010	54.2
					B	0.1155	0.2198	0.1043	63.5
					A	0.1191	0.2199	0.1008	63.2
				CON	P	0.0850	0.2160	0.1310	70.1
					B	0.0914	0.2244	0.1330	71.8
					A	0.0902	0.2246	0.1344	72.3
					J	0.0703	0.2078	0.1375	65.1
95	2	15	0.1721	UNC	P	0.0483	0.2339	0.1855	81.5
					B	0.0735	0.2689	0.1954	88.5
					A	0.0818	0.2693	0.1875	88.6
				CON	P	0.0395	0.2848	0.2452	93.6
					B	0.0445	0.2937	0.2492	95.1
					A	0.0408	0.2948	0.2539	95.3
					J	0.0207	0.2727	0.2520	88.1

UNC: unconditional bootstrap method, CON: conditional bootstrap method, P: percentile method,
B: bias-corrected percentile method, A: accelerated bias-corrected percentile method,
J: conditional jackknife method

4. Conditional bootstrap confidence interval for missing values

We will extend the conditional bootstrap confidence interval for γ^* to the case that the training samples contain missing values. We will not consider the conditional jackknife confidence interval in this case since the conditional jackknife method does not improve the bootstrap method when training samples do not contain missing values. We will consider the bootstrap confidence interval for the conditional error rate in (2.6) using W_c. The conditional error rate can be estimated by substituting the estimates $\widehat{\Sigma}$, $\widehat{\mu}^{(i)}$ for Σ, $\mu^{(i)}$ in (2.4) and (2.5), respectively. Let $\widehat{\mu}^{(i)}=\left[\overline{Y}^{(i)}, \overline{Z}^{(i)}\right]'$ be the estimate of $\mu^{(i)}$ from (2.1) and (2.2). For the covariance matrices, let

$$\widehat{\Sigma}^{(i)}_{xc} = \begin{bmatrix} \widehat{\Sigma}^{(i)}_{yyc} & \widehat{\Sigma}^{(i)}_{yzc} \\ \widehat{\Sigma}^{(i)}_{zyc} & \widehat{\Sigma}^{(i)}_{zzc} \end{bmatrix}$$

be the estimate from the complete observations of sizes m_i. Also let $\widehat{\Sigma}^{(i)}_{yyi}$ 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,

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

for $\Sigma^{(i)}, i = 1, 2$.

Now the pooled estimate of the covariance matrices is given by

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

We will use these estimates in the construction of the bootstrap confidence intervals for the conditional error rate β^* in (2/6) when the training samples contain missing observations. Basically the same procedure described for γ^* 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 conditional error rate in (1.5).

Basically the same procedure described for γ^* is applied in this situation for getting the three types of the bootstrap confidence intervals for β^*. We generated bivariate normal random deviates from π_1: N(0, I) and π_2 : N($[\triangle_y, \triangle_z]'$, I) by using IMSL subroutines, where $\triangle^2_y$ and $\triangle^2_z$ are Mahalanobis distance based on the variable Y and the variable Z respectively. Note that

$$\triangle^2_x=\triangle^2_y + \triangle^2_z \text{ for } X =[X,Z]',\ R =\triangle^2_y / \triangle^2_x, \text{ where } 0 \leq R \leq 1.$$

For each Monte Carlo study, 1,000 iterations will be obtained. In each iteration, B=10,000 bootstrap samples are generated.

From the Table 4.1, we also recommend the conditional bootstrap confidence intervals since the unconditional bootstrap confidence interval cannot control the confidence level. Among the three conditional bootstrap methods we recommend the bias-corrected percentile method to obtain the confidence interval for β^* in (2.6).

Table 4.1 Comparison of confidence intervals for β^*

%	P	k	n	m	R	$\triangle_x^2$	$\overline{\beta^*}$	Method†		Average Lower Limit	Average Upper Limit	Average Length	Coverage Prob.
70	2	1	15	10	0.8	4	0.1574	UNC	P	0.0725	0.1761	0.1036	45.8
									B	0.1218	0.2256	0.1038	64.8
									A	0.1295	0.2256	0.0961	63.9
								CON	P	0.0911	0.2242	0.1332	68.9
									B	0.0972	0.2324	0.1352	70.3
									A	0.0968	0.2325	0.1357	70.2
95	2	1	15	10	0.8	4	0.1574	UNC	P	0.0346	0.2218	0.1872	72.3
									B	0.0764	0.2716	0.1953	89.9
									A	0.0829	0.2717	0.1888	89.7
								CON	P	0.0437	0.2930	0.2493	92.3
									B	0.0498	0.3021	0.2523	94.0
									A	0.0476	0.3024	0.2547	94.0

UNC: unconditional bootstrap method, CON: conditional bootstrap method, P: percentile method,
B: bias-corrected percentile method, A: accelerated bias-corrected percentile method, J: conditional jackknife method

5. Numerical example

Application of the bootstrap method to estimate the error rate, β^* in (2.5) 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 5.1. 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, TOEFL score is missing.

Using this data set, we obtain the discriminant function $W_c = cW_x+(1-c)W_y$, where

$$W_x = a^{'}X + b,\ a^{'} = [-1.9957 - 0.0170 - 0.00040.00340.0242],\ b = -2.5252,$$

$$W_y = d^{'}X + e,\ d^{'} = [0.5302 - 0.0042 - 0.00230.2406],\ e = 0.2846,\ c = 0.7532.$$

For this example, we generate 10,000 bootstrap samples to construct the conditional bootstrap confidence interval for β^*. The result of using $c^* = 0.7532$ is that the estimate of β^* in (2.6) is 0.4627. The 95% confidence interval of β^* is (0.3500, 0.5750) which is obtained by the bias-corrected percentile method.

Table 5.1 Success and failure group

Population 1 : Success					Population 2 : Failure				
X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
2.97	420	800	600	497	3.75	250	730	460	513
3.80	330	710	380	563	3.11	320	760	610	560
2.50	270	700	340	510	3.00	360	720	525	540
2.50	400	710	600	563	2.60	370	780	500	500
3.30	280	800	450	543	3.50	300	630	380	507
2.60	310	660	425	507	3.50	390	580	370	587
2.70	360	620	590	537	3.10	380	770	500	520
3.10	220	530	340	543	2.30	370	640	200	520
2.60	350	770	560	580	2.85	340	800	540	517
3.20	360	750	440	577	3.50	460	750	560	597
3.65	440	700	630		3.15	630	540	600	
3.56	640	520	610		2.93	350	690	620	
3.00	480	550	560		3.20	480	610	480	
3.18	550	630	630		2.76	630	410	530	
3.84	450	660	630		3.00	550	450	500	
3.18	410	410	340		3.28	510	690	730	
3.43	460	610	560		3.11	640	720	520	
3.52	580	580	610		3.42	440	580	620	
3.09	450	540	570		3.00	350	430	480	
3.70	420	630	660		2.67	480	700	670	

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

6. 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 conditional bootstrap confidence interval for the conditional error rate β^* in (2.6) 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 and Sons, New York.

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. *Journal of the Royal Statistical Society C*, 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. and Han, C. (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 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 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 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, New Jersey.

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.